Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

December 2016

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

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

Contents

1 Subproject A1 **4**

1.1 Sebastian Buschjäger 5

1.2 Nico Piatkowski 9

2 Subproject A2 **14**

2.1 Amer Krivošija 15

2.2 Chris Schwiegelshohn 19

3 Subproject A3 **24**

3.1 Helena Kotthaus 25

3.2 Jakob Richter 29

4 Subproject A4 **34**

4.1 Markus Buschhoff 35

4.2 Robert Falkenberg 39

4.3 Mojtaba Masoudinejad 43

4.4 Olaf Neugebauer 47

4.5 Janis Tiemann 51

4.6 Aswin Karthik Ramachandran Venkatapathy 55

5 Subproject A6 **60**

5.1 Andre Droschinsky 61

5.2 Elena Erdmann 65

5.3 Martin Mladenov 69

5.4 Christopher Morris 73

6 Subproject B1 **78**

6.1 Salome Horsch 79

7 Subproject B2 **84**

7.1 Wen-Hung Kevin Huang 85

7.2 Thomas Kehrt 89

7.3 Jan Eric Lenssen 93

8 Subproject B3 **98**

8.1 Hendrik Blom 99

8.2 Benedikt Konrad 103

8.3 Marco Stolpe 107

8.4 Mario Wiegand 111

9 Subproject B4 **116**

9.1 Merlin Becker 117

9.2 Lars Habel 121

9.3 Florian Liedmann 125

9.4 Alejandro Molina 129

9.5 Benjamin Sliwa 133

10 Subproject C1 **138**

10.1 Sibylle Hess 139

10.2 Marc Schulte 143

10.3 Henning Timm 147

11 Subproject C3 **152**

11.1 Kai Brügge 153

11.2 Jens Björn Buß 157

11.3 Mathis Börner 161

11.4 Maximilian Meier 165

11.5 Thorben Menne 169

11.6 Maximilian Nöthe 173

11.7 Philipp Schlunder 177

12 Subproject C4 **182**

12.1 Leo Geppert 183

12.2 Alexander Munteanu 187

13 Subproject C5 **192**

13.1 Ulrich Eitschberger 193

13.2 Michael Kußmann 197

13.3 Thomas Lindemann 201

13.4 Frank Meier 205

13.5 Ramon Niet 209

13.6 Margarete Schellenberg 213

13.7 Holger Stevens 217

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

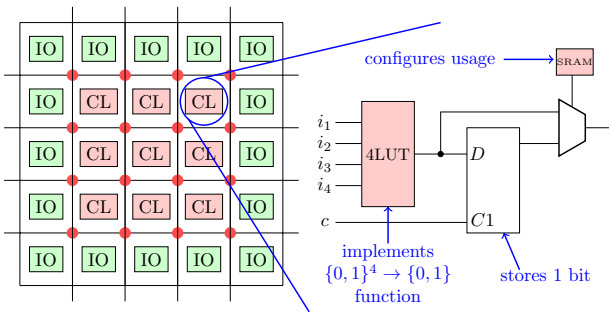
Machine Learning on FPGAs

Sebastian Buschjäger
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
sebastian.buschjaeger@tu-dortmund.de

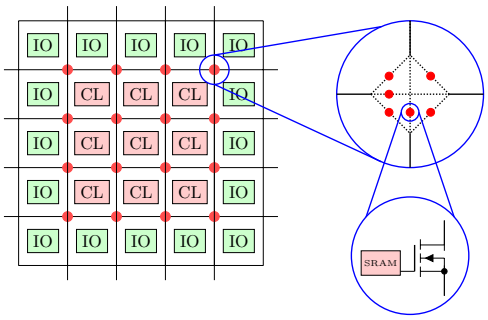
With increasing volumes in data and more sophisticated Machine Learning algorithms, the demand for fast and energy efficient computation systems is also growing. The combination of classical CPU systems with more specialized hardware such as FPGAs offer one way to meet this demand. FPGAs are fast and energy efficient reconfigurable hardware devices allowing new design explorations for algorithms and their implementations. This report briefly discusses FPGAs as computational hardware and their application in the domain of Machine Learning, specifically in combination with Gaussian Processes.

1 Introduction

With increasing integration of information processing systems in every part of everyday life, the amount of gathered data is steadily growing. To further support this integration, the amount of small, energy efficient devices employed in an ubiquitous way is also rapidly increasing.
In order to derive meaningful knowledge from the vast amount of data produced today, Machine Learning algorithms can be used. Once meaningful information is extracted, one is usually interested in performing certain actions based on these new information. Therefore, it is desirable to perform analysis while the observed physical process is still happening opposed to the classical approach of doing analysis afterwards. This effectively means, we need to employ Machine Learning algorithms in small, resource constrained, embedded devices near the physical process.



(a) Configurable logic blocks for FPGAs.



(b) Configurable signal paths of FPGAs.

2 Field Programmable Gate Arrays

Field Programmable Gate Arrays (FPGAs) are programmable hardware units, meaning their functionality is realised in hardware, but this functionality can be altered by programming. This way, FPGAs offer a synthesis between fast and resource efficient execution of application specific integrated circuits and the flexibility of general purpose processing units. This makes FPGAs not only a good research and development tool for hardware-oriented algorithms and the implementation of special purpose hardware, but also for general processing tasks in embedded devices.
In their core, FPGAs consist of three main building blocks. Reconfigurable logic cells (CL) are used to implement simple boolean functions or simple flip-flop memory (cf. figure 1a). Input-output (IO) blocks are used to perform I/O operations and provide a standard interface to a common clock or more sophisticated hardware devices. Logic blocks are connected with configurable signal paths, where each path crossing is configured by a simple transistor set-up, which again is programmable (cf. figure 1b). FPGAs are functional complete.

3 FPGAs in Machine Learning

FPGAs have mainly been used as co-processors in desktop and server systems so far, but less in embedded systems. Additionally, most literature focuses on the possibility to speed-up computation with FPGAs, but not necessarily on ways for energy reduction.
This observation led to my master thesis "Online Gaussian Processes on FPGAs" [2], where I specifically studied the realisation of Gaussian Processes (GP) on FPGAs in an embedded devices context, rather than as co-processor in full desktop or server systems. In order to use Machine Learning algorithms on FPGAs as system-on-a-chip one has to provide basic functionality for communicating training data and predictions. Therefore, I instantiated a virtual CPU on the FPGA and designed a simple chip architecture (cf. figure 2). Based on the `AXI` interface, this system can be extended in a modular fashion.

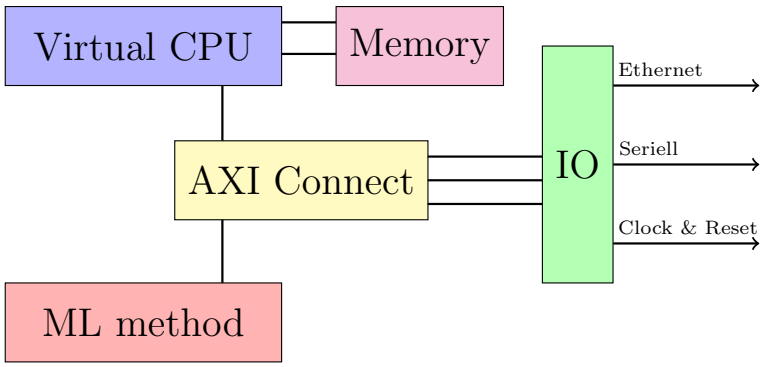


Figure 2: FPGA system architecture.

By careful configuration of the hardware involved and the software running on the virtual CPU, it is possible to abandon RAM memory and only use local memory registers. The resulting system only needs around 0.5W to run a full TCP/IP stack and control a machine learning accelerator block.

Gaussian Processes (GP) are a probabilistic method often used for localization tasks, which frequently occur in the context of embedded systems. GPs assume that every function can be represented with an infinite dimensional vector which is normal distributed. Then, by following the properties of a normal distribution, every marginal - thus every N dimensional sample the function - $\vec{f} = (f(x_1), \ldots, f(x_N))$ is also normal distributed. In this framework, predicting new function values $f(x^*)$ means to compute the marginalised, conditionalized normal distribution for $f(x^*)$ given the training data $\mathcal{D} = (X, \vec{y})$. We can model the variance of such a gaussian distribution with a kernel function $k(x_i, x_j)$ measuring the similarities between x_i and x_j. By introducing a small error σ_n we can account for noise in our data, which leads to:

$$\begin{aligned} f^*|X, \vec{y}, x^* &\sim \mathcal{N}(\widehat{f_*}, cov(f_*)) \\ \widehat{f^*} &= \mathbb{E}[f_*|X, \vec{y}, x^*] = K(x^*, X)[K(X, X) + \sigma_n^2 I]^{-1}\vec{y} \\ cov(f^*) &= K(x^*, X^*) - K(x^*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, x^*) \end{aligned}$$

To compute a standard GP, one has to invert the kernel matrix which is computationally too expensive on embedded devices. Therefore, I used the ProjectionGP algorithm by Csató etal. [3]. This algorithm builds a fixed sized set of basis vector approximating the kernel matrix $K(X, X)$ in an online fashion. Thus, this algorithm can be split into two phases: In the first phase, basis vectors will be added until the set of basis vectors is full. Then, every basis vector is evaluated with respect to the approximation and the least helpful vector is deleted. Counterintuitively, the second phase can be implemented much more efficient on FPGAs than the first phase, because loop unrolling can be performed leading to a speed-up of around 7.5. The complete implementation including the virtual CPU achieves a comparable throughput to embedded ARM devices, but only uses half the energy required by ARM processors.

4 Further research

FPGAs have a long history in Machine Learning, as well as in embedded systems. However, the combination of all three seems to be a new exciting direction. Embedded systems only offer limited memory resources, so that online Machine Learning algorithms seem to be necessary to deal with this challenge. Therefore, online algorithms in the context of embedded systems need to be investigated more. The mentioned ProjectionGP algorithm shares some interesting properties with the informative vector machine (IVM) [4], which seems to belong to a broader set of algorithms utilizing submodular function maximization [1]. The combination of submodular function maximization with the probabilistic interpretation of GPs might lead to new online algorithms with probabilistic guarantees beyond classical worst case analysis.
The presented thesis also laid the ground for further research with FPGAs. Neural Networks are currently one of the most interesting methods in Machine Learning. However, currently available libraries to build neural networks usually target server and desktop systems with GPUs, but not so much embedded systems with hardware accelerators. In order to leverage the knowledge about neural networks, a comprehensive course ("Fachprojekt: DeepLearning on FPGAs") with large practical focus has been developed to introduce students of the TU Dortmund university to embedded system / FPGA programming and deep neural networks. Ultimately, the knowledge gained in this semester will lead to a library for building neural networks on embedded devices and possibility also FPGAs.
Last, FPGAs might also be used in different parts of the SFB876, such as for camera calibration and image processing for FACT telescope data such as in project C3.

References

[1] A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, and A. Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th international conference on Knowledge discovery and data mining*. ACM, 2014.

[2] S. Buschjäger. Online gaussian processes on FPGAs, 2016.

[3] L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

[4] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, 2003.

Inference via Stochastic Quadrature

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

The partition function is fundamental for probabilistic models—it is required for inference, parameter estimation, and model selection. Evaluating this function corresponds to discrete integration, namely a weighted sum over an exponentially large set. This task quickly becomes intractable as the dimensionality of the problem increases. In this report, properties of a recently proposed quadrature-based approximation are discussed.

Graphical Models

Let $\boldsymbol{X}$ be a multivariate random variable where each $\boldsymbol{X}_i, 1 \leq i \leq n$ takes values $\boldsymbol{x}_i$ from space $\mathcal{X}_i$, i.e., $\boldsymbol{X}$ has joint state space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$. The conditional independence structure of $\boldsymbol{X}$ is represented as undirected graph $G = (V, E)$ with $V = [n]$[1]. Assignments to subsets $U \subseteq V$ are denoted by $\boldsymbol{x}_U$ with $\mathcal{X}_U = \bigotimes_{v \in U} \mathcal{X}_v$. We restrict the probability mass function (p.m.f.) of $\boldsymbol{X}$ to belong to an exponential family [6], i.e., it can be written as

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})) = \frac{1}{Z(\boldsymbol{\theta})} \psi(\boldsymbol{x}) \ .$$

Beside its *parameter vector* $\boldsymbol{\theta} \in \mathbb{R}^d$, $\mathbb{P}_{\boldsymbol{\theta}}$ consists of two major ingredients, namely[2] the *potential* $\psi(\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle)$ with *sufficient statistic* $\phi : \mathcal{X} \to \mathbb{R}^d$ and the *partition function* $Z(\boldsymbol{\theta}) = \exp(A(\boldsymbol{\theta}))$. The partition function

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \psi(\boldsymbol{x}) \tag{1}$$

[1] For ease of notation, $[n]$ is an abbreviation for the set of the first n integers $\{1, 2, \ldots, n\}$.
[2] $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denotes the inner product of two d-dimensional vectors.

accumulates the potential of every instance in $\mathcal{X}$ and ensures normalization of $\mathbb{P}_{\boldsymbol{\theta}}$. It can be shown that $Z(\boldsymbol{\theta})$ is in **FP** when G is sufficiently simple, e.g., a tree or fully disconnected. If no special property of the underlying structure can be exploited, evaluating $Z(\boldsymbol{\theta})$ is hard—in fact, **#P**-complete [5]. Hence, there is almost no hope in finding an efficient algorithm for $Z(\boldsymbol{\theta})$. State-of-the-art variational inference procedures approximate $Z(\boldsymbol{\theta})$ by considering simplified distributions [6], but such procedures suffer from unbounded error. We now discuss properties of a quadrature-based, stochastic approximation to (1) formerly proposed in [4].

Stochastic Quadrature-based Inference

If integrating a function f is not tractable, one has to resort to numerical methods in order to approximate the definite integral. The basic idea of a *quadrature rule* is to replace the integrand f by an approximation $g \approx f$ that admits tractable integration. Let $\tilde{\psi}_{\boldsymbol{a}}$ be a degree-k polynomial approximation to the potential function with coefficient vector $\boldsymbol{a} \in \mathbb{R}^{k+1}$ (here, indices of $\mathbb{R}^{k+1}$-vectors start from 0) and worst-case error $\varepsilon > 0$, i.e., $\tilde{\psi}_{\boldsymbol{a}}(\boldsymbol{x}) = \sum_{i=0}^{k} \boldsymbol{a}_i \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle^i$ and $\max_{\boldsymbol{x} \in \mathcal{X}} |\tilde{\psi}_{\boldsymbol{a}}(\boldsymbol{x}) - \psi(\boldsymbol{x})| \leq \varepsilon$. It can be shown that $\varepsilon \leq 4 \exp(\|\boldsymbol{\theta}\|_1)/((k-1)k!)$ when $\tilde{\psi}_{\boldsymbol{a}}$ is a degree-k Chebyshev approximation to ψ [7]. Based on an approximate potential function, we construct an approximate partition function $\tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \tilde{\psi}_{\boldsymbol{a}}(\boldsymbol{x})$, where the error is bounded by $|\tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| \leq \varepsilon |\mathcal{X}|$. Approximating $Z(\boldsymbol{\theta})$ via $\tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta})$ introduces an error without gaining any computational benefit. However, based on this quadrature, we derive a stochastic procedure that allows us to trade off error against complexity without altering the underlying conditional independence structure. We start by defining an important property of sufficient statistics.

Definition 1. *The sufficient statistic $\phi : \mathcal{X} \to \mathbb{R}^d$ is called χ-integrable if, for all $k \in \mathbb{N}_0$, the function $\chi_{\phi}^{k} : [d]^k \to \mathbb{R}$,*

$$\chi_{\phi}^{k}(\boldsymbol{j}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{i=1}^{k} \phi(\boldsymbol{x})_{\boldsymbol{j}_i}$$

can be computed by a deterministic polynomial time algorithm.

Note that $\chi_{\phi}^{0}(\cdot) = |\mathcal{X}|$. It can be shown that overcomplete sufficient statistics of discrete Markov random fields are χ-integrable [4]. We will induce randomness into the quadrature via random index tuples which follow a specific probability mass function.

Definition 2. *Let $\phi : \mathcal{X} \to \mathbb{R}_{+}^{d}$ be a non-negative, χ-integrable sufficient statistic. Let further $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{k+1}$ with $\boldsymbol{b}_i = \|\chi_{\phi}^{i}\|_1$ and $\tau = \langle |\boldsymbol{a}|, \boldsymbol{b} \rangle$, where $\|\chi_{\phi}^{i}\|_1$ denotes the 1-norm of χ_{ϕ}^{i} and $|\boldsymbol{a}|$ denotes the element-wise absolute value of $\boldsymbol{a}$. Then, the probability mass of a random index tuple $\boldsymbol{J}$ with length I is given by $\mathbb{P}_{\boldsymbol{a},\phi}(\boldsymbol{J} = \boldsymbol{j}, I = i) = \mathbb{P}_{\phi}(\boldsymbol{J} = \boldsymbol{j} \mid I = i) \mathbb{P}_{\boldsymbol{a}}(I = i)$ with*

$$\mathbb{P}_{\boldsymbol{a}}(I = i) = \frac{|\boldsymbol{a}_i| \boldsymbol{b}_i}{\tau} \qquad \text{and} \qquad \mathbb{P}_{\phi}(\boldsymbol{J} = \boldsymbol{j} \mid I = i) = \frac{\chi_{\phi}^{i}(\boldsymbol{j})}{\boldsymbol{b}_i} \ .$$

Notice that computing $\boldsymbol{b}$ requires $\mathcal{O}(d^k)$ evaluations of χ, but it depends only on ϕ and $\mathcal{X}$ (and not on $\boldsymbol{\theta}$). It will turn out that the computation of $\boldsymbol{b}$ is the most complex part of the here proposed inference method. Now, we define our estimator and establish its unbiasedness.

Theorem 3. *Let $\phi : \mathcal{X} \to \mathbb{R}_+^d$ be a non-negative, χ-integrable sufficient statistic, $\boldsymbol{a} \in \mathbb{R}^{k+1}$, sgn the signum function, and random variables $(\boldsymbol{J}, I)$ with joint p.m.f. $\mathbb{P}_{\boldsymbol{a},\phi}(\boldsymbol{J} = \boldsymbol{j}, I = i)$ (Def. 2). Then $\tilde{Z}_{I,\boldsymbol{J}}(\boldsymbol{\theta}) = \tau \operatorname{sgn}(\boldsymbol{a}_I) \prod_{r=1}^{I} \boldsymbol{\theta}_{\boldsymbol{J}_r}$ is an unbiased estimator for $\tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta})$.*

Proof. By Defs. 1 and 2:

$$\begin{aligned}
\mathbb{E}\left[\tilde{Z}_{I,\boldsymbol{J}}(\boldsymbol{\theta})\right] &= \sum_{i=0}^{k} \sum_{\boldsymbol{j} \in [d]^k} \mathbb{P}_{\boldsymbol{a},\phi}(\boldsymbol{J} = \boldsymbol{j}, I = i) \tau \operatorname{sgn}(\boldsymbol{a}_i) \prod_{r=1}^{i} \boldsymbol{\theta}_{\boldsymbol{j}_r} \\
&= \sum_{i=0}^{k} \boldsymbol{a}_i \sum_{\boldsymbol{j} \in [d]^k} \left(\prod_{r=1}^{i} \boldsymbol{\theta}_{\boldsymbol{j}_r} \right) \sum_{\boldsymbol{x} \in \mathcal{X}} \left(\prod_{r=1}^{i} \phi(\boldsymbol{x})_{\boldsymbol{j}_r} \right) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=0}^{k} \boldsymbol{a}_i \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle^i \;=\; \tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta})
\end{aligned}$$

■

Definition 4. *(Stochastic Quadrature). Let $\phi : \mathcal{X} \to \mathbb{R}_+^d$ be a non-negative, χ-integrable sufficient statistic, $\tilde{\psi}_{\boldsymbol{a}}$ a degree-k Chebyshev approximation to ψ with coefficient vector $\boldsymbol{a} \in \mathbb{R}^{k+1}$ and worst case-error $\varepsilon > 0$, and $\mathcal{S}$ a set that contains N independent samples from $\mathbb{P}_{\boldsymbol{a},\phi}(\boldsymbol{J} = \boldsymbol{j}, I = i)$. The output S of stochastic quadrature is*

$$S = \frac{\tau}{N} \sum_{(i,\boldsymbol{j}) \in \mathcal{S}} \operatorname{sgn}(\boldsymbol{a}_i) \prod_{r=1}^{i} \boldsymbol{\theta}_{\boldsymbol{j}_r} \,.$$

The following theorem specifies under which conditions we get a bounded relative error with high probability, and shows how to control the approximation quality via N and k.

Theorem 5. *Let S be the output of the stochastic quadrature (Def. 4). Furthermore, let $\zeta \in (0, 1]$, $\epsilon > 0$, $L = \min_i \boldsymbol{\theta}_i^k$, $U = \max_i \boldsymbol{\theta}_i^k$, $N = {}^1\!/_2 (\log {}^2\!/_\zeta) \tau^2 (U - L)^2 \varepsilon^{-2} |\mathcal{X}|^{-2}$ and $(k-1)\,k! \geq 8 \exp(2\|\boldsymbol{\theta}\|_1)/\epsilon$. Then,*

$$\mathbb{P}[|S - Z(\boldsymbol{\theta})| < \epsilon Z(\boldsymbol{\theta})] \;\geq\; 1 - \zeta \,. \tag{2}$$

Proof. By Hoeffding's inequality [1], Theorem 3 and the triangle inequality:

$$\begin{aligned}
\mathbb{P}[|S - \tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta})| \geq \varepsilon |\mathcal{X}|] &\leq 2 \exp\left(-N \frac{2\varepsilon^2 |\mathcal{X}|^2}{\tau^2 (U - L)^2} \right) \;=\; \zeta \\
\Rightarrow \mathbb{P}[|S - \tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta})| + |\tilde{Z}_{\boldsymbol{a}}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| \geq 2\varepsilon |\mathcal{X}|] &\leq \zeta \\
\Rightarrow \mathbb{P}[|S - Z(\boldsymbol{\theta})| \geq 2\varepsilon |\mathcal{X}|] &\leq \zeta
\end{aligned} \tag{3}$$

Applying the error bound on Chebyshev approximations [7] and Hölder's inequality [2]:

$$\varepsilon 2|\mathcal{X}| \;\leq\; \frac{4\exp\|\boldsymbol{\theta}\|_1}{(k-1)\,k!}2|\mathcal{X}| \;\leq\; \epsilon\frac{|\mathcal{X}|}{\exp\|\boldsymbol{\theta}\|_1} \;<\; \epsilon Z(\boldsymbol{\theta}) \tag{4}$$

The last step is also known as *naive mean field lower bound* [6] which follows from Jensen's inequality [3] applied to $\log Z(\boldsymbol{\theta})$. The statement of the theorem is derived by plugging (4) into the probability that is complementary to (3). ■

In some situations, e.g., maximum-likelihood estimation, we might want to estimate $\log Z(\boldsymbol{\theta})$ instead. Hence, it is useful to translate the bound on $|S - Z(\boldsymbol{\theta})|$ into an error bound on $|\log S - \log Z(\boldsymbol{\theta})|$.

Proposition 6. *Assume the preconditions of Theorem 5 hold. If further $S > 0$, then*

$$\mathbb{P}[|\log S - \log Z(\boldsymbol{\theta})| < \epsilon Z(\boldsymbol{\theta})/\min\{S, Z(\boldsymbol{\theta})\}] \;\geq\; 1-\zeta\,.$$

Proof. Let $l = \min\{S, Z(\boldsymbol{\theta})\}$ and $u = \max\{S, Z(\boldsymbol{\theta})\}$. According to the mean-value theorem, there is $\xi \in [l, u]$ such that $\xi|\log l - \log u| = |l - u|$. Plugging this into (2) and using $\xi \geq l$ implies the desired result. ■

References

[1] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[2] Otto Ludwig Hölder. Ueber einen Mittelwerthssatz. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften und der Georg-August-Universität Göttingen*, 2:38–47, 1889.

[3] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[4] Nico Piatkowski and Katharina Morik. Stochastic discrete clenshaw-curtis quadrature. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 3000–3009. JMLR.org, 2016.

[5] Leslie Gabriel Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

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

[7] Shuhuang Xiang, Xiaojun Chen, and Haiyong Wang. Error bounds for approximation in Chebyshev points. *Numerische Mathematik*, 116(3):463–491, 2010.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A2

Algorithmic aspect of learning methods in embedded systems

Christian Sohler Jens Teubner

Some Properties of Time Series in $\mathbb{R}$ Under the Fréchet Distance

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

The possibilities of extension of signatures – the curve simplification technique under continuous Fréchet distance, introduced in [4] to the streaming settings are considered. Also, the approximation error to the Fréchet distance of two curves from $\mathbb{R}^d$ that occures while projecting the curves to the randomly chosen onedimensional space is considered, both for the case of the discrete and the continuous Fréchet distance.

Introduction

Given time series τ with m vertices from $\mathbb{R}^d$, we can represent it as a mapping $\tau : [0, 1] \to \mathbb{R}^d$ and call it (polygonal) curve with complexity m. The notion of δ-signatures as the curve simplification tool was introduced in [4], where $\delta > 0$ measures the similarity of the original and the simplified curve. It is shown to be an useful tool for the clustering of the input consisting of n onedimensional ($d = 1$) time series of complexity m under the (continuous) Fréchet distance, by extracting the key information from the input while keeping the complexity ℓ of the simplified curve constant, or keeping the simplification error δ small. We know that the signatures always exist and have the efficient algorithms to compute them [4].

As the follow-up we wanted to explore the possibilities to extend this approach into the streaming settings, under the two possible scenarios defined by the relevance of the vertices of the input curve τ seen up to some moment. These are the complete input and the sliding window settings. We want to know how large can be the distance between the original curve and the simplified curve if the complexity ℓ of the simplified curve is

given, as well as how many vertices of the input curve we have to store, in order to keep the error δ small.

The second set of question we have considered is related to the projections of the curves from $\mathbb{R}^d$ to randomly chosen straight line (i.e. onedimensional subspace). We want to know how large is the ratio of the Fréchet distance $d_F(P, Q)$ between given two curves $P, Q \in \mathbb{R}^d$ and their projections, and which approximation factors can be obtained. These considerations are of probabilistic nature. These problems are considered both for the case of the discrete and the continuous Fréchet distance.

Time series in $\mathbb{R}$ under the streaming settings

The definition of the range of the time stamps of the input time series is easily extended from $[0, 1]$ to $[0, \infty)$ to obtain *time series stream*. Given time series stream $\tau : [0, \infty) \to \mathbb{R}$, we want to find the δ-signature $\sigma : [0, \infty) \to \mathbb{R}$ of τ, both for the case when the error measure $\delta > 0$ or the goal complexity ℓ is given, and for two different settings: (a) when the whole curve τ is considered (complete stream), and (b) in the sliding window model. Under the sliding window model, we consider at time t only the vertices of τ whose time stamp is within $[t - \varphi, t]$, i.e. the subcurve $\hat{\tau} = \tau[t - \varphi, t]$, with $\varphi > 0$ being the sliding window size.

Let $\ell \in \mathbb{N}$ be given constant. Abam *et al.* [1] have considered simplification of general paths in $\mathbb{R}^2$ under the Fréchet distance, obtaining an algorithm that for any fixed $\varepsilon > 0$ produces $4\sqrt{2}+\varepsilon$-competitive streaming algorithm that uses $O(\ell^2/\sqrt{\varepsilon})$ additional storage and processes each input point in $O(\ell \log(1/\varepsilon))$ amortized time. They have used the technique of Imai and Iri [5], given originally for Hausdorrf distance.

In our simplification approach with signatures, we can maintain the calculated δ-signature σ of the input time series stream τ for the complete stream settings. This is claimed by following lemma, which is proven by extending the algorithms for the computing of the minimum size and minimum error signatures respectively to the streaming settings.

Lemma 1. *Given a (possibly infinite) streaming curve $\tau : [0, \infty) \to \mathbb{R}$ with vertices in general position, then*

i) given a parameter $\delta > 0$, we can maintain a δ-signature σ of τ of minimum size using time and space $O(1)$ per update;
ii) given a parameter ℓ, we can maintain a minimum-error δ-signature σ of τ with $\ell' \leq \ell$ vertices using time $O(\ell \log \ell)$ per update and space $O(\ell)$ to store the vertices of the signature σ.

For the sliding window settings we adapt the scenario from the general streaming: the vertices of τ come in a stream, but after some time the oldest ones are no longer valid

and have to be ignored, i.e. τ is reduced to the last p vertices, such that their time stamp is within the sliding window of width φ. We have to consider the newly coming vertices (as in Lemma 1), but also the effect of excluding the currently oldest vertex $w_j \in \tau$. Unfortunately the construction of (small) signatures is not possible, since we could have to keep all the vertices of τ in memory to keep the exact signature $\hat{\sigma}$ of the curve $\hat{\tau}$.

Lemma 2. *Given curve τ and its subcurve $\hat{\tau} = \tau[t - \varphi, t]$, then for any $\delta, \ell, \varphi, t > 0$ it holds that*

i) the calculation of the δ-signature of the curve $\hat{\tau}$ needs linear update time and space in complexity of $\hat{\tau}$;
ii) the calculation of the minimum-error signature of the curve $\hat{\tau}$ with at most ℓ vertices needs the space at least as big as complexity of $\hat{\tau}$.

Instead of keeping the whole curve $\hat{\tau}$ in the memory, we show that the good approximation can be obtained by keeping the signature vertices only (Lemma 3).

Lemma 3. *Given $\delta > 0$, let $\tau = \{w_1, \ldots, w_m\}$ be the streaming curve and σ its δ-signature. Let $\hat{\tau} = \{w_i, \ldots, w_m\}$ be the curve obtained after the vertices $\Phi = \{w_1, \ldots, w_{i-1}\}$ expired and no longer belong to the part of the curve we observe. Let $\hat{\sigma}$ be the optimal δ-signature of $\hat{\tau}$. Let $\sigma' = \sigma \setminus \Phi$. Then the curves σ' and $\hat{\sigma}$ differ it at most first three vertices.*

It is still an open problem to find an efficient approximation to $\hat{\sigma}$, which can be found while keeping only the constant number of the curve vertices in the memory.

Projections to the onedimensional space

Given two curves P and Q from $\mathbb{R}^d$ of complexity m, Bringmann and Mulzer [3] have shown that there is no strongly subquadratic algorithm to calculate neither the discrete nor the continuous Fréchet distance between two curves, unless Strong Exponential Time Hypothesis (SETH) fails, and this claim holds even in onedimensional case. They have also shoved that there is the simple linear time $2^{\Theta(m)}$-approximation algorithm to $d_F(P, Q)$, as well as an α-approximation algorithm to $d_F(P, Q)$ in strongly subquadratic time. Both results hold for any dimension $d \geq 1$.

Maheshwari *et al.* [6] have considered the problem of deciding whether a given curve is similar to some part of a larger curve under given Fréchet distance error, and obtained the linear time algorithm for this problem, improving the classical result of Alt and Godau [2] by a logarithmic factor.

Inspired by these results we wanted to investigate the relation between the Fréchet distance of two given curves P and Q from $\mathbb{R}^d$, and the Fréchet distance of their projections

P' and Q' to some onedimensional space, whose unit vector $\mathbf{u}$ is chosen uniformly at random. In one special case the following lemma holds:

Lemma 4. *Given curves P of complexity m and Q of complexity 2, and let P' and Q' be their projections to some onedimensional space chosen uniformly at random. Then $d_F(P,Q)/d_F(P',Q') \leq 6$ with positive constant probability.*

The discrete Fréchet distance of P and Q matches the vertices of P only to the vertices of Q (and not to some point along Q, as it is the case for the continuous version). For the general curves P and Q of complexity m and both the discrete and the continuous Fréchet distance, there is a simple $O(m^2)$-approximation to $d_F(P,Q)$ using the free-space diagrams of Alt and Godau [2].

We have shown that using projections to the onedimensional space the factor of at least $O(m)$ may be needed to approximate $d_F(P,Q)$ by $d_F(P',Q')$. We want to match this lower bound with the equal upper bound, as some preliminary observations suggest. But to obtain this, some additional assumption on P and Q has to be made. One such assumption could be that these curves are c-packed, for some $c > 0$. For the curve P this means that for any point $p \in \mathbb{R}^d$ and any radius $r > 0$, the total length of the curve P inside the ball centered at p and with radius r is at most $c \cdot r$. This work is currently in progress.

References

[1] M. Abam, M. de Berg, P. Hachenberger, and A. Zarei. Streaming algorithms for line simplification. *Discrete & Computational Geometry*, 43:497–515, 2010.

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

[3] K. Bringmann and W. Mulzer. Approximability of the discrete Fréchet distance. *JoCG*, 7(2):46–76, 2016.

[4] A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the Fréchet distance. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 766–785, 2016.

[5] H. Imai and M. Iri. Polygonal approximations of a curve – formulations and algorithms. In G. Toussaint, editor, *Computational Morphology*, pages 71–86. North-Holland, Amsterdam, 1988.

[6] A. Maheshwari, J. Sack, K. Shahbaz, and H. Zarrabi-Zadeh. Improved algorithms for partial curve matching. *Algorithmica*, 69(3):641–657, 2014.

Finding the Jaccard Center

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

The Jaccard index is a widely used similarity measure on item sets. Given two sets X and Y over a ground set U, the similarity is defined as $J(X,Y) = |X \cap Y|/|X \cup Y|$ and the distance is $D(X,Y) = 1 - J(X,Y) = |X \triangle Y|/|X \cup Y|$. In this paper we study the problem of finding the center of a given set of item sets under the Jaccard distance, i.e. given a set of sets $N = \{X_1, \dots, X_n\}$ finding a set C such that $\max D(X,C)$ for $X \in N$ is minimized. The Jaccard index is arguably the oldest and best known similarity measure on binary data. The to this paper most related previous research is on finding the Jaccard median, i.e. finding a item set that minimizes the sum of Jaccard distances, see [2, 3]. More recently, Chierichetti et al. [1] showed that the Jaccard median problem is NP-hard but also admits a PTAS.

Hardness We reduce the problem of finding the optimum Jaccard center from vertex cover with minor constraints that do not affect the hardness of the problem. Given a graph $G(V,E)$, a vertex cover is a set $K \subset V$ such that for any $e \in E$ $e \cap K \neq \emptyset$. The minimum vertex cover is the vertex cover of smallest cardinality.

Theorem 1. *Computing the optimum Jaccard center is NP-hard.*

Proof. Let K be a minimum vertex cover of cardinality at most $\frac{|V|}{2} - 2$ in a graph with no isolated nodes. We claim that a collection of vertexes C is an optimum Jaccard center if and only if C is a minimum vertex cover.

Now consider a collection of vertexes C. We have for any edge $e \in E$

$$D(e,C) = \begin{cases} 1 & \text{if } |C \cap e| = 0 \\ \frac{|C|}{|C|+1} & \text{if } |C \cap e| = 1 \\ \frac{|C|-2}{|C|} & \text{if } |C \cap e| = 2. \end{cases}$$

Note that the distance for some edge is 1 if and only if C is not a vertex cover. Note also that $\frac{|C|}{|C|+1} > \frac{|C|-2}{|C|}$, i.e. if $C \neq V$ then $\max_{e \in E} D(e, C) = \frac{|C|}{|C|+1}$. Now for any collection of vertexes C that are vertex covers such that $|C| > |K|$, we have two cases. First $C \neq V$. Then $\max_{e \in E} D(e, C) = \frac{|C|}{|C|+1} \geq \frac{|K|+1}{|K|+2} > \frac{|K|}{|K|+1} = \max_{e \in E} D(e, K)$. If $C = V$ then $\max_{e \in E} D(e, V) = \frac{|V|-2}{|V|} = \frac{\frac{|V|}{2}-1}{\frac{|V|}{2}} \geq \frac{|K|+1}{|K|+2} > \frac{|K|}{|K|+1} = \max_{e \in E} D(e, K)$. □

Core-Covers Our algorithms are based on the existence of a small subset M of input points such that a high-quality center can be extracted from M. The construction is somewhat inspired by coresets for the Euclidean minimum enclosing ball problem, albeit with a weaker guarantee.

Definition 1. Let N be a collection of subsets of a ground set U, let OPT be the value of an optimal Jaccard center and let $\alpha \geq 1$ be a parameter. A subset $M \subseteq N$ is called an *α-core-cover* if there exists an optimal center C with $K = C \cap \left(\bigcup_{X \in M} X \right)$ and

$$\max_{X \in N} D(X, K) \leq \alpha \cdot \mathrm{OPT}.$$

A subset $M \subseteq N$ with $A_M = \bigcap_{X \in M} X$ and $O_M = \bigcup_{X,Y \in M} X \triangle Y$ is called an *anchored α-core-cover* if there exists an optimal center C with

$$\max_{X \in N} D(X, A_M \cup (O_M \cap C)) \leq \alpha \cdot \mathrm{OPT}.$$

We are especially interested in the size of core-covers with $\alpha = 1 + \varepsilon$. For the remainder of this section, we will give (non-constructive) upper and lower bounds on the number of points required to satisfy both guarantees. Our proofs are essentially based on the following observation.

Observation 1. $D(X, K) \leq D(X, K \cap C) + \frac{|K \setminus C| - 2|(X \cap K) \setminus C)|}{|X \cup K|}$ for any three sets $C, K, X \subseteq U$.

If X is an arbitrary input point, K is our possible solution and C is an optimal center, this observation implies that it is sufficient to show that $D(X, K \cap C)$ is a good approximation to $D(X, C)$ and $\frac{|K \setminus C| - 2|(X \cap K) \setminus C)|}{|X \cup K|}$ is smaller than $\varepsilon \cdot D(X, C)$. We state the result for core-covers and give a proof for anchored core-covers.

Lemma 1. *For any collection of subsets N, there exists an $(1 + \varepsilon)$-core-cover of size $\lceil 1/\varepsilon \rceil + 1$.*

Lemma 2. *For any collection of subsets N, there exists an anchored $(1+\varepsilon)$-core-cover $M \subset N$ of size $O(1/\varepsilon)$.*

Proof. Assume we have some optimal center C. Lemma 1 gives a set M such that $K \cap C$ for $K = A_M \cup (O_M \cap C)$ is an $(1+\varepsilon)$-approximate solution. Using Observation 1, the distance between K and some arbitrary point X is

$$\begin{aligned} D(X,K) &\leq D(X, K\cap C) + \frac{|K \setminus C| - 2\cdot|(X\cap K)\setminus C)|}{|X \cup K|} \\ &\leq (1+\varepsilon)\cdot \mathrm{OPT} + \frac{|A_M \setminus C| - 2\cdot|X \cap (A_M \setminus C)|}{|X\cup K|} \end{aligned}$$

If for every $X \in N$ we have $2\cdot|X\cap(A_M\setminus C)| > |A_M \setminus C|$, the right summand is negative and $D(X,K) \leq D(X, K\cap C) \leq (1+\varepsilon)\cdot\mathrm{OPT}$. Otherwise, there exists an X such that $|(A_M \cap X)\setminus C| \leq |A_M \setminus C|/2$. In the following, assume that we augment a set M satisfying the space and approximation bounds of Lemma 1 with additional points.

