technische universität
dortmund

SFB 876

Technical report for
Collaborative Research Center
SFB 876

Providing Information by Resource-Constrained Data Analysis

September 2012

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

SFB 876 Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

Ensuring k-anonymity in Smartphone Utilization and Mobility Data

Orwa Nassour
Fakultät für Informatik, Lehrstuhl 12
Technische Universität Dortmund
orwa.nassour@tu-dortmund.de

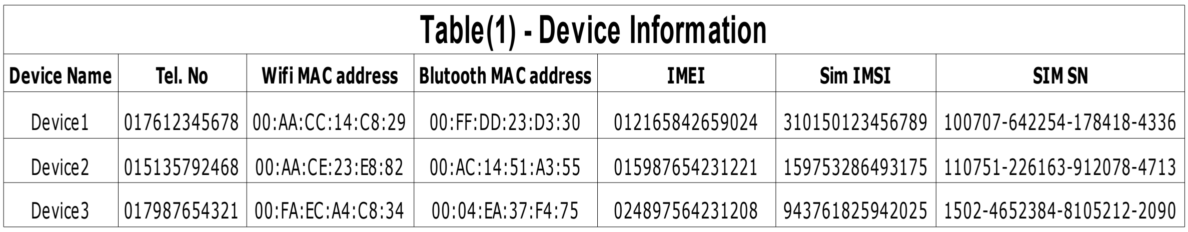
Protecting privacy in datasets that include personal information has become a crucial issue in the field of data mining. The identity of any individual should not be definitively recognized. Therefore, it should be anonymized before the containing dataset will be published. Anonymization is the process of converting private data into non-identifiable data. Therefore it is used to protect participants privacy from the public. Simple anonymization techniques such as suppression or generalization can be used to anonymize the personal data. However, applying these techniques just on the sensitive data cannot totally hide the user identity since other information which are considered as non-sensitive could be linked with external data sources, thus re-identifying the user identity. Therefore, to prevent such privacy attacks we need to identify the sensitive data which may be under attack, and the non-sensitive data which can be used to re-identify the users, then anonymize all these data properly before it is released.

1 Introduction

In fact, anonymizing personal data cannot totally protect the user since other information which are considered as non-sensitive could be linked with external data sources, thus re-identifying him. Suppression can anonymize data by replacing the sensitive attributes values with a meaningless character like stars (*). Whereas generalization does it by replacing the attribute values with less informative ones. The dataset may identify attributes that also appear in external data sources. These attributes are candidates for linking by attackers. The subset of these attributes are called "Quasi-Identifiers", and it is essentially the combinations of these quasi-identifiers that must be protected [3]. "A set of personal records is said to be k-anonymous if every record is indistinguishable from at least k-1 other records over given quasi-identifiers subset of attributes" [3]. This work focuses on implementing a k-anonymity protection model on a data collection tool called MobiDAC in order to guarantee the privacy of the users' data against external attacks. MobiDAC is developed by our department to be run on android mobile devices. Gathering data from the these devices is done using python scripts. The task of MobiDAC is to manage the lifecycle of these scripts and to transport the collected data to the operator. The collected data are used for various purposes, such as analysing and improving the power consumption and response time for these devices [2]. In this report, I will suppose that the data are collected in one database located on one machine. i.e. the operator. Since it is learned from experience that k-anonymity is difficult to inforce before all data is collected in one trusted place [1].

2 Analysing the attributes

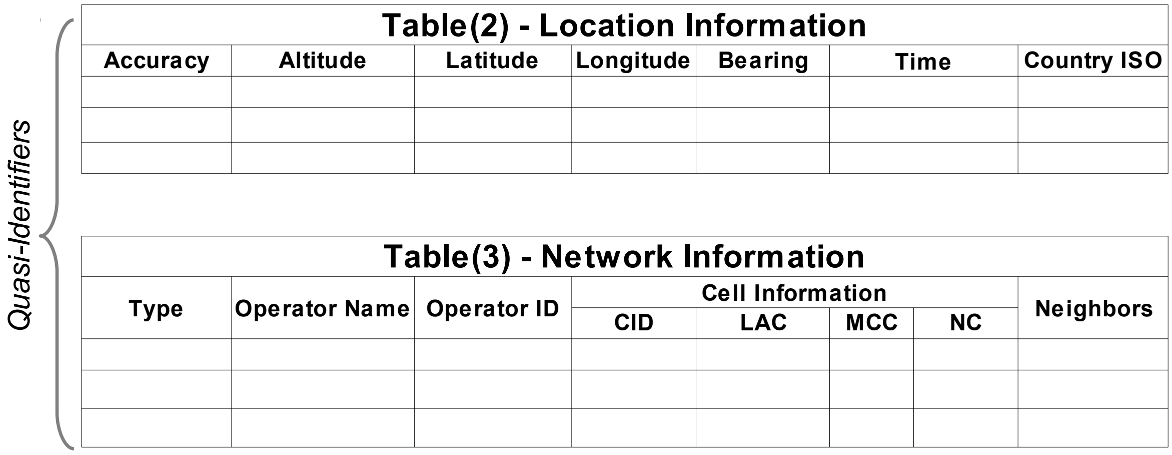
Before anonymizing the attributes of the collected data, it should be decided which attributes should be completely and which should be partially suppressed or generalized. In other words, what are the sensitive attributes and the quasi-identifiers of our dataset? To answer this question, the type of the collected data should be checked. Table(1) shows the sensitive attributes collected by MobiDAC. Any single attribute can exclusively identify the user from others. These attributes should be completely suppressed before publishing the dataset in order to hide the users' identities.


Table(1) - Device Information

Device Name	Tel. No	Wifi MAC address	Blutooth MAC address	IMEI	Sim IMSI	SIM SN
Device1	017612345678	00:AA:CC:14:C8:29	00:FF:DD:23:D3:30	012165842659024	310150123456789	100707-642254-178418-4336
Device2	015135792468	00:AA:CE:23:E8:82	00:AC:14:51:A3:55	015987654231221	159753286493175	110751-226163-912078-4713
Device3	017987654321	00:FA:EC:A4:C8:34	00:04:EA:37:F4:75	024897564231208	943761825942025	1502-4652384-8105212-2090

Other information such as location and network information can be linked by attackers to external datasets to re-identify the user. These attributes can be considered as Quasi-Identifiers for MobiDAC, and shown in tables (2) and (3). Other attributes can be added to this set such as Device type (GSM, UMTS, LTE..), MAC-addresses and SSIDs of the connected Access Points, applications and process names, and the discovered blutooth devices.

These information should be suppressed or generalized to prevent re-identifying the users. After choosing the sensitive attributes and the quasi-identifiers, we should anonymize



Quasi-Identifiers

Table(2) - Location Information

Accuracy	Altitude	Latitude	Longitude	Bearing	Time	Country ISO

Table(3) - Network Information

Type	Operator Name	Operator ID	Cell Information				Neighbors
			CID	LAC	MCC	NC	

them based on the degree of anonymization required, which is k. The greater the k, the better the protection. Suppression depends on the dataset size and the distribution of the collected data, i.e. the distance between the similar attributes, as well as the possible background knowledge that the attacker may use to link with the quasi-identifiers.

3 An algorithm for provideing k-anonymous protection

After analysing the attributes, an algorithm can be applied on the quasi-identifiers in order to anonymize its values based on the required protection level k. I wrote this algorithm as a first draft to the desired k-anonymous model.

For each attribute (column) in the dataset:

1. Sort the dataset based on the selected attribute
2. Take the first value
3. Compare this value with other values under the same attribute
4. If the number of matchings >= k-1:
    a) While this value is not the last one and the next value is matching it: Go to the next one
    b) If this value is the last one: Exit
    c) Else: Go to 3
5. Else:
    a) Calculate the distance between this value and other values under the same attribute (the number of different digits)

b) Take the smallest distance (biggest match), called n

c) Suppress the n different digits of this value and all values with distance <= n

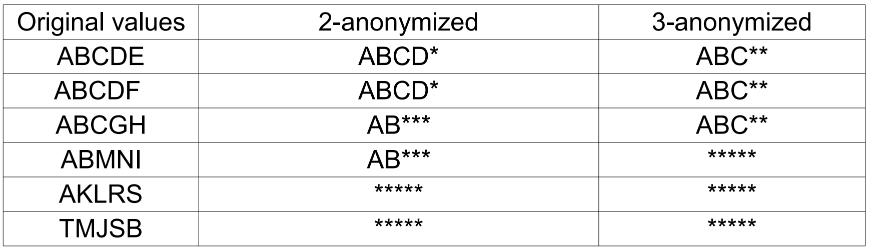
d) If the number of matchings >= k-1: Go to (4.1)

e) Else:

  i. n = n+1

  ii. go to (5.3)

As an example how this algorithm works, lets take an attribute with the following values: ABCDE ABCDF ABCGH ABMNI AKLRS TMJSB. The distances between the first and the other values are: 1, 2, 3, 4, 5 respectively. with k=2 and k=3, the output regarding the algorithm would be:



Original values	2-anonymized	3-anonymized
ABCDE	ABCD*	ABC**
ABCDF	ABCD*	ABC**
ABCGH	AB***	ABC**
ABMNI	AB***	*****
AKLRS	*****	*****
TMJSB	*****	*****

For the near future I am planning to improve this algorithm, and develop an application which can apply it on the collected data from MobiDAC in order to generate a k-anonymous dataset, that can be published safely.

References

[1] Tyrone Grandison and Alexandre Evfimievski. Pivacy preserving data mining. IBM Almaden Research Center.

[2] MobiDAC website.
`http://www.eclipse.org/modeling`.

[3] School of Computer Science, Carnegie Mellon University, Pittsburg, Pennsylvania. *A model for proecting privacy*, 10(5), 2002; 557-570, May 2002.

Spatio-Temporal Probabilistic Models for Sensor Networks

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

Streams of sensor measurements arise from twitter, mobile phone networks, internet traffic, road traffic, home automation systems, seismic motion and sea level - to mention just a few. The exploration and modelling of such measurements from multiple sensors induces the need for algorithms that are capable of processing the data as it becomes available and that can quickly provide partial results based on the data seen so far. Beside these requirements, the algorithm should obey resource constraints and capture the inherent spatio-temporal structure within sensor data. For this reason, we present Spatio-Temporal Markov Random Fields (ST-MRF) to model the dynamics of sensor networks and perform a predictive analysis on arbitrary subsets of sensors. ST-MRF tracks the empirical distribution of each sensor and concurrently updates a Maximum Likelihood estimate of the underlying distribution. We applied our method to model mobile phone network cells, freeway traffic of a German Autobahn network, sea levels of the pacific ocean and temperatures in an office building.

In order to model pairwise interactions of nearby sensors with respect to space and time, we developed the framework of Spatio-Temporal Markov Random Fields (ST-MRF) which is now described briefly followed by exemplary applications and an outlook. For a more detailed discussion, see [6].

Spatio-Temporal Markov Random Fields. ST-MRF do enhance the general framework of MRF by considering a sequence of graphs $G_1, G_2, \dots, G_T$ where each graph $G_t = (V_t, E_t)$ is called *layer*. Each layer serves as an undirected spatial dependency structure of all sensors at time t and each vertex from the set V_t corresponds to a sensor measurement at time t. Each edge $\{v,u\} \in E_t \subset V_t \times V$ with $V := \bigcup_{t=1}^{T} V_t$ encodes conditional independence assumptions among sensor measurements, i.e. a sensors value at time t is fully determined by its spatio-temporal neighborhood, which may contain values from several different $V_{t'}$. In the basic setting of ST-MRF, the spatial structure is stationary over time, that is $(v_t, u_{t'}) \in E_t \Leftrightarrow (v_{t+1}, u_{t'+1}) \in E_{t+1}$ for all $1 \leq t \leq T$ with $v_{T+t} := v_t$. For this kind of model, the projection from real time, e.g. 2:00am, to time index t usually depends on the period that should be modeled and the physical sampling rate of the sensors. As an example, consider a building with n rooms, each equipped with a temperature sensor that measures the temperature every minute. This results in 1440 sensor readings per day per sensor, which means $T = 1440$. It is also possible (and common) to build models with a temporal resolution that is lower than the physical sampling rate of the sensors. If we choose $T = 144$ in the above, measurements within consecutive intervals of 10 minutes have to be aggregated (min, max, mode, average, ..) or simply ignored. The particular projection is data and task dependent. Although we want to learn a model from a stream of sensor measurements, assume for now that the data is given in a set $\mathcal{T} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where each $x^{(i)}$ is the i-th joint reading of all sensors in the network over a period of length T. We denote the measurement of a single sensor s at time t as x_{s_t}. Without loss of generality, we assume that sensor measurements are discretized, i.e. each sensor measures a discrete state from finite set $\mathcal{X}$. The model parameters $\theta \in \mathbb{R}^d$ consist of one weight vector $\theta_{s_t} \in \mathbb{R}^{|\mathcal{X}|}$ per spatio-temporal vertex and one weight matrix $\theta_{\{s_t, v_{t'}\}} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ per spatio-temporal edge. This results in a total of $d = |V| \cdot |\mathcal{X}| + |E| \cdot |\mathcal{X}|^2$ parameters, with $E := \bigcup_{t=1}^{T} E_t$. Instead of using discretized measurements, we could assume that measurements have a Gaussian distribution. In this case, θ_{s_t} would store the mean and $\theta_{\{s_t, v_{t'}\}}$ the partial correlation coefficients. In either case, the parameters can be obtained by Maximum Likelihood Estimation, whereby the Likelihood of a particular θ given a data set $\mathcal{T}$ is defined as

$$\mathcal{L}(\theta; \mathcal{T}) := \prod_{i=1}^{N} p_\theta(X = x^{(i)}). \tag{1}$$

Here, $p_\theta(X = x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$ is the density of an exponential family member [7] which obeys the conditional independence structure given by G, $\phi(x)$ is a function (sufficient statistic) that maps x into a d-dimensional binary vector space and $A(\theta)$ normalizes the density. The components of vector ϕ represents a joint measurement of all sensors and all edges in the network as binary values. For sensors $v, w \in V$ and states $a, b \in \mathcal{X}$ the entries are defined by

$$\phi_{v,a}(X) := \begin{cases} 1 & \text{if } X_v = a \\ 0 & \text{otherwise,} \end{cases} \quad \phi_{vu,ab}(X) := \begin{cases} 1 & \text{if } X_v = a \text{ and } X_u = b \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

If the model is implemented directly into a sensor network, we do not want it to store its complete history of measurements. Therefore, we take the logarithm of the Likelihood (1) and rearrange, such that it only depends on the average value or the *empirical expectation* $\tilde{\mathbb{E}}[\phi(x)]$ of our sufficient statistic. Note that the extra $\frac{1}{N}$-factor does not change the optimal solution.

$$\ell(\theta; \mathcal{T}) := \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x^{(i)}) = \langle \theta, \tilde{\mathbb{E}}[\phi(x)] \rangle - A(\theta) \tag{3}$$

Note also that this fits to a streaming scenario, since we only have to count how often a certain combination appears in the stream. Taking derivatives of (3) it follows that

$$\frac{\partial \ell(\theta; \mathcal{T})}{\partial \theta_{v,a}} = \tilde{\mathbb{E}}\left[\phi_{v,a}(x)\right] - \hat{\mathbb{E}}\left[\phi_{v,a}(x)\right],$$

whereby $\theta_{v,a}$ denotes the weight of sensor v and state $a \in \mathcal{X}$. The *estimated expectation* $\hat{\mathbb{E}}\left[\phi_{v,a}(x)\right]$ is computed by Belief Propagation (BP) [3], which is also known as Sum-Product algorithm in the context of factor graphs. The objective $\max_\theta \ell(\theta; \mathcal{T})$ can be solved by any first-order optimization method. The prediction, by means of per-node maximum marginal probabilities or maximum a posteriori assignments, is also computed with BP. Note that graphs of ST-MRF models contain many loops per definition and that BP is an approximate method in this case.

The following experiments should give a rough impression of which kinds of predictive analysis can be done with ST-MRF, they serve as simple examples on how spatial-temporal data can be analyzed. Additional experiments on data that is collected by sensors in the Autobahn of North Rhine-Westphalia and data from buoys that measure the sea levels in the pacific ocean can be found in [6].

Next network cell prediction. For the first task, we applied ST-MRF to a network cell prediction task. The data consists of trajectories in terms of mobile network cell identifiers for several users, whereas the provided trajectories contained many missing values for some users. In this case, a sensor is identified with a user and a measurement is identified with the network cell identifier. The task was to model a users behavior over 24h. Due to missing values, we set the sampling rate to 10 minutes, which results in $T = \frac{24h}{10min} = 144$. We choose a temporal first order Markov dependency without any spatial dependency, which results in a non-stationary Markov chain per user. In the experiments, we reached $> 70\%$ prediction accuracy for certain users. See [2] for a detailed description of our experiments on mobile phone user data.

Modelling temperature sensor network. The second task is based on data collected in March 2004 from temperature sensors deployed in the Intel Berkeley Research lab[1] A nearest neighbor graph was constructed as basic spatial dependency structure, whereas an

[1] The data set is available online at `http://db.csail.mit.edu/labdata/labdata.html`.

edge is considered disconnected whenever it is blocked by walls. We also excluded sensors reported faulty in [1]. The final spatial graph contained 48 nodes and 150 edges. We reached a normalized absolute error of $< 15\%$ in the task of predicting a (discretized) room temperature for given points in time. See [4] for a detailed description of our experiments on temperature sensor network measurements.

All experiments are done with our own implementation of ST-MRF[2] called `iST-MRF` [4]. Current research is about reducing the resource consumption of graphical models in general. Results on reducing the memory complexity of ST-MRF are already submitted [5]. We also try to reduce the energy consumption of smart phones by means of probabilistic user models. We hence collect usage data from multiple smart phone users, which are currently analyzed with respect to their individual and common resource consumption.

References

[1] Daniele Apiletti, Elena Baralis, and Tania Cerquitelli. Energy-saving models for wireless sensor networks. In *Knowledge and Information Systems*, volume 28, pages 615–644. Springer London, 2011.

[2] Stefan Michaelis, Nico Piatkowski, and Katharina Morik. Predicting next network cell IDs for moving users with Discriminative and Generative Models. In *Mobile Data Challenge by Nokia Workshop in conjunction with International Conference on Pervasive Computing*, June 2012.

[3] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[4] Nico Piatkowski. iST-MRF: Interactive Spatio-Temporal Probabilistic Models for Sensor Networks. In Jilles Vreeken, Nikolaj Tatti, Bart Goethals, Anton Dries, Matthijs van Leeuwen, and Siegfried Nijssen, editors, *Proceedings of the ECML PKDD 2012 Workshop on Instant Interactive Data Mining*, September 2012.

[5] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Probabilistic Graphical Models with Compressed Temporal Dynamics. In *Submitted to the International Conference on Data Mining (ICDM)*. IEEE, Submitted in June 2012. (Still under review).

[6] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Spatio-Temporal Models For Sustainability. In Manish Marwah, Naren Ramakrishnan, Mario Berges, and Zico Kolter, editors, *Proceedings of the SustKDD Workshop within ACM Conference on Knowledge Discovery and Data Mining (SIGKDD) 2012*. ACM, August 2012.

[7] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2007.

[2] The software is available online at `http://sfb876.tu-dortmund.de/ist-mrf`.

MobiDAC: A Flexible Data Collection Infrastructure for Android Devices

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

Today's mobile devices are equipped with various external sensing hardware. Furthermore they are subject to complex interaction patterns and internal state changes. Much research is centered around the data that can be collected from these sources. The software responsible for data collection is usually developed from scratch. This report summarizes the current state of MobiDAC, a flexible data collection infrastructure for Android-based devices that allows rapid development of data collection campaigns via scripting languages, while taking care of their deployment and the transmission of generated data.

1 Introduction

Today's mobile phones are able to collect data about their environment, by determining their position, sensing the presence of other hardware (other phones, Bluetooth devices or network cell towers) or sensing the physical state via sensors for light, magnetism, acceleration, or even air pressure and temperature. Besides the external sensing possibilities, also the software running on these devices serves as a valuable data source regarding the usage patterns for the phone. Especially the system software is well-suited for this purpose, since the available data sources can be used independently of the concrete set of installed applications.

Much research has been conducted to use smartphone sensor and utilization data for specific purposes, like learning about the dynamics of social networks [1], distributed traffic monitoring, or calculating personal carbon footprint [3], just to name a few examples.

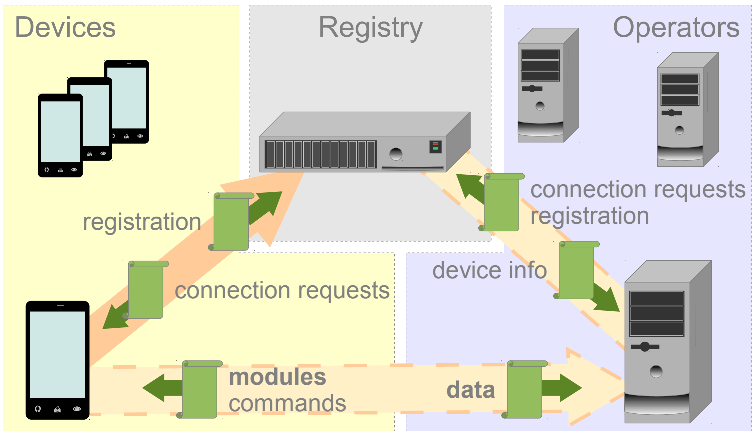


Figure 1: The MobiDAC infrastructure.

The data collection, however, is usually implemented from scratch for the respective purpose.

The Cambridge DeviceAnalyzer [4] covers a broad range of available data sources in order to generate a less specific but rich data set that is expected to reveal insights on utilization patterns. Over 10k participants have collected or are still collecting data.

A notable exception is *Campaignr* [2], a Software for *participatory sensing* that allows to configure the data collection by sending an XML document to the devices.

2 MobiDAC

The idea of MobiDAC is to provide an infrastructure for configurable and purpose-specific data collection campaigns comparable to Campaignr, however without being constrained to physical sensing only. Experimenters (called *operators)* write *sensing modules* that determine which data sources are used and how the data is preprocessed or aggregated. Sensing modules might range from simple descriptions as used by Campaignr to programs written in Turing-complete languages.

An operator can upload *sensing modules* to multiple devices and can remotely start or stop them. When a module is running, it collects, possibly preprocesses and saves data locally on the device. Data is transmitted, when the device is plugged into a charger or on explicit request from operators, but only when connected via WiFi.

Infrastructure MobiDAC is designed to allow devices and operators[1] to arbitrarily change their Internet connection status and address. This is especially important for

[1] For better readability, we do not distinguish between operators and operators' machines.

the unreliable connections of mobile devices. Therefore, operators and devices register themselves to the *Registry*, another part of the infrastructure that keeps track of registered operators and devices and mediates communication between them.

Mobile phones generally do not have public IP addresses and are rather subject to network address translation. Since, they are not reachable from the outside, every connection from operators to a device has to be initiated by the device itself (light orange arrow in Figure 1). However, the operators and their addresses are not necessarily known to the devices in advance. To allow operators to establish connections to devices, the registry implements a *push service*: An operator wishing to connect to a device does so by issuing an according connection request to the Registry. The request is forwarded to the respective device via a permanently maintained connection from every device to the registry (dark orange arrow).

A device wishing to connect to an operator simply asks the registry for the operator's availability and necessary information for a temporary connection.

Data Collection Modules Currently, MobiDAC uses the Scripting Layer for Android [2] to execute data collection modules. Thus, they can be written in any of the supported scripting languages. For these, the scripting layer provides a simplified copy of the Android API that allows to use services of the *Application Framework* (e.g., positioning, telephony, or speech recognition). The API calls on the script side are translated into remote procedure calls (RPCs) that are transmitted to SL4A's language-independent core via local TCP/IP. The core consists of an Android service that invokes the Android API calls that correspond to the received RPCs. If the called API method has a return value, it is transported back to the script via the same mechanism.

Apart from the Android API, scripts can use the Linux kernel's virtual file systems (`/proc` and `/sys`), which also provide useful information (e.g., about running processes).

Security and Privacy An operator can upload and start instrumentation modules only if the respective participant (device owner) explicitly granted access. MobiDAC uses TLS for authenticated and encrypted connections between devices and operators. This prevents unauthorized access to devices and *man-in-the-middle* attacks targeted to access the data that is sent from device to operator.

In order to keep the Registry simple and lean, communication with it is not secured.

The participant may pause the data collection at any time. Furthermore, locations may be automatically filtered according participant-specified criteria.

[2] `http://code.google.com/p/android-scripting/`

3 Future Work

Besides being used for actual data collection, MobiDAC will serve as an experimentation platform for novel data collection approaches. The current data collection interface (Android and VFS) has two major drawbacks: 1.) Not all potentially interesting data is accessible this way, e.g., context switch times of processes, 2.) some parts of the system state have to be checked regularly for changes, because there is no according notification mechanism, and 3.) capturing OS kernel or hardware state (e.g., sensors) at Android API level means, that we unnecessarily cross additional software layers, which needs additional CPU cycles and thus more energy. This is also true for SL4A. which adds another layer, but allowed for rapid and comfortable development in the first iteration. MobiDAC will be extended by support for SystemTap[3], which allows dynamic instrumentation at OS kernel level.

Further, we plan to enhance participants' direct control over privacy and energy consumption, by adding configurable restrictions to data sources as well as an energy quota that is enforced for data collection campaigns. Default settings that confine data collection and energy consumption into the limits of DeviceAnalyzer could encourage the use of MobiDAC also outside the SFB.

References

[1] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[2] A. Joki, D. Estrin, and J.A. Burke. Campaignr: a participatory sensing software architecture for cellphones. `http://escholarship.org/uc/item/8v01m8wj.pdf`, 2007.

[3] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 55–68. ACM, 2009.

[4] D.T. Wagner, A. Rice, and A.R. Beresford. Device analyzer. `http://www.cl.cam.ac.uk/~acr31/pubs/wagner-daabstract.pdf`.

[3] `http://sourceware.org/systemtap/`

Subproject A2
Algorithmic aspects of learning methods in embedded systems

Christian Sohler Jan Vahrenhold

Distributed MEB

Alexander Munteanu, Alexander Skopalik
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
{alexander.munteanu, alexander.skopalik}@tu-dortmund.de

We investigate the *minimum enclosing ball (MEB)* problem in a distributed setting and emphasize on communication complexity of approximating MEB. We show that there exists a distributed algorithm that approximates MEB within a factor of $\sqrt{2} \approx 1.414$ in which every node communicates not more than a d-dimensional point and a radius.

Introduction

We consider the *minimum enclosing ball (MEB)* problem. Given a set P of points in $\mathbb{R}^d$, the task is to compute a ball of minimum radius containing the points in R. In the literature it is also known under various names such as *smallest enclosing ball, minimum covering sphere,* 1*-center, minimum bounding sphere* problem.

The MEB problem is a well studied problem with many areas of applications. For example in computer graphics it is used for collision detection and visibility culling, in machine learning for support vector clustering or similarity search, and in optimization for planning the optimal position of a base stations in the facility location problem.

Nimrod Megiddo [5, 6] gave a linear time algorithm for solving linear programs of fixed dimensions and extended it to the MEB problem. However, the running time grows exponentially with the dimension which makes the algorithm impractical for solving problems in high dimensional space. Welzl [7] presented an elegant randomized algorithm that runs in linear expected time and improves upon the dependence on the dimension albeit it remains exponential. In practice it is limited to $d \approx 30$ dimensions.

Approximation algorithms and heuristics have been proposed to overcome this limitation. Badoiu and Clarkson [2] present a coreset based $(1 + \epsilon)$-approximation in time

$O(nd/\epsilon + 1/\epsilon^5)$. Agarwal and Sharathkumar [1] developed a one-pass streaming algorithm based on the coreset construction of Badoiu and Clarkson with linear space in d (and polylogarithmic in $1/\epsilon$) which yields an approximation factor of 1.3661. Chan and Pathak [3] improved this ratio to less than 1.22 by more involved analysis, which pushed the upper bound a step closer towards the lower bound of Agarwal and Sharathkumar [1] of 1.207.

All the algorithmic approaches mentioned above have in common, that the involved computations are performed sequentially. In a distributed setting, these sequential algorithms can be simulated in the sense that every node simulates the algorithm on his subset of the input and sends its internal state to the next node, which can resume the computations on the next subset of the data. The simulation proceeds in this distributed but still sequential manner until all nodes are done. Such an adaption of sequential optimization algorithms to distributed scenarios has been proposed before in [4]. Our approach in contrast, is capable of real parallel computation in the distributed setting as described more formally below.

Our Contribution

Consider the scenario of k nodes and an additional dedicated master node. Let $P = \bigcup_{i=1}^{k} P_i \subset \mathbb{R}^d$ be the input to MEB, where each P_i is available on exactly one of the k nodes. Every node is allowed to send data to every other node where the cost of sending a message is quantified by the number of bits which are actually sent. Our goal is to approximate the MEB of P while minimizing the communication cost.

Our first algorithm can be shown to be a $\sqrt{2}$-approximation for MEB. First, every node i computes the MEB B_i of his point set P_i and communicates its center and radius to the master node. The master then computes the MEB B of all balls B_i which is the output of the algorithm.

Algorithm 1 Distributed-MEB($P_1, P_2, \ldots, P_k$)

```
$\forall i \in [k]$  // every node in parallel
    send $B_i(c_i, r_i) = MEB(P_i)$ to master node
return  $B(c, r) = MBB(B_1, \ldots, B_k)$  //master computes minimum ball of balls
```

Algorithm 1 needs to communicate exactly one d-dimensional point and a radius per node and therefore has a cost of communicating $O(kd)$ floating point numbers. Moreover if $B_{opt}(c_{opt}, r_{opt})$ is the MEB of P, it holds that $r \leq \sqrt{2} \cdot r_{opt} \approx 1.414 \cdot r_{opt}$.

Future Work

Preliminary results indicate that the communication complexity of our approximation algorithm is optimal or at least tight up to logarithmic factors. More precisely, we conjecture that there is no constant factor approximation of MEB in the distributed setting if the nodes are restricted to communicate strictly less than the bit-representation of a ball center, i.e. a d-dimensional point.

A natural further question is how additional communication can be used to increase the quality of solutions. Our plan here is to design an algorithm which iteratively improves the quality of the computed MEB. Such a process can be used to quantify the amount of communication needed to achieve a given approximation ratio and reveal the trade-off behavior between cost and quality for distributed MEB.

With respect to energy consumption models, it would be interesting to consider scenarios in which communication going from nodes to a master is expensive whereas the master can send arbitrary amounts of data for free or very cheaply to all nodes. We also intend to consider different types of network topologies and their impact to communication complexity.

References

[1] Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. In *SODA*, pages 1481–1489. SIAM, 2010.

[2] Badoiu and Clarkson. Smaller core-sets for balls. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[3] Timothy M. Chan and Vinayak Pathak. Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. In Frank Dehne, John Iacono, and Jörg-Rüdiger Sack, editors, *WADS*, volume 6844 of *Lecture Notes in Computer Science*, pages 195–206. Springer, 2011.

[4] Hal Daumé III, Jeff M. Phillips, Avishek Saha, and Suresh Venkatasubramanian. Efficient protocols for distributed classification and optimization. *CoRR*, abs/1204.3523, 2012.

[5] Megiddo. Linear-time algorithms for linear programming in $\mathbb{R}^3$ and related problems. *SICOMP: SIAM Journal on Computing*, 12, 1983.

[6] N. Megiddo. Linear programming in linear time when the dimension is fixed. *Journal of Association for Computing Machinery*, 31(1):114–127, January 1984.

[7] Emo Welzl. Smallest enclosing disks (balls and ellipsoids). In *Results and New Trends in Computer Science*, pages 359–370. Springer-Verlag, 1991.

Online Admission Control for Cloud Computing: The Impact of Integrality

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

Background

In the Infrastructure-as-a-Service (IaaS) service model of Cloud Computing, users can request virtual processors from a service provider producing a typical online admission control problem. The system maps the virtual processors of accepted request on real processors or cores such that several virtual processors can share the same hardware using multitasking and context switching. As a specified processing power is typically guaranteed to the user, the provider cannot accept a request if this acceptance produces a violation of one or more service guarantees. The requests are typically subject to quantization of start times and request sizes. In addition, there are often minimum and maximum request sizes. To allow a high degree of flexibility, several types of services levels with different restrictions may be offered at different costs. The provider is generally interested in maximizing his return, see Schwiegelshohn and Tchernykh [14]. This kind of problem belongs to the class of online admission control problem that have been addressed frequently in the 90s for network infrastructures, see Awerbuch et al. [1] [2], Bar-Noy et al. [3], Garay et al. [7] [8] and Plotkin [13].

In this paper, we suggest a simple basic model from real time scheduling for this problem if only a single service level is offered. We allow preemption of the execution of requests to model the sharing of hardware resources. The service guarantee in such an environment is expressed with the help a stretch factor $f > 1$ and deadlines, see, for instance, DasGupta and Palis [6]: A request of size p with release date r is given the deadline $d = t + f \cdot p$. The restrictions on the request packages are represented by using integer request sizes ranging from 1 to a maximum value $p_{\max}$. The hardware infrastructure of the provider typically consists of large homogenous clusters suggesting a parallel identical machine (P_m) environment. For a single service level, the return to the provider can be represented by system utilization. Note that provider strives to balance user requests and its hardware infrastructure such that only few requests must be rejected and similarly only few processors are idle on average. Therefore, we can expect schedules with occasional idle times.

The provider can only influence the utilization of his system by a suitable selection of the service guarantee, the restrictions on the request packets, and appropriate acceptance and allocation algorithms. Bounds on the achievable utilization in dependence of selection parameters may help to support the decision process. To determine such bounds, we use competitive analysis, see, for instance, Borodin and El-Yaniv [5].

First we address a conventional online problem on a single machine and give upper bounds for all possible values of f and $p_{\max}$ taking into account the integer restriction on request packets sizes. We show that a simple greedy acceptance approach matches these bounds as it does in the unconstrained case. Then we relax the conventional online restriction by allowing to first collect all requests with the same release date before making a decision. This so called *integral online problem* allows more flexibility leading to an improvement of the upper bounds. Although greedy acceptance comes again close to match these bounds there is still a gap that cannot be overcome using this approach. This result is in contrast to the corresponding conventional problem. Moreover, integrality of the restrictions produces discontinuous competitive factors that depend on the stretch factor and on different processing times. Finally, we transfer the single machine results to the parallel identical machine environment and show that we may achieve better bounds by using suitable acceptance algorithms.

Only one proof is given in the main part of the paper while all other proofs can be found in the appendix.

Related Work

Baruah and Haritsa [4] discuss online scheduling of sequential independent jobs with stretch factors and preemptions on real time systems. In their paper, they present the ROBUST (Resistance to Overload By Using Slack Time) algorithm that guarantees a minimum slack factor for every task. The slack factor f of a task is defined as the ratio of the difference between deadline and the submission time over the execution time requirement. It is a quantitative indicator of the tightness of the task deadline. Therefore, it matches our definition. ROBUST guarantees an effective processor utilization (EPU) of $\frac{f-1}{f}$ during the overload interval. Das Gupta and Palis [6] base their hard real-time scenario on the concept of Baruah and Haritsa. They address the same problem as we do but they neither consider processing time limitations nor integrality. On a related note, some theoretical scheduling papers address online scheduling and the rejection of jobs, while incorporating some form of processing time restrictions. A close match especially in term of processing time constrictions to our problems can also be found in the work of Lipton and Tomkins [12], Goldwasser [10], Goldman et al. [9], and Lee [11], though unlike our models, these papers prohibit preemption. Goldman et al. generalized Lipton and Tomkins' model and gave analysis for setting with multiple different slack factors scenarios and jobs with either length 1, $\{1, k\}$ or powers of 2. Goldwasser distinguished between the case of exactly two distinct processing times or arbitrary processing times and gave tight competitive factors of $\frac{1}{1+\max\left(\frac{\lceil f-1\rceil+1}{\lceil f-1\rceil},\frac{\lfloor f-1\rfloor+1}{f-1}\right)}$ and $\frac{1}{2+\frac{1}{f-1}}$, respectively. Lee further studied the case of small stretch factors $f \leq 2$ and gave a randomized algorithm with a competitive factor of $\frac{1}{O\left(\lceil\log\frac{1}{f-1}\rceil\right)}$.

Our Contribution

We extend the scenario studied by DasGupta and Palis [6] by additionally considering integral release dates and processing times. In order for these limitations to have impact, all our upper and lower bounds are subjected to the maximum possible processing time of a job. To our best knowledge, such limitations have not been previously studied in any online scheduling problem. While greedy allocation guarantees tight competitive factors in both the unconstrained and the constrained case, processing time limitations are also responsible for some significant changes:

1. Improved competitive factors including special results for small values of the maximum processing time
2. Discontinuity and even partially lack of monotony of the upper bounds of the competitive factors as function of the stretch factor
3. Necessity to consider several processing times when determining the competitive factor for a given stretch factor and maximum processing time.

Additionally, we introduce an integral online model where all jobs with the same (integral) release date are scheduled together. This model produces improved competitive factors and more variety in a single machine environment. But contrary to the conventional online model, greedy allocation cannot guarantee tight competitive factors anymore. Finally, we exemplarily consider the parallel machine environment for both online versions and a small maximum processing time ($p_{\max} = 2$). We transfer a randomization approach and present algorithms that are not greedy and guarantee competitive factors with either no or a small gap. Although we discuss worst case results in this paper, they may also be useful in practice as they provide relations for different selections of stretch factors and processing time limitations. Note that in IaaS, the provider is free to select these parameters and therefore requires some information regarding the consequences of his choice. Therefore, the results of this paper may serve as a starting point for further studies on management configurations in an IaaS environment. Details can be found in the paper.

References

[1] B. Awerbuch, Y. Azar, and S.A. Plotkin. Throughput-competitive on-line routing. In *FOCS*, pages 32–40, 1993.

[2] B. Awerbuch, Y. Bartal, A. Fiat, and A. Rosén. Competitive non-preemptive call control. In *SODA*, pages 312–320, Philadelphia, PA, USA, 1994. Society for Industrial and Applied Mathematics.

[3] A. Bar-Noy, R. Canetti, S. Kutten, Y. Mansour, and B. Schieber. Bandwidth allocation with preemption. In *STOC*, pages 616–625, New York, USA, 1995. ACM.

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

[5] A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.

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

[7] J.A. Garay and I.S. Gopal. Call preemption in communication networks. In *Proceedings of the eleventh annual joint conference of the IEEE computer and communications societies on One world through communications (Vol. 3)*, IEEE INFOCOM, pages 1043–1050, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.

[8] J.A. Garay, I.S. Gopal, S. Kutten, Y. Mansour, and M. Yung. Efficient on-line call control algorithms. *Journal of Algorithms*, 23(1):180 – 194, 1997.

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

[10] M.H. Goldwasser. Patience is a virtue: the effect of slack on competitiveness for admission control. In *SODA*, pages 396–405, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.

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

[12] R.J. Lipton and A. Tomkins. Online interval scheduling. In *SODA*, pages 302–311. Society for Industrial and Applied Mathematics, 1994.

[13] S. Plotkin. Competitive routing of virtual circuits in atm networks. *IEEE Journal on Selected Areas in Communications*, 13:1128–1136, 1995.

[14] U. Schwiegelshohn and A. Tchernykh. Online scheduling for cloud computing and different service levels. In *IPDPSW, HPGC*, pages 1–10. IEEE, 2012.

Subproject A3
Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jörg Rahnenführer Peter Marwedel

A JVM-based Compiler Strategy for the R Language

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

The R language is known to have performance problems when handling large datasets. Our goal is to optimize the execution runtime of R programs. Therefore, we analyze the performance issues of machine learning programs written in R and present our development plans for a new JVM-based compiler strategy for the R Language.

Characteristics of R Programs

The R language [6] is designed as a dynamically typed, functional language for processing vector data. It uses call by value semantics and lazy evaluation.

As in R every value is a vector, even scalar values are boxed into single-element vectors, which results in often unnecessary vector allocations. On function calls, the vectors are passed by value, hence each vector is copied and the called function only sees and works on the copy. This results in a waste of memory resources, especially if the vector is not modified by the callee.

All variables and function symbols in an R program are stored in environments. Environments represent the interpreter state and are allocated on the heap of the interpreter. This even applies to variables that are referenced in the current environment only. However, determining which variables are local to a function is a complex task, as functions further below in the control flow can arbitrarily modify environments of their callers.

Function calls are evaluated lazily: a function is bound to the parameters at the call site, but evaluation is not performed until those values are used during program execution.

For each function call a *promise*, which represents a function closure with all parameters, is allocated on the heap. These features make R a very expressive language, however, they come at the cost of high memory management overhead.

Morandat et al. [4] analyzed R programs from different fields of statistics and were able to show major performance issues of the current R interpreter implementation. We used their *traceR* tool to analyze programs specifically from the field of machine learning to verify if those bottlenecks apply also to our programs. We examined machine learning programs which perform different types of regression for models of the survival analysis. Figure 1 shows the time profiles and Figure 2 shows the memory allocation profiles for two of those programs. One can see that about 30% of the runtime is spent on memory management and vector copies. Memory management includes the time for allocation as well as garbage collection and for operations on the interpreter stack.

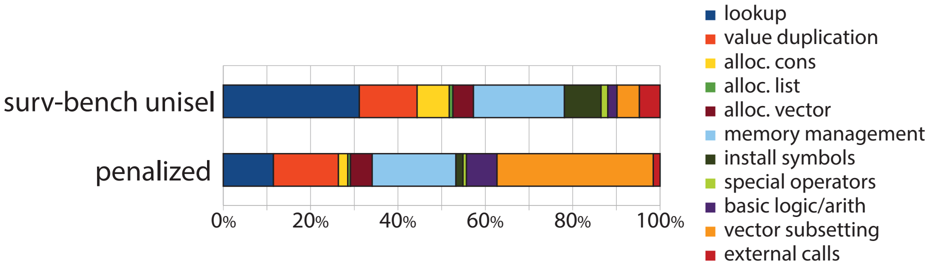


Figure 1: Runtime Profiles of Machine Learning R programs

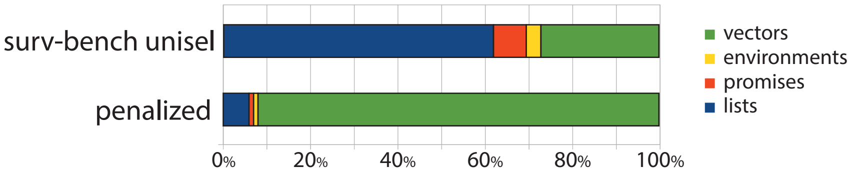


Figure 2: Memory Allocation Profiles of Machine Learning R programs

The memory profile of *surv-bench unisel* shows that about 25% of the runtime is spent on looking up functions from the environment. Those lookups have to be performed at all callsites in the program, even for builtin operators. About 10% of the time is used for symbol installation, which happens when R expressions are assigned to symbolic names within the interpreter. Memory management alone makes up to 20% of the runtime. The memory profile (Figure 2) of *surv-bench unisel* shows that over 60% of the allocated memory is used for list datastructures. There is also a good amount of memory used for promises, which corresponds to the large percentage of function lookups shown in Figure 1. Promises are created for every function call for lazy evaluation.

The *penalized* runtime profile shows that about 35% of the time is spent in subsetting operations. Subsetting is used for gathering or putting data at specific index ranges of vectors. The memory profile (Figure 2) of *penalized* shows that over 90% of the allocated memory is used for vector data structures. This corresponds to the runtime profile of *penalized* where most of the time is spent on the allocation of vectors for subsetting operations and vector copying.

Optimizations that can be applied to counter the mentioned bottlenecks are function specialization, the use of unboxed scalar values, reduction of vector copies as well as avoiding the creation of promises that are evaluated locally. To speed up basic vector operations, runtime specialization can be used to transform vector operations to SIMD instructions. To enable those optimizations we plan to develop a JVM-based compiler strategy.

JVM-based Compiler Strategy

R programs are usually processed by the native R interpreter written in ANSI C. This interpreter does not include a just-in-time (JIT) compiler, thus every single program line has to be evaluated separately. Our JVM-based compiler strategy for the R language includes an AST-interpreter for R written in Java and an extensible JIT compiler [3]. Figure 3 shows the components of our compiler toolchain.

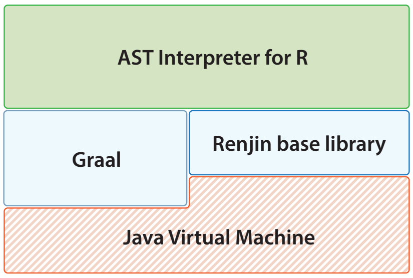


Figure 3: Components of the JVM-based Compiler Toolchain

With the use of a Java-based *AST-interpreter* for R, the execution of R programs could be optimized: By targeting the JVM, dynamic compilation is enabled within the interpreter code. However, transferring the R interpretation process to the JVM does not automatically lead to high optimization potential, because R programs still need to be interpreted.

Although dynamic compilation could already speed up the interpretation process, the native JIT compiler of the JVM is not aware of the R language and its specific optimization needs. In order to push more aggressive optimizations, the dynamic compiler should be extended by knowledge about R characteristics to enable language specific low-level optimizations and generate highly optimized machine code. For this purpose, the *Graal* compiler [5, 7], which is specifically designed for extensibility, should be employed.

Amongst others, Graal uses a program dependence graph [2] in static single assignment (SSA) form as intermediate representation (IR), which models both data- and control-flow dependencies. This Graal IR is extensible, and also additional optimization phases could be added to the compiler.

For the core library of the R language which is written in ANSI C, we looked at *Renjin* [1], an alternative R implementation written in Java. Renjin does not aim to be an exact clone of the original R interpreter, but it shares most of the structural design of R while using the object oriented features of Java to achieve a better modularity. Renjin contains an implementation of most of the R base library. By using and extending this library instead of rewriting it in Java from scratch should ease the implementation of the basic R functionality within our runtime system, but as Renjin is still under development, a certain amount of features is still missing.

To reduce the execution runtimes of R programs, optimizations on different levels should be applied. On the R *AST-Interpreter* level, source level optimizations could be applied to the IR produced by the Java-based R parser. During program execution R-specific runtime information could be gathered within the dynamic compilation process, which is required to apply the R optimizations mentioned above. To implement such a compiler strategy, not only new AST nodes for the intermediate representation have to be constructed; extending existing optimizations and implementing new R-specific optimizations is also challenging.

References

[1] Bertram, A.: Renjin: JVM-based Interpreter for the R Language for Statistical Computing. `http://code.google.com/p/renjin`, 2012.

[2] Ferrante, J. et al.: The Program Dependence Graph and its use in Optimization, ACM Transactions on Programming Languages and Systems, pp. 319-349, 1987.

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

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

[5] The Graal Project: `http://openjdk.java.net/projects/graal/`, 2012.

[6] The R Project for Statistical Computing: R Language Definition, `http://cran.r-project.org/doc/manuals/R-lang.html`, 2012.

[7] Wuerthinger, T.: Extending the Graal Compiler to optimize Libraries, In Proceedings of the ACM international conference companion on OOP systems languages and applications, pp. 41-42, 2011.

Integrative analysis of multiple genomic datasets

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

The major problem in the analysis of data from high-throughput technologies like microarrays is the typically small number of observations compared to the large number of features. But in the past years many microarray datasets have been made publicly available in databases such as the Gene Expression Omnibus (GEO) database [3]. Larger sample sizes can be achieved by combining multiple studies. This may lead to improved prediction performance and a robust feature selection. However, heterogeneity between the datasets implies severe problems for the later analysis. We want to identify and describe dataset specific differences and furthermore improve and develop new techniques for the integrative analysis of multiple studies.

For this purpose a comprehensive experimental framework is needed. Besides the rather complex data structures resulting from the multiple data sources, fitting statistical models on high dimensional data is computational expensive. For this scenario and in principle for most of the problems statisticians encounter in applied data analysis, the `R` [5] packages `BatchJobs` and `BatchExperiments` [1] have been developed.

We have extracted nine breast cancer datasets from the GEO database with the following characteristics: (a) survival times of the patients are included, (b) raw gene expressions are measured using the Affymetrix HGU133a chip, and (c) the sample size exceeds 50 patients. Considering the different definitions of survival endpoints [4], different treatments and missing values of important clinical covariates, up to five datasets with 889 patients linked to 22 283 probe sets remain for further analysis.

Pooling datasets requires in general great care and good knowledge of both the data and possible pitfalls. Normalization of the input data is basically always necessary but sometimes not sufficient. Identifying the latter cases is especially challenging in high dimensional settings like microarray analysis with more than 20 000 features.

We compared four different approaches to identify possible heterogeneity between the datasets and get more insight into resulting effects. All approaches rely on the Cox Proportional Hazard model [2]. The model determines the influence of p features on the hazard function $h(t)$ which is an appropriate surrogate for survival times. Let $h_0(\cdot)$ denote an arbitrary baseline hazard function and let $\beta = (\beta_1, \ldots, \beta_p)$ a the vector of regression coefficients, then the Cox model has the form

$$h(t \mid x) = h_0(t) \exp(\beta' x). \tag{1}$$

All compared approaches operate univariate on the datasets, i. e., each model is build using only a single gene and mandatory clinical covariates. Extending the comparison to models which include many or all genes at once is planned for the future.

The sadly most frequently used and at the same time most optimistic method to pool datasets is to just normalize the input datasets to the same scale and ignore any further concerns about heterogeneity. This straightforward but naive approach is compared with slight modifications of the Cox model (1):

- Include a dummy variable which encodes the originating dataset as a predictor variable. This allows the model to balance out differences through shifting the baseline hazard by an estimated amount for each dataset separately.
- Allow for different baseline hazards, also known as stratifying. For each dataset the baseline hazard $h_0(t)$ is estimated independently. This approach is much more flexible as it can even out effects over time, but still implies the constraint that the estimated effects are on the same scale for each strata.

A different approach for the integrative analysis of multiple datasets originates in the field of meta analysis. Instead of constructing one big model incorporating all the data, models are build separately for each dataset. The heterogeneity of the results can be examined with forest plots and be tested using fixed and random effects models. An exemplary forest plot in Figure 1 visualizes the estimated hazard rate of a single gene for each dataset separately as well as the aggregated hazard rate across datasets using fixed and random effects models.

This rather preliminary analysis resulted in some interesting observations which now need to be confirmed using simulation studies. Furthermore the analysis of the real data must be extended to examine the influence of different mandatory clinical covariates or definitions of the survival endpoint. For a complete comprehensive analysis, including the simulation, the packages `BatchJobs` and `BatchExperiments` will be used.

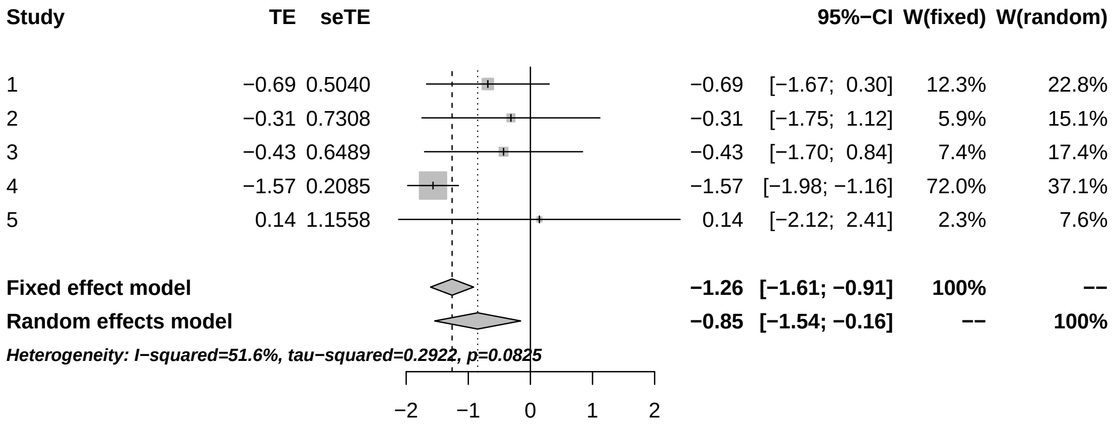


Figure 1: Exemplary forest plot for a single gene and five data sets

The package BatchJobs implements the core infrastructure to communicate with high performance clusters (HPC), also referred to as batch systems, from within R. Implementations of the popular higher order functions Map, Reduce and Filter can be used to define batch jobs, or to collect and filter results. The communication with the cluster system is handled transparently and hassle-free for the user. The computational state is always persistent in a database and subsets of jobs can be extracted on criteria matching the computational state or job parameters. Built-in debug features assist in tracking down bugs which is especial tedious in parallel computing.

BatchJobs is in principle independent from the underlying batch system and can easily be extended using a simple API and templates. As of writing, implementations for the three batch systems TORQUE/PBS[1], Load Sharing Facility[2] and Oracle Grid Engine[3] are already built-in. The support for the last two remaining popular batch system, Condor[4] and Slurm[5], is planned and will be included into the package soon. Furthermore we provide local single-core and multi-core execution as well as the possibility to merge several loosely connected machines into a makeshift cluster.

The second package, BatchExperiments, extends BatchJobs with an abstraction for computer experiments. Most everyday tasks in data analysis can be broken down to "ap-

[1] http://www.adaptivecomputing.com/products/open-source/torque/

[2] http://www.platform.com/workload-management/high-performance-computing/lp

[3] http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html

[4] http://research.cs.wisc.edu/condor/

[5] https://computing.llnl.gov/linux/slurm/

plying algorithms on problems". Consequently `BatchExperiments` offers flexible functions to define problems and algorithms, connect them with experimental designs and then submit them to the underlying batch system. Seeding mechanisms automatically ensure reproducibility and the effective usage of `BatchJobs`' design renders complete portability: Experiments developed on a local machine can easily be moved to a cluster whenever computational times demands for it – and moved back to analyse the results. Many optimizations to support large scale experiments have already proven useful in self-conducted computer experiments with more than 500 000 jobs. All the functionality provided by `BatchJobs` is of course also available in `BatchExperiments`.

The possibility to move from a HPC to a cloud computing service whenever the HPC's computing power is not sufficient anymore is a feature we want to have in the long run. An implementation for Amazon's EC2[6] is scheduled for next year.

References

[1] Bernd Bischl, Michel Lang, Olaf Mersmann, Jörg Rahnenführer, and Claus Weihs. Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments. Technical report, TU Dortmund, 2012.

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

[3] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–10, January 2002.

[4] Clifford a Hudis, William E Barlow, Joseph P Costantino, Robert J Gray, Kathleen I Pritchard, Judith-Anne W Chapman, Joseph a Sparano, Sally Hunsberger, Rebecca a Enos, Richard D Gelber, and Jo Anne Zujewski. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(15):2127–32, May 2007.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[6] `http://aws.amazon.com/ec2/`

Classification methods for high dimensional genomic data

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

Problem statement

In the last years the number of available datasets with high-dimensional measurements from molecular biology has increased drastically. A typical challenge is the presence of a large number of variables, often in the thousands, compared to a small number of experiments (samples), typically at most a couple hundred. In these experiments the expression (activity) or abundance of thousands of genes or proteins is measured on a genome-wide scale. The resulting data enable a better understanding of the underlying biological processes that cause diseases or influence disease progression. Many applications in this field originate from cancer research. Here, the goal is to obtain improved models for cancer development and new algorithms for improved disease diagnosis and therapy selection. For the selection of the statistical methods it is critical to consider the type of the target variable. For the choice of the therapy typically a binary or categorical decision is required. In this context classification methods have been studied extensively in the literature. However, a meaningful alternative is to directly use the survival time of a patient at target variable or categorize patients into two groups according their survival time. In this scenario it is important to appropriately deal with the problem of missing data due to censoring.

Goal

The major goal of our research is to identify new markers from genetic data of tumor samples that allow a classification of the corresponding patients according to different survival time prediction. An important point is to find a tradeoff between model interpretation, model stability, and prediction accuracy.

Prediction and classification methods

For this purpose, we have compared several modern competitive classification methods that allow the use of high dimensional data as variables in Cox models [1]. We have selected a number of cancer data sets and applied the methods in suitable cross-validation scenarios. To address the overfitting problem, a popular approach is to use penalizes likelihood methods. Ridge regression and lasso regression are the most prominent representatives of this class of algorithms. We have compared the results to a new approach called survival SVM [2,3]. Kernel based methods, especially support vector machines (SVMs) have been successfully applied in many domains in biostatistics and bioinformatics, in particular also in the area of high dimensional data. The survival SVM makes use of the concordance index as a measure of association between the predicted and observed failures in case of right censored data. Two samples are called concordant if the order of failures is the same between the predicted and the observed times. The SVM tries to maximize the concordance index.

Evaluation study

In our studies we have used support vector machines both for classification of patient subgroups as well as for directly modelling the survival time. In order to classify patients into prognostic groups we have used on the one hand clinical covariates such as tumor stage or WHO grade and on the other hand binarized event times. The latter means that a cutoff (e.g. 5 years after surgery) is determined and patients are divided into two groups according to the fact if their event time is smaller or larger than this cutoff. We have applied several penalized regression methods and the kernel based methods regarding their prediction quality on several cancer data sets. More precisely, in one study we used three data sets that contained survival times of breast cancer patients. For evaluation we used various measures like a score based on the Logrank test, the Brier score, and the concordance index. It turned out that the SVM based methods are highly competitive for this application compared to established penalized regression methods. An advantage of the SVM based methods is that the variables can be ordered according to their weight in the resulting decision function, i.e. according to their expected importance in the classification. These weights can then be a guide for variable selection by filtering. Two alternatives exist. One can choose a prefixed number of genes or one can select genes with a weight above a quantile of the distribution of the weights. An overview of the applied analysis is presented in Figure 1 on page 3.

New toxicology project

A second research topic in the last months is based on a cooperation with Prof. Dr. Jan Hengstler from IfADo (Leibniz-Institut für Arbeitsforschung an der TU Dortmund). In a toxicological study nerve cells were treated in vitro with different compounds in different doses. Then genome gene expression was measured in the treated cells. We have applied a large analysis pipeline to evaluate the influence of compound and dose

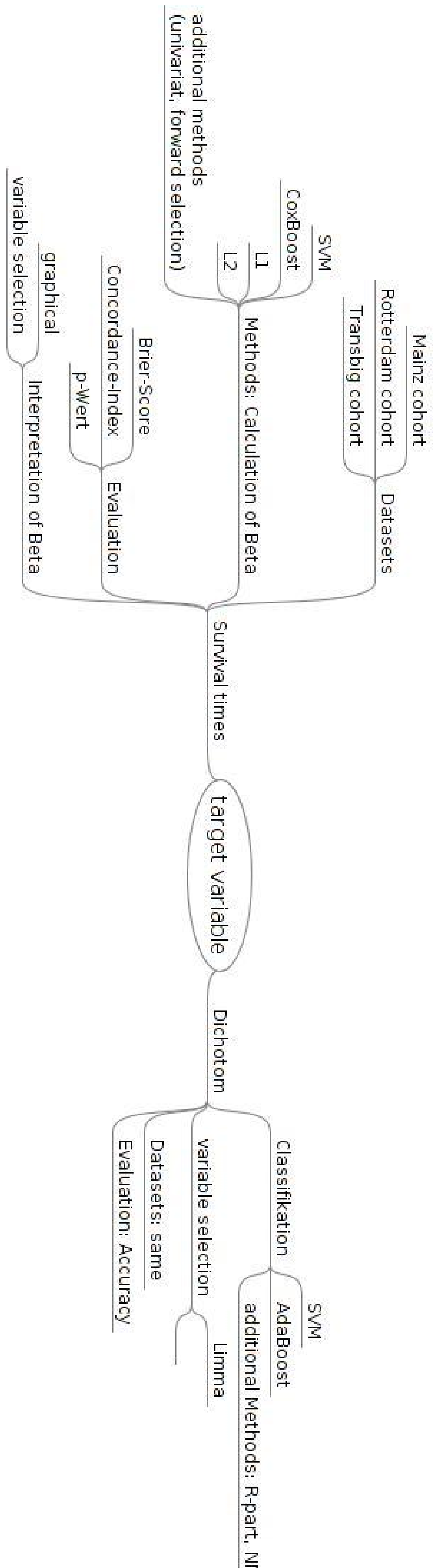


Figure 1: Overview of the analysis scheme

on the gene expression changes, including exploratory data analysis (principal component analysis, cluster analysis), identification of differentially expressed genes with adjusted t-test like statistics, and identification of differential pathway activity (gene set overrepresentation analysis). Future goals are the development of methods for classifying the compounds according to toxicity classes or according to the dose of the corresponding compound. Again, the value of SVM based methods are to be investigated.

Literature

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

[2] Van Belle V., Pelckmans K., Van Huffel S., and Suykens J.A.K. Support vector methods for survival analysis in clinical applications: a combined ranking-regression approach. Technical report, 09-235, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) 2009, submitted for publication. Downloadable from ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanbelle/reports/09-235.pdf.

[3] Evers, L., and C.-M. Messow. Sparse Kernel Methods for High-dimensional Survival Data. Bioinformatics, 14(24): 1632–1638, 2008.

Subproject A4
Resource efficient and distributed platforms for integrative data analysis

Peter Marwedel Olaf Spinczyk Christian Wietfeld

Configurability in Embedded System Software: An Opportunity for Optimizing Energy-Consumption

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

System software for embedded devices has to be extremely efficient. Memory usage, processing overhead and energy consumption should be minimized in order to meet the physical constraints of typical embedded systems. Such a multi-objective optimization is difficult to achieve with existing software that was not intended for this. In a case study of networking stacks for embedded devices, I will show how to minimize energy consumption, which is often the most scarce resource on battery-powered devices.

1 Introduction

There is a clear trend towards smart (computerized) devices, which often need to interact with the Internet, such as modern mobile phones. Hence, system software for these devices, in particular operating systems, implements the TCP/IP protocol suite, which is the de-facto standard for information exchange between computer systems.

However, the software-intensive Internet protocols were not intended (primarily) for deeply-embedded, resource-constrained devices, so that engineers likely run into efficiency problems when using TCP/IP. For example, memory capacities and energy are often tightly constrained, especially for battery-powered devices.

Most TCP/IP implementations were designed as a component of a PC or server operating system – such as BSD, Linux, or Windows – and are quite resource intensive. Prominent examples for TCP/IP stacks for embedded devices are *micro-IP* and *lightweight-IP* [2]. uIP proved successfully that memory constraints for TCP/IP stacks can be met by reducing the functionality to an absolute minimum by leaving out all optional communication-protocol features. In the following sections, I will examine both uIP and lwIP in more detail, in particular with regards to energy efficiency.

2 State of the Art

As mentioned before, uIP implements the TCP/IP protocol suite in a very minimal way, whereas lwIP offers much richer functionality for embedded devices, which may be required for more-demanding application scenarios. Consequently, there is a gap in between, which is filled out by my TCP/IP implementation *CiAO/IP* [1]. CiAO/IP is a highly-configurable TCP/IP stack that can be statically configured to meet the application developers' requirements by leaving everything off that is *not* needed - in order to achieve minimal resource consumption for a particular use case. This configuration process is done at compile time, when the application developer specifies a set of required protocol features. CiAO/IP differs from uIP and lwIP in a very high degree of configurability - *every* feature is optional and, thus, configurable by the application developer.

Table 1 gives an overview on the features of CiAO/IP and those of uIP and lwIP. A configurable feature is denoted by ✓and a fixed feature, which is always present and not removable by the end-user, is denoted by $*$. Unimplemented features are shown as whitespace.

3 Optimization Potential

The main benefit of the high configurability offered by CiAO/IP is optimization for *nonfunctional properties*, such as energy consumption. To assess the potential of energy savings, I compared the energy consumption of uIP and CiAO/IP on the *BTNode*[1] sensor-network platform, which basically consists of an 8-bit ATmega 128L AVR microcontroller with 128KiB ROM and 4KiB RAM plus a sub 1 GHz CC1100 radio device.

A typical duty for sensor nodes is a firmware update: a file transfer of 32 kB over TCP in this case. For energy measurements[2], I use the *B-MAC* [3] protocol for collision avoidance and preamble sampling with an interval of 25 ms (*Low Power Listening*).

Figure 1 presents the mean energy consumption and the sample standard deviation of uIP and CiAO/IP. For fairness, both IP stacks were configured with exactly equivalent features.

[1] `http://www.btnode.ethz.ch/`

[2] *Hitex PowerScale* with *ACM probes* used for energy measurements

Feature	**uIP**	**lwIP**	**CiAO/IP**
Multiple Networking Interfaces		*	✓
Checksum Offloading per Interface			✓
Multiple Connections (Concurrency)	*	*	✓
Buffers per Connection (Isolation)			✓
Operating System Support		✓	✓
IPv4	✓	✓	✓
IPv4 Tx	*	*	✓
IPv4 Rx	*	*	✓
IPv4 Fragment Reassembly	✓	✓	✓
IPv4 Fragmentation		✓	
IPv6	(✓)[a]	(✓)[a]	(✓)[b]
ARP	✓	✓	✓
ARP Reply	*	*	✓
ARP Request	*	*	✓
ARP Cache Timeout	*	*	✓
Static ARP Cache Entries			✓
ICMP	*	✓	✓
UDP	✓	✓	✓
UDP Tx	*	*	✓
UDP Rx	*	*	✓
UDP Checksumming	✓	✓	✓
TCP	*	✓	✓
Client (Connect)	✓	*	✓
Server (Listen)	*	*	✓
Sliding Window (Tx)		*	✓
Sliding Window (Rx)		*	✓
Silly Window Syndrome Avoidance (Tx)		*	✓
Silly Window Syndrome Avoidance (Rx)		*	✓
Round-Trip Time Estimation		*	✓
Congestion Control (Slow-Start)		*	✓
TCP Urgent Data	✓	*	
Limit Excessive Retransmissions	*	*	✓
MSS Option	*	*	✓
Timestamp Option		✓	

[a] provided by a different code base, no dual stack

[b] under development

Table 1: Features provided by uIP, lwIP and CiAO/IP.

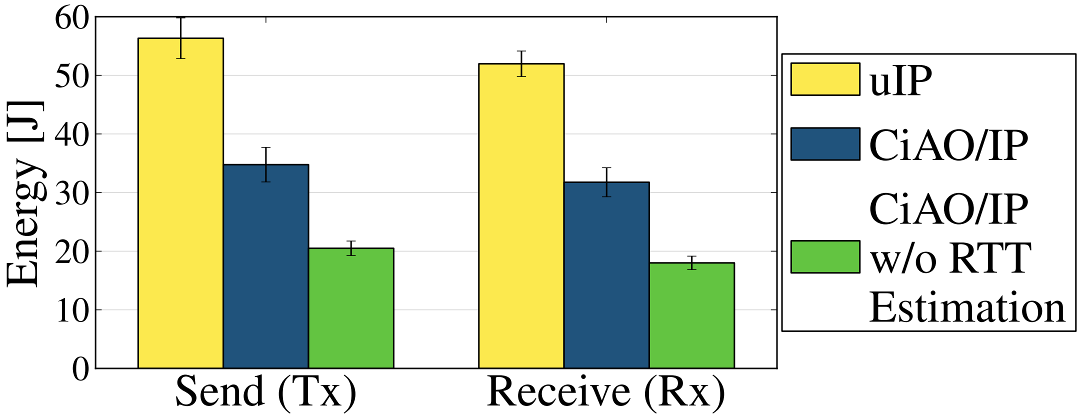


Figure 1: Shown is the mean energy consumption and the sample standard deviation of unidirectional TCP data transfers (32 kB) between two wireless sensor-networking nodes. *CiAO/IP w/o RTT Estimation* uses a fixed retransmission timeout (200 ms) for lost packets.

CiAO/IP consumes significantly less energy: This is caused by the coarse granularity of uIP's round-trip time estimation, which can only be multiple of 500 ms. In CiAO/IP, this estimation is configurable in units of 1 ms. Additionally, this feature of CiAO/IP can be deactivated at all, which further reduces the energy consumption.

This baseline assessment reveals the high potential of fine-grained configurability for optimization of system software. However, the optimization has to be carried out in a smart and automatic way by inferring which features should be activated and which not. This calls for algorithms of Data Mining and Machine Learning.

References

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

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

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

Energy measurement and -models for embedded systems

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

Basic research on measurement methods for low-power devices was done in the last year within the SFB 876 A4 project. Mathias Meier introduced a measuring method for DC current using a shunt resistor[1] [4]. As this method is very simplistic, it lacks the necessary preciseness, because for low-power applications, a shunt with a relatively high resistance is needed to achieve a voltage-drop which is large enough for measurement. But a high voltage-drop means a significant change in the voltage supplied to the device under test. Additionally, further effort was needed to achieve long-time measurements with enough time precision for our purposes. We developed a measurement device which can deal with low-power situations, integrates the current between two samples and allows long-time measurements.

In the field of power modeling, we combined approaches to model hard- and software to gain information about the average power consumption of a system and its components. This information can be used to implement an API for an online power estimation. This approach is to be published in [1].

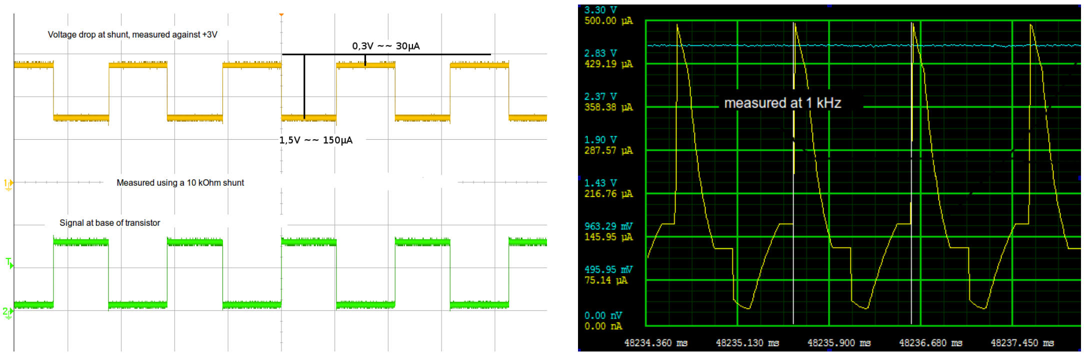
1 Measurement methods for low-power devices

Before trying to develop own tools, we investigated commercial products and tested the Hitex Powerscale system [2] in combination with a Hitex ACM probe. The manufacturer claims this device to be able to measure currents from 200 nA up to 500 mA, automatically switching between several shunt resistors to enlarge its measurement range, while

[1] Low-ohmic and precise resistor

keeping the voltage drop below 100 mV. All measurements can be done at a sample rate of 100 kHz and are supported by an additional voltage measurement and a graphical Windows application. As we could show, the Powerscale system was not able to fulfill all requirements.

Fig. 2(a) shows a test circuit which simulates a device with periodically changing power consumption. A rectangle generator is used to trigger a transistor circuit, which changes between two load resistors at a preset frequency. By using a shunt and an oscilloscope, we were able to prove the correctness of this circuit. Fig. 1(a) shows a captured image of the oscilloscope display. In comparison, Fig. 1(b) shows the results of the Powerscale system with ACM probe. Even the basic, rectangular shape of our load curve cannot be identified anymore.



(a) Reference measurement at 1 kHz

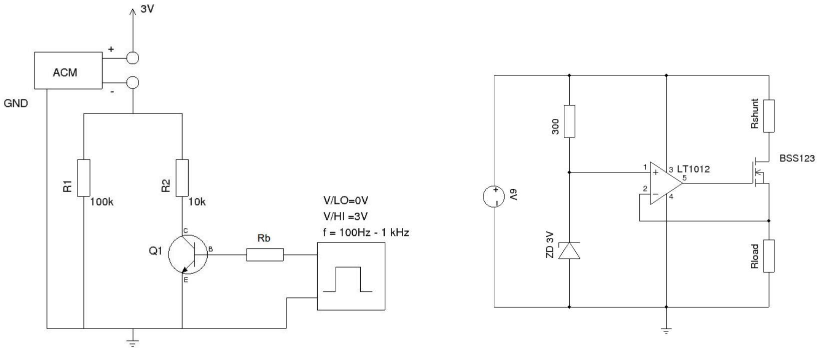
(b) Rectangle low-power signal at 1 kHz measured with Powerscale and ACM probe

Figure 1: Reference- and PowerScale results

In search for a solution to our problem – gaining reliable data on power consumption for further evaluations – we decided to develop a custom circuit to fulfill our needs.

Fig. 2(b) shows the basic idea of the core of the circuit. It consists of an operational amplifier (OpAmp), which keeps the voltage supply of the test device at a constant level, which is preset by a Zener diode. This means, the circuit is a power-supply and a measuring device at the same time. The needed electrical current still flows over a shunt, which now can have a high resistance, because the resulting voltage-drop will be compensated by the regulating OpAmp circuit.

This measuring circuit had to be extended further for long-time measurements. Previously, an oscilloscope was used to obtain short snapshots of the power consumption at a high time resolution, thus approximating the energy consumption (the integral of the power consumption) well due to this resolution. To measure over a longer period of time, an ADC and a USB interface were added to the circuit and analog integration is used to calculate the energy consumption between two samples. Within intervals of 10 µs, the ADC samples the *integral* of the power consumption function at 100 kHz. Since it would require an indefinitely high voltage to perform DC integration over time, three analog integrators based on OpAmps and a capacitance are used, and at any time, one integrator is integrating, one is holding the integrated value of the previous interval for sampling by the ADC and the remaining integrator is discharging to prepare for integrating the next interval. Thereby the circuit can measure the energy consumption over an indefinitely long period of time at a reasonable time resolution.[2]



(a) Test circuit

(b) Basic OpAmp circuit for measuring low power current

Figure 2: Test- and measurement circuits

2 An energy model for embedded systems

The most common approach for hardware energy modeling uses priced finite-state machine (FSM) models, which describe the states of a hardware component and the corresponding energy draw. These models can be extended by a timing behavior, describing intrinsic state transitions of the hardware component. Such models are usually referred to as *priced timed automata* (PTA). PTAs have the advantage that they do not enforce a certain level of detail, but they can get very complex in a fine grained model.

As PTAs are finite-state machines, they have inputs and outputs, which represent hardware signals and interaction with the physical environment. As an embedded system consists of multiple hardware components and a controlling software layer, the PTAs have to be interconnected accordingly. While the interconnection of hardware components might be trivial, as the hardware design of the system should be well-known, there is still a need for a fitting signaling model of the controlling software layer. Assuming such a model existed, it would become possible to determine the percentage of time each hardware component resides in each of its PTA states. Knowing the electric current draw in each state, it would facilitate the calculation of the average power consumption of each component over time.

[2] The author would like to thank Christian Günter for ideas and work on the integration circuit.

The idea behind the software model is to control the automata models of the hardware components. To achieve this, a straight-forward approach is the execution of the actual code in a specialized target-platform emulator, as shown by Landsiedel et al. in [3]. While a target-platform emulator is timing-correct in itself and does not need a code-transformation, it is a very specialized solution. For each considered target system, there must be a dedicated emulator that is able to deal with PTAs. In [1] we showed an approach that utilizes a standard simulation engine (OmNeT++), to process a behavioral and timing-correct model of the software layer of a basic sensor node. Such a model lacks intrinsic timing-correctness, but with a given timing model added to the behavioral model, the simulator can be timing-correct while still being target-platform independent. One drawback is the need to transform software into a simulator model. As no good automatisms exist for this kind of transformation, we propose to reverse the design flow: With model-driven software development (MDSD) techniques, it should be possible to generate both simulator code and target-platform code based on a behavioral model.

Often, the results of the simulation will be dependent on the environmental situation, when user-interactions or sensor events change the control flow of the software. So it becomes necessary to define what we called "environmental scenarios", which are sets of environmental conditions, e.g. how many user interactions are supposed to occur, or how much packet-loss is expected. With a broad set of environmental scenarios, a map of expected energy consumption can be created. Such a map represents the resulting energy usage either per-device or averaged for the whole system. We showed in [1] that these result sets can be condensed into an array for use in a run-time application interface.

References

[1] Markus Buschhoff. A unified approach for online and offline estimation of sensor platform energy consumption. In *Proceedings of the 4th International Wireless Communicaion and Mobile Computing Conference (IWCMC), Workshop on Energy Aware Computing and Communication Networks*, Limassol, Cyprus, 2012. To appear.

[2] HiTex. Energy optimization: Powerscale.
`http://www.hitex.com/index.php?id=powerscale`.

[3] O. Landsiedel, K. Wehrle, and S. Gotz. Accurate prediction of power consumption in sensor networks. In *Proceedings of the 2nd IEEE workshop on Embedded Networked Sensors*, EmNets '05, pages 37–44, Washington, DC, USA, 2005. IEEE Computer Society.

[4] Matthias Meier. Measurement methods and prediction of resource consumption. In Katharina Morik and Wolfgang Rhode, editors, *Technical report for collaborative research center sfb 876 - graduate scool*, number 4, pages 40–43. TU Dortmund University, October 2011.

Empirical Profiling of LTE UE for Energy-Efficient Data Transmission

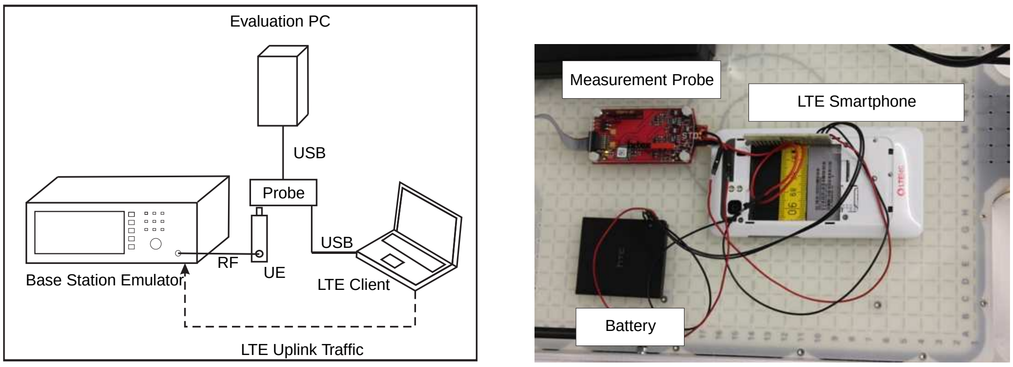
Björn Dusza
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
bjoern.dusza@tu-dortmund.de

In this report a methodology for empirical power consumption modeling of Long Term Evolution (LTE) User Equipment (UE) as well as an example application of such approach are shown. It is illustrated that major power savings of up to 27.5% are possible if physical resources above the required minimum are provided to the subscriber. Therefore, a scheme is presented, that allows network operators for trading network capacity for battery lifetime.

1 Motivation

The overall energy-efficiency of mobile networks has become a major research focus for the design and implementation of current and next generation wireless networks. This is due to the fact that the amount of energy that can be stored in modern lithium-ion batteries is not developing as fast as the energy hunger of current smartphones which leads to dramatically decreasing battery lifetimes, especially for heavy users. Nevertheless, the time that an UE can be operated with one filling of the battery is one of the most important decision parameters for the customers of new devices. For the performance evaluation of power efficient protocols and algorithms it is of major importance to have detailed knowledge on the relationship between the system parameterization (allocated bandwidth, transmission power, UE characteristics) and the actual power consumption. In the following, an empirical approach addressing this issue as well as one example application of power consumption measurements for energy-efficient Voice over IP (VoIP) communication will be presented.

2 Measuring the Power Consumption of LTE UE



(a) Measurement Setup for Power Measurements of USB enabled UE

(b) Power Consumption Measurement for LTE Smartphone

Figure 1: Power Consumption Measurements for Various Types of LTE UEs

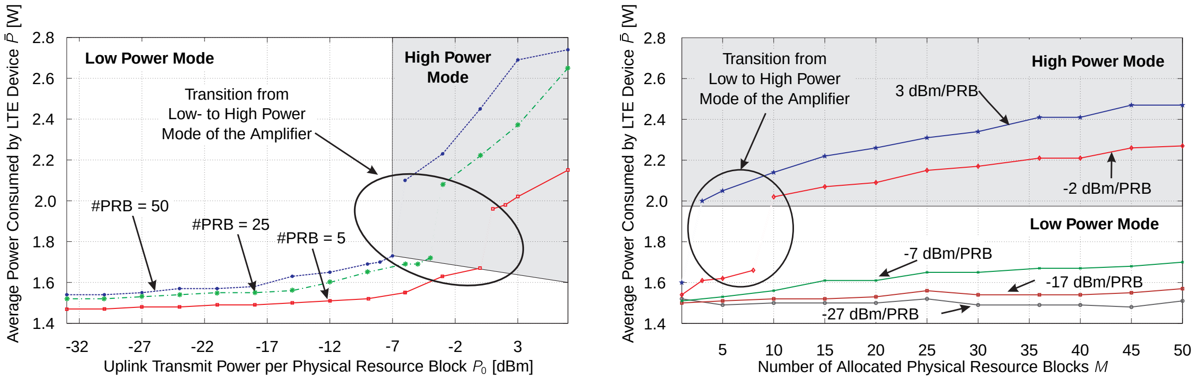
The measurement setup shown in Fig. 1 illustrates how the power consumption of a commercially available LTE UE can be reliably measured in a laboratory environment. An LTE Base Station Emulator (BSE) is used for the creation of a standard conform radio cell in the lab. This device allows for full control of the uplink transmission power P_{Tx} without any impact of the radio channel. The device under test is then attached to the BSE and a data transfer under a given system parameterization is established. The average power that the UE consumes during this data transfer is measured by means of a specialized power logging device. In case of an USB enabled UE, the voltage and current is measured between the host PC and the USB data-stick (cf. Fig. 1(a)) while for smartphones the power consumption is determined directly at the battery (cf. Fig. 1(b)).

3 Quantification of the UE Power Consumption

The uplink transmission power of an LTE UE can be expressed as

$$P_{Tx} = P_0 + 10 \cdot log_{10}(M) \qquad (1)$$

with the transmission power per Physical Resource Block (PRB) P_0 and the number of allocated PRB M. For the determination of the power consumption it is assumed the the UE is operating at a fixed value of P_0 that is independent of the actual path loss. The impact of P_0 and M on the average power consumption $\bar{P}$ of a Samsung GT-B 3730 data-stick is shown in Fig.2 [4]. As one can see from Fig. 2(a), the power consumption curve can be devided into two pieces. For low values of P_0 (low power mode) the curve is only slightly increasing and almost independent of M. If a P_{Tx} dependent threshold is



(a) Influence of the UL Transmit Power per PRB on the Power Consumption

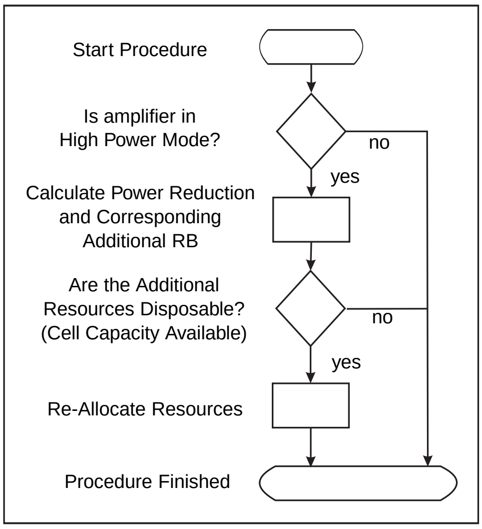
(b) Influence of the Number of Allocated PRB on the Power Consumption

Figure 2: Influence of the UL Transmit Power on the Energy Consumption

reached the slope of the curve immediately becomes steeper (high power mode). This effect is due to an additional power stage in the amplifier. A comparable effect can also be observed in Fig. 2(b). Here, the value of $\bar{P}$ is shown for various values of M and five dedicated values of P_0.

4 Trading Network Capacity for Battery-Lifetime

As it was shown in the previous section, significant power savings and therefore battery-lifetime enhancements are possible by avoiding the disadvantageous high power mode. For real-time applications (e.g. VoIP), which are characterized by the demand for a continuously available data rate, this can be achieved as illustrated in Fig. 3 [4]. If the network is not completely utilized, e.g. because it is over-dimensioned, it should be considered by the network operator to allocate more PRB than the minimum requirement to the users. These additional PRB can than be used for a reduction of P_{Tx} and therefore $\bar{P}$. If a user is just at the edge of the high power mode, already a minor reduction of P_{Tx} may cause a significantly increased battery lifetime. At the x-axis of Fig. 4, selected combinations of Modulation and Coding Scheme (MCS), number of PRB



Start Procedure
Is amplifier in High Power Mode?
no
yes
Calculate Power Reduction and Corresponding Additional RB
Are the Additional Resources Disposable? (Cell Capacity Available)
no
yes
Re-Allocate Resources
Procedure Finished

Figure 3: Flow Chart Illustrating Generic Power Reduction Scheme

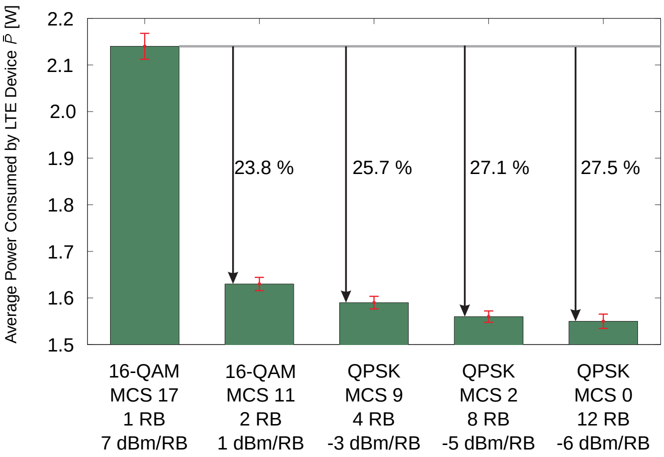


Figure 4: Power Consumption of LTE UE for Different MCS/RB Constellations

and transmission power that is needed for achieving an MCS dependent target SNR at the base station are given. All of these combinations allow for the same throughput. As one can see from Fig. 4 a reduction of the average power consumption of up to 23% is possible if only one additional PRB is allocated and a more robust MCS is used [4]. The use of a stronger error correction code comes along with a higher redundancy and therefore the need for additional PRB. On the other hand the transmission power can be reduced without effecting the user experience.

References

[1] Björn Dusza, Christoph Ide, and Christian Wietfeld. A Measurement Based Energy Model for IEEE 802.16e Mobile WiMAX Devices. In *Proc. of IEEE 75th Vehicular Technology Conference (VTC 2012-Spring)*, Yokohama. Japan, May 2012. IEEE.

[2] Björn Dusza, Christoph Ide, and Christian Wietfeld. Interference Aware Throughput Measurements for Mobile WiMAX over Vehicular Radio Channels. In *Proc. of IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Paris, France, April 2012. IEEE.

[3] Björn Dusza, Christoph Ide, and Christian Wietfeld. Measuring the Impact of the Mobile Radio Channel on the Energy Efficiency of LTE User Equipments. In *Proc. of the 21st International Conference on Computer Communication Networks (ICCCN)*, Munich, Germany, July 2012. IEEE.

[4] Björn Dusza, Christoph Ide, and Christian Wietfeld. Utilizing Unused Network Capacity for Battery Lifetime Extension of LTE Devices. In *Proc. of the IEEE International Conference on Communications Workshops (ICCW)*, Ottawa, Canada, June 2012. IEEE.

Low Power Femtocells for Complex Fading Environments

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

The demand of broadband services for increasing data rates cannot be handled by the existing macrocell infrastructure. Femtocells counter this problem by small coverage areas combined with high spatial reuse and Long Term Evolution (LTE). However, femtocells come along with severe Inter-Cell Interference (ICI) as they use the same frequency resources as macrocells. Therefore, a new spectrum allocation policy for femtocells based on self-organizing random frequency hopping is proposed. The achieved gain is evaluated in different frequency non-selective as well as frequency selective fading environments by an analytical model for the Bit Error Ratio (BER). In this way, a reduction of the BER up to a factor of 7 can be achieved compared to systems where the interferer transmits in the same frequency bands as the femtocell user. The results are validated by independent simulations.

1 Motivation and Problem Statement

As the traffic in mobile radio systems increases approximately 30% each year, macrocells with large coverage areas are no longer able to handle such amounts of data. Therefore, small low power cells with high spatial reuse have been introduced, called femtocells. They are typically applied to extend indoor coverage or to offload traffic from macrocells. As femtocells normally share the same frequency resources as macrocells, such applications result in two-tier networks with severe interferences. In case of a downlink transmission, the femtocell users are interfered by the macrocell base station, while for an uplink transmission, the femtocell access point is interfered by macrocell users, cf. Fig. 1 and Fig. 2.

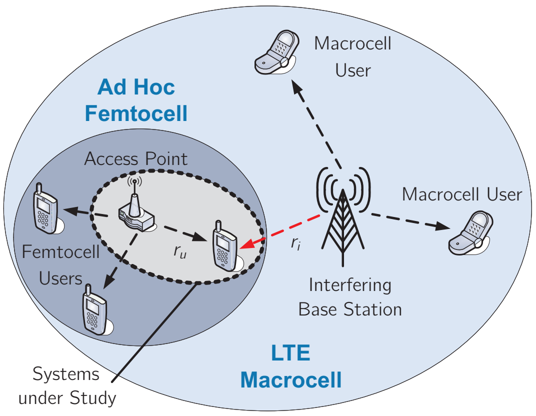


Figure 1: Interference in the downlink

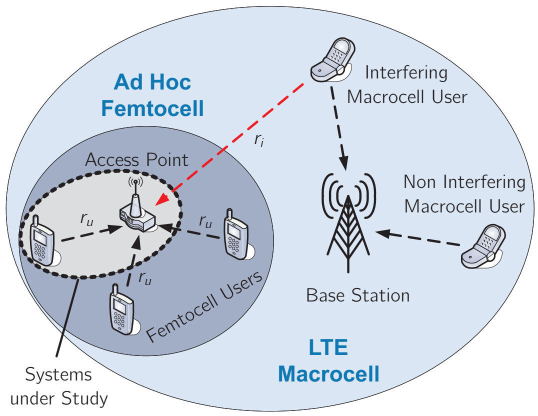


Figure 2: Interference in the uplink

Since the positions of femtocells within the existing infrastructure are random, the required interference mitigation has to be performed in a self-organizing way. Many different self-organizing ICI mitigation techniques have been published, which can be categorized into the groups Time and Frequency ICI cancellation [1]. As centralized orthogonal planning is not possible for femtocell networks, due to their random locations, existing interference mitigation approaches are commonly based on knowledge of the used resources of surrounding macrocells. In this way, the femtocell resources, e.g. frequencies or transmit powers, can be adjusted to reduce the interference.

Since all of the known ICI mitigation techniques require knowledge of the cell environment, we propose to use random carrier frequencies (RFH) for femtocells. Each femtocell user allocates a random hopping pattern according to a given probability density function before transmission and informs the receiver about the selected pattern. In this way, femtocells are able to integrate themselves into macrocells without exchange of information between the cells. Furthermore, no time is consumed for sensing the cell environment, which makes random frequency hopping highly attractive in cell outage scenarios, where macrocell base stations fail and coverage has to be immediately provided by femtocells.

2 Analytical Model for the BER

In order to quantify the introduced interference of Orthogonal Frequency Division Multiple Access (OFDMA) femtocells applying random frequency hopping, we derive an analytical model for the BER in different fading environments. On condition of Binary Phase Shift Keying (BPSK) for modulation, the BER is equivalent with the Symbol Error Ratio (SER). On the other hand, the SER can be deduced from the interference signal at the output of the decision device in the femtocell user equipment [2]. By applying Laplace transform techniques and a subsequent Gauss-Chebyshev approximation, the BER per subcarrier s can be calculated for fixed fading amplitudes as [3]

$$\mathrm{BER}_s \approx \frac{1}{\nu} \sum_{k=1}^{\nu/2} \mathrm{Re}\left\{\Phi\left(cr_k\right)\right\} + \tau_k \mathrm{Im}\left\{\Phi\left(cr_k\right)\right\} \tag{1}$$

with

$$\Phi(x) = \mathrm{e}^{\frac{\sigma_n^2 x^2}{2}} \left[\mathbb{E}\left\{\mathrm{e}^{-x i_s}\right\}\right]^M \mathbb{E}\left\{\mathrm{e}^{-x g_s}\right\} \ , \tag{2}$$

where r_k is a complex numerical parameter according to [3], c the Chernoff bound of $\Phi(c)$, σ_n the variance of Additive White Gaussian Noise, g_s the frequency domain channel gain, and M the number of interferers. In case of non-frequency selective fading, equation (1) has to be averaged according to the distribution of fading amplitudes $p_r(x)$

$$\Phi(x) = \mathrm{e}^{\frac{\sigma_n^2 x^2}{2}} \left[\int_0^\infty \mathbb{E}\left\{\mathrm{e}^{-x i_s}\right\} p_r(x) \mathrm{d}x\right]^M \mathbb{E}\left\{\mathrm{e}^{-x g_s}\right\} \ , \tag{3}$$

whereas for frequency selective fading, the interference signal i_s has to be summed up according to the number of multipaths L

$$\Phi(x) = \left[\mathrm{e}^{\frac{\sigma_n^2 x^2}{2}}\right]^L \left[\int_0^\infty \mathbb{E}\left\{\mathrm{e}^{-x \sum_{l=1}^{L} i_s}\right\} p_r(x) \mathrm{d}x\right]^M \left[\mathbb{E}\left\{\mathrm{e}^{-x g_s}\right\}\right]^L \ . \tag{4}$$

3 Reduction of BER by Random Frequency Hopping

In the following, a performance analysis is presented based on the analytical models of the previous section. All results are based on typical system parameters of LTE cells in [3] and represent a worst case analysis, as both cells serve the maximum number of users. Fig. 3 depicts the BER according to (1) and (2) as a function of the distance to the macrocell interferer and different hopping distributions. In order to evaluate the introduced gain, the figure also shows the BER of systems using centralized orthogonal planning (best case) and of systems where the macrocell interferers allocate the same frequency bands as femtocell users (worst case).

Fig. 3 demonstrates that an increasing distance r_i reduces the BER regardless of whether uniform or Gaussian random hopping is applied. However, uniform hopping always achieves a lower BER compared to Gaussian one. This is due to the fact that uniform hopping uses the whole bandwidth for transmission, while Gaussian hopping prefers the center of the channel. In general, random hopping benefits from free bandwidth within the cells, which arises when the users are not allocating the complete cell spectrum. This free bandwidth cannot be used by cells with centralized planning, but by cells deploying random frequency

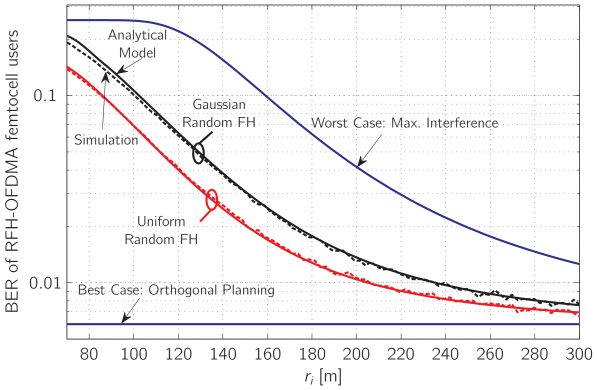


Figure 3: BER versus distance of the interfering systems

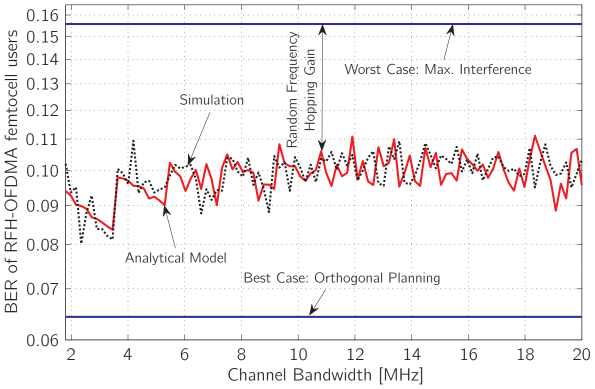


Figure 4: BER in a Rayleigh frequency-nonselective channel

hopping. The results in the figure are validated by simulations, which are composed of a BPSK modulation with random input bits, an inverse Fourier transform, a random carrier frequency oscillator, a channel with corresponding fading and a receiver reverting all processing of the transmitter.

In case the mobile radio channel is characterized by non-frequency selective Rayleigh fading, the evaluation of (3) as a function of the channel bandwidth results in Fig. 4. Although Rayleigh fading leads to simultaneous degradation of the received power from the femtocell access point as well as from the macrocell interferer, random frequency hopping still introduces a relative gain of 30%. Note that a rising channel bandwidth results in more interferers, since the bandwidth per interferer is fixed to 1.8 MHz, which increases the BER. Hence, the unused bandwidth, which is the difference between the channel bandwidth and occupied bandwidth of all users, varies as a sawtooth function.

Our future work will include a combination of random frequency hopping and fractional frequency reuse as well as an analytical model for the transmission power of femtocells.

References

[1] D. Lopez-Perez, I. Guvenc, G. De La Roche, M. Kountouris, T.Q.S. Quek, and J. Zhang. Enhanced Intercell Interference Coordination Challenges in Heterogeneous Networks. *IEEE Wireless Communications*, 18(3):22–30, 2011.

[2] S. Rohde, M. Putzke, and C. Wietfeld. Ad Hoc Self-Healing of OFDMA Networks using UAV-Based Relays. *Ad Hoc Networks (Elsevier)*, http://dx.doi.org/10.1016/j.adhoc.2012.06.014, 2012.