We assume $\mathrm{OPT} < 1/(1+\varepsilon)$ as otherwise any point is a $(1+\varepsilon)$ approximation. Let $X \in N$. We have

$$|A_M \setminus C| \leq \mathrm{OPT}\cdot\frac{|C|}{1-\mathrm{OPT}} \leq \mathrm{OPT}\cdot\frac{|X|}{(1-\mathrm{OPT})^2} \leq \mathrm{OPT}\cdot\frac{4}{\varepsilon^2}\cdot|X|.$$

After adding $\log\frac{4}{\varepsilon^3}$ points such that $|A_M \setminus C|$ is halved with each point, we have $|A_M \setminus C|/|X\cup K| \leq \varepsilon\cdot\mathrm{OPT}\cdot|X|/|X\cup K| \leq \varepsilon\cdot\mathrm{OPT}$. Rescaling ε completes the proof. □

PTAS The algorithm consists of two main steps. Let $\widehat{\mathrm{OPT}}$ be a candidate objective value, note that there exist $O(d^2)$ distinct objective values for the Jaccard center problem with a ground set of size d. We aim to find a real vector C' such that $D(X,C') \leq \widehat{\mathrm{OPT}}$ is satisfied for all $X \in N$. By multiplying both sides with $|X\cup C|$, we obtain

$$|X \triangle C| \leq \widehat{\mathrm{OPT}}\cdot|X\cup C|. \tag{1}$$

Observe that $|X\triangle C| = \sum_{i=1}^{d} X_i - 2X_iC_i + C_i$ and $|X\cup C| = \sum_{i=1}^{d} X_i - X_iC_i + C_i$. Hence, we obtain a set of linear inequalities which we can be test for feasibility if we relax the integrality constraints on C. Note that if there exists a feasible integral solution of Equation 1, there also exists a feasible relaxed solution C'. We interpret the C'_i as probabilities, i.e. we obtain a binary vector C by rounding each C'_i to 1 with probability C'_i. Using Chernoff bounds, this approach yields a good solution if $\mathrm{OPT}\cdot|X| > s\cdot\log n/\varepsilon^2$ for all X and some constant s.

If $\mathrm{OPT}\cdot|Y|$ is smaller than this threshold for at least one $Y \in N$, then we could employ a naive brute force algorithm by iterating over all $\binom{d}{s\cdot\log n/\varepsilon} \in O(d^{s\cdot\log n/\varepsilon})$ subsets S and outputting the best $Y \triangle S$. To eliminate the dependency on d, we first show that a

bound on $\mathrm{OPT} \cdot |Y|$ implies that $|X_1 \triangle X_2|$ for any two points $X_1, X_2 \in N$ is bounded. Then we compute an anchored core-cover M by iterating over all sets of $O(1/\varepsilon)$ input points. Having determined M, computing the optimum $A_M \cup S$ with $S \subseteq O_M$ becomes feasible.

Theorem 2. *Given a set of n item sets N from a ground set U of cardinality d and any $\varepsilon > 0$, there exists an algorithm computing a $(1+\varepsilon)$-approximation to the optimal Jaccard center. The algorithm runs in time $d^2 \cdot (n^{O(\varepsilon^{-6})} + LP(n, d))$, where $LP(n, d)$ is the time required to solve a linear program with n constraints and d variables.*

Proof. In the following, we always assume that $\mathrm{OPT} < 1/(1+\varepsilon)$, as otherwise any solution is a $(1+\varepsilon)$ approximation.

Lemma 3. *Let S be a random binary vector obtained by rounding a fractional feasible solution of the set of Equations 1 and let $\varepsilon > 0$ be a constant. Assume that $OPT \cdot |X| \geq \frac{27 \ln(4n)}{\varepsilon^2}$ for all $X \in N$. Then with probability at least $1/2$, the rounding procedure produces a binary solution S with $\max_{X \in N} D(X, S) \leq (1+\varepsilon) \cdot OPT$.*

If $\mathrm{OPT} \cdot |X| > \frac{27 \ln(4n)}{\varepsilon^2}$ for all $X \in N$, we can use the LP-based rounding scheme analyzed in Lemma 3. For the other cases, we will utilize Lemma 2 as follows. Assume that there exists at least one point Y with $\mathrm{OPT} \cdot |Y| \leq \frac{27 \ln(4n)}{\varepsilon^2}$. We have $\mathrm{OPT} \cdot |C| \leq \mathrm{OPT} \cdot |Y|/(1 - \mathrm{OPT}) \leq \frac{27 \cdot (1+\varepsilon) \cdot \ln(4n)}{\varepsilon^3}$. For any two points $X_1, X_2 \in N$, we then have $|X_1 \triangle X_2| \leq \frac{108 \cdot (1+\varepsilon)^2 \cdot \ln(4n)}{\varepsilon^4}$. Let M now be a set of points satisfying the guarantee of Lemma 2. Such a set can be determined in time $n^{O(\varepsilon^{-1})}$ by iterating through all subsets of cardinality $O(\varepsilon^{-1})$. Since $|O| \leq \sum_{X_i \in M} \sum_{X_j \in M} |X_i \triangle X_j| \leq |M|^2 \cdot \max_{X_i, X_j \in M} |X_i \triangle X_j| \in O(\log n/\varepsilon^{-6})$, we can compute an optimal solution of $\max_{X \in N} \min_{S \subseteq O} D(X, A \cup S)$ in time $2^{|O|}$. The total running time amounts to d^2 calls to the LP given via Equations 1 or d^2 applications of Lemma 2 with a running time of $2^{O(\log n \cdot \varepsilon^{-6})} = n^{O(\varepsilon^{-6})}$. □

References

[1] Flavio Chierichetti, Ravi Kumar, Sandeep Pandey, and Sergei Vassilvitskii. Finding the Jaccard median. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 293–311, 2010.

[2] H. Späth. The minisum location problem for the Jaccard metric. *Operations-Research-Spektrum*, 3(2):91–94, 1981.

[3] G. A. Watson. An algorithm for the single facility location problem using the Jaccard metric. *SIAM Journal on Scientific and Statistical Computing*, 4(4):748–756, 1983.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A3

Methods for Efficient Resource Utilization in Machine Learning Algorithms

Peter Marwedel Jörg Rahnenführer

Resource-Aware Scheduling Strategies for Parallel Model-Based Optimization

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

This report presents resource-aware scheduling strategies for parallel R applications. One important application is hyperparameter optimization of machine learning algorithms. For hyperparameter optimization, model-based optimization is used to optimize the resampled performance measurement of a machine learning algorithms on a given dataset. Here, resource demands like CPU utilization or memory usage heavily vary depending on the type and configuration of the applied learning algorithm. Running such an application in a naive parallel way leads to inefficient resource utilization and thus to long runtimes. Compared to standard mechanisms our scheduling approach improves the utilization of parallel computer architectures by estimating resource demands.

1 Parallel Model-based Optimization

The R package `parallel` [1] offers a scheduling strategy, called load balancing. It dynamically allocates tasks to worker processes. This option is recommended when tasks have widely different computation times or if computer architectures are heterogeneous. We analyzed memory and CPU utilization of parallel applications with our TraceR [2] profiling tool and found that the load balancing mechanism is not sufficient for parallel tasks with high variance in resource demands [3]. A scheduling strategy needs to know resource demands of a task before execution to efficiently map applications to available resources. Therefore, we build a regression model to estimate resource demands based on previous evaluated tasks. Resource estimates like runtime are then used to

guide our scheduling strategies. Those strategies are integrated in our Resource-Aware Model-Based Optimization Framework (RAMBO) [4] [5]. One of our goals is to obtain best possible predictions regarding diagnosis or therapy response for patients within a given time budget. For improved prediction we need to find the best machine learning configuration, therefore we use hyperparameter optimization. Here, model-based optimization is applied to reduce necessary evaluations of configurations. Additionally we evaluate configurations proposed by the model in a resource-aware parallel way to get better performance within the same time budged for a fast diagnosis.

Model-based optimization (MBO) is an iterative optimization algorithm that starts on an initial set of already evaluated configurations. In each step a regression model is fitted on the so far available evaluations. It serves as a surrogate model to predict the outcome of yet unseen configurations. The infill criterion of the model guides the search to a new configuration which is usually a compromise between good predicted performance and uncertainty of the search space region. The new configuration is evaluated, appended to the current data and the next MBO-iteration step starts until the time budget is exhausted [6]. Each configuration forms a job with different resource demands. For efficient scheduling of parallel point proposals we estimate the resource demands of configurations with an additional surrogate model. Resource-aware scheduling is an active field of research which is often tailored specifically for different hardware platforms, from small embedded systems [7] up to heterogeneous clusters [8]. In contrast to these classical scheduling problems, we are in control of the job generation as we can query the resource model for jobs with suitable resource requirements and postpone or skip suggested jobs.

2 Resource-Aware Scheduling Strategies

For scheduling, we use a synchronous approach, where the results of all jobs (configurations) within one MBO-iteration are gathered before each model update. One of the naive approaches for synchronous MBO is qLCB. qLCB is a multipoint LCB infill criterion to propose q points for one iteration [9]. At each sequential step the number of evaluated jobs is equal to the number of available CPUs. As one of our first resource-aware scheduling strategies for the synchronous approach we used a first fit heuristic [5]. Here, in each iteration we generate a list of proposed jobs with the help of the qLCB infill-criterion. We then determine the job with the highest priority (based on infill-criterion) and run this job on the first CPU exclusively. Its estimated runtime serves as a deadline for all other parallel running jobs and thus as the deadline for the current MBO-iteration. We expect that the result of the job with the highest priority is most profitable for the model update. If any CPU is left without a job we query the surrogate model for a new job. After the jobs are evaluated the surrogate model is updated and the next iteration starts over.

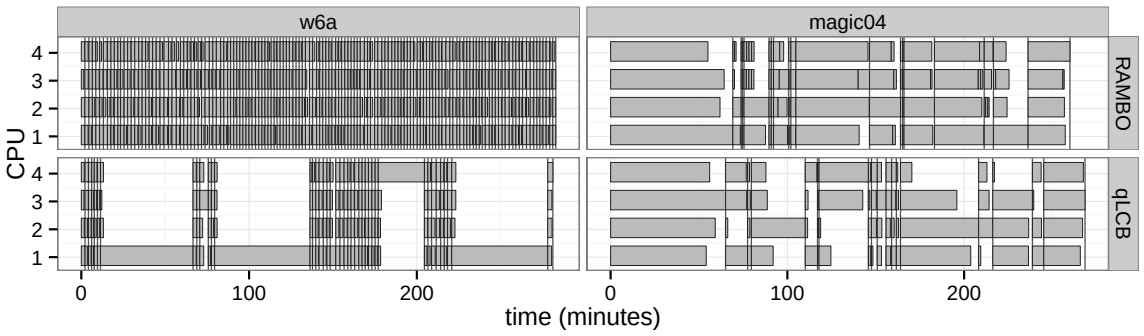


Figure 1: Scheduling visualization for w6a and magic04 data sets. Vertical lines indicate the end of one MBO iteration

Figure 1 visualizes the mapping of parallel jobs for two optimization runs for the naive approach qLCB and for our new RAMBO approach, where the first fit scheduling is integrated. Boxes indicate mapping of jobs on CPUs. Vertical lines indicate the end of one MBO-iteration and empty spaces indicate CPU-idle time. Whereas RAMBO balances long execution times more evenly qLCB has only a low CPU utilization. RAMBO achieves higher CPU utilization, meaning more evaluations which yields better knowledge of the hyperparameter space and thus higher confidence in the optimization result. It also shows that our scheduling strategy not only prefers short jobs but is also able to schedule long jobs more efficiently. On the `w6a` dataset RAMBO is capable of evaluating twice as many configurations as the naive qLCB approach.

3 Future Work

As a next step we will evaluate new scheduling strategies [10] for the synchronous RAMBO approach. We already found, that the current qLCB infill-criterion is not perfect for producing scheduling priorities. Therefore we plan to design a multi-objective approach with a so called group infill-criterion based on the underlying hardware architecture, the distance of configurations within the model space and the resource demands. We currently also compare the synchronous RAMBO approach with an asynchronous approach. In the asynchronous approach the model is updated each time a job has finished its computation. This approach increases runtime overhead for model-update and point proposal but also decreases the CPU idle-time. To evaluate energy consumption of RAMBO on heterogeneous architectures we consider to transfer our framework to the Odroid XU-3 platform which was already used in the part project B2 [11].

References

[1] Ripley B, Tierney L., Urbanek S., parallel: Support for Parallel Computation. R package included in the R-core 2.14.0. 2015. URL http://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf

[2] TraceR: A Profiling Tool for R. TU Dortmund University. 2016. URL https://github.com/allr/traceR-installer

[3] Kotthaus H., Korb I., Marwedel P., Performance Analysis for Parallel R Programs: Towards Efficient Resource Utilization. Technical Report 01/2015, Department of Computer Science 12, TU Dortmund University July 2015, SFB876 Project A3

[4] RAMBO: Resource-Aware Model-Based Optimization. 2016. URL https://github.com/mlr-org/mlrMBO/tree/smart_scheduling

[5] Richter J., Kotthaus H., Bischl B., Marwedel P., Rahnenführer J., Lang, M., Faster Model-Based Optimization through Resource-Aware Scheduling Strategies. Proceedings of the 10th Learning and Intelligent OptimizatioN Conference (LION'10) (accepted for publication). Ischia Island, Italy, 2016

[6] Haftka, R.T., Villanueva, D., Chaudhuri, A., Parallel surrogate-assisted global optimization with expensive functions - a survey. Journal of Structural and Multidisciplinary Optimization. 54(1), pp. 3-13, 2016

[7] Tillenius, M., Larsson, E., Badia, R.M., Martorell, X., Resource-Aware Task Scheduling. ACM Trans. on Embedded Computer Systems (TECS). 14(1), 5:1–5:25, 2015

[8] Delimitrou, C., Kozyrakis, C., Quasar: Resource-efficient and QoS-aware Cluster Management. Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14). pp. 127–144, 2014

[9] Bischl, B., Wessing, S., Bauer, N., Friedrichs, K., Weihs, C., MOI-MBO: Multiobjective Infill for Parallel Model-Based Optimization. Proceedings of the 8th Learning and Intelligent OptimizatioN Conference (LION'8). Florida, 2014

[10] Kotthaus H., Richter J., Lang A., Lang M., Rahnenführer J., Marwedel P., Resource-Aware Scheduling Strategies for Parallel Machine Learning R Programs through RAMBO. RIOT: R Implementation, Optimization and Tooling Workshop. Stanford University, Palo Alto, California. 2016.

[11] Neugebauer O., Libuschewski P., Engel M., Müller H., Marwedel P., Plasmon-based virus detection on heterogeneous embedded systems. In Proceedings of Workshop on Software and Compilers for Embedded Systems (SCOPES) 2015

Faster Model Based Optimization through Resource Aware Scheduling and asynchronous Evaluations

Jakob Richter
Faculty of Statistics
TU Dortmund University
jakob.richter@tu-dortmund.de

Optimizing expensive black-box functions requires many resources, mainly CPU time. To save overall time, main goals are to reduce and to parallelize necessary black-box evaluations efficiently. To reach the first goal a state-of-the-art method is the model-based optimization (MBO) [8] which is still subject to research [3, 6]. In my ongoing work in cooperation with Helena Kotthaus we develop a new MBO method for highly parallel computing infrastructures that reduces idle time and thus leads to a faster optimization result. Furthermore we investigate on existing methods with focus on their scalability. MBO in itself is a sequential algorithm which makes parallelization a non-trivial task that has been subject of further research [4, 7, 2]. Existent solutions can be split into the categories: synchronous and asynchronous. For the synchronous approach a master process proposes a set of black-box configurations to be evaluated by the slaves before making the next proposal. In the asynchronous case each process autonomously generates a single proposal for itself based on the available result of all previous runs.

Problems to solve are: When does the calculation overhead for the asynchronous approach affect optimization negatively? How can we reduce the idle time most efficiently in the synchronous case? Can we improve the set of proposed points in the synchronous case?

The aim of optimization is to find a configuration for the given black-box that minimizes a loss-function defined on a black-box evaluation. MBO fits a regression model on the set of known configurations and known outcomes of the loss-function to prevent evaluations

of unpromising configurations and thus saves resources. In each iteration, based on the regression model, a promising candidate is proposed and then evaluated.

1 The synchronous approach

The synchronous approach is closely related to the classical MBO approach. In ordinary MBO, only the one most promising candidate configuration is evaluated per iteration. In the parallel MBO approach a set of q candidate configurations is evaluated in parallel. Promising configurations are determined by optimizing a so called infill criterion. Infill criteria generally try to balance exploitation: the regression model expects a low outcome of the loss-function, and exploration: the regression model indicates uncertainty. In equation 1 the definition of the *Lower Confidence Bound* (LCB) is given:

$$LCB(\boldsymbol{x}) := \hat{y}(\boldsymbol{x}) - \lambda \cdot \hat{s}^2(\boldsymbol{x}), \tag{1}$$

whereas $\hat{y}(\boldsymbol{x})$ denotes the predicted outcome of the black-box for the candidate configuration $\boldsymbol{x}$ and $\hat{s}^2(\boldsymbol{x})$ denotes uncertainty of the regression model. Proposals on how to choose q candidates simultaneously have been made in [2]. Usually q is chosen equivalent to the number of available cores (k) and accordingly k candidates are evaluated in each iteration. To obtain q different values for LCB λ is drawn from an exponential distribution. This is then called $qLCB$. When all evaluations are finished, the model is be updated with the results and the next q candidates will be proposed according to the updated regression model. If the runtimes of the black box evaluations depend strongly on the chosen configuration it happens that some evaluations take longer then others and thus, some cores are waiting for new proposals. This leads to high amounts of unused CPU-time which means unused resources.

To tackle this problem for the synchronous approach we use a further regression model to predict the runtime of each configuration based on past configurations and their measured runtime. This enables us to schedule the evaluations beforehand so that idling is reduced.

We implemented a basic first fit scheduling heuristic in our RAMBO-Framework [9]. There we propose $q = 3 \cdot k$ configurations and give highest priority to the one that minimize the loss (small value of λ). This has shown to be an improvement over the unscheduled $qLCB$ approach as idle time was reduced and thus the number of evaluations went up and the optima could be reached faster.

One challenge is to find a set of points, for which not each point itself is seen as promising but the set as a whole. This should prevent points close to each other being evaluated in one iteration. Hence, we focus our research on a meaningful group infill criterion

that balances the distance in connection with their infill criteria. A draft is given in in equation 2:

$$Inf(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}) = \sum_{i=1}^{q} [\hat{y}\boldsymbol{x}_i - \lambda \cdot \hat{s}^2(\boldsymbol{x})] * \gamma \cdot d(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}) - s(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}), \quad (2)$$

whereas d is the normalized mean distance between the configurations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$ and s is the idle time estimated by the scheduler for these points. It is important to note that q is not fixed. The challenge is the high cost of optimizing this criterion. But a first experiment has shown that already an roughly approximated optimization leads to well suited sets of configurations to lead the search.

2 The asynchronous approach

Another method reduces the idling times to zero but comes at a cost. In the asynchronous approach each core runs independently and starts a new iteration directly once a previous one has finished. Hence idling does not occur anymore but the overhead for model building and candidate proposal increases. As suggested in [4] to propose a new point for one process it has to take into account the configurations ($\boldsymbol{x}_{busy}$) that are still being evaluated in the other processes. Indeed their outcome is unknown but their posterior distribution can be estimated by the regression model. This leads to an Expected Expected improvement (EEI) [4] which can be calculated using monte carlo simulation. For the $\boldsymbol{x}_{busy}$ configurations we draw the outcomes out of the posterior distribution. For each draw the EI [8] is calculated. The EEI is then just the arithmetic mean over all expected improvements. For each monte carlo iteration a separate regression model has to be calculated. Hence the overhead increases drastically as many iterations (~ 100) are necessary to estimate the EEI sufficiently exact. Our studies indicated that the increased overhead is only negligible for long black-box runtimes.

3 Further Collaborations

In collaboration with the B2 project and Pascal Libuschewski we use heavily parallelized multi criteria MBO [5] to optimize the parameters of a simulated GPU to solve various benchmarks. Depending on the task the GPU has to solve, different configurations are necessary to optimize the run-time and the energy consumption, making it an multi criteria optimization task. Besides the obvious solution to always take *the best* GPU to solve a task fast it can be observed that for some benchmarks there is no performance increase for a higher number of GPU cores. So to keep hardware and energy costs low it is of interest to know the best configuration under given constraints. The result of

a multi criteria optimization is a pareto front giving a set of points, of which each is optimal, meaning that further increasing one target criterion will decrease at least one other criterion. The user can then decide which trade off he is willing to take.

I participate actively in the development of the R-Package **mlr** [1] and **mlrMBO** amongst other contributions.

References

[1] Bernd Bischl et al. "mlr: Machine Learning in R". In: *Journal of Machine Learning Research* 17.170 (2016), pp. 1–5. URL: http://jmlr.org/papers/v17/15-066.html (visited on 11/07/2016).

[2] Bernd Bischl et al. "MOI-MBO: Multiobjective Infill for Parallel Model-Based Optimization". In: *Learning and Intelligent Optimization Conference*. Florida, 2014. DOI: 10.1007/978-3-319-09584-4_17.

[3] Alexander IJ Forrester and Andy J Keane. "Recent advances in surrogate-based optimization". In: *Progress in Aerospace Sciences* 45.1 (2009), pp. 50–79.

[4] David Ginsbourger, Janis Janusevskis, and Rodolphe Le Riche. "Dealing with asynchronicity in parallel Gaussian Process based global optimization". In: *4th International Conference of the ERCIM WG on computing & statistics (ERCIM'11)*. 2011.

[5] Daniel Horn et al. "Model-based multi-objective optimization: taxonomy, multi-point proposal, toolbox and benchmark". In: *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2015, pp. 64–78. URL: http://link.springer.com/chapter/10.1007/978-3-319-15934-8_5 (visited on 11/07/2016).

[6] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Sequential model-based optimization for general algorithm configuration". In: *Learning and Intelligent Optimization*. Springer, 2011, pp. 507–523.

[7] Janis Janusevskis et al. "Expected improvements for the asynchronous parallel global optimization of expensive functions: Potentials and challenges". In: *Learning and Intelligent Optimization*. Springer, 2012, pp. 413–418.

[8] Donald R. Jones, Matthias Schonlau, and William J. Welch. "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4 (1998), pp. 455–492. DOI: 10.1023/A:1008306431147.

[9] Jakob Richter et al. "Faster Model-Based Optimization through Resource-Aware Scheduling Strategies". In: *Proceedings of the 10th Learning and Intelligent OptimizatioN Conference (LION'10)*. Italy, (accepted for publication).

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A4

Resource efficient and distributed platforms for integrative data analysis

Michael ten Hompel Olaf Spinczyk
Christian Wietfeld

Energy Adaptation for Ubuquitous Computing Systems in Practice

Markus Buschhoff
Department of Computer Science 12
Technische Universität Dortmund
markus.buschhoff@tu-dortmund.de

Ubiquitous appliances often suffer from strict resource constraints: In a flexible, wireless environment, these devices are required to harvest energy from their environment and to adapt their energy consumption to environmental conditions. In a practical approach to ubiquitous indoor systems, the *Solar Doorplate* project group was initiated within the project A4 context to identify and encounter the problems that typically arise in these system environments. Here, components like the Kratos Operating System [4] and the MIMOSA measurement platform were committed to the project group.

1 Solar Doorplate

The *Solar Doorplate* project group was initiated as a student project according to the guidelines of the TU Dormund for project groups. It consisted of 12 students and was led by members of the A1 project (Alexander Lochmann), A4 project (Markus Buschhoff), and the corresponding project lead Olaf Spinczyk.

The target of the project group was to design a prototype for wireless, radio updated door plate / name tag, that satisfies its energy demands solely by harvesting from environmental indoor light.

The scientific aim of initiating the Solar Dooplate project group was to research the usability of our formerly presented operating system and measurement approaches in a practical scenario. The Kratos operating system, the Mimosa measurement platform [1] and the Phynode platform were contributed to the project group and were constantly

updated and adapted to suit their needs during the project phase. A brief summary of the major contributions to the project group and the results of the project group are presented in the following.

1.1 Tickless Scheduling

One contribution to the project group was the operating system Kratos, which was carefully adapted to suit the needs of the underlying hardware architecture, the PhyNode platform constructed in the A4 project. For the purpose of reducing the idle power consumption of the platform, an event-triggered, tickless scheduling system was implemented. This scheduler reduces the need to wake up the system from sleep modes for preemtion checks, if no concurring task is in the running queue. Instead, a hardware timer is preset to wake up the system to the next pre-calculated system event. This implies, that non-deterministic system events can wake up the system as well, e.g. by interrupt requests, which is a valid prerequisite for the given platform.

1.2 Interrupt Stack Separation

The tight memory limits of the used hardware platform enforce a static memory allocation and by that a predictable stack utilization. This imposed a large memory overhead, as each thread needed a stack space reserve for unpredictable events (interrupt handling). To encounter this waste of memory, a separate stack space for interrupt handling was implemented into Kratos. All interrupt activities are handled after switching to this separated stack (which uses all remaining memory), while the application threads can use a bounded stack area. Assuming a predictable stack behavior (e.g., no recursions, no stack utilization within indeterministic loops), the worst-case stack consumption of each thread can be analyzed at compile time by code analysis tools.

1.3 Project Group Results

With the given system environment, the project group was able to implement the requested Solar Doorplate application and to fulfill the minimum targets and several optional targets. The project group could show that the energetic behavior of the system was predictable, that the proposed energy modeling techniques are practically applicable and that they were able to successfully utilize Kratos and Mimosa for this purpose. By that, the system behavior could be shaped to fit into the energetic constraints imposed by harvesting from artificial light and communicating over wireless links.

Table 1 shows the energy bilance of a solar doorplate device per hour, categorized by energy consumption and harvest. The idle current is the base current of the device that

	Power [µW]	Duration [s]	Count	Energy per hour [mJ]
Consumption				
Idle	-18	3600	1	-64.8
Upd. Display	-22000	2.546	2	-112.02
Radio send		127	2	-1.32
Redio recv(eWOR)	-27000	4	2	-216
Light sens.	-4783	0.021	2	-0.2
Harvest				
Solar cell	200	3600	1	720
Sum				**325.75**

Table 1: Energy bilance of a solar doorplate with ePaper display, as shown in the final report of the Project Group 595.

is consumed at any time. It is remarkable, that the idle power was lowered from 960 µW to 18 µW by a careful system analysis and a diversity of optimizations, like the tickless scheduler shown above.

2 Sevice-Level Based Scheduling Strategies

While common CPU scheduling strategies demand to utilize a CPU as much as possible to avoid CPU idling and to reduce latencies, energy aware approaches might consider different, as the energy consumption of a system is not only determined by the CPU, but also by the utilization of peripheral devices. For wireless sensor networks and a subset of ubiquitous systems, energy consumption even is dominated by periphery. Especially for wireless network connections, indeterministic conditions (network status, distance to communication partner in mobile systems, etc.) have a strong influence on the overall energy consumption. Here we find new scheduling demands that can be considered as a trade-off between the quality of an application's service (e.g. its latency) and its energy consumption.

In an approach to let an application developer decide about these trade-offs in a wireless sensor network (WSN) operating system like Kratos, we analyzed an approach that defines discrete *service levels*. Under the valid assumption that a typical WSN task is cyclically run, has a close-to-constant energy consumption during each run, and a task run can be delayed within given boundaries, we can define a service level as a subset of interdependent tasks to run and their respective delays before start. A set of service levels can be structured as the states of a finite state machine. The transitions of this state machine are formed by energy events, e.g., the accumulator state, the harvest income or the most recent energy consumption.

This approach was evaluated within a master thesis [2], and we showed that service level automaton driven scheduling can be efficiently implemented in an operating system. A carefully modeled service level automaton can by that reduce the application's service-quality and energy consumption in low-energy situations. We assume that the system stability and service availability can be raised by this form of adaptation, but we still have to prove this assumption in our future work with actual applications and real hardware.

3 Scheduling for Non-Linear Battery Effects in Lithium Thin-Film Batteries

Non-linear capacitor effects were already researched within a former master thesis [3]. The positive results led to follow-up research for other energy sources, namely lithium based thin-film batteries, as they are a popular choice in recent technologies because of their high energy density and physical flexibility.

We could show that using energy aware offline scheduling has less an effect for batteries than it has for the formerly evaluated super-capacitors [5]. The low self-discharge of lithium based batteries make them generally more efficient, so that the possible efficiency gain becomes smaller. Anyhow, considering the still present non-linear effects still makes sense, especially regarding further conditions like the battery age and temperature.

References

[1] Markus Buschhoff, Christian Günter, and Olaf Spinczyk. *MIMOSA, a Highly Sensitive and Accurate Power Measurement Technique for Low-Power Systems*, pages 139–151. Springer International Publishing, Cham, 2014.

[2] Sebastian Engels. Energiegewahres Service-Level-Management in eingebetteten Betriebssystemen. Masterarbeit, Universität Dortmund, Germany, February 2016.

[3] Michael Hesse. Berücksichtigung nicht-linearer Energiequelleneffekte im eingebetteten Betriebssystem Kratos. Diplomarbeit, Universität Dortmund, Germany, September 2015.

[4] Katharina Morik and Wolfgang Rhode (Editors). Technical report for collaborative research center sfb 876 - graduate school. Technical Report 10, TU Dortmund University, December 2014.

[5] David Schmidt. Modellierung von Lithium-Dünnfilmbatterien für die Ablaufplanung im Betriebssystem Kratos. Masterarbeit, Universität Dortmund, Germany, September 2016.

Client-Based Data Rate Estimation in LTE Mobile Networks

Robert Falkenberg
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
robert.falkenberg@tu-dortmund.de

Clients in LTE mobile networks lack of the knowledge about the cell's current load. This greatly impairs clients to reliably forecast their expected data rate for an upcoming transmission. In our work we performed statistical analysis of LTE control channels to uncover this missing information.

1 Introduction

Mobile networks become increasingly popular for connecting cars to exchange telemetry and other assistant data. Besides small embedded devices progressively head for an Internet connection by mobile networks forming the Internet of Things (IoT). In both cases the need for an estimation of the quality of the radio link in terms of expected data rate arises. While cars may use this information for the choice of the momentarily best link in case of multiple available connections (LTE, local Wi-Fi, car-2-car), embedded devices might delay their transmission until the radio conditions improve and the transfer can be performed more energy-efficiently.

In LTE, like other mobile networks as well, the achievable data rate of the link highly depends on the radio channel conditions and the overall cell load induced by other participants. As already described in [3] the radio conditions and their influence on the transfer duration can be estimated at client-side from passive indicators provided by the LTE modem.

Unfortunately LTE provides no indicator to the client about the current load situation of the cell. As Fig. 1 points out, the available radio resources are being shared among the active users. Under congested conditions the assigned resource fraction to a single user drops out significantly smaller than in an unloaded case. A spectrum analyzer or a User Equipment (UE) can only detect the overall activity of the cell derived from the signal power on the channel. But the current LTE system design gives no possibility to break

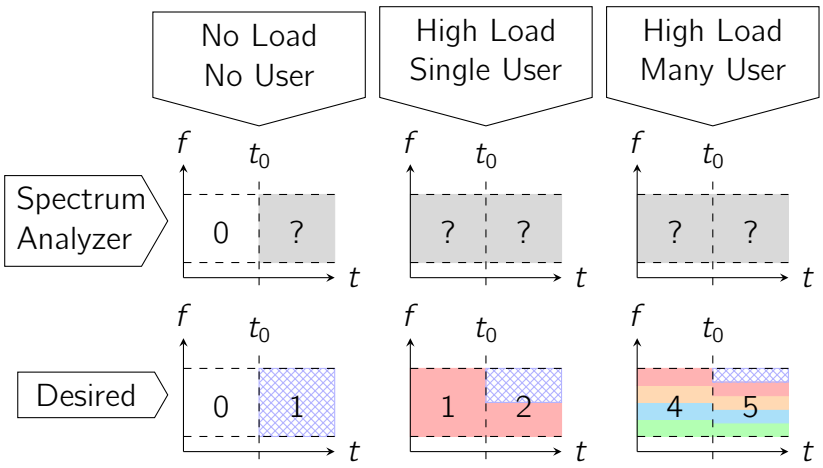


Figure 1: Existing and Desired View of the LTE Channel Utilization at Client-Side

this activity down into single users as shown in the lower part of the figure. Hence, any channel-based connectivity estimation is affected by uncertainties in congested scenarios. To close this gap, we managed to decode control information of all participants in a cell and uncovered their individual radio resource allocations [2]. This allows a rating of the total cell load and a more reliable forecasting of the data rate.

2 Approach and Implementation

In LTE the base station (evolved NodeB (eNodeB)) manages all radio resources within its covered area and assigns them to the particular UEs according to the internal scheduling strategy. Those allocations are transmitted in a dedicated section, the Physical Downlink Control Channel (PDCCH), at the beginning of every subframe once per millisecond. They are not encrypted but the 16-bit checksum of each struct is scrambled (XOR) with the physical-layer address (Radio Network Temporary Identifier (RNTI)) of the addressed UE. Consequently a device must know a valid RNTI to validate or discard a decoded allocation struct. Except for some reserved RNTIs for broadcast information, a normal UE has only knowledge of its own RNTI as it is assigned by the eNodeB over an encrypted channel when establishing a connection.

However, statistical analysis of blind decoded data still enables an identification of valid RNTIs: First, all possible slots within the PDCCH are decoded and the checksum is calculated assuming the resulting checksum to be an RNTI. Second, this RNTI is inserted into a histogram of recently appeared RNTIs as shown in Fig. 2. If the RNTI occurrences exceed a certain threshold, this identifier is assumed as valid. This approach works due to the fact that a UE usually stays active over multiple subframes maintaining the same RNTI during that transmission. Due to space restrictions the calculations of appropriate threshold values and the miss-probability are being omitted and can be read in [2].

A slight drawback of this approach is a small detection latency of joining transmissions

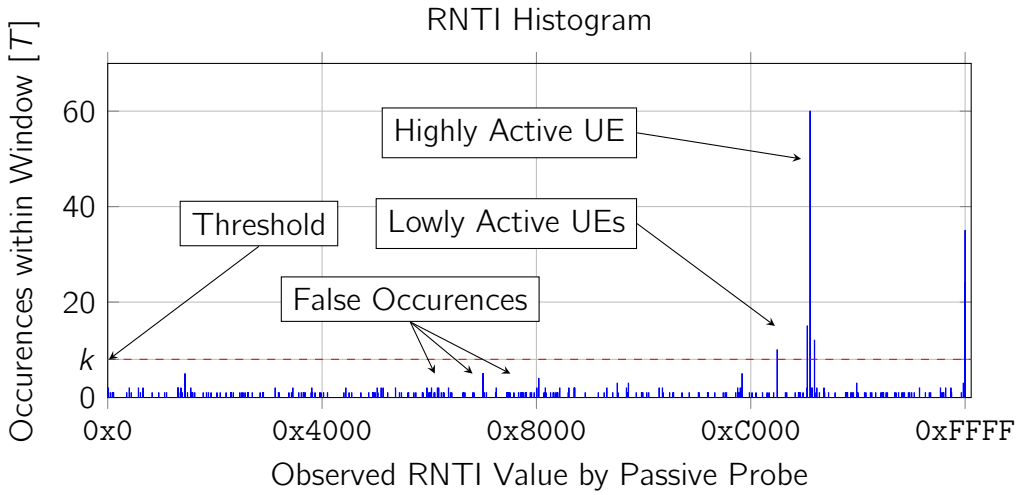


Figure 2: Snapshot of an RNTI Histogram

as the occurrence of the particular RNTI must first exceed the threshold value k. In case of a joining transmission in one direction the latency is at least k milliseconds if the corresponding UE gets resources assigned in k consecutive subframes. Consequently, if a UE keeps his activity below that threshold within the histogram depth (e.g., 500 ms), its RNTI will not be considered. On the contrary, its activity will consume only few resources of the eNodeB and therefore decrease the impact on the total estimation accuracy.

We implemented our approach into OpenAirInterface [1], an open source software turning a general purpose computer into an LTE UE. The software is capable of attaching to any LTE network over the air and works as a passive probe which continuously detects active users (see Fig. 3).

Figure 3: Passive Probe

The filtering process has only very low complexity and a moderate memory consumption. Most of the computational complexity is caused – in contrast to a normal UE– by decoding the entire PDCCH with all possible contents. Compared to the effort required to decode actual user payload, which is usually the lion's share during a transmission decoding, this amount can still be neglected. Besides, maintaining the histogram simply consists of enqueueing and dequeueing an integer into a circular buffer, one addition and subtraction and a comparation with the threshold value per incoming RNTI. Our implementation requires 256 kB of memory for the occurence counter of any possible RNTI and additional 164 kB for the circular buffer of recent RNTIs over 500 ms.

Assuming a Round-Robin scheduling of radio resources over all active attendees, we used the detected number of active users N to perform an estimation of the expected share in resources for a joining user according to the momentary load situation. The estimated data rate r_{UE} is

$$r_{UE} = \frac{1}{N+1} \cdot r_{\max}\left(\mathrm{MCS}_{UE}\right), \tag{1}$$

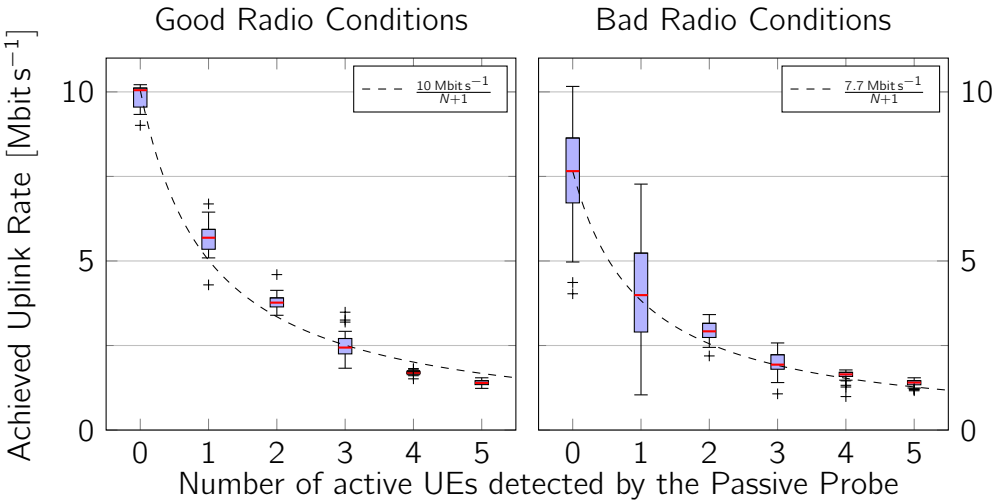


Figure 4: Data Rate vs. Detected Number of Active Participants by the Passive Probe

where $r_{\mathbf{max}}(\mathrm{MCS}_{UE})$ is the maximum achievable data rate by the UE at the current radio conditions with the chosen Modulation and Coding Scheme (MCS).

3 Measurements and Results

In order to evaluate our approach, we performed measurements with varying numbers of active users where N is being detected by the passive probe and is used to forecast r_{UE} for an additional joining UE. Subsequently the actual achieved data rate by the joint UE is being logged, too. Fig. 4 shows the results of the measurements: The dashed line corresponds to the theoretical data rate according to Eq. 1 at the current radio conditions of the joining UE. In addition, the actual data rate is visualized as box plots showing a good match with the estimation.