[3] M. Putzke and C. Wietfeld. Self-Organizing Ad Hoc Femtocells for Cell Outage Compensation Using Random Frequency Hopping. In *IEEE 23st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2012.

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

Gabriele Kern-Isberner Joachim Biskup

Knowledge-based Agents and Secrecy

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

A brief summary of the authors work within project A5 is given covering the development of an agent model for confidentiality preservation, adequate belief change operations, the implementation of the model and applications.

Project A5 aims at developing theories of confidentiality for multiagent systems whereby a defending agent $\mathcal{D}$ maintains a view of a potential attacking agent $\mathcal{A}$. Agents shall be knowledge based or epistemic, i. e. equipped with symbolic knowledge representation formalisms and in particular with non-monotonic ones with advanced inference and change operators. This reasoning component shall be embedded in a multiagent system that is based on the Beliefs, Desires, Intentions (BDI) model. This model allows for the design of autonomous agents in dynamic multiagent systems. Epistemic agents with 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.

Realistic scenarios of autonomous intelligent agents in dynamic, uncertain environments, do often not meet the prerequisites for strong global definitions of secrecy. The latter are too strict for such scenarios. In [6] we considered secrecy from the point of view of an autonomous epistemic agent with incomplete and uncertain information which is situated in a multiagent system. The agent pursues its goals by performing actions in its environment which naturally includes communication with other agents. In realistic settings the information to be kept secret is neither global, i. e. uniform, nor static. Secrets are not global in their content as an agent has different secrets with respect to different agents. They are also not global with respect to their strength. That is, an agent wants to keep some information more secret than other. These differences in strength of secrets arise naturally from the value of the secret information which depends on the severeness of the negative effects, or the cost, for the agent resulting from disclosure of the secret information. These costs can differ widely and consequently the agent is interested in not revealing secret information to different degrees. Secrets are also not static, they arise, change and disappear during runtime of an agent such that it has to be able to handle these changes adequately. We developed an adequate notion of secrecy for epistemic agents in multiagent systems that satisfies the requirements laid out above. This notion is based on a complex epistemic state $\mathcal{K}_{\mathcal{D}}$ of an agent. Agent $\mathcal{D}$'s view on the world is given by $V_W(\mathcal{K}_{\mathcal{D}})$. The view agent $\mathcal{D}$ has on the view on the environment of agent $\mathcal{A}$ is given by $V_{\mathcal{A}}(\mathcal{K}_{\mathcal{D}}) \subseteq \mathcal{L}_V$. The secrets of an agent are defined as the set $\mathcal{S}(\mathcal{A}) = \{(\Phi_1, Bel_1, \mathcal{A}_1), \ldots, (\Phi_n, Bel_n, \mathcal{A}_n)\}$ The intuitive semantics of a secret is that if agent $\mathcal{D}$ holds the secret $(\Phi, Bel_{\mathcal{A}}, \mathcal{A})$, it does not want that agent $\mathcal{A}$ believes Φ by use of the belief operator $Bel_{\mathcal{A}}$, i. e. $\Phi \notin Bel_{\mathcal{A}}(V_{\mathcal{A}}(\mathcal{D}))$. This formulation allows for the specification of agent specific secrets with varying strength. The strength is defined by the chosen belief operator from a *Belief-Operator Family* which is a complete lattice $(\mathcal{B}, \preceq_{\mathcal{B}})$ of belief operators. The demanded dynamics of secrets is realized by use of a change operator $\circ$ which changes an epistemic state $\mathcal{K}$ given some information τ, such that $\mathcal{K} \circ \tau = \mathcal{K}'$. We instantiated and compared this framework to state of the art frameworks and especially to [3], [1] and [2]. We showed that our framework generalizes many other notions of secrecy and satisfies the requirements postulated initially. Extensions of the epistemic secrecy framework published in [6] are currently being developed. One extension realizes a full fledged BDI agent model for confidentiality preservation. Hereby we continue the epistemic approach towards agents and treat the agent's desires, intentions and know-how as part of the epistemic state of the agent, i. e. on the same level as its beliefs. An agent should be able to reason about its intentions and plans, in our case in particular with respect to confidentiality concerns. To this end we developed a functional component of an agent allowing for this type of reasoning. Another extension is considering the aspect of uncertainty of secrecy in multiagent scenarios with incomplete and uncertain information. Hereby the interrelation of different dimensions of uncertainty are investigated within the framework of [6] such that complex composed belief operators reflecting different aspects of confidentiality can be used in the framework. This composition allows for a granular specification of levels of protection of a secret.

The concept of change operations for confidentiality is crucial in the dynamic scenarios we are considering. In particular a strong foundation of change operations for non-monotonic logics is needed. A focus of this project is the use of Answer Set Programming (ASP) as one candidate for non-monotonic knowledge representation. ASP allows 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 theory of classic belief revision has well formulated notions of theory change. In general the object of change are belief sets or belief bases, i. e. sets of classical propositional sentences which are deductively closed or not, respectively. Belief bases especially, are interesting for an efficient implementation. For both well defined sets of rationality postulates for different change operations have been defined. For logic programs most approaches to dynamics are very pragmatic and lack a formal representation of requirements of the change process. Few works examined the formal properties of approaches to change logic programs and the connection to belief base revision has not been considered in detail yet. In [5] we presented a base revision description and construction for belief bases represented by extended logic programs under the answer set semantics. We defined a *multiple base revision operator* $* : \mathcal{P}_{\mathcal{A}} \times \mathcal{P}_{\mathcal{A}} \to \mathcal{P}_{\mathcal{A}}$ which is applicable to ASPs. For this problems involving the lack of a definition of negation, disjunction or deductive closure of programs and the effects of the non-monotonic semantics had to be handled appropriately. The results presented in this paper show that the base revision approach to belief revision is applicable to revision of logic programs and formalizes adequate properties such an operation should satisfy. We defined new postulates and a construction of an operator which satisfies all proposed postulates. This work, in combination with previous works, e.g. [4], forms a strong basis for the implementation of adequate resource bounded inference and change operations for our confidentiality scenario.

In [7] and [8] we considered an extended concept of belief revision in which the process of belief change is seen as a two step process, such that $\mathcal{B}_{\mathcal{D}} \circ \Phi = \mathcal{B}_{\mathcal{D}} * f_{\mathcal{B}_{\mathcal{D}}}(\Phi)$ with a transformation function $f_{\mathcal{B}_{\mathcal{D}}}$ and some (prioritized) multiple base revision $*$ operator. First the new information is evaluated by $f_{\mathcal{B}_{\mathcal{D}}}$ based on its credibility and the agents current beliefs, then the belief base of the agent is revised by the result of the evaluation process. For the evaluation process we considered an argumentation theoretic approach in a multiagent scenario and showed that it satisfies various desirable properties.

We implemented a multiagent system framework called *Angerona* which is based on a versatile plugin architecture. Agents within this framework are epistemic BDI agents based on [6] and its BDI extension whereby both, the knowledge representation and concrete agent cycle are flexible. The knowledge representation is based on the interfaces from the *Tweety*[1] library and thereby allows for the use of a variety of formalisms. The plugin architecture allows us to easily define and compare different types of agents and evaluate their performance. The ASP library has been greatly extended in the *Tweety* library and used to implement an ASP plug-in for *Angerona*. An extended BDI cycle for secrecy preservation based on [6] and its extensions have been implemented as well as an ASP plug-in realizing different belief and change operators. As part of the RISE internship program of the DAAD an intern from the USA was modeling and evaluating secrecy scenarios using *Angerona* for three month.

A cooperation with Project B4 has been initialized in which privacy issues of Project B4 are tackled with the methods of Project A5. Vehicles providing floating car data are modeled as agents in a multiagent system and implemented using the *Angerona* system in combination with a traffic simulator. The multiagent approach allows for a preprocessing of the floating car data such that the resulting data send to the traffic prognosis server is highly anonymized. This preprocessing would not be possible for

[1] http://tweety.sourceforge.net/

a single agent. Current work aims at the elaboration of adequate confidentiality and availability policies, simulation, evaluation and comparison to the state of the art.

In summary, various aspects for the development of confidentiality preserving agents have been investigated and elaborated on a theoretical and practical level. First theoretical results have been shown and a promising prototype system have been implemented. We will continue the started threads and complete and extend the ASP instantiation of the framework.

References

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

[2] Joachim Biskup and Cornelia Tadros. Policy-based secrecy in the runs & systems framework and controlled query evaluation. In *Proc. of the 5th Int'l Workshop on Security (IWSEC 2010)*. Information Processing Society of Japan (IPSJ), 2010.

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

[4] Patrick Krümpelmann. Dependency semantics for sequences of extended logic programs. *Logic Journal of the IGPL*, doi: 10.1093/jigpal/jzs012, 2012.

[5] Patrick Krümpelmann and Gabriele Kern-Isberner. Belief base change operations for answer set programming. In *Proceedings of the 13th European Conference on Logics in Artificial Intelligence (JELIA'12)*, volume 7519 of *Lecture Notes in Artificial Intelligence*. Springer, 2012.

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

[7] Patrick Krümpelmann, Matthias Thimm, Marcelo A. Falappa, Alejandro J. Garcia, Gabriele Kern-Isberner, and Guillermo R. Simari. Selective revision by deductive argumentation. In *Formal Argumentation - First International Workshop on Theory and Application, (TAFA'11), 2011, Revised Selected Papers*, volume 7132 of *Lecture Notes in Computer Science*, page 281, 2012.

[8] Luciano H. Tamargo, Alejandro J. Garcia, Matthias Thimm, and Patrick Krümpelmann. Argumentative credibility-based revision in multi-agent systems. In *Proceedings of the 13th Argentine Symposium on Artificial Intelligence (ASAI 2012)*. 2012.

On the Inference-Proofness of Materialized Views and their Generation

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 creating an approach to generate so-called inference-proof materialized views with the help of refusals is outlined. In this context the progresses made are discussed as well as the loose ends to which attention will be drawn next year. Finally, it is discussed briefly how the goals formulated in last year's technical report have been implemented within last year's research or why they could not have been implemented as proposed.

My research during the last year again dealt with the generation of inference-proof materialized views of database instances. Given an original database instance, an inference-proof materialized view of this original instance is an alternative database instance not containing any information which has to be kept secret according to a confidentiality policy – neither directly nor indirectly by enabling a user to deduce such information by drawing inferences, possibly with the help of his a priori knowledge. Having done own research about inference-proof materialized views with the help of fragmentation of database instances the year before [6] and knowing about an approach to create inference-proof materialized views with the help of lies [7], the next idea was to develop an approach to create inference-proof materialized views with the help of refusals.

To develop such an approach, considerations about the modelling of refusals within a materialized view have to be made. Beside the possibility to suppress certain components of tuples of the original instance (similar to the approach dealing with fragmentation discussed in [6]), another possibility is to refuse the truth values of certain (complete) tuples of the original instance, which might enable a user to infer some knowledge to be kept secret. At the conceptual level a refusal of a truth value of a tuple can be expressed by extending the set of truth values of the original instance by a distinguished truth value denoting a refusal. This technique is known from previous approaches to establish inference-proofness by distorting answers to queries dynamically at run-time with the help of refusals [2,3], in which the distinguished truth value "mum" is given as the answer to harmful (closed) queries. Similarly, in a materialized view a tuple is associated with the distinguished truth value "refused", if the truth value associated with this tuple in the original instance must not be revealed to the user. So, although being aware that "refused" is not the original truth value of a refused tuple, a user should not be able to determine the original truth value of this tuple.

At the operational level it is necessary to decide for which tuples the corresponding truth values should be refused. While in terms of availability a minimum number of refusals is favoured, in terms of confidentiality no combination of non-refused tuples may offer the possibility to infer information to be kept secret. If the confidentiality policy is restricted to potential secrets expressed by select-project sentences and the user's (assumed) a priori knowledge is restricted to functional dependencies and full join dependencies (cf. [1]), one possibility to compute all options to infer harmful knowledge with the help of so-called template dependencies is presented in [5]. If this approach is adapted such that only fully instantiated template dependencies are computed, a set of hypothesis rows of such a template dependency corresponds to a harmful combination of tuples – i.e. a set of tuples enabling a user to infer knowledge which has to be kept secret – and at least one tuple of such a set has to be refused.

As different template dependencies might have overlapping sets of hypothesis rows, it is desirable in terms of availability to determine a minimum set of tuples containing at least one hypothesis row of each set of hypothesis rows. This problem is proved to be equal to the well known NP-complete combinatorial optimization problem of finding a "Minimum Hitting Set" [8] and can therefore be modelled as a so-called Integer Linear Program (ILP). Moreover, if not all tuples of the original database instance are of the same importance and less important tuples should be refused preferably to increase availability, this can be achieved by setting an importance weight for each tuple of the original database instance and by slightly modifying the objective function of the ILP.

Regarding efficiency the generation of inference-proof materialized views has the advantage, that queries can be answered safely without employing costly mechanisms of (dynamic) inference control at run-time: 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 maybe this complexity can be reduced with the help of approximation algorithms (e.g., for computing a "Minimum Hitting Set").

Unfortunately, this promising approach is not ready for publication, yet. Until now it is not examined satisfactorily how the semantics of functional dependencies and full join dependencies changes when extending the set of truth values of the original instance by an additional truth value denoting refusals. Depending on these results the semantics of the original database instance and maybe also the exact modelling of refused database tuples may have to be adapted suitably in order to form a coherent overall approach. For this research the ideas proposed in [9] might provide a possible basis. Moreover, it should be analysed whether availability can be increased by reducing the number of refusals without compromising confidentiality. In this context the number of so-called additional refusals to eliminate so-called meta-inferences is of interest. Deduction of knowledge is called a meta-inference if a user succeeds in exploiting an explicit refusal notification to infer information to be kept secret by simulating the behaviour of the algorithm used for generating the inference-proof views [2]. Of course, the elimination of those meta-inferences must be analysed with scrutiny.

As already announced in last year's technical report, during the last year the inference-proofness of fragmentation again was a topic of research. In [6] only one specific approach to vertical fragmentation – splitting a relational instance into one externally stored part and one locally-held part – was considered. But as there are more approaches to achieving confidentiality by vertical fragmentation surveyed in [10], analysing the inference-proofness of these approaches is of interest, too. 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 is currently 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 within the framework of controlled query evaluation are supposed to be similar to the ones presented in [6].

Although in last year's technical report it was proposed that my research for the last year should have dealt with inference-proof materialized views created with existing algorithms for dynamic controlled query evaluation surveyed in [2, 4], this research has to be postponed until next year. Important parts of this research should have been based on experimental evaluations by processing the (open) identity query asking for a full relational instance with the help of a prototype for controlled query evaluation. Though meanwhile this prototype is able to handle open queries, the processing of those open queries needs too much time to compute examples of relevant size because of the high complexity of theorem-proving on which dynamic controlled query evaluation is based. So, hopefully some optimizations can be made to increase the efficiency of this prototype, such that inference-proof variations of an original instance can be computed by using algorithms for dynamic controlled query evaluation.

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 and Piero A. Bonatti. Controlled query evaluation for enforcing confidentiality in complete information systems. *International Journal of Information Security*, 3(1):14–27, 2004.

[4] Joachim Biskup and Piero A. Bonatti. Controlled query evaluation with open queries for a decidable relational submodel. *Annals of Mathematics and Artificial Intelligence*, 50(1–2):39–77, 2007.

[5] Joachim Biskup, Sven Hartmann, Sebastian Link, Jan-Hendrik Lochner, and Torsten Schlotmann. Signature-based inference-usability confinement for relational databases under functional and join dependencies. In Nora Cuppens-Boulahia, Frédéric Cuppens, and Joaquín García-Alfaro, editors, *Data and Applications Security and Privacy XXVI, DBSec 2012*, volume 7371 of *Lecture Notes in Computer Science*, pages 56–73. Springer, 2012.

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

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

[8] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1979.

[9] Leonid Libkin. Incomplete information and certain answers in general data models. In Maurizio Lenzerini and Thomas Schwentick, editors, *Principles of Database Systems, PODS 2011*, pages 59–70. ACM, 2011.

[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 Revision as Nonmonotonic Reasoning under Confidentiality Constraints

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

In multiagent systems, several agents (i.e., autonomous computing systems) share information for the purpose of achieving a joint goal, e.g., a sale contract, arrangement of a meeting etc. Whilst sharing of information is a necessary means for the cooperation among the agents it is subject to obligations or interests of individual agents to hide sensitive information from others. In our ongoing work in project A5 we augmented an agent with additional components for declaring its confidentiality interests and, complementary, components for controlling its interaction with other agents and effectively enforcing its declared interests. As one focus of my own research I studied belief revision under confidentiality constraints where revision is considered as the process of nonmonotonic reasoning from the information available to the agent. To provide for a flexible implementation of an agent, its nonmonotonic reasoning may be an instance of any class of nonmonotonic consequence relations with appropriate restrictions. Based on axiomatizations of such classes, as one main result of the joint work with my supervisor we give means to simulate the skeptical reasoning of one agent about sensitive belief of another agent.

Like in my last year's research, my work focused on a scenario of an isolated interaction between two agents, a requesting agent $\mathcal{A}$ and a reacting agent $\mathcal{D}$, outlined by Fig. 1. The exchange of information relies on a common propositional language $\mathcal{L}_{pl}$. In this scenario, agent $\mathcal{D}$ has the *role of a defender* against agent $\mathcal{A}$ having the *role of a (potential) attacker* who attempts to obtain confidential information against $\mathcal{D}$'s obligations

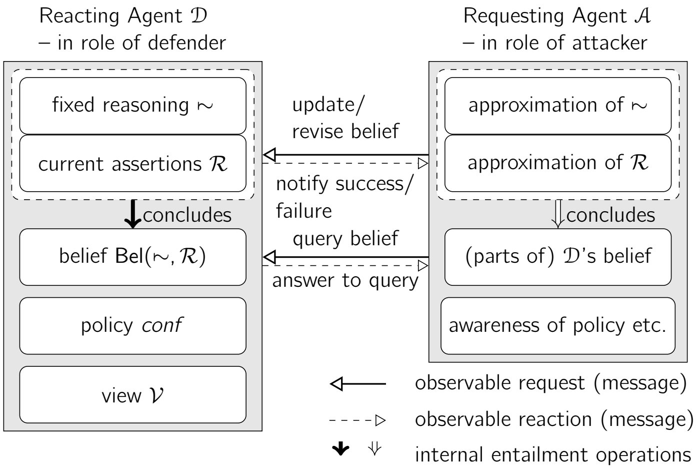

Reacting Agent $\mathcal{D}$
– in role of defender
fixed reasoning $\mid\!\sim$
current assertions $\mathcal{R}$
concludes
belief $\mathsf{Bel}(\mid\!\sim, \mathcal{R})$
policy *conf*
view $\mathcal{V}$
update/
revise belief
notify success/
failure
query belief
answer to query
Requesting Agent $\mathcal{A}$
– in role of attacker
approximation of $\mid\!\sim$
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

or interests. Agent $\mathcal{D}$ is devised with the usually desired functionality of an interacting agent: $\mathcal{D}$ faces generally incomplete information (*current assertions* $\mathcal{R} \subset_{fin} \mathcal{L}_{pl}$) about its environment, e.g., trading stock and offers in an e-commerce scenario, etc. Moreover, in order to plan how to achieve its personal and the joint goals and to guide its decisions in the process of planning, agent $\mathcal{D}$ must be capable to draw reasonable conclusions from the available information $\mathcal{R}$. This capability is modeled by a *consequence relation* (*fixed reasoning* $\mid\!\sim\; \subseteq \mathcal{L}_{pl} \times \mathcal{L}_{pl}$) that maps assertions to conclusions. Gathering all conclusions from its current assertions, $\mathcal{D}$ forms its *belief* $\mathsf{Bel}(\mid\!\sim, \mathcal{R})$ about the environment: $\mathsf{Bel}(\mid\!\sim, \mathcal{R}) := \{B \in \mathcal{L}_{pl} \mid \mathsf{con}(\mathcal{R}) \mid\!\sim B\}$ with $\mathsf{con}(\mathcal{R})$ denoting the conjunction of the assertions $\mathcal{R}$.

In the primary work [4] with my supervisor, we implement agent $\mathcal{D}$'s consequence relation by the very general semantic approach of ordinal conditional functions (OCF) [1, 7]. In our subsequent work [3], we extend this scenario in the way that the designer of agent $\mathcal{D}$ may implement its reasoning component with a structure such as OCFs or others suiting the functionality of the agent. The kind of structures chosen by the designer usually defines a class $\mathcal{C}$ of consequence relations. The members of this class are all consequence relations that can be respresented by these structures. In both works, we assume that the requesting agent $\mathcal{A}$ is a aware of the kind of the structure used in the implementation. This assumption follows the common rationale in computer security of “No security by obscurity” because the implementation of agent $\mathcal{D}$ may be publicly known (except, of course, the concrete instances $\mid\!\sim$ and $\mathcal{R}$ of its belief components). The next paragraphs will review these works.

In both works, we elaborate on the previously outlined scenario and consider that agent $\mathcal{A}$ requests answers to 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 process of nonmonotonic reasoning from the assertions $\mathcal{R} \cup \{A\}$ [5].

For confidentiality preservation, agent $\mathcal{D}$ is devised with two additional components, the policy *conf* (declaration of confidential belief) and the view $\mathcal{V}$ (agent $\mathcal{A}$'s presumable view on $\mathcal{D}$'s fixed reasoning and current assertions), following the proposal in [2] for BDI agents. These components are used by the control procedures to enforce $\mathcal{D}$'s confidentiality interests. In [3], the procedures additionally need an axiomatization of the class $\mathcal{C}$ of consequence relations underlying $\mathcal{D}$'s fixed reasoning.

Confidentiality guarantees can only be made under assumptions about agent $\mathcal{A}$'s capabilities and behavior when reasoning about $\mathcal{D}$'s sensitive belief. In both works, we assume agent $\mathcal{A}$ to be a skeptical reasoner about sensitive belief so that it desires to conclude with certainty that $\mathcal{D}$ believes a fact contained in the confidentiality policy *conf*. Agent $\mathcal{A}$'s skeptical entailment (1) is 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 $\mathcal{A}$.

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

Informally, $\mathcal{A}$ considers a consequence relation possible iff this relation agrees with its approximations and is an instance of class $\mathcal{C}$.

In the primary work [4], we consider the class $\mathcal{C}^{\kappa}$ that defines nonmonotonic consequence relations by ordinal conditional functions κ. The key issue of skeptical reasoning is that agent $\mathcal{D}$ might have hidden some of its current assertions from agent $\mathcal{A}$. Agent $\mathcal{A}$ must take hidden assertions into consideration in its process of reasoning about sensitive belief. We show that the skeptical entailment operator under the hidden assertions assumption can be reduced to the flat operator (1) where we use semantic arguments, constructing an appropriate ordinal conditional function as a witness of non-entailment. The flat skeptical entailment operator is thus the essential means of the control procedures the effectiveness of which in enforcing the confidentiality policy we formally prove.

In the subsequent work [3], we consider that $\mathcal{D}$'s fixed reasoning is implemented as an instance of a general class $\mathcal{C}$ of consequence relations. However, such classes should have an axiomatization that allows to establish our results on the reduction of skeptical entailment to the flat operator and on the effectiveness of the control procedures for this general case. For this purpose, we characterize appropriate axiomatizations in a conditional language [6] by imposing simple syntactical restrictions on the shape of schema formulas and by requiring some basic axiom schemes. When implementing agent $\mathcal{D}$, the agent's designer may choose some well-known class of nonmonotonic consequence relations and some of its axiomatizations from the literature and show that the axiomatization meets our restrictions and requirements. This way, existing decidability and complexity results for these classes may be applied to the computation of skeptical entailment as well as existing feasible solvers for deduction under the chosen axiomatization.

In my ongoing research, the agents will be implemented with the outlined features of nonmonotonic reasoning under confidentiality constraints in the multiagent Java-based platform Angerona developed in the collaboration of this project. Further, I started studying the enhancement of the revision process by merging assertions under confidentiality constraints. Merging removes contradictions from a set of assertions by dropping assertions with least priority. With merging, more complicated notifications about its results must be controlled likely resulting in more involved control procedures. Further, in the collaboration of this project we work on a richer language for declaring the confidentiality policy and its semantics. In particular, sensitive pieces of information may need different levels of protection. The level of protection is modelled by the nature of the operator by which $\mathcal{A}$ reasons about sensitive information. Skeptical entailment provides the least level of protection.

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

[3] Joachim Biskup and Cornelia Tadros. Revising belief by nonmonotonic reasoning without revealing secrets, 2012. Submitted to AMAI special issue for FoIKS 2012.

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

[5] Didier Dubois. Three scenarios for the revision of epistemic states. *J. Log. Comput.*, 18(5):721–738, 2008.

[6] Nir Friedman and Joseph Y. Halpern. Plausibility measures and default reasoning. *J. ACM*, 48(4):648–685, 2001.

[7] Wolfgang Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In Brian Skyrms and William L. Harper, editors, *Irvine Conference on Probability and Causation*, volume II of *Causation in Decision, Belief Change, and Statistics*, pages 105–134. Kluwer, Dordrecht, 1988.

Subproject B1
Analysis of Spectrometry Data with Restricted Resources

Sven Rahmann Jörg Ingo Baumbach

Pairwise Alignment of MCC/IMS measurements

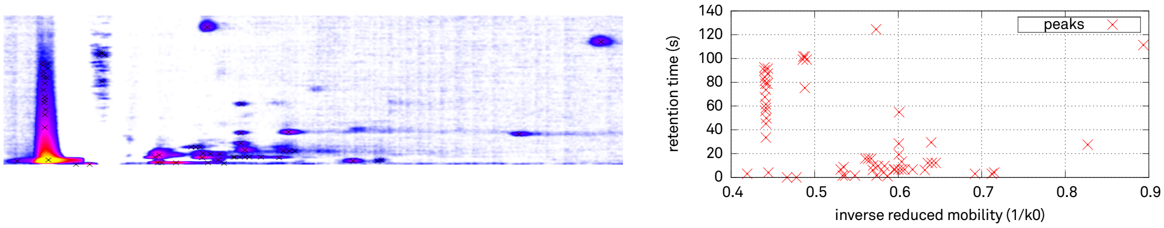
Marianna D'Addario
Computer Science XI
Technische Universität Dortmund
marianna.daddario@tu-dortmund.de

The interest in noninvasive breath analysis in medical prognostics is increasing due to an innovative technique that couples Ion Mobility Spectrometry (IMS) with a Multicapillary Column (MCC) [2]. This report summarizes the efforts to automate the pairwise alignment of MCC/IMS measurements, as part of the project TB1.

The MCC/IMS devices measure volatile organic compounds (VOCs) in the air or in exhaled breath. The data produced by these devices is visualized as a heat map, shown in Figure 1(a), where every elevated area indicates a measurable intensity of a organic compound and is called a peak.

The MCC/IMS measurements are already used to identify patterns that signalize known diseases, e.g. lung cancer or diabetes [7], [1]. The need emerges to compare measurements taken from different patients, at different times and with different devices of different ages. In consequences of these differences, measurements have to be aligned before the data can be compared.

One attempt is the pairwise alignment of a reference peak set and a new MCC/IMS measurement which we refer to as data peak set. The *reference peak set* contains the positions in terms of inverse reduced mobility and retention time of known VOCs. The *data peak set* contains every detected peak of a measurement and can be computed by an arbitrary method. Generally, a *peak set* contains for every peak of an MCC/IMS measurement several parameters depending on the used peak detection method, e.g. volume, shape and the peak's mode that is the point within a peak of highest intensity. However, the peak's mode is always contained in a peak set. Therefore, the presented method is independent of the peak detection method and uses only modes of peaks to determine the alignment.



(a)

(b)

Figure 1: An MCC/IMS measurement and its peaks: (a) MCC/IMS heat map (×: peak modes), (b) MCC/IMS peak modes

The general idea of the pairwise alignment is to compute a matching between the modes of the two peak sets. We consider the peaks' modes as points in a coordinate system with inverse reduced mobility ι on the x-axis and retention time r on the y-axis (Figure 1(b)). Hence, we refer to the modes of the reference peak set as *reference points* (P_{ref}) and to the data peak set ones as *data points* (P_{data}).

The pairwise alignment results in an assignment of data points to reference points. Note that the number of points in the sets differs and not every data point can be assigned to a reference. Moreover, it is possible that no assignment can be found hence no VOC appears in both measurements, except for two markers that are present in every measurement due to the used technique. A first step is to adjust P_{data} regarding to these known markers. Due to noise and external influences this adjustment is not yet sufficient to deliver a matching. Therefore, a distance is calculated on which a statement can be done on whether a point of P_{data} can be assigned to one of P_{ref} or not. Then we compute a score regarding the distance of the points in an L^p space. Finally, to achieve the matching we use the *Hungarian Algorithm* for the previous calculated distances.

Adjustment of $\mathbf{P_{data}}$ The points of P_{data} are adjusted with respect to the reactant ion peak (RIP) and the benzothiazole peak. The RIP and the benzothiazole peak are present in every MCC/IMS measurement, since the first is caused by the cleaning gas (nitrogen) and the second by evaporation from the plastics compounds of the device.

A shift s_ι in reduced inverse mobility and linear factor c for the retention time are calculated. This shift is computed as difference between the inverse mobilities of the RIP. Let ι_{rRIP} be the reduced inverse mobility of the reference's RIP and ι_{dRIP} that of the data's RIP, then $s_\iota = \iota_{\text{dRIP}} - \iota_{\text{rRIP}}$.

According to Perl et al. [6] and Cumeras et al. [3] for the retention time a linear alignment is appropriate. Let r_{rBenzo} be the retention time of benzothiazole in P_{ref} and r_{dBenzo} that of benzothiazole in P_{data}, then we gain the factor $c = r_{\text{dBenzo}}/r_{\text{rBenzo}}$.

The shifted data points P'_{data} have coordinates $(\iota - s_\iota, r/c)$. These coordinates are the positions where the points of P_{ref}, if any present in reference measurement, are expected.

L^p norm score function In a scaled L^p space, the distance between two points $v_1 = (\iota_1, r_1) \in P_{\text{ref}}$ and $v_2 = (\iota_2, r_2) \in P'_{\text{data}}$ is defined as:

$$d_p(v_1, v_2) = \left(\left(\frac{\iota_1 - \iota_2}{a} \right)^p + \left(\frac{r_1 - r_2}{b} \right)^p \right)^{1/p}$$

We scale each dimension by adding a and b to the calculation of d_p and assume the independence of ι and r.

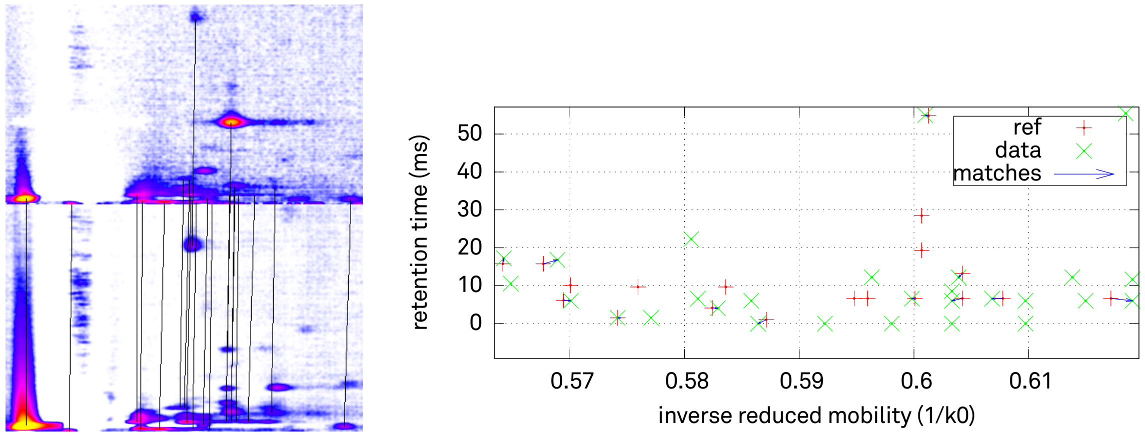
The score of two points $v_1 \in P_{\text{ref}}$ and $v_2 \in P'_{\text{data}}$ is calculated using the defined distance d_p. All points of P'_{data} that have $d_p(v_1, v_2) \leq 1$ to a point of P_{ref} should have positive score, while those with distance $d_p(v_1, v_2) > 1$ should have zero as score. The latter case indicates that a matching is not feasible. The score function $s(v_1, v_2)$ should return the maximum score s_{max} if the distance is $d_p(v_1, v_2) = 0$. The score is computed as follows:

$$s(v_1, v_2) = (s_{\text{max}})^{1 - d_p(v_1, v_2)}$$

Hungarian Algorithm The Hungarian Algorithm solves the assignment problem in polynomial time [5]. Considering a number of agent and a number of tasks, any agent can complete any task with some defined costs. The assignment problem consists of assigning exactly one agent to every task so that the sum of the upcoming costs is minimal. Formally, given two sets, A and T, and a cost function $c : A \times T \to \mathbb{R}$ find a bijection $f : A \to T$ such that the sum of the costs $\sum_{a \in A} c(a, f(a))$ is minimized.

The assignment problem properly fits our needs to match the reference points to the data points using the above presented score function and maximizing the scores instead of minimizing costs. We solve the assignment problem given the two sets P_{ref} and P'_{data}. The point set with fewer points among P'_{data} and P_{ref} is chosen as task set, given that every task must be assigned. Finally, we ignore the matches with score zero. Using the open source library *dlib* [4], the problem is solved with the `max_cost_assignment` function, which implements the Hungarian algorithm.

Conclusion The Hungarian algorithm optimizes the global alignment considering for every point of P_{ref} the nearest point of P'_{data}. As can be seen from the heat map visualization in Figure 2, the found alignment is reasonable. The next step is to explore the L^p norm for different p and to check which norm is most appropriate to compute the alignment of VOCs. Considering the amount of data this could be a good opportunity to cooperate with project A2.



(a) (b)

Figure 2: Alignment by Hungarian algorithm with $p = 2$ (Euclidean distance). (a) peaks connected by lines are matched (b) subsection of the alignment as scatter plot

References

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

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

[3] R. Cumeras, T. Schneider, P. Favrod, E. Figueras, I. Gracia, S. Maddula, and JI Baumbach. Stability and alignment of MCC/IMS devices. *International Journal for Ion Mobility Spectrometry*, pages 1–6, 2012.

[4] King, Davis. dlib C++ library. `http://dlib.net/`, February 2012.

[5] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[6] T. Perl, B. Bödeker, M. Jünger, J. Nolte, and W. Vautz. Alignment of retention time obtained from multicapillary column gas chromatography used for VOC analysis with ion mobility spectrometry. *Analytical and bioanalytical chemistry*, 397(6):2385–2394, 2010.

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

Supervised statistical learning of metabolic ion mobility spectrometry profiles for chronic obstructive pulmonary disease

Anne-Christin Hauschild
Department Computational Systems Biology,
Max Planck Institute for Informatics, Saarbrücken, Germany
Department Microfluidics and Clinical Diagnostics,
KIST Europe, Saarbrücken, Germany a.hauschild@mpi-inf.mpg.de

Exhaled air carries information on human health status. Ion mobility spectrometers combined with a multi-capillary column (MCC/IMS) is a well known technology for detecting volatile organic compounds (VOCs) within human breath. This technique is relatively inexpensive, robust and easy to use in every day practice. However, the potential of this methodology depends on successful application of computational approaches for finding relevant VOCs and classification of patients into disease-specific profile groups based on the detected VOCs. We developed an integrated state-of-the-art system using sophisticated statistical learning techniques for VOC-based feature selection and supervised classification into patient groups. We analyzed breath data from 84 volunteers, each of them either suffering from chronic obstructive pulmonary disease (COPD), or both COPD and bronchial carcinoma (COPD+BC), as well as from 35 healthy volunteers, comprising a control group (CG). We standardized and integrated several statistical learning methods to provide a broad overview of their potential for distinguishing the patient groups. We found that there is strong potential for separating MCC/IMS chromatograms of healthy controls and COPD patients (best accuracy COPD vs. CG: 94%). However, further examination of the impact of bronchial carcinoma on COPD/no-COPD classification performance is necessary (best accuracy CG vs. COPD vs. COPD+BC: 79%).

1 Introduction

Multi-capillary column-ion mobility spectrometry (MCC/IMS) is a comparatively inexpensive, sensitive high through-put method to analyze human exhaled air carrying information about health status. The resulting MCC/IMS chromatograms, contain this information. Sophisticated computational approaches can be utilized for classifying patients into disease-specific profile groups and identifying the volatile organic compounds (VOCs) that are important.
First, a brief introduction to chronic obstructive pulmonary disease (COPD) is provided. Various preprocessing steps for the analysis of MCC/IMS data and the different machine learning methods applied in this study are also elucidated. Finally, results obtained from various machine learning methods are presented and followed by a discussion and comparison with the state-of-the-art techniques. For more details on the MCC/IMS technique, we refer to the technical report of Kathrin Rupp.
COPD: COPD is an inflammatory lung disease characterized by a permanent blockage of airflow from the lungs. The primary cause of COPD is tobacco smoke (through smoking or second-hand smoke). The disease is widely under-diagnosed, although it is a life-threatening lung disease which is not fully reversible. The World Health Organization (WHO) reported it to be one of the most frequent causes of death. According to a WHO report in 2008, an estimated 64 million people worldwide suffered from COPD in 2004, and more than 3 million people died of COPD in 2005 [4].

2 Material and Methods

The data set consists of three different groups of volunteers: healthy controls (class: HC), COPD patients (class: COPD), and COPD with bronchial carcinoma patients (class: COPD+BC). It was preprocessed utilizing the VisualNow software, followed by an expert-driven component detection. This included standard methods for preprocessing and data reduction, e.g denoising, smoothing and peak detection, such as log-normal detailing and wavelet transformation. We used several standardized statistical supervised learning methods in this project, to get a broad overview of the potential of the data and the different classification techniques. Thus, we included some rather simple methods, decision tree, naive Bayes, and linear support vector machine (SVM), for instance. On the other hand, we used more recent and sophisticated techniques, such as neural net, random forest and radial SVM. The R language package (Version 2.13.1) was used to implement the statistical analysis and feature selection [1]. The accuracy of the different statistical learning techniques was evaluated in a 10-fold cross validation environment. In settings where the data set is small, in our case restricted to 119 samples, a simple splitting into training and test set can lead to relatively noisy estimates of predictive performance.

Therefore, cross validation is used to give an estimate for the actual accuracy of the predictive model. To ensure that each subset covers the variety of all classes, the classes are balanced within each subset, for the two-class as well as the three-class-problem. To assess the information content within the breath data and to avoid overtraining the statistical learning methods, simpler methods as well as more sophisticated methods were applied without further tuning of the parameters.

3 Results and Discussion

The more simplistic methods, i.e., decision tree and naive Bayes, achieved an accuracy between 82% and 85% and an AUC of around 80%. The linear SVM performed slightly better with an AUC of 83% and an accuracy of around 87%. While the more sophisticated methods, i.e., neural net and radial SVM, gave an accuracy of 89% and AUCs of 86% and 87%, respectively. The best performing method of the classification, distinguishing between COPD patients, was random forest, which had the best prediction accuracy of 94% as well as high values for AUC (92%), sensitivity (98%) and specificity (86%). As expected, the more sophisticated methods performed best, having a relatively low bias, which means they do infer less than the simpler methods. On the other hand, the basic methods performed surprisingly well in terms of AUC and accuracy, which indicates that the data provided some information to distinguish the two classes. However, due to the unbalanced data set (COPD $\approx$ 70% vs. HC $\approx$ 30%), one has to take a closer look at the sensitivity and specificity. While the sensitivity of both types of methods was good (between 87% and 98%), the specificity of the enhanced methods (80% to 85%) was in general higher than the specificity of the simplistic methods (71% to 74%).

Prediction quality of the three-class problem was comparably low (accuracy $\approx$ 70%) for each of the applied machine learning techniques, except random forest (accuracy $\approx$ 79%). The AUC dropped by at least ten percent for all of the methods except naive Bayes, which may be due to its simplicity and its robustness. Therefore, it remains unclear whether the data's information content is high enough for distinguishing the three groups of volunteers. All tested methods showed a very low sensitivity for the COPD class in contrast to a high sensitivity for the BC class, which indicates that the differentiation between class COPD and COPD+BC is difficult. In fact most of the measurements of COPD patients were falsely predicted to suffer from both COPD and bronchial carcinoma, which might be reducible to the characteristic of COPD as a common and important independent risk factor for lung cancer.

Other studies, e.g. Baumbach *et al.* [2], Finthammer *et al.* [3] or Westhoff et al. (2011) [5], that also used supervised statistical learning or relational probabilistic learning methods, showed a much better prediction performance for classification of bronchial carcinoma or COPD. Nevertheless, despite the good results, one has to consider that (1) the prediction was done on a comparatively large feature set, and (2) the accuracy and AUC were evaluated on the training set, without cross validation.

3.1 Summary and Conclusion

Ion mobility spectrometry data of human breath can generally be utilized for distinguishing between lung diseases if used properly with statistical learning environments. To demonstrate this, we evaluated sophisticated machine learning techniques on MCC/IMS chromatograms regarding their classification performance and ability to identify the most important molecular compounds. Therefore, the breath of 84 patients either suffering from COPD or both COPD and bronchial carcinoma was processed and compared with 35 healthy volunteers. The by far best test error estimates (AUC 91%, ACC 94% for COPD vs. HC; AUC 79%, ACC 67% for COPD vs. COPD+BC vs. HC) were achieved with the random forest method. These results pinpoint a strong potential to separate healthy from COPD, but also suggest that a further examination of the differences between COPD and COPD+BC is needed. In the future, we plan to optimize and enhance the standardized learning methods for enhancing prediction performance, on the one hand, and the identification of the smallest discriminating set of molecules, on the other hand. This will be a tremendous progress in the field of COPD and cancer diagnostics. However, larger COPD+BC data sets are necessary here.

References

[1] R: A language and environment for statistical computing, April 2011.

[2] J. Baumbach, A. Bunkowski, S. Lange, T. Oberwahrenbrock, N. Kleinboelting, S. Rahmann, and J.I. Baumbach. Ims2 – an integrated medical software system for early lung cancer detection using ion mobility spectrometry data of human breath. *Journal of Integrative Bioinformatics*, 4(3):75:1–12, 2007.

[3] M. Finthammer, Chr. Beierle, J. Fisseler, G. Kern-Isberner, B. Mller, and J.I. Baumbach. Probabilistic relational learning for medical diagnosis based on ion mobility spectrometry. *Int. J. Ion Mobility Spectrom.*, 13(2):83–92, 2010.

[4] World Health Organization. The global burden of disease, 2004 update, 2008.

[5] M. Westhoff, P. Litterst, S. Maddula, B. Bödecker, and J.I. Baumbach. Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (copd) including lung cancer from healthy control by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, pages 139–149, 2011.

Intensity Reconstruction in Ion Mobility Spectrometry Measurements

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

To measure the presence and concentration of compounds and especially volatile organic compounds (VOC) in exhaled breath, ion mobility spectrometry (IMS) becomes more and more attractive due to its construction and working properties. A problem that occurs is that (provided by the ionization source an IMS device is using) only relative intensities of peaks within a spectrum are captured. We present a novel method to reconstruct the intensities by modeling the spectrum with a mixture of weighted inverse Gaussian distributions.

Overview

Pre-separating analytes with a multicappilary column (MCC) before they enter the IMS device yields in an extremely increased resolution and accuracy for detecting compounds. The technique of an IMS device is explained by Baumbach et al. [1]. Because of equidistant distances of capturing in retention time r (time to cross the MCC) as well as in drift time t (time to cross the IMS device) we assume for all retention times R and drift times T that $R = \{1, \dots, |R|\}$ and $T = \{1, \dots, |T|\}$. We obtain a two-dimensional matrix, which can be visualized as a heat map, see Figure 1. The presence of a certain compound provides a high signal region called a peak. To get rid of biasing influences like the length of the drift tube or the electric field, inverse reduced mobility $1/K_0$ can be calculated. However, for further computation, we are using the index t as a value in drift time. One weakness of the device is that not the whole concentration of analytes is ionized within the ionization chamber. The more distinct compounds residing within the chamber, the smaller the probability for a molecule getting ionized becomes. Hence, only relative intensities of peaks are visible in a spectrum.

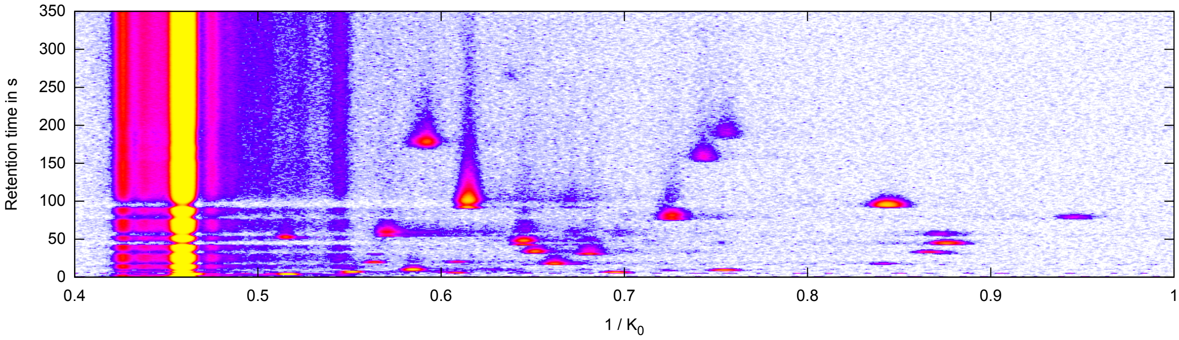
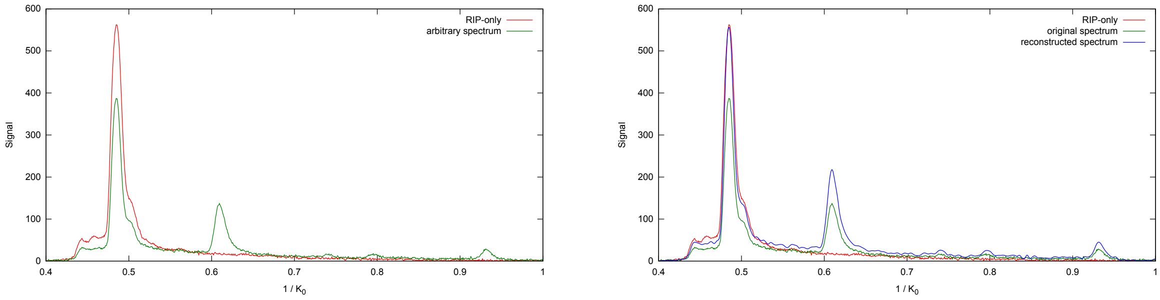


Figure 1: Heat map of an MCC/IMS measurement. X-axis: inverse reduced mobility $1/K_0$ in Vs/cm^2; Y-axis: retention time r in seconds; signal: white (lowest) $<$ blue $<$ purple $<$ red $<$ yellow (highest), *reaction ion peak* (RIP) at $1/K_0 = 0.46$Vs/cm^2



(a) Red: RIP-only (provided by drift gas); green: an arbitrary IMS spectrum with reduced RIP (at 0.49Vs/cm^2) caused by appearance of additional compounds (at $0.61, 0.71, 0.79, 0.93$Vs/cm^2)

(b) After parametrization the peak intensities are reconstructed

Figure 2: An arbitrary spectrum before and after processing containing spectrum parametrization and intensity reconstruction

Assuming that the concentration of drift gas is constant over all IMS measurements, peak intensity of the reference peak must be determined, consider Figure 2(a) (red spectrum). Typically, synthetic air or nitrogen N_2 is used for drift gas [2]. Let S be the matrix of the measurement, S_r the spectrum (row) at retention time r, $S_{\cdot,t}$ the chromatogram (column) at drift time t, $s_{r,t} \in S$ the value at retention and drift times (r, t). Let $v_r = \sum_{t=1}^{|T|} s_{r,t}$ be the sum of all values in S_r. In an arbitrary spectrum that contains several additional peaks, a reduced RIP appears (see Figure 2(a), green line). To avoid gaining background noise by reconstructing the intensities, we extract the peaks of the spectrum by modeling them with statistical functions.

Peak Description

We are given an arbitrary one-dimensional spectrum. Now we describe the peaks with shifted inverse Gaussian distributions g. To estimate the three parameters for g and the area under the curve, we use inverse Gaussian descriptors [3] for an easier estimation. The descriptors μ', σ, m can be easily transformed into the original parameters μ, λ, o. Under the condition that s_t has the highest value with respect of his neighbor points, we set the mode $m = t$. The area $a \in \mathbb{R} \geq 0$ is defined as follows: $a = s_t / g_{\mu,\lambda,o}(t)$. Now, we only have to estimate two parameters μ' and σ. Changing the parameter values has following effects: The lower we set σ, the narrower the curve becomes and vice versa. The larger μ' than m, is set the more skewed the curve forms.

Our approach is to use two nested loops in which both values are more and more adjusted. In the first step an outer loop μ' is considered and is in the inner loop σ. To determine the mean parameter μ', we start with an exponential search as the outer loop and a binary search for parameter σ in the inner loop.

In the second step, we also use two nested loops. However, this time we use a binary search instead of an exponential search for the outer loop to determine μ' since we only found valid powers of two. After the two steps, we obtain μ', σ, m, a. Now, we subtract the model from the spectrum and start a new searching and estimating iteration until no value in the spectrum exceeds a defined threshold.

This decomposition, unfortunately, prefers the first i.e. the left peak which results in a biased estimation of the further peaks in an overlapping clump. To correct these biases, we finally execute an EM algorithm [4] similar to our description in [3], but one-dimensional. Thereby we recompute the values of the model parameters using the previously estimated parameters as start parameters.

Intensity Reconstruction

A naive method for spectrum reconstruction is gaining the spectrum until the RIP in the spectrum is as high as in the RIP-only spectrum. A disadvantage of this method is that, in the worst case, the background noise extremely increases. A subsequent peak detection could potentially find a high amount of false positives (consider Figure 3, red line).

For a better reconstruction, we consider S_r again. First, we need a scaling factor that determines how much the RIP shrinked in S_r compared to S_0. We require that $\alpha' = \arg\min_{\alpha \in \mathbb{R}} \sum_{t \in T} (S_r - \alpha \cdot S_0)^2$. After setting the first derivative to zero and solving for α, we obtain $\alpha = (\sum_{t \in T} S_r \cdot S_0) / (\sum_{t \in T} S_0^2)$. To avoid a biased parametrization due to the RIP in S_r, we cut out the RIP by subtracting, let $S_r' = S_r - S_0 \cdot \alpha$.

Now, we perform a parametrization on S_r' and obtain the relative weights of the peaks ω_j among others. To obtain their uncorrected intensity f_j', we simply multiply the sum of all values v_r' with their weights, let $f_j' = \omega_j \cdot v_r'$. The last step to obtain the corrected

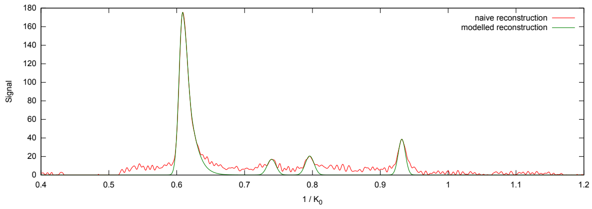


Figure 3: Difference between the naive method (multiply all values with scale factor) and using peak descriptors to reconstruct intensities

intensities is to divide f'_j by the scaling factor, thus we yield $f_j = f'_j \cdot \alpha$. A reconstructed spectrum is defined as follows $\forall t \in T : L_{r,t} = \sum_{j \in c} f_j \cdot g_{\mu_j, \lambda_j, o_j}(t)$.

Figure 3 demonstrates the capital difference between the naive method, where the original spectrum is divided by the scale factor and subtracted by the RIP-only spectrum ($S_r/\alpha - S_0$), and the parametrized spectrum L. While most of the background noise is also gained, only the "important" models increased their area or intensity.

Outlook

Our IMS pipeline provides several points: A new MCC/IMS measurement runs the peak detection first. After that, the detected peaks are modeled to obtain peak descriptors like position and intensity. Subsequently, the measurement is aligned to a reference data set. The intensity reconstruction shall be performed in a separate preprocessing step before the peak detection. Especially overlapping peaks and the monomer dimer phenomenon provide problems for analysis. Omitting these phenomenons, the last two steps of the pipeline are already implemented. Our next approach will be the implementation of a robust peak detection method. We will also investigate in the prediction of peak descriptors in measurements. This computation could simplify the registration and labeling of unknown compounds into the reference data set.

References

[1] J. Baumbach and G. Eiceman, *Appl. Spectrosc.* **53**, 338A (1999).

[2] D. Collins and M. Lee, *Analytical and bioanalytical chemistry* **372**, 66 (2002).

[3] D. Kopczynski, J. Baumbach, and S. Rahmann, Peak Modeling for Ion Mobility Spectrometry Measurements, in *20th European Signal Processing Conference*, 2012.

[4] A. Dempster, N. Laird, and D. Rubin, *Journal of the Royal Statistical Society. Series B (Methodological)* , 1 (1977).

Exomate – An exome sequencing pipeline

Marcel Martin
Bioinformatics for High-Throughput Technologies
Technische Universität Dortmund
marcel.martin@tu-dortmund.de

The subset of all regions of the genome that encode proteins is called the *exome*. Most hereditary diseases are caused by mutations within the exome. While the human genome contains three billion DNA "characters" (nucleotides or basepairs), the exome accounts for only about 30 million of them, making it much less expensive to sequence an exome than an entire genome. Current standard practice for trying to find disease-causing mutations is therefore to sequence exomes only.

Due to sequencing errors, each base of a DNA sample needs to be sequenced more than once, resulting in an n-fold *coverage*. Taking into account further losses, a raw exome sequencing dataset, as output by the sequencing machine, typically has a compressed size of 1–10 GiB. This includes some metainformation and also per-base quality scores.

A typical task is to find, given two or more sequencing datasets, common mutations that may point to the cause of a hereditary disease. Or, samples may originate from tumor and healthy tissue and comparing them can reveal mutations that cause or influence the development of the tumor.

While many standard bioinformatics tools exist that help in the analysis of sequencing data, these are usually command-line tools, many of which need to be run for a full analysis, which results in a cumbersome workflow with repetitive tasks. We have therefore developed a pipeline called *Exomate* that automates most of the tasks and provides an interactive web interface to the medical researcher who can easily get analysis results and is also able to re-run parts of the analysis with changed parameters.

Exomate is split up into three parts. First, there is a highly resource-intensive compute backend, which should preferably run on a multicore system or possibly on a cluster. The second part is a PostgreSQL database backend, which should run on a system with high single-core performance. The third is a web application that connects to the database and needs few resources itself.

Compute backend The compute backend is a GNU Makefile, supplemented by a set of "glue" scripts. The backend performs all steps necessary to get to a final set of mutations (*variant calls*) per sample. For most computations, we rely on standard tools and file formats.

Quality control is done on the raw sequencing data in FASTQ format. The program FastQC [1] computes basic statistics such as read counts and the distribution of quality values. This step of the pipeline is necessary in order to determine whether the sequencing process was successful and in particular whether the results of the following steps can be trusted.

Read mapping is the process of aligning the sequencing reads (length of approx. 100 nucleotides each) error-tolerantly to the reference genome. We use a wrapper script for BWA [1] that reduces disk I/O by sending the results of one step of the program (`bwa aln`) to the next (`bwa sampe`) through a Unix pipe. BWA is multi-threaded, but since it does have a non-parallelized part in its computations, we use at most four threads and instead analyze multiple samples at the same time if there are more cores available. Mapped reads are stored in the standard BAM format.

Variant calling is the process of finding differences between the sample and the given reference, done in our case with the Genome Analysis Toolkit (GATK) [2]. Since there are always sequencing errors, variants are assigned a quality value by the GATK, which gives an estimate of the authenticity of a variant call. Usually, variants below a given threshold are filtered, but we lower this threshold in order to increase sensitivity (at the cost of specificity). Variants are output in variant call format (VCF).

Database entry As last step, the VCF files are parsed and the contained variants are added to the database.

Database In addition to the variant calls, the database contains patient metadata that are needed in order to track where the variants come from. This allows queries such as: "Retrieve all variants from patient *x*." or "Retrieve all variants of patient *x*'s tumor that do not occur in his/her blood." The tables in the database describe the following entities.

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Person Each person has a unique accession number such as “P00543”. To maintain privacy, no names, birthdays etc. are kept in the database. Instead, persons are always referred to using their accession number. The association between names and accession numbers is only available to the medical researchers, who maintain a separate table (usually a spreadsheet) on their own secured systems with that information. Family relationships between persons are modelled through self-referencing foreign keys.

Sample A sample is a piece of tissue (blood, tumor) from a single patient. Multiple samples from a single patient are possible.

Library The genetic material that has been extracted from a sample is whose exome was captured and has been prepared for sequencing is called a library. It needs to be modelled in the database since there are multiple “exome capture kits”, where the more recent ones capture larger parts of the exome.

Unit A library can be sequenced more than once with different parameters. We call one instance of a sequenced library a “unit”.

From one entity to the following in the above listing, the relationship is one-to-many in all cases.

The actual mutation data uses the same conventions as the VCF input files. An entry contains the sample id, the chromosome name, the position of the mutation, the sequence of the reference at that position and the altered, observed sequence, such as: Sample 5, chromosome 3, pos. 242897, `A`→`AGC`, which is an insertion of the two bases `GC`. Since variants often occur in more than one sample, we normalize the data by splitting variants up into a “Variant” and a “Call” table. A variant is the four-tuple (chromosome, position, reference seq., altered seq.), without sample id. A call is identified by its sample and variant ids, but also contains meta information from the VCF file such as quality of the call and read depth.

Further tables in the database store gene and transcript locations, pre-computed scores that help estimate whether a mutation is detrimental or not [3], and an annotation table that, for each variant, gives the names of the gene it is located on and which amino acid and how gets altered by that variant.

Indel normalization The representation for a variant chosen in VCF files is not always unique for insertions and deletions (indels).

For example, the insertion of the sequence `AC` at position 99 could be described by 99 `G`→`GCA`, which is this alignment:

```
G--C
GCAC
```

A different algorithm may have determined that the alignment should look like this:

```
GC--
GCAC
```

This results in a variant encoded as 100 C→CAC.

Both alignments are optimal and there is no reason to prefer one over the other. In order to avoid this ambiguity, we will normalize insertions and deletions by moving them as far to the left as is possible.

An insertion described by $r_1 \to a_1, \ldots, a_n$ with $r_1 = a_1$ and $n > 1$ can be moved one position to the left without changing the alignment score or cost if $r_1 = r_n$. A deletion described by $r_1, \ldots, r_m \to a_1$ with $r_1 = a_1$ and $m > 1$ can be moved one position to the left if $r_m = r_n$. Together, an insertion or deletion described by $r_1, \ldots, r_m \to a_1, \ldots, a_n$ with $r_1 = a_1$ and either $n > 1, m = 1$ or $n = 1, m > 1$ can be moved one position to the left without changing alignment score if $r_m = r_n$.

Web frontend The web frontend that connects to the database contains the main logic of Exomate as SQL statements. It is implemented in Python with Flask[2] and uses the object-relation mapper SQLalchemy[3] to generate database queries dynamically, allowing a high level of interactive analysis.

References

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

[2] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010.

[3] Pauline C. Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, Jul 2003.

[2] http://flask.pocoo.org/

[3] http://www.sqlalchemy.org/

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).

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

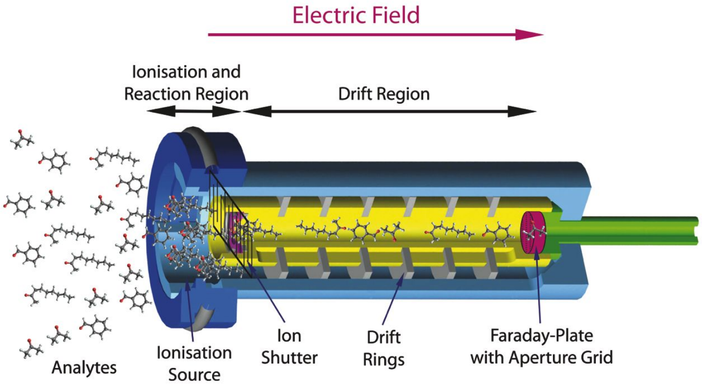


Figure 1: Working principle of an ion mobility spectrometer

medical diagnostics. As also disturbing substances based on nutrition, smoking and other individual behavior are measured in human breath; another aspect is to identify and pool these in a negative list.

Parallel gas - chromatography (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 due to the electron ionization characteristic fragments are determined.

One of the most common disturbing substances in breath analysis is toothpaste. Therefore another project is a toothpaste study, aiming to find out how long toothpaste can be detect in human breath. This is important to know the disturbances in patient breath.

At KIST Europe, I demonstrated that in exhaled breath some of the VOCs are related to the toothpaste rather than to the appropriate disease under investigation.

In IMS measurements most of the time there are many peaks not related to the exhaled breath itself. They might be from the room air or plastic devices used to collect the sample etc. It is important to distinguish between the peaks arising from the human breath, where we hope to identify analytes related to certain disease, and analytes which are disturbances that can distort or hide the peaks relevant to the study. Therefore where breath samples were taken before brushing and 1,2 and 6 hours after brushing. 10 mL of breath were used for each sample.

The temperature in the MCC as well as in the drift tube IMS was held at 40°C. The peaks were characterized using the software VisualNow.

Different brand of toothpaste with different ingredients were tested. One exemplary breath measurement is shown in Figure 2, [1]. All measured brand of toothpaste showed a similar pattern. All test persons already showed a peak before brushing. This might be some residual substances from brushing the evening before the measurement was conducted. These initially present peaks increase after brushing and then decrease again over the day. After 6 hours, at the end of the measuring cycle they are still present (see Figure 3).

In summary, it is virtually impossible to eliminate all toothpaste related peaks from any given breath sample, if the test subject has an average oral hygiene. Similarly, many VOCs show significant day-to-day variation in the signal intensities, which are related to various nutrients consumed by the individual under study for breath analysis. Finally, it can inferred that, systematic and environmental variabilities must be taken into consideration in order to relate the outcomes to medical question.

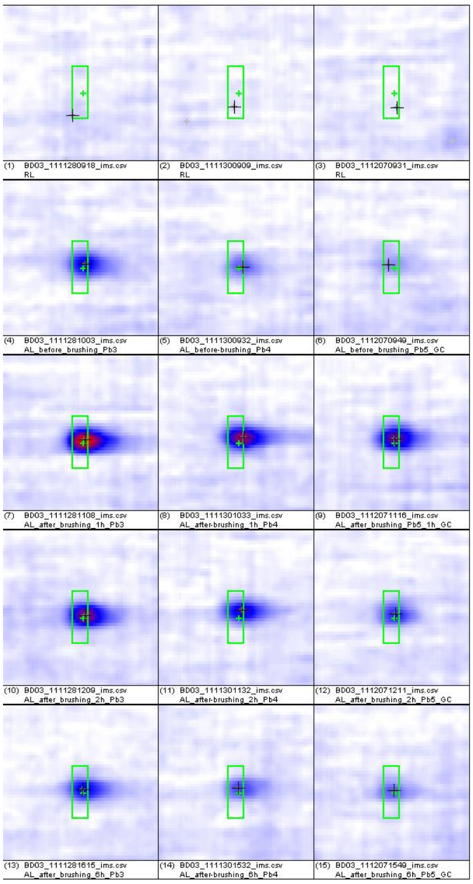


Figure 2: Toothpaste signals during the day

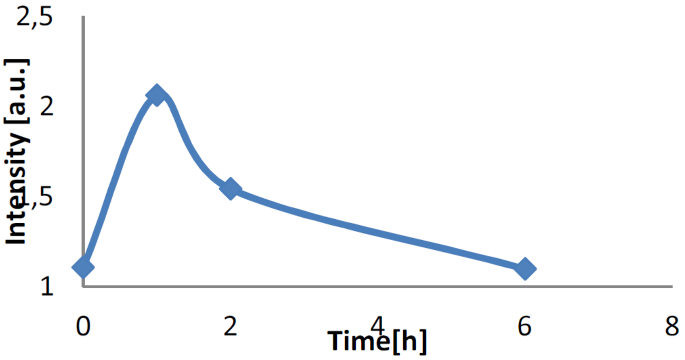


Figure 3: Timeseries of Eucalyptol

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). The samples for the GC-measurements are taken using TDS tubes. Several Peaks should be identified and verified by comparing the spectra of the IMS and the GC/MSD.

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

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.

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.

Another project is the rat sepsis model. In this project a multimicrobial sepsis is induced in rats. The breath of the anaesthetized rats is collected using the MCC/IMS and GC/MS tenax tubes. The samples will be compared and differences in the breath before and after the sepsis are expected. The aim of this study is to find peaks which indicate the sepsis, and finally assign these peaks to the metabolic pathway.

References

[1] K. Rupp, S. Maddula, and J.I. Baumbach. Good Breath - Bad Breath. *Poster 4. Anwendertreffen, 2012, Berlin.*

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

Peter Marwedel
Heinrich Müller
Alexander Zybin

ILP-based memory-aware mapping optimization for MPSoCs

Olivera Jovanovic
Computer Science 12
Technische Universität Dortmund
D - 44221 Dortmund, Germany
olivera.jovanovic@tu-dortmund.de

The design of modern embedded is a complex task containing several steps. One important step is the mapping of application tasks onto a homogeneous or heterogeneous multiprocessor systems. Here, the goal is to reduce the runtime or energy consumption. However, the influence of the memory hierarchy is not included in this optimization step, even though it is a well known fact that it directly influences the energy and runtime of the system. Therefore, the mapping optimization step has to include, analyze and utilize the underlying memory subsystem. We developed an ILP-based memory-aware mapping optimization tool which integrates the memory subsystem and is able to either optimize the runtime or energy consumption. Our first results show that the runtime based ILP optimization reduced runtime by 18% and the energy based ILP optimization reduced the energy by 21% in average compared to a state-of-the-art mapping optimization which does not utilize the memory hierarchy.

The memory subsystem has a significant contribution to the energy consumption and runtime of embedded systems [2, 3]. Concerning the runtime, it was known that the speed of a processor doubles each eighteen months while the memory speed grows only by 7% per year [1]. This fact is also well known as the memory wall problem. This problem still exists for multicore and embedded systems. Memories are slowing down the performance and large memories consume lot of energy per access. Therefore, small, fast and energy efficient memories like caches or scratchpad memories were introduced in order to close this huge performance and energy gap. Therefore, recent architectures are expanded by memory hierarchies featuring multiple levels of memories. The goal is to keep small and fast memories close to the processor in order to overcome the problem of high energy consumption and excessive runtime. Mapping application tasks onto the different processors in MPSoCs is an important step in the design of embedded systems. Here, the memory hierarchy has a significant influence. Unfortunately, the memory subsystem is not an integral part of this optimization step. Our ILP-based memory-aware optimization tool is integrating the memory subsystem in the optimization decision. We consider only scratchpad memories in our optimization. Contrary to caches, they have to be allocated by the designer. Also, they are predictable in terms of runtime and energy since their content is known in advance.

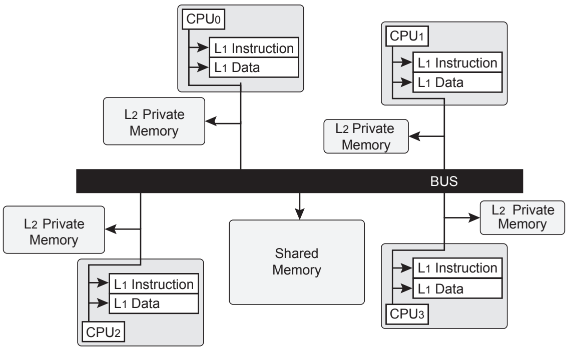


Figure 1: Heterogeneous MPSoC architecture with multi-level memory hierarchy

Our architecture model is shown in Figure 1. Each processor can have different levels of memories with different sizes. Here, each processor has a level one instruction and data scratchpad memory and one private memory at the second level. Only the processor has exclusive access to these memories. Further, each processor has access to a large shared memory. Each memory has different performance and energy values which are mainly dependent on the size and type of the memory.

The thread based application model that is used in our mapping optimization is shown in Figure 2. The main thread creates new threads which run in parallel in a so called parallel section. After the threads have finished their computation, they are joined again. Here, the memory requirements of each thread have to be analyzed and matched to the available system resources. Each memory object has its own size and number of read and write accesses. Depending on the underlying memory hierarchy and the threads' characteristics, for example if it is more data or computation intensive, it could be more efficient to map a thread onto a slower processor with a bigger local memory than on a slightly faster processor with a very small local memory, or vice versa. The problem gets even more complex when several threads have to be mapped to the available, different processor and its accessible memories.

We have set up an integer linear programming (ILP) for an application-to-architecture mapping with integrated memory-awareness. Depending on the requirements of the

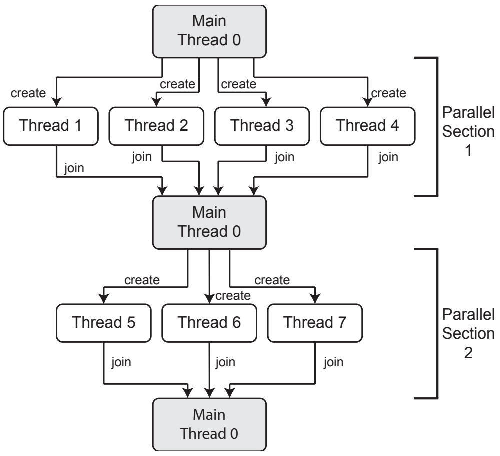


Figure 2: Thread-based application model

underlying embedded system, the designer chooses the proper ILP either for the minimization of runtime or energy consumption. The underlying values are based on an analytical model and are later verified by the cycle-accurate CoMET simulator.

The ILP formulation for the minimization of runtime, minimizes the overall runtime. This means that we minimize the finishing time of the last executed thread:

$$min(EndTimeThread_n) \tag{1}$$

The equations for constraints were also set up and will not be presented here, due to limited page size. Therefore we will describe them shortly. Next to the definition of the finishing time of a thread and its execution time on a certain processor, a binary variable is introduced, which defines if thread i is executed on processor j. Further, the access time to memories is defined in detail. As first, a binary variable defines if a memory object m_{obj} is mapped to a memory m which is accessible by processor p. It is very important to generate only valid mapping, i.e. if thread i is mapped onto processor j, then its memory objects can be only mapped to memories, which are accessible by this processor j. For the right determination of the memory access time, it is also important to integrate the number of read, write accesses and memory object size in this equation. Also, a constraint is added which proofs if the memory size of a memory is not exceeded. Further, dependencies in the taskgraph are considered as well as the constraint that no two threads can be executed on the same processor at the same time.

We have also set up a separate ILP optimization which minimizes the energy consumption. Here, the energy consumed on the processor, the buses and the memories is included in the equations. The energy spent on memory accesses has also to include the different read and write access as well as the memory object size and access width in order to determine the proper energy consumption. The energy spent on processors is separated into the energy spent in active mode and idle mode. Since the time that is spent in both modes has to be known, all equations from the runtime ILP are also included in the ILP for the energy minimization.

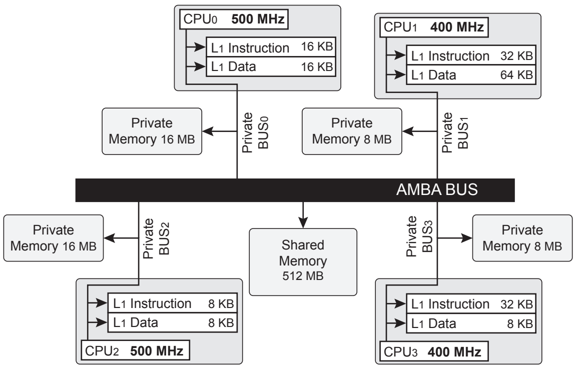


Figure 3: Heterogeneous MPSoC Architecture for Evaluation

In order to validate our mapping solutions, a simulation was performed on a cycle-accurate CoMET simulator by Synopsis [4]. For this, we implemented the architecture shown in Figure 3 in the CoMET simulator. Here, we compared ourselves to a state-of-the-art mapping tool which implements the same ILP but without the memory hierarchy awareness, which is the common practice for a state-of-the-art tool. We use benchmarks from the UTDSP suite [5] and a real-life benchmark (MPEG4). In average, the runtime could be reduced by 18% with the ILP optimization for runtime and further energy could be reduced by 21% with the ILP optimization for energy with our memory-aware tool. The average runtime of the ILP optimization takes about 11 seconds on an AMD Opteron 2.46 GHz.

References

[1] P. Machanick. Approaches to Addressing the Memory Wall. Technical report, School of IT and Electrical Engineering, University of Queensland, 2002.

[2] M. Verma and P. Marwedel. Advanced Memory Optimization Techniques for for Low-Power Embedded Processors. Springer-Verlag, 2007

[3] M. Kandemir and A. Choudhary Compiler-directed scratch pad memory hierarchy design and management. In *Proc. of DAC*, New Orleans, USA, 2002.

[4] Synopsys. CoMET. Virtual Prototyping Solution. `http://www.synopsys.com`, 2012.

[5] C. G. Lee. UTDSP Benchmark Suite. `http://www.eecg.toronto.edu/~corinna/DSP/infrastructure/UTDSP.html`, 2012.

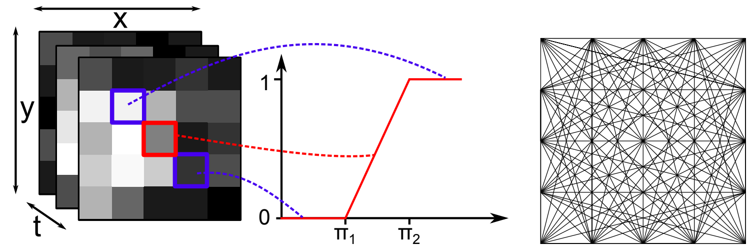
Detection of Nanoobjects in Optical Biosensor Data

Pascal Libuschewski
Lehrstuhl 7 & 12
Technische Universität Dortmund
pascal.libuschewski@tu-dortmund.de

This report presents the novel concepts of the real-time detection of nano-size objects in optical biosensor data, made by the project B2. Basis of the field of research is the PAMONO sensor which is capable to visualize nano-size objects, e.g. 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. The constraints speed and energy consumption arise in embedded systems that are the target platform for the application.

1 Introduction

The PAMONO sensor (Plasmon assisted Microscopy of Nano-Size Objects) [3] is a novel sensor to detect nano particles. A liquid containing the particles is passed through the sensor. The particles bind to the sensor surface and the plasmon resonance effect changes the reflection characteristics of the laser light illuminating the sensor. The change in intensity near an adhesion could be used as an indirect evidence for the particle. The task of the B2 project, among other, is to detect particles in real time and provide a platform based design and a platform design for the applications. This report shows the progress made in the real-time processing pipeline and sketches the further steps towards an automatic platform design with resource consumption as objective.



(a) Fuzzy detection: Three detection values (left) are mapped to a fuzzy set (right).

(b) Beamlets on a 4×4 dyadic square.

Figure 1: Fuzzy detection and beamlets.

2 Fuzzy Detection

To detect particles, the reflected light from the sensor surface is inspected. As a particle binds to the surface a characteristic step or jump in intensity can be observed. Smaller nano-particles cause the plasmon resonance effect to be less in intensity. The problem that raises from this, is that the signal-to-noise ratio is low for particles less than 150nm. To improve the signal-to-noise ratio a 2D+t fuzzy noise reduction, a 1D and 2D median filtering and a 1D wavelet denoising have been developed. The fuzzy noise reduction was evaluated in [2]. Even with most of the noise removed, the detection for small particles is still too low. Especially the fringe of the adhesion is problematical, because the size of the step drops significantly at the fringe. The result from this is, that only few pixels in the center get detected reliable.

A fuzzy approach was used to make use of the information in the detection values below or near to the noise level. Also domain specific knowledge was formed into fuzzy rules. To improve the detection of the small steps, the result of the different detection methods are transferred to a fuzzy set “step” in a fuzzification step (see Figure 1(a)). The fuzzy sets for “virus”, “background”, “artifact” and “noise” are derived from fuzzy set “step”. These fuzzy sets are then refined with fuzzy rules. The fuzzy rules combine the detection at one pixel with detections in the 2D+t neighborhood. It is for example possible to detect a low step at the fringe of the virus adhesion by calculating a fuzzy set of the neighborhood and combine it with the current detection within a fuzzy rule.

3 Beamlet Detection and Segmentation

Preliminary work has been made on beamlets [1], which provide a robust detection of blobs in noise affected data and also the ability to represent 3D objects. Both characteristics could be beneficial for the given problems. A beamlet is a line segment which connects two points on the border of a multiscale dyadic square (c.f. Fig. 1(b)). By chaining

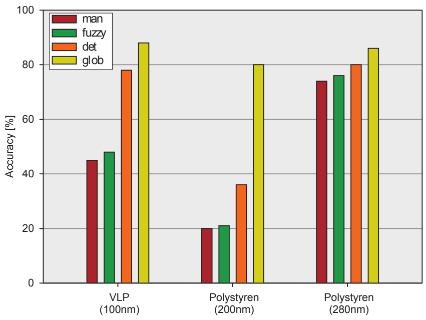


Figure 2: Detection results before (red) and after (yellow) automatic optimization

beamlets together, curves or other objects could be formed. To detect a blob in a noisy image a so called cost-to-time ratio cycle optimization problem is solved. Several options are possible for beamlets: The blob detection could be used on its own to detect the particles directly. It could be combined with the existing detection methods to provide a more robust detection or to decrease the search region on the images. It could be used as feature for the machine learning approaches. Or 2D or 3D beamlets could be used to replace the marching squares algorithm. Further research has to be done to decide which approach is suited best for the given problem.

4 Automatic Optimization

The automatic parameter optimization has been further developed to a multistep approach with an overall global optimization to decrease the run time and increase the detection result. Every processing step is first optimized with a genetic algorithm run. The final global optimization uses the chromosomes from the single steps and optimizes the whole pipeline at once. In an automatic process synthetic test data is generated and evaluated. It has been shown that the global optimization can improve the detection result significantly. It has also been shown, that the local optimization is beneficial for the overall run time, as they provide a useful start configuration for the global optimization. The results have been published in [2].

5 Future Work

The focus for the future work will be on the hardware with the objectives speed and energy consumption, which is important to transfer the application from a desktop to a mobile device. Therefore the platform design will be investigated using a design space exploration to explore which hardware requirements fit the problem best. For this step the hardware will be fully simulated giving control over many hardware parameters, e.g. clock speed, global/local memory size, energy consumption for different operations or work group sizes. A genetic-algorithm approach from section 4 is used to find the parameters for a real-time and energy-saving hardware platform. The best possible hardware could then be selected without expensive testing.

To adapt the beamlet blob detection to a particle detection, some problems have to be solved. First, the optimization assumes that there is one and only one blob present in the whole image, in the PAMONO sensor signal particles appear often with a rate of more than one particle per frame. Second, the ability to perform the detection in parallel and in real-time has to be analyzed. Third, a data structure to efficiently store and access beamlets on the GPU has to be developed.

6 Conclusion

The evaluation has shown that the automatic optimization outperforms every manual optimization, as could be seen in Figure 2. The fuzzy noise reduction could improve the detection rate (accuracy), even without adapting the manually chosen detection parameters. The global optimization was also beneficial in every test case. The optimization of the detection parameters in Figure 2 does not include the new fuzzy detection enhancement, but it could be shown that the the fuzzy detection enhancement is advantageous especially for the data sets with small particles and poor signal-to-noise ratio (e.g. 76% detection rate improved to 83% or 92% to 97%). Results are submitted for publication.

A collaboration with the project C3 is planned. As the FACT data shows similar problems, methods from the B2 project could also be applied to the FACT data with only minor changes. The ability to process the data online and in real time could be used by A3 e.g. for a fast preprocessing.

References

[1] David L. Donoho, Xiaoming Huo, Ian Jermyn, Peter Jones, Gilad Lerman, Ofer Levi, and Frank Natterer. Beamlets and multiscale image analysis. In *Multiscale and Multiresolution Methods*, pages 149–196. Springer, 2001.

[2] Pascal Libuschewski, Frank Weichert, and Constantin Timm. Parameteroptimierte und gpgpu-basierte detektion viraler strukturen innerhalb plasmonen-unterstützter mikroskopiedaten. *Bildverarbeitung für die Medizin, Springer Verlag*, pages 237–242, 2012.

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

Automating the Analysis of PAMONO Biosensor Data

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

This technical report summarizes the progress made in automating the analysis of PAMONO biosensor data, as part of project B2. A nearly completed automatic processing pipeline is introduced, that uses data-driven synthesis as a mean to define an objective function for parameter optimization and validation of analysis results. Future work in image processing and nano-object detection is sketched, as well as collaborations with other projects of the collaborative research center.

1 Introduction

Project B2 develops methods for detecting biological viruses by means of the PAMONO[1] biosensor [9]. The overall task is to provide the foundation for an inexpensive, real-time-capable, mobile virus detection device. The computer science aspect of this task is developing methods for automatic analysis of PAMONO sensor data. These methods aim at optimizing detection quality while providing results in real-time and minimizing energy consumption. These properties make them suitable for on-site diagnosis which can be carried out on portable computers.

The focus of this technical report lies on methods for automating the virus detection process. This automation is driven by the optimization of detection quality with respect to synthetic ground-truth data, cf. Section 2. Details on the processing pipeline are covered in Section 3, while Section 4 describes the progress in validating the overall methodologies. Section 5 outlines ongoing and planned collaborations with other projects.

[1] PAMONO: Plasmon Assisted Microscopy Of Nano-Objects

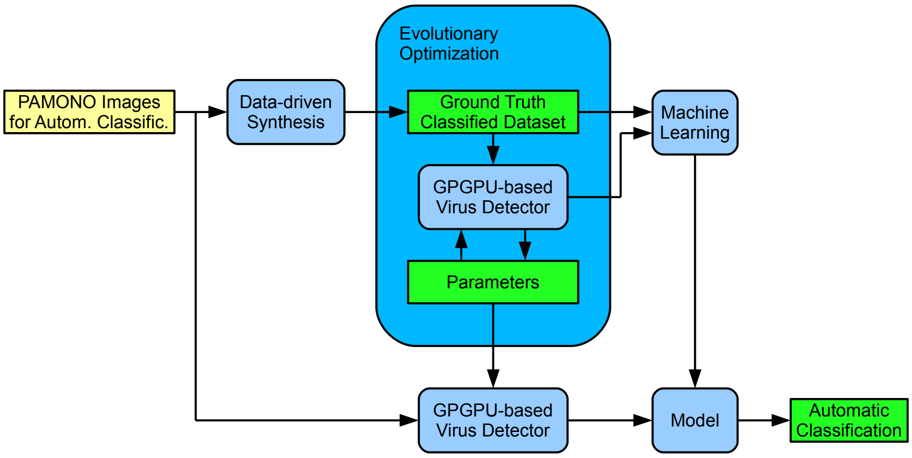

Evolutionary Optimization
PAMONO Images for Autom. Classific.
Data-driven Synthesis
Ground Truth Classified Dataset
Machine Learning
GPGPU-based Virus Detector
Parameters
GPGPU-based Virus Detector
Model
Automatic Classification

Figure 1: Automated Workflow

2 Automation by Synthesize/Optimize Pipeline

The analysis pipeline realized in 2011 [8] was automatic with respect to the detection of virus candidate areas and their classification after supervised learning. Manual input was required in two points: The parameters for the detector had to be hand-crafted to fit a given dataset, and the training data had to be labeled manually. The current pipeline, as depicted in Figure 1, aims at automating these aspects. The new workflow is as follows: The dataset (i.e. the time-series of images) generated by the sensor is input into a data-driven synthesis process, generating a synthetic ground-truth dataset with the same signal properties as the real-data input, but with known segmentation and classification. Using this ground-truth dataset, suitable parameters for the detector can be determined: An Evolutionary Algorithm (EA) [6] maximizes the positive agreement [1] between detection result and ground-truth with respect to these parameters. As the synthetic dataset has the same signal properties as the real-data input, the optimized parameters can set up the detector for that dataset as well. The detector is applied to both datasets, providing ground-truth classified detection results for the synthetic dataset and unclassified detection results for the real one. Thus a model can be trained [8] on the labeled ground-truth, which can then be applied to classify the unlabeled input.

3 Pipeline Components

Besides the synthesis component, the presented pipeline comprises components for image reconstruction and enhancement, which will be the topic of Section 3.1. These processing steps are followed by components for nano-object detection and classification, to be described in Section 3.2. In both cases, last year's advances are presented, followed by outlooks on future work.

3.1 Image Reconstruction/Enhancement

The images provided by the PAMONO sensor are polluted by a background signal, exhibiting approximately 20 times larger magnitude than the desired signal. Background removal relies on the assumption that the background is approximately constant, and thus its efficacy improves with the degree of validity of this assumption. To raise this validity, a variational image stabilization [5] has been implemented to account for inevitable micro-concussions of the sensor that cause sub-pixel intensity shifts. Stabilizing these shifts significantly improves the amount of artifacts remaining after background removal. However, sensor noise on a scale similar to that of the desired signal resides, necessitating denoising. To this end, temporal [3] and spatial [4] wavelet denoising techniques are to be evaluated with respect to the objective function defined by synthesis.

3.2 Detection/Classification

Concerning detection and classification, the existing GPGPU virus detector has been integrated into the overall MATLAB pipeline depicted in Figure 1. Integration of existing RapidMiner processes is in progress, providing automatic classification by machine learning. This constitutes the last remaining step in the implementation of Figure 1. Consequently, a quantitative evaluation can be conducted soon. Future work will extend the set of regarded features: Their domain will be 2D+t volumes of intensities, thus accounting for the spatio-temporal nature of the input data.

4 Validation

PAMONO is a very recent technique, making sufficient validation a crucial criterion in its acceptance. The conducted/planned validations divide into three categories:

1. **Validation of the sensor:** The samples analyzed using the sensor are compared to ground-truth obtained from electron microscopy. This is currently in progress.
2. **Validation of manual classification:** Observer agreements [1, 2, 7] of the manual classifications have been measured and are to be published in an upcoming paper.
3. **Validation of automatic classification:** Validation of the synthesize/optimize pipeline from Figure 1 will be conducted by synthesizing separate full datasets for training and testing. This avoids overfitting and alleviates problems with too small sample sizes as may arise in Split- and Cross-Validation.

5 Collaborations

A bidirectional collaboration with project C3 has been established. While C3 will analyze the noise in PAMONO data, B2 will examine FACT data because the underlying signal processing and classification problems bear similarity to those of B2. Furthermore, C3 and B2 plan to share methods for time-series analysis.

Further collaborations are possible with project B1, concerning blob-detection in IMS spectra, and with A1, concerning random fields as a sensor model for step detection and for segmentation.

References

[1] D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551 – 558, 1990.

[2] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

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

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

[5] J. Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[6] R. Poli, W. B. Langdon, and N. F. McPhee. *A Field Guide to Genetic Programming*. University of Essex (as PDF under Creative Commons License), 2008.

[7] P.E. Shrout and J.L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420 – 428, 1979.

[8] 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.

[9] 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.

Design of an Optimizing Compiler Enabling Multi-Objective Code Optimizations for GPGPU Applications

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

Latest chip design trends make GPGPU computing extremely interesting for embedded and cyber-physical systems and especially for sensor/actuator networks with high computational requirements. A typical sensor in this area is the PAMONO biosensor. The efficient utilization of many-core chips and the restricted availability of energy make code optimizations important.

Many-core architectures for computationally expensive applications in science and industry are becoming more and more important [6, 7]. An interesting chip technology for acceleration of parallel applications at desktop/server level and in the embedded and cyber-physical system domain are GPUs (Graphics Processing Units). These GPUs can execute general purpose applications by utilizing GPGPU (General Purpose Computing on GPUs) features of modern chip designs. The utilization of GPUs as acceleration devices in highly specialized application areas such as the automotive [5] or medical sector – such as the PAMONO biosensor [2] – demand for optimizing a GPGPU application to a certain specific platform type [8, 10]. Due to the use of GPGPU in mobile devices, energy consumption awareness is an important objective which has to be considered during GPGPU application design. The GPGPU application design process should therefore – in the face of restricted resources – simultaneously aim towards the lowest possible energy consumption and best possible runtime performance.

Traditional GPGPU application design is dominated by a trial-and-error-based optimization process. This process comprises code optimizations as well as load optimization for the different processing cores of graphics cards in order to achieve the optimal acceleration of an application. In addition to that, the GPGPU application code must be tuned towards static parameters in the workflow of GPGPU programming, such as the number of parallelly allocatable threads on each core. *Code optimizations*, *load optimizations* and *GPGPU static parameters* should therefore be taken into account simultaneously in an automatic optimization process. The automatic optimization process can exploit the following peculiarities in the GPGPU application design process:

- Modern graphics card chips support SIMD (Single Instruction Multiple Data) techniques with an enormous number of threads (SIMT – Single Instruction Multiple Threads). The SIMT architecture utilizes a shared register file for an efficient register access. By decreasing the number of registers allocated to a thread, it is possible to increase the number of threads which are running on single processor and thereby increases the performance of the GPGPU application [1].
- Modern graphics card chips comprise a large number of processing cores which share a common main memory. Therefore, a memory wall problem [3] exists, i.e. the speed of processing is much higher than the memory speed. In the face of this memory wall, an efficient utilization of memory-related instructions is mandatory for the GPGPU application design process.

The two peculiarities presented above should be targeted in an optimizing compiler for GPGPU applications. One of the most powerful and most interesting compiler optimization is instruction scheduling [4], because it can

- change the liveness of register values and alter the register utilization. This is beneficial for allocating more threads to a GPU and to boost the performance of GPGPU applications.
- improve the execution order of instructions. Instruction scheduling can place memory-related instructions in the code of GPGPU applications in a more efficient manner, meaning that the available memory bandwidth is better utilized.

Based on the fact that the parameter space is quite large and compiler support is required, an automated instruction scheduling optimization method must be designed. Most of the average case optimization techniques at compiler level have the drawback that they lack the capability to optimize the code in an elaborated way since they have no knowledge of the actual hardware platform. Therefore, the automated instruction scheduling optimization must utilize a profiling-based approach which feeds profiled performance indicators back into the compiler backend in order to achieve better solutions. Multi-objective instruction scheduling methods towards the optimization of the energy consumption and runtime performance were addressed in [9, 11].

References

[1] David B. Kirk and Wen-mei W. Hwu. *Programming Massively Parallel Processors: A Hands-on Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2010.

[2] Pascal Libuschewski, Frank Weichert, and Constantin Timm. Parameteroptimierte und GPGPU-basierte Detektion viraler Strukturen innerhalb Plasmonenunterstützter Mikroskopiedaten. In *Proccedings of the Workshop Bildverarbeitung für die Medizin (BVM)*, Lectures Notes on Computer Science, pages 237–242. Springer, 2012.

[3] Philip Machanick. Approaches to Addressing the Memory Wall. Technical report, School of IT and Electrical Engineering, University of Queensland, 2002. `http://www.itee.uq.edu.au/~philip/Publications/Techreports/2002/Reports/memory-wall-survey.pdf`.

[4] Steven S. Muchnick and Phillip B. Gibbons. Efficient instruction scheduling for a pipelined architecture. *SIGPLAN Not.*, 39:167–174, April 2004.

[5] Nvidia Corporation. NVIDIA and Audi Marry Silicon Valley Technology with German Engineering. `http://www.nvidia.com/object/io_1262839759949.html`, 2010.

[6] Nvidia Corporation. CUDA Toolkit. `http://developer.nvidia.com/cuda-downloads`, 2012.

[7] 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.

[8] Constantin Timm, Andrej Gelenberg, Peter Marwedel, and Frank Weichert. Energy Considerations within the Integration of General Purpose GPUs in Embedded Systems. In *Proceedings of the Annual International Conference on Advances in Distributed and Parallel Computing (ADPC)*. GSTF, 2010.

[9] Constantin Timm, Markus Görlich, Frank Weichert, Peter Marwedel, and Heinrich Müller. Feedback-Based Global Instruction Scheduling for GPGPU Applications. In *Proceedings of the ICCSA Workshop on Advances in High Performance Algorithms and Applications (AHPAA)*, Lecture Notes on Computer Science. Springer, 2012.

[10] Constantin Timm, Frank Weichert, Peter Marwedel, and Heinrich Müller. Design Space Exploration Towards a Realtime and Energy-Aware GPGPU-based Analysis of Biosensor Data. In *Special Issue "International Conference on Energy-Aware High Performance Computing (ENA-HPC)"*, Computer Science - Research and Development, pages 1–9. Springer, 2011.

[11] Constantin Timm, Frank Weichert, Peter Marwedel, and Heinrich Müller. Multi-Objective Local Instruction Scheduling for GPGPU Applications. In *Proceedings of the International Conference on Parallel and Distributed Computing Systems (PDCS)*. IASTED/ACTA Press, 2011.

Subproject B3
Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Processing Data Streams

Christian Bockermann
Lehrstuhl für künstliche Intelligenz
Technische Universität Dortmund
christian.bockermann@tu-dortmund.de

Today's masses of data have pushed the development of processing data using one-pass or online algorithms. Several libraries exist for data analysis and machine learning, providing algorithms to train prediction models on high-volume streams.

In this work we present the *streams* library, a flexible environment for defining data stream experiments that incorporate various of the existing approaches into a modular framework for processing data streams. Along with the *streams* library, the *streams* -Plugin enables the analysis of data streams within the RapidMiner toolbox.

Introduction

More and more applications rely on dynamic data that is produced in realtime and at a high volume. Scientific experiments, network traffic, sensor networks in manifacturing processes or message services are examples of such applications. Often the data in these applications is outdated quickly and reactions need to be applied in near to realtime. An example is given by Google's news search: the old batch-wise indexer has been replaced by a dynamic indexing engine to capture even more recent news articles. In other scenarios an on-time analysis might save resources as irrelevant data can quickly be detected and discarded.

Continuous data poses several challenges for data analysts: The data are often produced at large volume and require continuous processing to provide up-to-date prediction models or summaries. Such models or statistics need to be accessible at anytime. For preprocessing that data only limited resources with regard to memory, CPU and I/O is available. Recent advances such as Google's Map/Reduce paradigm address these by large scale parallelization of batch processes. While this scales well with the large amounts of data at hand, it does not tackle the problem of processing data *continuously*.

To catch up with the reqirements of large scale and continuous data, online algorithms have recently received a lot of attention. Various algorithms have been proposed for online quantile computation, frequent itemset mining, clustering or classification.

Our Contributions In this work we introduce the *streams* library, a small software framework that focuses on online processing of data. It allows for modelling data stream processes within XML files and supports instance extensions with custom classes. In summary, the proposed library supports

1. Modelling of continuous stream processes following the *single-pass* paradigm,

2. Anytime access to services that are provided by the continuous processing and the online algorithms deployed in the process setup,

3. Processing of large data sets using limited memory resources.

Based on the *streams* library, we implemented a RapidMiner *Streams Plugin* [2] to allow for the integration of streaming capabilities into the RapidMiner toolbox. This integration is based upon the RapidMiner-Beans plugin [1], a plugin for generic creation of RapidMiner operators.

An Abstract Stream Processing Model

In this section we introduce the basic concepts and ideas that we model within the *streams* framework. This mainly comprises the data flow (pipes and filters), the control flow (anytime services) and the basic data structures and elements used for data processing. The objective of the very simple abstraction layer is to provide a clean and easy-to-use API to implement against.

Data Items, Streams and Processors Figure 1 illustrates an abstract data process flow following the widely accepted pipes-and-filters pattern. A stream provides access to single elements (instances, events or examples) which are sequentially processed by one or more processing units. Each such data item represents a tuple, i.e. a set of (*key*,*value*) pairs and is required to be an atomic, self contained element. Data items from a stream may vary in their structure, i.e. may contain different numbers of (*key*,*value*) pairs, supporting sparsity. A *data stream* is essentially a possibly unbounded sequence of data items. In the pipeline model, a *processor* is some processing unit that applies a function or filter to a data item. This can be the addition/removal/modification of (*key*,*value*) pairs to the current item or an update of some model/state internal to the processor. Then the outcome is delegated to the subsequent processor for further computation. A set of processors is wrapped in a *process*, which itself is an active component that reads from a stream and applies all inner processors to each data item. The process will be running until no more data items can be read from the stream. Multiple streams and

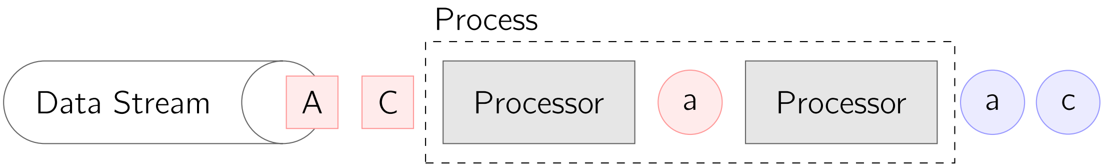


Figure 1: The general pipeline model for data processing.

processes can be defined and executed in parallel. These five basic elements (*stream*, *data item*, *processor*, *process* and *queue*) already allow for modelling a wide range of data stream processes with a sequential and multi-threaded data flow.