4 Conclusion and Further Research

The results are very promising allowing a more reliable data rate forecasting at client-side. However, this approach is limited to a Round-Robin scheduling without any prioritization. Hence, our further research focuses on improving the general applicability in real deployed mobile networks. This includes a detection and consideration of prioritized users which consume a larger fraction of the available spectrum.

References

[1] EURECOM Mobile Communications Department. OpenAir5G LAB. `http://openairinterface.eurecom.fr`, October 2016.

[2] Robert Falkenberg, Christoph Ide, and Christian Wietfeld. Client-based control channel analysis for connectivity estimation in lte networks. In *IEEE Vehicular Technology Conference (VTC-Fall)*, Montréal, Canada, September 2016.

[3] Christoph Ide, Robert Falkenberg, Dennis Kaulbars, and Christian Wietfeld. Empirical analysis of the impact of lte downlink channel indicators on the uplink connectivity. In *IEEE Vehicular Technology Conference (VTC-Spring)*, Nanjing, China, May 2016.

Photovoltaic Energy Harvesting Characteristics under Ultra-Low Light

Mojtaba Masoudinejad
Lehrstuhl für Förder- und Lagerwesen
Technische Universität Dortmund
mojtaba.masoudinejad@tu-dortmund.de

PhyNode as an intelligent cyber-physical system operating in the PhyNet-Lab Wireless Sensor Network (WSN) is empowered by Photovoltaic (PV) energy harvesting. To understand the harvesting behavior of thy PhyNode, analysis for different lighting type and intensities are presented here. Based on the collected data, different analysis are provided to optimize the harvesting.

1 Introduction

PhyNetLab [5] made of 350 PhyNode [4] agents in the operation field and 6 access points in the upper tier (Fig. 1) is a WSN for the warehousing application.

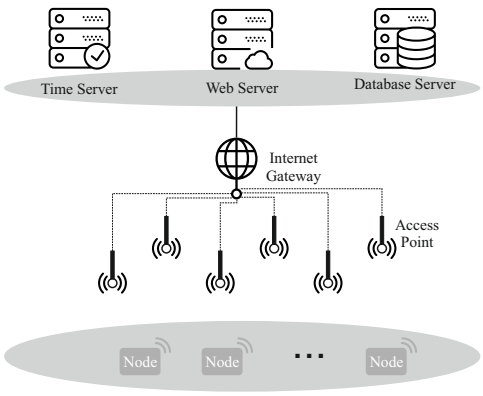


Figure 1: The overall structure of the PhyNetLab

PhyNode is an electronic board with processing, communication, ambient sensing and user interaction capabilities mounted on a typical materials handling bin. PhyNode has to be energy neutral without the need for maintenance of the energy storage components or recharging. To fulfill the energy neutrality, energy harvesting has to be integrated within the PhyNode. Among possible harvesting solutions, [1] showed that in spite of very low lighting situation in a warehouse, PV harvesting is still the best possible solution. However, the extreme low lighting situation in addition to the lack of sunlight brings the indoor PV harvesting into its limit. To conquer the crucial energy situation that the PhyNode deals during its lifetime, its hardware and software should be specially configured to collect the maximum available energy while reduces the demand.

PhyNode's energy module is made of three parts. In addition to the PV harvester it has a Texas Instrument BQ25505 Maximum Power Point Tracking (MPPT) chip and a storage element. For proper selection of the MPPT technique and its appropriate tuning, analysis of the PV harvester is necessary. It should be tested under standard lighting condition. Therefore, the measurement platform presented in [2], [3] is used to collect the PV cells' behavior under different indoor lighting situations. Among multiple light sources available, tests under halogen and florescent lighting are considered here.

2 Measurement

The first and simplest analysis of each PV harvester is the measurement of its V-I curve. This can be done using a Source Measurement Unit (SMU). It forces different voltages at the harvester and measures the current for each working point. Using the V-I curve, the power for each voltage (or current) can be calculated as well. Subsequently, the MPP can be found using the measured V-P curve. Since both these curves are dependent on the lighting intensity, light spectrometry is used to define the exact light intensity. A typical data-set of the measurement would be as presented in the Fig. 2. The overall shape of the V-I and V-P curves are the same as the normal outdoor PV cells though the values are much smaller.

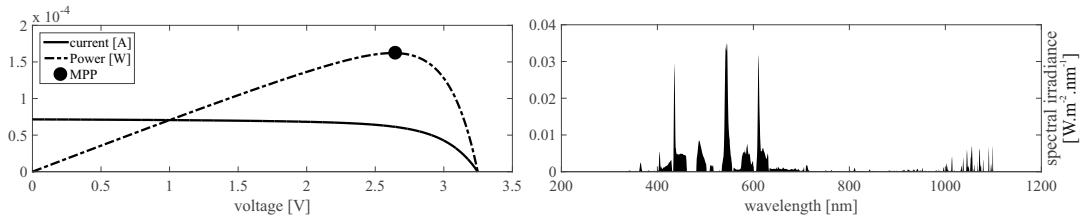


Figure 2: Left: VI/VP curves, right: light spectrum of the florescent lighting

Not only the indoor area has much lower light intensity compared to the outdoor, but also is not constant all over the space. Moreover, warehouses are mostly designed in a vertical

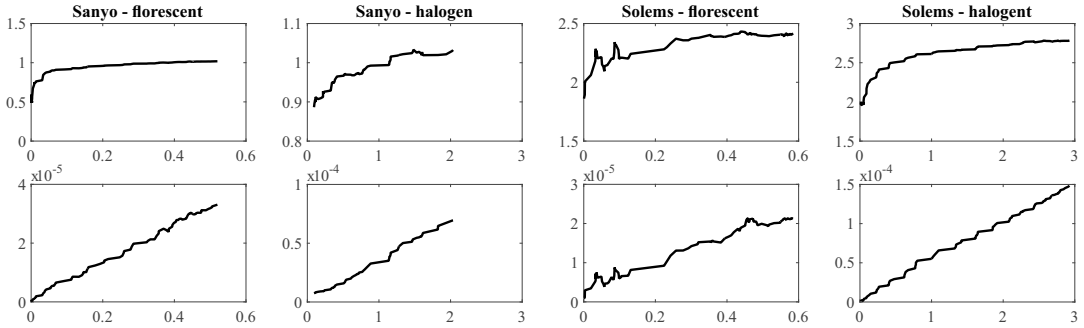


Figure 3: Irradiance $[W/m^2]$ vs. top: MPP voltage $[V]$, bottom: MPP current $[A]$

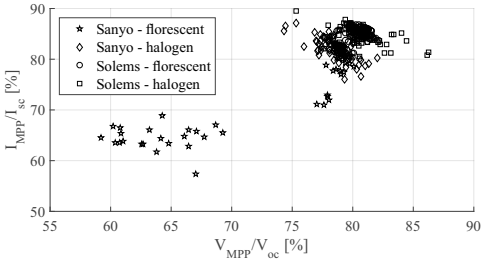


Figure 4: MPP voltage/current ratios to the system values for different combinations

manner and the distance between bins and light source can change largely according to the height, leading to large intensity deviations. Therefore, effect of the irradiance change should be analyzed properly. To collect the necessary data, four combination of hardware (two PV harvesters under two different light sources) are tested in different light intensities measuring the V-I and V-P curves in addition to the spectrometry. Using this data, different analysis can be done on the harvester's behavior under different intensities. At first, change in the MPP voltage and current can be seen in Fig. 3.

As can be seen, the MPP voltage increases exponentially according to the irradiance while the MPP current is changing linearly. Since fractional MPPT algorithms use a fixed portion of the measured parameters (V_{oc} or I_{sc}), analyzing their relations would provide an evaluation factor. Using the measured datasets, distribution of the two main relations (V_{Mpp}/V_{oc} and I_{Mpp}/I_{sc}) for all four test combinations is shown in the Fig. 4.

Generally, fractional algorithms are simple to implement with minimum energy requirements, though the data presented in the Fig. 4 shows lack of optimum efficiency when a fixed fraction constant is used. In case a method be able to fine-tune these ratios according to the irradiance, it would improve the performance. To do so, analysis of ambient factors and their effect on the ratio is necessary. As example, changes of the ratio for two cases are presented in Fig. 5.

As can be seen, there is an exponential relation between increase in the irradiance and the

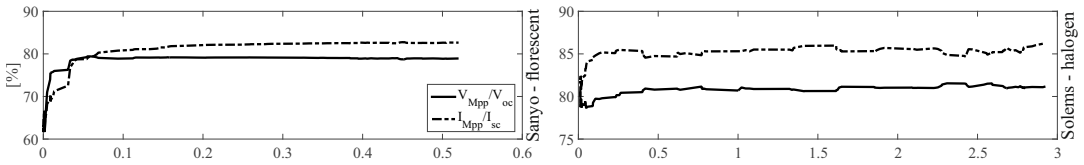


Figure 5: MPP voltage/current ratios change according to the irradiance $[W/m^2]$

voltage/current fraction constant. This can be used to improve the fractional MPPT algorithm using measurement of the irradiance and map it into the related fractional constant or even directly to the optimum voltage or current.

3 Future works

Using characteristics of the PV harvesters under different lighting situation, MPPT systems should be optimized to fit the low light application behaviors.

References

[1] A. Kamagaew, T. Kirks, and M. Ten Hompel. Energy potential detection for autarkic Smart Object design in facility logistics. In *IEEE International Conference on Control System, Computing and Engineering, ICCSCE*, pages 285–290. IEEE, 2011.

[2] M. Masoudinejad, J. Emmerich, D. Kossmann, A. Riesner, M. Roidl, and M. ten Hompel. Development of a measurement platform for indoor photovoltaic energy harvesting in materials handling applications. In *6th International Renewable Energy Congress*, 2015.

[3] M. Masoudinejad, J. Emmerich, D. Kossmann, A. Riesner, M. Roidl, and M. ten Hompel. A measurement platform for photovoltaic performance analysis in environments with ultra-low energy harvesting potential. *Sustainable Cities and Society*, 25:74–81, 2015.

[4] A.K. Venkatapathy, A. Riesner, M. Roidl, J. Emmerich, and M. ten Hompel. PhyNode : An intelligent , cyber-physical system with energy neutral operation for PhyNetLab. In *Proceedings of Smart SysTech; European Conference on Smart Objects, Systems and Technologies;*, pages 1–8. VDE-Verl, 2015.

[5] A.K.R. Venkatapathy, M. Roidl, A. Riesner, J. Emmerich, and M. ten Hompel. PhyNetLab: Architecture design of ultra-low power Wireless Sensor Network testbed. In *IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks*, pages 1–6. IEEE, 2015.

Approximate Computing on Embedded Systems
-
Big Applications for Small Devices

Olaf Neugebauer
Lehrstuhl für Technische Informatik und Eingebettete Systeme
Technische Universität Dortmund
olaf.neugebauer@tu-dortmund.de

Approximate computing has the potential to increase the performance of existing embedded application or to enable new applications previously not feasible on these systems. These benefits do not come for free, downsides need to be analyzed carefully and considered during the introduction of approximate computing. In this report we discuss the benefits and downsides of approximate computing and present strategies to cope with this issues.

1 On Approximate Computing

Growing demands for processing power, especially in high-performance embedded systems application domains such as autonomous cars and media signal processing, increase the pressure on hardware and software developers to create well performing systems which are also power and energy efficient.

Traditionally, hardware manufacturers solved the performance challenge by increasing the frequencies of processors and memories, whereas the reduction of power and energy consumption was mostly achieved by shrinking semiconductor structure sizes. More recently, multiple, potentially heterogeneous, cores are being combined onto a single chip to tackle the needs of versatile performance requirements. However, frequency increase is limited by thermal constraints.

One increasingly popular method to mitigate the rising impact of these performance and power walls is to use *approximate computations* which are no longer bit-perfect to satisfy the quality of service (QoS) requirements of a given application. Here, QoS describes the calculated output quality of an application. Approximate computing techniques can also be applied to recognition, mining and machine learning applications.

Approximate computing promises to be a worthwhile approach to overcome or at least shift the influence of performance and power walls. In recent years, related techniques have seen significant adoption [1, 2, 4, 9, 10]. Using principles from the approximate computing paradigm, we were able to execute the PAMONO pipeline on an embedded system [7].

2 On Quality Metrics

When analyzing the effectiveness and efficiency of different approximate computing approaches, it becomes obvious that it is required to ensure that approximations stay within an acceptable range, i.e., the results have to meet a certain QoS. As a consequence, suitable quality metrics are required to evaluate the usefulness of a given approximation technique. The selection of such quality metrics is complex, error-prone, and highly application dependent. Especially if metrics lead to contrary or insufficient conclusions, it is necessary to take multiple metrics into account during the development and application of approximation techniques.

(a) Original (b) MSE = 460, SSIM = 0.95 (c) MSE = 459, SSIM = 0.50

Figure 1: Quality Comparison according to MSE and SSIM.

During the development and analysis of applications using approximate computing, a selection of proper quality metrics is of high importance. Figure 1 shows three versions of the same picture with Figure 1(a) being the reference picture. In Figure 1(b) the image has been brightened, and in Figure 1(c) the image is distorted with Gaussian blur. Applying a signal fidelity metric like mean square error (MSE) results in almost identical values for both pictures. Thus, judging which picture is *better* is complicated. In contrast to signal fidelity metrics, perceptual visual quality metrics (PVQM) consider the human

as receiver of the results. The resulting structural similarity index metric (SSIM) values indicate that Figure 1(b) is much *better* than Figure 1(c). Not only for humans, also for pattern recognition applications these results are relevant. For example, an edge detection algorithm would usually perform better on images with sharp edges such as the image from Figure 1(b).

This example highlights the importance of quality metrics and their selection for approximate computing. Selecting proper metrics is complex, error-prone and usually requires deep application knowledge.

3 Quality Assessment Framework

As the previous section highlights, a holistic software approach is required to assist developers during the design and assessment of approximation techniques. In this section we outline the requirements of such a software which we are currently developing. The main features can be summarized as 1) a flexible open source framework, 2) allowing analysis of applied approximation techniques 3) with respect to run time, energy consumption, output quality and 4) user defined objectives like file size. Especially for the output quality metrics, several metrics should be included and applied in parallel.

The open source approach allows others to use this framework to asses their approximation techniques and ease the comparison of different techniques. Thus, the framework integrates several state of the art metrics like SSIM or PSNR. Traditional evaluation strategies mostly focus on run time, quality and sometimes energy. We think, that for example the file size of the output is also an important factor especially in the embedded systems domain. Thus, in our opinion a holistic view on performance is necessary to evaluate the approximation techniques, therefore, we allow the users to add new objectives and metrics which are relevant for a specific scenario.

As a target platform we consider the Odroid XU-3 which was used in our previous work. This platform allows us to measure the energy consumption of the key components. Beside the onboard energy measurement facilities we plan to add new external sensors to measure the complete energy consumption of the prototype board.

4 Future Work

We would like to combine the quality assessment framework with our PA4RES framework [6] and especially with PICO [5]. This enables us to evaluate the benefits of approximation techniques during the parallelization process targeting embedded systems. Several papers show the potential for relaxes synchronization [8] or best-effort parallel execution [3].

We like to investigate on how existing and new techniques can be used in the embedded system context to gain performance but still meet a certain output quality. An interesting application to study the benefits is the PAMONO pipeline developed in part project B2, where several processing steps are executed on a stream of pictures. Here, each step could be targeted by approximation techniques and then their influence on the final results could be analyzed.

References

[1] Jie Han and Michael Orshansky. Approximate Computing: An Emerging Paradigm For Energy-Efficient Design. In *Proc. of ETS*, 2013.

[2] Logan Kugler. Is "Good Enough" Computing Good Enough? *Communication of the ACM*, 58(5), 2015.

[3] Jiayuan Meng, Srimat Chakradhar, and Anand Raghunathan. Best-effort parallel execution framework for recognition and mining applications. *IPDPS 2009 - Proc. 2009 IEEE Int. Parallel Distrib. Process. Symp.*, 2009.

[4] Sparsh Mittal. A survey of techniques for approximate computing. *ACM Comput. Surv.*, 48(4), 2016.

[5] Olaf Neugebauer, Michael Engel, and Peter Marwedel. Multi-objective aware communication optimization for resource-restricted embedded systems. In *Proceedings of Architecture of Computing Systems. Proceedings, ARCS*, 2015.

[6] Olaf Neugebauer, Michael Engel, and Peter Marwedel. A parallelization approach for resource-restricted embedded heterogeneous MPSoCs inspired by OpenMP. *Journal of Systems and Software*, pages –, 2016.

[7] Olaf Neugebauer, Pascal Libuschewski, Michael Engel, Heinrich Müller, and Peter Marwedel. Plasmon-based virus detection on heterogeneous embedded systems. In *Proceesings of Workshop on Software & Compilers for Embedded Systems, SCOPES*, 2015.

[8] Lakshminarayanan Renganarayana, Vijayalakshmi Srinivasan, Ravi Nair, and Daniel Prener. Programming with relaxed synchronization. In *RACES*, page 41, New York, New York, USA, 2012. ACM Press.

[9] Swagath Venkataramani, Srimat T Chakradhar, Kaushik Roy, and Anand Raghunathan. Computing approximately, and efficiently. In *Proc. of DATE*, 2015.

[10] Qiang Xu, Todd Mytkowicz, and Nam Sung Kim. Approximate Computing: A Survey. *IEEE Design Test*, 33(1), 2016.

Wireless Positioning based on IEEE 802.15.4a Ultra-Wideband in Multi-User Environments

Janis Tiemann
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
janis.tiemann@tu-dortmund.de

Wireless positioning based on IEEE 802.15.4a gained attention for precise localization recently. However, for large scale logistic applications many aspects are not considered. This work aims to focus on optimizing the positioning accuracy, multi-user scalability and energy efficiency of such systems. Due to the simple nature of the most commonly used two-way ranging based systems, a significant amount of messages is required to obtain a position in a set of infrastructure anchors. This work proposes, implements and analyzes time-difference of arrival based positioning to improve the system capabilities.

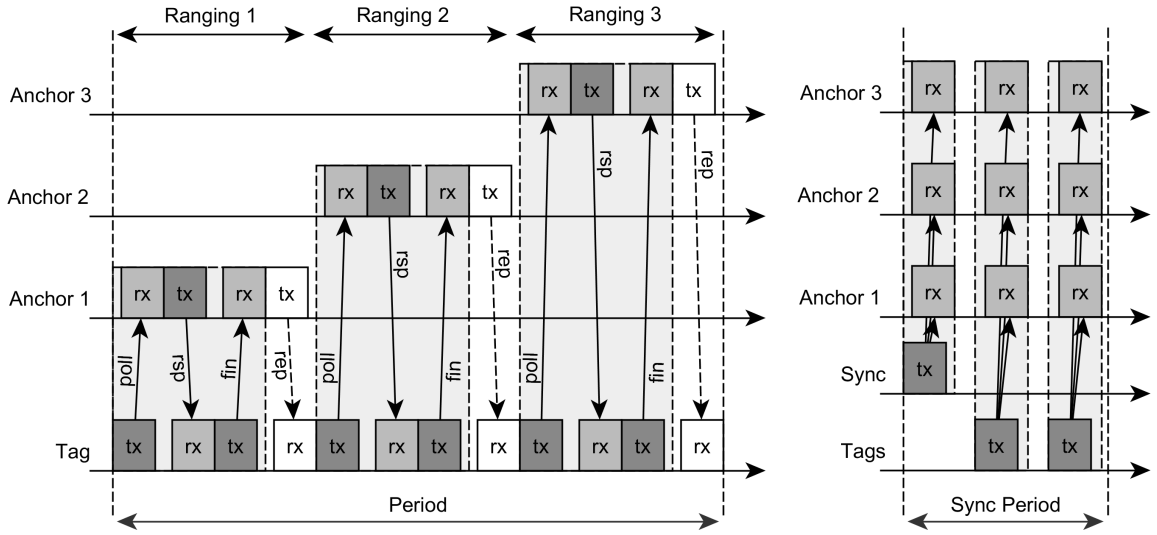
1 Introduction

Current developments in ultra-wideband (UWB) chips compatible to IEEE 802.15.4a challenged a new range of applications for wireless positioning systems. Previous work [1] enabled indoor unmanned aerial vehicle (UAV) flights by the emulation of commonly used global navigation satellite system (GNSS) receivers. In other work [2], the capabilities of UWB based positioning for precision parking of inductive charging electric vehicles is analyzed. However, the abovementioned work is based on the use of symmetric double-sided two-way ranging (SDS-TWR). Though, yielding accurate ranging and keeping system complexity low, two-way ranging based systems have significant limitations when it comes to multi-user capabilities.

Therefore, this work proposes the use of time-difference of arrival (TDOA) based positioning to enable an infrastructure based system with high multi-user capabilities while maintaining a system accuracy comparable to SDS-TWR.

2 Approach

As depicted in Fig. 1, TDOA based positioning as proposed in this work occupies significantly less time on the channel then traditional SDS-TWR based positioning. A detailed multi-user capability analysis is given in [3].



(a) Symmetric Double-Sided Two-Way Ranging (b) Time Difference of Arrival

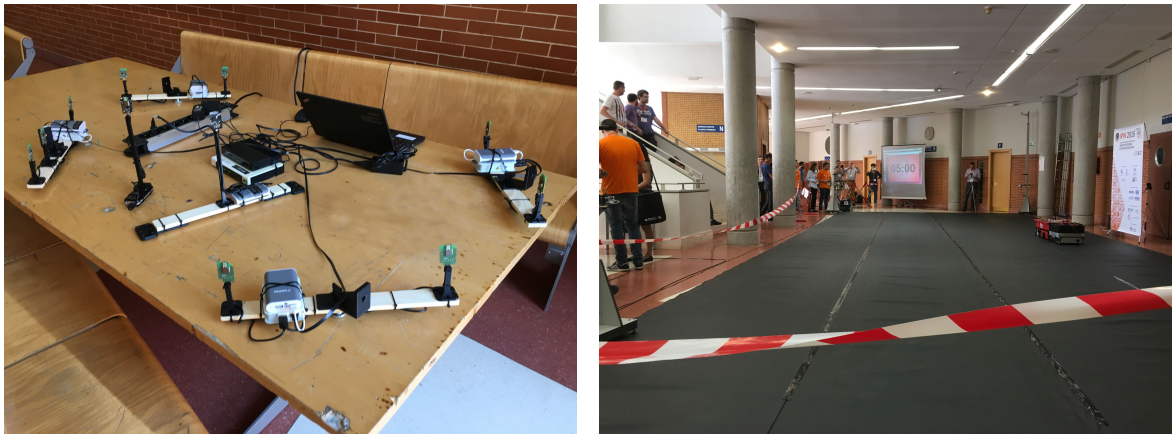
Figure 1: Schematic comparision of traditional symmetric double-sided two-way ranging (SDS-TWR) with the proposed time-difference of arrival (TDOA) approach.

The multi-user capability of TDOA based systems is achieved on the expense of system complexity. As depicted by Fig. 1(b), clock synchronization for the anchor nodes is necessary as individual anchor clocks drift significantly over time. In the proposed system wireless clock synchronization over a precise periodic synchronization beacon is applied. In order to achieve accurate positioning results, clock drift extrapolation, outlier detection, static offset calibration and constrained extended Kalman filtering is required.

The key components of the positioning system and the raw positioning data are provided open-source alongside this work for the sake of comparability. Therefore, the methods and results achieved in this work are transparent to a broad audience of research and industry.

3 Experimental and Competitive Accuracy Analysis

In order to comparably analyze the systems capabilites, we participated in Track 4 of the annual EvAAL competition at the Seventh International Conference on Indoor Positioning and Indoor Navigation (IPIN2016) in Alcalá de Henares, Madrid, Spain. During the competition preparation, the system developed was benchmarked and detailed in [4].



(a) On-site preparations (b) Competition area

Figure 2: Pictures of the EvAAL competition preparation and execution.

The goal of the competition depicted in Fig. 2 was to obtain a previously unknown trajectory of a magnetic line-follower robot, provided by the organizers. Each competitor had a setup time of 45 min including node placement, measurement and calibration.

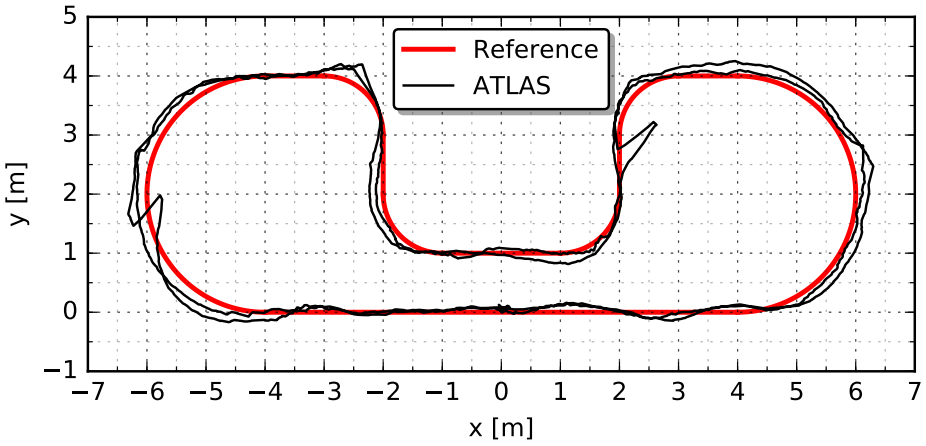


Figure 3: Competition positioning results.

Therefore, not only the theoretical accuracy in the laboratory setup was key to success, but also the practical manageability of the system. The competition positioning results

are depicted in Fig. 3. The 75 % quantile of the horizontal positioning error was below 15 cm. Although the proposed TDOA-based system has more unknowns than comparable TWR-based systems, similar or better accuracies compared to TWR-based systems were achieved.

4 Conclusion and Further Research

In this work, a scalable positioning method for IEEE 802.15.4a based UWB transceivers was proposed and evaluated. Although the system accuracy is sufficient for most applications, multi-path rejection might be improved in future work, to allow for robust operation in crowded environments. The multi-user capabilities are improved significantly compared to TWR. However, a guaranteed quality of service in terms of update frequency is not possible with the current medium access scheme. Therefore, future work may include advanced scheduling to enable positioning services guarantees. Not only multi-user capabilities are improved by the use of TDOA, but also energy efficiency of the mobile nodes. Another aspect that needs to be covered by future work is a detailed energy model for UWB systems and a system topology comparison for different system topologies.

References

[1] J. Tiemann, F. Schweikowski, and C. Wietfeld. Design of an UWB indoor-positioning system for UAV navigation in GNSS-denied environments. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, Oct 2015.

[2] J. Tiemann, J. Pillmann, S. Böcker, and C. Wietfeld. Ultra-wideband aided precision parking for wireless power transfer to electric vehicles in real life scenarios. In *IEEE Vehicular Technology Conference (VTC-Fall)*, Montréal, Canada, Sep 2016. IEEE.

[3] J. Tiemann, F. Eckermann, S. Böcker, and C. Wietfeld. Multi-user interference and wireless clock syncronization in TDOA-based UWB localization. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Alcala de Henares, Madrid, Spain, Oct 2016.

[4] J. Tiemann, F. Eckermann, S. Böcker, and C. Wietfeld. ATLAS - an open-source TDOA-based ultra-wideband localization system. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Alcala de Henares, Madrid, Spain, Oct 2016.

Communicating Industry 4.0 Scenarios: Energy Neutral Industrial WSN

Aswin Karthik Ramachandran Venkatapathy
Lehrstuhl für Förder- und Lagerwesen
Technische Universität Dortmund
aswinkarthik.ramachandran@tu-dortmund.de

This report presents a communication strategy for energy neutral Wireless Sensor Network (WSN) devices that are designed for inventory and order management in an industry 4.0 warehouse. The nodes are energy neutral and have the ability to display product information using an E-INK display and a button for user interaction. Since the devices are modelled to be energy neutral, any existing protocols cannot be used for communicating various scenarios of this use case. This report starts with discussing about the limitations from the use case and the hardware followed by explaining the architecture of the industrial implementation after devising the communication strategy.

1 Introduction

The factor of flexibility is a very important criteria in an industry 4.0 material handling facility [1]. Smart objects or cyber physical systems (CPS) with embedded intelligence that can sense the environment in which they are deployed in and can react to those stimulus, these kinds of intelligent CPS account to a major role in achieving this flexibility [1]. In this report, an order management scenario in a warehouse is considered. The end effect of such a system in a warehouse not only helps in order management but also with the data they produce inventory management as a direct effect by understanding the data. Complex data processing algorithms can then provide other important indicators that are required for providing insight on the supply chain. For such a WSN in a warehouse with energy neutral properties, the communication strategy about medium access and protocol is reported in this article.

2 Hardware platform

The hardware platform of this implementation is classified into two levels of hardware in an IoT architecture. The end nodes, which are the ULP WSN nodes are called *Label* which mimic the printed label etiquette in a warehouse that provide the product information. The hardware in the lowest level of the architecture are the energy neutral nodes which are used in the material handling facility for order management and visualisation of the components that is being handled at the facility. This device is modelled after the Phynode which are also energy neutral ULP WSN node that have the same hardware components [3]. The CC1200 chip is a fully integrated sub-1 GHz band transceiver designed for high-performance at low power and low voltage operation, in cost-effective wireless systems. The access-point is the interface between the application layer and the field nodes. The application layer is where the generation data from the field nodes are monitored, analysed and processed. In this industrial implementation scenario an order management application is connected using an API depending on the management platform.

3 Energy neutral communication protocol

The energy neutral protocol is designed to accommodate large-scale industrial deployment. For the communication to be reliable between the energy neutral node and the access-point with occasional transmission of large data over the network, a multi-channel communication strategy is designed. The colour codes are used in the following illustrations to depict the datagrams of each specific communication case. The use case with the industrial implementation follows the experiences in developing a large scale test bed for ULP WSN with energy neutral properties [2].

3.1 Multi-channel communication protocol

The necessity for a multi-channel communication protocol is due to the fact that the devices are energy neutral and are capable to be awake only for short duration of time. This also means that the periodicity of the device highly depends on the availability of energy. This is one of the main reasons for developing a special communication protocol for this devices in such use cases. Since the devices cannot wake up on a specified time for synchronisation with the access-point or the devices cannot have complex protocol messages to acquire channel access for communicating the different states of the device. Therefore, a clean channel for pairing with the access-point and joining the network and three other dirty channels where the devices wake up periodically, if possible, and communicate their statuses.

3.2 Protocol specification

From the requirements of the application and the initial analysis of energy neutral hardware, a detailed protocol specification is made for reliable communication of the energy neutral nodes. In the initial definition, the protocol consists of four distinct messages for communication. The following are the message types in their chronological order: 1) Pair message, 2) Data message, 3) Status message, 4) Order message. The com-

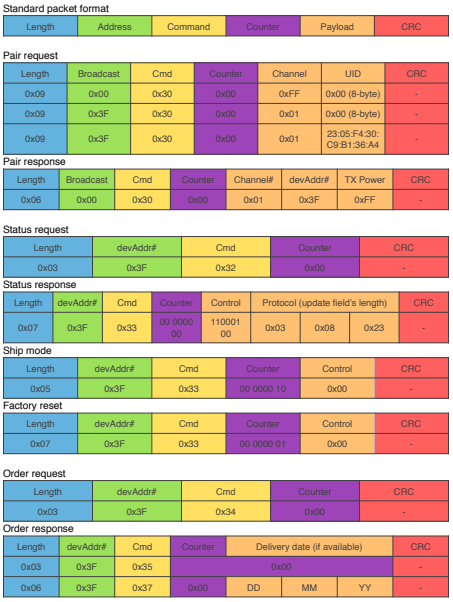


Figure 1: Message request to the access-point and its response

munication is planned as a stateless protocol where the request and response are given sequential numbers that any latest message received is always the right message. Since the labels are energy constrained and they do not have the guarantee to wake up in a period duty cycle, the requests are always from the labels i.e. those labels always initiate communication with the access-point. The access-point resolves the requests and sends back the requested data within a guaranteed period of time considering the processing delays.

4 Conclusion and Future Works

The multi-channel communication protocol has been designed and developed for the energy neutral labels. The software platform of the access-point which is not in focus of the article is designed in a very reliable manner that the data that is available on the access-point is always fetched in time that the labels are always communicated with latest information in the order management platform. A installation of 120 labels has

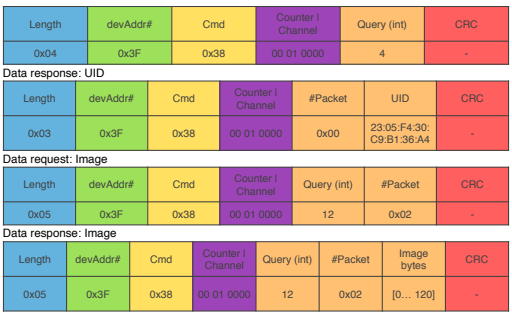


Figure 2: Message request to the access-point and its response

been implemented in two collocated places and were tested for the reliability of the use case. The energy harvesting nature of the labels are clearly characterised on the communication logs by the following: the labels stop communicating at a certain time in the evening when there is no light available and the energy harvested is depleted, the device consistently communicated again when there was availability of sun light and enough energy has been harvested for communication. The communication was very stable in a qualitative manner but a comprehensive analysis of the multi-channel protocol has to be made to understand the qualities of the communication strategy and also to set limits for the number of devices per access-point and to determine device density for a warehouse depending on different channel occupancy rates.

References

[1] Andreas Kamagaew, Thomas Kirks, and Michael ten Hompel. Energy potential detection for autarkic smart object design in facility logistics. In *Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on*, pages 285–290. IEEE, 2011.

[2] Aswin Karthik Ramachandran Venkatapathy, Moritz Roidl, Andreas Riesner, Jan Emmerich, and Michael ten Hompel. Phynetlab: Architecture design of ultra-low power wireless sensor network testbed. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a*, pages 1–6. IEEE, 2015.

[3] Ramachandran Venkatapathy, Aswin Karthik, Andreas Riesner, Moritz Roidl, Jan Emmerich, and Michael ten Hompel. Phynode: An intelligent, cyber-physical system with energy neutral operation for phynetlab. In *Smart SysTech 2015; European Conference on Smart Objects, Systems and Technologies; Proceedings of*, pages 1–8. VDE, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A6

Resource-efficient Graph Mining

Kristian Kersting Petra Mutzel
Christian Sohler

Finding Largest Common Substructures of Molecules in Quadratic Time

Andre Droschinsky
Chair of Algorithm Engineering (LS11)
TU Dortmund
andre.droschinsky@tu-dortmund.de

Finding the common structural features of two molecules is a fundamental task in cheminformatics. Most drugs are small molecules, which can naturally be interpreted as graphs. Hence, the task is formalized as maximum common subgraph problem. Albeit the vast majority of molecules yields outerplanar graphs this problem remains NP-hard.

We consider a variation of the problem of high practical relevance, where the rings of molecules must not be broken, i.e., the block and bridge structure of the input graphs must be retained by the common subgraph. We present an algorithm for finding a maximum common connected induced subgraph of two given outerplanar graphs subject to this constraint. Our approach runs in time $\mathcal{O}(\Delta n^2)$ in outerplanar graphs on n vertices with maximum degree Δ. This leads to a quadratic time complexity in molecular graphs, which have bounded degree. The experimental comparison on synthetic and real-world datasets shows that our approach is highly efficient in practice and outperforms comparable state-of-the-art algorithms.

The results, including an extensive experimental comparison, have been accepted at the *43rd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2017)* [3].

1 Introduction

The maximum common subgraph problem arises in many application domains, where it is necessary to elucidate common structural features of objects represented as graphs.

In cheminformatics this problem has been extensively studied [5, 6] and is often referred to as maximum or *largest common substructure problem*. Two variants of the problem can be distinguished: The maximum common induced subgraph problem (MCIS) is to find isomorphic induced subgraphs of two given graphs with the largest possible number of vertices. The maximum common edge subgraph problem (MCES) does not require that common subgraphs are induced and aims at maximizing the number of edges. Both variants can be reduced to a maximum clique problem in the product graph of the two input graphs [5] and are NP-hard. However, there are several restricted graph classes which render polynomial time algorithms possible. The seminal work in this direction is attributed to J. Edmonds [4], who proposed a polynomial time algorithm for the maximum common subtree problem. Here, the given graphs and the desired common subgraph must be trees. Recently, we showed that this problem can be solved in time $\mathcal{O}(\Delta n^2)$ for (unrooted) trees on n vertices with maximum degree Δ [2]. Unfortunately, already for outerplanar graphs the problem is NP-hard. On the other hand, if both graphs are biconnected and outerplanar, we can compute a biconnected MCS in polynomial time. These complexity results and the demand in cheminformatics lead to the consideration of MCES under the so-called *block and bridge preserving* (BBP) constraint [6], which requires the common subgraph to retain the local connectivity of the input graphs. BBP-MCES is not only computable in polynomial-time, but also yields meaningful results for cheminformatics.

We took up the concept of BBP to propose a novel BBP-MCIS algorithm with running time $\mathcal{O}(\Delta n^2)$ in outerplanar graphs with n vertices and maximum degree Δ. Our approach supports to solve BBP-MCIS w.r.t. a weight function on the mapped vertices and edges. The experiments have shown that BBP-MCIS in almost all cases yields the same results as BBP-MCES for molecular graphs under an adequate weight function. Our method outperforms in terms of efficiency the BBP-MCES approach of [6] by orders magnitude, cf. Fig. 1.

2 Computing a maximum common subtree isomorphism

We recently showed, that for arbitrary weights a maximum common subtree isomorphism between two trees G and H of order n with $\Delta(G) \leq \Delta(H)$ can be computed in time $\mathcal{O}(n^2(\Delta(G) + \log \Delta(H))$ [2]. The previously proven time bound was $\mathcal{O}(n^5)$ [1]. We obtained the improvement by (i) considering only a specific subset of subproblems that we showed to be sufficient to guarantee an optimal solution; (ii) exploiting the close relation between the emerging matching instances. We showed that for general trees any further improvement of this time bound would allow to solve the assignment problem in $o(n^3)$, and hence improve over the best known approach to this famous problem for more than 30 years. For trees of bounded degree the running time bound of $\mathcal{O}(n^2)$ is tight.