Data Flow and Control Flow An additional requirement of data stream processing is given by the *anytime paradigm*, which allows for querying processors for their state, prediction model or aggregated statistics at any time. We refer to this anytime access as the *control flow*. Within the *streams* framework, we model these anytime functions as *services*. A service is a set of functions that is usually provided by processors and which can be invoked at any time. It is also possible to define standalone services, e.g. for lookup tables on static data. Processors may also consume services. A simple example is given by a learning algorithm, that provides predictions based on its current model as shown in Figure 2. Here, an *Add Prediction* processor acts as service consumer, adding a prediction to the data based on the prediction service provided by the learner. The data flow and control flow define two orthogonal views of the stream processing. All the processors are working on distinct, single data items, limitting the overall memory to only the model and some aggregated model error.

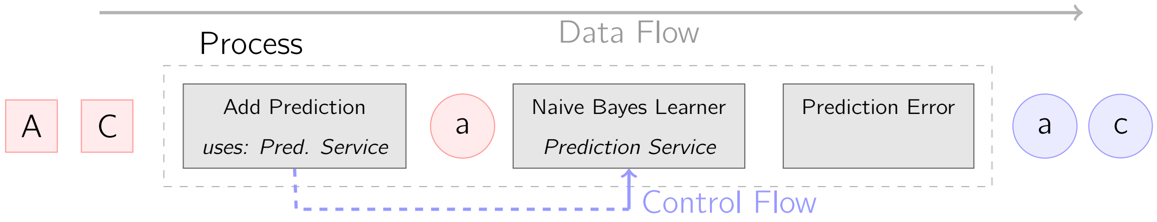


Figure 2: *The* data flow *and* control flow *in the use case of the general* test-then-train *evaluation scheme.*

Defining Stream Process

The definition of processes like the one shown in Figure 2 is done by XML elements for the stream, the process and the processes required. The following XML snippet shows the XML for the process above. The overall process is wrapped into a *container* element, which contains definitions of a *stream* and a *process* that will handle that stream.

Each element inside the *process* is a processor and corresponds to a simple Java class acting upon a single data item. Custom classes can be added by implementing the *Processor* interface and adding an appropriate XML element to the *process* definition.

```
<container>
   <stream id="data" class="stream.io.CsvStream"
           url="file:/tmp/test-data.csv" />
   <process input="data">
      <stream.learner.AddPrediction learner-ref="NaiveBayes" />
      <stream.learner.NaiveBayes id="NaiveBayes" label="play" />
      <stream.learner.evaluation.PredictionError label="play" />
   </process>
</container>
```

Figure 3: A simple experiment process in the *streams* runtime definition. Each XML element within the `process` element directly coresponds to a Java class implementing the `Processor` interface.

Applications and Future Work

The *streams* library does provide a variety of Java classes and a generic configuration for modelling stream processes with XML. The flexible nature and its easy extension-capabilities allow for designing stream processes for various application domains. Current research projects using *streams* library are the ViSTA-TV EU project for pre-processing user-log files and the SIEMAG/LS8 collaboration for online-monitoring of steel melt processes. Within the SFB-876, the *streams* library is used for processing high-volume data files of the FACT telescope.

Future work will be further extensions of the library to distributed processing of streams. In addition the integration of existing online-learning libraries (MOA) is to be finished. Based on the existing library, a correlated analysis of multiple streams is subject to further research. Such settings are found in the ViSTA-TV project as well as the correlated analysis of SQL-log-data with web-server access traces.

References

[1] Christian Bockermann and Hendrik Blom. Get some coffee for free - Writing Operators with RapidMiner Beans. In *RCOMM 2012: RapidMiner Community Meeting And Conference*. Rapid-I, 2012.

[2] Christian Bockermann and Hendrik Blom. Processing Data Streams with the RapidMiner Streams Plugin. In *RCOMM 2012: RapidMiner Community Meeting And Conference*. Rapid-I, 2012.

Leveling of Low Volume and High Mix Production

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

Conventional leveling approaches only focus on large scale production, especially in form of mixed-model assembly lines. This report presents a methodology for leveling of low volume and high mix production. It uses clustering techniques to group product types into product families. Based on these families, a production schedule is created which describes a repetitive sequence of capacity slots considering all families. According to this schedule each family is manufactured within a periodic interval.

1 Introduction

Production leveling is an essential element of the Toyota Production System and lean production respectively [6]. It aims at balancing production volume as well as production mix. Conventional leveling approaches distribute production volume and mix to equal-sized periods. The sequence of these periods constitutes the so called leveling pattern. According to this pattern every product type is manufactured within a periodic interval. This interval is represented by the key figure EPEI (every part every interval).

Literature dealing with production levling can be devided into two classes. While the first class focuses on procedure models (c.f. [7] for example), the second class describes leveling as an optimization problem in context of production sequencing (c.f. [3]). The latter are also referred to as level scheduling approaches. Both, procedure and optimization models focus on large scale production, especially in form of synchronized mixed-model assembly lines. Nevertheless, leveling can be implemented in low volume and high mix production by means of a methodology presented in this report. This methodology is composed of two fundamental steps. In the first step, clustering techniques are used to group product types into families referring to their manufacturing similarity (c.f. section 2). Based on these families, a family-oriented leveling pattern can be created in the second step (c.f. section 3).

2 Product family formation for leveling

The problem of forming product families for leveling is to group a set of *n* product types into the most adequate partition with *k* families. The partition size *k* is not known exactly ex ante. Each product type is described by p attributes representing manufacturing oriented grouping criteria. These attributes have to be weighted according to their impact on the grouping result. In addition to the formal objective of maximizing the similarity of product types within each family, objectives deduced from the application context have to be considered [2], [9]. One of these objectives is to consider only a defined partition size interval which is expected to include the optimal partition size. Thus, the solution space is limited to an interval that is defined using expert knowledge. Furthermore, roughly equal-sized families and only few very small families are aimed at.

Among the large number of approaches for product family formation which can be found in literature, cluster analysis is the most flexible and therefore the most adequate method [10]. To solve the problem of forming product families for leveling and find an attribute weight vector which optimizes all objectives simultaneously, a multi-objective optimization approach is employed. This approach is used in another application context in project B3 of Collaborative Research Center 876 (c.f. [8]). In this context it optimzes attribute weights based on probabilistic information about the allocation of objects to groups. For product family formation for leveling the optimization approach integrates a conventional clustering algorithm and uses the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [4]. It delivers a set of Pareto-optimal attribute weight vectors for every partition size *k*. To evaluate these solutions, the objectives described before are merged into an overall objective function in form of a desirability index [9]. More details on this grouping approach are given in [2].

3 Leveling pattern creation

To create a family-based leveling pattern the formed product families are segmented according to their overall volume share and their variation in volume. Based on this segmentation, product families characterized by a high volume, high order frequency, and low variation in demand are identified. These so called leveling families are chosen to be scheduled cyclically in the leveling pattern. While the pattern includes a capacity slot for each leveling family, the remaining families are considered in an aggregated way, i.e. in form of a single capacity slot per leveling cycle. In the next step the sequence of the selected leveling families in the pattern has to be determined. This sequence represents a leveling cycle in which every leveling family is manufactured once. To find a sequence which minimizes the overall changeover time, the sequencing problem is transferred to the Traveling Salesman Problem (TSP). It describes the problem of constructing a shortest tour passing exactly once through each of the n vertices of a graph. This tour represents a leveling cycle with minimal overall changeover time. Based on the assumption of a symmetric changeover matrix, the TSP is solved using a state-of-the-art neighborhood search heuristic, the Helsgaun variant of the Lin-Kernighan algorithm [5].

To determine capacity slots for each leveling families, the frequency in which the pattern repeats has to be calculated. This frequency is represented by the EFEI-value (every family every interval). The EFEI-value depends on the number of shifts in the considered planning period, the available capacity for changeover, and the overall changeover time required for one leveling cycle. Using the EFEI-value, if necessary adapted regarding minimal lot sizes, capacity slots in the leveling pattern can be determined for each leveling family. Besides, the required overall capacity for stranger families is also divided into equal-sized parts according to the number of leveling cycles per month. More details about the creation of a family-based leveling pattern are given in [1].

4 Real life application

The methodology described before was applied to level an assembly line with characteristics of low volume and high mix production. The considered assembly line is used to manufacture about 300 customized products for the engineering industry. In the first step, seven roughly equal-sized product families were formed referring to work content and share of identical components. Regarding volume share and fluctuation four families were classified as leveling families. For these four families a sequence with minimal overall changeover time was determined. Considering available and required capacities an EFEI-value of one day (two working shifts) was chosen. Capacity slots were determined according to this EFEI-value. The first implementation of the leveling pattern showed positive effects on inventory levels and delivery reliability.

5 Conclusion and future research

This report presents a methodology for leveling low volume and high mix production. This methodology uses clustering techniques to subsume the large number of product types into a manageable number of product families. These families are utilized to create a family-oriented leveling pattern. An aspect for future research will be to transfer the step of leveling pattern creation to a multi-objective level scheduling problem.

Literatur

[1] Fabian Bohnen, Matthias Buhl, and Jochen 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] Fabian Bohnen, Marco Stolpe, Jochen Deuse, and Katharina Morik. Using a clustering approach with evolutionary optimized attribute weights to form product families for production leveling. In Katja Windt, editor, *Robust Manufacturing Control*, volume 1 of *Lecture Notes in Production Engineering*. Springer, 2012. Accepted for Publication.

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

[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[5] Keld Helsguan. An effective implementation of the lin-kerninghan traveling salesman heuristic. *European Journal of Operational Research*, 126(1):106–130, 2010.

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

[7] Mike Rother and Rick Harris. *Creating continuous flow*. Lean Enterprise Institute, Brookline, 2001.

[8] Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In *Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, Michalis Vazirgiannis, editors, ECML PKDD 2011, LNAI vol. 6913*, pages 349–364. Springer, Berlin, 2011.

[9] Claus Weihs and Gero Szepannek. Distances in classification. In *Petra Perner, editor, Advances in Data Mining*. Springer, Berlin, 2009.

[10] Yong Yin, Ikou Kaku, Jiafu Tang, and JianMing Zhu. *Data Mining*. Springer, London, 2010.

A Quality Based Process Control Concept for Rolling Mills

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

Steel production processes are renowned for being energy and material demanding. In these processes, the intermediate product's quality cannot be assessed which may cause waste of energy and material resources as well as unnecessary machine wear. This Techreport gives a brief summary of some aspects that were analyzed in project B3 in order to develop an intelligent process control approach. Beyond that, first results of statistical analyses on the quality-related significance of process parameters are disclosed.

1 Introduction of Intelligent Manufacturing Process Control in Rolling Mills

Within the collaborative research center 876, project B3 focuses on assessing intermediate product quality in hot rolling processes by means of data mining in process data. In current state, material is pushed through the entire process chain without monitoring its quality properties. This production control concept leads to unnecessary waste of material and energy resources, in the case of processing material with minor quality [2]. To improve resource efficiency, the process's sustainability and material quality a new production control scheme has to be set up that incorporates information on quality currently produced.

The production control concept developed in project B3, called Intelligent Manufacturing Process Controll (IMPC), is depicted in figure 1. The basic idea is derived from process industry, where process control schemes known as "advanced process control" are state of

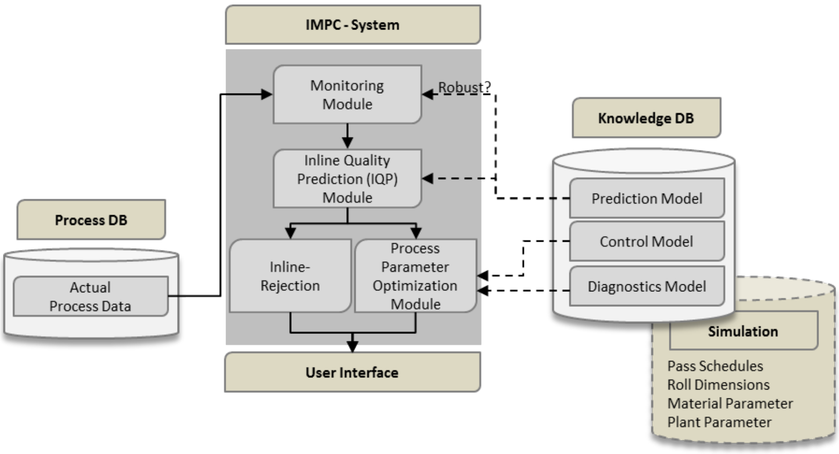


Figure 1: Intelligent Manufacturing Process Control Model (IMPC) [1]

the art [4]. The IMPC concept consists of four separate modules: (1) a data acquisition and storage module, that records all relevant process data and stores it in a database, (2) a monitoring module, keeping track of current process metrics, (3) the inline quality prediction (IQP) module, assessing each intermediate product's quality properties at all points in the process chain, and (4) a control module, that can either reject products of inferior quality or in an advanced version adapts process parameters so that quality deviations are corrected. [1] [2]
Core module of the entire concept is the IQP module. It relies on knowledge of previous processes' parameters as well as the physically measured product quality at the end of the process chain. Incorporating data mining as well artificial intelligence algorithms it identifies process parameter features that cause inferior quality at the end of the process chain. As real time process parameter optimization is yet to develop, the IMPC in its current version is designed to constantly analyze process parameters in order to eject those products showing process parameter features related to insufficient quality as early as possible. Consequently, an indicator of product quality has to be developed for the IQP module. Beyond that, quality determining process parameters have to be identified which deliver data for the quality prediction in the IQP module.

2 Quality Representation and Statistical Analyses

Each final product's quality is checked at the very end of the process chain. For this purpose a wide variety of product properties is tested. All this different information has to be condensed to a single quality indicator for the IQP module. For this purpose the quality level of a steel bar is defined. It is computed from a specified set of quality properties of all steel rods resulting from the bar according to formula 1 [1].

$$Q_b = 1 - \frac{1}{R_b}\sum_{r=0}^{R_b}\sum_{p=1}^{P}(W_p.\frac{\lambda_{p,r,b}}{\lambda_{p,max}}); \forall b \in B, where \tag{1}$$

$$Q_b \in [0,1], \text{quality level of bar b} \tag{2}$$

$$W_p \in [0,1], \sum_{p=1}^{P} W_p = 1, \text{weight of quality property p} \tag{3}$$

$$\lambda_{p,r,b} \in [1, \lambda_{p,max}], \text{value of quality property p of rod r in bar b} \tag{4}$$

$$b \in \{1, ..., B\}, \text{steel bar b in set of all bars B} \tag{5}$$

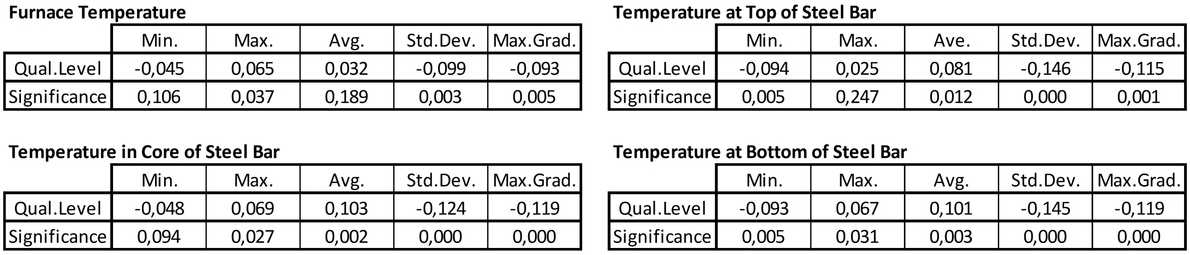
$$r \in \{1, ..., R_b\}, \text{steel rod r in set of all rods resulting from bar b} \tag{6}$$

$$p \in \{1, ..., P\}, \text{quality property p in set of all properties} \tag{7}$$

$$B, R_b, P \in \aleph \tag{8}$$

In the next step process parameters are analyzed by linking them to the resulting quality levels. By means of correlation as well as logistic regression models, parameter metrics of about 800 steel bars collected at the rotary hearth furnace are analyzed regarding their influence on produced quality [3] [5]. Bars with quality levels ≤ 0.75 were labeled 0, i.e. defective, the remainder 1 indicating sufficient quality. The results of the correlation analyses are shown in figure 2, those of the logistic regression models in figure 3. Based on these results it can be determined, that the furnace temperature's standard deviation and maximum gradient have a statistically significant impact on quality levels, judged on a significance level ≤ 0.05. Although the regression models contain only 20 parameters collected at the first process step with more than 100 parameters at other process steps remaining unconsidered, the coefficient of determination ($Nagelkerkes - R^2$) indicates that the bar models already account for about 7-9% of variability. The remaining 92% of variability have to be explained by the remaining process parameters. [1]

These results show that the quality levels derived from quality metrics can be employed in the IQP module as statistically significant dependencies can be found between those and process parameters.



Furnace Temperature

	Min.	Max.	Avg.	Std.Dev.	Max.Grad.
Qual.Level	-0,045	0,065	0,032	-0,099	-0,093
Significance	0,106	0,037	0,189	0,003	0,005

Temperature at Top of Steel Bar

	Min.	Max.	Ave.	Std.Dev.	Max.Grad.
Qual.Level	-0,094	0,025	0,081	-0,146	-0,115
Significance	0,005	0,247	0,012	0,000	0,001

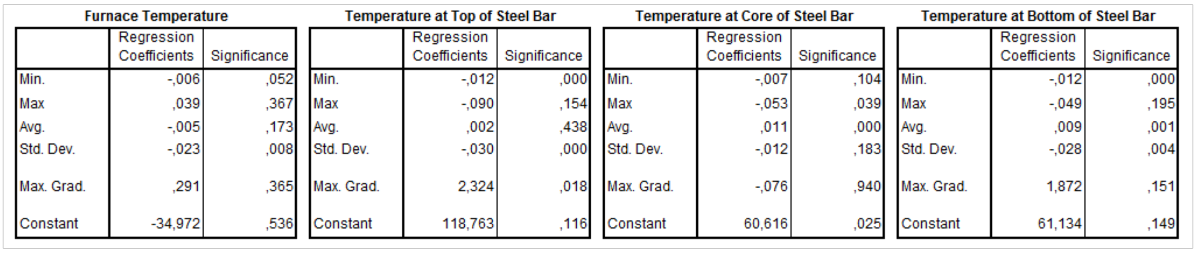
Temperature in Core of Steel Bar

	Min.	Max.	Avg.	Std.Dev.	Max.Grad.
Qual.Level	-0,048	0,069	0,103	-0,124	-0,119
Significance	0,094	0,027	0,002	0,000	0,000

Temperature at Bottom of Steel Bar

	Min.	Max.	Avg.	Std.Dev.	Max.Grad.
Qual.Level	-0,093	0,067	0,101	-0,145	-0,119
Significance	0,005	0,031	0,003	0,000	0,000

Figure 2: Correlation Coefficients of quality level and characteristic parameters [1]



Furnace Temperature

	Regression Coefficients	Significance
Min.	-,006	,052
Max	,039	,367
Avg.	-,005	,173
Std. Dev.	-,023	,008
Max. Grad.	,291	,365
Constant	-34,972	,536

Temperature at Top of Steel Bar

	Regression Coefficients	Significance
Min.	-,012	,000
Max	-,090	,154
Avg.	,002	,438
Std. Dev.	-,030	,000
Max. Grad.	2,324	,018
Constant	118,763	,116

Temperature at Core of Steel Bar

	Regression Coefficients	Significance
Min.	-,007	,104
Max	-,053	,039
Avg.	,011	,000
Std. Dev.	-,012	,183
Max. Grad.	-,076	,940
Constant	60,616	,025

Temperature at Bottom of Steel Bar

	Regression Coefficients	Significance
Min.	-,012	,000
Max	-,049	,195
Avg.	,009	,001
Std. Dev.	-,028	,004
Max. Grad.	1,872	,151
Constant	61,134	,149

Figure 3: Regression Coefficients and Significances from Logistic Regression [1]

3 Conclusion and Future Work

This Techreport presents a production control concept that incorporates information on product quality. Moreover, the reduction of various quality properties to a single quality level proved to produce valid results and the regression analyses give first hints on relevant process parameters to include in the IQP module.

The next steps in Project B3 will include conducting the statistical test described above on the process parameters remaining as well as identifying features in process parameters' series data that are related to certain quality properties. Moreover, the IMPC concepts will be detailed. Therefore, quality thresholds have to be identified for each process to specify the minimum quality level of a product accepted for further processing.

References

[1] Benedikt Konrad, Daniel Lieber, and Jochen Deuse. Striving for zero defect production: Intelligent manufacturing control through data mining in continous rolling mill processes. In Katja Windt, editor, *Robust Manufacturing Control*, volume 1 of *Lecture Notes in Production Engineering*. Springer, 2012. Accepted for Publication.

[2] Daniel Lieber, Benedikt Konrad, Jochen Deuse, Marco Stolpe, and Katharina Morik. Sustainable interlinked manufacturing processes through real-time quality prediction. In *Leveraging Technology for a Sustainable World*, Proceedings of the 19th CIRP Conference on Life Cycle Engineering, pages 393–398. Springer, 2012.

[3] Scott Menard. Applied logistic regression analysis. Number 07-106 in Sage University Papers Series on Quantitative Applications in Social Sciences. Sage, 2001.

[4] Dale E. Seborg, Thomas F. Edgar, and Duncan A. Mellichamp. *Process Dynamics and Control*. Wiley, 2. edition, 2004.

[5] Ishwar K. Sethi. Data mining: An introduction. In Dan Braha, editor, *Data Mining for Design and Manufacturing*, pages 1–40. Kluwer Academic, 2001.

Distributed Data Mining on Sensor Measurements in Interlinked Production Processes

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

In interlinked production processes, e.g. common in the steel industry, the quality of a product can often only be physically assessed at the end of the process. However, certain kinds of errors might already be introduced much earlier in the process, leading to unnecessary costs in subsequent processing steps. It would thus be beneficial if errors could be predicted in real-time as early as possible, based on sensor measurements of the current and previous processing stations. This report presents our efforts on the development and application of distributed data mining methods in such a scenario.

Introduction Though the sensing in interlinked production processes is local and thereby distributed, currently all measurements are transferred to a central server and stored in a database for offline analysis. While such an infrastructure might ease a first analysis, it is costly in terms of the required hardware and maintenance. Incremental approximation methods like the Core Vector Machine (CVM) [8] often only require a small sampled fraction of training examples, rendering the central storage of all data unnecessary. Moreover, the already distributed nature of the scenario might be exploited for parallel computations at the local stations. This distribution is not straightforward, however, since the quality potentially can be influenced by more than a single processing station. Further challenges are what features to extract from multivariate series of sensor values and how to deal with orders for which only the proportions of labels are known.

Distributed Learning of a Global Model One approach to cover the influence of multiple processing steps on the resulting quality is to learn a separate prediction model at each processing station. The model would need to be based also on the measurements of predecessor stations and thus not only be local, but global.

As was shown by Tsang et al. [8], many learning tasks can be reduced to determining a minimum enclosing ball (MEB) around all training examples, for which the Core Vector Machine (CVM) can compute a $(1+\epsilon)^2$-approximation in only constant time and space. Although these constants can be large in the worst case, often a much smaller fraction of training examples needs to be sampled in practice to achieve similar or even better accuracy as the SVM. In a network setting, using the sampling CVM might thus already reduce the number of training examples that need to be transferred from the predecessor stations to the current one. In an international collaboration [6] with Kanishka Bhaduri and Kamalika Das, we investigated if certain calculations of the CVM could be moved directly to the local (sensor) nodes, reducing communication costs and total training time further. The data is vertically partitioned in our case, i.e. the attributes of single examples (the sensor measurements related to the processing of a single steel block) are distributed across processing stations. This scenario is especially hard for normalized kernels as required by the CVM, like the commonly used non-linear RBF kernel. However, by using the Epanechnikov kernel whose form is similar, we were able to distribute the CVM's furthest point calculation. This comes at the expense of having to broadcast the current MEB to all local nodes, but instead of all feature values, only a single real value needs to be sent per example. For iteration numbers appearing in practice, we could prove that communication costs are orders of magnitude lower as long as more than a single attribute is stored per sensor node. The theoretical results were also verified empirically on synthetically generated Gaussian mixture data as well as on real-world datasets. Moreover, for the synthetic data and a varied number of attributes, it was shown that the Epanechnikov kernel yields similar accuracy as the RBF kernel. Our vertically distributed CVM (VDCVM) has been implemented in Java and can be run in a real network. With the sampling approach, it was already run on several gigabytes of training data in only a few seconds. The Java code is also integrated in a prototypical RapidMiner operator. It can work on subsets of attributes in parallel. However, inside RapidMiner, the algorithm is not yet distributed and the size of the training data is currently limited by the available main memory.

Distributed Learning of Local Models In cases where the quality of the end product only depends on local features, the predictions at predecessor stations could be combined to yield a single prediction at the current station. In a joint work with Sangkyun Lee [3], the primal SVM optimization problem is separated in such a way that the optimization problem can be solved in parallel locally at the particular processing stations. Local kernels are working only on attributes from the local stations, respecting the vertical partitioning of the data. Solving in the primal and the separation become possible by projecting the vectors in feature space to a finite lower dimensional space with the help of random projections. The optimization at the local stations is then done by a stochastic gradient descent approach called ASSET [4]. The prediction at the current station is a weighted summation of the individual predictions from predecessor stations. Optionally, the weights can be optimized globally by solving a linear or quadratic program, depending on the type of regularization. In each outer iteration, only a fixed number of updated weights needs to be communicated to the local nodes. The optimization algorithm has been implemented by modifying the existing C++ ASSET software. Different nodes can be simulated on a single machine by starting multiple threads. It has been empirically verified on synthetically generated as well as standard data sets that the separable optimization leads to similar prediction accuracy as when done centrally, but only in cases where the label does not depend on features from different nodes. In the factory setting, the separable optimization therefore may only be used to predict errors that can be detected locally. Then the parallel optimization at predecessor stations would allow for much faster training.

Data Management and Feature Extraction As first experiments in a rolling mill case study have shown [5], meta data like the duration of steel blocks staying in a rotary hearth furnace don't suffice to predict the quality correctly. From our industrial partner we are now regularly getting sensor measurements recorded at multiple processing stations. A relational database schema has been developed for storing all meta data about orders, quality information and the measurements in a normalized way. A command line tool implemented in Java imports the original source files (CSV and Excel) into the database, fitting it to the aforementioned schema. The tool can also export subsets of the data as necessary, based on user defined SQL filters. The database contains around 23 GB of data recorded from January 2011 to June 2012. In first experiments, the focus is on a small subset of about 500 labeled blocks, with measurements from all processing stations. For training, appropriate features have to be extracted from the raw value series. A RapidMiner preprocessing process has been designed that calls subprocesses for each type of value series. Each subprocess works on a single time series at a time and since is not constrained by main memory limitations. All subprocesses can work in parallel. The architecture is highly modular and allows for the easy addition of new value series types and accompanying extraction processes. In close collaboration with our project partners, we have defined extraction processes for all recorded value series types. Our current work is on the correct splitting of the series. We expect to finish this work at the end of August 2012, resulting in a first labeled data set for experiments.

Conclusion and Future Work Support vector methods have been distributed in two different ways: The VDCVM can learn a global model from vertically distributed sensor measurements incrementally, while separating the primal SVM optimization problem allows for the independent training of local models, combining their results to a global prediction. Both methods have low communication costs. We have begun to define processes for the feature extraction from value series. Once these are finished and a first labeled data set exists, both distributed methods can be evaluated on this data set. The one class model learned by the VDCVM may further be used as input for an efficient Support Vector Clustering (SVC) algorithm by Jung et al. [2]. It is intended to work on the distribution of this algorithm, too. Using SVC would allow for extending the prediction model, also including data for which only the label proportions of whole customer orders are known, and compare this to LLP using k-Means [7]. Clustering results largely depend on the correct weighting of attributes, as was shown with the LLP algorithm and could also be verified in a case study on forming product families for production leveling [1]. An interesting question is if the evolutionary optimization used for the weighting can be replaced by more efficient non-convex optimization methods and if these methods can also work distributed. The collaboration with Kanishka Bhaduri is continued in form of a book chapter about distributed data mining in sensor networks.

References

[1] F. Bohnen, M. Stolpe, J. Deuse, and K. Morik. Using a clustering approach with evolutionary optimized attribute weights to form product families for production leveling. In *Robust Manufacturing Control*, volume 1. Springer, 2012.

[2] K.-H. Jung, D. Lee, and J. Lee. Fast support-based clustering method for large-scale problems. *Pattern Recogn.*, 43:1975–1983, May 2010.

[3] S. Lee, M. Stolpe, and K. Morik. Separable approximate optimization of support vector machines for distributed sensing. In *ECML PKDD 2012*. Springer, 2012.

[4] S. Lee and S.J. Wright. ASSET: Approximate stochastic subgradient estimation training for support vector machines. In *Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM)*, pages 223–228, 2012.

[5] D. Lieber, B. Konrad, J. Deuse, M. Stolpe, and K. Morik. Sustainable interlinked manufacturing processes through real-time quality prediction. In *19th CIRP Conference on Life Cycle Engineering*, pages 393–398. Springer, 2012.

[6] M. Stolpe, B. Kanishka, D. Kamalika, and K. Morik. Anomaly detection in large data sets by vertically distributed core vector machines, 2012. Submitted to the Int. Conf. on Data Mining (ICDM).

[7] M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *ECML PKDD 2011, Part III*, pages 349–364. Springer, 2011.

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

Subproject B4
Analysis and Communication for dynamic traffic prognosis

Michael Schreckenberg Christian Wietfeld

Channel-Aware Floating Car Data Transmission via LTE

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

The collection of Floating Car Data (FCD) is very important for dynamic traffic forecasts. For this purpose, sensor nodes in cars are used to transmit traffic information via Long Term Evolution (LTE). In this report, we evaluate this data transmission with regard to several Key Performance Indicators (KPIs) for a channel-aware transmission. The introduced indicators are the negative impact on human communication, the power consumption of the mobile devices and the local distribution of cars sending FCD in the scenario. As methodology for performance evaluation a close to reality parameterized Markovian model, laboratory data rate measurements as well as power consumption measurements and ray tracing simulations are used. By applying the channel-aware transmission, the Quality of Service (QoS) level of human communication can be obtained and simultaneously the power consumption is significantly reduced. In addition, the fraction of active FCD devices on the highway can be increased.

1 Motivation

Highly precise and real-time traffic forecasts became a major research topic in the last years. Hereby, the precision of prognosis depends on the quantity, quality and reliability of available information of the traffic flow [1]. The communication aspects of such approaches are as important as the actual prognosis. The dynamic traffic forecast is based on FCD which contain vehicular sensor data. This data can be transmitted from sensing cars to a server via public cellular communications systems such as LTE. Thereby, the negative impact on the existing human communication has to be as small as possible. To ensure this, we use a Channel Sensitive Transmission (CST) for the FCD [2]. The goal is to analyze the performance of this approach regarding different KPIs. In Fig. 1 a system overview for performance analysis of channel sensitive transmission including methodology for the evaluation of the different KPIs is illustrated.

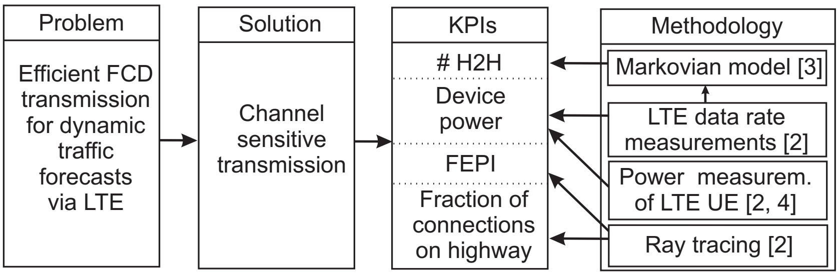

Problem
Efficient FCD transmission for dynamic traffic forecasts via LTE
Solution
Channel sensitive transmission
KPIs
# H2H
Device power
FEPI
Fraction of connections on highway
Methodology
Markovian model [3]
LTE data rate measurements [2]
Power measurem. of LTE UE [2, 4]
Ray tracing [2]

Figure 1: System Overview for Performance Evaluation of Channel Sensitive Transmission

2 Channel Sensitive Transmission Scheme

By applying the channel sensitive transmission scheme to the FCD transmission, many Machine-to-Machine (M2M) devices with good channel conditions transmit FCD. This is guaranteed by a transmission probability $p_{i,j}$:

$$p_{i,j} = \frac{\left(\frac{SNR_i}{SNR_{max}}\right)^{\alpha} \cdot \left(\frac{v_{max}}{v_j}\right)^{\beta}}{\sum\limits_{l=1}^{N} \sum\limits_{k=1}^{M} \left(\frac{SNR_l}{SNR_{max}}\right)^{\alpha} \cdot \left(\frac{v_{max}}{v_k}\right)^{\beta}}, i = 1...N, j = 1...M$$

Here, $p_{i,j}$ is the normalized transmit probability for class i, j. The index i separates the different Signal-to-Noise Ratios (SNRs) and j the different velocities for the classes of the Markovian model [3]. The arrival rate for each class is multiplied with $p_{i,j}$. SNR_{max} is the SNR for which the highest data rate can be achieved and v_{max} is the highest velocity in the scenario. The parameters α and β control the intensity of the channel sensitive transmission scheme. The number of devices which transmit data should be independent of the coefficient α. Therefore, we normalize the transmit probability.

3 Key Performance Indicators

In this section, the KPIs which are important for the performance analysis of the FCD transmission are described.

- **Number of servable H2H connections**: The negative impact of FCD transmission on the Human-to-Human (H2H) communication should be as small as possible. By means of a Markovian model which is parameterized by laboratory data rate measurements [3], the number of H2H connections with a certain QoS level is evaluated for different transmission strategies of the FCD. The QoS level is set to 10 % blocking probability and for the FCD transmission we assume that 5 % of all cars transmit FCD.
- **Power consumption** of the LTE User Equipment (UE). For the transmission of the FCD, common LTE USB sticks or smartphones can be used. For these devices the power consumption is a major KPI, because the development in battery technology can not feed the power hungry devices. Therefore, saving power of mobile devices is a relevant topic over the last years. We measured the power consumption in the laboratory and derived an average power for FCD devices [2, 4].
- **FCD Entropy Position Index (FEPI)**: The distribution of the active FCD users is very important for the dynamic traffic forecast. We introduced the FEPI [2], which is based on the Theil Index T. Thereby, the scenario is divided into K parts. We used $K = 16$. $\bar{x}$ is the average number of users per part and x_p is the number of users in part p. From this value we derived the FEPI:

  $$T = \frac{1}{K}\sum_{p=1}^{K}\left(\frac{x_p}{\bar{x}} \cdot ln\frac{x_p}{\bar{x}}\right); \qquad FEPI = \sqrt{\frac{T_0}{T}}$$

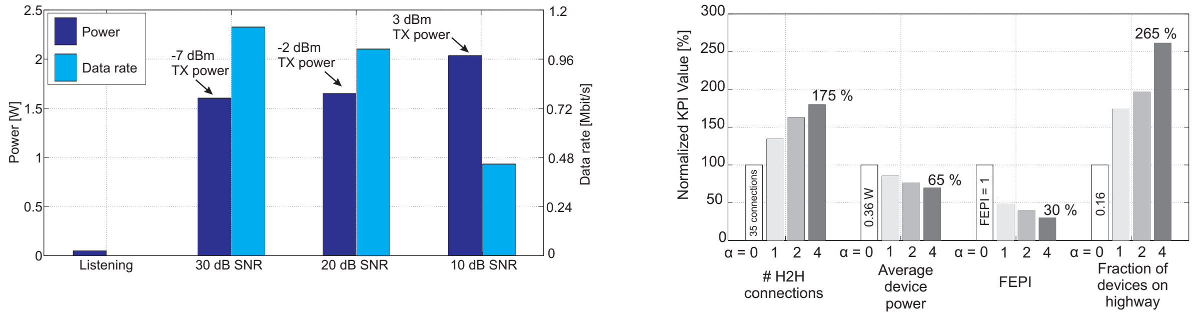
  T_0 is the Theil Index without channel sensitive transmission. This KPI is derived from ray tracing simulations and describes how homogeneously the users are distributed in the scenario.
- **Fraction of active FCD vehicles on the highway**: For the dynamic traffic forecast FCD from cars which are traveling on the highway are of special interest. This KPI is also gained from ray tracing simulations [2].

4 Results for KPIs under Channel Sensitive Transmission

In Fig. 2a the data rate and the corresponding power consumption $P_{i,j}$ for the transmission and listening mode for different SNRs are presented. These values are needed to parameterize the Markovian model and to evaluate the energy-efficiency of FCD transmission.

An overview of the KPIs for channel sensitive transmission is given in Fig. 2b. For reasons of comparison, the values for each KPI are normalized by the KPI values for $\alpha = 0$, which represent a transmission without channel awareness. It can be seen from the figure that the number of H2H connections can be increased by 75 %. This result is taken from the Markovian model. The average power consumption decreases by 35 %. This means that the negative impact of the FCD transmission on the H2H communication decreases although the FCD devices save power. The costs for this gain are the smaller FEPI. The devices transmit FCD dependent on the channel conditions. Hence, the distribution of the users gets organized. This drawback is not that critical for the dynamic traffic forcast, because the fraction of cars which are on the highway and transmit FCD increases by over 150 %.

In the future, we will validate the results by event-based simulations. These simulations make also investigations of different traffic characteristics possible.



(a) Consumed Power of LTE Device and Data Rate for Different Channel Conditions

(b) KPIs under Channel-Aware FCD Transmission

Figure 2: Laboratory Measurements and KPI Results

References

[1] C. Ide, Timo Knaup, B. Niehöfer, Daniel Weber, Lars Habel, Michael Schreckenberg, C. Wietfeld. *Efficient Floating Car Data Transmission via LTE for Travel Time Estimation of Vehicles*, Proc. of the IEEE 76th Vehicular Technology Conference, Québec City, Canada, Sep 2012

[2] C. Ide, B. Dusza, C. Wietfeld. *Performance Evaluation of V2I-Based Channel Aware Floating Car Data Transmission via LTE*, Proc. of the 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, USA, Sep 2012

[3] C. Ide, B. Dusza, M. Putzke, C. Müller, C. Wietfeld. *Influence of M2M Communication on the Physical Resource Utilization of LTE*, Proc. of the 11th Wireless Telecommunications Symposium, London, UK, Apr 2012

[4] C. Ide, B. Dusza, C. Wietfeld. *Energy Efficient LTE-Based Floating Car Data Collection for Dynamic Traffic Forecasts*, Proc. of the IEEE International Conference on Communications Workshop on Novel Approaches to Energy Measurement and Evaluation in Wireless Networks, Ottawa, Canada, Jun 2012

Travel Time Estimation of Vehicles with Floating Car Data

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

The travel time estimation of vehicles is an important factor in the area of dynamic traffic prognosis. Currently, the traffic data base mainly consists of information from stationary detectors distributed over the road network. Our approach is to improve this data base by using Floating Car Data (FCD) including travel time information transmitted to a server via Long Term Evolution (LTE). In this report, the benefit of FCD on the accuracy level of travel time estimation is analyzed. An enhanced Nagel-Schreckenberg cellular automaton model is used and it is shown that a penetration rate of a few percent is sufficient for a realistic travel time estimation.

Motivation

Due to the gradual deterioration of the road traffic situation an innovative, high precision and real time traffic forecast became increasingly necessary in the past years. Thereby, the quality of the forecast depends on the quantity, quality and reliability of the available information on the current traffic flow [2]. Currently, traffic information which is used for traffic prognosis is based on stationary loop detectors, TMC (Traffic Message Channel) and police messages. An improvement of this incomplete data base can achieved by including FCD transmitted via LTE. As with many other things, the costs must be balanced against the benefits. On the one hand the additional FCD should be a benefit for the traffic forecast, on the other hand the influence of the FCD transmission on the LTE network should as small as possible [2]. The travel time estimation is a major factor in the area of traffic prognosis and is the focus of this approach.

Traffic Simulator Generating Virtual FCD

Our simulation is based on an enhanced Nagel-Schreckenberg model with velocity depended randomization [3], [1]. A one dimensional array of 1000 cells represents a single lane track, whereas a cell length of 7.5 m is used (see Fig. 1). Thereby, each cell can be occupied by at most one vehicle. The velocity v of each vehicle is an integer with a value between 0 and v_m, in our case the maximum velocity v_m is set to 4.

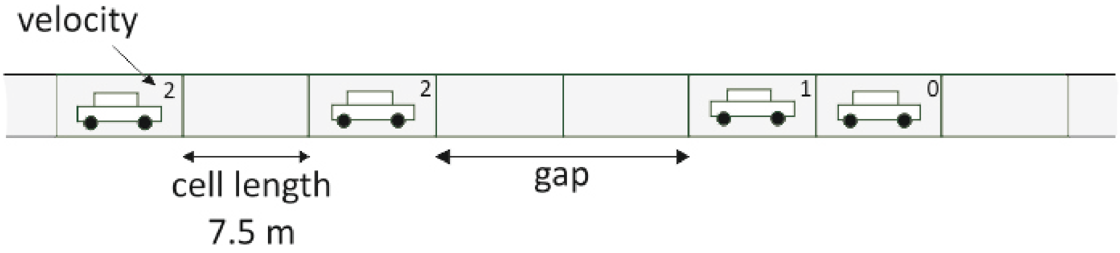


Figure 1: Discrete vehicle positions according to the Nagel-Schreckenberg model for generating virtual FCD travel times

The update rules of the model are performed in parallel for all vehicles, with one time step set to 1 second. This corresponds to velocity bins of 27 km/h and a maximal velocity v_m of 108 km/h. Open boundary conditions and two bottleneck situations, one within the track (cells 500 to 550) and one at the end (last 4 cells) are used. If the leftmost cell of the array is empty, a vehicle will be inserted with a probability γ and a velocity of v_m. If a vehicle is near the end of the track and the velocity is high enough to reach it in the next time step, δ is the probability that it can leave the track. To simulate the bottlenecks within and at the end of the track the dawdling parameter p is increased, respectively the parameter δ is smaller than 1 (δ = 0.5). To realize the velocity depended randomization, the dawdling parameter p is a function of the velocity (slow-to-start rule):

$$p(v) = \begin{cases} 0.5 & \text{for } v = 0 \\ 0.2 & \text{for } v > 0 \end{cases} .$$

In the bottleneck situation within the track, $p(v)$ is always set to 0.5. In order to get a high flow situation, on average 1800 vehicles per hour are inserted onto the track and $\gamma = 0.5$ is used.

After a sufficiently long relaxation time of 5000 time steps the simulation runs for further 15000 steps. Every vehicle which passes the track transmits its travel time as FCD. To evaluate the benefit of the upcoming prognosis, we simulate different penetration rates of FCD vehicles (5 %, 1 % and 0.1 %). Thereby, the corresponding vehicles are chosen randomly.

Results

To evaluate the necessary penetration rate of FCD vehicles providing their travel time to get a reliable traffic state estimation, a Monte Carlo Simulation is used. Fig. 2 shows one representative result of the simulations out of 3000 iterations. In this simulation, we use 100 % FCD penetration rate as reference. Thereby, every vehicle announces its travel time after passing the track. The obtained data is grouped into 5 minute intervals. To clarify the difference of varying penetration rates, the figure also illustrates the estimated travel time using 5 %, 1 % and 0.1 % penetration rate of FCD vehicles in comparison to the ideal simulated travel time. If no FCD vehicle announced its travel time within a 5 minute interval the missing value is calculated by linear interpolation.
The 5 %-curve is quite similar to the reference. It is obvious, that the shape of the 1 %-curve is still roughly the same but it diverges in some cases significantly. A estimation based on 0.1 % FCD differs clearly from the ideal travel time. However, the results suggest, that a reliable travel time estimation with a few percent FCD vehicles should be possible.

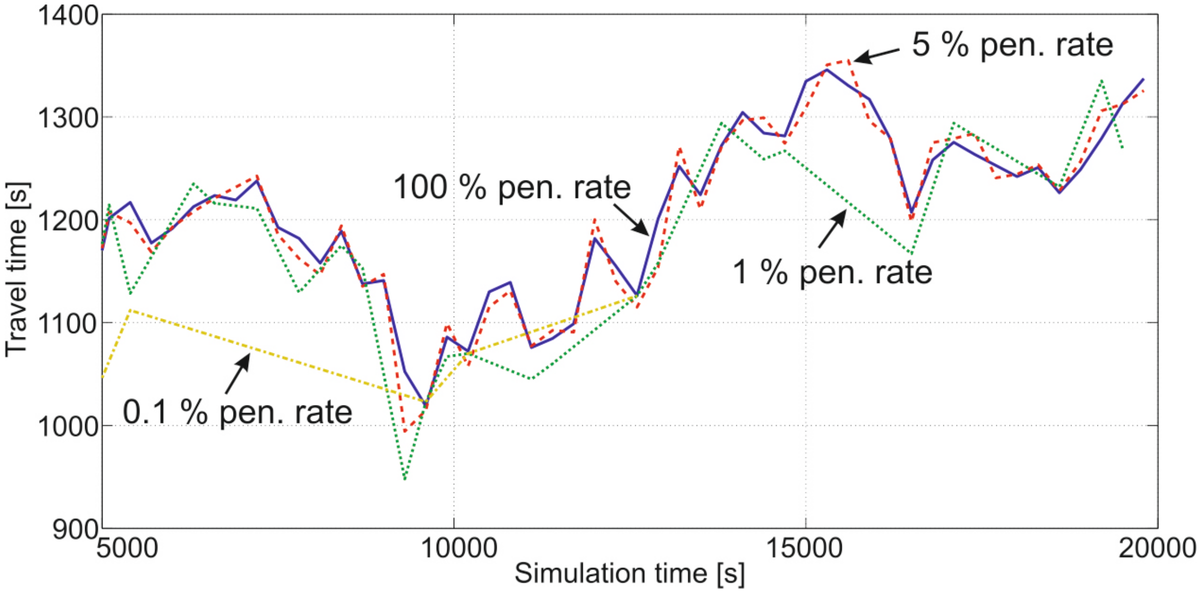


Figure 2: Traffic simulator: Travel time estimation of total track for different FCD penetration rates

To validate these results, the empirical Cumulative Distribution Function (CDF) over all 3000 iterations illustrates the quality of the travel time estimation for the different penetration rates in the simulation. Therefore, the relative deviation between the travel times measured at 5 %, 1 % and 0.1 % and the reference was calculated (see Fig. 3).

It clearly indicates that a reliable travel time estimation with a minor FCD penetration rate is possible. In 90 % the travel time deviation at 5 % penetration rate is lower than

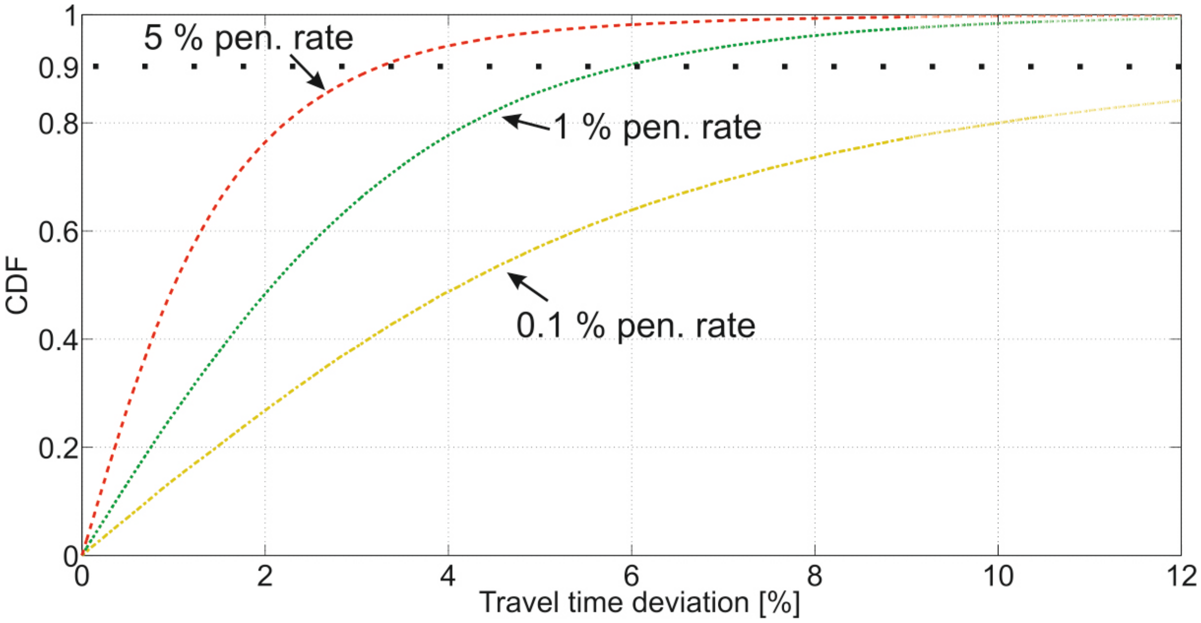


Figure 3: Traffic simulator: CDF of the travel time deviation

3.3 %, respectively 6 % at a 1 % penetration rate. That means, only a few percent penetration rate is sufficient for a reliable travel time estimation.

Conclusion and Outlook

In this report it has been shown that only a few percent FCD penetration rate is needed for a significant travel time estimation. This result could be strengthened with additional, more complex simulations, e.g. a simulation of a two-lane track with lane changing rules and different vehicle types.

References

[1] R. Barlovic, L. Santen, A. Schadschneider, and M. Schreckenberg. Metastable states in cellular automata. *Eur. Phys. J. B 5*, 1998.

[2] C. Ide, T. Knaup, B. Niehöfer, L. Habel, C. Wietfeld, and M. Schreckenberg. Efficient floating car data transmission via lte for travel time estimation of vehicles. *Proc. of the IEEE 76th Vehicular Technology Conference*, 2012.

[3] K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *J. Physique I 2, 2221-2229*, 1992.

Time-Efficient Evaluation of Position-Specific Communication and Localization Aspects

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

The accuracy of satellite positioning systems is increasing continuously. But, despite all progress in satellite and localization technology, there are still many elements of uncertainty in evaluating the accuracy of satellite positioning systems for a specific point on earth at a given time, like ionospheric refraction or effects due to the direct receiver's surrounding. To overcome these limitations, the Communication Networks Institute has developed a highly accurate simulation model which allows to determine the localization accuracy using raytracing technology and ionospheric considerations for a given satellite constellation at a random position on earth. In addition a Simulation-based decision support was developed, to enable a trade-off between minimizing the simulation time and maximizing the accuracy of the modeled scenario.

1 Simulation Approach

Based on the idea that it is not suitable to use one tool to cover all aspects to satellite signals, a combined framework of separated tools called *Multiscale Simulation Environment (MSE)* and a post-processing Matlab engine called *Satellite Positioning Accuracy Determination (SPAD)*.

Aiming on a position- and time-specific statement, the inclusion of ionospheric effects is necessary within an extension of the existing Matlab engine to not only rely on statistical

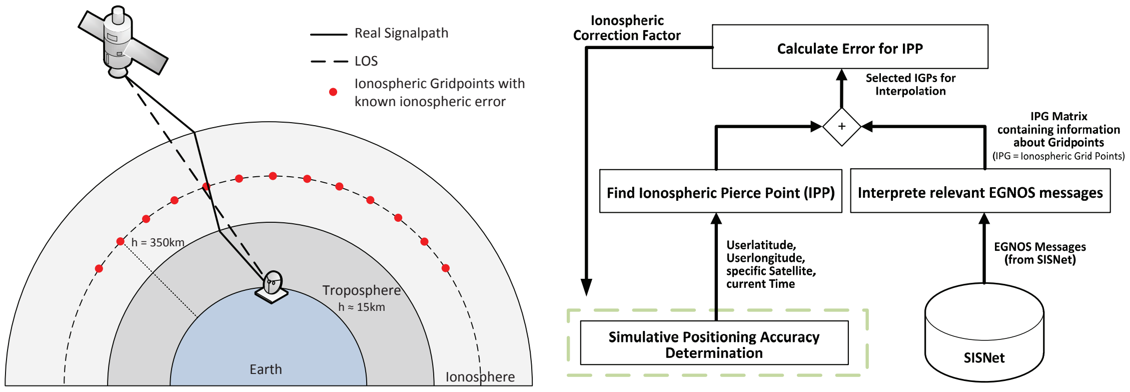

Real Signalpath
LOS
Ionospheric Gridpoints with known ionospheric error
h = 350km
Troposphere
h ≈ 15km
Earth
Ionosphere
Ionospheric Correction Factor
Calculate Error for IPP
Selected IGPs for Interpolation
+
IPG Matrix containing information about Gridpoints
(IPG = Ionospheric Grid Points)
Find Ionospheric Pierce Point (IPP)
Interprete relevant EGNOS messages
Userlatitude, Userlongitude, specific Satellite, current Time
EGNOS Messages (from SISNet)
Simulative Positioning Accuracy Determination
SISNet

Figure 1: Physical explanation of the ionospheric delay (left) and simulative implementation within the SPAD [1]

models and descriptions. Figure 1 left clarifies the evolution of the ionospheric delay. Factors delaying ionospheric radio signal propagation can include sporadic-E, spread-F, solar flares, geomagnetic storms, ionospheric layer tilts and solar proton events [1]. Specialized satellite payload enables measurements at the ground for defined points within the ionosphere, the *Ionospheric Grid Points (IGPs)*. Using this grid, a prediction of the additional delay at any given ionospheric intersection point, the so-called *Ionospheric Pierce Point (IPP)*, becomes possible by interpolating between the nearest given IGPs. Figure 1 right visualizes the calculation within the existing Matlab routine. The basic idea of the SPAD is already published in [3] but was enhanced for this contribution with detailed ionospheric considerations. Based on the time and the simulated position of the object and the satellites, a function calculates the IPP. In parallel, the latest EGNOS messages are downloaded from SISNeT, an Internet service for the latest IGP values. Afterwards, the nearest IGPs are extracted and their values are interpolated to reach a realistic statement for the calculated IPP. As a last step, the corresponding error for the given satellite/earth-receiver constellation is returned to the SPAD.

Using this knowledge for a specific position and time, a compensation via post-processing become possible [1].

2 Optimizing Computing-Time for Raytracing Analyses

The usage of raytracing technology within the positioning accuracy prediction is obvious, since satellite-based positioning is a kind of signal processing at a given point. Hence, every impact to any analyzed signal also affects the results. On the other hand, the assignment of raytracing is aggravated, based on the required computing time for high-detailed scenario models, which are necessary to evaluate all occurring effects. In addition, satellite communication signals have to overcome long distances. This affects the corresponding simulation settings. Raytracing technology is primarily designed to evaluate local scenarios and by that it is dimensioned for short or medium range distances. Hence, some modifications had to be fulfilled to guarantee a realistic satellite signal simulation and to enable a trade-off between minimizing the simulation time and maximizing the accuracy of the modeled scenario with two optimizing strategies explained below [2].

2.1 Optimizing Strategies

One obvious modification to reduce the computation time is the decrease of the used receiver grid granularity. Less receiver fields come along with less computation complexity, but also with a minor result resolution. This strategy, called Grid Up or GU(x) for short, will be used in different manifestations ('x' clarifies the enlargement of each side of one receiver field in meter). To extract the necessary computation time prognosis for a given scenario and the arising error, the results of the modified grid are compared to those of the non-modified one. To compensate the differing sizes of the receiver fields a corresponding weighting based on the percentage of the overlap for each originated/original field pair has been integrated.

The second strategy is based on the elimination of single or multiple receiver fields at the edges of the simulation scenario. By eliminating some outer fields of the scenario before starting the raytracing simulation, the complexity will be decreased. This strategy is called Receiver Reduction, or RR(y) for short, whereby the 'y' indicates the percentage of non-erased original receiver fields. To enable a calculation of the arising error a comparison to the original grid is necessary, whereby the modified one is always supplemented with NC fields at the position of every erased one.

Figure 2 shows a conditioning scatter plot of the possible time-gain and the corresponding wrongly determined receiving fields (*Receiver Error Ratio (RER)*) for the different scenarios using stand-alone strategies in different manifestation as well as combinations of those. Thereby, every deviation from reference results is counted. To enable a better comparison, five different scenario types (sizes S to XL) are plotted side by side. For each scenario and each used strategy, the measured time gain (circles) and the corresponding relative error (rectangles) are shown. Thus, those strategies should be preferred which display a small error and a huge gap between measured error and time gain. It is obvious that even with a tenable simulation error of $< 5\%$ a high economy of time is possible. Just to give an example: In a test-scenario of the exhibition hall in Cologne, Germany, a computing-time reduction of $> 75\%$ was possible, with an additional simulation error of just 3.5%. Of course, those possibilities increases with the size of the scenario, hence the possible savings in small scenarios are respectivly small. But nevertheless, using those comparable simple and non-ressource-intensive routines, even complex evaluation tools like raytracers may be used in time-critical applications.

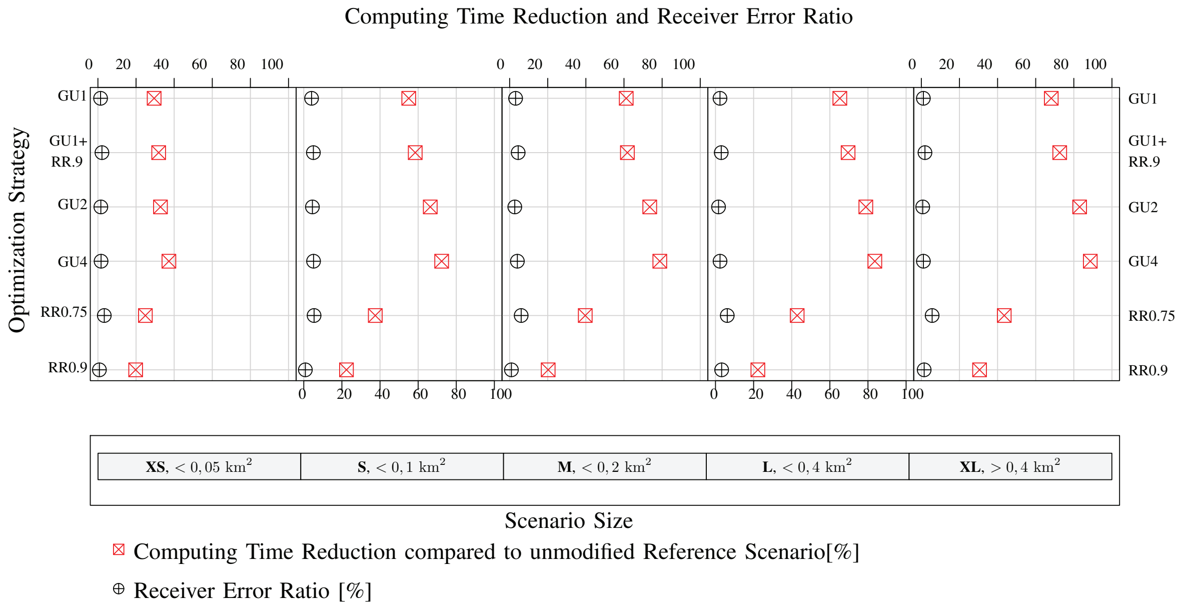


Figure 2: Conditioning Plot of possible Computing Time Gain and corresponding Simulation Error [2]

References

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

[2] Niehoefer, B., Lehnhausen, S. and Wietfeld, C. , *Optimizing-Strategies for a Time-Efficient Evaluation of Position-Specific Communication Aspects in Disaster Relief Scenarios*, accepted for publication at IEEE European Space Telecomunication (IEEE ESTEL), 2012, Rome, Italy

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

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

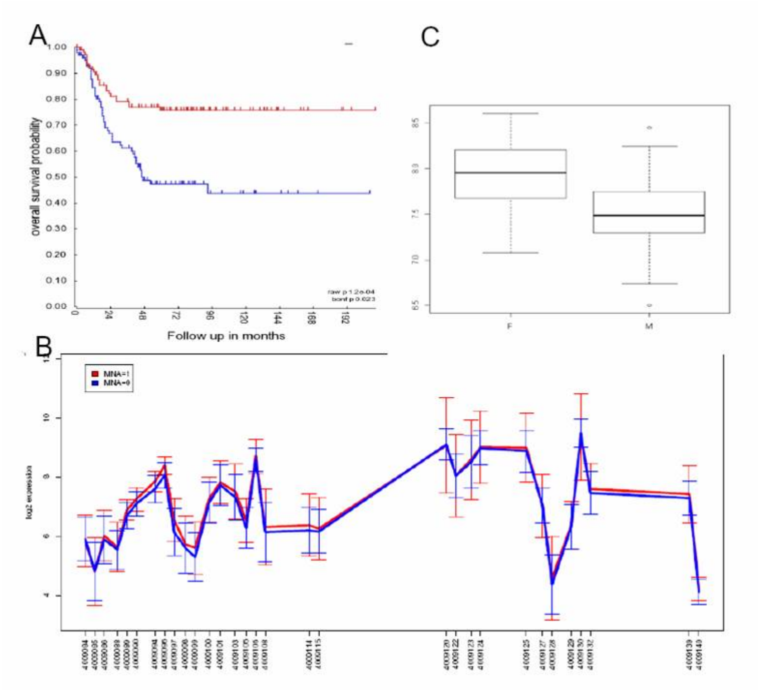
Katharina Morik   Alexander Schramm

The JARID1C histone demethylase is upregulated in aggressive neuroblastomas independent of MYCN amplification

Kathrin Fielitz
Oncology Lab – Children's Hospital Essen
Universität Duisburg-Essen
KaFielitz@googlemail.com

Neuroblastoma is the most common solid extracranial malignancy of childhood, accounting for 15% of the deaths attributed to malignancies in children. Since neuroblastoma shows a large clinical heterogeneity, with unfavourable tumors proceeding quickly and favourable tumors regressing, we are in need of identifying genes usable for outcome prediction. In this project we currently focus on the histone demethylase JARID1C, since it is upregulated in neuroblastomas with poor outcome independent of the MYCN oncogene. We were able to down-regulate JARID1C in neuroblastoma cell lines demonstrating that this protein is essential for cell survival.

Neuroblastoma is the most common extracranial malignancy of childhood, showing a large heterogeneity in clinical course. Unfavorable neuroblastomas proceed quickly, aggressively and fatally, while tumors with a favorable biology sometimes even regress without treatment. In spite of the consistently ongoing advancement in developing and improving diagnostic and therapeutic possibilities, it remains difficult to make a reliable prediction for the course of the disease or the patient's outcome. Routine diagnostics include determination of the amplification status of the MYCN oncogene, which is associated with poor outcome. In order to find the best possible therapy for each patient, new medical target structures have to be identified. In the last few years, there has been the approach of identifying genes, which could serve as medical target structures, through microarry and real-time PCR analyses [1] [2]. This study will introduce the Jumonji AT rich interactive domain 1C (JARID1C/ KDM5C/ SMCX) as one of the genes identified through



1.jpg

Figure 1: JARID mRNA expression and correlation with clinical features in neuroblastoma. A) Kaplan-Meyer analyses of patients with high (blue) and low (red) JARID1C expression in the primary tumor. B)Comparison of JARID1C expression in MYCN single copy (blue) and MYCN amplified (red) tumors C) comparison of JARID1C expression in female (F) and male (M) patients

microarray analyses and will further point out the biological relevance of this protein for neuroblastic tumors.

JARID1C was first identified as a cause of X-linked mental retardation [3]. It is a 26 exon gene, encoding for a 180 kDa protein from the JARID1 family [4].

Histones are highly alkaline proteins around which DNA winds, forming nucleosomes, that can undergo a posttranslational modification such as methylation or acetlylation. 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].

After a long lasting believe that histone methylation was enzymatically irreversible, the members of the JARID1 family were identified to have histone demethylating capacity [8] [9]. JARID1C in particular was identified to demethylate histone 3 lysine 4 (H3K4) from tri- to monomethyl [9].

We analyzed JARID1C expression in primary neuroblastoma. The Affymetrix exon arrays for this purpose were prepared from 113 patients [10]. Kaplan-Meier analyses revealed that patients with a high expression of JARID1C in the tumor, had a worse prognosis than

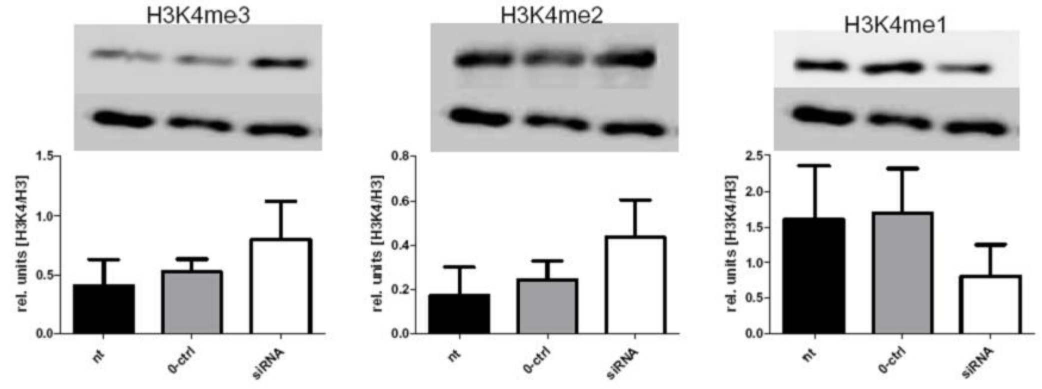


Figure 2: Results of western blot analysis for the methylation states of H3K4 after JARID1C knock down.The graphs show average values +/- SEM

patients with low levels of JARID1C (p= 1,2 x 10-4, Bonferroni correction for multiple testing p=0,0234). Yet there was no correlation to MYCN amplification, but a significant correlation to gender – female patients had significantly higher rates of JARID1C expression than males (p=3x10-16) (Fig.1), which can be attributed to the X-chromosomal location of the coding gene [3].

As already mentioned in the previous report, the protein expression of JARID1C is cell line specific. All the experiments were performed on the cell lines with the highest JARID1C expression in MYCN single copy or MYCN amplified cell lines. We analyzed molecular mechanisms induced through a siRNA mediated knock-down of JARID1C. We began investigating the effect of siRNA on RNA expression by the use of RT-qPCR. Our results show that mRNA levels of JARID1C were significantly decreased in both cell lines (p<0,0001). The protein levels were decreased significantly as well (p<0,005). Iwase and coworkers showed that JARID1C mediates the demethylation of H3K4me3 and H3K4me2 [9]. In order to show the enzymatic capacity of JARID 1C after the knock-down, we performed western blot analyses against three methylation states of H3K4. We were able to show that a transfection with siRNA directed against JARID1C leads to a shift in methylation. We see an increase of H3K4me3, no significant change in dimethylation of H3K4 and a decrease of the amount of momomethyl in H3K4 (Fig.2). Long non-coding RNAs (lncRNA) are non-protein coding transcripts longer than 200 nucleotides. MALAT1 (metastasis associated lung adenocarcinoma transcript 1) is one of these lncRNAs. We found an inverse correlation between JARID1C and MALAT1 expression by siRNA experiments, which was confirmed by RT-qPCR. In the further course we will try to find out through functional analyses, if MALAT1 has a significant contribution to neuroblastoma biology.

References

[1] Schramm, A.; Schulte, J.H.; Klein-Hitpass, L.; Havers, W.; Sievers, H.; Berwanger, B.; Christansen, H.; Warnat, P.; Brors, B.; Eils, J.; Eils, R.; Eggert, A.: Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. In Oncogene vol.: 24, 7902-7912, 2005
[2] Schramm, A.; Vandesompele, J.; Schulte, J.H.: Translating expression profiling into a clinically feasible test to predict neuroblastoma outcome. In Clnical Cancer Research, vol. 13: 1459 – 1465; 2007
[3] Jensen, L.R.; Amende, M.; Gurok, U.; Moser, B.; Gimmel, V.; Tzschach, A.; Janecke, A.R.; Tariverdian, G.; Chelly, J.; Fryns, J.P.; Turner, G.; Reinhardt, R.; Kalscheuer, V.M.; Ropers, H.H.; Lenzner, S.: Mutations in the Jarid 1c gene, which is involved in transcriptional regulation and chromatin remodelling causes X-linked mental retardation. In The American Journal of Human Genetics, 76, 227 – 236; 2005
[4] Klose , R.J.; Kallin, E.M.; Zhang, Y.: JmjC-domain containing proteins and histone demethylation. In Nature Reviews – Genetics, Vol. 7, pgs. 715 – 727; 2006
[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. In 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. In 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. In Proceedings of the National Academy of Sciences vol. 101; no. 19, pgs. 7357 – 7362, 2004
[8] Christensen, J.; Agger, K.; Cloos, P.A.C.; Pasini, D.; Rose, S.; Sennels, L.; Rappsilber, J.; Hansen, K.H.; Salcini, A.E.; Helin, K.: RBP2 belongs to a family of demethylases, specific for tri- and dimethylated lysine 4 on histone 3. In Cell, Vol. 128, 1063 – 1076, 2007
[9] 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/Jarid 1i defines a family of histone H3 lysine 4 demethylases. In Cell, 128, 1077 – 1088; 2007 [10] Schramm, A.; Schowe, B.; Fielitz, K.; Heilmann M.; Martin M.; Marschall, T.; Köster, J.; Vandesompele, J.; Vermeulen,J.; de Preter, K.; Koster, J.; Versteeg, R.; Noguera, R.; Speleman, F.;Rahmann, S.; Eggert, A.; Morik, K.; Schulte, J.H.: Exon-level expression analyses identify MYCN and NTRK1 as major determinants of alternative exon usage and robustly predict primary neuroblastoma outcome. In British Journal of Cancer. 2012

Functional validation of transcripts with alternative exon usage in neuroblastoma

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

Neuroblastoma is an embryonal cancer of the sympathetic nervous system and diagnosed in early childhood. The tumor originates from precursor cells of the peripheral nervous system and arises in a paraspinal location in the abdomen or chest. The clinical presentation of this tumor can be very heterogeneous. The international Neuroblastoma staging system (INSS) classified 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 therapy targets will be identified to enable a better outcome prediction. It will especially concentrate on transcripts with alternative exon usage because this modification has been shown to contribute to various diseases including cancer. One candidate gene already identified by our studies is cyclin B1, a protein with two different isoforms exhibiting different correlation with outcome.

Neuroblastomas (NB) are the most common and deadly solid tumors in childhood. They account for 7-10 % of all childhood cancers. NBs derive from the neural crest tissue and usually arise in a paraspinal location in the abdomen or chest [1,2]. The median age at diagnose is 17 months and the incidence of neuroblastoma is 10.2 cases per million children under 15 years [3,4]. This kind of cancer exhibits diverse and often dramatic clinical behavior. To enable an individual therapy, it is essential to extend the knowledge of the molecular medicine. 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 associated with a poor outcome whereas a high expression of the neurotrophin

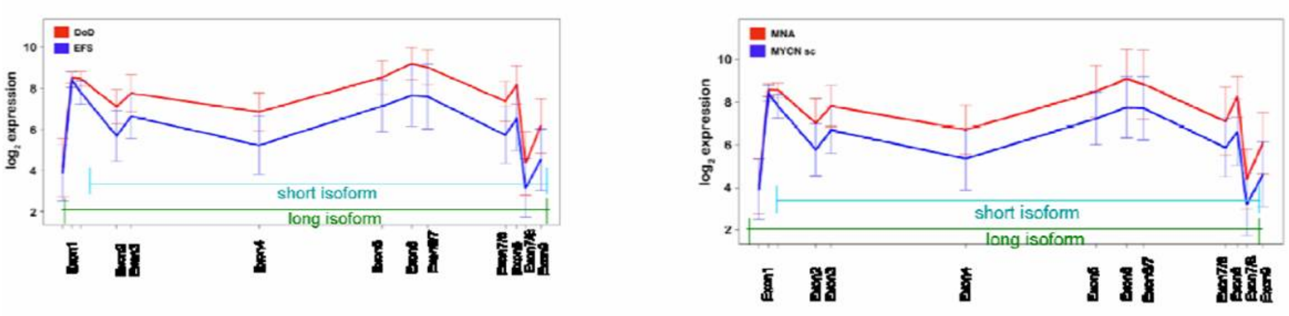


Figure 1: Analysis of Exon Arrays for CCNB1 [9] The expression of the short isoform of CCNB1 is increased in patients, who died of disease and in Neuroblastoma with MYCN amplification.

receptor TrkA is a favorable indicator [1]. To contribute to a better risk assessment this project will deal with the identification of NB relevant genes and transcripts. A special focus is set on genes with an alternative exon usage. The usage of different exons could be established by alternative splicing or by alternative promotor usage. This change in mRNA composition causes the synthesis of different protein variants or the unbalanced expression of normal protein isoforms and this could initiate or sustain tumor growth [5]. Guo et al. 2011 could show that alternative splicing plays a significant role in high stage NB and they suggested a MYCN-associated splicing regulation pathway [6]. The search of such genes is based on Affymetrix Exon array (HuEx 1.0 ST) data, providing high resolution expression profiles of NB.

Among the transcripts associated with patient outcome, we identified cyclin B1 (CCNB1), a cell cycle regulator activating the cyclin dependent kinase1 (cdk1) and promoting the passage through G2-phase to M-phase. Deregulated expression of this gene is supposed to be involved in neoplastic transformation [7]. So it could be a very interesting target for cancer therapies. The microarray shows that NB express both a long and a short isoform. According to the literature the long isoform is constitutively expressed whereas the short form is cell cycle dependent and predominantly expressed during G2/M-phase. These different transcripts are due to alternative promotors [8]. Previous analyses indicate that the short isoform is higher expressed in patients who died of disease and in NB with MYCN amplification (Fig.1). RT-PCRs showed that various NB cell lines express different levels of these isoforms and also the ratio between these isoforms is cell line dependent, although a correlation between MYCN and CCNB1 expression could not be observed [9]. It can be hypothesized that the isoforms have different functions in the cell so in the next steps we will check the functions of both transcripts. Therefore we knocked-down the long isoform and both isoforms by siRNA and perform cell viability assays. In both cases a decrease in cell viability is recognisable (Fig.2). Furthermore 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. First I determined the IC50-values, the concentration of the inhibitor at which the cell viability

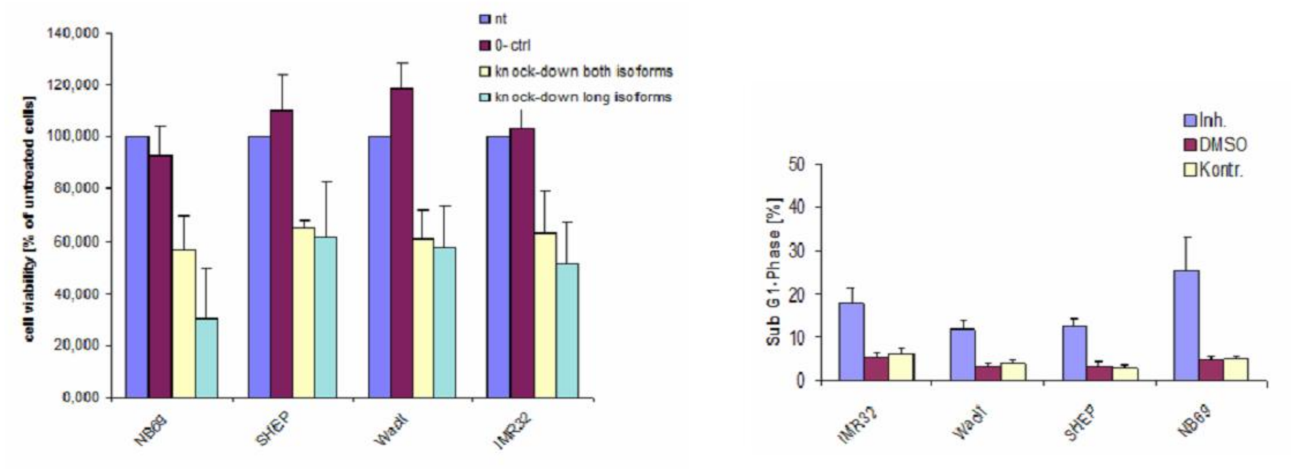


Figure 2: Knock-down of CCNB1 causes a decrease in cell viability and the treatment of the cdk1/cyclinB1 inhibitor causes an increase in apoptosis.

is reduced to 50 %, but the values do not correlate with the MYCN expression or the CCNB1 isoforms expression pattern. Moreover I investigated the rate of apoptosis of the inhibitor treated cells. The sub G1-phase, a proof for apoptopic cells, is increased under treatment of the cdk1-inhibitor (Fig.2). Kensuke Kojima et al. suggested that the inhibitor enhances mitochondrial apoptosis and Kreis et al. observed an enhanced p53 protein level and an increase in its downstream target p21 [10,11]. An increase in p21 expression could also be observed by PCR. The next steps will be checking the p53 expression as well as the mitochondrial apoptosis.

References

[1] Brodeur G.M. Neuroblastoma biological insight into a clinical enigma. Nature Reviews 2003, 3:203-216
[2] Hoehner JC, Gestblom C., Hedborg F., Sandstedt B., Olsen L., Pahlman S. A developmental model of neuroblastoma: differentiating stroma-poor tumors' progress along an extra-adrenal chromaffin lineage. Lab Invest 1996, 75:659-75
[3] London W.B., Castleberry R.P., Matthay K.K., Look A.T., Seeger R.C., Shimada H., Thorner P., Brodeur G., Maris J.M., Reynolds C.P., Cohn S.L. Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the Children's Oncology Group. J Clin Oncol 2005, 23:6459-65
[4] Maris J.M. Recent advances in neuroblastoma. N ENG J MED 2010, 362:2201-11
[5] 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
[6] 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
[7] Yuan J.; Yan R.; Krämer A.; Eckerdt F.; Roller M.; Kaufmann M.; Strebhardt K. Cyclin B1 depletion inhibits proliferation and induces apoptosis in human tumor cells. Oncogene, 2004, 34: 5843–5852
[8] Hwang A., McKenna W.G., Muschel R.J. Cell cylce-dependent usage of transcriptional start sites. The Journal of Biological Chemistry 1998, 47:31505-31509
[9] Schramm A., Schowe B., Fielitz K, Heilmann M., Martin M., Marschall T., Köster J., Vandesompele J., Vermeulen J., de Preter K, Koster J., Versteeg R., Noguera R., Speleman F., Rahmann S., Eggert A., Morik K., and Schulte J. H. Exon -level expression analyses identify MYCN and NTRK1 as major determinants of alternative exon usage and robustly predict primary neuroblastoma outcome. BJCancer 2012
[10] Kojima, Kensuke; Shimanuki, Masaya; Shikami, Masato; Andreeff, Michael; Nakakuma, Hideki Cyclin-dependent kinase 1 inhibitor RO-3306 enhances p53-mediated Bax activation and mitochondrial apoptosis in AML. Cancer Science 2009, 6: 1128–1136
[11] Kreis, N. N.; Sanhaji, M.; Krämer, A.; Sommer, K.; Rödel, F.; Strebhardt, K.; Yuan, J. Restoration of the tumor suppressor p53 by downregulating cyclin B1 in human papillomavirus 16/18-infected cancer cells. Oncogene 2010, 41: 5591–5603

Algorithms for the Investigation of Genotype and Phenotype

Dipl.-Inf. Johannes Köster
Genome Informatics, Institute of Human Genetics
Faculty of Medicine, University of Duisburg-Essen.
johannes.koester@tu-dortmund.de

We continue the development of protein hypernetworks, especially toward aquisition of the required data. Further, we work on the implementation of a GPGPU-based read mapping algorithm, and on a novel text-based workflow management system, called Snakemake.

The genome contains the hereditary information of an organism. It consists of genes, that encode proteins that are built by the translational machinery. Individual organisms can be differentiated by their genome, or genotype. Even between two individuums of the same species the genotype differs. By encoding proteins, the genotype determines the phenotype of a cell to a large extent. In this scope, the phenotype is the entirety of physical and functional properties that are expressed by a cell. Rather than emerging directly from individual genes, these properties are generated by the cooperation of multiple proteins in large networks. On the one hand, these networks show complex regulation mechanisms among proteins, for example by allosteric regulation or competition on binding sites. On the other hand, they can regulate genes, allowing the cell to react on external signals by changing the expression of genes. Moreover, gene regulation again can have an effect on the regulatory network itself. Modern high throughput technologies allow large scale studies of both views. We aim to investigate solutions to improve genotype and phenotype analysis in various ways.

Genotype analysis using high-throughput sequencing is still in its infancy. The thesis may address several weaknesses of current analysis pipelines. E.g. mapping RNA-reads using a GPGPU to provide increased performance. In phenotype – i.e. protein network – analysis, predictive models suffer either from being too detailed to stay feasible for large-scale data (e.g. differential expression based models) or they are not able to capture the functional implications by regulatory mechanisms (e.g. graph based models). Based on a

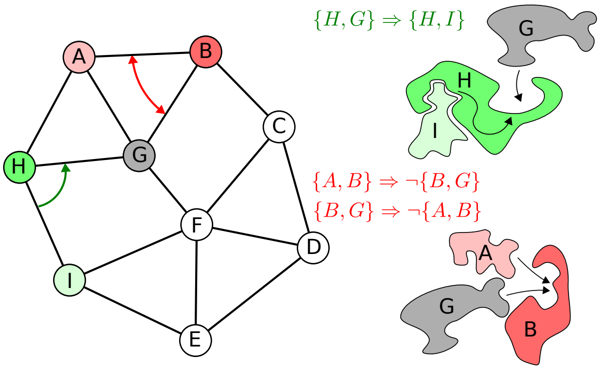

A
B
C
D
E
F
G
H
I
$\{H,G\} \Rightarrow \{H,I\}$
G
H
I
$\{A,B\} \Rightarrow \neg\{B,G\}$
$\{B,G\} \Rightarrow \neg\{A,B\}$
A
G
B

Figure 1: Protein hypernetwork. A graph based protein network representation (left) is extended by propositional logic constraints, that model interaction dependencies like allosteric activation of interactions and competition on binding domains (right).

diploma thesis, a novel functionally predictive model for protein networks, called "protein hypernetworks" shall be improved and established. Ultimately genotype and phenotype analysis may be combined to provide a deeper insight into biological mechanisms.

The algorithms and models are mainly investigated in two cooperative projects. First, protein hypernetworks are used to analyse the human adhesome network together with the Max-Planck-Institute of Molecular Physiology Dortmund (Dr. Eli Zamir). Here, model-based predictions are verified in cooperations with system biologists using orthogonal experimental techniques like Fluorescence Microscopy. Second, genotype approaches are verified in cooperation with the University Hospital Essen (Dr. Alexander Schramm, Prof. Johannes Schulte) and applied to the analysis of neuroblastoma cancer.

Protein hypernetworks have been shown to improve the prediction of protein complexes and the functional importance of proteins. Protein hypernetworks extend protein network graphs with propositional logic constraints (see Figure 1). The constraints are able to capture a major regulatory mechanism, that we call protein interaction dependencies. Among these are allosteric effects, competition on binding domains, steric hindrance and phosphorylations. In the past months focus was put on harvesting these constraints from literature. For the human adhesome network, 71 new interaction dependencies were extracted from 59000 related full-text scientific publications (see Figure 2). For this purpose, a combination of tokenization of relevant words and regular expression patterns was used [2]. In a current bachelor's thesis, the Quine-McCluskey-Algorithm is used to infer boolean logic constraints from truth tables filled with observations of simultaneously measured protein interactions. The measurement of more than one interaction simultaneously using fluorescence spectroscopy is possible in theory, but will need some time to be reliably established. Therefore, in addition we try to infer above truth tables from the integration of protein complex measurements with binary protein interaction networks.

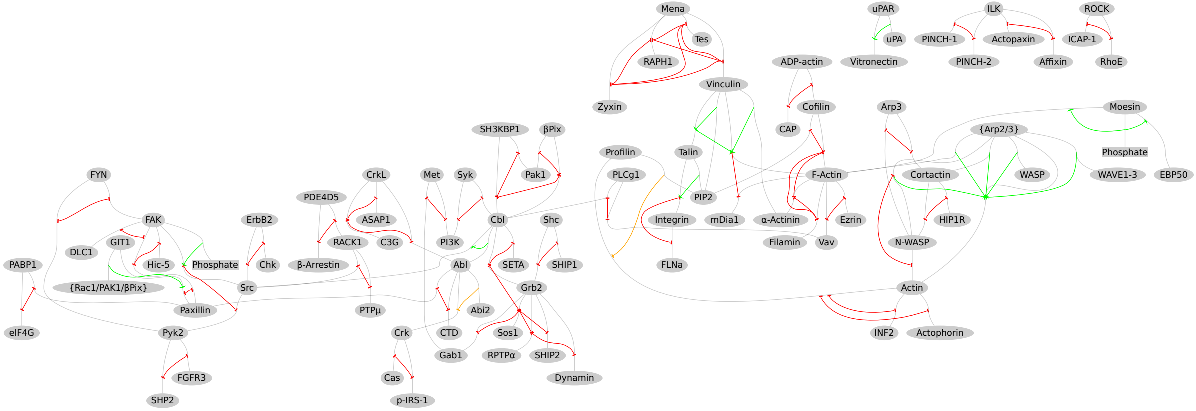


Figure 2: Network of human adhesome interactions with interaction dependencies gained by our text-mining approach [2].

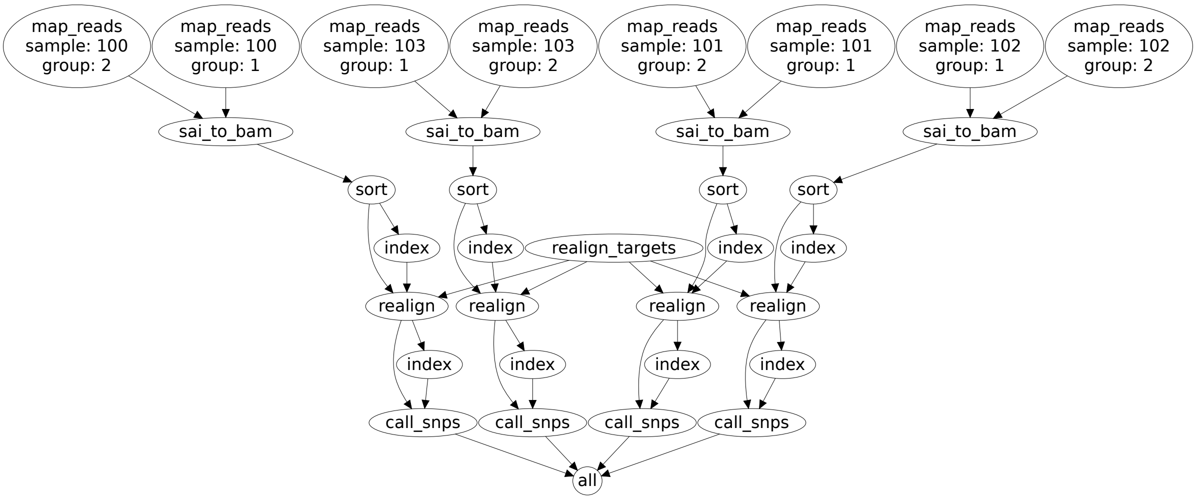

map_reads sample: 100 group: 2
map_reads sample: 100 group: 1
map_reads sample: 103 group: 1
map_reads sample: 103 group: 2
map_reads sample: 101 group: 2
map_reads sample: 101 group: 1
map_reads sample: 102 group: 1
map_reads sample: 102 group: 2
sai_to_bam
sai_to_bam
sai_to_bam
sai_to_bam
sort
sort
sort
sort
index
index
realign_targets
index
index
realign
realign
realign
realign
index
index
index
index
call_snps
call_snps
call_snps
call_snps
all

Figure 3: A sample workflow generated using Snakemake [1].

Bioinformatics, especially the analysis of genotype information from high-throughput sequencing, requires the chained execution of many different analysis tools and format conversions. To ensure reproducibility and documentation, a workflow management is important. Execution and modification of workflows on headless servers and clusters and collaborative editing via SVN are important arguments for a text-based workflow management. In 2012, Snakemake [1], a novel text-based workflow language and execution environment was developed to replace the dated an insufficient GNU Make that was used previously in our group. From atomic rules, that execute a specific task and produce output files from input files while allowing wildcards, Snakemake automatically determines a DAG of rule jobs to be executed in order to obtain the desired output files (see Figure 3). Thereby, independent paths in the DAG can be executed in parallel. Snakemake scales well from single core machines to clusters without the need to adapt the workflow.

Finally, a GPGPU based read-mapper using OpenCL was further developed. Read mapping was split into mapping 32 bit seeds rapidly on the GPU, and performing a full Smith-Waterman alignment for only few matches on the CPU. In contrast to current read-mappers like BWA, the algorithm promises to perform linearly with the number of allowed errors.

References

[1] Johannes Koester and Sven Rahmann. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*, August 2012.

[2] Johannes Köster, Eli Zamir, and Sven Rahmann. Efficiently mining protein interaction dependencies from large text corpora. *Integrative Biology*, June 2012.

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

Roland Fried Wolfgang Rhode

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 [13]

including some updates. As well as new catalogues like massive binary systems based on the Einstein catalogue [14].

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 [12] [11].

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 [15] for RapidMiner [10] 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 480 data events is shown in Figure 2. The events were selected by the stacking algorithm out of scrambled data, w.r.t. a catalogue consisting out of six Milagro sources. Scrambled means the azimuth angle has been randomized to ensure the blindness. As this results in only using atmospherical events, one can compare the result with the theoretical predictions for that flux.

The data points align perfectly with the prediction by Honda [9] for an atmospherical flux. Usual unfolding analyses in IceCube use approx. 40 000 Events and are therefore able to produce smaller bins and cover a larger energy range.

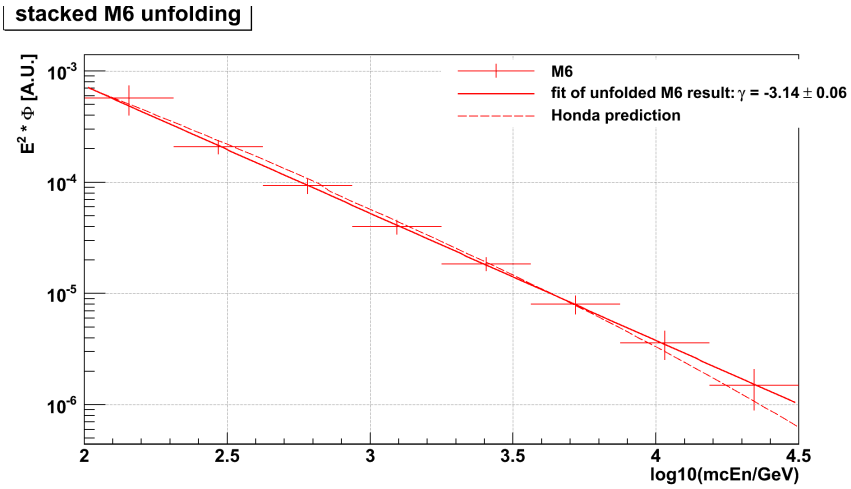


Figure 1: The unfolding result of 480 scrambled data events is presented in red, as well as the corresponding fit. The distribution is proportional to a power law with a slope $\gamma = -3.14$. The theoretical prediction is shown in dashed red.

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] Morihiro Honda et al. Calculation of the flux of atmospheric neutrinos. *Physical Review D*, 52:4985, 1995.

[10] 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.

[11] N Milke, M Doert, and W Rhode. Solving time-dependend inverse problems with truee: Examples in astroparticle physics. *Paper submitted to NIM*.

[12] 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*.

[13] 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.

[14] A. M. T. Pollock. The Einstein view of the Wolf-Rayet stars. *Astrophysical Journal*, 320:283–295, September 1987.

[15] 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.

Energy loss mechanisms in PROPOSAL

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

In this report a short summary of the Monte Carlo program PROPOSAL is given. Particular attention is paid to the implemented cross sections and the role of PROPOSAL in the Monte Carlo chain.

1 Introduction

IceCube is a large scaled neutrino detector located near the geographic south pole. Within an area of 1 km^3 it consists of 86 strings with a distance of 125m between each of them [4]. Digital Optical Modules (DOMs) with which Cherenkov light can be detected are fixed to every string [2]. By measuring this light, it is possible to reconstruct the path of a neutrino. IceCube uses the physical effect that neutrinos produce charged leptons while they propagate through the ice. These leptons can be detected in the DOMs via the Cherenkov light that they emit on their way through the ice.
Monte Carlo simulations demonstrate an approved and successful method to analyse the huge amount of data consisting of much more background than real signal events. Within the Monte Carlo simulation chain, the path and all interactions from the neutrino to the detected Cherenkov photonare calculated.

2 The IceCube Monte Carlo chain

The IceCube Monte Carlo chain consists of three parts:

1. the generator which describes the propagation of the cosmic particles from the top of the atmosphere down to the surface of the ground,

2. the propagator which simulates the propagation of the leptons from the surface down to and through the detector
3. and within the detector interaction the Cherenkov photons, produced by the leptons in the environment of the detector, and their interaction with the detector himself are simulated.

At the moment, IceCube uses the propagator Muon Monte Carlo (MMC) [1] written in Java. The sucessor program of MMC is PROPOSAL. It is written in C++ and provides the same precision and a better velocity than MMC based on the same numerical procedures. An advantage of a program written in C++ is the better possibility to integrate it into the IceCube Monte Carlo chain, which is basically written in C++. Furthermore this programming language guarantees easier applicability in other experiments besides IceCube. To ensure a high precision of the program, the physical interactions must be understood and investigated correctly. So, a detailed investigation of the used interactions and their cross sections was made in my diploma thesis and a short overview will be given in the next chapter.

3 Energy loss mechanisms in PROPOSAL

The energy loss of a particle depends on its initial energy, the medium through which the particle propagates, the type of the particle and the type of the interaction. Ionization, bremsstrahlung, pair production and photonuclear interaction are the four types of energy loss interactions considered in PROPOSAL.

3.1 Ionization

Charged particles with low initial energy loose their energy while passing through a medium especially by ionization. This energy loss can be described by the Bethe Bloch equation, which depends on the charge of the particle, the charge of the nuclei of the medium and the energy of the initial particle [3]. Ionization has the smallest contribution to the total energy loss for high energies ($> 10\,\mathrm{TeV}$) and it is the only significant energy loss for small energies ($< 1\,\mathrm{GeV}$).

3.2 Bremsstrahlung

Bremsstrahlung occurs when a charged lepton moves through the field of a nucleus. As a result of this interaction the lepton emits a photon. In PROPOSAL there are four different parametrizations of the Bremsstrahlung cross section implemented. The choice of these parametrizations depends on the energy and the type of the initial particle. One of these parametrization fits best for electrons, the other ones are more suitable for muons.

3.3 Photonuclear interaction

The photonuclear interaction describes the inelastic interaction of a lepton with an atomic nucleus. Three different parametrizations of this interaction are implemented in PROPOSAL. Compared to [3] the relative errors of the used default parametrization are less than 5% for energies $E > 100\,\text{GeV}$ but they become bigger for energies of $E \approx 100\,\text{TeV}$. Accordingly, this interaction has the biggest errors compared to the others.

3.4 Pair production

The production of an electron pair in the field of an atomic nucleus is described by this cross section. It is the major contribution of all energy loss mechanisms for high energies ($> 1\,\text{TeV}$). Another possible interaction is the muon pair production which is not considered in PROPOSAL, because the cross section is estimated to be $2 \cdot 10^4$ [5] times smaller than the direct electron pair production cross section. Therefore it can be neglected.

4 Conclusion and outlook

Figure 1 shows a comparison of the simulated data to reference values [3]. It is obvious that the PROPOSAL values fit very well to the reference values in the whole energy range for every interaction.

In the next future PROPOSAL will be included in the IceCube Monte Carlo chain as the default propagator. A further step is to run PROPOSAL on GPUs to reduce the total runtime by parallelization. The first part to accomplish this task is already achieved. Finally PROPOSAL could also be used by other experiments where leptons are propagated through a medium.

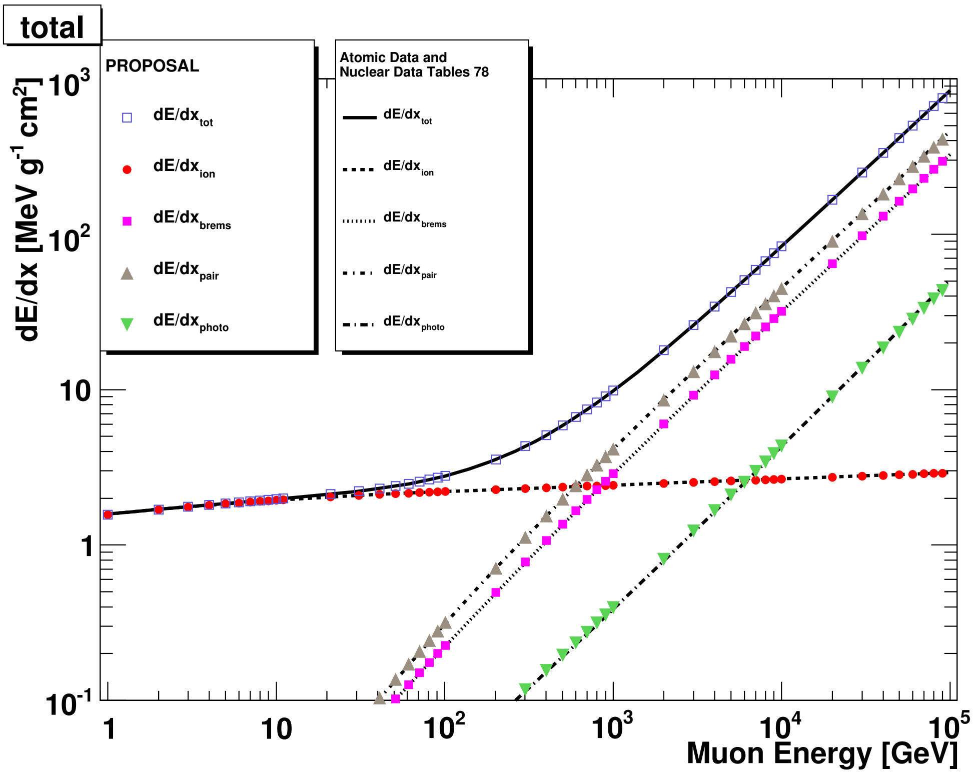


Figure 1: Continuous energy losses of a muon by each interaction and totally compared to reference values [3]

References

[1] D. Chirkin and W. Rhode. *arXiv:hep-ph/0407075*, July 2004.