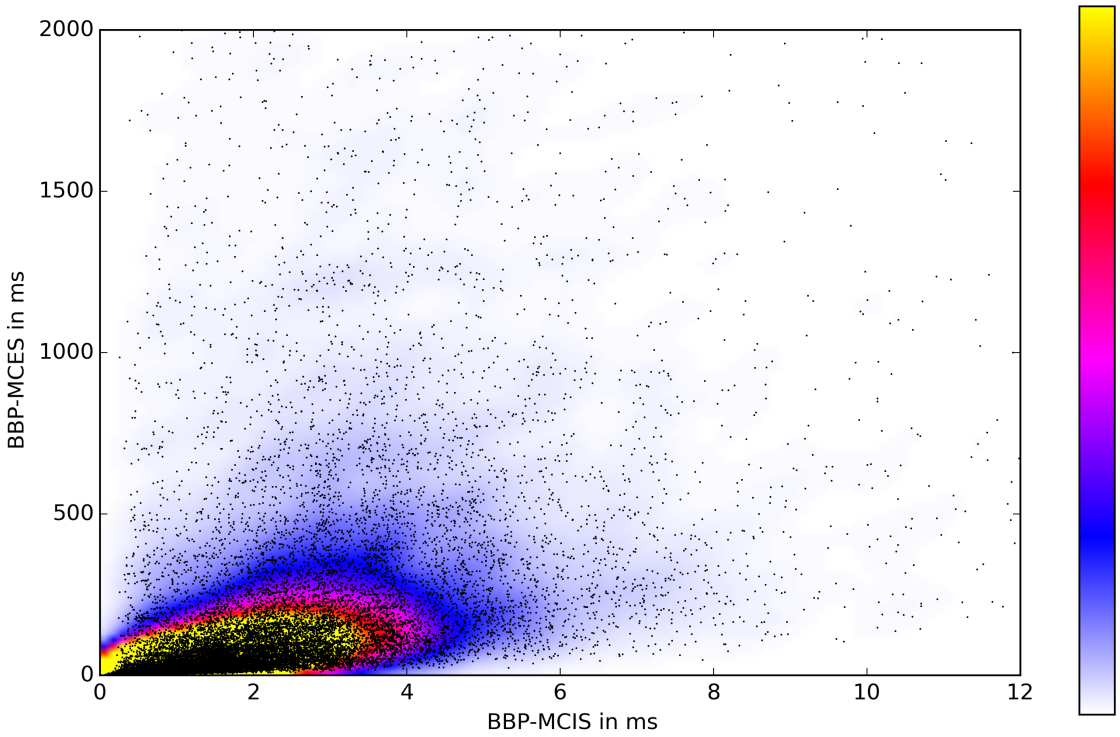
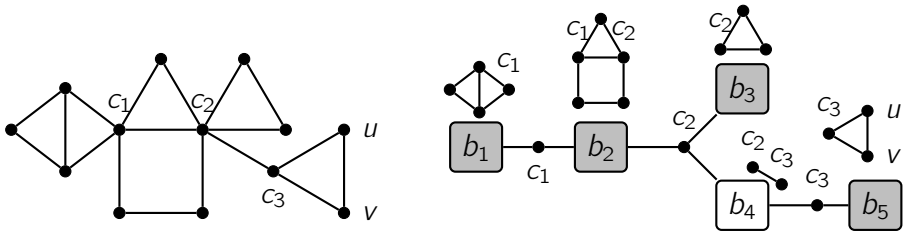


Figure 1: Running times in ms for 28 399 BBP-MCS computations. Each black dot represents a BBP-MCS computation on two randomly chosen outerplanar molecular graphs. It directly compares the running time of our algorithm (MCIS, x-axis) and the implementation from [6] (MCES, y-axis). The running times of another 601 BBP-MCS computations did not fit into the borders.

3 Largest Common Substructures of Molecules

We took up the concept for molecular graphs with its rings, which may not be broken, and proposed a novel block and bridge preserving maximum common subgraph algorithm (BBP-MCIS) with running time $\mathcal{O}(\Delta n^2)$ for outerplanar graphs with n vertices and maximum degree Δ. The degree of the vertices in molecular graphs is bound by $\mathcal{O}(1)$, therefore we achieve quadratic running time in this case. We obtained this result by combining ideas of [2] for the maximum common subtree problem with a new algorithm for finding biconnected MCIS in biconnected outerplanar graphs. For this subproblem we developed a quadratic time algorithm, which exploits the fact that the outerplanar embedding of a biconnected outerplanar graph is unique. One important data structure to combine the ideas are BC-trees, which allow a representation of the molecular graphs as trees, cf. Fig. 2. Our approach supports to compute a BBP-MCIS w.r.t. a weight function on the mapped vertices and edges. Our experiments suggest that BBP-MCIS in almost all cases yields the same results as BBP-maximum common edge subgraph



(a) Input graph (b) BC-tree of the input graph

Figure 2: A connected outerplanar graph (a) and its BC-tree (b). Block nodes are gray, while bridge nodes are white. The solid black nodes are cutvertices. The corresponding subgraphs of G are shown above the block and bridge nodes.

(BBP-MCES) for molecular graphs under an adequate weight function. Moreover, our method outperforms the BBP-MCES approach of [6] by orders magnitude.

References

[1] Andre Droschinsky, Bernhard Heinemann, Nils Kriege, and Petra Mutzel. Enumeration of maximum common subtree isomorphisms with polynomial-delay. In Hee-Kap Ahn and Chan-Su Shin, editors, *Algorithms and Computation (ISAAC)*, LNCS, pages 81–93. Springer, 2014.

[2] Andre Droschinsky, Nils Kriege, and Petra Mutzel. Faster Algorithms for the Maximum Common Subtree Isomorphism Problem. In *MFCS 2016*, volume 58 of *LIPIcs*, pages 34:1–34:14, 2016. `arXiv:1602.07210`.

[3] Andre Droschinsky, Petra Mutzel, and Nils Kriege. Finding largest common substructures of molecules in quadratic time. 43rd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2017), to appear.

[4] David W. Matula. Subtree isomorphism in $O(n^{5/2})$. In P. Hell B. Alspach and D.J. Miller, editors, *Algorithmic Aspects of Combinatorics*, volume 2 of *Annals of Discrete Mathematics*, pages 91–106. Elsevier, 1978.

[5] John W. Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7):521–533, 2002.

[6] Leander Schietgat, Jan Ramon, and Maurice Bruynooghe. A polynomial-time maximum common subgraph algorithm for outerplanar graphs and its application to chemoinformatics. *Annals of Mathematics and Artificial Intelligence*, 69(4):343–376, 2013.

Discovering hidden structure in news articles

Elena Erdmann
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
elena.erdmann@tu-dortmund.de

While big data makes available vast amounts of texts and news, researchers in social and communicational studies are just beginning to understand the potential of large scale analysis for their fields. In a highly interdisciplinary approach, we aim to bridge the gap between graph analytics, Natural Language Processing and Journalism studies.

We developed a novel topic models that provides additional information by deconstructing each topic as a directed graph of words. We applied our technique alongside several state of the art methods to compare the discourse about the trade agreement TTIP in the USA and Germany. Finally we give an overview on the current work about improving topic classification by adding human interaction to the machine learning process.

Admixtures of PDNs capture directed dependencies in topics [2] Inouye et al.'s Admixtures of Poisson MRFs (APM) [5] can model undirected word dependencies within a topic and in turn better capture the semantics of a topic than independent topic models such as LDA. Often, however, word dependencies are asymmetric: If the word 'growth' appears in a text (specifically a financial text), the appearance of (increasing) 'bank' might be more likely, whereas the appearance of the word 'bank' does not make 'growth' more likely. To capture such asymmetric word dependencies, we introduce a novel topic model based on an admixture of Poisson Dependency Networks (APD). APDs combine Innouye's generalized admixtures [5] with Poisson Dependency Networks [4]. This is an appealing representation—both distributional and graphically— since the directedness of dependency networks naturally captures asymmetric word dependencies, and even symmetric ones when using two directed ones. The generative process of Admixtures of

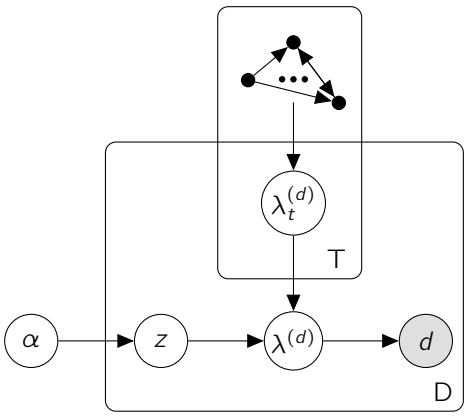


Figure 1: Plate model of APDs illustrating the two layers of hidden variables z and λ_t. Fixing z allows to estimate λ_t efficiently, and vice versa.

PDNs (APD) is inspired by LDA [1] and displayed in figure 1. We empirically demonstrate that our method compares to APM and LDa in all standard measurements and further allow to detect directed connections between words that give additional insight to the researcher.

Towards an International View on TTIP [3] TTIP affects millions of people living in the U.S. and the E.U. and its negotiations have been controversial. However, content analysis of newspapers indicates that the issue is of diverging national importance. We compare coverage in New York Times (NYT) and in Süddeutsche Zeitung (SZ), both high-circulation dailies from metropolises that exhibit a rather liberal orientation. Despite these similarities, coverage on TTIP varies significantly (see Fig. 2(a)). It is notable that coverage on TTIP increases considerably from 2014 in SZ, whereas the number of articles in the NYT remains unaltered. In order to shed light on this disparity the general coverage on the U.S.A. and Europe, respectively, was analyzed.

(Topic Models) LDA was used to find 100 topics in each sub-corpus. The topics were labeled by journalists using top words and top articles. A glance at the results shows distinctly the different perspectives on TTIP in the public spheres. In SZ TTIP appears on the top word list of a topic covering articles on policies of the European Commission along with conjoined words like customs, arbitration and investor protection. In NYT TTIP is not among the top words of the European Commission topic which is dominated by stakeholders dealing with financial issues around the euro crisis.

(Directed word dependencies) A PDN [4] trained on the articles of Süddeutsche Zeitung shows that the constitution of the E.U. as a politico-economic union of 28 states places the question of parliamentary participation in the foreground (see Fig. 2(b)). In the U.S., this question does not arise.

(a) Comparison of number of articles containing 'TTIP' in New York Times and Süddeutsche Zeitung.

(b) PDN created from the articles containing 'TTIP' in Süddeutsche Zeitung. (German words were translated to English)

(c) Attentional curves capture the development of topics in news articles over time, here illustrated for the war on Ukraine.

Figure 2: Boosting international understanding and transparency in news coverage using machine learning and data-driven journalism.

(Attentional Topic Models) When analyzing the discourse on Europe in NYT through Attentional Topic Models [6], we found no attentional topic corresponding to the TTIP. Instead, the discussion focused on different topics such as the war in the Ukraine (see Fig. 2(c)).

(word2vec) Analyzing word2vec results highlights diverse reciprocal perception: U.S. and German newspaper both coincide covering Germany mainly considering the recent migration. However, on the coverage on the U.S. SZ and NYT drift apart: In the SZ the importance of U.S. as an economic partner is demonstrated, whereas the NYT covers the U.S. in a broader range of topics including several sports. Comparing the use of TTIP indicates that the NYT uses more matter-of-fact words in connection with TTIP while SZ seems to include more commenting and evaluating words including 'chlorinated chicken'.

Applying machine learning to understand the cross-nationally diverse discourse on TTIP highlights the benefit which can be derived from an interdisciplinary approach. Yet, this interdisciplinary approach is still in its infancy.

Towards Topic Classification by Human Machine-Learning Interaction When studying the discourse about a given topic, such as TTIP, a simple keywordsearch can already provide most articles that are linked to the topic. In our cooperation with researchers from journalism studies, we found that they preferred topics generated by topic models such as LDA. What they specifically liked about these topics was the more comprehensive overview, given by the presentation as a multinomial distribution over words. However, when investigating one specific topic, unsupervised methods do not necessarily provide the topic of interest.

We believe that the key to a trustworthy topic classification of documents lies in an entirely interpretable model for classification. Our main objective in our future work

is to find all documents related to a certain topic in a large corpus of documents. In her mind, the user has a very detailed idea what these documents look like. When she reads a document, it comes easy to her to label the document as belonging to the topic or not. However, the user cannot formalize the topic to enable the computer to classify documents just as easily. We therefore propose an interactive approach to classify documents. We build a classifier to label documents and ask feedback from the user on both the classification and the features used by the classifier.

In a way, we regard the topic in the user's mind as a classification algorithm, that we now try to reverse-engineer. We start by training a classifier on a rough initial labeling of documents, such as given by a keywordsearch. We then distinguish the most influential features for our classifier. We present those features to the user and ask her to label these features as indicative, irrelevant or contradictory to the topic. We then use features that have been rated differently by user and machine, to actively sample those documents that are most likely to be misclassified by the algorithm. We ask the user to label these documents and retrain a new classifier on the user-labeled corpus. We repeat this process until the user is satisfied by the model for the topic as provided by the algorithm and the accuracy of the predictions.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning (ICML 2006)*, pages 113–120, 2006.

[2] E. Erdmann, A. Molina, and K. Kersting. Topic models with asymmetricword dependencies. manuscript under revision, 2016.

[3] Elena Erdmann, Karin Boczek, Lars Koppers, Gerret von Nordheim, Christian Pölitz, Alejandro Molina, Katharina Morik, Henrik Müller, Jörg Rahnenführer, and Kristian Kersting. Machine learning meets data-driven journalism: Boosting international understanding and transparency in news coverage. *arXiv preprint arXiv:1606.05110*, 2016.

[4] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, pages 1–31, 2015.

[5] D. Inouye, P. Ravikumar, and I. Dhillon. Admixture of poisson mrfs: A topic model with word dependencies. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 683–691, 2014.

[6] C. Pölitz, E. Erdmann, C. Bauckhage, H. Müller, K. Morik, and K. Kersting. Attentional topic models: Gompertz captures growth and decline of popular topics. manuscript under revision, 2016.

Lifted Convex Quadratic Programming

Martin Mladenov
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
martin.mladenov@cs.tu-dortmund.de

Symmetry is the essential element of lifted inference that has recently demonstrated the possibility to perform very efficient inference in highly-connected, but symmetric probabilistic models models. This raises the question, whether this holds for optimization problems in general. We show that for a large class of optimization methods this is actually the case. More precisely, we introduce the concept of fractional symmetries of convex quadratic programs (QPs), which lie at the heart of many machine learning approaches, and exploit it to lift, i.e., to compress QPs. These lifted QPs can then be tackled with the usual optimization toolbox (off-the-shelf solvers, cutting plane algorithms, stochastic gradients etc.). If the original QP exhibits symmetry, then the lifted one will generally be more compact, and hence their optimization is likely to be more efficient.

As our main contribution, we formalize the notion of symmetries of convex quadratic programs (QPs). Specifically, we first show that unlike for graphical models, where the notion of indistinguishability of variables is that of exact symmetry (automorphisms of the factor graph), QPs admit a weaker (partitions of indistinguishable variables which are at least as coarse) notion of indistinguishability called a fractional automorphism (FA) resp. equitable partition (EP). This implies that more general lifted inference rules for QPs can be designed. This is surprising, as it was believed that FAs apply only to linear equations. Second, we investigate geometrically how FAs of quadratic forms arise. The existing theory of symmetry in convex quadratic forms states that an automorphism of $\boldsymbol{x}^T Q \boldsymbol{x}$ corresponds to a rotational symmetry of the semidefinite factors of Q. We generalize this in that FA of $\boldsymbol{x}^T Q \boldsymbol{x}$ can be related not only to rotations, but also to certain scalings (as well as other not yet characterized properties of the semidefintie factors).

1 Main Results

We now state the main results of the research. First, we show the existence of the variables of a quadratic program such that there exist a solution of the quadratic program whose values are the same within an equivalence class of the partition. We call these lifting partitions.

Theorem 1. *Let $\boldsymbol{QP} = (Q, \boldsymbol{c}, A, \boldsymbol{b})$ encoding the convex quadratic program* minimize $\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}$ *s. t.* $A\boldsymbol{x} \leq \boldsymbol{b}$. *If $\mathcal{P}$ is a partition of the variables of $\boldsymbol{QP}$ with averaging operator $X^{\mathcal{P}}$, such that: (a) $X^{\mathcal{P}}Q = QX^{\mathcal{P}}$ and $\boldsymbol{c}^T X^{\mathcal{P}} = \boldsymbol{c}^T$, (b) there exists a partition $\mathcal{Q}$ of the constraints of $\boldsymbol{QP}$ with averaging operator $X^{\mathcal{Q}}$ such that $X^{\mathcal{Q}}\boldsymbol{b} = \boldsymbol{b}$ and $X^{\mathcal{Q}}A = AX^{\mathcal{P}}$, then $\mathcal{P}$ is a lifting partition for $\boldsymbol{QP}$.*

Proof. Found in [3]. □

Next, we show that these partitions can be found efficiently.

Theorem 2. *Given $\boldsymbol{QP} = (Q, \boldsymbol{c}, A, \boldsymbol{b})$, the coarsest partition $\mathcal{P}, \mathcal{Q}$ satisfying the above conditions can be found via the color refinement algorithm in quasilinear time.*

Proof. Found in [3]. □

The computation of such partitions works by converting the matrices and vectors involved in the QP to a coefficient graph and running colour refinement on it. An example of this conversion is given on Fig. 1

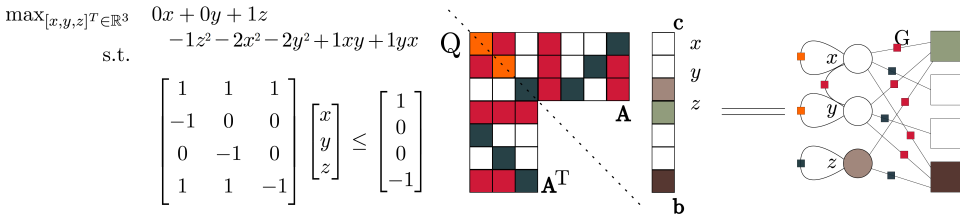


Figure 1: Example for partition detection.

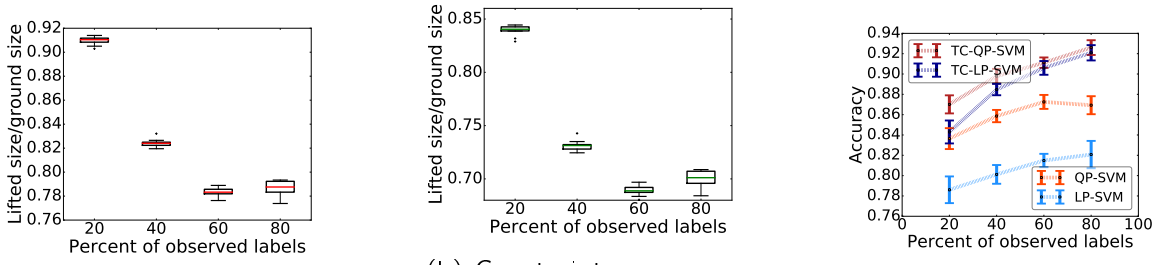
Our final investigation is inspired by the characterization of automorphisms of semidefinite matrices and quadratic forms. One way to think about a semidefinite matrix Q is as the Gram matrix of a set of vectors, i.e. $Q = BB^T$ where B is an $n \times k$ matrix and $k \geq \text{rank}(Q)$. In this light, the quadratic form $\boldsymbol{x}^T Q\boldsymbol{x}$ can be seen as the squared Euclidean norm of a matrix-vector product. That is, $\boldsymbol{x}^T Q\boldsymbol{x} = \boldsymbol{x}^T BB^T\boldsymbol{x} = (B^T\boldsymbol{x})^T(B^T\boldsymbol{x}) = ||B^T\boldsymbol{x}||^2$. It is a basic fact that the Euclidean norm is invariant under orthonormal transformations, that is, for any orthonormal matrix O and any vector $\boldsymbol{y}$, $||O^T\boldsymbol{y}|| = \boldsymbol{y}$ as $\boldsymbol{y}^T OO^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y}$. Thus, suppose we have a rotational autmorphism of B,

i.e., a pair of orthonormal matrix O and permutation matrix Π, such that $\Pi B = BO$ or also $\Pi BO^T = B$. That is, rotating the tuple of vectors that are the rows of B together yields same tuple back, but in different order. Observe then, that Π would be a renaming automorphism for Q, since $\Pi Q \Pi^T = \Pi BO^T OB\Pi^T = BB^T = Q$, implying $\Pi Q = Q\Pi$. Moreover, if the right dimension (number of columns) B is held fixed, the converse is true as well [1]. That is, not only do rotational symmetries of B correspond to renaming symmetries of Q, but vice-versa, as for fixed k, the semidefinite factors of Q are unique up to rotations. It turns out that a relaxed form of this argument holds for fractional symmetries.

Theorem 3. *Let X be a symmetric and X is idempotent (as our usual color-refinement automorphisms are) matrix, and $Q = BB^T$ be a positive semidefinite matrix with B having full column rank. Then $XQ = QX$* if and only if *there exists a symmetric matrix R such that $XB = BR$.*

Proof. Found in [3]. □

To evaluate the usefulness of this approach, we considered a relational classification task on the Cora dataset [4]. The Cora dataset consists of 2708 scientific papers classified into seven classes. Each paper is described by a binary word vector indicating the absence/presence of a word from a dictionary of 1433 words. The citation network of the papers consisting of 5429 links. The goal is to predict the class of the paper. For simplicity, we converted this problem to a binary classification problem by taking the largest of the 7 classes as a positive class. We compared four different learners on Cora. The base classifiers are an ∞-norm regularized SVM (LP-SVM) [6] and a conventional SVM (QP-SVM) [5] formulated as a convex QP. Both use the word feature vectors and do standard linear prediction (no kernel used). Additionally we considered transductive, collective versions of both of them following [2], denoted as TC-LP-SVM resp. TC-QP-SVM. Both transductive approaches have access to the citation network and implement the following simple rule: whenever we have access to an unlabaled paper i, if there is a cited or citing labeled paper j, then assume the label of j as a label of i. To account for contradicting constraint (a paper citing both papers of and not of its class), we introduced separate slack variables for the transductive constraints and add them to the objective with a different penalty parameter. In order to investigate the performance, we varied the amount of labeled examples available. That is, we have four cases, where we restricted the amount of labeled examples to $t = 20\%$, 40%, 60%, and 80% of size of the dataset. For all SVM models, we also ran a ground and a lifted version. The results are summarized in Fig. 2: QP-SVM outperforms LP-SVM for each setting, both are outperformed by TC-L/QP-SVM, and TC-QP-SVM outperforms TC-LP-SVM. While there was no appreciable symmetry in either QP-SVM or LP-SVM, TC-QP-SVM exhibited significant variable and constraint reduction: the lifted problem was reduced to up to 78% of the variables, resp., 70% of the constraints of the ground problem, while computing the same labels and in turn accuracy.



(a) Variable compression

(b) Constraint compression

(c) Accuracy

Figure 2: Relational experiments on the Cora dataset. (Best viewed in color)

References

[1] D. Bremner, M. Dutour Sikrić, and A. Schürmann. Polyhedral representation conversion up to symmetries. In *Proceedings of the 2006 CRM Workshop on Polyhedral Computations*. AMS, Providence, 2009.

[2] K. Kersting, M. Mladenov, and P. Tokmakov. Relational linear programming. *Artificial Intelligence Journal (AIJ)*, OnlineFirst, 2015.

[3] M. Mladenov, L. Kleinhans, and K. Kersting. Lifted inference for convex quadratic programs. In *, Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2017.

[4] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[5] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications and control series. John Wiley & Sons, New York. A Wiley-Interscience Publication, 1998.

[6] W. Zhou, L. Zhang, and L. Jiao. Linear programming support vector machines. *Pattern recognition*, 35(12):2927–2936, 2002.

Ressource-efficient Graph Kernels

Christopher Morris
Chair of Algorithm Engineering
TU Dortmund University
christopher.morris@tu-dortmund.de

We report about our recent progress in developing resource-efficient graph kernels. In the first part of this report, we present the *hash graph kernel framework*, which makes the comparison of large graphs with continuous labels feasible.

In the second part, we outline some work in progress, i.e., randomized algorithms for approximating graph kernels so that they can applied to very large graphs.

1 The Hash Graph Kernel Framework

Kernel methods are a broad class of algorithms for pattern analysis, which rely on a positive-semidefinite function, which measures the similarity between data objects. In several domains like chemo- and bioinformatics as well as social network and image analysis structured objects appear naturally. Graph kernels are a key concept for the application of kernel methods to structured data and various approaches have been developed in recent years, see [12, 9], and references therein. The considered graphs can be distinguished in

1. graphs with discrete labels, e.g., molecular graphs, where nodes are annotated by the symbols of the atoms they represent, and
2. attributed graphs with (multi-dimensional) real-valued labels in addition to discrete labels.

Attributed graphs often appear in domains like bioinformatics [2] or image classification [5], where attributes may represent physical properties of protein secondary structure elements or RGB values of colors, respectively. Taking the continuous information into account has been proven empirically to be beneficial in several applications, see, e.g., [2].

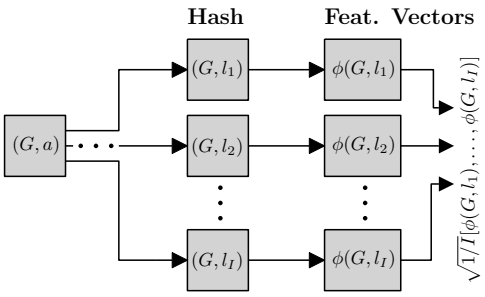


Figure 1: Illustration of the idea of hash graph kernels.

Graph kernels supporting complex annotations use implicit computation schemes and do not scale well, see, e.g., [4, 6, 10]. Whereas graphs with discrete labels can be compared efficiently by graph kernels based on explicit feature maps. This is what we made use of to develop a unifying treatment.

In [7, 8] we introduced *hash graph kernels* for graphs with continuous attributes. This family of kernels is obtained by a generic method, which iteratively hashes continuous attributes to discrete labels in order to apply a base kernel for graphs with discrete labels. This allows to construct a single combined feature vector for a graph from the individual feature vectors of each iteration, see Figure 1 for an illustration. We exemplified this for two established graph kernels:

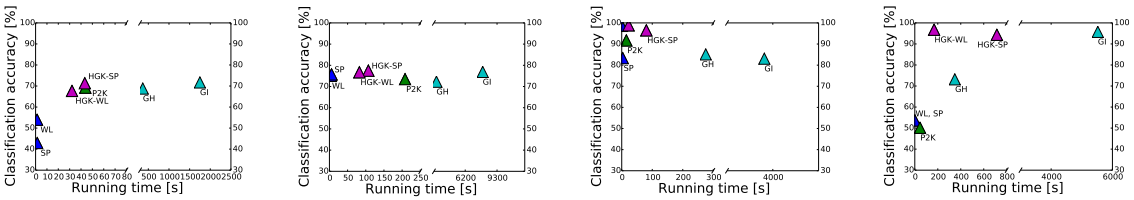
- We obtained a variation of the Weisfeiler-Lehman subtree kernel [11], which implicitly employs a non-trivial kernel on the node and edge annotations and is suitable for continuous values.
- Moreover, we derived a variant of the *shortest-path kernel* [1] which also supports continuous attributes while being efficiently computable by explicit feature maps.

For both kernels we provided a detailed theoretical analysis, i.e., the explicit hash graph kernel approximates the above implicit kernels arbitrarily close with any constant probability δ, i.e.,

$$\sup_{G,H \in \mathcal{G}} \left| \Phi(G)^\top \Phi(H) - k(G, H) \right| \leq \epsilon,$$

where $\mathcal{G}$ denotes a set multiset of graphs, $\Phi(G)$ for G in $\mathcal{G}$ denotes the feature vector computed by the hash graph kernel algorithm, and k denotes the kernel we want to approximate. The additive error ϵ and the success probability δ can be fine-tuned by the number of iterations of the hash graph kernel algorithm (Variable I in Figure 1).

We demonstrated the effectiveness of these kernel in an extensive experimental study on real-world and synthetic data sets. The results show that hash graph kernels are orders of magnitude faster than state-of-the-art kernels for attributed graphs without drop in classification accuracy, see Figure 2 for an illustration. We refer to [8] for further details.



(a) Enzymes (b) Proteins (c) SyntheticNew (d) Synthie

Figure 2: Overview of running times versus classification accuracies for the employed graph kernels and data sets (Purple triangles represent the hash graph kernel with shortest-path kernel and the Weisfeiler-Lehman subtree kernel, respectively.). Note the break in the x-axis.

2 Randomized Graph Kernels for the Comparison of Very Large Graphs

In the following we outline some work in progress. In very large networks, e.g., social networks or graph databases, we have to deal with networks with millions or even billions of nodes. Hence, even graph kernels that can be executed in polynomial-time are not feasible. Therefore, we are interested in deriving approximate graph kernels which rely on sampling.

Especially, we are interested in approximating the k-dimensional Weisfeiler-Lehman algorithm, which has deep connections to the graph isomorphism problem, e.g., see [3], and developing a corresponding graph kernel.

For $k = 1$ the algorithm has been successfully employed as a graph kernel [11]. Directly developing a graph kernel based on the k-dimensional Weisfeiler-Lehman algorithm for $k > 2$ is not feasible since its running time is in $\mathcal{O}(k^2 n^{k+1} \log n)$ for a single graph where n is the number of vertices [3]. Hence, we resort to a sample-based approximation algorithm.

We can already show that the feature vector of the 1-dimensional Weisfeiler-Lehman kernel can be approximated with an additive error with regard to the ℓ_1 norm. This also implies an additive error for the corresponding kernel. For bounded-degree graphs and a fixed number iterations this approximation can be computed in constant time.

Moreover, we want to implement our approximate kernels and evaluate them on large graphs stemming from social networks and chemo-informatics.

References

[1] K. M. Borgwardt and H.-P. Kriegel. "Shortest-path kernels on graphs". In: *Proceedings of the Fifth IEEE International Conference on Data Mining*. 2005, pp. 74–81.

[2] K. M. Borgwardt, C. S. Ong, S. Schönauer, S.V.N. Vishwanathan, A. J. Smola, and H.-P. Kriegel. "Protein function prediction via graph kernels." In: *Bioinformatics* 21 Suppl 1 (2005), pp. i47–i56.

[3] J. Cai, M. Fürer, and N. Immerman. "An optimal lower bound on the number of variables for graph identifications". In: *Combinatorica* 12.4 (1992), pp. 389–410.

[4] A. Feragen, N. Kasenburg, J. Petersen, M. D. Bruijne, and Borgwardt K. M. "Scalable kernels for graphs with continuous attributes". In: *Advances in Neural Information Processing System*. Erratum available at `http://image.diku.dk/aasa/papers/graphkernels_nips_erratum.pdf`. 2013, pp. 216–224.

[5] Z. Harchaoui and F. Bach. "Image Classification with Segmentation Graph Kernels". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8.

[6] N. M. Kriege and P. Mutzel. "Subgraph Matching Kernels for Attributed Graphs". In: *Proceedings of the Twentieth International Conference on Machine Learning*. 2012.

[7] C. Morris, N. M. Kriege, K. Kersting, and P. Mutzel. "Faster Kernels for Graphs with Continuous Attributes via Hashing". In: *Proceedings of the Fifth IEEE International Conference on Data Mining*. 2016.

[8] C. Morris, N. M. Kriege, K. Kersting, and P. Mutzel. "Faster Kernels for Graphs with Continuous Attributes via Hashing". In: *CoRR* abs/1610.00064 (2016). url: `http://arxiv.org/abs/1610.00064`.

[9] M. Neumann, R. Garnett, C. Bauckhage, and K. Kersting. "Propagation kernels: Efficient graph kernels from propagated information". In: *Machine Learning* 102.2 (2016), pp. 209–245.

[10] F. Orsini, P. Frasconi, and L. De Raedt. "Graph Invariant Kernels". In: *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*. 2015, pp. 3756–3762.

[11] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. "Weisfeiler-Lehman Graph Kernels". In: *Journal of Machine Learning Research* 12 (2011), pp. 2539–2561.

[12] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. "Graph Kernels". In: *Journal of Machine Learning Research* 11 (2010), pp. 1201–1242.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B1

Analysis of Spectrometry Data with Restricted Resources

Jörg Ingo Baumbach Jörg Rahnenführer

Automated versus semi-manual peak detection in MCC-IMS

Salome Horsch
Statistical Methods in Genetics and Chemometrics
Technische Universität Dortmund
salome.horsch@tu-dortmund.de

The analysis of breath gas for supporting diagnosis of diseases offers great prospects. Using the Multi-capillary-column-Ion-mobility-spectrometry (MCC-IMS) technology includes several advantages, particularly considering the current technical progresses, making the devices smaller and thus more mobile. That way MCC-IMS could be used in a broad manner evaluating data directly where it has been collected without important time delay. Currently, preprocessing still requires manual intervention, keeping the technology from this aim. Several methods for automatic preprocessing have been carried out in the past but reliable answers to which method to prefer are still missing. We compare 25 combinations of automated algorithms to the manual approach on three datasets and suggest a combination for future use.

1 Methods

The MCC-IMS-device accomplishes two main tasks. It leads the breath gas through the Multi-capillary-column measuring the time (retention time) different compounds take to pass it and afterwards hands them over to the Ion-mobility-spectrometer, where the molecules arrive successively. The molecules are then ionized and lead through the drift tube also measuring the time (drift time) the ions need. In the end an electrical potential is quantified (intensity) representing the amount of arriving ions. As a consequence pairs of times can be collected in data matrices with the ordered retention times as rows and a transformation of the drift times as columns, each cell containing the correspondent intensity. The matrices can be visualized as heatmaps, showing peaks, where similar peak

and retention times deliver related signal intensities. Those peaks need to be identified (peak picking) and afterwards compressed (peak clustering) among several measurements into a well defined set of peak locations and peak intensities for all measurements. In up-to-date practice, the peak picking and clustering step are accomplished in just one step under manual supervision, utilizing the commercial software VisualNow. This is considered the current gold standard. Here we show comparisons of several automated peak picking and peak clustering methods. To evaluate the performance of this process we apply all combinations on three sets of two-class data. We call a combination in preprocessing successful, when it achieves high AUC values in classification. Since this could be dependent on the classification algorithm we execute six different classification algorithms here as well. Adding the manual gold standard, 156 comparisons in total are performed on each dataset. Each combination is evaluated by classification results from a 50 times repeated 10-fold cross-validation (CV).
The competing algorithms for the three different tasks are not introduced here in detail, for more information the preliminary results on one dataset can be viewed in [1]. For peak picking six algorithms were tested, namely automated detection in VisualNow (VN^a), Local Maxima (LM), Peak Model Estimation (PME), Peak Detection by Slope Analysis (PDSA), Savitzky-Golay Laplace-operator filtering thresholding Regions (SGLTR) and Online Peak Model Estimation (OPME). Those were combined with the five peak clustering methods Automated Clustering in VisualNow (VN^a), Grid Square (GS), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Cluster Editing (CE) and EM Clustering (EM), where technically feasible. Note that the VN^a peak clustering could not be combined with other peak picking methods than its own. For classification six algorithms, namely linear and rbf- Support Vector Machines (SVM), K-Nearest-Neighbor (kNN), Classification Tree (CT), Generalized Boosted Models (GBM) and Random Forest (RF) were used. The setting has slightly been changed for GBM compared to [1], in the way that an additional tuning parameter was used.
We use three real world datasets, each of them containing a healthy control group and a group of diseased probands. The first dataset consists of 127 persons (97 cases and 35 controls), the second dataset contains 67 (37 cases and 30 controls) and the third 67 (37 cases, 30 controls) persons. The numbers of variables and thus features for classification vary strongly among the different combinations of preprocessing between 11 and 265.

2 Results

We compare 26 methods for peak detection, each one repeatedly with six classification algorithms and applied to three different datasets. To merge the results from all datasets, the 156 combinations were ranked on each dataset, the combination achieving the highest median AUC (the median of 50 replications of 10-fold CV) ranked 1 and the worst combination ranked 156. Table 1 shows the 20 best combinations with their respective

rank sums and the mean of the median AUCs of the three datasets. The top three methods all include SGLTR peak picking and a RF classification with very low ranks on the second and third dataset (ranks 1-7) and a little worse ranks on the first dataset (12-14) with correspondent mean AUC values of 0.947-0.957 among the three datasets. Rank four is occupied by the current gold standard VN^m combined with RF classification, showing, that indeed this seems to be a reliable method. In general, RF is clearly performing best in classification, occuring 13 times in the table. Each time another method is used in the table (6 times GBM, once SVM^{rbf}) the corresponding combination performed better using RF. As far as the peak picking methods are concerned the best automated algorithms include SGLTR, LM and VN^a, occurring 7, 4 and 6 times in the top 20 table, respectively. OPME is included once. For the peak clustering methods there is no simple rule that shows which one is best in general. But overall we find DBSCAN 7 times, EM 4, CE 3 and VN^a and GS 2 times among the top 20 combinations.

Table 1: Ranks of median AUCs for each combination of peak picking, peak clustering and classification algorithms, ranksum over the three datasets and corresponding mean AUC.

					Rank			
No.	Peak	Cluster	Classif	$\overline{AUC}$	D_1	D_2	D_3	$\sum D_i$
1	SGLTR	DBSCAN	RF	0.957	12	3	1	16
2	SGLTR	EM	RF	0.950	12	2	5	19
3	SGLTR	CE	RF	0.947	14	7	3	24
4	VN^m	VN^m	RF	0.936	6	19	2	27
5	LM	DBSCAN	RF	0.916	2	15	18	35
5	VN^a	VN^a	RF	0.925	3	9	23	35
7	LM	EM	RF	0.913	5	11	26	42
8	VN^a	DBSCAN	RF	0.919	8	31	4	43
9	SGLTR	EM	GBM	0.927	26	5	15	46
10	LM	GS	RF	0.914	4	4	44	52
11	SGLTR	GS	RF	0.907	17	12	32	61
12	SGLTR	CE	GBM	0.919	48	7	16	71
13	VN^a	VN^a	SVM^{rbf}	0.897	27	27	20	74
14	VN^a	EM	RF	0.889	1	60	17	78
15	VN^a	DBSCAN	GBM	0.914	50	20	10	80
16	OPME	DBSCAN	RF	0.885	15	46	29	90
16	VN^m	VN^m	GBM	0.901	31	51	8	90
18	LM	DBSCAN	GBM	0.884	18	13	62	93
19	SGLTR	DBSCAN	GBM	0.921	82	6	11	99
19	VN^a	CE	RF	0.894	33	54	12	99

Looking at the performance of the top 20 combinations in detail, we can see that there are great differences in the rank sums. The best combination (SGLTR, DBSCAN, RF) scores with a rank sum of just 16 whereas the worst top 20 combination reaches a rank sum of 99. This shows that there is a great variety in the performances on the three datasets. Mostly there is one dataset on which the constellation of algorithms performs a lot worse than on the others. It is rare that the performances are equal in all three

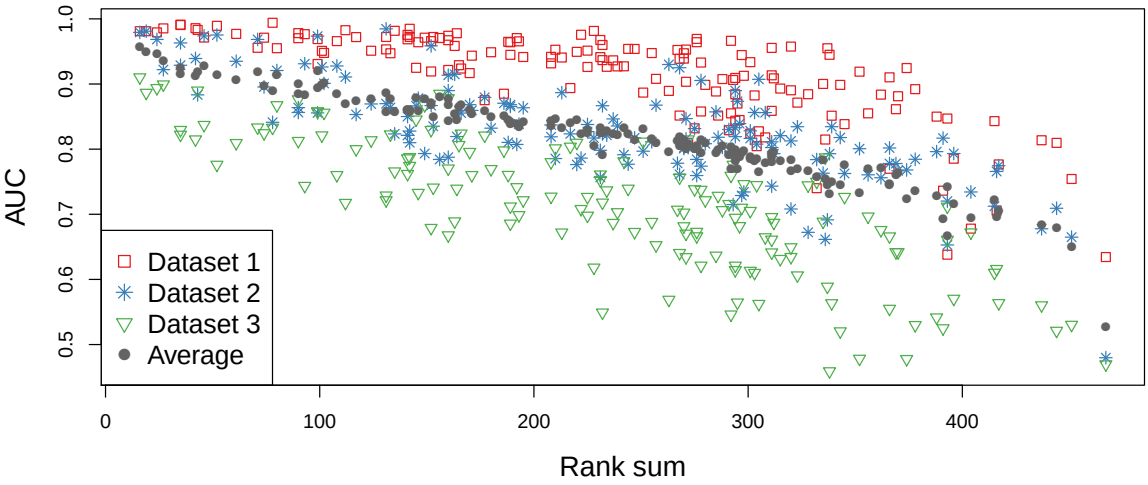


Figure 1: Correlation between the rank sum of all combinations and the median AUC values in the three datasets and their respeactive arithmetic mean.

settings (as for VN^a, VN^a, SVM^{rbf} with ranks 27, 27, 20 for the three datasets, respectively). Figure 1 displays on the y-axis the median performances in terms of AUC for the combinations in all three datasets as well as their respective arithmetic mean, related to the rank sum over all datasets on the x-axis. In general it can be seen that the median AUC values are of course highly correlated to the rank sum. The first dataset is easiest to classify and also there are no big differences in AUC for the combinations with the best rank sums for this setting. For the second and especially the third dataset a strong negative correlation is visible between the rank sum and the median AUC. So an increase in the ranks goes along with a decrease of median AUC. This relativizes the higher ranks for the combination of the SGLTR variants in the first dataset since the correspondent AUC values are only slightly lower than the ones with lower ranks. More important is the success for these combinations on the other two datasets where they achieve also much better AUCs than the rival combinations.
Therefore we suggest SGLTR peak picking with DBSCAN (or possibly EM) peak clustering and a Random Forest classification for future classification tasks in MCC-IMS.

References

[1] S. Horsch, D. Kopczynski, J. I. Baumbach, J. Rahnenführer, and S. Rahmann. From raw ion mobility measurements to disease classification: a comparison of analysis processes. *PeerJ PrePrints*, 3, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B2

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

Jian-Jia Chen Roland Hergenröder
Heinrich Müller

Schedulability Analysis for Resource Access Tasks with Control Flow Graph

Wen-Hung Kevin Huang
SFB 876, Project B2
Computer Science XII, TU Dortmund
wen-hung.huang@tu-dortmund.de

Resource access is becoming a prominent characteristic in real-time systems such as: (i) I/O-intensive systems (ii) multi-core processors, and (iii) computation offloading systems with coprocessors, like Graphics Processing Units (GPUs) or mobile cloud computing. Traditionally, the Worst-Case Execution Time is derived in a way that the CPU presumably stalls during such accesses. This may result in enormous underutilizations of systems. In this report, we propose an approach based on control flow graph to avoid any unnecessary utilization loss in abstract interpretation.

1 Introduction

Traditionally, timing analysis consists of two separate steps: (*i*) worst-case execution time (WCET) analysis, which computes an upper bound on the execution time of a single job of a task running in isolation, and (*ii*) schedulability analysis, which determines whether multiple tasks are guaranteed to meet their deadlines, while sharing a processor.

As far as resource access is concerned, the delay of access must be taken into account of calculating the WCET. Traditional approaches assume that each memory access takes constant time. For instance, SymTA/P assumes that each memory access takes constant time. This approach is sensible only if resource access time is relatively short. However, in practice each resource access can range from a few microseconds (e.g., a write operation on a flash drive) to a few hundreds of milliseconds (e.g., offloading computation to Graph Processing Units (GPUs)). As a result, subsuming resource consumptions into WCET can result in a huge underutilization of systems.

2 Control Flow Graph (CFG), Directed Acyclic Graph (DAG), and Multi-Path Suspension Task

The task to be analyzed attains its WCET on one (or sometimes several) of its possible execution paths. In general, however, this worst-case input and initial state are not known and hard or impossible to determine. A data structure, the task's control-flow graph, CFG, describes a superset of the set of all execution paths. There are a number of

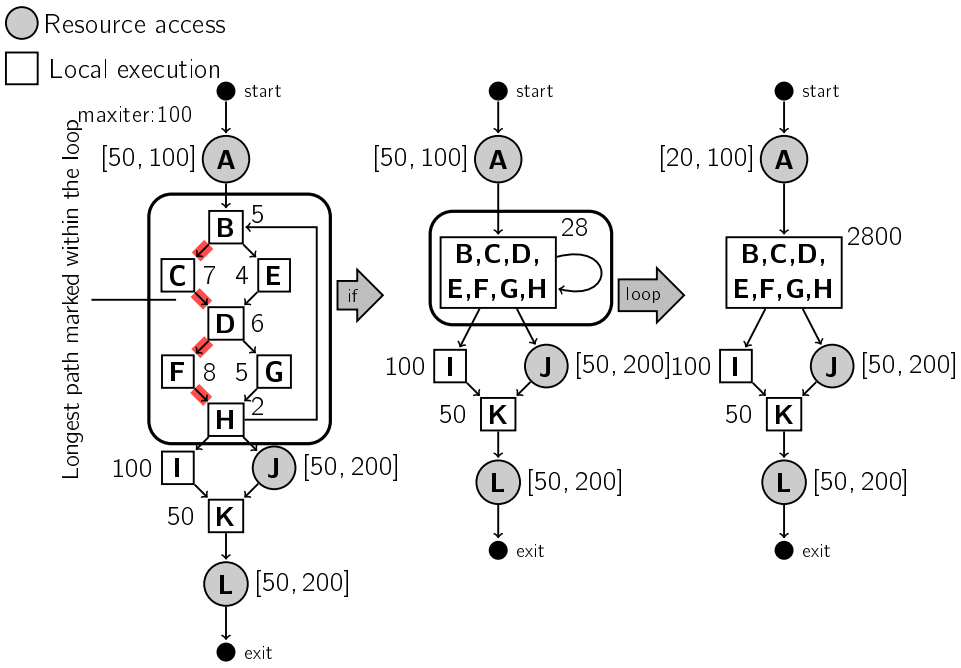


Figure 1: The longest path within a loop is first identified. Collections of local execution nodes within a loop are collapsed into single nodes, simultaneously deriving a timing for the new node.

approaches to automatic flow analysis. Some of the methods are general, while others are specialized for certain types of code constructs.

Definition 1 *A* basic execution block b^e *is a sequence of instructions which contain no instruction to access the shared bus.*

Definition 2 *A* basic access block b^r *is a sequence of instructions of resource access.*

Definition 3 *A task τ_i consists of a set of basic blocks B_i together with a starting block b_i^{start} at which the execution of task τ_i starts and an exit block b_i^{exit} at which task τ_i is left, each of which is categorized as basic execution block b^e or basic access block b^r. Functions of task τ_i are a* control flow graph*(CFG) where $G_i = (V_i, E_i)$*

$$V_i = \bigcup_{b_x \in B_i} b_x^e \bigcup_{b_x \in B_i} b_x^r \tag{1}$$

and for all b_j, b_k contains an edge $(b_x, b_y) \in E_i$ if immediate flow of control from $b_x \in \{b^e, b^r\}$ to $b_y \in \{b^e, b^r\}$ is possible.

A task τ_i is characterized by a three-tuple $\tau_i = (G_i, T_i, D_i)$: G_i is its control flow graph; T_i denotes the minimum inter-arrival time (also known as period) of τ_i; and each job of τ_i has a relative deadline D_i. We here further assume the CFG contains no loops. The loops would be preproceeded by using the techniques in the literature [3]. To put it more precisely, the loop in the graph is first identified and the longest path within the loop is found. The time for the longest path is combined with flow information about the loop bound to extract an upper bound for the nodes with the loop, as shown in Figure 1.

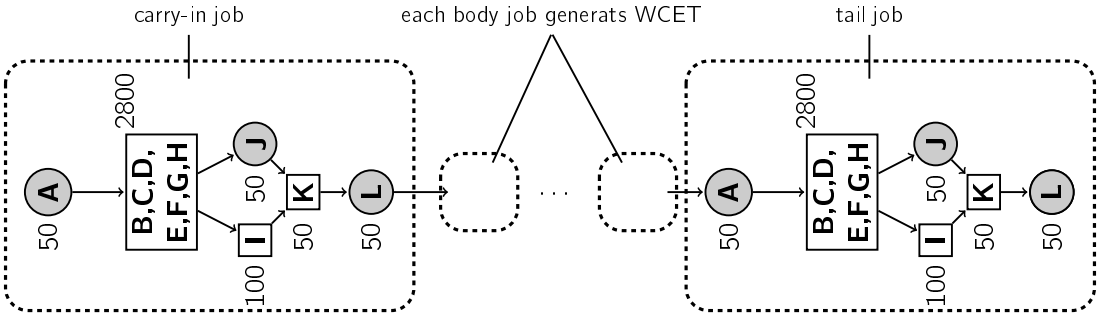


Figure 2: The work of higher-priority task τ_i over an interval is composed of carry-in job, body jobs, and tail job, in which each resource access presumably takes its **lower** bound.

Definition 4 *A* path *through a control flow graph* $G = (V, E)$ *is a sequence of nodes* $P = (v_0, v_1, \cdots, v_\ell)$ *such that* $(v_{j-1}, v_j) \in E$ *for all* $j \in \{1, \cdots, \ell\}$

Profit and Cost. Let a profit p_x be associated with each vertex $v_x \in V_i$ and a distance $c_{x,y}$ be associated with each edge $(v_x, v_y) \in E_i$. The profit of each execution block b^e is given an upper bound on the total execution time of that block every time it is locally executed, whereas the profit of each resource block b^e is *null*; On the other hand, the cost of each edge is given an upper bound on the total execution time of that block every time it is locally executed, if the head of the edge is a basic execution block; otherwise, that every time it accesses shared.

3 Schedulability Analysis

The response time of a task can be decomposed into two parts (*i*) the times spent itself execution and resource, and (*ii*) the times interfered by higher-priority tasks.

Computing (*i*). In this case, the contribution must be no more than the maximum time among all paths task τ_k goes through, irrespective of the types of basic blocks.

Definition 5 *The Worst-Case Cost Path (*$WCCP_i$*) of a task* τ_i *is a path* $b_{start} \rightarrow b_{exit}$ *taking the* **upper** *bound on resource access through* G_i *with the maximum cost* $\psi(P)$ *among all such paths. The Worst-Case Cost Time (WCCT) of a task* τ_i *is equal to* $\psi(WCCP_i)$.

Computing (*ii*). In this case, we need to bound the cumulative execution time of each task τ_i that can possibly generate over an interval of length t. Roughly speaking, we can classify those jobs as three different types: (*a*) head job: the job having release time prior to the beginning of the interval and absolute deadlines within the interval, (*b*) body jobs: the jobs having release times and absolute deadlines within the interval, and (*c*) tail job:the last job release. This is also illustrated in Figure 2.

We first have the following observation: each body job can possibly contribute for the path with the maximum execution time. To this end, we define the Worst-Case Execution Path and the Worst-Case Execution Time (WCET) of a task, which excludes resource access blocks.

Definition 6 *The Worst-Case Execution Path ($WCEP_i$) of a task τ_i is a path $b_{start} \rightarrow b_{exit}$ through G_i with the maximum profit $\lambda(P)$ among all such paths. The Worst-Case Execution Time (WCET) of a task τ_i is equal to $\lambda(WCEP_i)$.*

The calculation of head and tail job is equivalent to the *Selective Travelling Salesman Problem* (STSP) [2], where it is not necessary to visit all vertices. Given a weighted graph with profits associated with the vertices, the STSP (or *orienteering problem*) consists of selecting a simple circuit of maximal total profit (computation time), whose length does not exceed a prespecified bound (total time). In fact, this problem can be solved by using *dynamic programming* [1].

Definition 7 (Concatenation) *The* concatenation *is* $G_1; G_2 = (V_1 \cup V_2, E_1 \cup E_2 \cup (v_{G_1}^{end}, v_{G_2}^{start}))$

Putting these pieces together, we have the following theorem:

Theorem 1 (TDA) *If the worst-case response time of task τ_k is $\leq T_k$ and there are no deadline misses from higher-priority tasks $hp(k)$, then the worst-case response time of CFG task τ_k is upper bounded by the smallest R_k satisfying the following recurrence:*

$$R_k = \psi(WCCP_k) + \sum_{\tau_i \in hp(k)} W_i(R_k) \tag{2}$$

where $hp(k)$ denotes the set of the tasks with higher priority than task τ_k and

$$W_i(t) = STSP(G_i; G_i, t_i) + n_i(t) \cdot \lambda(WCEP_i) \tag{3}$$

and

$$t_i = t - \left\lceil \left\lfloor \frac{t - 2T_i}{T_i} \right\rfloor \right\rceil^{\dagger} T_i \tag{4}$$

and

$$n_i(t) = \left\lceil \left\lfloor \frac{t - 2T_i}{T_i} \right\rfloor \right\rceil^{\dagger} \tag{5}$$

where $[x]^{\dagger}$ is defined as $max(0, x)$

References

[1] M Hayes and JM Norman. Dynamic programming in orienteering: route choice and the siting of controls. *Journal of the Operational Research Society*, 35(9):791–796, 1984.

[2] Gilbert Laporte and Silvano Martello. The selective travelling salesman problem. *Discrete Applied Mathematics*, 26(2-3):193–207, 1990.

[3] Reinhard Wilhelm, Jakob Engblom, Andreas Ermedahl, Niklas Holsti, Stephan Thesing, David B. Whalley, Guillem Bernat, Christian Ferdinand, Reinhold Heckmann, Tulika Mitra, Frank Mueller, Isabelle Puaut, Peter P. Puschner, Jan Staschulat, and Per Stenström. The worst-case execution-time problem - overview of methods and survey of tools. *ACM Trans. Embedded Comput. Syst.*, 7(3), 2008.

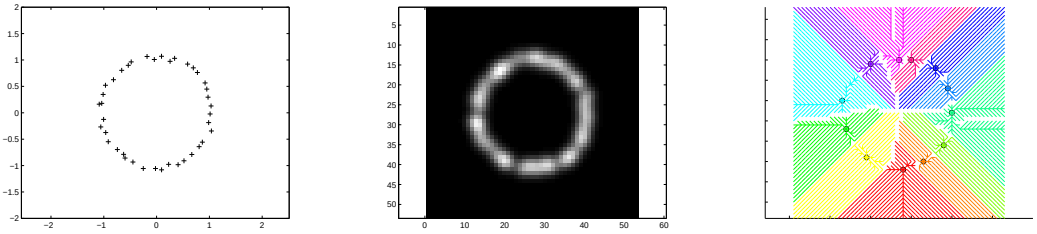
Dimension-recursive reconstruction of manifold hyper-surfaces from scattered point data

Thomas Kehrt
Lehrstuhl Informatik VII (Graphische Systeme)
Technische Universität Dortmund
thomas.kehrt@tu-dortmund.de

This report concerns the reconstruction of manifolds from noisy point clouds. We present an approach that is basically suitable for point clouds of any dimension. Starting from the point cloud a smooth function is obtained by low-pass filtering, which then is decomposed as a Morse-Smale-like cell complex by recursion over the spatial dimension. For each cell, a surface patch contributing to the resulting manifold is determined locally.

1 Introduction

Retrieving structure from scattered point data is crucial to a wide variety of applications. A considerable number of algorithms such as regression, interpolation, clustering, principle component analysis and many more have been developed in the past. We aim to provide algorithmic means to reconstruct $(N-1)$-manifold hyper-surfaces of arbitrary algebraic topology embedded in the N-dimensional space from finite sets of points sampled from a surface of this kind. As a particular challenge, the requirements to the point cloud in the form of signal-to-noise-ratio, sampling density, outliers etc. may be low. In this report, an approach is outlined which takes data quality (sampling density, systematic errors and noise) into account by a single parameter. Its value, however, has currently to be chosen manually.



(a) Input point set. (b) Function D applied on a grid. (c) 2D cells by color with integral curves.

Figure 1: Cell complex.

2 Method

In the following the method is illustrated for point clouds in two dimensions in order to ease understandability. To see the details in the figures, the zoom-in facilities of a document reader should be employed.

Given a point set P (see fig. 1a), where $P \subset \mathbb{R}^N$, the first step is to replace it with a sum of Dirac functions,

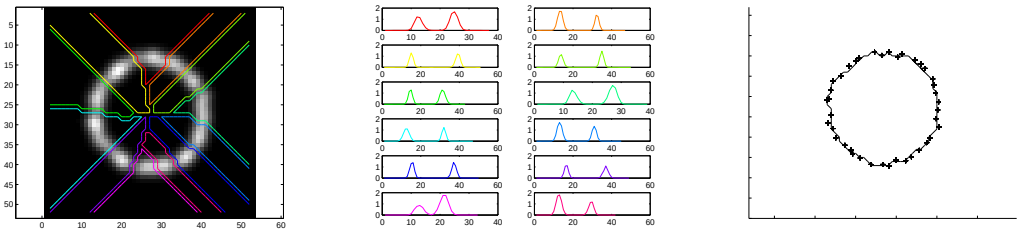
$$\hat{D}(\mathbf{x}) = \sum_i \delta(\mathbf{x} - \mathbf{p}_i),$$

by placing an impulse at every point $\mathbf{p}_i \in P$. Next, a continuous function is generated by convolution with a smooth low-pass filtering kernel $K(\mathbf{x})$ (fig. 1b):

$$\begin{aligned} D(\mathbf{x}) &= \hat{D}(\mathbf{x}) * K(\mathbf{x}) \\ &= \sum_i K(\mathbf{x} - \mathbf{p}_i). \end{aligned}$$

This resembles well-known approaches of signal processing such as mean-shift, as well as of probability theory such as Kernel Density Estimation (KDE). Basically K is a reconstruction filter and, when choosing K as a radial Gaussian filter, it has just one aperture parameter.The desired manifold is induced by the "ridges" of the function D considered as a hyper-surface over $\mathbb{R}^N$. Ideally, the ridges would be of constant "height". However, because of varying sampling densities and noisy sample points, the height of ridges changes along a ridge thus inducing so-called critical points, i.e. local extrema and saddle points of varying index. The critical points, as well as the unknown algebraic topology of the hyper-surface represented by the given sample points, makes fitting of explicit surface prototypes inappropriate. Our approach is based on Morse theory by considering D as a Morse function [3]. Thus, the next step is the construction of a Morse-Smale-like complex of D.

A Morse-Smale complex is a cell complex [1, 2]. Let ∇D be the gradient of D and let $\gamma(t)$ be an integral curve, i.e. a curve tangential to ∇D at each of its points. Let further



(a) Cell boundaries. (b) Function values of D on the cell boundaries. (c) Stitched solution along with input points.

Figure 2: Recursion.

be $\mathbf{a}_u$ and $\mathbf{a}_l$ critical points of D and $D(\mathbf{a}_u) > D(\mathbf{a}_l)$. Then γ is open ended and has limits

$$\lim_{t\to-\infty} \gamma(t) = \mathbf{a}_l, \quad \lim_{t\to\infty} \gamma(t) = \mathbf{a}_u.$$

Every point in $\mathbb{R}^N$ has exactly one integral curve passing through it and the curve has a set of upper and lower limits $\mathbf{a}_u$ and $\mathbf{a}_l$, respectively. The desired cell complex is composed of cells consisting of all points in $\mathbb{R}^N$ whose integral curves have the same upper limit $\mathbf{a}_u$.

This definition deviates from the usual definition in that a cell is only formed by the upper limits. The reason is that integral curves of points far away from the given point set terminate at points at infinity – i.e. pseudo-points – which cannot be found robustly or distinguished numerically. Furthermore, saddle-points of D cannot be handled numerically in an easy way since the vector field ∇D is always tangent at saddle points. However, they are approximated in a recursion which is performed next by our approach.

The recursion is as follows. Let ∂C^d be the boundary of a cell C^d of dimension d of the cell complex. Observing that ∂C^d is a manifold of dimensionality $d-1$ (fig. 2a) and that D can be computed on ∂C^d, a cell complex can be computed on ∂C^d as well (fig. 2b). That is

$$\partial C^d = \bigcup_i C_i^{d-1}.$$

Hence it is possible to produce cell complexes which are recursive in dimension. Every step of recursion contributes a part to the solution. This process terminates at $d = 1$. Let $\mathbf{p}_{max}$ be the location of a local maximum of $D(\partial C^2)$ on the boundary of a C^2-cell. It is the location of a saddle-point of $D(C^2)$ and lies on a separatrix of $D(C^2)$. The separatrix intersects the desired surface orthogonally. Because, therefore $\mathbf{p}_{max}$ is a point of the solution, the upward part of its integral curve γ in C^2 must be, too, since it follows the ridge. This process can be continued in a C^3-cell of which the C^2-cell is a boundary. Starting from all points of the integral curve γ and following the integral curves of that points in C^3. In general the d-solutions are concatenated by integral curves

emanating from the $(d-1)$-solutions. All points visited are part of the solution. Finally, the concatenation of two cells is done by connecting each point to the nearest neighbor of the adjacent cell solution (fig. 2c).

3 Discussion

Implementation of the method requires a finite representation of the geometric structures. In two dimensions ($N = 2$), the vertices of a regular grid placed in the region of interest, e.g. an axis-aligned rectangle, are chosen as starting points for the integral curves (1b and 1c). The implementation for $N \leq 2$ has been successfully performed. Ongoing work focuses on $N = 3$ and application to real-life data, e.g. 3D scan data. We hope to gain insights leading to a general and generic approach based on N-dimensional grids also for $N > 3$.

Regular grids are a canonical choice to retrieve the topological structure of D. However, they have the drawback of high memory requirements. A point set P on an N-dimensional hyper-cube grid of M points in every dimension requires $|P| \cdot M^N$ evaluations of K and memory of size $\mathcal{O}(M^N)$.

Furthermore, while the N-dimensional step can be performed on a hyper-cube-grid, the $(N-1)$ cell boundary manifolds are no longer an $(N-1)$-hyper-cube grid. They are graphs on N-dimensional grid nodes (see 2a). This holds for all subsequent steps as well. So, data structures have to be designed which scale in dimension for all subsequent downward steps while at the same time facilitate concatenation of particulate solutions. Ideally, this would not further increase the memory requirements.

References

[1] Herbert Edelsbrunner, John Harer, Vijay Natarajan, and Valerio Pascucci. Morse-smale complexes for piecewise linear 3-manifolds. In *Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, SCG '03. ACM, 2003.

[2] Herbert Edelsbrunner, John Harer, and Afra Zomorodian. Hierarchical morse complexes for piecewise linear 2-manifolds. In *Proceedings of the Seventeenth Annual Symposium on Computational Geometry*, SCG '01, pages 70–79, New York, NY, USA, 2001. ACM.

[3] J Milnor. *Morse Theory*. Princeton University Press, 1963.

Towards Resource-constrained Convolutional Neural Networks for Mobile Biosensing

Jan Eric Lenssen
Lehrstuhl für Graphische Systeme
Technische Universität Dortmund
janeric.lenssen@tu-dortmund.de

Convolutional neural networks (CNNs) are an essential method in the field of deep learning, achieving award-winning performance in several image processing tasks. Despite being computable on graphic processing units (GPUs) efficiently, CNNs often require more graphics memory and computing power than mobile GPUs provide, especially when high-resolution images shall be processed in real-time. Our research focuses on reducing runtime and memory consumption of CNN inference as well as on finding approximations of network computation that lead to similar results. We apply these resource-constrained CNNs on image data gathered by the PAMONO-sensor, a mobile biosensor for nano-object detection. Preliminary results described in this work consists of a platform-independent, fast and memory-saving CNN inference library based on OpenCL, as well as results for nano-particle size distribution estimation from PAMONO-sensor data with real-time capable CNNs.

1 Platform-independent CNN Inference Library

We developed a new CNN inference library instead of using existing deep learning frameworks like Caffe, Tensorflow or Theano because of three main reasons: First, most of these frameworks rely on CUDA as backend, which is not platform-independent and not applicable on most mobile systems. Second, to provide efficient inference algorithms, we need access to the lower levels of the GPU implementation. In CUDA however, all

low level neural network computations are encapsulated in the CUDNN library, which is not open for modifications. Third, these frameworks focus on training performance. Redundant information is processed and kept in memory during inference, for example intermediate results for gradient computation.

Our library follows a data flow graph paradigm and works as follows: First, a (circle-free, directed) graph from a set of implemented nodes is defined. This can be done manually, by importing a trained tensorflow model or by a combination of both. Then, an automatic scheduler is run, determining order of execution, the creation and assignment of a minimal number of required OpenCL buffers and OpenCL specific parameters. Finally, the prepared graph can be applied to input images. All steps before the last can be executed in a preprocessing step, once for each graph. Currently, the library contains convolutional neural network nodes (convolutional layer, pooling, fully-connected layer, softmax layer, normalization layer), pre-/postprocessing nodes (transformations, region of interest extraction, fixed image filter application, up- and downscaling) and some basic mathematical operations (addition, multiplication, modulo, max/min, argmax/argmin). New nodes can be added by implementing a fixed set of functions. In general, every algorithm that can be described by OpenCL kernels and maps a fixed-sized 3-dimensional input tensor to a 3-dimensional output tensor can be added.

The library is implemented in plain C and provides interfaces for C and C++. Dependencies are only *Glib 2.0* and *OpenCL 1.1*, allowing it to run on a large number of different mobile and desktop platforms.

2 Real-time Analysis of SPR Signals using CNNs

The PAMONO-sensor is a biosensor that allows the detection of nano-particles (e.g. viruses, virus-like particles or fine dust) utilizing the effect of Surface Plasmon Resonance (SPR) [6]. A suspension with nano-particles is piped along a gold surface, to which specific antibodies are applied. A laser lights the gold surface while a CCD-Camera captures images of the surface over time. At places of particle bindings, the refractive index of the surface changes, which leads to intensity steps and plasmon waves at points of particle binding in the captured image sequences. Theses wave patterns can be detected and classified using methods from image processing.

We complemented our existing real-time image processing pipeline (c.f. [3]) by two different convolutional neural networks: a fully convolutional network (c.f. [4]) for signal detection and a convolutional network (c.f. [1]) for particle size classification [2,5]. Both CNNs are integrated into the pipeline using the library described in Section 1. The whole pipeline, including example images after different steps, is shown in Figure 1.

The FCN for particle detection solves a binary, pixel-wise classification problem. Our network outputs a map of binary softmax distributions, containing probabilities for the

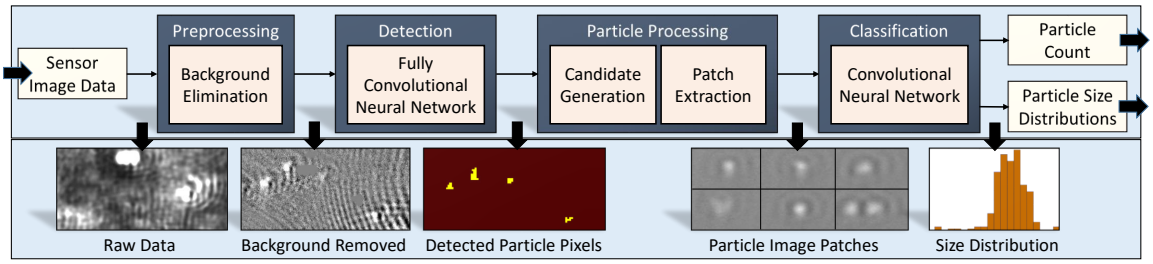


Figure 1: Our real-time GPU image processing pipeline containing both CNNs.

two classes. For the size estimation network, we found that formulating the problem as an ordered regression problem (using a classification network with size intervals as classes) provides better results than using a regression network with one output value. Therefore, the classifying neural network computes softmax over the 27 output classes, where each represents a particle size interval of 10 nm.

Training of the CNNs is performed using synthesized, labeled data. Training data and the binary ground truth for the segmenting FCN is synthesized using the PAMONO signal model provided by previous research [7]. We evaluate the network using manually labeled real PAMONO image sequences. For the size classification network, training data acquisition is more complicated, since we do not have exact labels for real SPR signals. Therefore, we developed a synthesis procedure to generate training data using random sampling and interpolation of signals.

Figure 2 shows a part of our distribution estimation results for whole image sequences [2]. This is the first time, particle size distributions could be automatically extracted from SPRi sensor data [5]. The bottom row contains our pipelines outputs for four different PAMONO image sequences. In the top row, the reference measurement using a LM10 device (Malvern, UK) is shown. It is evident that information about particle size could be extracted and used for precise classification of particle size.

3 Conclusion and Future Work

We provided preliminary work in the field of resource-constrained convolutional neural network that provides infrastructure for further research and evaluations. These includes a platform-independent library for CNN inference on GPUs, which is open for low-level algorithmic improvements. Also, we showed that our library and CNNs in general are able to provide contributions in the field of medical biosensing and extend our existing image processing pipeline for PAMONO data. In the future, we will focus mainly on methodological aspects of fast, memory-saving CNN inference. The results will be used, amongst other things, to further improve the analysis of PAMONO-sensor data.

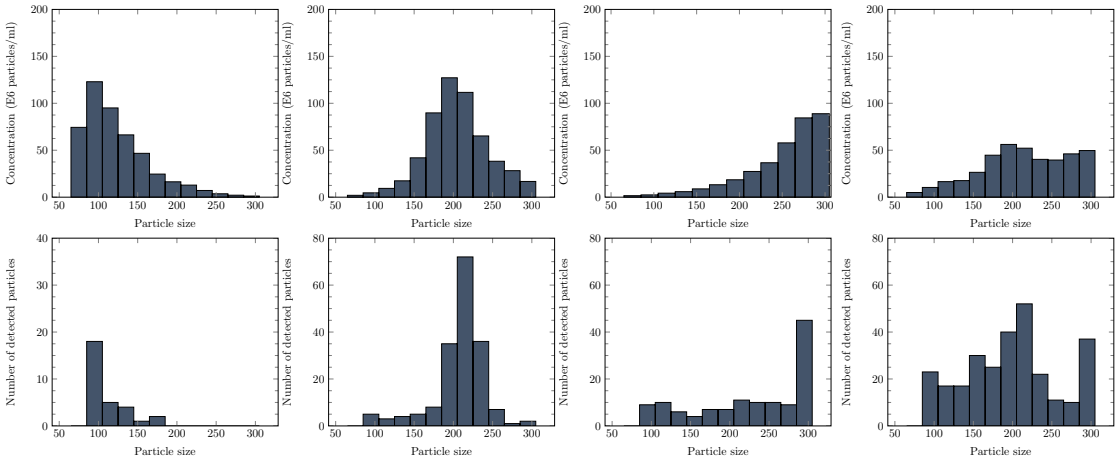


Figure 2: Reference measurements (top) and our results (bottom). Particle sizes (from left to right): 80 nm, 200 nm, 300 nm and a mixture of these three.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[2] Jan Eric Lenssen, Victoria Shpacovitch, and Frank Weichert. Real-time virus size classification using surface plasmon resonance and convolutional neural networks. *Manuscript under revision*, 2017.

[3] Pascal Libuschewski, Dominic Siedhoff, Constantin Timm, Andrej Gelenberg, and Frank Weichert. Fuzzy-enhanced, real-time capable detection of biological viruses using a portable biosensor. *Biosignals*, pages 169–174, 2013.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3431–3440, 2015.

[5] Victoria Shpacovitch, Irina Sidorenko, Jan Eric Lenssen, et al. Application of the pamono-sensor for quantification of microvesicles and determination of nano-particle size distribution. *Manuscript under revision*, 2017.

[6] Victoria Shpacovitch, Vladimir Temchura, Mikhail Matrosovich, et al. Application of surface plasmon resonance imaging technique for the detection of single spherical biological submicrometer particles. *Analytical Biochemistry*, 486:62 – 69, 2015.

[7] Dominic Siedhoff, Pascal Libuschewski, Frank Weichert, Alexander Zybin, Peter Marwedel, and Heinrich Müller. Modellierung und optimierung eines biosensors zur detektion viraler strukturen. *Bildverarbeitung für die Medizin*, pages 108–113, 2014.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B3

Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Domain Knowledge and Data Analytics of Industrial Processes

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

When we try to improve industrial processes solely by an analysis of sensor data, we miss an important source of data and information to simplify the feature extraction and the the model learning and to improve the overall prediction quality.
Information about the structure of the machine, internal states of components and commands, the relations between components and commands and the execution order of commands can help in combination with the state of the to extract and select better features and to build better models. Without this information, one would have to execute a brute force analysis of the correlations of all possible intervals of the time series data or the extensive use of domain knowledge, which would restrict the application of the analysis process to the given industrial process.

Nevertheless, domain knowledge about how the machinery is executing the given tasks and the primary physical process, helps to increase the quality and reduces the possibility

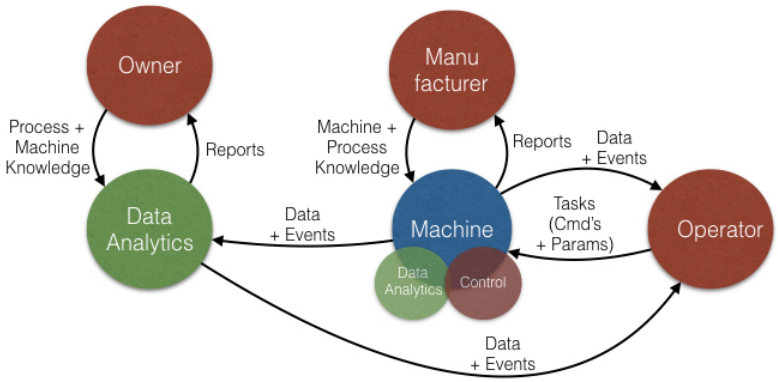


Figure 1: Roles of a common data analysis process for industrial processes.

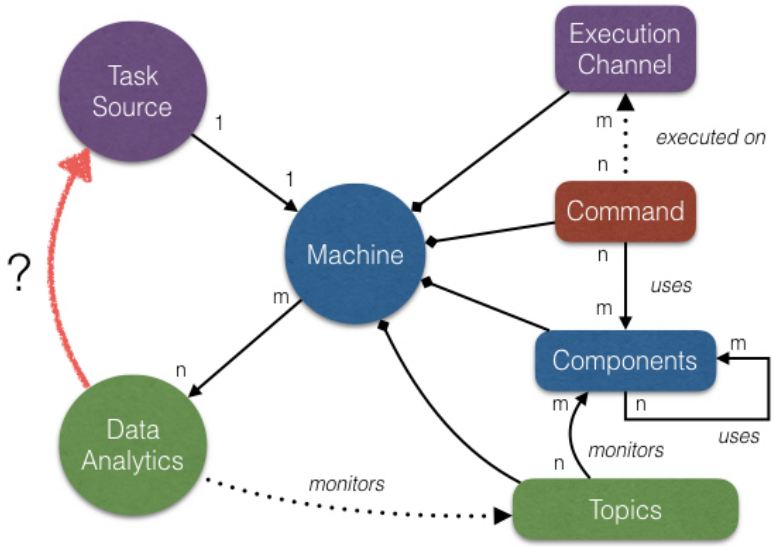


Figure 2: Machine model

of errors. Sadly, the domain knowledge is distributed between different roles that are involved in the operation of the industrial process (cf. fig. 1).

The owner of the machinery is usually a managed entity, like a group or department in a company or institution. It will define the requirements for the specific industrial process. Depending on the nature of the entity, there will be theoretical knowledge about the physical process and the maintenance and the operation of the machinery. In addition, experienced operators have a deeper understanding of the operation and maintenance of the machine, but usually no theoretical knowledge about the physical process or the machine control. Also, there is commonly no knowledge management process implemented to collect and preserve the knowledge of the operator and combine it with the domain knowledge of the entity. The third large set of domain knowledge is only available to the manufacturer of the machinery. A deeper understanding of the machine, it's components and the internal state is only possible with direct access to the components and the machine control. On the other hand, knowledge about details of the executed process (input materials, target quality etc.) are usually limited to the customer.

Let's assume, we would have access to all of the domain knowledge and the internal state of the machine. How would we integrate it into the data analysis and what are the advantages?

machine and execution model Even though there a hundreds of different manufacturing processes with specific supply chains or networks and processing steps, we will concentrate on a single processing step and the corresponding machine. Let's assume a simple

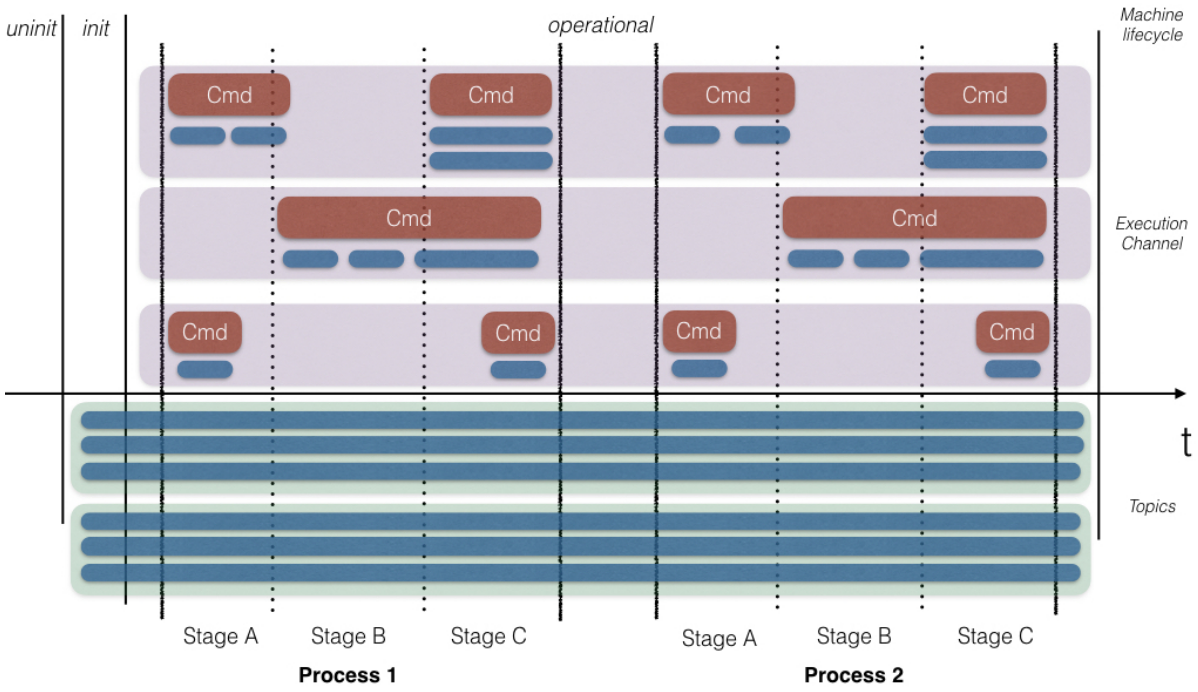


Figure 3: Example of a machine execution and the corresponding states of the machine.

machine model (cf. fig 2) to describe the operation of the process.

The executed process is a sequence of tasks, that is provided by a task source. The task source could either be an operator or another machine. Every task describes a command with parameters. The command parameters and the base parameters of the components describe the reference state of the machine. Every command will use a set of components to execute on an execution channel, which describes an execution pipeline. Topics will collect sensor data, meta data, state data and events to analyze the behavior of the components and commands. Every topic has it's own sample rate. The data analytics infrastructure will only have access to the data streams that are produced by the topics.

Advantages

The advantages of this model can be described better, when we have a look at an example of the execution of multiple processes on a machine (cf. fig 3).