[2] IceCube Collaboration. *Astrophysical Journal*, 745:45, January 2012.

[3] D. Groom and S. Klein. *The European Physical Journal C - Particles and Fields*, 15:163, 2000.

[4] A. Karle et al. IceCube - the next generation neutrino telescope at the South Pole. *Nuclear Physics B Proceedings Supplements*, 118:388–395, April 2003.

[5] S. Kel'ner, R. Kokoulin, and A. Petrukhin. *Physics of Atomic Nuclei*, 63:1603, 2000.

Development of a Monte-Carlo simulation for lepton propagation

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

IceCube is a large scale neutrino detector located at the South Pole. The one cubic kilometer detector volume consists of the South Pole ice, which has excellent optical properties [1]. IceCube uses the physical effect, that neutrinos interacting with a media produce charged leptons such as muons, electrons and taus. These leptons propagate through the detector and emit Cherenkov light, which is detected by high sensitve photon sensors [3].

The complexity to analyse the data is, that leptons coming from the atmosphere produce a similar signal in the detector. The outcome of this is a huge amount of background which overlaps the neutrino signal. The ratio of signal to background is about one to a million.

To analyse the data and to find neutrino signals, Monte-Carlo simulations are essential.

1 Simulations in IceCube

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

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

Propagators take the generated particles and simulate their behaviour while propagting through the detector. 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 written 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**opagator with **O**ptimal **P**recision and **O**ptimized **S**peed for **A**ll **L**eptons).

2 Status and Plans

The development of PROPOSAL is completed. Also a detailed documentation of the C++ code was included. The excellent agreement of PROPOSAL and MMC is shown in figure 1.

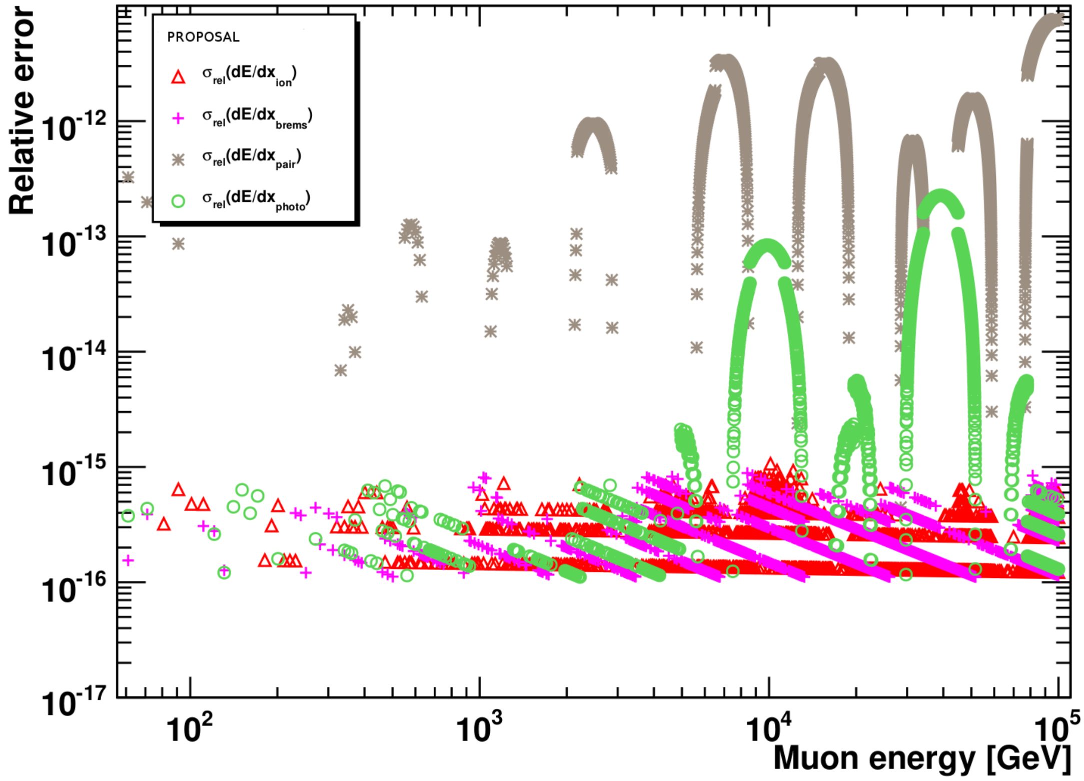


Figure 1: The relative error of PROPOSAL and MMC is shown using the example of the energy loss per distance. Due to a different double precision in Java and C++ the result of MMC and PROPOSAL are slightly different.

The first tests using PROPOSAL inside the IceCube Monte-Carlo chain are done. Figure 2 shows the agreement of MMC and PROPOSAL using the example of the muon track length inside the IceCube detector. According to different random number generators in Java and C++ small but statistical not significant differences appear.
The next step is to make PROPOSAL the default propagator in the IceCube simulation.

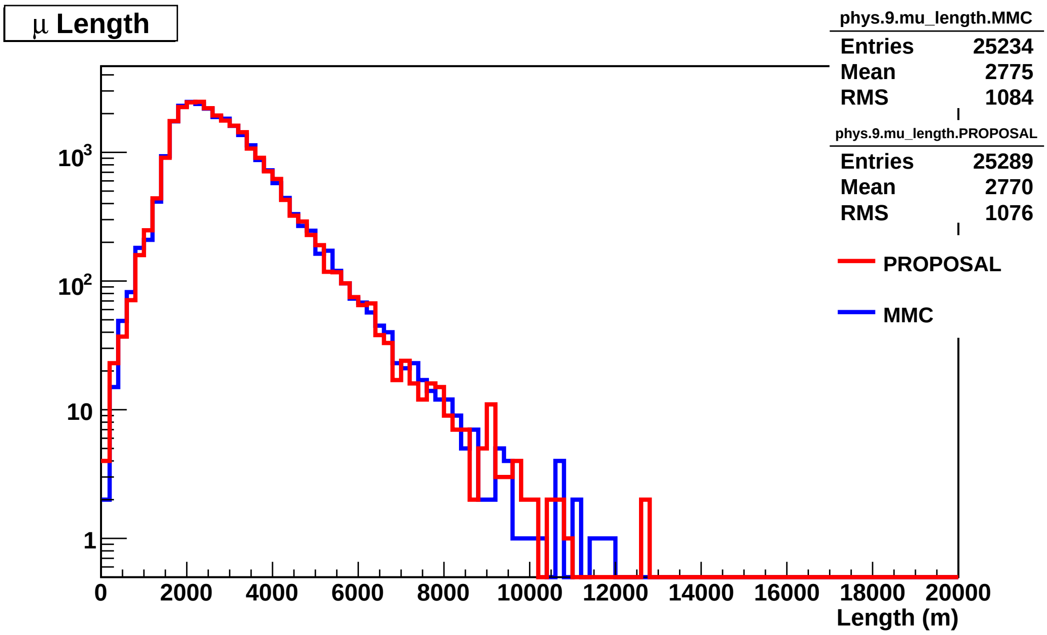


Figure 2: The simulated track length of muons is shown. PROPOSAL and MMC produce the same results within statistical uncertainties

Because parallelization will speed up the simulation a lot and therefore save computing time and money, it is planed to implement PROPOSAL for GPUs. Here the first tests were done by Tomasz Fuchs.

References

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

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

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

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

Event classification and spectral reconstruction of the IceCube 59 full year data

Natalie Milke, Martin Schmitz
Lehrstuhl für Experimentelle Physik 5b
Technische Universität Dortmund
natalie.milke@tu-dortmund.de, martin.schmitz@tu-dortmund.de

The determination of the atmospheric neutrino flux spectrum is important for the neutrino astronomy as its distribution at high energies can shed light on the predicted flux of extragalactic neutrinos. IceCube is a cubic kilometer large neutrino telescope located at the geographic South Pole and is well suited for the detection of high energy neutrinos. Due to the low neutrino cross section the ratio between detected neutrinos and the undesired atmospheric muons is unfavorable. To obtain a pure neutrino sample the neutrino events are selected by applying a sophisticated classification algorithm Random Forest within the framework Rapid Miner. The spectral reconstruction is carried out by using the unfolding algorithm TRUEE. While in the last report the atmospheric neutrino analysis of 10 % of IceCube 59 data has been introduced, here we use the full data sample and show the results of Natalie Milkes analysis. The range of the neutrino spectrum has been extended to higher energies. Further preliminary studies on 10 % of the data taken with the larger IceCube 79 have been performed by Martin Schmitz.

The atmospheric neutrinos are produced in the interactions of cosmic rays with the Earth's atmosphere and represent a background for the expected extragalactic neutrinos. Those would help to trace the sources of the cosmic rays, but have not been detected yet. The atmospheric neutrino flux spectrum is steeper than the spectrum of extragalactic neutrinos [5]. Hence, the flux of extragalactic neutrinos would cause a flattening of the spectrum at high energies. Therefore, the precise estimation of the neutrino flux spectrum at high energies is essential to identify a possible extragalactic contribution to the whole spectrum.

The one cubic kilometer large neutrino telescope IceCube [1] has been constructed to detect neutrinos with energies beyond 10^6 GeV by measuring the properties of their secondary muons. IceCube is located at the geographic South Pole and in its final configuration consists of 5160 digital optical modules (DOM) arranged along 86 strings in the depth between 1450 m and 2450 m in the glacial ice. While travelling 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. During its construction over several years, the partially built IceCube detector took data in different configurations. In the previous technical report of 2011 only 10% of the IceCube data has been used, taken with the 59 string configuration (IC 59), to develop and optimize the analysis procedure [10]. Here we present the analysis of the full 100 % measurement of IC 59 and give an outlook on the atmospheric neutrino analysis using data, measured with the larger IceCube detector configuration IC 79.

During the measurement of the relevant neutrino-induced signal events a big amount of the undesired events from atmospheric muons are recorded and have to be suppressed. For this purpose the cut on the zenith distribution and the velocity of the muon track is applied. The remaining atmospheric muons are removed by the multivariate classification method Weka-Random Forest [3], [6], included in the framework RapidMiner [8]. The output is the confidence for an event to be signal and is a value between zero and one. We apply the cut on confidence at one to obtain a neutrino sample with minimum purity of 99.4 % and thus a negligible amount of atmospheric background muons. Since the analysis has been developed using 10 % of data, Fig. 1 shows the full confidence range for 10 % and a zoom into high confidence region for the full data sample. For more details

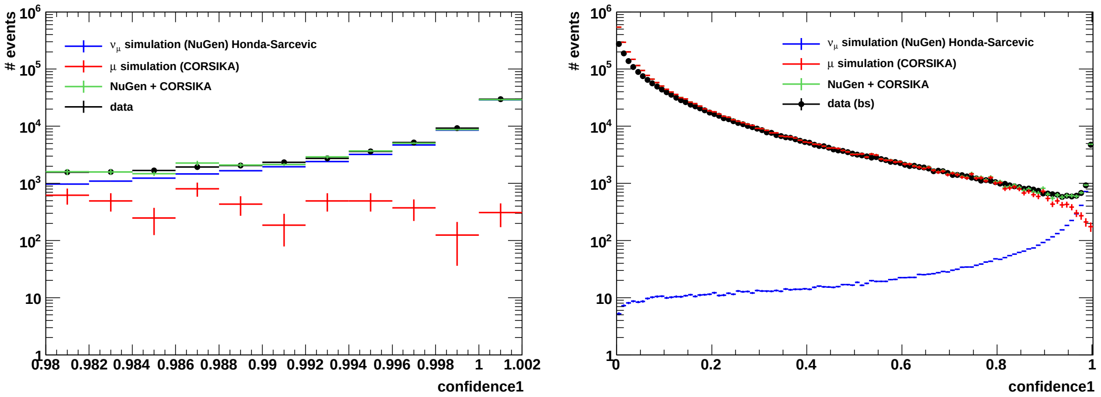


Figure 1: Output of the Random Forest as confidence for signal. Left picture shows 10 % of IC 59 data and the right picture the zoom of the high confidence region of the full data sample. The final neutrino sample is obtained after the cut at confidence = 1. The comparison of data with the simulated signal events (NuGen) and background muons (CORSIKA) shows a good agreement.

about the application of Weka-Random Forest in the IceCube analysis see also the report of Tim Ruhe.

The spectral reconstruction of the neutrino flux is performed using the TRUEE unfolding program [9], which is developed within the Collaborative Research Center SFB 823 and has been introduced in the previous report of 2011.

Three measured energy-correlated attributes (observables) are used together with the unfolding software configuration, developed with the 10 % of data, as presented in the previous report of 2011. The final result of the spectral reconstruction of the atmospheric neutrino flux is shown in Fig. 2. As can be seen from the color-coded error bars, the statistical uncertainties estimated by TRUEE are small. This means, that the event classification enables to obtain high statistics at high energies and the unfolding with TRUEE provides statistically independent data points with the developed software configuration. The previous neutrino flux spectrum, determined with the IC 40 configuration is shown as well and demonstrates the extension of the energy range by the new analysis.

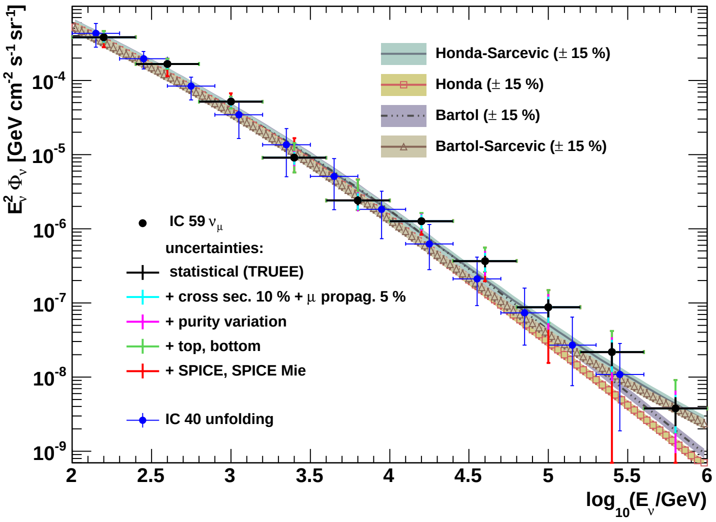


Figure 2: Unfolded atmospheric neutrino flux spectrum of the full IC 59 data. The presented uncertainties are color-coded due to their origin. The black error bars indicate the statistical uncertainties determined by TRUEE. For comparison the result of the previous analysis with the IC 40 data is shown in blue. Four different theoretical flux model combinations, predicted by Honda [7], Sarcevic [4] and Bartol [2], are shown as well. The spectrum is weighted by squared energy for a better illustration.

The analysis of 10 % of the IC 79 data has started. Comparing IC 59 and IC 79 you can see that we have a more symmetric detector and better reconstruction algorithms. This results in the fact that we can release the velocity and zenith angle precut. This and the bigger detector leads to much higher size of the neutrino sample. To get a final sample the cut on the signalness needs to be chosen carefully to get enough events with a very high purity. The unfolding is currently tested on Monte-Carlos. First tests are showing that higher energies than 10^6 GeV are unfoldable by having as least the same resolution than the IC 59 unfolding. The actual energy bounds and resolution is under investigation and depends on the chosen energy correlated variables.

References

[1] R Abbasi et al. IceCube - Astrophysics and Astroparticle Physics at the South Pole. 2011.

[2] G D Barr et al. A Three - dimensional calculation of atmospheric neutrinos. *Phys.Rev.*, D70:023006, 2004.

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

[4] R Enberg, M H Reno, and I Sarcevic. Prompt neutrino fluxes from atmospheric charm. *Phys.Rev.*, D78:043005, 2008.

[5] E Fermi. On the origin of the cosmic radiation. *Phys. Rev.*, 75:1169 – 1174, 1949.

[6] M Hall et al. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[7] M Honda et al. Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data. *Phys.Rev.*, D75:043006, 2007.

[8] I Mierswa, M Wurst, R Klinkenberg, M Scholz, and T Euler. Yale: Rapid prototyping for complex data mining tasks. pages 935–940, New York, NY, USA, August 2006. ACM.

[9] N Milke, M Doert, S Klepser, D Mazin, V Blobel, and W Rhode. Solving inverse problems with the unfolding program truee: Examples in astroparticle physics. *submitted to NIM*, 2012.

[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*, 2011.

FACT - a fact

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

The construction of the novel Imaging Air Cherenkov Telescope FACT [1] was finished autumn 2011. Since first data taking, the custom designed FPGA based data acquisition system has shown its stability and reliability. First results show a homogeneity of the detector over the time of opration and a homogeneous response of all pixels, which is essential for the operation of a long-term monitoring telescope such as FACT (see Figure 1).

FACT is the first Cherenkov telescope making use of novel semiconductor based Geiger-mode avalanche photodiodes (G-APDs) as photon detectors. The G-APDs were able to show their comparability if not superiority to common photon detectors such as photomultiplier tubes (PMTs). However, the development of methods for a full calibration and an automated analysis of the data is still ongoing.

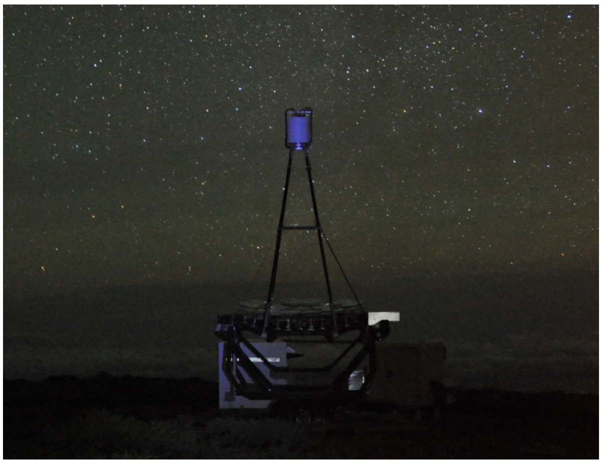


Figure 1: The FACT telescope.

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 relativistic charged particles emit Cherenkov radiation, while propagating through the atmosphere. This process lasts a few nanoseconds, thus an adequate photodetector is necessary to detect these blue flashes of light. [2]

The FACT camera has been installed on the Canary Island La Palma in October 2011. Already in the first night of datataking, which was a full moon night, Cherenkov showers were detected. Since that time the collaboration has taken a vast amount of data of different known sources of ultra high energetic Cherenkov sources, such as the Supernova Remnant Crab Nebula or the Active Galactic Nuclei Markarian 501 and Markarian 421.

The data acquisition system consists of 40 digitizer boards each comprising 36 channels of a total of 1440 channels. The signal of each channel is sampled at a frequency of 2 GHz and stored in an analog ring buffer until a trigger condition has been fulfilled. In case of a trigger, up to 1024 samples can be digitized with a resolution of 12 bit and are transmitted via ethernet to a data storage PC. The overall noise performance of the data acquisition system allows a single photon resolution, thus providing an excellent method for an absolute gain calibration of the system by looking at the single photon spectrum (see Figure 2). The knowledge of the total number of incident photons is needed for the energy reconstruction of the primary particle.

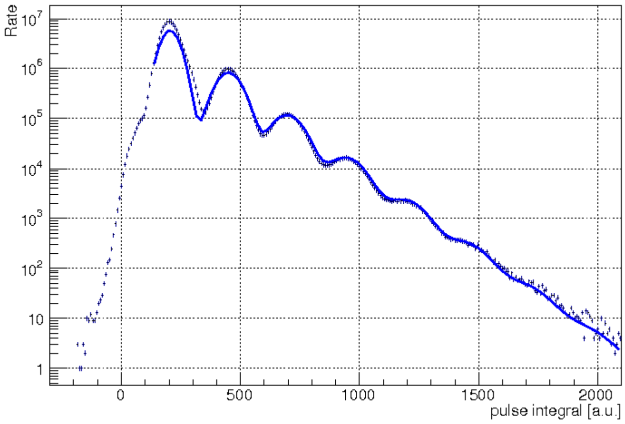


Figure 2: Single photon spectrum for all pixels as a function of the pulse integral.

In contrast to other Cherenkov telescopes, FACT's photon detectors are not harmed by bright light, such as the light of the full moon, thus the FACT telescope can in principle operate during full moon as well and prolongate its observation time. However, due to the increase of background noise, the analysis of data taken during full moon needs to be done with special care.

The FACT camera operates at a trigger rate of 60 to 70 Hz. This leads to 4 to 8 TB of raw data per month, depending on the weather conditions. This data is currently compressed and send to the ISDC data center in Geneva for further analysis.

The recorded raw data needs certain steps of calibration, before the actual analysis can be performed.

Prior to any data analysis, the current camera needs to be understood. This is necessary to correct certain inhomogeneities in the camera plane. In order to find such inhomogeneities two different approaches are undergone. The first one makes use of the detectors intrinsic noise signal generation. Measuring this dark noise amplitudes gives hints for inhomogeneities in the detector gain. The second approach involves a calibration light pulser, external of the camera. This pulser is capable of producing flashes of about 20 photons with a duration of less than 150 ns with a high degree of stability. By analyzing the detectors answer to these calibration pulses, one can determine the overall camera performance very well. However, certain features in the calibration pulser performance are still under investigation.

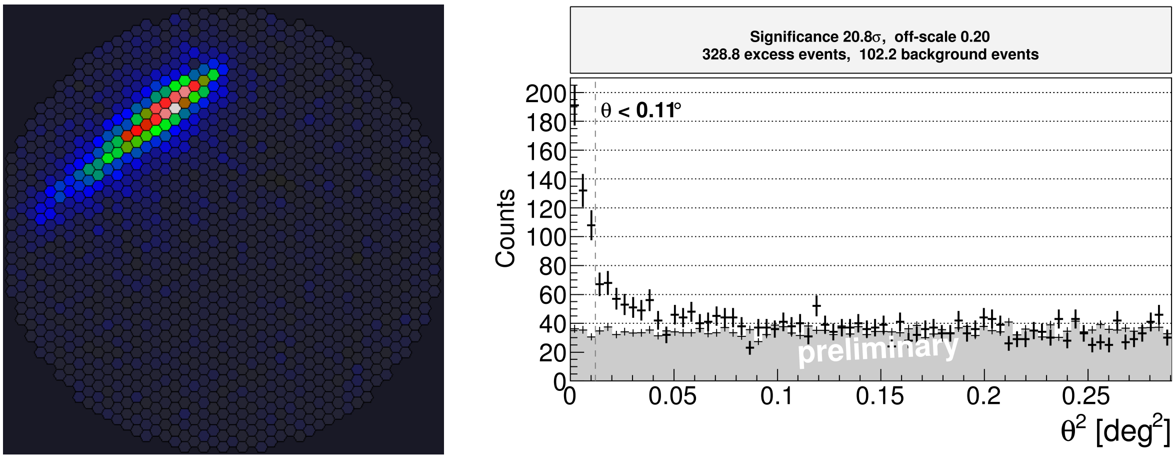
The first calibration step, the so-called DRS calibration, is related to the hardware imperfections and requires certain calibration data to be taken on regular intervals. After this calibration has been performed, the data can be understood as a movie sequence showing where in the camera and at what time Cherenkov photons caused a signal.

The next step includes the determination of the arrival time and the quantity of the Cherenkov signal in each camera pixel. Different methods of signal reconstruction and noise reduction are currently under investigation. However, the most straight forward methods already give good results and allow further analysis.

The aim of the Cherenkov data analysis is to reconstruct the type, direction and energy of the primary incident particle.

This analysis involves image cleaning, which is the suppression of pixels unrelated to the triggering Cherenkov shower, the image parameter calculation, helping to classify the type of the incident primary particle and finally the classification and energy estimation of the particle itself.

In Figure 3(a) a recorded Cherenkov shower can be seen. Figure 3(b) shows the detected signal of the Crab Nebula with a significance of 20.8 σ.



(a) A typical Cherenkov shower as it is recorded by the FACT camera.

(b) Detected signal of the Crab Nebula of data recorded with the FACT telescope.

Figure 3:

Outlook

Several steps of raw data calibration don't require only computing time but also valuable disk space. Some of these steps could be performed in future system upgrades already on the data acquisition electronics. The precalibrated data can be more efficiently analyzed, which allows an easier implementation of a next level trigger. It appears even possible to calculate important diagnostic data already on the data acquisition electronics inside the camera, and thus implement sophisticated system status monitoring.

References

[1] 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.

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

Improvement of Monte Carlos for the FACT telescope

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

In October 2011 the First G-APD Cherenkov Telescope (FACT) started observing very high energy gamma-rays from galactic and extragalactic sources [1]. It is the first Cherenkov telescope using semiconductors, the so-called Geiger-mode Avalanche PhotoDiodes (G-APD), as photon detection devices [3].

For the analysis of the recorded data, especially the reconstruction of its energy and its particle type, the production and processing of Monte Carlo data is necessary. One step in the processing is the simulation of the reflecting system of the telescope. After the installation and the adjustment of the real mirrors, their position and alignment were determined. This information was integrated into the processing of the Monte Carlos and tested.

1 Cosmic rays and their detection

There exist different locations in our universe where cosmic rays are produced, accelerated and emitted. Possible sources are e.g. supernovae, their remnants and pulsars as galactic sources and Active Galactic Nuclei (AGN) as extragalactic sources. The different particles that are emitted from these sources are hadrons, e.g. protons, electrons, photons and neutrinos. Due to the electric charge of the protons and electrons, they are deflected from their original travel path by intergalactic magnetic fields. Photons and neutrinos are electrically neutral and thus they are not influenced by magnetic fields. It is an advantage of these particles that they point back to their origin so that the source positions can be determined.

Since the Earth's atmosphere absorbs all electromagnetic radiation above 10 eV [5], the high energetic gamma-rays cannot be detected directly. The only possible way to detect these gamma-rays on the Earth's surface is indirect by observing the Cherenkov light of extensive air showers, an avalanche of secondary particles. Depending on the energy and the type of the incident particle the development of the shower differs. This feature is used to distinguish between gamma-rays and other particles.

2 FACT - The First G-APD Cherenkov Telescope

As an Imaging Air Cherenkov Telescope (IACT), FACT has a primary mirror, consisting of several segments, that reflects the Cherenkov light flashes of the showers onto the camera in the focal plane. In this way an image of the shower can be taken. The number of photons and their arrival time in a pixel are used to define the time evolution of the shower and to reconstruct the source position and the energy of the primary particle.

For FACT the refurbished mount of the former HEGRA CT3 and 30 hexagonal mirrors of the former HEGRA CT1, which have been refurbished, are used. The mirrors have been adjusted and with the aid of laser reflections their final alignment was determined. The area of the single mirrors adds up to a total reflective area of 9.51 m^2. The mean focal length is 4.89 m [2].

3 Monte Carlos for FACT analysis

The analysis of the FACT data starts with the calibration of the raw data where e.g. the conversion of electronic units into photoelectrons takes place. For each triggered event it comes to the decision if a pixel belongs to the shower or not. Essential criteria are the number of photoelectrons in each pixel and the temporally correlation with its neighboring pixels. After these cuts, image parameters are calculated for all events.

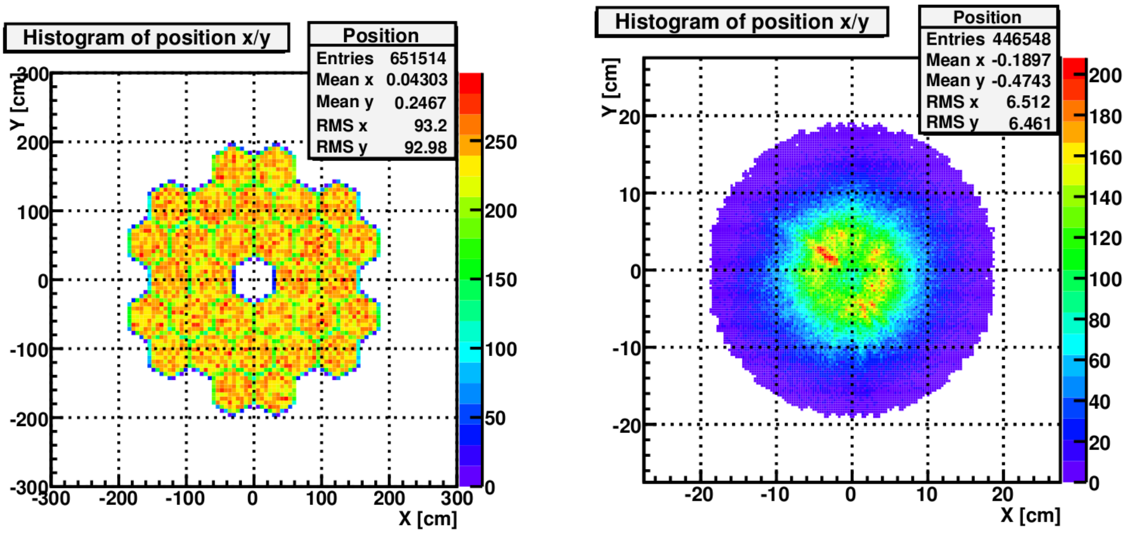
The separation of the gammas from hadrons is the next step. It is important to know the differences between the features of the air showers because the ratio of signal (gamma) to background (hadron) is 1:1000 [5]. Monte Carlos, simulated air showers induced by gamma-rays, and OFF-Data, representing the hadronic showers, are used to solve this task. For the Monte Carlos the development of the air showers is simulated and they are afterwards processed with the same analysis chain as the real data [4]. By using a random forest the features of gamma and hadronic induced showers are studied and the real data can be characterized by using the random forest. Also the energy of the primary gamma can be calculated by comparing the real data to the Monte Carlos.

In order to process the Monte Carlos with the same analysis steps they need to have the same information as the raw data. After the simulation of the air shower and the Cherenkov photons they have to be processed with a reflector and camera simulation program. In order to have the best possible Monte Carlos all the characteristics of the telescope system have to be known and simulated as realistic as possible.

For the simulation of the reflection it is important to know e.g. the reflectivity of the mirrors in dependency of the wavelength and the point spread function which describes the quality of the image. It is also necessary to know the geometry of the reflector. It can be described with the position and the alignment of the mirror segments.

In figure 1 the simulation of the reflecting system is shown. The used Monte Carlo file contained the information of the Cherenkov photons of 40000 air showers induced by gamma-rays. In figure 1(a) the distribution of the Cherenkov photons in the reflector plane is shown. Only the photons which were reflected by the mirrors remain. The arrangement of the mirrors can be clearly seen.

After the reflection the path of the photons to the camera is calculated. In figure 1(b) the photon distribution in the camera is shown. The distribution is centered around zero. The simulation demonstrates that the mirrors are well aligned and that this information could be included into the processing of the Monte Carlos for the FACT telescope.



(a) Photon distribution in reflector plane.

(b) Photon distribution in camera plane.

Figure 1: Distribution of simulated Cherenkov photons in the reflector and the detector plane of the FACT telescope. The color code indicates the number of photons at a certain position. In (a) it can be seen that the diameter of the reflector is between 3.6 and 3.9 m. (b) shows that the camera diameter is ≈ 40 cm.

4 Conclusion

For the analysis of very high energy gamma-ray data for FACT the use of Monte Carlo simulations is necessary. It is needed for particle determination and energy reconstruction. The improvement of the reflector simulation for Monte Carlos and the approach towards reality makes the analysis better. The more realistic the simulation of the reflector and the other components of the telescope the more reliable is the comparison of the real data with the Monte Carlos. Furthermore the separation of signal and background and the energy reconstruction are more precise.

References

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

[2] H. Anderhub et al. FACT—The first Cherenkov telescope using a G-APD camera for TeV gamma-ray astronomy. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 639(1):58 – 61, 2011.

[3] H. Anderhub et al. Electronics for the camera of the First G-APD Cherenkov Telescope (FACT) for ground based gamma-ray astronomy. *Journal of Instrumentation*, 7(01):C01073, 2012.

[4] T. Bretz and D. Dorner. MARS -The Cherenkov Observatory edition. *AIP Conference Proceedings*, 1085:664 – 667, 2009.

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

Systematic Studies For An IceCube Atmospheric Neutrino Analysis

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. Significant improvements of the IceCube atmospheric neutrino analyses were achieved by utilising machine learning techniques such as Mininmum Redundancy Maximium Relevance and Random Forests. Within this study previous results could be confirmed by means of detailed systematice studies.

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, 3, 7].

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 [6].

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.

The low signal to background ratio in combination with the large number of attributes available in an IceCube analysis makes this task well suited for a detailed study within the scope of machine learning.

In previous studies [9] the improvements considering the event selection part of an atmospheric neutrino analysis in IceCube have been pointed out. These improvements mainly based on using machine learning techniques such as an MRMR feature selection [4] and a Random Forest [2].

The background of an atmospheric neutrino analysis, however, can consist of several components. The main components are single, double and triple coincident muons. Double and triple conincincident events are detected in case two or more muons enter the detector from different directions within a certain time window. It is, however, important to distinguish coincident events from muon bundles, which is another possible background component originating mainly from the interaction of heavy nuclei in the Earth's atmosphere.

Monte Carlo simulations for each individual component were generated using CORSIKA [5] and processed by a Random Forest. A signalness was assigned to the events. The signalness distribution is depicted in figures 1, 2 and 3.

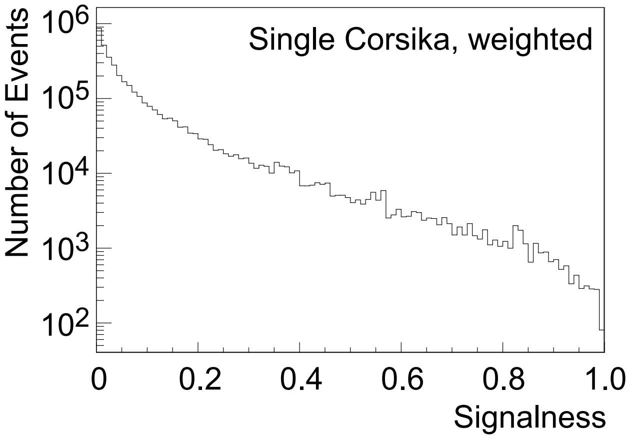


Figure 1: Single Corsika events weighted to the livetime of the burn sample. One finds that the number of background events decreases as the signalness increases

Figure 1 depicts the signalness of single background events weighted to the livetime of the burn sample. One finds that the number of events decreases as the signalness increases.

Figure 2 shows the signalness of double coincident background events weighted to the livetime of the burn sample. As in case of the single background events the number of events decreases as the signalness increases. Statistical fluctuations become prominent for $signalness \geq 0.7$.

Figure 3 depicts the signalness of triple coincidents background events weighted to the livetime of the burn sample. Statistical fluctuations start to dominate the histogram around $signalness \geq 0.7$. No triple coincindent background events are found at a signalness considered as the final cut level.

In table 1 the number of number of background events for single, double and triple coincident Corsika are shown. The sum of all three background components and the

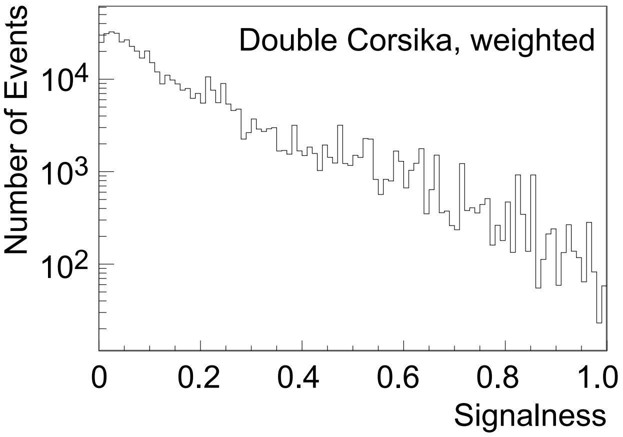


Figure 2: Double Corsika events weighted to the livetime of the burn sample.

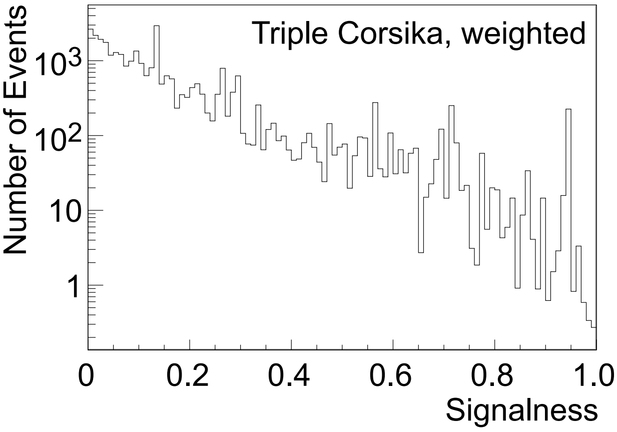


Figure 3: Triple Corsika events weighted to the livetime of the burn sample. A prominent peak is observed around 0.94.

Cut	Single Ev.	Double Ev.	Triple Ev.	Σ	Prediction
0.990	92.0	59.0	0.3	151.3	204 ± 52
0.992	75.8	58.1	0.3	134.2	160 ± 32
0.994	62.7	11.9	0	74.6	109 ± 31
0.996	44.5	11.8	0	56.3	64 ± 28
0.998	35.9	11.8	0	47.7	28 ± 12
1.000	9.5	0	0	9.5	14 ± 9

Table 1: Estimated number of background events for single, double and triple conincident Corsika. For comparison the sum of all three background components and the expected number of background events estimated from a 5-fold cross validation are shown as well.

total number of expected background events are shown for comparison. One finds that the sum of the three components agrees well with the background predictin estimated from a 5-fold cross validation. As few as 0.3 triple coincident background events are expected to enter the final event selection.

References

[1] J. Ahrens *et al.* Sensitivity of the IceCube detector to astrophysical sources of high energy muon neutrinos, Astropart. Phys. **20** (2004)

[2] L. Breiman, Random Forests, Machine Learning 45 (2001)

[3] T. DeYoung, Neutrino Astronomy with IceCube, Modern Physics Letters A, Vol. 24, Iss. 20 (2009)

[4] 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)

[5] D. Heck, CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers, Forschungszentrum Karlsruhe Report RZKA 6019 (1998)

[6] M. Honda *et al.*, Calculation of the flux of atmospheric neutrinos, Phys. Rev. D **52**,9 (1995)

[7] E. Resconi, Status and prospects of the IceCube neutrino telescope, Nucl. Instr. and Meth. A **602**, 7 (2009)

[8] T. Ruhe and K. Morik, Data Mining IceCube, Proceedings of the ADASS XXI conference 2011 (2011)

RapidMiner for IceCube

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

The IceCube neutrino telescope is located beneath the glacial ice at the South Pole. IceCube data shows a unfavourable signal to background ratio for atmospheric muon neutrinos. IceCube data is not homogeneous, therefore the data contains events with different pattern. These different patterns can be used for separation between atmospheric muon neutrinos and background for example with random forest algorithms. A further increase in separation power can be reached by forming clusters in the data according to prior knowledge and trained random forest algorithms on each cluster. An increase of recall an purity could reached with this ensemble of random forest.

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

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

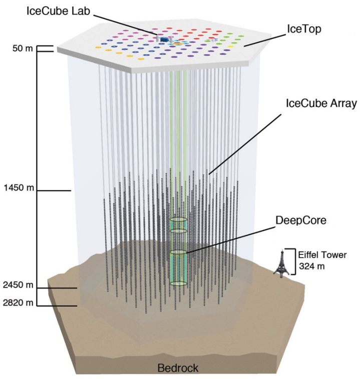


Figure 1: The schematic structure of the IceCube detector. The picture shows the 86 Strings in light grey and the 5160 DOMs in dark grey. [10]

theenergy spectrum of the atmospheric neutrinos. [4], [5], [7]

We know that each events in the IceCube data is different, but nonetheless we can find groups of data with very similar patterns. We can than use theses groups in the goal to increase the separation power between signal and background whilst training a random forest on data from a cluster with very similar event patterns. The idea is that the resulting random forest will be more sensitive to these patterns compared to one random forest trained on the whole dataset. From physical background knowledge one can think of a big amount of such clusters. For this analysis where eight clusters chosen. The first four clusters where chosen by the calculated vertex position. The vertex position is the position on which the neutrino has decayed. A schematic drawing of these four clusters can be found in figure 2. It is easy to imagine that an event with its vertex in the middle of the detector has a very different event pattern to one with an vertex on the edge of the detector.

- I: MPEFit_Fit_Pos_Z $\in [350\,\mathrm{m}, \mathit{max}]$
- II: MPEFit_Fit_Pos_Z $\in]min, 350\ \mathrm{m}] \wedge$ radius $\in [0\,\mathrm{m}, 300\,\mathrm{m}[$
- III: MPEFit_Fit_Pos_Z $\in]min, 350\,\mathrm{m}] \wedge$ radius $\in [300\,\mathrm{m}, 450\,\mathrm{m}[$
- IV: MPEFit_Fit_Pos_Z $\in]min, 350\,\mathrm{m}] \wedge$ radius $\in [450\,\mathrm{m}, \mathit{max}]$

The next four clusters were chosen by the number of string with data taking DOMs in the events. This observable shows a strong correlation to the neutrino energy. Again it is easy to understand that an event triggered by a low energetic neutrino will show very different patterns than one with very high energy.

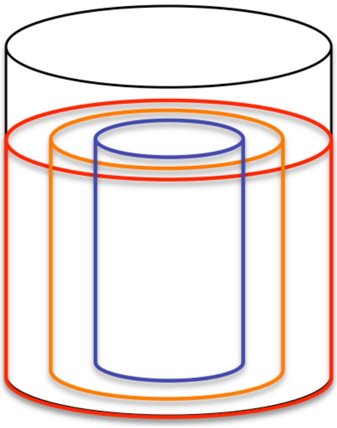


Figure 2: The schematic structure of the clusters I - IV.

- V: NString = 1
- VI: NString = 2
- VII: NString = 3
- VIII: NString $\geq$ 4

For each cluster in the data a set of observables has been derived using the MRMR feature selection algorithm [3]. With this set of observables one can optimize a random forest for each cluster and combine their results. In one can find the combinations which yielded the highest purity and recall of all possible combinations. More over in the table is a comparison to an analysis which uses the same detector configuration and also random forest but not an ensemble of random forests. This shows that this new method can increase purity and recall in comparison to using only one random forest for the whole data.

data set	$N_{neutrinos}$	purity
highest purety	3909	$(99{,}9\,^{+0{,}1}_{-0{,}2})\,\%$
highest recall	4815	$(98{,}7 \pm 0{,}4)\,\%$
analysis by [9]	2833	$(99{,}8\,^{+0{,}2}_{-0{,}4})\,\%$
analysis by [9]	3638	$(99{,}4 \pm 0{,}6)\,\%$

Table 1: This table shows a comparison between this analysis and an other analysis. Both analysis were maid on the same detector configuration. The main difference between the two is that this analysis utilises multiple random forest. $N_{neutrinos}$ is the number of neutrino events.

References

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

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

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

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

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

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

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

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

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

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

Gamma-Hadron Separation of Monte Carlo Simulations with RapidMiner for FACT

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

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

The so-called Imaging Air Cherenkov Telescopes (IACTs) are able to detect very high energy gamma-rays of galactic or extragalactic objects like supernovae or Active Galactic Nuclei (AGN). Due to the neutral electric charge gamma-rays are not influenced and deflected by intergalactic magnetic fields. Thus the direction they are coming from points directly to the astrophysical source. When very high-energetic gamma or hadronic particles are hitting the upper atmosphere layers of Earth, they induce an extensive air shower which consists of secondary relativistic charged particles. This shower emits a blueish light, the so-called Cherenkov light [6].

FACT is the first IACT which uses Geiger-mode Avalanche PhotoDiodes (G-APDs) instead of photomultiplier as photosensors to detect this light. It is located on the canary island La Palma at 2200 m a.s.l. and was commissioned for the first time during 11th October 2011 [4]. Due to a signal to background ratio of 1:1000 the separation of gamma showers from hadronic showers is very important to increase the sensitivity of the telescope and thus the effective observation time.

The building and testing of the separation model is done with a Random Forest (RF) algorithm [5], which is available in the data mining environment RapidMiner [2]. The model is trained on all available Monte Carlo simulations for FACT, which were produced by CORSIKA [1]. For this purpose gamma and proton showers were used and further processed by the analysis software MARS CheObs ed [3]. After data processing about 16000 events remained, respectively, and were used for the training. The Random Forest was trained by parameters which describe the shower images and thus allow to distinguish between gamma showers and hadronic showers. For the RF 500 trees were built and six randomly chosen features taken out of a total amount of 50 parameters. Furthermore a signalness cut from S=0.5 to S=1.0 as well as a five fold cross validation were applied to the RF model to determine statistical mean and error values and to estimate the stability of the model. In Fig.1 the first results of the testing on the simulation data are shown. The green data points show the efficiency against the signalness cut. Here the efficiency

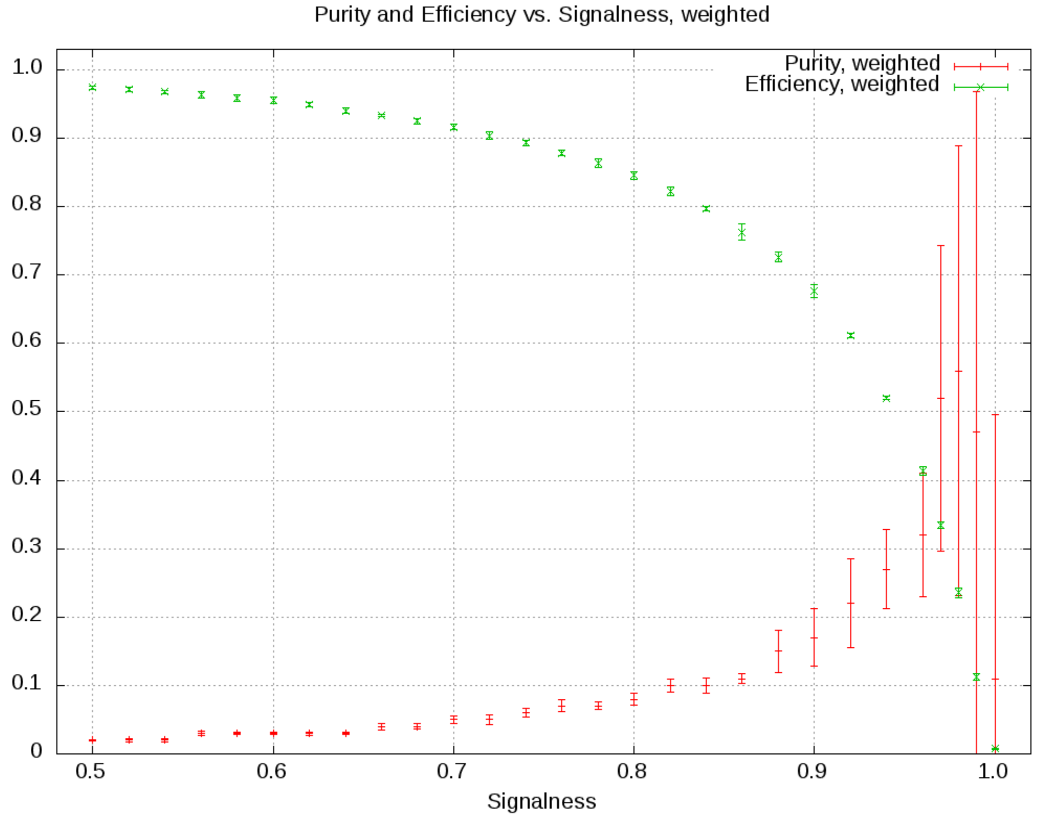


Figure 1: Displayed are the results of a trained Random Forest model on Monte Carlo simulation. Red data points indicate the weighted purity increasing with a higher signalness cut, while green data points show the decreasing efficiency. The high error bars are resulting from a small amount of simulation data.