Every machine has a life cycle. Severe Problems will usually occur directly in the initialization phase or in the operational phase of the machine. For the sake of simplicity, we define the machine as operational when all commands and components are operational. When at least one execution channel is executing an command, the state of the machine is busy. The search for failure reasons is largely simplified by this state information, be-

cause we can model the expected behavior of the components in the different stages of operation.

It also simplifies to identify the different stages of the process. As we can see in fig. 3, it is assumed that the process and the stages of the process are defined by the start and the end of a command. With this information, the data analytics could directly differentiate between process synchronous and asynchronous sensor streams, and extract features for every process step automatically. This knowledge can either reduce the required domain knowledge or simplify the integration of domain knowledge. This will reduce the required efforts for the preprocessing, feature extraction and possibly the selection of features of the specific industrial process.

When the data analytics infrastructure has access to the states and scheduling of tasks, it could automatically differentiate between different operation modes and types of processes. In combination with sensor data about the used materials and unique hardware identifications within the meta data, the data analytics infrastructure would have the capability to model the expected behavior of the machine and possible behavior drifts of the operation.

It would also be possible to differentiate between expected and unexpected behavior of components. If a component shows an unexpected behavior, the data analytics framework would directly know that either something is influencing the component or a failure inside the component is happening. Either the processed material behaves differently, and therefore the complete process, or there is something in the environment of the machine, that is influencing the operation.

When quality metrics are measured inline and are also available, the learning of quality prediction models would be simplified, too. It would be possible to identify the most important stages of the process and commands and therefore the most important features.

conclusion
Hopefully, the description of the machine and execution model could help to show the importance and advantages of a deeper integration of the data analysis, machine control and domain knowledge of all involved parties of the industrial process. It should also be obvious, that all involved parties could benefit. The machine manufacturer can improve the maintenance cycles and construction of machines for the most common use cases of the machine. The operator of the machine could get better insights into error and warning messages. The owner of the process could better optimize the average process and product quality of the industrial process.

Product family based line balancing

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

Assembly Line Balancing (ALB) approaches aim at optimally assigning tasks to workstations. While traditional ALB solution procedures lead to increasing idle times and therefore inefficiencies in high mix, low volume production settings, alternative approaches have to be designed. This paper presents an approaches for family based line balancing as a solution to this question.

1 Assembly Line Balancing

The Assembly Line Balancing Problem (ALBP) is the problem of assigning assembly tasks to the stations of an assembly line. Research on ALBP can be traced back to Helgeson et al. [5] and Jackson [6] who present the first formulation of the problem setting and a first solution procedure, respectively.

In general, ALBP can be differentiated with respect to the complexity of the problem statement into Simple Assembly Line Balancing Problems (SALBP) and General Assembly Line Balancing Problems (GALBP). SALBP are focusing assembly lines on which only a single product is assembled according to a well-defined list of tasks. The corresponding assembly line consists of serially linked line station, all of which are identical, such that each task may be assigned to each station. Moreover, each assembly task's duration is deterministic and independent of the sequence of the tasks' execution. [1] If any of these conditions is relaxed, the problem is called a GALBP. Most relevant in terms of practical application of ALBP is its extension to those assembly lines on which a wide variety of different products or product variants are produced. These types of GALBP are referred to as Multi Model Assembly Line Balancing (MuMoALBP) or Mixed Model Assembly Line Balancing (MiMoALBP). While in MuMoALBP assembly is organized in

batches with a batch consisting of identical products, MiMoALBP allows to assembly all different products or product variants in an arbitrary order. [1]

In order to solve MiMoALBP it is common practice to reduce the problem to the Single Model Assembly Line Balancing Problem or if passible to SALBP. This is achieved by combining all different products or variants. In line balancing products are represented by their precedence graph. Consequently, combining the different products requires deriving the joint precedence graph of all products. In order to account for each product's individual demand, tasks are weighted according to the demand. [2] By this, the variety of different products is reduced to one averaged one. Once the MiMoALBP is reduced to the single model one, solution approaches of the latter are applied. A comprehensive review of these can be found in [1].

Although this approach allows for efficiently solving the problem stated, the line balance computed is only efficient if customer orders are similar to the average product. As customer demand tends to be erratic and varies significantly according to volume as well as mix, the computed line balance might be almost arbitrarily efficient or inefficient. Some approaches address this issue: Gel et al. [4] present a dynamic approach to assembly line balancing in which tasks are split into fixed and shared ones. While fixed tasks are imperatively conducted by the operator of the very station, shared tasks may be executed by an operator from a neighboring station. Reinhart and Proepster [7] address flexibility in planning and organization of assembly lines by a scenario driven approach to account for type-volume-mix changes. To further increase flexibility, variant-specific work stations a introduced.

2 A product family based approach for line balancing

In contrast to the approaches introduced previously, Deuse et al. [3] as well as Siedelhofer et al. [8] propose to combine MiMoALBP and MuMoALBP to add to balancing efficiency. The underlying principle is to analyze product and process related data in order to define homogenous product families for which line balancing can be computed. Instead of creating a balance for all products, family-specific balances are derived with each family-specific balance being a MiMoALBP.

As said before, different products might require individual assembly tasks having an individual work content, station equipment, material etc. Consequently, homogeneous product families contain products with resemblance in these characteristics. In case of required tasks, equipment or material, structural resemblance of different products is defined by the presence or absence of each characteristic, i.e. binary data. On the contrary, work content is measured in time units, which adds ratio scaled variables. Preprocessing these different types of data is the mandatory first step in establishing homogeneous product families for balancing. Secondly, outliers are identified and removed. In this context,

outliers are those customer orders that differ significantly from other orders with respect to tasks performed or parts needed so that they cannot be produced together with other orders efficiently on a flow line. The subsequent step in preprocessing aims at reducing redundant information. Thinking of the data described before, it is obvious that many attributes are highly correlated. E.g., pre-assembling a standardized module such as a dashboard will always call on the same set of tasks regardless the exact specification of the remaining car, while other data on the module might still vary (e.g., color of the dashboard representing different material numbers). Correlated attributes are identified and removed using Principle Component Analysis (PCA) with a variance threshold of 0.95. In following grouping process k-Means Clustering is employed. The number of clusters k for which clustering is performed is determined based on by assessing the results from hierarchical clustering. The majority vote of different algorithms (Single, Complete and Average Linkage as well as WARD) for the 'ellbow' yields the expectation for the optimal number of clusters k*. To account for specific influences of the case study, k-Means is conducted iteratively on an interval [k*-n, k*+m]. The clustering results are evaluated using a desirability index which contains different quality measures such as homogeneity of clusters built or size of clusters. The labeled customer orders, each one assigned to exactly one cluster, is passed on to the line balancing methodology.

As in the state-of-the-art approach to MiMoALBP combining the different precedence graphs into a merged one is the first step of the methodology, but here this step is done for every family. Consequently, each product family can be treated as a single MiMoALBP, while the overall view is similar to the MuMoALBP. The second step of the methodology aims at solving the MiMoALBP for each product family. This is accomplished conducting the following heuristic approach: firstly, an initial balancing solution is generating for a given takt time using a single pass, best-fit greedy heuristic. Starting off from an empty assembly line, it will keep on opening new assembly stations as long as not all assembly tasks are assigned to exactly one station. In each iteration a set of candidate tasks is generated with regard to the precedence constraints. Each candidate task is evaluated according to the following three criteria: (a) maximization of current station's utilization, (b) number of successors and (c) maximization of current station's utilization in combination with a successor? The task with the highest score is selected and assigned to the station. Upon completion of the initial solution, an improvement heuristic is executed that incorporates guided local search to first minimize the number of stations required (and therefore maximizing the utilization) and second optimize the balancing result with respect to horizontal and vertical smoothness. By this, for each product family an optimum solution is computed. Yet, the sum of these group-based optima might not yield the overall optimum line balancing, when paying attention to change over costs to switch the line setup from one family to another one. Striving for an overall optimum balancing solution, a third heuristic is launched that tries to find the best compromise of locally best line efficiencies per family and lowest change over cost for a cycle in which each family is exactly respected once. To achieve this, the heuristic firstly identifies those line balances of individual product groups that are similar to each

other. Afterwards, each of these sets of line balances is optimized separately using Tabu Search. To guide the search, a weighted average line balance for all line balances in the determined set is computed. This serves as a reference and the algorithm tries to adjust each line balance according to the average.

3 Conclusion

The sketched approach provides a powerful tool to increase line efficiency of mixed-model assembly lines for large numbers of individualized products. Though it relies on the data that is required for state-of-the-art balancing approaches it requires more effort in preprocessing this data to establish product families. But first results from case studies show the potential to significantly reduce losses due to low line efficiency.

References

[1] I. Baybars. A survey of exact algorithms for the simple assembly line balancing problem. *Management Science*, 32(8):909–932, 1986.

[2] N. Boysen, M. Fliedner, and A. Scholl. Assembly line balancing: Joint precedence graphs under high product variety. *IIE Transactions*, 41(3):183–193, 2009.

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

[4] E. S. Gel, W. J. Hopp, and M. P. van Oyen. Factors affecting opportunity of work-sharing as a dyn. line balancing mechanism. *IIE Transactions*, 34(10):847–863, 2002.

[5] W. B. Helgeson, M. E. Salveson, and W. W. Smith. How to balance an assembly line. *Management Report*, 7, 1954.

[6] J. R. Jackson. A computing procedure for a line balancing problem. *Management Science*, 2(3):261–271, 1956.

[7] G. Reinhart and M. Pröpster. Handlungsfelder der dynamischen austaktung: Ein prozess zur reaktion auf nachfrageschwankungen im nutzfahrzeugbau. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 107(6):404–408, 2012.

[8] C Siedelhofer, T. Henke, B. Konrad, J. Deuse, and J. Litterscheidt. Konzepte zur effizienzsteigerung von variantenfließlinien bei hohen variantenanzahlen. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, (10):594–597, 2016.

Distributed Analysis of Vertically Partitioned Sensor Measurements under Communication Constraints

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

Nowadays, large amounts of data are automatically generated by devices and sensors which form the Internet of Things (IoT). Understanding such data creates opportunities to substantially improve our way of living. For instance, the information obtained from embedded data analysis could help the optimization of processes, leading to more sustainable systems saving resources in sectors such as manufacturing, logistics, energy and utilities, the public sector, or healthcare. However, IoT's inherent distributed nature, dynamism and resource constraints challenge even the most advanced automated data analysis methods know today. Especially challenging are highly communication-constrained settings which require decentralized algorithms that process most of the data on the generating devices themselves. Over the last few years, we have developed different approaches and algorithms for the scenario which are presented in the following.

1 Introduction

The collection and transmission of data in the IoT is enabled by electronics, software, sensors and network connectivity embedded into physical objects. Devices measure, for instance, parameters of production processes, environmental conditions of transported goods, energy consumption of smart homes, traffic volume, air pollution and water consumption, or pulse and blood pressure of individuals. In 2010, already 12.5 billion devices

were connected to the IoT, a number about twice as large as the world's population at that time. The biggest benefits may come from the analysis of data from heterogenous sources (see [1]). However, the combined data may pose a big data problem. Currently, there is a strong research focus on the centralization of all data in the cloud, processing it according to the paradigm of parallel high-performance computing. However, the resources of devices and sensors at the data generating side might not suffice to transmit all data. For instance, pervasive distributed systems such as wireless sensor networks are highly communication-constrained, as are streaming high throughput applications. Hence, the IoT requires a new generation of distributed algorithms which are resource-aware and intelligently reduce the amount of data transmitted and processed throughout the analysis chain. The scenario of vertically partitioned data is particularly challenging. Here, not observations, but their feature values are stored across different physical nodes. The learning of accurate prediction models may thus require the combination of information from different nodes. In the following, we present algorithms for the scenario which are communication-efficient, at the same time yielding sufficient accuracy.

2 Local Preprocessing of Value Series

In a hot rolling mill case study from the smart manufacturing sector, different physical machines assess parameters describing the processing of single steel blocks. Based on such distributed sensors measurements, the quality of individual blocks should be predicted in real-time as early as possible. Each machine creates not single value series, but many of them. Their heterogeneity leads to challenging questions concerning the steps of preprocessing and finding a good representation for learning. The proposed solution [2] extracts statistical information from different segments of the value series and projects it to fixed-length propositional feature vectors. Thereby the number of up to 60,000 raw measurements per block is reduced to only 218 features, speeding up all following operations. Meaningful patterns, like processing modes for the rolling of different block sizes, can still be extraced from the reduced data, for instance by dimensionality reduction or clustering. Each value series can be preprocessed independently from other series. Since statistics like mean and variance may be calcuted in a streaming fashion with a few number of addition and multiplication operations, a local preprocessing at sensor nodes seems feasible.

3 Learning from Label Proportions by Clustering

In the aforementioned case study, quality information is usually not given for individual blocks, but charges of blocks. How can we nevertheless predict the quality of individual

blocks? Learning from aggregated label information is a relatively novel problem in machine learning research. We propose the Learning from Label Proportions by Clustering (LLPC) algorithm [3]. It first clusters the dataset and then searches for an assignment of class labels to clusters which minimizes the distance between the model-based and given label proportions. Bias and variance can be traded off against each other by varying the number of clusters k, which is an improvement over existing clustering algorithms for the problem. The algorithm's performance is compared to three other state-of-the-art approaches, in terms of accuracy and running time. It can be shown that LLPC results in lower asymptotic (linear) running time, while accuracy is comparable to that of its competitors, or significantly higher. The proposed algorithm comes with many other benefits, like ease of implementation and a small memory footprint. It therefore has a much better chance than existing solution to run on small resource-constrained devices.

4 Training of Local Models from Label Counts

Time constraints, as in smart manufacturing, lead to questions typical for the vertically partitioned data scenario. Which data should be analyzed locally, to match the constraints, and which should be sent to a central server? For highly decentralized systems, like wireless networks of topologically close sensors, the Training of Local Models from (Label) Counts (TLMC) algorithm is proposed [4]. The method builds on LLPC, reducing communication by transferring only label counts for batches of observations. For predicting a target value at node j, local LLPC models are trained on the sensor measurements of a fixed number of topologically close neighboring nodes. Feasibility of the approach is demonstrated by evaluating the algorithm's performance in the context of traffic flow prediction. It is shown that TLMC is more communication-efficient by about an order of magnitude in comparison to the centralization of all data, but that accuracy can nevertheless compete with that of a centrally trained global model.

5 Vertically Distributed Core Vector Machine

Finally, a communication-efficient distributed algorithm for anomaly detection is proposed, the Vertically Distributed Core Vector Machine (VDCVM) [5]. It improves on the vertically distributed 1-class SVM by replacing the central model with the Core Vector Machine (CVM), which incrementally samples only as many batches of observations as needed to reach a $1+\varepsilon$ solution with high probability. Using the CVM with a combined RBF kernel further allows for distributed kernel calculations across local nodes, resulting in the transmission of a single scalar value per observations and current model parameters per iteration, instead of all feature values. Despite a growing number of model parameters, the proposed algorithm communicates up to an order of magnitude less data

than centralization as long as the core set stays small enough. Despite the reduction, the VDCVM achieves similar or even higher accuracy on several controlled and benchmark datasets than existing less communication-efficient solutions.

6 Conclusions

Communication-efficient learning is possible in cases where features from different nodes are conditionally independent, given the target value to be predicted. Most efficient are local preprocessing and the training of local models, which exchange label information instead of measurements between nodes. Communication can be even further reduced by learning from the counts of labels. In the context of traffic flow prediction, the accuracy achieved is still sufficient in comparison to centralizing all data and training a global model. In the case of anomaly detection, similar results can be achieved by utilizing core sets and distributed kernel calculations.

The developed approaches have many applications in communication-constrained settings, in the sectors mentioned above. It has been shown that data can be reduced and analyzed already before it enters the cloud. Decentralized processing might thus enable the analysis of big data masses when centralization of all data becomes infeasible.

References

[1] M. Stolpe. The Internet of Things: Opportunities and Challenges for Distributed Data Analysis. *SIGKDD Explorations)*, pages 15–34, 2016.

[2] M. Stolpe, H. Blom, and K. Morik. Sustainable Industrial Processes by Embedded Real-Time Quality Prediction. In *Computational Sustainability*, pages 201–243. Springer, 2016.

[3] M. Stolpe and K. Morik. Learning from Label Proportions by Optimizing Cluster Model Selection. In *Proc. of the ECML/PKDD*, pages 349–364. Springer, 2011.

[4] M. Stolpe, T. Liebig, and K. Morik. Communication-efficient learning of traffic flow in a network of wireless presence sensors. In *Proc. of the PDCKDD 2015*, CEUR Workshop Proceedings, page (to appear). CEUR-WS, 2015.

[5] M. Stolpe, K. Bhaduri, K. Das, and K. Morik. Anomaly Detection in Vertically Partitioned Data by Distributed Core Vector Machines. In *Proc. of the ECML/PKDD*, pages 321–336. Springer, 2013.

Automating Ultrasonic Inspections by Means of Supervised Learning

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

This report presents a concept for automating ultrasonic inspections to detect material defects in rolled steel rods. For this reason ultrasonic scans as well as process and product data are preprocessed and used as an input for a multiclass classification task with multiple labels representing various quality parameters. In this report the classification task is conducted by the use of standard learning algorithms for training individual models for each label. Prospectively, algorithms for structured output learning should be applied to train one single model for predicting all labels.

1 Introduction

Within the Collaborative Research Center 876 project B3 focuses on the development of algorithms for the time-constrained analysis of sensor data by means of data mining. The analysis of recorded data is conducted using the example of hot rolling processes in the steel industry. In hot rolling processes technological and organizational restrictions obstruct physical quality inspections of intermediate products [2]. The objective of project B3 is to utilize process and product data to train a model for assigning quality information to each steel rod at an early stage of the manufacturing process. With the help of this information process control decisions could be derived to increase the first pass yield. In the first period of the SFB 876 the prediction of a binary label (OK/ not OK) was focused [2]. In the second period more detailed quality information should be assigned to every steel rod to allow for more advanced control decisions. Therefore,

detailed properties of material defects, like type, dimensions and intensity of defects, are used as multiple labels.

To determine the inner quality of a steel rod an ultrasonic inspection is conducted at the end of the process chain [1]. Commonly, this inspection is conducted semi-automatically. While the test itself is usually executed in an automated plant, test results (i.e. ultrasonic scans) are analyzed manually. This report presents a concept for assigning various quality parameters to each steel rod by processing ultrasonic test data as well as product and process information by means of supervised learning. With this approach it will be possible to automate the time-consuming manual analysis of ultrasonic test results.

2 Ultrasonic Testing

Ultrasonic testing is a non-destructive testing method for detecting inhomogeneities in materials by transmitting and measuring ultrasonic waves. The most common ultrasonic testing technique is the pulse-echo method where the probe acts as pulser and receiver. Initially, the probe sends ultrasonic pulse-waves into a test object. The waves are partially reflected at the interface of two materials in the object, such as defects like blowholes or the back wall of the object. The reflected waves are received by the probe and measured and recorded by a transducer. A diagnostic machine displays the intensity of reflection as a function of the arrival time of reflected waves. This graphic display is called A-scan (Figure 1).

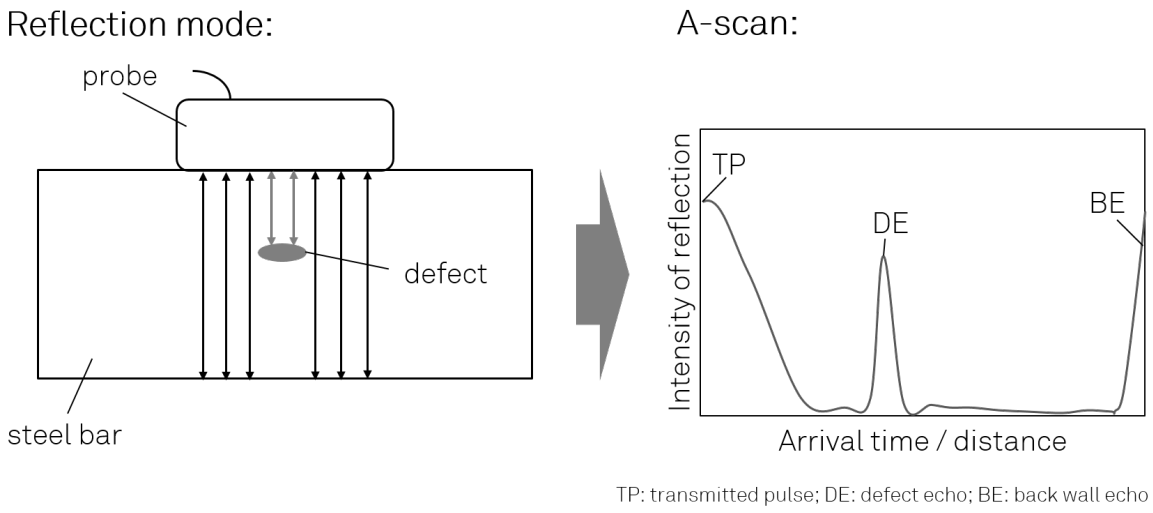


Figure 1: Procedure of ultrasonic testing

3 Supervised learning for automated analysis of ultrasonic test results

Ordinarily, the A-scan is analyzed manually and is used together with additional product and process information to manually create a quality report containing different quality parameters like type of the defect, defect dimensions and defect intensity. In order to automate the analysis and the creation of quality reports we apply supervised learning based on a historical data set. The data set contains A-scans as well as product and process information for each steel rod. We used the manually derived quality parameters in the quality reports as labels. In a first step we cleansed the data and extracted relevant features. For this purpose we filtered out the initial pulse of the probe and the back wall echo as well as noise. Furthermore, we shortened A-scans with multiple back wall echoes due to technological properties of the testing method to ensure the examples' comparability. Based on the cleansed data set we extracted statistical values like mean, variance and median of the A-scan as well as number, value and slope of the peaks and additional characteristics.

Subsequently, we applied different base learners to train individual models for each quality parameter to have a reference for more advanced methods. We used Support Vector Machines, Decision Trees, Random Forest, Naive Bayes and Neural Nets with various parameter settings for each method. For finding the best parameter settings and evaluating the models' final performance we executed one 10-fold cross validation each. In order to improve accuracy, we selected the most relevant features by an evolutionary feature selection. Our experiments were conducted in the RapidMiner analytics platform.

As a result, all of the base learners performed well on the data set and they only showed small performance differences. By using the defect's type as label we achieved an accuracy between 0.73 and 0.83 with the Support Vector Machine as best learner. The results using other quality parameters as labels were slightly poorer.

The defect's type could be interpreted as an aggregated quality indicator being composed of other quality parameters of the quality report. This implies that dependencies between the quality parameters exist. Hence, we assume that algorithms for structured output learning could improve the accuracy by considering these dependencies. Structured output learning is a family of supervised learning methods which predict structured objects (e.g. a parse tree) instead of scalar quantities. Typical techniques of structured output learning are classifier chains [4], probabilistic classifier chains [3] and the structured support vector machine [5]. In the next step we want to examine, how these techniques could be adapted to our use case and how they compare with the aforementioned base learners.

4 Conclusion and further research

The experiment results are very promising allowing a complete automation of ultrasonic inspections. In our further research we have to analyze the suitability of the techniques of structured output learning in order to gain further performance improvements.

References

[1] Jochen Deuse, Mario Wiegand, Loga Erohin, Daniel Lieber, and Ralf Klinkenberg. Big Data Analytcis in Produktion und Instandhaltung. In Hubert Biedermann, editor, *Instandhaltung im Wandel. Herausforderungen und Lösungen im Zeitalter von Industrie 4.0*, pages 33–48. TÜV Media, 2014.

[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] Elena Montañes, J Robin Senge, Jose Barranquero, José Ramón Quevedo, Juan José del Coz. Dependent Binary Relevance Models for Multi-label Classification. In *Pattern Recognition 47 (3)*, pages 1494–1508, 2014.

[4] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier Chains for Multi-label Classification. In *Machine Learning 85.3*, pages 333-359, 2011.

[5] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun. Large Margin Methods for Structured and Interdependent Output Variables. In *Journal of Machine Learning Research 6*, pages 1453–1484, 2005.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B4

Analysis and Communication for Dynamic Traffic Prognosis

Kristian Kersting Michael Schreckenberg
Christian Wietfeld

Analytical Discussion of Delay Propagation in Urban Public Transport Networks

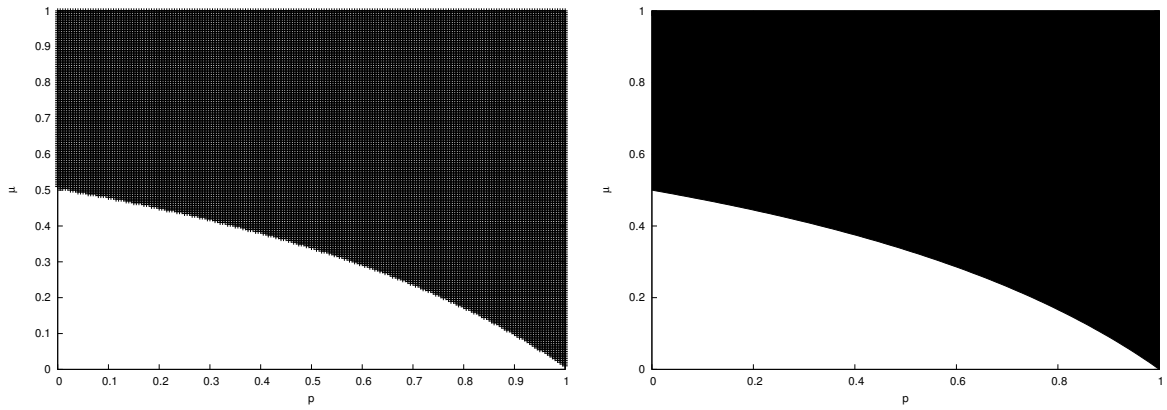
Merlin Becker
Physics of Transport and Traffic
Universität Duisburg-Essen
merlin.becker@uni-due.de

In urban public transport networks, many modes of transport are possible. Especially in rail-based environments, delays can have a long-range impact. The following report deals with the discussion of the delay propagation in such networks using a mean-field approach.

1 Introduction

The mean field theory is widely used to describe the impact of large-scale structures on single elements qualitatively. There are many applications especially in network science. For example the epidemic spreading of diseases has been broadly discussed. [1], [3]

The dynamics of the model are developed in discrete timesteps t. It should not to be confused with delay Δt. The quote of trains with a delay Δt at a timestep t and all trains is $n(t, \Delta t)$. To describe the dynamics of delay propagation we introduced the parameter p, a probability to transmit a delay by a conflicting (i.e. delayed) train. The mechanism of delay reduction μ (e.g. due to buffer times) completes this model. Hence follows formula (1):



(a) Simulated. (b) Analytical.

Figure 1: Comparison of simulation (subfigure (a)) and equation (3) (subfigure (b)). The simulation was driven with $n(t,1) = 20$ and $n(t,0) = 980$. Black-marked sections show combinations of parameters where trains have no delay for $t \to \infty$.

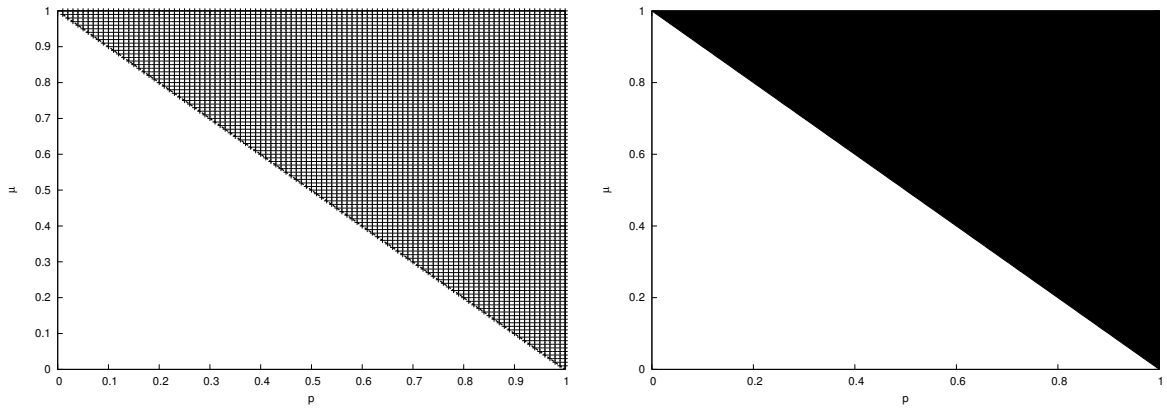
$$\begin{aligned} n(t+1,\Delta t) = {} & \mu[pn(t,\Delta t+1) + (1-p)n(t,0)n(t,\Delta t+1)] \\ & - (1-\mu)(1-p)(1-n(t,0))n(t,\Delta t) \\ & + (1-\mu)(1-p)(1-n(t,0))n(t,\Delta t-1) \\ & - \mu[pn(t,\Delta t) + (1-p)n(t,0)n(t,\Delta t)] \\ & + n(t,\Delta t) \end{aligned} \tag{1}$$

In this model there are no trains earlier than the timetable. That means, there is no $\Delta t < 0$. At the boundary $\Delta t = 0$, formula (2) follows for $n(t+1,0)$.

$$\begin{aligned} n(t+1,0) = {} & \mu[pn(t,1) + (1-p)n(t,0)n(t,1)] \\ & - (1-\mu)(1-p)(1-n(t,0))n(t,0) \\ & + n(t,0) \end{aligned} \tag{2}$$

Case 1: Small Delays Small delays in reference to the timetable are common during daily operation. Simulation of such delays shows, that the influence on the whole systems for times $t \to \infty$ depends on the set of parameters (see Figure 1). It can be shown, that this boundary is described by formula (3) [2].

$$\mu = \frac{1-p}{2-p} \tag{3}$$



(a) Simulated. (b) Analytical.

Figure 2: Comparison of simulation (subfigure (a)) and equation (4) (subfigure (b)). The simulation was driven with $n(t, 35) = 1000$ and $n(t, 0) = 0$. Black-marked sections show combinations of parameters where trains have no delay for $t \to \infty$.

Case 2: Long Delays The simulation shows that, even when every train has a high delay, there are parameter sets, lead to a delay-free system for times $t \to \infty$ (see Figure 2). It follows that the influence of long delays depends on the probabilities of delay-reduction and delay-transmission. This boundary can be described by formula (4) [2].

$$\mu = 1 - p \tag{4}$$

2 Section Model for Delayed Trains

Combining the results from both cases, the section model for delayed trains was constructed. There a three sections (see Figure 3) for times $t \to \infty$. In section 3 there is no intial condition allowing $n(\infty, 0) = 0$, i.e. all trains are on time for large t. Similarly in section 1, there is just one possible result $n(\infty, 0) = 0$. In section 2 it depends on p and μ if trains are on time or delayed for large t.

3 Conclusion

This simple model shows the interaction between delay transmission and delay reduction. Without any modifications there are just two possible system states for $t \to \infty$. All trains can be delayed with an infinitely large delay or all trains are on time. There is explicitly no equilibrium state containing trains on time and delayed trains. Thus there must be some kind of a delay-introducing rate. This will be discussed in a further publication.

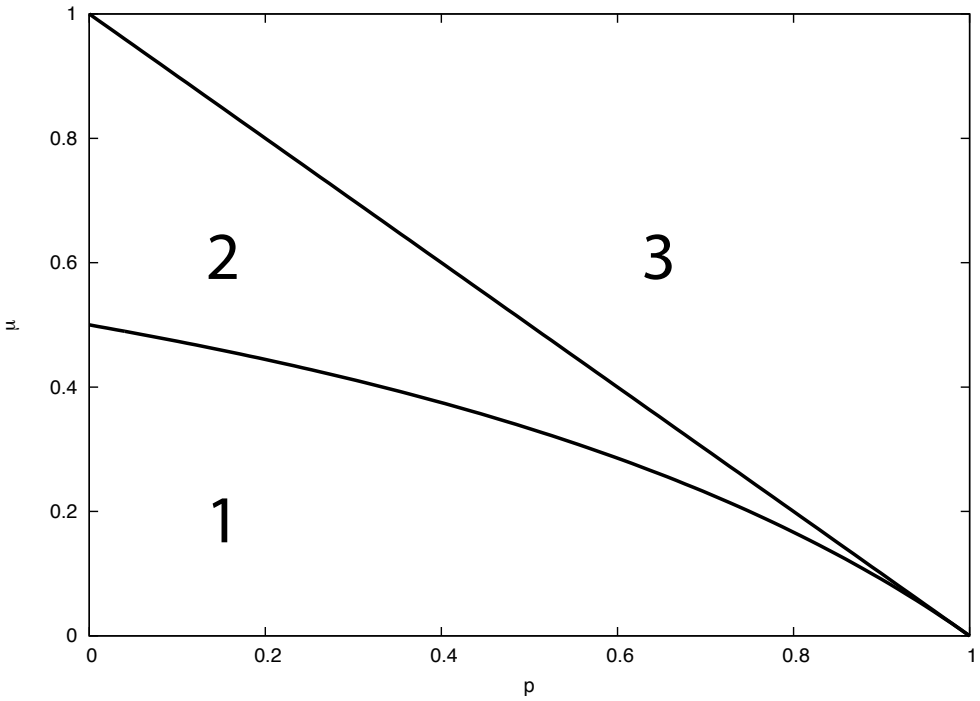


Figure 3: Three Sections of the section model for delayed trains: section 1 contains all sets of parameters which induce a non-reducable number of delayed trains. In section 2, it depends on the parameter set if delays can disappear or not. Section 3 contains all sets of parameters, where all trains are on time at $t \to \infty$.

References

[1] David Barsky and Chris Wu. Critical exponents for the contact process under the triangle condition. *Journal of Statistical Physics*, 91(1-2):95–124, 1998.

[2] Merlin Becker. Dynamische konfliktlösung für die stammstrecke des s-bahn-netzes münchen. *Masterthesis*, 2016.

[3] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

Modeling and Simulation of Heterogeneous Vehicular Traffic Conditions

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

This report summarizes recent research to improve microscopic vehicular traffic simulations under heterogeneous real-world and real-time conditions. As an example, it is shown how breakdowns caused by adverse environmental conditions can be modeled according to the three-phase theory of vehicular traffic.

1 Introduction

Heterogeneous conditions in terms of roads, participants and data availability are typical challenges for vehicular traffic simulations of real-world road networks. In [2], we already addressed dealing with incomplete traffic data at simulation runtime.

To improve simulations under difficult environmental conditions, we recently have proposed a new concept [1] for transmitting environmental *extended Floating Car Data* (xFCD) from driving vehicles as well as stationary measured environmental data to a server, which continuously feeds the data into a cellular-automaton (CA) based microscopic vehicular traffic simulation that we have developed and analyzed in [3, 4].

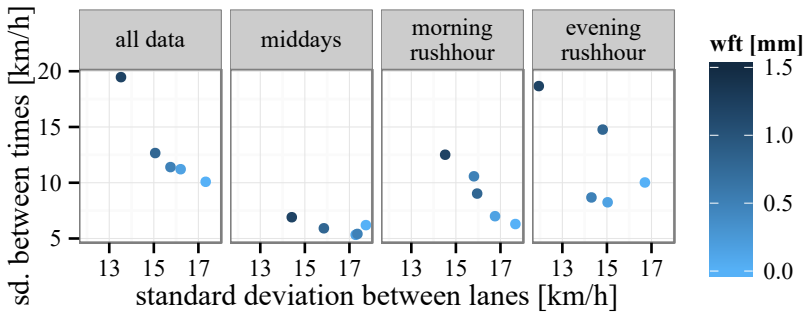


Figure 1: Comparison of standard deviations of the passenger car velocity: Each point represents measurements at a certain water film thickness. From [1].

2 Empirical Analysis

From theoretical approaches in the framework of the three-phase traffic theory, it is known that bad weather conditions can cause traffic breakdowns as other bottlenecks like on-ramps.

Fig. 1 visualizes the empirical relationship between the homogeneity of velocities and their predictability at a given water film thickness. The homogeneity of velocities can be described by calculating the standard deviation of the passenger car velocities from all lanes at a certain point of time and water film thickness. The averages of these values are used as x-value of the points in Fig. 1. High values represent high velocity differences between the lanes, which typically can be found at dry surface conditions. When the road surface is wet, the velocities synchronize like in the synchronized flow phase of the three-phase traffic theory. The predictability of a velocity can be expressed using the standard deviation of velocities measured at different times, but the same water film thickness. Based on the average passenger car velocity of the whole roadway, the y-values of the points in Fig. 1 represent this predictability for the given water film thickness. High values indicate that the set of average velocities contains a wide span of velocities, meaning that the velocity at a certain point of time is less predictable. This is often the case when the surface is wet, especially during the rush hours.

According to the three-phase traffic theory, there is an infinite number of road capacities

Table 1: Maximum vehicular traffic flows depending on surface wetness. From [1].

surface	range of water film thickness w	maximum flow
dry	$w \leq 0.15\,\text{mm}$	1960 vehs/(h, lane)
damp	$0.15\,\text{mm} < w < 0.9\,\text{mm}$	1720 vehs/(h, lane)
wet	$w \geq 0.9\,\text{mm}$	1320 vehs/(h, lane)

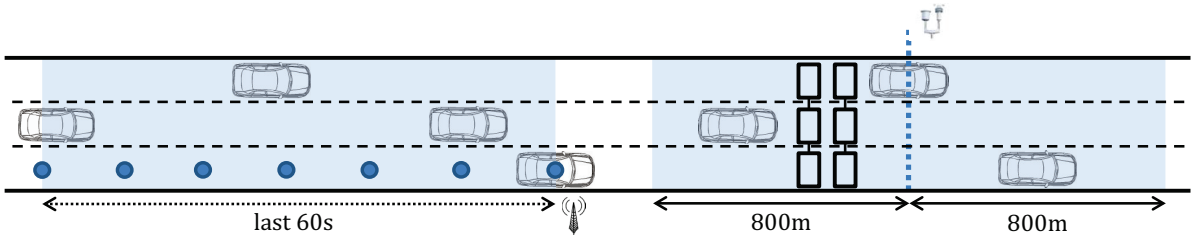


Figure 2: Visualization of different strategies for incorporating environmental data into the simulation. Left: xFCD, right: stationary sensor. From [1].

(i.e. traffic flows measured downstream of a bottleneck) between a minimum capacity where the breakdown probability is 0 and a maximum capacity where it is 1. In Table 1, we show how much the maximum capacity is influenced by surface wetness by means of maximum measured vehicular traffic flows, which have been obtained from free flow traffic within the 11-month data set. These values effectively are lower bounds of the maximum capacities of each type of surface wetness. Because of the limited amount of data and possible detection errors, measuring a maximum capacity itself is not possible. We therefore defined a threshold, meaning that Table 1 only shows flows that were reached more than ten times within 11 months.

3 Simulations

Connection with real-time data For stationary environmental detectors, a fixed area around the simulated detector is sufficient. In the given data set, correlations between neighboring water film detectors stay very high for detector distances between 650 m and about 1000 m. Hence, we choose an extend of 800 m on both sides of the detector. Environmental data is transmitted every 60 s.

In case of xFCD, a simulated moving area behind a transmitting vehicle should be preferred. A transmitting vehicle driving with $v \approx 108\,\mathrm{km/s}$ covers an area of about 1800 m within 60 s. This coverage is in very good agreement in space and time with that of stationary detectors.