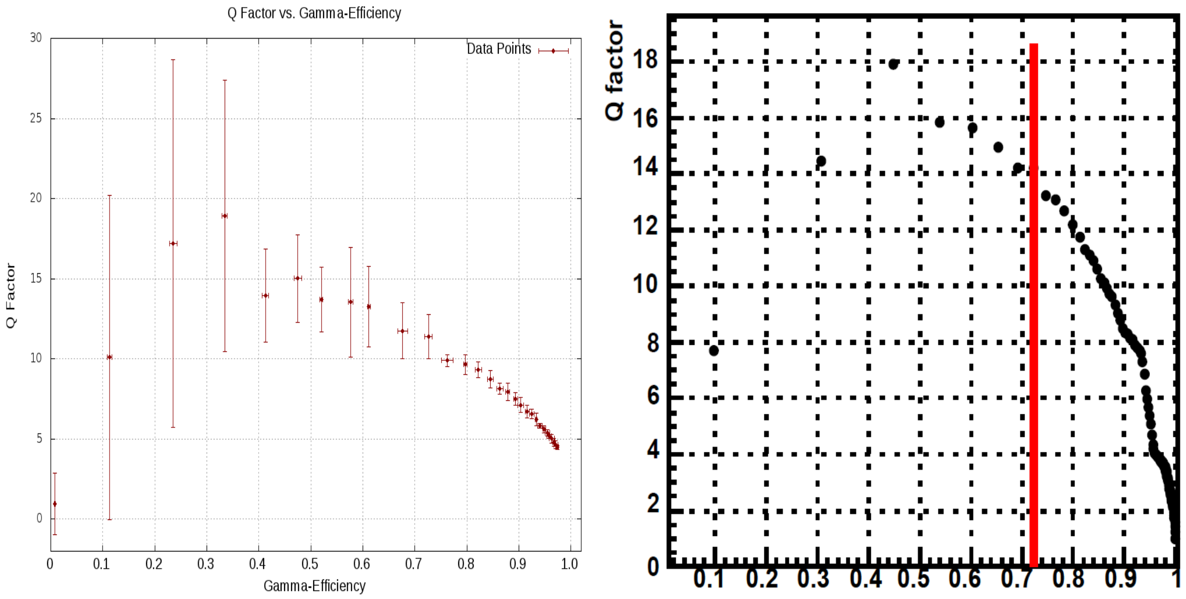
can be described as

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

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

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

whereas N_{fp} is the amount of false positive classified events after the signalness cut. For a detailed description of the classification see [8]. While the purity is increasing with an increasing signalness cut, the efficiency is decreasing in the same time. The large errorbars of the last two data points result from the small amount of simulation data. The challenge is to find a signalness cut at which not too much data is cutted away while the purity of the dataset is still high. One can find an efficiency of E=23 % - 41 % and a purity of P=9 % - 33 % between a signalness cut of S=0.98 and S=0.99 . The big range between the values shows that the signalness cut has to be trimmed. To compare the built model with an RF separation model for the MAGIC Telescopes [7], a quality factor Q has to be determined.



(a) Q factor vs. gamma efficiency of the RF model for FACT simulation data.

(b) Q factor vs. gamma efficiency of the RF model for MAGIC data [7].

Figure 2: Q factors against gamma efficiencies for RF models of FACT simulation data and MAGIC data

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

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

In Fig.2 the Q factor is shown for each signalness cut against the gamma efficiency for the FACT simulation data (left) and for MAGIC data (right). The curve shape is almost identical, which shows that both RF models have the same performance and both models produce nearly the same values. For the MAGIC data a cut at Q=14 is applied, which results in an efficiency of E=70 % and a purity of P=22 %, while a cut at Q=13 can be applied for FACT data, which results in an efficiency of E=61 % and P=22 %. These values show that the here presented unoptimized RF model for FACT data can nearly achieve a separation as good as the model used in the MAGIC experiment.

Further improvements like a study of the number of used attributes and trees used by the Random Forest is ongoing. The hadron simulations will be replaced by real hadronic data and new attributes for separation are developed and will soon be included.

References

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

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

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

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

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

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

[7] Marcos López Moya. *Astronomía Gamma con el Telescopio MAGIC: Observaciones de la Nebulosa y Pulsar del Cangrejo*. PhD thesis, Universidad Complutense de Madrid, July 2006.

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

Robust and weighted regression to calculate periodograms for disturbed irregularly sampled time series

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

An important task in astroparticle physics is the detection of periodicities in irregularly sampled time series, called light curves. Periodogram methods for light curves often may be characterized as fitting periodic models using least squares regression. We illustrate advantages and disadvantages of using robust and/or weighted regression instead. For more background, mathematical details and references see [3].

Typical time series occuring in astroparticle physics are light curves, which exhibit special properties like a periodic sampling scheme, heteroscedastic uncorrelated observational noise and so-called red noise. Moreover, the experiments suffer from a varying measurement precision, which can be estimated for each observation time. Typically, an estimate for the standard deviation of the white noise is given for each time point. We will call those estimates measurement accuracies. A detailed mathematical description of the data assumed is given in [3] and [2].

Fourier Analysis, as a standard method to find periodicities in an observed signal, cannot cope with irregular sampling. Many methods for detecting periodic fluctuations in an irregularly and periodically observed signal have been developed and work as follows:

- Choose a model for the periodic fluctuation (e.g. a sine, a periodic step function, a Fourier sum or periodic splines).
- For different trial periods, fit the periodic fluctuation. If it is intended to take the measurement accuracies into account, use weighted regression for this step.

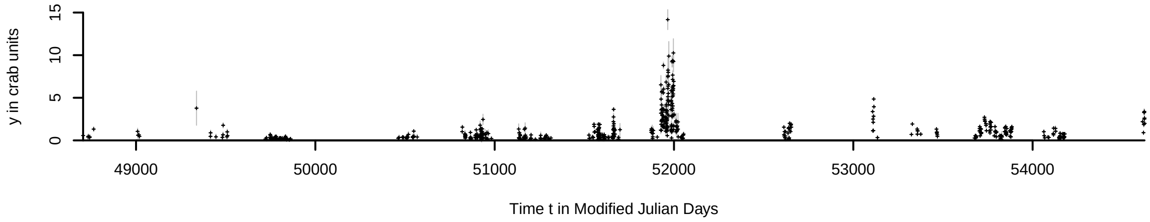


Figure 1: Observations for the very high energy gamma-ray source Mrk 421. Grey vertical lines at each point show the measurement accuracies.

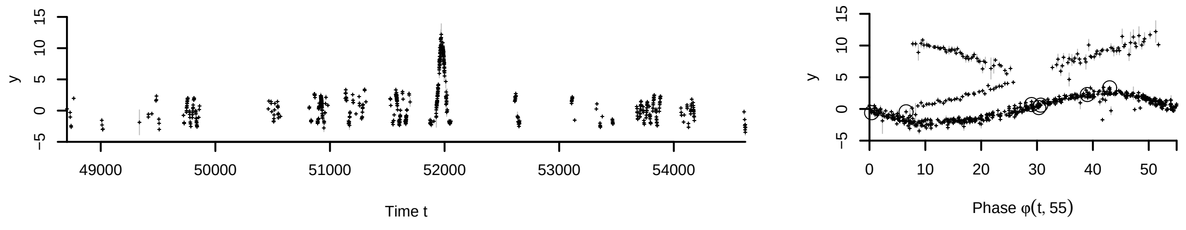
- Calculate a criterion value for the goodness of fit and call it periodogram bar. The set of periodogram bars is called periodogram. When using weighted regression, we will refer to the periodogram as weighted periodogram.

As criterion value for the periodogram bar we use the coefficient of determination, which should be large if the trial period is the true period of the fluctuation. Most values found in the literature are equivalent to this. Another typical choice is the squared amplitude when the model for the periodic fluctuation is a sine.

The most popular periodograms as for example the Lomb-Scargle Periodogram [1] are based on L_2 regression. A few attempts have been made to apply robust regression techniques in this context. We illustrate that it may be quite sensible to use them with an example based on the sampling and measurement accuracies of observations for the astrophysical very high energy gamma-ray source Mrk 421 (see [4] and references therein), see Figure 1. The signal of the artificial light curve shown in Figure 2 is generated combining normal white noise with a sine-shaped periodic fluctuation of period 55 and an interval of atypically behaving observations of a peak shape, as it was similarly observed for Mrk 421. Figure 3, Panels (a) to (d) show the periodograms obtained fitting the true model by different unweighted regression techniques: Least squares (L_2) and the robust approaches least absolute deviation (L_1), M-regression using Tukey's biweight function (Tukey) and M-regression using Huber's function (Huber). Other choices are possible. All four periodograms show a peak at the true period 55, but it is much more outstanding in the robust periodograms.

This example motivates the use of robust regression in periodograms and a more comprehensive comparison. It also poses the question, how outstanding a peak in a periodogram needs to be to be regarded as valid, i.e. supposed to belong to a true period. The typically used thresholds apply only for least squares regression periodograms. They are represented by horizontal grey lines in Figure 3. It is obvious that this criterion is very liberal and regards lots of periods as valid. The horizontal black line comes from an adaptive approach based on outlier detection and seems to be more suitable.

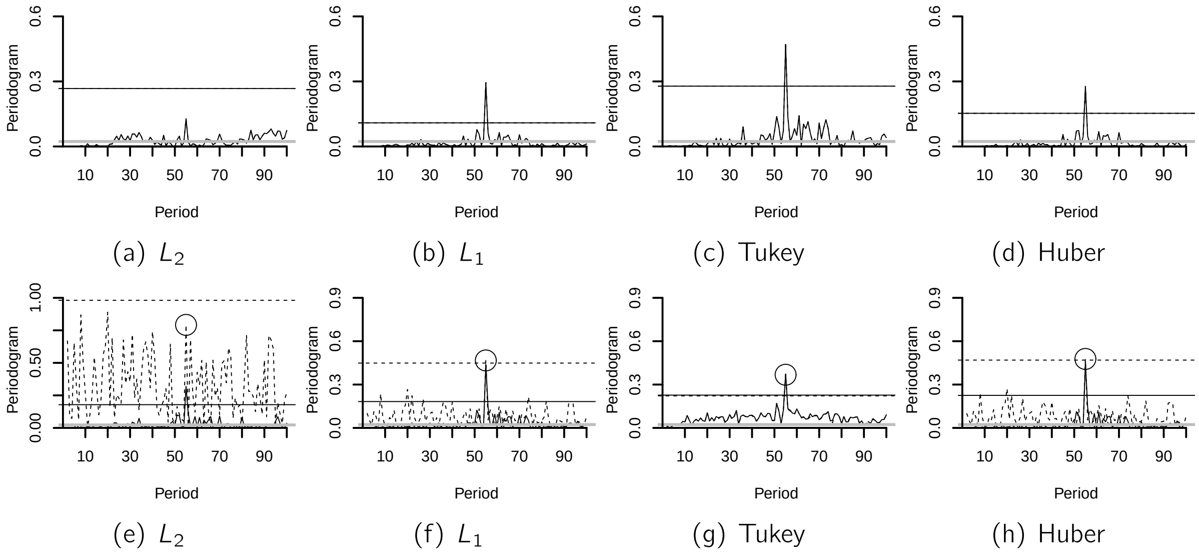
Robust regression techniques can be helpful in the case of disturbed light curves, for example if the observations in some interval show a high peak, as it was observed for real



(a)

(b)

Figure 2: (a) Artificially generated time series based on the observation times and measurement accuracies observed for Mrk 421. Five missing measurement accuracies s_i are substituted by the median of the remainder. White homoscedastic normal noise is added, but no red noise is added. (b) The same time series folded to its fluctuation period $p_f = 55$. The 7 circles mark the observations that are modified for the second example.



(a) L_2

(b) L_1

(c) Tukey

(d) Huber

(e) L_2

(f) L_1

(g) Tukey

(h) Huber

Figure 3: Periodgrams of the light curve shown in Figure 2 obtained by fitting a sine with an intercept using different regression techniques. The grey line corresponds to a conventional predefined $1 - \alpha$ =0.95-threshold, the black line (solid and dashed) correspond to the respective adaptive threshold. (a) to (d): Unweighted periodograms. (e) to (h): Weighted periodograms using the true weights (solid line) and altered weights (dashed line). The circle marks the height for the dashed periodograms at period 55.

data. However, in [3] the methods are applied to further data and it is found that when the peak is part of the periodic fluctuation, robust methods have problems.

We considered both weighted and unweighted regression and it turned out that least squares methods are very sensible to very small measurement accuracies. The solid curves in Figure 3, Panels (e) to (h) show the weighted periodograms of the light curve displayed in Figure 2. All methods except the one using Tukey regression profit from the weights, as the peak is more outstanding than in the unweighted periodograms. However, now we set 7 out of 655 measurement accuracies to 0.001 (minimum before: 0.05). The chosen observations are marked with a circle in Figure 2. They are not large before resetting and are not part of the peak shaped disturbance. The dashed periodograms in Figure 3, Panels (e) to (h), are calculated using these modified weights. While Tukey regression behaves robustly against the disturbed measurement accuracies, the periodograms using L_1 and Huber regression are much more affected and the least squares periodogram does not provide usable information any more. We observe the same effect for the other models different from a sine, too. For this reason we recommend unweighted regression when using Tukey regression, and also for the other regression methods in case of uncertainty about the quality of the measurement accuracies.

References

[1] 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.

[2] A.M. Thieler. Periodicity detection in irregularly sampled light curves by robust regression and outlier detection. In K. Morik and W. Rhode, editors, *Technical report for Collaborative Research Center SFB 876 – Providing Information by Resource- Constrained Data Analysis*, pages 154–157. TU Dortmund University, SFB 876, 2011.

[3] 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*, 2012. under revision.

[4] M. Tluczykont, E. Bernardini, K. Satalecka, R. Clavero, M. Shayduk, and O. Kalekin. Long-term lightcurves from combined unified very high energy gamma-ray data. *Astronomy and Astrophysics*, 524:A48, 2010. Data available: http://nuastro-zeuthen.desy.de/magic_experiment/projects/light_curve_archive.

The variable VHE γ-ray sky explored by MAGIC - BLAZAR variability study

Malwina Uellenbeck
Lehrstuhl für Astroteilchenphysik E5b
Technische Universität Dortmund
malwina.uellenbeck@tu-dortmund.de

This report is a short summary of a variability 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 variable behavior of such extragalactic sources. Also clues on a potential periodic behavior of the sources might be drawn from a study of the obtained lightcurves.

1 Strong variability from VHE γ-rays Blazars

High-energy γ-ray above an energy of few 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 [8]. 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 second one in the GeV range). Furthermore, these objects are dominated by a highly variable component of non-thermal radiation produced in such relativistic jets. Such flux variability can occur on different time scales: from fast flares lasting few minutes to high states of several months [6]. Thus, such AGN γ-ray analysis can be very time consuming. Furthermore, this variability in the γ-range may be also periodic, which in turn could be the signature of a binary black hole. Systems of black holes are strong emitters of gravitational waves whose amplitude depends on the binary orbital parameters such as the component mass, the orbital semi-major-axis and eccentricity [7].

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. [4]. Being much more sensitive than satellite experiments, like e.g. Fermi-LAT, 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.

3 The MAGIC telescopes

MAGIC is a system of two 17 m dish Imaging Atmospheric Cherenkov Telescopes (IACTs) located at the Roque de los Muchachos observatory (28.8°N, 17.8°W, 2200 m a.s.l.), in the Canary Island of La Palma, see Fig.1. Since 2009 the MAGIC telescopes are carrying out stereoscopic observations with a sensitivity of $< 0.8\%$ of the Crab Nebula flux, for energies above ~ 300 GeV in 50 hr of observations [2]. The trigger energy threshold of the system is the lowest among the current operating IACTs, giving the possibility to observe γ-rays between 50 GeV and several TeV.

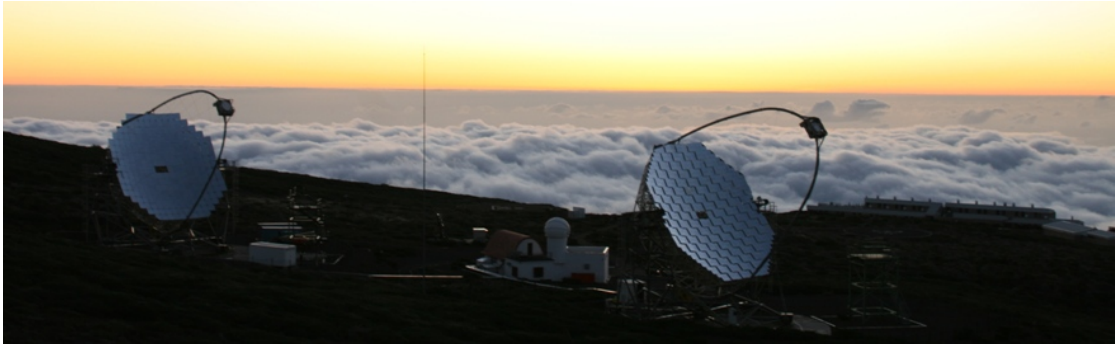


Figure 1: MAGIC telescopes located at the Canary Island of La Palma

4 The MAGIC monitoring program

IACTs like MAGIC can give a valuable input for understanding of the acceleration mechanism in blazars. Not only by participation in multiwavelength observations, but also by performing a source state independent, long term monitoring of the most interesting brighter γ-ray emitters. There are many advantages of such observations. They allow to obtain an unbiased distribution of flux states and perform any statistical study which requires high statistics on various flux levels [5].

The MAGIC data have been processed with the standard MAGIC analysis tools [1]. A fraction of the data has been removed due to poor observation conditions. From the MAGIC analysis results depicted in Fig. 2 one can clearly see the strong variability feature of the most prominent AGNs: Mrk 421 & Mrk 501.

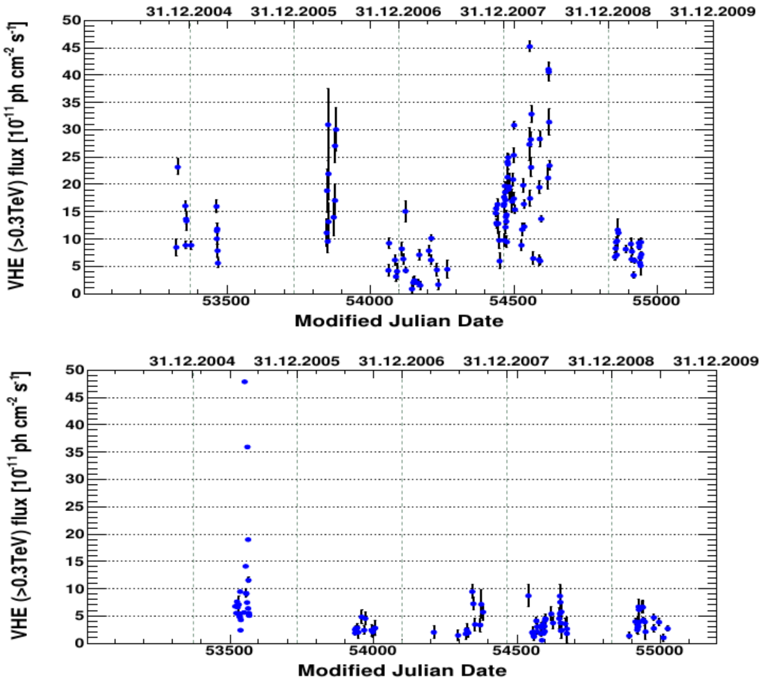


Figure 2: MAGIC lightcurve of Mkn 421 (top) and Mrk 501 (bottom) observed above 0.3 TeV from 2005 until 2009 [9]. A clear variable behavior of this most prominent AGNs during five observation years is visible.

5 Conclusion and Outlook

In this report it has been shown that the measured flux levels during 5 years of MAGIC observations for both sources (Mrk 421 and Mrk 501) were found in very variable state (see Fig.2). This results provide a valuable material for further statistical studies e.g "Periodicity detection in irregularly sampled light curves by robust regression and outlier detection" investigated by Anita Monika Thieler [3] (Project of part C3) or others.

References

[1] J. Albert et al. (MAGIC Coll.) . *Astrophys. J, L23*, 2008.

[2] Aleksićat all. Performance of the MAGIC stereo system obtained with Crab Nebula data. *Astroparticle Physics*, 35:435–448, February 2012.

[3] R. Fried A.M. Thieler, M. Backes and W. Rhode. Periodicity detection in irregularly sampled light curves by robust regression and outlier detection.

[4] M. Catanese and T. C. Weekes. Very High Energy Gamma-Ray Astronomy. , 111:1193–1222, October 1999.

[5] 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.

[6] E. Resconi, D. Franco, A. Gross, L. Costamante, and E. Flaccomio. The classification of flaring states of blazars. , 502:499–504, August 2009.

[7] C. Rödig, T. Burkart, O. Elbracht, and F. Spanier. Multiwavelength periodicity study of Markarian 501. , 501:925–932, July 2009.

[8] R. 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.

[9] R. Wagner. Monitoring of bright, nearby Active Galactic Nuclei with the MAGIC telescopes. In *International Cosmic Ray Conference*, volume 8 of *International Cosmic Ray Conference*, page 143, 2011.

Modeling Images of Gamma Air Showers for Classification in the MAGIC Experiment

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

The MAGIC telescopes on the canary island of La Palma are imaging atmospheric Cherenkov telescopes. Their purpose is to detect Cherenkov light [3] 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). 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 [8]. So the gammas have to be separated from the hadrons via classification.

The standard by now for this classification is to use Hillas Parameters [6] as variables in a random forest [1]. Hillas Parameters have some drawbacks though, with the most important one being that they were not created as best possible separators of gamma and hadron events, but only to describe the shape of air showers. Our aim is to find variables other than the Hillas Parameters, by using prior knowledge of what the theoretical differences between gammas and hadrons are.

To find new variables, which separate better than the Hillas Parameters, we use the same concept as in statistical testing. Consider a pair of hypotheses

H_0 : Observation i is a gamma event

vs.

H_1 : Observation i is a hadron event

for each $i \in 1, ..., n$, where n is the number of observations in a data set. As we know that air showers induced by gamma events have a distinctive shape which distinguishes them from hadrons and the shape of an air shower directly translates into the distribution of arrival coordinates of photons in the camera, it is plausible to say that the vector of such coordinates (X, Y) is a random vector with a two dimensional distribution $P_{\theta_0} \in (P_\theta)_{\theta \in \Theta}$. This only holds, if the the air shower inducing particle is a gamma event, so the pair of hypotheses above can be rewritten as

H_0 : The arrival coordinates of the m observed photons of observation i, $(x_1, y_1)_i, ..., (x_m, y_m)_i$, are realizations of a two dimensional distribution belonging to the family $(P_\theta)_{\theta \in \Theta}$.

vs.

H_1 : The arrival coordinates of the m observed photons of observation i, $(x_1, y_1)_i, ..., (x_m, y_m)_i$, are **not** realizations of a two dimensional distribution belonging to the family $(P_\theta)_{\theta \in \Theta}$.

There are several test procedures for such a pair of hypotheses, for example a two dimensional Kolmogorow Smirnow Test [5] or a Chi-Squared Test. The p-value of such tests should make good tools for the separation of gammas and hadrons, if the distribution of gamma-showers in the camera are represented well by the family $(P_\theta)_{\theta \in \Theta}$.

However, there is a major obstacle in this procedure. As we have only information about the number of photons in each pixel, but not about the exact arrival coordinates of the photons, the above mentioned tests become very liberal, often rejecting the null hypothesis, even if it is correct. In other words, in most cases the p-value is very small, regardless of what the true event class is.

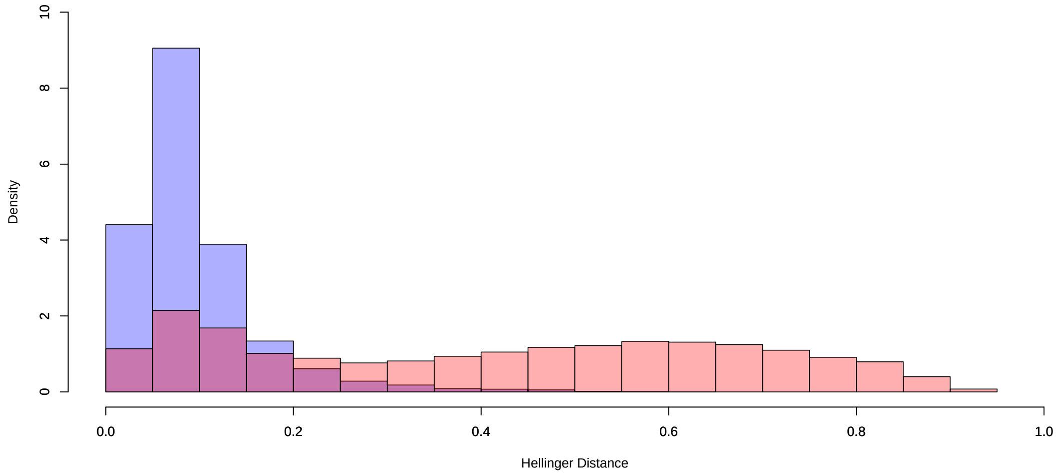


Figure 1: Histograms of the Hellinger Distance of observed gamma (blue) and hadron (red) air showers to a fitted bivariate normal distribution.

To overcome this, instead of trying to determine if the empirical distribution and the theoretical distribution are the same, we only try to measure the distance between the two. Even if tests reject the null hypothesis, it is valid to assume, that for gammas the empirical distribution is closer to the theoretical one, than for hadrons.

To measure this distance, we use the Hellinger Distance [2], as we need a standardized measure, like the Hellinger distance, as opposed to for example the test statistics of the tests above or the Kullback Leibler Divergence. The Hellinger Distance between two distibutions P and Q with density functions f and g is given by

$$H(P, Q) = 1 - \int \sqrt{f(x)g(x)}dx.$$

Although there are attempts to exactly describe the Cherenkov light distribution of gamma air showers $((P_\theta)_{\theta \in \Theta})$ [4], there is no closed form of it. Therefore, a distribution family has to be found, which has a closed form and fits the true distribution well. Considering the nearly elliptical shape of gamma showers, the first idea is to describe the showers through bivariate normal distributions, although this cannot be an optimal solution, as gamma showers are known to be skewed.

Figure 1 shows histograms of the hellinger distance between the empirical distribution and a fitted bivariate normal distribution for 7821 gamma events and 7412 hadron events. As can be seen, there is a large difference between the two histograms. Gamma showers seem to have a much smaller distance to a normal distribution than hadron showers. The difference seems to be larger than for most of the Hillas Parameters (reviewed in [7]).

Although a normal distibution does not seem to be the best distribution to describe gamma showers, this result is promising for future work. Next steps in this work include finding other distributions than the normal distribution, which might better fit the skewness of gammas as well as a study on how well the hellinger distance is suited to be used in a classification method to seperate gamma and hadron events.

References

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

[2] L.M.L. Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer series in statistics. Springer-Verlag, 1990.

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

[4] Mathieu de Naurois and Loïc Rolland. A high performance likelihood reconstruction of gamma-rays for imaging atmospheric cherenkov telescopes. 32, 2009.

[5] G. Fasano and A. Franceschini. A multidimensional version of the Kolmogorov-Smirnov test. , 225:155–170, March 1987.

[6] A.M. Hillas. Cherenkov light images of eas produced by primary gamma. In *Proceedings of the 19th International Cosmic Ray Conference ICRC*, volume 3, page 445, 1985.

[7] T. Voigt. Exploration und Vorverarbeitung von MAGIC-Daten zur Gamma-Hadron-Separation. Diplomarbeit, Technische Universitaet Dortmund, Germany, June 2010.

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

Demixing of empirical distribution functions

Max Wornowizki
Statistik in den Biowissenschaften
Technische Universität Dortmund
wornowiz@tu-dortmund.de

We propose an algorithm checking whether given simulated data fits observed real data good enough. If this is not fulfilled a distribution function is computed describing which values are over- respectively underrepresented in the current simulation in comparison to the real data. The fast and intuitive algorithm can be based on the Kolmogorov-Smirnov-test, a widely-used nonparametric test procedure, its weighted versions or other procedures providing confidence bands for empirical distribution functions. The method is applied to preprocessed IceCube data in order to identify insufficient simulated variables and illustrate the problematic regions.

IceCube is a cubic kilometer large neutrino detector located at the geographic South Pole [1]. It consists of 86 strings with 60 digital optical modules each embedded in glacial ice between 1450 and 2450 m below the surface. Unfortunately IceCube does not detect only atmospheric neutrinos. The measurements consist mostly of atmospheric muons, which are not of interest in the given context. Classification algorithms like Random Forests are obvious tools for the separation of the relevant from irrelevant observations. However, no correctly labeled real data is available to train the methods. Simulations have to be used instead at this step of the analysis. Since conclusions drawn from wrong simulations are useless or even worse, good agreement of simulated and observed data must be guaranteed [4].

The typical approach of testing is not completely satisfying in this situation. In case of rejection of the null hypothesis of identical distributions the way to improve the simulation remains unclear. Now since statistical tests correspond to respective confidence bands one is interested in regions where the bands are violated. Some tests based on distribution functions are considered first due to the following reasons: they provide well computable confidence bands and are distribution free, that is, they do not assume Gaussian or other distributions.

Consider n_1 observations $x_1, \ldots, x_{n_1} \in \mathbb{R}$ of a continuous variable X and denote the probability law of X by F. Simulated data points $y_1, \ldots, y_{n_2}$ corresponding to a variable Y with probability law G are also available. A standard approach to test $H_0 : F = G$ without making further restrictions to F or G is the two-sample Kolmogorov-Smirnov test [3]. Denote the empirical distribution functions of the first respectively second sample by F_e respectively G_e and define $N := \frac{n_1 \cdot n_2}{n_1 + n_2}$. H_0 is rejected if G_e lies outside the confidence band defined by the upper boundary function $u = min(1, F_e + \frac{K_\alpha}{\sqrt{N}})$ and the lower boundary function $l = max(0, F_e - \frac{K_\alpha}{\sqrt{N}})$ with an appropriate critical value K_α.

If simulated data in regions outside the expected are excluded before the analysis there is a $s \in (0, 1]$ and a probability law H such that $F = s \cdot G + (1 - s) \cdot H$. Thus F can be considered as a mixture of G and H. The idea of our approach is to find a second simulation with probability law H such that a proper mixture of values simulated from both resembles the observed data. If s equals one the current simulation is correct and H is not of interest. However, H is unique for any fixed $s < 1$. Therefor demixing F, that is estimating s and H, provides all information of how to improve the simulations. Picking up the idea of the Kolmogorov-Smirnov test one solution is to identify a monotone step function $\tilde{H}$ and a factor $\tilde{s} \in (0, 1]$ such that the function $\tilde{F} := \tilde{s} \cdot G_e + (1 - \tilde{s}) \cdot \tilde{H}$ lies within the confidence band. $\tilde{H}$ then reflects properties of a second simulation which is necessary to justify that the data coming from both simulations, combined in proportions $\tilde{s}$ respectively $1 - \tilde{s}$, fit the observed data. Obviously neither $\tilde{H}$ nor $\tilde{s}$ are unique without further constraints. The following two demands are reasonable and guarantee an unique solution: the simulation shall be changed minimally that is $\tilde{s}$ chosen maximally. On the other hand given $\tilde{s}$ the resulting $\tilde{F}$ should resemble F the best.

An intuitive greedy algorithm for finding $\tilde{s}$ and $\tilde{H}$ under the above constraints was elaborated. We claim that it converges to the optimum of the given problem. The method constructs a decreasing sequence of minimal shrinkage values $s_0, s_1, \ldots, s_k$ and an increasing sequence of step functions $H_0, H_1, \ldots, H_k$ and stops if the candidate $F_k := s_k \cdot G_e + (1 - s_k) \cdot H_k$ lies within the confidence band. In principle three basic operations are applied: on the one hand candidates lying above the upper bound u somewhere have to be shrunk to lie below or on u. On the other hand falling below the lower bound l must also be corrected by adding probability mass to appropriate regions. At last candidates lying within the confidence band may not be proper distribution functions anymore and have to be adjusted by adding further probability mass. The operations are applied whenever necessary in the presented order. However, due to the converse nature of the first and the last two steps some data situations require multiple executions of some of them. The method is quite fast since it operates solely on the set $\{x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}\}$ and typically converges within 10 operations. It is directly applicable for other tests based on empirical distribution functions. One just has to determine the confidence band by inverting the ordinary test procedure. For example in the case of the weighted Kolmogorov-Smirnov test with weight function

$w(F_e, G_e) = \left[\frac{n_1F_e+n_2G_e}{n_1+n_2} \cdot (1 - \frac{n_1F_e+n_2G_e}{n_1+n_2})\right]^{-\frac{1}{2}}$ [2] solutions of quadratic equations provide the upper and lower bounds of the confidence band. In comparison to the ordinary version weighted Kolmogorov-Smirnov tests result in confidence bands with nonconstant distances between F_e and the bounds. This is desirable for testing since those bounds tend to lie closer to F_e except in the center of the distribution. On the other hand a tighter confidence band gives less space to lie in and thus requires more correction. So $\tilde{s}$ will be smaller and $\tilde{H}$ will be higher in most cases compared to the ordinary Kolmogorov-Smirnov version.

Both the ordinary as well as the weighted version of the algorithm mentioned before were applied to IceCube 59 data (observations based on 59 of the final 86 strings and corresponding simulations). The preprocessing of the data reduced the initial 996 attributes to 394 by removing constant or almost constant ones, excluding variables filled with missing values for the most part and filtering highly correlated groups. The method compared the observed data with simulations of the atmospheric mouns and other background events for the remaining 262 continuous variables. As an example the following figure shows the situation for the MPE_TT1_HighNoise_Zd attribute:

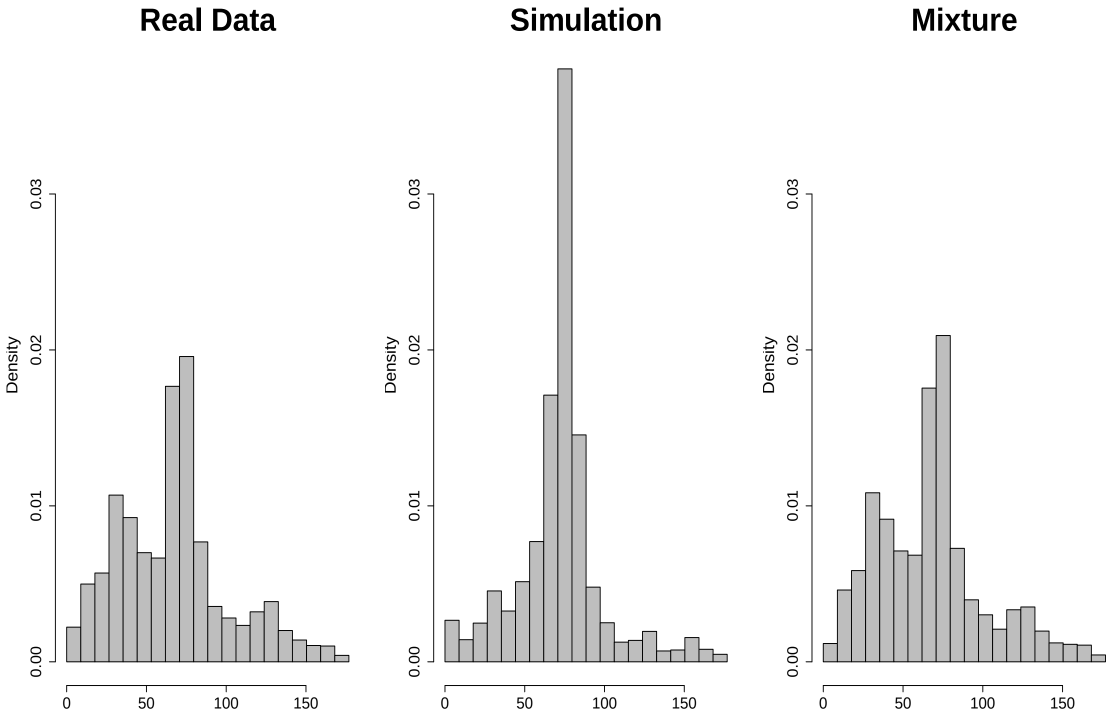


Figure 1: Histrograms of the observed data, the simulated data and the proposed mixture for the variable MPE_TT1_HighNoise_Zd

In comparison to the observed data the simulated ones are overrepresented in the middle between 50 and 100 and at the same time underrepresented for values below 50. The mixture proposed by the ordinary Kolmogorov-Smirnov version of the algorithm corrects the simulation quite well at the cost of a shrinkage factor $\tilde{s} = 0.44$. The results for the IceCube data in general were as follows: for most variables either no correction is needed or the shrinkage parameter $\tilde{s}$ takes a value above 0.9 indicating reasonably good simulations. However some attributes and attribute groups produced by the same reconstruction technique are not simulated appropriately as demonstrated in the above example. As expected tighter confidence bands lead to smaller shrinkage parameters. Both extensions to multivariate techniques and a different modelling of the problem by moving mass instead of shrinking may be studied in further work.

References

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

[2] P. L. Canner
A Simulation Study of One- and Two-Sample Kolmogorov-Smirnov Statistics with a Particular Weight Function
Journal of the American Statistical Association, Vol 70, No. 349.

[3] D. A. Darling
The Kolmogorov-Smirnov, Cramer-von Mises Tests
The Annals of Mathematical Statistics, Vol 28, No. 4.

[4] T. DeYoung
Neutrino Astronomy with IceCube
World Scientific Publishing Company
Modern Physics Letters A, 2009

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

Katja Ickstadt Christian Sohler

Combining Dimensionality Reduction Techniques and Bayesian Regression

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

We combine the dimensionality reduction techniques principal component analysis, factor analysis, and lasso regression with Bayesian regression to achieve an efficient algorithm. Combining principal component analysis and Bayesian regression leads to the fastest computing time. Combining lasso regression and Bayesian regression leads to a considerable reduction in computing time and only a minor reduction in the model fit.

Consider a data set with $n \gg p$, where n is the number of observations and p is the number of variables. We want to analyse data sets with this structure using Bayesian regression analysis. Bayesian methods allow incorporating 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 obtaining the posterior distribution of the parameters of interest. This is a computationally demanding task in all but the simplest cases. The standard methods are Markov Chain Monte Carlo (MCMC) methods, which 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 propose reducing the dimensionality of the data set. While it is possible to reduce the number of observations, here we aim to reduce the number of variables to d (with $d < p$). To do this, we employ a dimensionality reduction technique first before carrying out Bayesian analysis. There are many different methods that can be used to reduce the dimension of a data set. This technical report considers principal component analysis

(PCA), factor analysis, and lasso regression. In the following, these techniques are introduced.

Principal Component Analysis and Factor Analysis

PCA transforms the data into so-called principal components, which are then used for further analysis instead of the original data. The principal components form an orthogonal basis of the data set. Each component explains a decreasing part of the variation in the data, with the first component explaining more variation than any other by definition.

The dimensionality of the data set can be reduced by only using the first d components and discarding the remaining $(p-d)$. There are different criteria which help determining the number of components to be kept. Most of them rely on the eigenvalues of the covariance matrix.

PCA has already been used in combination with classical linear regression. For principal components regression, the components are used as independent variables. This approach yields good results when dealing with multicollinearity, because it ensures that all independent variables are uncorrelated.

Factor analysis is similar to PCA in spirit. Here, it is assumed that latent factors exert influence on the observations. The number of factors is often fixed before the analysis, although there are techniques that can be used if the number is unknown.

The resulting factors are often rotated, e.g. by the varimax rotation. The aim is achieving a factor loading that is either high or low for each variable, thus making interpretation of the results easier.

For further reading on both techniques as well as principal component regression and the number of components to keep, see Jolliffe (1986) [2] and Mardia et al. (1995) [5].

Lasso Regression

Lasso regression has been introduced by Tibshirani (1996) [8]. It belongs to the class of penalised regression techniques and aims for the vector β to contain a substantial amount of zero entries. Lasso regression minimises the term

$$\min_{\beta} \|(y - X\beta)\|_2 + \lambda \|\beta\|_1$$

with respect to $\beta \in \mathbb{R}^p$, where X is a $(n \times p)$-matrix which contains the independent variables and $y \in \mathbb{R}^n$ is the depedent variable. $\lambda \geq 0$ is a positive real number which controls the amount of shrinkage. $\lambda = 0$ results in the least squares estimator, while high amounts of λ tend to lead to sparser estimates of β.

In this context, we only consider the variables, whose corresponding estimates are unequal to zero, for the Bayesian regression, thus using lasso regression for dimensionality reduction.

Results and Outlook

We test the different dimensionality reduction techniques on a data set with $n = 507$ and $p = 22$. This is a rather small data set, but it ensures comparability of the results with standard Bayesian regression.

To obtain the posterior distribution of the parameters, we employ MCMC methods using the software OpenBUGS version 3.2.1 [4], R version 2.15.1 [6] and the R-package R2WinBUGS version 2.1-18 [7]. For lasso regression, the R-package lasso2 version 1.2-12 [3] is used.

We compare the results by their running time and the value of the Deviance Information Criterion (DIC). The DIC is based mainly on the fit of the model, but penalises overly complex models. Lower DIC-values indicate that the model is preferable compared to competing models with higher DIC-values for the same data set. For more details, see Gelman et al. (2004) [1]. For all techniques, we simulated 200 000 MCMC iterations, of which 50 000 were discarded as burn-in, after which convergence was reached in every case.

dimensionality reduction technique	MCMC computing time (sec.)	DIC
none (standard Bayesian regression)	281	2241
PCA (three components)	25	2531
factor analysis (three factors)	25	2555
lasso regression (eight variables)	62	2295

Table 1: Comparison of the different regression types

Table 1 gives an overview of the results. Please note that the MCMC computing time purposely does not include the time needed for carrying out the initial PCA, factor analysis or lasso regression. We plan to include these costs after some possible changes to the algorithms.

Standard Bayesian regression results in the best fit, 281 seconds are needed to conduct MCMC sampling. When applying PCA or factor analysis first, only three components and three factors are kept respectively. In both cases, only 25 seconds are needed. The resulting DIC is considerably higher. Lasso regression (with $\lambda = 0.5$) reduces the number of variables to eight. The analysis takes 62 seconds, while the DIC indicates a reasonable goodness of fit.

Bayesian regression based on PCA outperforms Bayesian regression based on factor analysis, while Bayesian regression based on lasso regression results in a better model fit, but needs more time. This result depends on the setting of the corresponding tuning parameter, which directly or indirectly influences the number of components, factors or variables to keep. We aim to analyse the role of these parameters in more detail in the future.

Future work will also include replacing the PCA with an efficient version based on sketches. This algorithm is currently under development. An important feature is to allow the user to choose the number of components to keep. For classical PCA this number is established after PCA has been carried out. We plan to estimate the number beforehand by efficiently calculating the biggest eigenvalues of the covariance matrix only. One version of an efficient algorithm might be to use efficient sketches instead of matrix X, which results in a greatly reduced number of observations n.

References

[1] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, 2. edition, 2004.

[2] Ian T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer, New York, 1986.

[3] Justin Lokhorst, Bill Venables, and Berwin Turlach. *lasso2: ℓ_1 constrained estimation aka 'lasso'*, 2011.

[4] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28:3049–3067, (2009).

[5] Kantilal V. Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. Academic Press, London, 1995. 10. printing.

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

[7] Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16, 2005.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, 58(1):267–288, 1996.

Efficient Sketches for Bayesian Regression

Alexander Munteanu
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
alexander.munteanu@tu-dortmund.de

In Bayesian regression analysis we are often confronted with the need of evaluating data dependent probability densities. These computations are significantly slowed down when we deal with large data sets, which cannot be stored in main memory. This drawback is driven by frequent access to external memory. We investigate the application of so called *Johnson-Lindenstrauss-Transform* to the data, which reduces the space requirements by storing only a small sketch of the input while preserving statistical structures up to small relative errors.

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 Analysis

In classical regression analysis we are given n data points $x_i \in \mathbb{R}^d$ and their target values $y_i \in \mathbb{R}$, where y is assumed to be chosen as an affine function of X with Gaussian noise, i.e. $y = X\beta + \tau, \tau \sim \mathcal{N}(0, \sigma^2)$. Here β denotes some parameter vector which we would like to estimate from the input data. We emphasize on the *overconstrained* case with $n \gg d$.

Given this scenario and furthermore assuming independence of the observed data, y is distributed according to a Gaussian distribution with mean $X\beta$ and covariance $\sigma^2 I$ which we call the Likelihood $\mathcal{L}(y|X,\beta)$. Computing the maximum likelihood estimator is then equivalent to minimizing the sum of squared errors over all parameter vectors [4], i.e. to computing

$$\hat{\beta} = \arg\min ||X\beta - y||^2. \tag{1}$$

In Bayesian regression analysis we are given some prior distribution $\pi_{pre}(\beta)$ over all parameter vectors and we would like to compute the posterior distribution

$$\pi_{post}(\beta) \propto \mathcal{L}(y|X,\beta)\pi_{pre}(\beta)$$

of the parameters given the observed data. Since for general priors, this distribution cannot be computed analytically, numerical *Markov-Chain-Monte-Carlo*-algorithms like the Metropolis-Hastings algorithm [8,10] or Gibbs sampling [7] are used to draw samples according to the posterior. The samples can subsequently be analyzed in order to estimate parameters of the underlying distribution. The aforementioned algorithms perform a random walk over the parameter space, i.e. the domain of β and need to compute the density for every candidate point proposed by the random walk for deciding whether it will be accepted or rejected. These calculations include the computation of the likelihood and of the prior. The ladder is often chosen to allow fast evaluation. Thus, the computational effort mainly consists of computing $||X\beta - y||^2$ for a very large number of points β throughout the run of the algorithm. Our aim is to approximate this expression for any given point β with high accuracy and much faster.

Random Linear Embeddings

Johnson and Lindenstrauss showed in 1984 [9] that there exists a random matrix $S \in \mathbb{R}^{k \times n}$, $k \in \Omega(\epsilon^{-2} \log n)$ such that for every n-dimensional vector x it holds that

$$(1-\epsilon)||x||^2 \leq ||Sx||^2 \leq (1+\epsilon)||x||^2 \tag{2}$$

with constant probability.

We will refer to such a matrix S as a *Johnson-Lindenstrauss-Transform (JLT)*. There has been extensive work on efficiently computing, storing and evaluating such JLT mappings [1–3] and their application to approximating classical ℓ_2-regression has been studied in the overdetermined [5, 11] as well as in the underdetermined case [6].

Our contribution

Our contribution is to study and apply JLT subject to Bayesian regression analysis or, more precisely, to the approximation of the involved Likelihood computations when dealing with Gaussian models. The choice of the prior distribution remains open, so that a large class of posterior models is still possible.

Similar to [5,6,11], our approach is to maintain sketches SX, Sy of the data while reading the input X, y from a sequential data stream (e.g. a sequential read from hard disk) and perform all computations on the sketches instead of the original data.

Note that evaluating $||SX\beta - Sy||^2$ as an approximation to the original log of the Likelihood gives rise to another Gaussian distribution. An approximation guarantee like (2) does not immediately imply that the statistical parameters of this distribution are preserved under the embedding. Specifically, we would like to have a linear transform such that the mode of the resulting Gaussian distribution as well as the covariances are approximately preserved.

This task requires embedding a whole d-dimensional subspace of $\mathbb{R}^n$ which can be done by discretizing the unit sphere and embedding all grid points leading to $k \in \Omega(d \log(d/\epsilon)/\epsilon^2)$ as shown in [11]. Note that the space complexity is consequently independent of n.

This yields the following approximation guarantees with constant probability. Let β_{opt} be the solution to 1 and let $\tilde{\beta}_{opt}$ be the optimal solution to the embedded instance, i.e. $\tilde{\beta}_{opt} = \arg\min ||SX\beta - Sy||^2$. Furthermore let σ_{min} be the smallest singular value of X, let σ be the vector composed of the singular values of X and let $\tilde{\sigma}$ be the corresponding vector for SX. Then

$$\begin{aligned} ||\beta_{opt} - \tilde{\beta}_{opt}||^2 &\leq \frac{\epsilon^2}{\sigma_{min}} ||X\beta_{opt} - y||^2 & (3) \\ ||\sigma - \tilde{\sigma}||^2 &\leq \epsilon^2 \cdot \sum_i \sigma_i. & (4) \end{aligned}$$

The first bound, given by (3) addresses the estimated mean and an intuitive way to describe it is that the approximation quality is good, when the observed data actually fits the linear model which is assumed. Experiments with more or less fitting data strongly indicate that the dependence on the least squares fit is actually present, not only as a theoretical upper bound.

Furthermore inequality (4) expresses the fact that the d-dimensional subspace spanned by the columns of X is not squeezed or stretched very much, i.e. the covariances are preserved, and is derived from Corollary 11 in [11] which states, that every singular value of X is preserved up to a factor of $(1 \pm \epsilon)$ when embedded using S.

References

[1] Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *JCSS: Journal of Computer and System Sciences*, 66, 2003.

[2] Ailon and Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 2006.

[3] Alon, Matias, and Szegedy. The space complexity of approximating the frequency moments. *JCSS: Journal of Computer and System Sciences*, 58, 1999.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[5] K. L. Clarkson 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.

[6] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *CoRR*, abs/1109.3843, 2011.

[7] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[8] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[9] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26::189–206, 1984.

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092, June 1953.

[11] 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.