Results In [1], we also provide a parametrization $p_w(w)$ for damp and wet road surfaces. Therefore, we performed simulations of traffic on a 10 km long three-lane highway section using the aforementioned rules. In each simulation, a constant inflow of vehicles at the upstream boundary between 600 vehs/(h, lane) and 2400 vehs/(h, lane) is used. 15 % of all vehicles are trucks, which are inserted in toto on the right lane. Between 6750 m and 8250 m, the simulated road is affected by surface wetness as described in [1] after 60 min of relaxation time. The parametrization is obtained by measuring breakdown

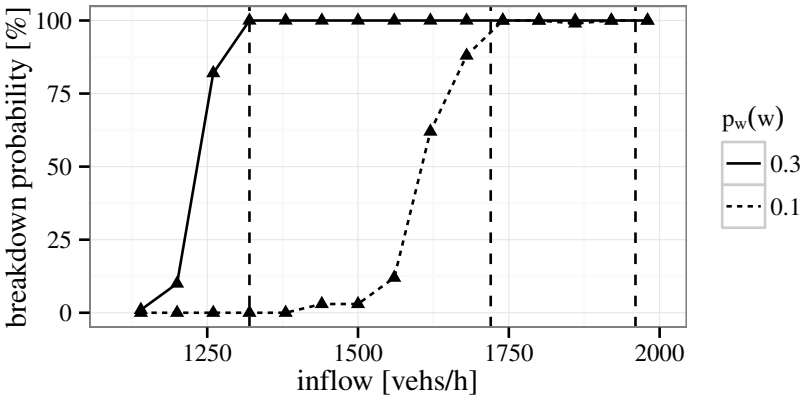


Figure 3: Simulated breakdown probabilities for the parametrization $p_w(w)$, measured at position 6850 m (averages from 100 simulation runs). From [1].

probabilities for different values of $p_w(w)$. Traffic breakdowns were counted using the following criterion: A breakdown has occurred, when the passenger car velocity was lower than 65 km/h during 5 min within the first 30 min after the surface wetness had been switched on.

A subset of the resulting breakdown probability curves is displayed in Fig. 3. To find an appropriate parametrization, we use the empirically measured maximum flows shown in Table 1 as an upper boundary, depicted by the vertical dashed lines.

References

[1] Lars Habel, Christoph Ide, Christian Wietfeld, and Michael Schreckenberg. Improving a Microscopic Traffic Simulation using Real-Time Information on Environmental Conditions. In *IEEE Veh. Technol. Conf.*, Montreal, 2016. IEEE.

[2] Lars Habel, Alejandro Molina, Thomas Zaksek, Kristian Kersting, and Michael Schreckenberg. Traffic Simulations With Empirical Data – How To Replace Missing Traffic Flows? In Victor L. Knoop and Winnie Daamen, editors, *Traffic Granul. Flow '15*, Delft, NL, in press. Springer.

[3] Lars Habel and Michael Schreckenberg. Asymmetric lane change rules for a microscopic highway traffic model. In Jarosław Wąs, Georgios Ch. Sirakoulis, and Stefania Bandini, editors, *Cell. Autom.*, volume 8751 of *LNCS*, pages 620–629. Springer, 2014.

[4] Lars Habel and Michael Schreckenberg. Analysis of asymmetric lane changing in a CA-based highway traffic model. *J. Cell. Autom.*, 11(4):351–361, 2016.

LOcal interferenCe compensATion (LOCATe) for GNSS-based Lane-Specific Positioning of Vehicles

Florian Liedmann, geb. Schweikowski
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
florian.liedmann@tu-dortmund.de

Precise traffic prediction and automated driving require accurate and reliable positioning information to be communicated to neighbouring vehicles as well as centralized traffic management centres. This contribution proposes and evaluates the so-called Local Interference Compensation (LOCATe) method, which is a cloud-aided solution to provide lane-specific positioning information of vehicles. With LOCATe, location errors of satellite-based positioning systems caused by shadowing and multi-path fading of the environment are predicted, quantified and compensated by identifying those impacts with ray-tracing techniques applied to a 3D model of the environment. This contribution introduces significant improvements of the original LOCATe approach which lead to both an enhanced accuracy and a reduction of the computation time to reach nearly real-time capability.

1 Introduction

As already introduced in [1] and shown in Figure 1 in the inner blue box, *LOCATe* is based on a satellite simulator combined with SiSNeT, UNB3 considerations and Ray-Tracing technology. To enable a compensation of the mentioned effects, the challenge is exactly the other way around. As input, *LOCATe* uses already distorted GNSS measurements and calculates the most probable real position on earth resulting in this kind of measurements after being influenced by all effects mentioned above. Hereby, the cloud-aided approach is used to circumvent the resource-limitations of mobile devices and leave the complex calculations (e.g. Ray-Tracing) to a backend without any resource-restrictions. Additionally, the communication delay and protocols are neglected in our considerations, due to the small amount of data for this kind of application per user equipment (cmp. [2]) and the existing permanent communication possibility in our time using smartphones. Summarized, the functionality of *LOCATe* is separated in the following three steps: 1.

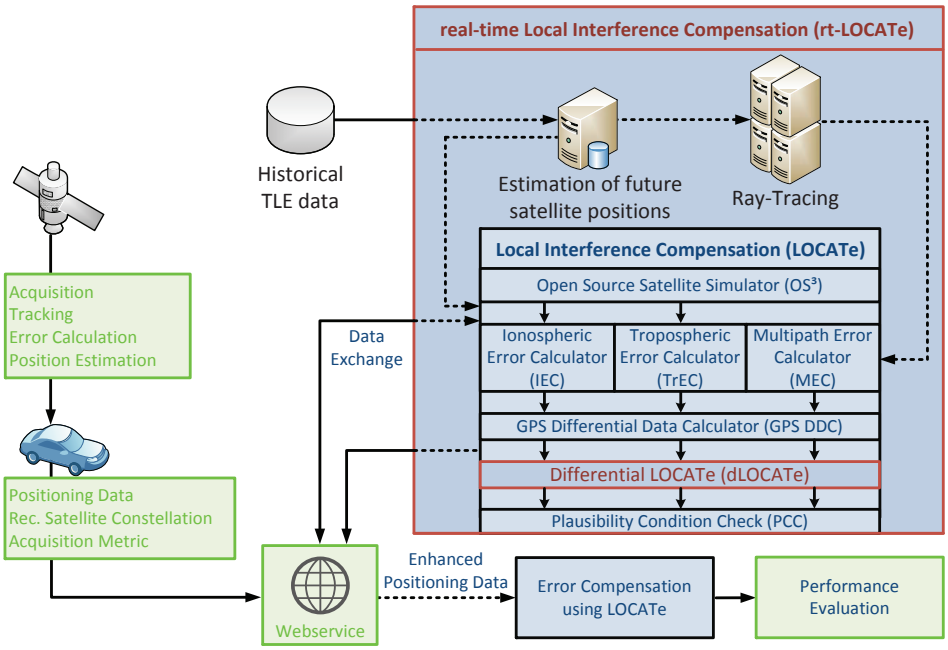


Figure 1: Architecture of LOCATe and Extensions to reach real-time capabilities

Predict all influences, 2. *Quantify* the accruing shift, 3. *Compensate* the overall error vector.

2 Extensions to reach Real-Time Capability

The first extension focuses on real-time capabilities. *LOCATe* includes resource-intensive Ray-Tracing technology and is not applicable to any tasks, which rely on real-time capability. The basic idea of *predictive LOCATe (pLOCATe)* is to use the latest set of Two Line Elements (TLE) data to predict the satellite positions for a short period of time in future. Afterwards, the predicted positions are forwarded to Multipath Error Calculator (MEC) to start a Ray-Tracing analysis for future constellations and the results are stored in relational databases to replace the complex Ray-Tracing analysis by a simple database access during run-time. On the one hand, out-dated TLE data causes a higher inaccuracy of a satellite's [3] and beyond that of the receiver's position, but on the other hand, this increases the time-efficiency significantly by a dramatically decreased process time of *LOCATe*. As a consequence, the accuracy loss should be analyzed and preferably compensated using the second extension *differential LOCATe (dLOCATe)* for instance.

Based on the basic approach of Multiple Antenna GNSS, the idea of *dLOCATe* is to work with at least two antennas with an exactly known distance x to each other. Position measurements in parallel on both points will cause an unknown error $\Delta err_{1,2}$, hence the calculated distance y between these defective positions may differ to the real distance x. By shifting both calculated positions simultaneously with the same value, distance y can be set to x. Thus, the influence of $\Delta err_{1,2}$ can be reduced in average. Although this method may also corrupt some measurements, this effect will be limited to significantly uneven maldistribution between the two measured points. In addition, scenario-specific considerations may also increase the positioning accuracy, but they limit the degree of freedom for possible results.

3 Performance Gain using LOCATe

First of all, an evaluation of the achieved accuracy gain using the basic *LOCATe* idea without any extensions is illuminated. Thereby, a developed *Advanced Software-Defined GNSS receiver (ASDR)* published in [4] was used to enable a *White Box*-analysis as well as a commercial GPS receiver from u-blox. The basic *LOCATe* idea increases the accuracy by an average of 45% in comparison with stand-alone GPS position measurements. In addition, and for some applications even more important, the maximum error can be reduced by 9m, which results in an overall improvement of more than 87% for all position values. By combining *predictive and differential LOCATe* to the so-called *real-time LOCATe (rt-LOCATe)* [5], the median of both error values shows approximately the same level facing only the accuracy of both systems. This is caused by Ray-Tracing results with predicted and less accurate satellite positions. But taking the CPU intensity of each task into account, the normalized key performance indicator E_{REL} of 1 (in case of GPS) can be significantly increased to 2,08 (*rt-LOCATe*), which implies at least an enhancement of $\approx 110\%$. Especially the inclusion of CPU usage clarifies the possible benefit of using *LOCATe* even for time-relevant application scenarios. The significant accuracy gain in combination with the negligible time-loss through extra CPU time clearly enables further applications to take huge advantages using this approach.

Because *LOCATe* is still under development and the testing equipment is not smooth-running, the campus of the university and its detailed 3D-model was used to create virtual lanes in direction of the widest dispersion to generate a first indication of the possible additional benefit in using these kind of considerations. Figure 2 shows the first results. Hereby, all measurements recorded on two highly accurate surveyed reference points are shown on a map in the middle, circuited by different virtual two-lane highways with 2.9m width each, and the gained *rt-LOCATe* results including *MapMatching*. It is obvious that the result depends on the used angle but also shows a significant accuracy gain. The lane-specific content increases to a value between 82% and 91% of the position measurements, which means an immense accuracy gain. Without increasing the accuracy before the *MapMatching* routine using *LOCATe*, only 64% of the gained measurements were allocated correctly, and by that may be count as lane-specific. Just the combination with the ideas presented in this paper increases the performance in a way, that nearly 90% of all improved measurements are mapped to the right lane.

4 Conclusion

This contribution provides the conceptional design, architecture and performance evaluation of the consistent further developed *Local Interference Compensation (LOCATe)* for GNSS to increase the accuracy especially for lane-specific applications in urban areas observing roughly real-time capabilities (*rt-LOCATe*). Using this post-processing approach, certain accuracy enhancements were identified using the three steps: Predict, quantify and compensate all influences on satellite signals at a specific point on earth. Hereby, a possible accuracy gain of more than 45% is detectable in consideration of the real-time capabilities and the constrained resources in the mobile devices using the cloud-based approach. By combining *LOCATe* with existing ideas like *MapMatching*, the accuracy gain increases even more and allows a lane-specific positioning with a probability of up to 90% by using GPS only. All results and models were evaluated with a prototypical

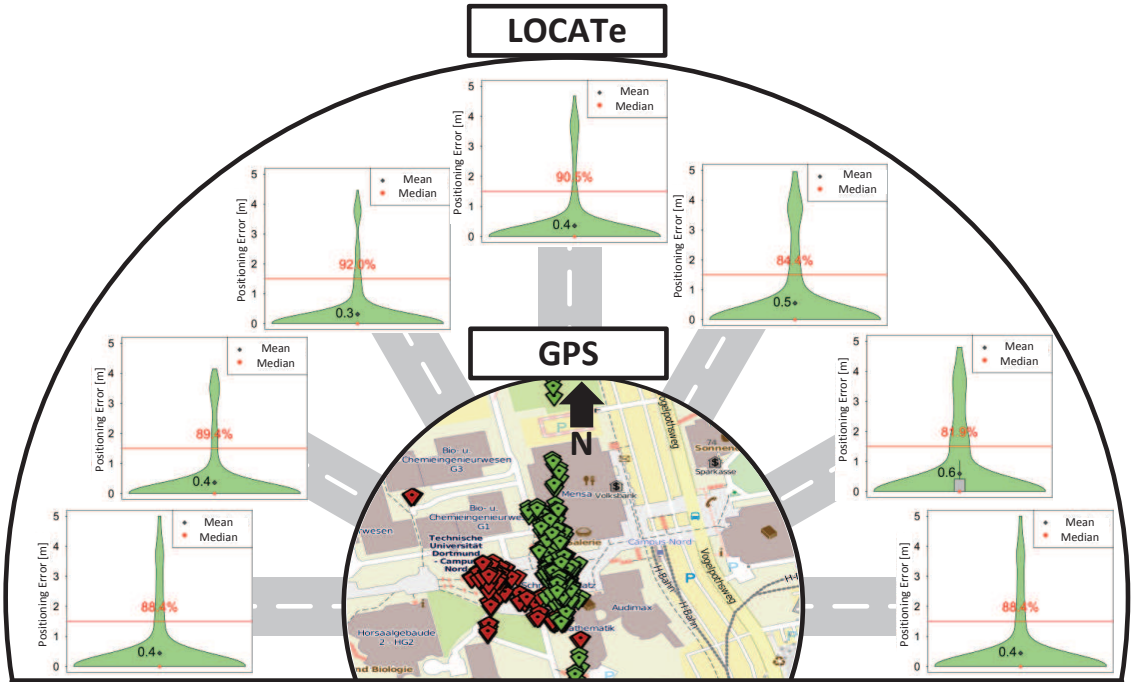


Figure 2: Accuracy Gain using real-time LOCATe on defective GPS measurements using different virtual highway directions

implementation on a SDR, underlining the applicability of the presented approach.

References

[1] B. Niehoefer, F. Schweikowski, S. Lehnhausen, and C. Wietfeld, "Cloud-aided sdr solution for lane-specific vehicle positioning via local interference compensation," in *Aerospace Conference, 2014 IEEE*, pp. 1–7, IEEE, 2014.

[2] B. Niehoefer, B. Ralf, and C. Wietfeld, "Gps community map generation for enhanced routing methods based on trace-collection by mobile phones," in *1st International Conference on Advances in Satellite and Space Communications (SPACOMM)*, pp. 156–161, IASTED, 2009.

[3] B. Niehoefer, S. Šubik, and C. Wietfeld, "The cni open source satellite simulator based on omnet++," in *Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques*, pp. 314–321, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013.

[4] B. Niehoefer, F. Schweikowski, and C. Wietfeld, "Smart constellation selection for precise vehicle positioning in urban canyons using a software-defined receiver solution," in *Communications and Vehicular Technology in the Benelux (SCVT), 2013 IEEE 20th Symposium on*, pp. 1–6, IEEE, 2013.

[5] B. Niehoefer, F. Schweikowski, and C. Wietfeld, "Local interference compensation (locate) for gnss-based lane-specific positioning of vehicles," in *Vehicular Technology Conference (VTC Spring), 2016 IEEE 83rd*, pp. 1–5, IEEE, 2016.

Poisson Graphical Models for Vehicular Traffic Prediction

Alejandro Molina
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
alejandro.molina@tu-dortmund.de

This report describes advances in graphical models for multivariate poison distributions and their applications to the analysis of traffic data. We present an improvement for the existing Poisson Dependency Networks (PDN) as well as a new Poisson Sum-Product Network model (PSPN) that offers tractable inference.

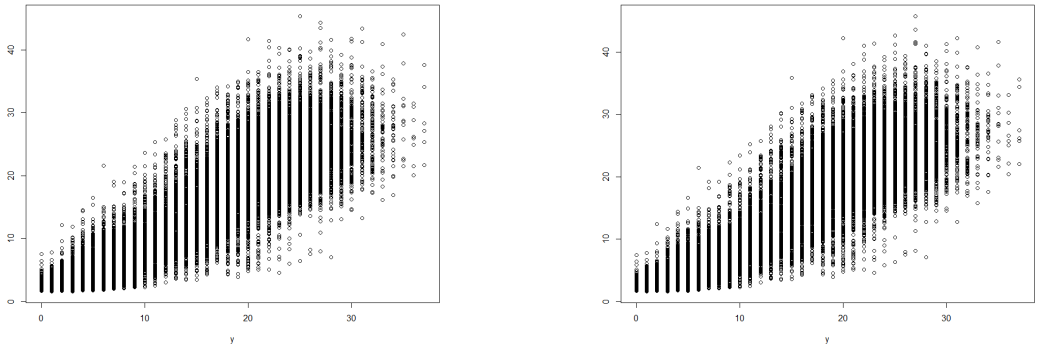
1 Coresets for Poisson Dependency Networks

One important aspect when modeling traffic data is that the number of measurements collected by the sensors rapidly grow with time and then we are in the setting where the number of instances is much larger than the number of dimensions ($n >> d$).

Based on the Poisson Dependency Networks (PDNs) [2] and in collaboration with C4 and their work on Random projections [1], we analyze how to reduce the number of measurements n and train a GLM [3] based PDN while maintaining the prediction power. This approach we call CoresetPDN.

To learn these CoresetPDNs, we obtain a weighted subset of the measurements through sub-sampling based on statistical leverage scores and then optimize the GLMs in the subset of measurements. By decoupling n from the optimization of the GLMs, we can train much faster and still make meaningful predictions.

An initial empirical analysis on the behavior of this approach is shown in figure 1, as it can be seen, the prediction results are comparable even when we take into account that CorsetPDN is using only 10% of the data



(a) CoresetPDN predictions

(b) PDN prediction

Figure 1: Comparison between Empirical and Predicted Vehicular Traffic Flows PDNs and CorsetsPDNs using 10% of the data. It can be seen that the results are very similar.

2 Poisson Sum Product Networks

PDNs offer great prediction capabilities, however, they are based on an inconsistent dependency network, and although it is still possible to recover a joint probability distribution $P(X)$ by means of sampling, however, it still shares a limitation with many other Poisson Graphical Models, in that they don't offer tractable inference.

We have therefore created the Poisson Sum Product Networks (PSPN), a Multivariate Deep Poisson model based on SPNs that offers exact tractable inference.

Sum Product Networks (SPNs) [4] are deep GMs capable of representing tractable distributions. The key benefit of SPNs over conventional graphical models such as MRFs is that common probabilistic inference tasks such as maximum-a-posteriori (MPE) and posterior marginal (MAR) estimation can be solved in time and space that scales linearly with the size of the representation.

Formally, an SPN as shown in Fig. 2 is a rooted directed acyclic graph model defined recursively as follows: (1) A tractable univariate distribution is an SPN. (2) A product of SPNs defined over different variables is an SPN. And (3), a weighted sum of SPNs with the same scope variables is an SPN. Thus, an SPN can be expressed as a rooted directed acyclic graph with univariate distributions as leaves, sums and products as internal nodes, and the edges from a sum node to its children labeled with the corresponding weights.

Tractable Inference in SPNs: To compute probabilities in an SPN, we simply have to compute the values of the nodes starting at the leaves. Since each leaf is an univariate distribution, one simply sets the evidence on those distributions, obtains the probabilities and then evaluate bottom up. On product nodes, we simply multiply the values of the

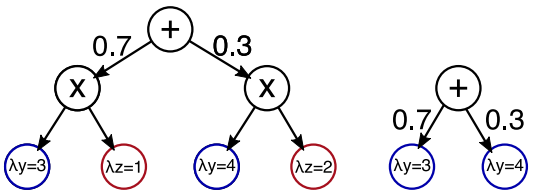


Figure 2: A 2D PSPN over y, z **(left)** and the corresponding marginalized PSPN for y **(right)**. Tractable inference can be performed using symbolic evaluation as explained later. **(left)** Resulting equation for $P(y,z) = 0.7\left((3^y e^{-3})/y!\ (1^z e^{-1})/z!\right) + 0.3\left((4^y e^{-4})/\ y!(2^z e^{-2})/z!\right)$. **(right)** Equation for $P(y) = 0.7\left((3^y e^{-3})/\ y!\right) + 0.3\left((4^y e^{-4})/y!\right)$. Univariate Poissons for y are shown in blue, and for z in red. (Best viewed in color)

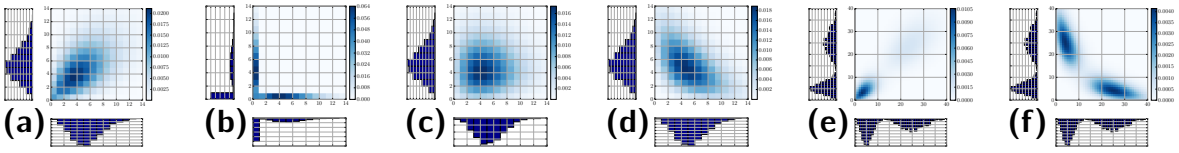


Figure 3: PSPNs are flexible multivariate Poisson distributions. Shown are the densities and marginals of six 2D Poisson SPNs with different dependencies between the two count variables. (a) Positive dependency suggests that two events often co-occur, (b) negative dependency that two events rarely co-occur, (c) zero dependency that events are not correlated, and (d) anti dependency that if one event occurs the other one does not occur. Mixture with positive (e) and anti (f) dependencies.

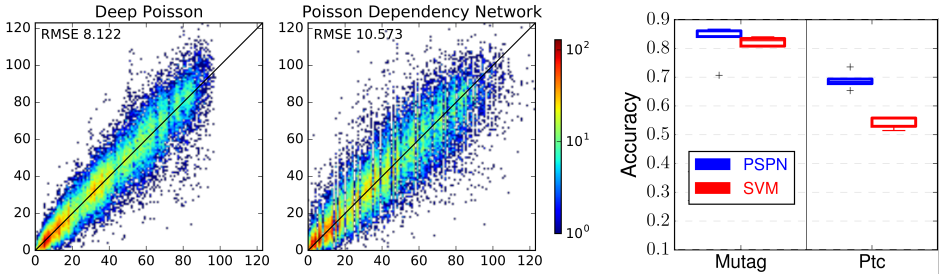
children nodes. On sum nodes, we sum the weighted values of the children nodes. The value at the root indicates the probability of the given configuration.

In terms of representation power, we can see that the PSPN is very flexible and can represent different types of dependencies between variables as shown in figure 3

When we apply the PSPN to the prediction of traffic, we see that a deep model helps in this setting, as can be seen in Figure 4 **(a)**. In cooperation with A6, we also analyzed how to classify graphs by first converting them to count features similar to the Weisfeiler-Lehman graph kernel [5], the classification results are comparable to state of the art as seen in Figure 4 **(b)**.

3 Hybrid Sum Product Networks

We are also working on extending the SPNs to compute a joint probability distribution for data wich contains different types of distributions.



(a) CoresetPDN predictions

(b) PDN prediction

Figure 4: **(a)** Comparison of traffic flow prediction by PSPN (left) with PDN (right), PSPNs prediction error is lower. **(b)** Graph classification by means of SVMs and PSPNs.

References

[1] Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for bayesian regression. *Statistics and Computing*, pages 1–23, 2015.

[2] Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning Journal (MLJ)*, 2015.

[3] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall / CRC, London, 1989.

[4] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690, 2011.

[5] N. Shervashidze, P. Schweitzer, E.J. van Leeuwen, K. Mehlhorn, and K.M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

Leveraging Mobility Control Knowledge for Intelligent Routing in Wireless Mesh Networks

Benjamin Sliwa
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
benjamin.sliwa@tu-dortmund.de

Providing reliable and efficient communication in Mobile Ad-hoc Networks (MANETs) is a challenging task for mesh routing protocols as the network topology is frequently changing. In this report, a cross-layer approach for utilizing mobility control knowledge from the application layer in order to optimize the forwarder selection process is presented. The results of in-depth simulation studies for our proposed protocol B.A.T.Mobile show that the Packet Delivery Ratio (PDR) of data transmissions can be significantly improved by the proposed method.

1 Mobility-aware Routing with B.A.T.Mobile

In order to provide reliable communication in highly dynamic networks, we have developed the mobility-aware routing protocol B.A.T.Mobile based on Better Approach To Mobile Adhoc Networking (B.A.T.M.A.N.) which uses a cross-layer approach to provide an interaction between the mobility control and routing layers. A novel mobility prediction mechanism is used to estimate future node positions in order to allow intelligent forwarder selections. For a detailed description of the calculations, see [1]. An example for the iterative prediction process is illustrated in Fig. 1. In each iteration i, the prediction method with the highest assumed precision is chosen.

- The *Steering Vector* $\vec{\sigma}_i$ is the estimated movement vector for the current iteration from the mobility control software. Since it is a weighted superposition of multiple

mobility algorithms, it cannot be predicted itself and is only available in the first iteration of the prediction process.

- If *Waypoints* $\vec{W}_i$ are available, they can be used as static orientation points for the general movement direction of the mobile node.
- As a fallback mechanism, the *Extrapolation* method can be used. The next vector is calculated as the mean movement vector of the last N_e positions.

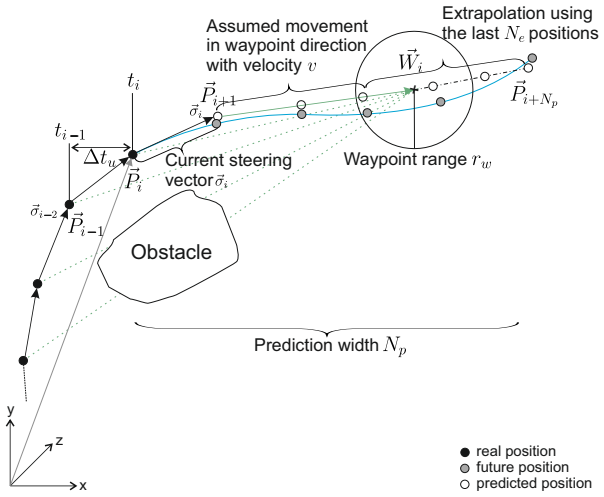


Figure 1: Cross-layer solution approach: utilizing predicted mobility information for the routing decisions

Instead of only using a single routing path, nodes maintain individual neighbor rankings for all destinations (see Fig. 2). The neighbor rankings contain all direct neighbors of the sender node which can theoretically reach the destination and a path score S as an estimation about the path quality which is computed from the link scores S_L. Using this approach the routing decision is simplified to selecting the node with the highest score as a gateway. A positive side-effect is the improved self-healing capability of the system as the multipath concept allows fast route switching in case of node failures. For the

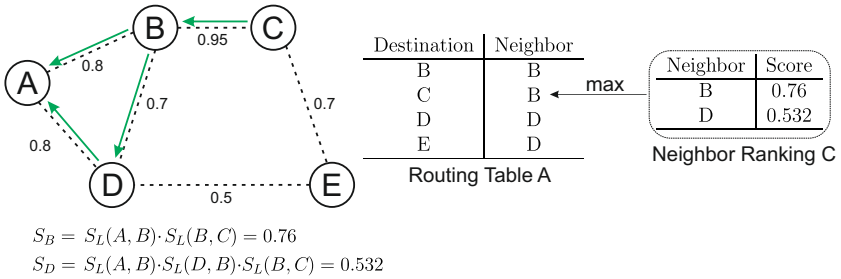


Figure 2: Score-based maintenance of multiple routing paths using neighbor rankings in an example network. Node *A* calculates the neighbor scores for the destination *C* from the received messages.

proposed B.A.T.Mobile protocol the score calculation is based on the current distance and on the predicted distance development between two nodes taking the relative mobility into account. The distribution of routing messages is handled according to the rules of the base protocol.

2 Simulative Performance Evaluation

In order to analyze the possible gains of using the cross-layer approach we use the discrete event-based network simulation tool Objective Modular Network Testbed in C++ (OMNeT++). The simulation setup as well as the implementations for mobility prediction and routing have been published as an Open Source framework as described in [2]. We compare our proposal to the reactive protocol Ad-hoc On-demand Distance Vector (AODV) and the proactive protocol Optimized Link State Routing (OLSR). Additionally the results for the base-version B.A.T.M.A.N. are shown as a reference.

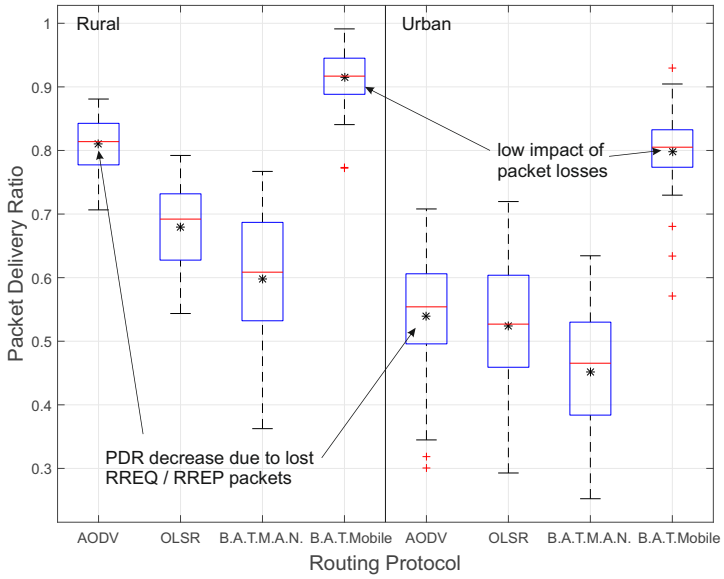


Figure 3: Comparison of the routing protocols for *Controlled Waypoint* mobility in two channel models

Fig. 3 shows the application-layer PDR for two different channel models. The impact of packet loss is taken into account considering the urban channel model. AODV suffers from the loss of route request (RREQ) and route reply (RREP) packets decreasing the routing performance. Analogously the path planning approach of OLSR fails because it requires current topology information of all network nodes. For B.A.T.Mobile the impact of lost packets is much lower, as every routing message contains an indicator about the quality of the reverse path to the originator. Fig. 4 shows the impact of individual mobility information for the overall PDR in relation to the positioning precision. The extrapolation approach is only useful if the location error is low, slight benefit can be achieved by

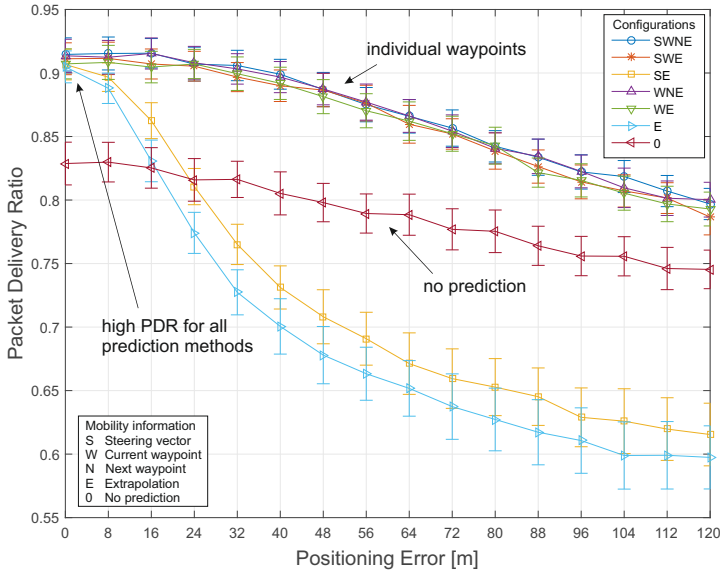


Figure 4: Sensitivity analysis: Impact of positioning errors on the routing performance for different mobility control information configurations

integrating the steering vector. In contrast to that, waypoints are not directly influenced by positioning errors and act as static orientation points which can significantly improve the PDR even under challenging conditions.

3 Conclusion and Further Research

The results show that mobility prediction is a suitable method for improving the routing performance of wireless mesh networks. In inner-city Vehicular Ad-hoc Network (VANET) scenarios the signal power is attenuated by shadowing effects from buildings in the area. In order to make more precise statements about the signal quality at predicted positions, we will integrate ray tracing methods in future work.

References

[1] Benjamin Sliwa, Daniel Behnke, Christoph Ide, and Christian Wietfeld. B.A.T.Mobile: Leveraging mobility control knowledge for efficient routing in mobile robotic networks. In *IEEE GLOBECOM 2016 Workshop on Wireless Networking, Control and Positioning of Unmanned Autonomous Vehicles (Wi-UAV)*, Washington D.C., USA, Dec 2016.

[2] Benjamin Sliwa, Christoph Ide, and Christian Wietfeld. An OMNeT++ based framework for mobility-aware routing in mobile robotic networks. In *OMNeT++ Community Summit 2016*, Sep 2016.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C1

Feature selection in high dimensional data for risk prognosis in oncology

Sangkyun Lee Sven Rahmann
Alexander Schramm

Punk Splits Tempers: On Characterizing Similarities and Differences of Labeled Data

Sibylle Hess
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
sibylle.hess@tu-dortmund.de

Finding the characteristics in data composed by samples of various classes is inter alia of interest for the development of personalized medicine in cancer treatment. Focusing on the identification of genes which make the difference between, e.g., an initial and a relapse tumor cell, we develop a general Boolean matrix factorization procedure, which takes class information into account. With synthetic and real-world data experiments, we show that the proposed method Punk is able to identify even finer differences between classes in comparison to state-of-the-art algorithms.

Introduction The task to "understand cancer" in the meaning of an identification of gene constellations which create the preconditions enabling the growth of a tumor, can be formulated in its general form as a Nonnegative Matrix Factorization (NMF) problem. Given a data matrix $D \in \mathbb{R}_+^{m \times n}$, NMF aims at finding nonnegative matrices $X \in \mathbb{R}_+^{n \times r}$ and $Y \in \mathbb{R}_+^{m \times r}$, such that $D \approx YX^T$. We call the columns in X patterns and refer to the columns in Y as usage. There are various possibilities to incorporate the knowledge about classes into this naturally unsupervised learning task.

Given the genomic information of a cell, e.g., in variant call format, together with further sample information such as cell characteristics (relapse/tumor/normal), we can store the information about deviations to the reference genome in a labeled binary row vector. Collecting these binary rows into a binary matrix D, we wish to find those patterns which discriminate the classes and those which are common over multiple classes. Since

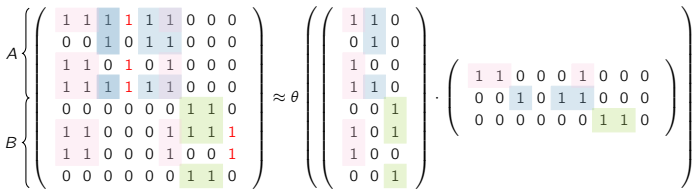


Figure 1: A Boolean factorization with overarching and class specific patterns (highlighted). Rows of the data matrix belong to one of the classes A and B.

the data matrix is binary, it is desirable to obtain a binary factorization as well, i.e., $X_{is}, Y_{js} \in \{0, 1\}$, not at least to obtain interpretable results.

Figure 1 shows such a Boolean factorization, utilizing the function $\theta(X) = (\mathbb{1}_{X_{ji}>0})_{ji}$ ($\mathbb{1}$ denotes the indicator function) to express the Boolean algebra in elementary terms. We can see that the data matrix on the left contains rows of two classes A and B. A possible factorization of this data matrix is showed on the right, containing three colored rank-1 factorizations. We observe that the pink factorization is overarching both classes, while the blue factorization is specific for class A and the green factorization is specific for class B. The identification of these class specific and general factors is of interest in order to understand the similarities and differences of the classes. However, the denoted factorization is not exact. We can see that the ones colored in red are not modeled – these entries are assigned to noise. Still, there is a class specific connection between these ones and the pink rank-1 factorization. Disregarding these details may lead to miss the crucial class discriminant information as we can not expect the data to contain clearly separable factors among the classes as highlighted in blue and green. Similarly, Kim et al. [4] find that approximate resemblance of overarching factors between the classes improve the quality of the obtained model in various aspects.

Consequently, we propose a Binary Matrix Factorization algorithm which identifies shared and specific rank-1 factorizations as well as the class specific variations of patterns which occur in multiple classes, i.e., the red ones in our example.

Punk The algorithm Punk is an extension of the algorithm Pimp [3] which incorporates the information of class-assignments. The algorithm performs a Proximal Alternating Linearized Minimization (PALM) [1] on a real-valued relaxation of the objective function. That is, we minimize a function

$$F(X, Y, D) + \phi(X) + \phi(Y), \tag{1}$$

where F is smooth and ϕ is proper and semicontinuous, by alternating updates

$$X_{k+1} = \text{prox}_{\alpha_k \phi}(X_k - \alpha_k \nabla_X F(X_k, Y_k, D)); \tag{2}$$

$$Y_{k+1} = \text{prox}_{\beta_k \phi}(Y_k - \beta_k \nabla_Y F(X_{k+1}, Y_k, D)). \tag{3}$$

The proximal mapping of ϕ is a function which returns a matrix satisfying the following minimization criterion:

$$\text{prox}_{\phi}(X) \in \arg\min_{X^\star} \left\{ \frac{1}{2}\|X - X^\star\|^2 + \phi(X^\star) \right\}.$$

Loosely speaking, the proximal mapping gives the argument a little push into a direction which minimizes ϕ. The variables α_k and β_k of Eq.(2) and (3) are the step sizes, which are computed under the assumption that the partial gradients $\nabla_X F$ and $\nabla_Y F$ are globally Lipschitz continuous with moduli $M_{\nabla_X F}(Y)$ and $M_{\nabla_Y F}(X)$.

While F reflects the relaxed version of the objective function, originally formulated for binary matrices, the function $\phi(X) = \sum_{i,j} \Lambda(X_{ij})$ for $X \in \mathbb{R}^{m \times n}$ penalizes non-binary values by utilizing the function

$$\Lambda(x) = \begin{cases} -|1 - 2x| + 1 & x \in [0, 1] \\ \infty & \text{otherwise.} \end{cases}$$

We derive a closed form for the calculation of the proximal mapping of $\phi(X) = \sum_{i,j} \Lambda(X_{ij})$ to sufficiently determine the argument of the minimum in Eqs. (2) and (3) [3].

In order to capture the class specific deviations of a pattern, we introduce the matrix $V \in \mathbb{R}^{n \times cr}$, where c is the number of classes. For each pattern $X_{.s}$ there are c columns $V_{.s}^{(a)}$ for $1 \leq a \leq c$ and $V^{(a)} = V_{.\{(a-1)r+1,\ldots,ar\}}$, one for each class, which reflects class characteristic deviations of $X_{.s}$. Therefore, the residual sum of squares is computed by

$$\sum_{a=1}^{c} \left\| D_{[a].} - Y_{[a].}(X + V^{(a)})^T \right\|^2,$$

where $[a] = \{j \mid \text{class of } D_{j.} \text{ is } a\}$. To ensure that the matrix V captures only class specific variations and no general patterns, we use a regularizer G penalizing the entries in $V_{.s}^{(a)}$ which seldom match the patterns in rows selected by $Y_{[a]s}$, and the entries in $V_{.s}^{(a)}$ which often match the patterns in rows selected by $Y_{[b]s}$ with $b \neq a$

$$\begin{aligned} G(Y, V) &= \sum_{s=1}^{r} \sum_{a=1}^{c} \left(\left|Y_{[a]s}\right| \left|V_{.s}^{(a)}\right| - Y_{[a]s}^T D_{[a].} V_{.s}^{(a)} + \sum_{b \neq a} Y_{[b]s}^T D_{[b].} V_{.s}^{(a)} \right) \\ &= c \,\text{tr}(Y^T D) + \sum_{a=1}^{c} \text{tr}\left(Y_{[a].}^T (\mathbf{1}_{m_a n} - 2 D_{[a].}) \right). \end{aligned}$$

All in all, we consider the minimization of the function

$$F(X, V, Y, D) + \phi(X) + \phi(V) + \phi(Y),$$

which is an expanded version of the objective introduced in Eq. (1) with

$$F(X, V, Y, D) = \frac{\mu}{2} \left\| D_{[a].} - Y_{[a].}(X + V^{(a)})^T \right\|^2 + H(X, V, Y, D) + G(Y, V)$$

by PALM. The function H derives from the relaxation of the objective function, formulated for binary matrices, as introduced by Pimp. Likewise, a rounding procedure is used to determine the factorization rank.

Experiments In first (synthetic) experiments, we verify the presumption that Pimp is able to identify class specific factors as displayed by the blue and green colored components in Fig. 1 – even more appropriate than specifically designed algorithms, although Pimp does not integrate any information about the class assignments.

In a second batch of experiments, we compare Punk, Pimp and state-of-the-art algorithms, which are designed to figure out similarities and differences between binary matrices of different classes. As such, we compare against the joint subspace Boolean matrix factorization by Miettinen [5] and the DiffNorm algorithm by Budhathoki and Vreeken [2] with respect to more subtle variations between the classes, as displayed by the red ones in Fig 1 on synthetical data. Furthermore, we will consider real-world data from applications in genome analysis and text-mining and investigate the benefit of taking the smaller deviations from larger patterns also into account.

References

[1] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[2] Kailash Budhathoki and Jilles Vreeken. The difference and the norm—characterising similarities and differences between databases. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 206–223. Springer, 2015.

[3] Sibylle Hess, Katharina Morik, and Nico Piatkowski. Pimping under the palm – tiling through proximal alternating linearized minimization (under review). *Data Mining and Knowledge Discovery*, 2017.

[4] Hannah Kim, Jaegul Choo, Jingu Kim, Chandan K Reddy, and Haesun Park. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 567–576. ACM, 2015.

[5] Pauli Miettinen. On finding joint subspace boolean matrix factorizations. In *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA*, pages 954–965, 2012.

Functional validation of mutations associated with neuroblastoma progression

Marc Schulte
Oncology Lab-Children's Hospital Essen
Universität Duisburg-Essen
marc.schulte@uk-essen.de

Neuroblastoma is one of the most common childhood cancers, with about 150 diagnoses per year only in Germany. A special feature of neuroblastoma is that the disease progression of individual patients varies significantly. Therefore some neuroblastomas regress after low-dose chemotherapy, or even without any treatment. However, in other cases the disease spreads very aggressively. Even if the tumour shows good response to an initial treatment, difficult to treat metastases and recurrent tumours are often developed. We previously identified several potential key genes by the analysing genome sequencing and expression data of paired primary and recurrent neuroblastoma. [5] These genes might play a yet unknown role in the development of difficult to treat metastases and recurrent tumours. In order to validate these in silico data in a biological system, CRISPR / Cas9 will be used to investigate the function of those genes in neuroblastoma cell lines. Currently we establish various CRISPR / Cas9 models for overexpression, knockout of individual genes, mimicking of specific mutations and mutation combinations. Our aim is to gain new insights into mechanisms underlying neuroblastoma recurrence to identify potential targets for a personalized therapy based on individual relapse-driving events.

Neuroblastom is the most common extra cranial solid tumour in childhood and accounts for 7 -10% of all childhood cancers. About 150 children are diagnosed with neuroblastoma in Germany every year. As a tumour of the autonomic nervous system, neuroblastoma derives from neural crest tissue and thus usually arises in a paraspinal location in the abdomen or chest [1]. Thanks to improved therapies neuroblastoma often initially responds

very well to the treatment. However, at relapse there is only very little to offer for the patients and hence relapses correlate with poor prognosis and fatal outcome (Figure1a). The median age at diagnosis is 17 months and the incidence of neuroblastoma is 10.2 cases per million children under 15 years [4].Thus, neuroblastoma develops in early life and therefore one can hypothesis that environmental influences are rather unimportant for the development of this type of cancer. The fact that the genetic background seems to be the major driving force in neuroblastoma, makes it an excellent model for the investigation of individual oncogenes.

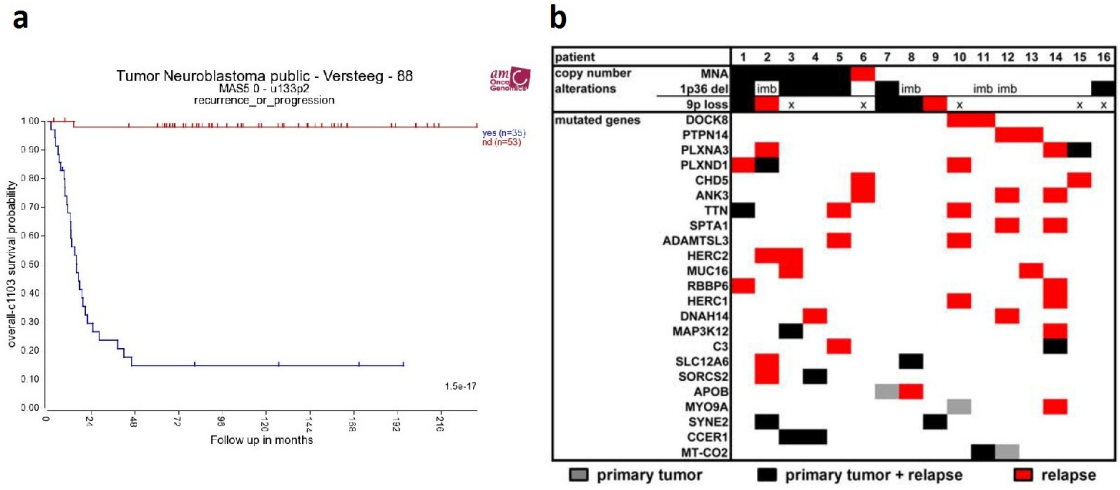


Figure 1: **Survival and gene mutations associated with neuroblastoma relapse.** (a) Kaplan Meier curve from "R2: Genomics Analysis and Visualization Platform" [3] comparing the survival of the patients with neuroblastoma reoccurrence or progression (blue) to those without (red). (b) Genomic alterations in matched primary and relapse neuroblastomas. Copy number alterations of MYCN (MNA), 1p36 and 9p were detected by array CGH (x = no samples for array CGH available) Mutated genes (SNVs) identified by exome sequencing are indicated as coloured boxes. (Adapted from Schramm et al., 2015 [7]).

In 2015, C1 subproject members used whole-exome sequencing, mRNA expression profiling, array CGH and DNA methylation analysis to characterize 16 paired samples at diagnosis and relapse from individuals with neuroblastoma. We were able to show that the mutational burden significantly increased in relapsing tumours, accompanied by altered mutational signatures and reduced subclonal heterogeneity. Furthermore, we found indications for clonal mutation selection during disease progression. In total, we identified 23 genes, which were mutated in at least two samples, as shown in Figure 1b. We hypothesize that these genes are involved in relapse formation. However, the relevance of these genes in the context of neuroblastoma needs to be validated biological. Therefore, the aim of this Project is to investigate the function of the identified genes

in neuroblastoma cell lines. For this purpose the CRISPR / Cas9 system will be used to create neuroblastoma knockout (KO) cell lines for the previously reported genes. The CRISPR / Cas9 system originated as a bacterial defence system, which cuts invading viral DNA at a defined position, and thus can effectively be used for genome editing. [2] This system consists of a protein (Cas9) and an RNA part. Cas9 binds to so-called single guide RNAs (sgRNAs), which direct the Cas9-sgRNA complex to a complementary DNA sequence. Here, Cas9 endonuclease induces double-strand breaks in the DNA. In order to create a gene knock out by using CRISPR / Cas9, we generated a neuroblastoma cell line expressing the Cas9 protein and designed gene specific guide RNAs. Further, we generated a Cas9 inducible Cell line, which allows to limit Cas9 expression to certain time frame. Here, Cas9 is fused to a green fluorescent protein (GFP), which allows to indirectly quantify Cas9 expression (Figure2).

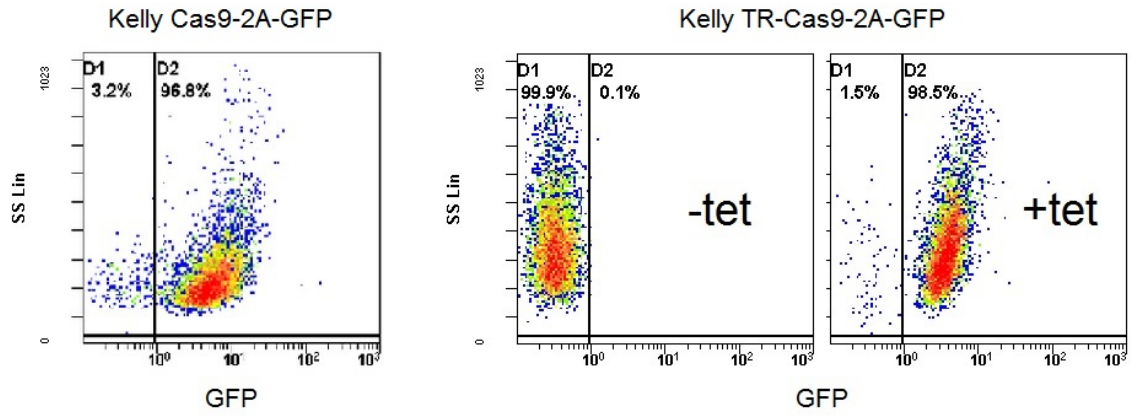


Figure 2: **FACS analyses of Kelly cells with constitutively and inducible Cas9 expression.** The neuroblastoma cell line Kelly was transfected with a vector coding for the Cas9 protein liked to GFP via a 2A sequence, leading to the generation of a constitutive expression system (left). Further, the same construct was used under the control of a tetracycline (tet) inducible promotor to create the inducible system (right). All cell lines were incubated with/without tetracycline and analysed by fluorescence flow cytometry (FACS)

To not only investigate the presents, but also the functionality of the Cas9 protein in the neuroblastoma cells. The Cas9 inducible cell line SHEP-TR-Cas9-GFP was transfected with guide RNAs targeting the GFP gene sequence attached to the Cas9. Therefore a functional Cas9 protein should cut the DNA at this position and thereby reduce the GFP signal. Consequently the reduction in GFP signal, which can be seen in Figure 3a, indicates a functional Cas9 protein.

However, this system can only be used to investigate loss of function mutations by KO. In order to also investigate gain of function mutations, CRISPRa will be used. Here a Cas9 protein, unable to cut DNA, is fused to an transcription activation domain (dCas9-VP160) and hence can be used to increase the expression of a gene of interest. To test

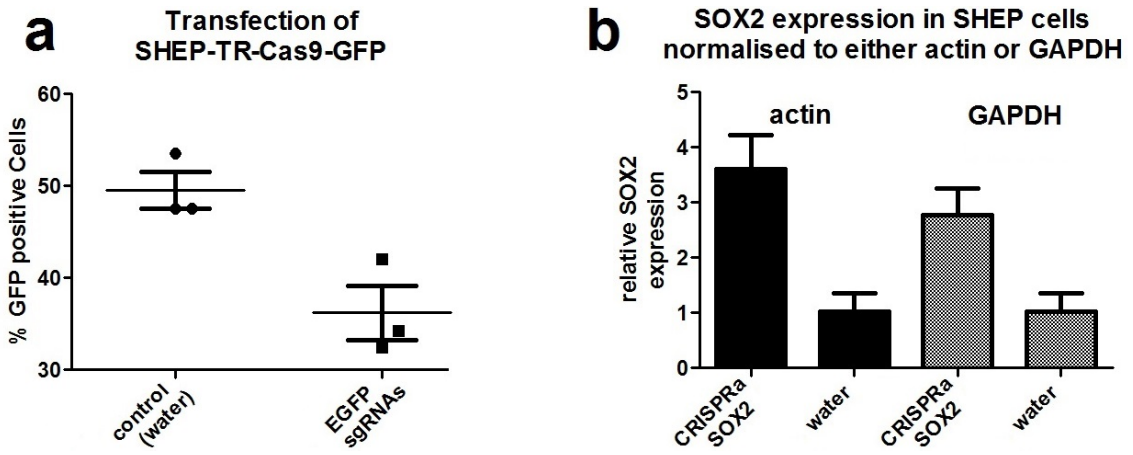


Figure 3: **Testing the CRISPR/Cas9 system.** (a)Cas9 inducible neuroblastoma cells were induced and transfected with guide RNAs targeting the GFP gene. Tin FACS analysis the transfected cells show decreased Fluorescence as compared to the water control. (b) The neuroblastoma cell line SHEP was transfected with a dCas9-VP160 vector containing three guides targeting the SOX2 gene. Real time PCR revealed increased SOX2 mRNA levels in transfected cells compared to the water control.

this system, we used guide RNAs for the SOX2 gene and were able to detect an increased gene expression as compared to the water control (Figure 3b). Further optimisation of the protocols is required to activate or inactivate the genes of interest and investigate the effects in in vitro experiments.

References

[1] Brodeur, G. (2003). Neuroblastoma: biological insights into a clinical enigma. Nature Reviews Cancer, 3(3), pp.203-216.

[2] Doudna, J. and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213), pp.1258096-1258096.

[3] Hgserver1.amc.nl, (2015). R2: Genomics Analysis and Visualization Platform. [online] Available at: http://hgserver1.amc.nl/cgi-bin/r2/main.cgi [Accessed 12 Nov. 2015].

[4] Maris, J. (2010). Recent Advances in Neuroblastoma. New England Journal of Medicine, 362(23), pp.2202-2211.

[5] Schramm, A. et al. (2015). Mutational dynamics between primary and relapse neuroblastomas. Nature Genetics, 47(8), pp.872-877.

Simulation of Double Digest RAD Sequencing Data

Henning Timm
Genome Informatics, Institute of Human Genetics
University Hospital Essen, University of Duisburg-Essen
henning.timm@tu-dortmund.de

This report documents the progress of RAGE, our simulation software for double-digest restriction site associated DNA sequencing (ddRAD) data. The structure of ddRAD data makes their analysis, as well as their simulation, a challenging task, since many of the biological and technological effects shaping the data interact with each other. Several steps of ddRAD analysis rely on coverage data, hence an accurate simulation of coverage values is required.

Introduction

The ability of our tool RAGE – ddRAD Data Generator – to accurately simulate ddRAD data, to provide a ground truth for validation, and to visualize the simulated data for the user, has greatly increased in the last year. The first part of this report gives an overview of the biological and technological effects influencing ddRAD data. The second part explores the simulation of these effects, focusing on the simulation of coverage values. Finally, an outlook on the further development of RAGE and its implications for the analysis of ddRAD data is given.

Structure of ddRAD data

As a reduced representation sequencing technology, ddRAD focuses the sequencing endeavor on a specific subset of the genome [4]. In the case of RAD sequencing, the

analysis is focused on DNA sequences adjacent to restriction sites. Restriction enzymes cut the DNA strand at specific motifs (restriction sites), thereby creating fragments for sequencing. While the family of RAD sequencing technologies contains several different technologies, we focus on ddRAD which offers unique advantages over RAD variants [1]. The ddRAD approach uses two enzymes, a rare cutter and a frequent cutter, to create fixed length fragments. This approach yields DNA from the same restriction sites across multiple individuals of the same species, which makes it ideal for large scale population analysis. A restriction site from which reads are generated is called a locus. One of the main challenges of ddRAD data analysis is reassembling these loci from the reads without a reference genome.

The structure of ddRAD data is influenced greatly by effects specific to the ddRAD approach. The most prominent effects are: highly covered loci (HCLs), null alleles (NAs), and singletons. A highly covered locus is formed when multiple loci from a repetitive region of the genome collapse into a single locus. These loci differ from valid loci insofar as they have a much higher coverage, but are unusable for the analysis. Null alleles are effects that partially or completely remove a locus from the data for an individual. This can be due to library preparation errors or an incomplete digestion of the DNA strand, but can also happen due to biological effects, like mutations of the restriction site. Singletons are singular reads that contain a sequence different from all other loci. These reads often are the result of errors or contamination of the sample during library preparation.

The structural simulation of these ddRAD specific effects has been further improved during the last year. Additionally, the ability to simulate more complex interactions of different effects, for example combinations of single nucleotide polymorphisms (SNPs) and null alleles, was added to RAGE. After the correct structure of the reads is achieved however, determining the numbers of reads to simulate remains a challenging task.

Simulation of Coverage

A main point of research in the last year has been the simulation of coverage values. In the context of ddRAD data, coverage describes the number of reads per individual associated with a locus. The target sequencing depth d_s, the aspired number of reads by design of the sequencing experiment, is often not reached during sequencing.

As illustrated by the example in Figure 1, the distribution of coverage values is expected to show two maxima: one at coverage 1, which is created by singletons and one at coverage d_s, created by valid reads. The peak for the valid reads is expected to be left-skewed, with coverage values scattering both below and above d_s. For HCL reads, no clear peak is visible, as the presence and structure of HCLs depends on many factors, including target genomic sequence and applied restriction enzymes. Constructing a model that

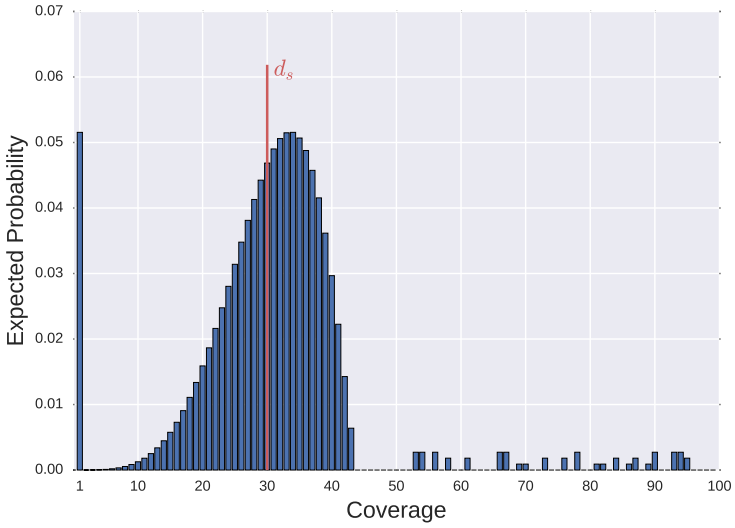


Figure 1: Exemplary distribution of coverage values for a dataset. The peak at coverage 1 is created by singletons, the coverages above 45 are created by HCLs. The coverage of HCLs can be much higher than displayed here.

can produce this distribution of coverages has been a major point of research in the last year. A combination of several coverage models, one for each the three classes described above, has proven to be the most effective way to simulate this behavior.

The total number of singletons is simulated as a fraction of the total number of valid reads. Since singletons only occur with a coverage of 1, this is sufficient to generate the first peak.

The expected left-skewed distribution for valid reads can be achieved by several distributions, but a Beta-binomial Distribution (BBD) has proven to be the most accurate match for the expected shape. The BDD requires three parameters: two shape parameters α and β and a parameter n defining the maximum number of events. Using the shape parameters, the BDD can be fitted to the expected coverage distribution and adapted to other shapes of distributions if necessary. Additionally the mean of the distribution, which is calculated as

$$\mathbb{E}(X) = \frac{n\alpha}{\alpha+\beta} \qquad X \sim \mathrm{BBD}(\alpha,\beta,n) \qquad \alpha > 0 \quad \beta > 0 \quad n \in \mathbb{N}_0 \tag{1}$$

has to be moved to d_s, by choosing the n parameter depending on α, β, and d_s. A Poisson Distribution, which has also been evaluated for this task, was not able to model the high variance that is expected in ddRAD data.

The coverage of HCLs is chosen from a Poisson Distribution (PD). HCL reads are often discarded from the analysis as noise by using the standard deviation of the observed

coverage distribution (after discarding singletons and small loci). The commonly used analysis tool Stacks [2] discards all reconstructed loci that contain more than $\mu+2\sigma$ reads, where μ and σ are the mean and standard deviation of the observed coverages [3]. Hence the λ parameter of the PD, which denotes the expected number of events occurring during a fixed time frame, is set to be above this threshold per default. For a more accurate simulation of HCL coverages however, more research is required.

Future Work

With RAGE being able to simulate realistic ddRAD data, our next step will be the creation of a ground truth dataset of simulated and real data. This dataset will be used to improve and evaluate ReClue, our software for ddRAD data analysis, and compare it with other tools, like Stacks. Knowing the detectable effects in the dataset will enable us to develop new and improved techniques for ddRAD analysis.

References

[1] Kimberly R Andrews, Jeffrey M Good, Michael R Miller, Gordon Luikart, and Paul A Hohenlohe. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2):81–92, 2016.

[2] Julian Catchen, Paul A Hohenlohe, Susan Bassham, Angel Amores, and William A Cresko. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11):3124–3140, 2013.

[3] Julian M Catchen, Angel Amores, Paul Hohenlohe, William Cresko, and John H Postlethwait. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, 1(3):171–182, 2011.

[4] Brant K Peterson, Jesse N Weber, Emily H Kay, Heidi S Fisher, and Hopi E Hoekstra. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, 7(5):e37135, 2012.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C3

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

Katharina Morik Wolfgang Rhode
Tim Ruhe

Astroteilchenphysik, TU Dortmund

FACT Real Time Analysis

Kai Brügge

January 17, 2017

FACT, The First G-APD Cherenkov telescope (FACT), is an Imaging Atmospheric Telescope (IACT) located at the Observatorio del Roque de los Muchachos on the Canary Island of La Palma. FACT is dedicated to monitor bright galaxies in the northern sky. IACTs record light flashes produced by gamma rays or cosmic particles that hit earth's atmosphere. Data from IACTs is typically strongly contaminated by background noise. Seperating the background from the the signal, cosmic gamma rays, is only possible using machine learning tools and large amounts of simulated data. This report presents some details about FACT's real time analysis (RTA). All of the software presented here is freely available on GitHub at `https://github.com/fact-project`.

1 The First G-APD Cherenkov Telecope

FACT, the First G-APD Cherenkov Telescope, is an Imaging Atmospheric Telescope (IACT) located at the Observatorio del Roque de los Muchachos on the Canary Island of La Palma.[1] At the time of its construction in 2011 it was the first IACT to use Silicon Photomultipliers (SiPMs) instead of the conventional Photomultiplier Tubes. With its 4 meter mirror it is rather small when compared to other experiments with similar technologies. Due to its small size however its the perfect testbed for automation technologies.

2 Raw Data Processing

FACT's raw data contains the voltage curves recorded in each individual pixel of the camera. Each of these curves has to be treated for calibration artifacts and other types of noise. From these voltage curves the number of photons and their arrival times can be estimated for each pixel. Figure 1 shows one exemplary voltage curve

on the left hand side and the estimated number of photons in each pixel on the right hand side. This is what one singe event looks like in FACT's data.

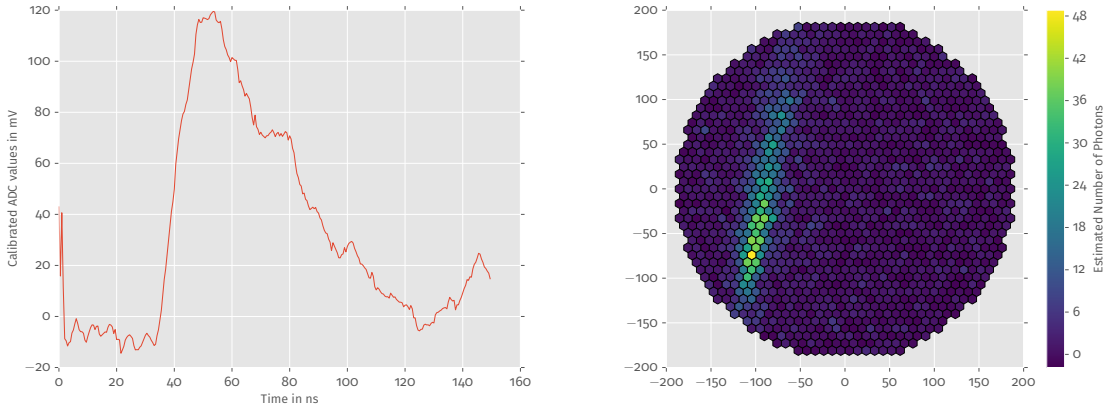


Figure 1: Light flashes in the atmosphere are recorded by FACT's camera. Voltage curves are stored for each pixel (left image) which are later used to estimate the number of photons which arrived in the camera (right image).

Raw data processing is achieved using the FACT-Tools. A collections of processors and services for the `streams-framework`[2]. The FACT-Tools reduce FACT raw data into image parameters that can be used for supervised machine learning methods[3]. In total 70 features are produced by the raw data processing.

3 Machine Learning on FACT Data

Labeled data is needed to perform classification into background and signal class using supervised machine learning methods. Since its technincally impossible to produce a test beam of large enough energy, simulations are needed to get labeled data. Machine learning is used for two major tasks when analysing IACT data: energy estimation and gamma-hadron separation.

3.1 Signal/Background classification

This classification task seperates recorded events by the type of the primary particle which entered the atmosphere and produced the recorded light flash. The negative, background, class are light flashes induced by cosmic rays. Charged nuclei which hit constantly the earth from all directions in the sky. IACTs are however only interested in the light flashes produced by gamma rays. Since gamma rays travel in straight lines from the source to the observer, one can infer things about the objects that produce the gamma rays.

Due to its robustness and the embarrassingly parallel nature of the algoroithm the Random Forest is the machine learning methods of choice for FACT and many other areas of high energy physics. A classifer with 150 trees is used as a model is used in FACT's Real-Time-Analysis. A total of 52 features are used while training the model. Figure 2 shows the receiver operating characteristic of the trained model. The recall and precisoen on of the model versus the prediction threshold is also plotted. Background events due to cosmic rays appear up to $10\,000$ more often in real recorded data. By adjusting the prediction theshold number of false positives can be reduced. At the same time however the recall gets worse. This trade-off has to be made each time one analyses data in this fashion.

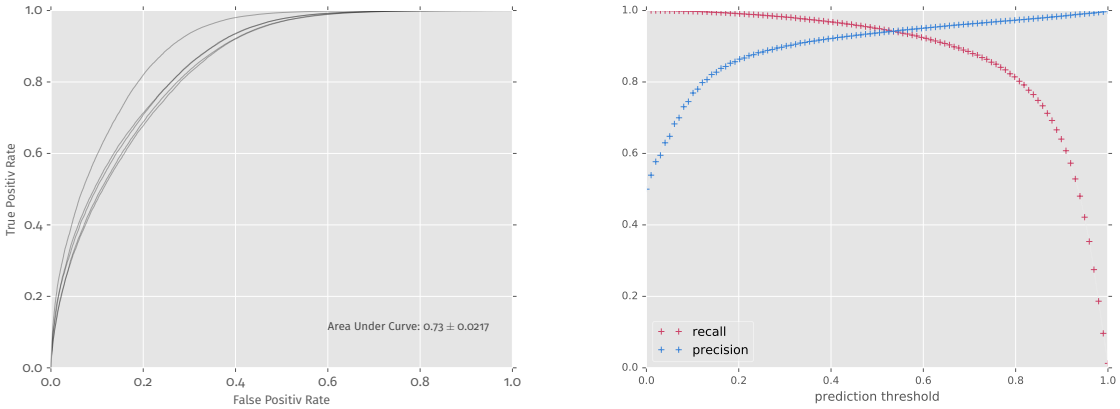


Figure 2: Cross Validated RoC curves for the Random Forest on the left and recall and precission versus the prediction threshold of the Random Forest on the right.

3.2 Energy Estimation

In a similar fashion to the classification model, a Random Forest regression model is trained. Its target variable is the energy of the primary particle that entered the atmosphere. Figure 3 shows the correlation between the true energy and the estimated energy as tests on simulated data. While the correlation is clearly visible, there is a clear smearing effect visible. This *resolution* of the estimator is plotted on the right hand side of the figure.

4 Monitoring Sources

One of FACT's main goals is the long-term monitoring of bright AGN type gamma ray sources. These sources can show variable behavior in terms of both total brightness and spectral properties. IACTs record so-called light curves which show the gamma

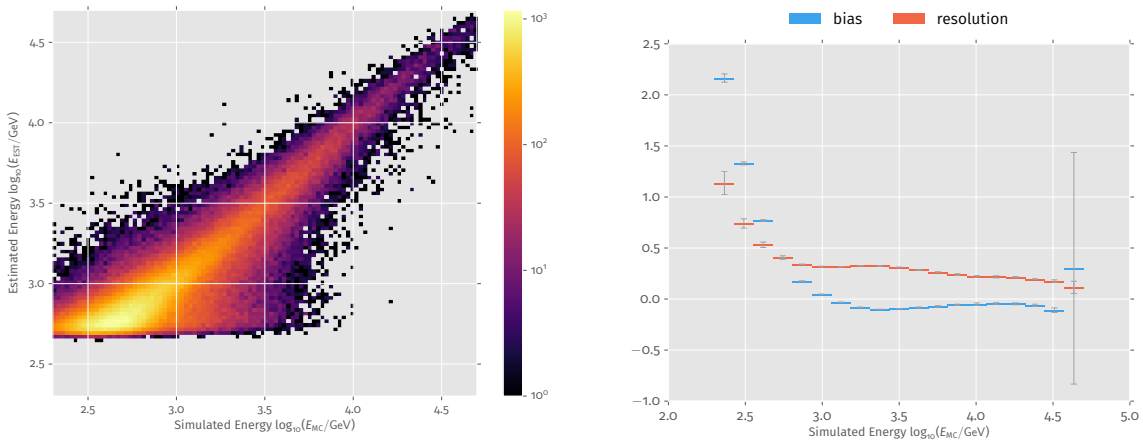


Figure 3: True vs estimated energy on the left and bias and resolution of the estimator in the right.

ray flux versus time. If a source shows a sudden increase in flux, it is considered to be in a flaring state. In case a flare is detected, FACT can alert other experiments to trigger multi-wavelengths observations.

5 The Real Time Analysis

The RTA combines all the steps needed to measure excess reates from observed sources. Starting with the raw telescope data, the RTA produces high level analysis results which can be stored in a database and displayed on a website. The single steps can be summarized as follows:

Telescope Data The telescope emits 50 events per second on average. This leads to a mean data rate of 43.2 MB per second and up to 1TB of collected raw data per night.

Calibration Calibration data is applied to the data stream to convert raw sensor output to physical units. This is the most time consuming step and is paralellized to several CPUs.

Image Parameter The number of photons and their respective arrival times are extracted from the raw data. Camera pixels not containing any cherenkov photons are dropped and parameters describing the final image are calculated.

Energy Estimation Use pre-trained regression models on the parameters to estimate the energy of signal events.

Classification Use pre-trained models to separate signal from background events.

Neighbor Pixel Correlation

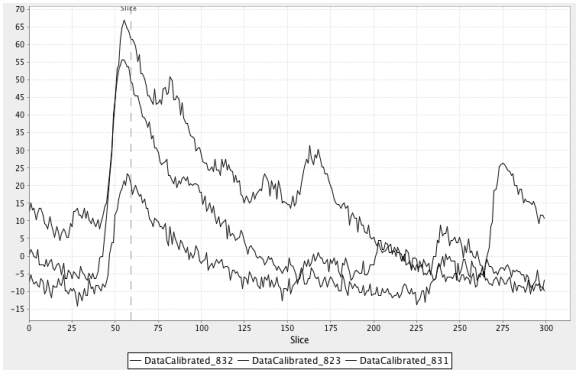
A new Feature for the Image Cleaning in FACT-Tools

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

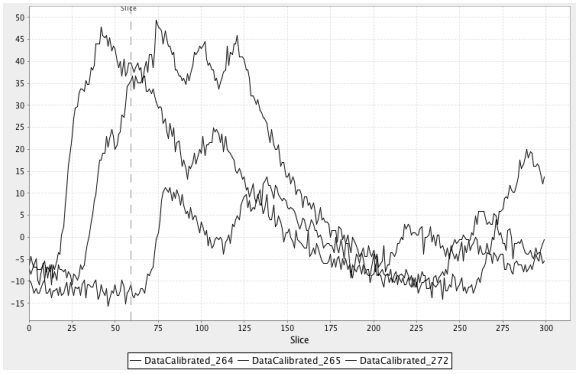
The First G-APD Cherenkov Telescope (FACT) is pioneering the application of silicon-based photo sensors for the imaging atmospheric Cherenkov technique. These sensors, allow to operate under more severe light conditions than the commonly used photomultiplier tubes. In order to analyse data from such light conditions, it is necessary to optimize the image cleaning, one of the first analysis steps in the FACT analysis chain. The current FACT-Tools standard analysis uses a threshold based image cleaning, which is based on the arrival time and quantity of photons in the camera pixels. In order to aim for a more sophisticated, machine learning based image cleaning, additional pixel based features are introduced in this report.

Introduction The first G-APD Cherenkov Telescope (FACT) is equipped with 1440 silicon photo multipliers (SiPM) as camera pixels. By use of these devices, the FACT collaboration successfully proved the application of semi-conductor based photon detectors (G-APDs aka. SiPMs) for the imaging atmospheric Cherenkov technique [1].

With this technique, the Cherenkov light from a shower, within the telescope's field of view, is reflected towards the camera plane. Here it appears as a two dimensional light distribution (see figure 2b). The incoming photons are measured in the hexagonal camera pixels as a time series of the SiPM's electronical signal. A typical example for this signal is given in figure 1a. Here the time series of four pixels within the shower is displayed. The typical waveform is defined by a steep rising edge and an about 150 slices long, exponentially falling edge. From this, the pixel content is conventionally parameterized as the quantity and the mean arrival time of Cherenkov photons. In a subsequent step, these two numbers are used to perform a basic image cleaning on the camera pixels.



(a) Signal of neighboring shower pixels.



(b) Signal of neighboring non-shower pixels.

Figure 1: Photon signals as time series of three neighboring pixels containing Cherenkov photons and night sky background photons.

The image cleaning has the goal to distinguish between pixels with Cherenkov photons from a shower and those that only contain night sky background photons (NSB). The Cherenkov photons are produced in the shower within a small time window in the order of 10 ns. Therefore, their temporal distribution can be considered as coherent within one pixel and small in between neighboring pixels. However, night sky background photons, which are mostly dominated by direct and scattered starlight aswell as the moon, are also reflected into the camera plane. This background has to be suppressed for further analysis steps. Nevertheless, their arrival times are homogeneously distributed over the time series. In consequence, they appear more incoherent in between neighboring pixels. This allows for an image cleaning based on their time difference and their neighborhood conditions.

In case of the FACT-Tools standard analysis [3], a basic two-level time-neighbor cleaning is applied. Pixels are considered as shower pixels if their photon content exceeds a certain threshold and the arrival time difference of neighboring pixels is smaller than a certain time interval of less than 5 ns.

These thresholds depend on the behavior of the NSB. Since, they have to be adapted if the background rate increases, e.g. due to moon light. In addition, not only Cherenkov photons are extracted but also NSB photons.

Nevertheless, the image cleaning is a classification task with the aim to distinguish pixels containing Cherenkov photons from those pixels that do not contribute to the shower. A more sophisticated approach to this problem would be a machine learning based classification of camera pixels. In preperation for this apporach a larger variety of features than the current two (photon charge and arrivaltime) is necessary. A promissing feature set, which is based on neighborhood conditions, is presented in the following.

Neighbor Pixel Correlation The time series of each pixel, as depicted in figure 1a, can be interpreted as a distribution of amplitudes. Since SiPMs signals have a typical shape, this distribution should have a recognizable shape for each detected photon.

Due to the coincident arrival of Cherenkov photons in neighboring shower pixels, their amplitude distributions appear to be more alike (see figure 1a) than the distributions of neighboring non-shower pixels (figure 1b). This leads to the assumption that the time series of neighboring shower pixels should have a higher correlation than those of neighbor pixels that do not contain coincident Cherenkov photons.

Demanding this hypothesis, the correlation of two adjacent pixels is calculated by the correlation coefficent [2]:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} \tag{1}$$

Since the time series of a pixel which is sampled with a bin width of 0.5 ns, a descretized formulation of the covariance is used:

$$\sigma_{xy} = \sum_{i}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle) \tag{2}$$

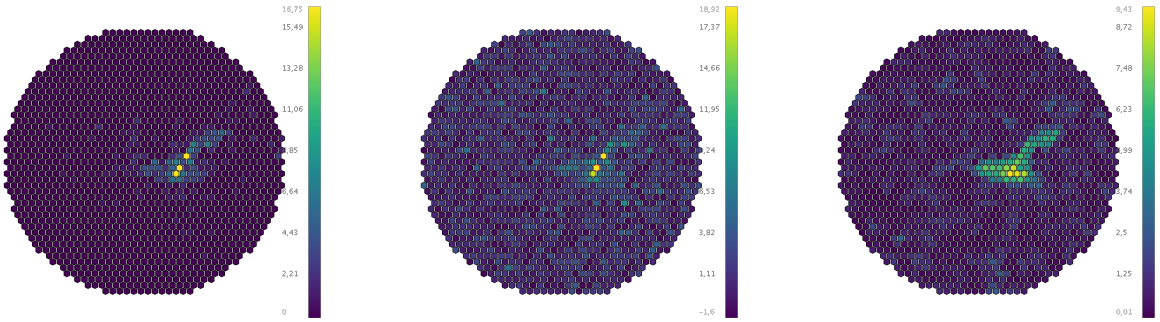
This allows to compute a correlation for each pixel with each of its up to six neighbors. In order to give a certain value to each pixel, the moments (mean, variance, etc.) of the distribution of a pixel's correlation coefficients with its neighbors are used. These add up six further features to parameterize a pixel's content:

1. Average neighbor pixel correlation (`ANC`)
2. Standard deviation of the neighbor pixel correlation (`StdNC`)
3. Maximum neighbor pixel correlation (`MaxNC`)
4. Minimum neighbor pixel correlation (`MinNC`)
5. Kurtosis of the neighbor pixel correlation (`KNC`)
6. Skewness neighbor pixel correlation (`SNC`)

Results The behavior of these new features is evaluated with simulated air showers from the FACT simulation chain. In order to have a first measure for these new feature's performance, they are compared to the simulation's truth of detected cherenkov photons in the pixels. Figure 2 shows the simulation truth (2a), the estimated quantity of photons (2b), and the `ANC`(2c) for a shower example.

The conventionaly used photon quantity has the tendency to also extract photons in pixels that do not contain any Cherenkov photons, since it cannot distinguish between cherenkov photons and NSB photons.

The `ANC`seems to be less prone to give a high value to pixels without Cherenkov photons. Also the shape of the shower is well represented if compared to the true photon quantity.



(a) True photon quantity (b) Estimated photon quantity (c) Average neighbor pixel correlation

Figure 2: Example of a Cherenkov shower from proton simulations. The true number of photons is displayed in comparison to the commonly used estimated quantity and newly developed average neighbor pixel correlation.

Also the `MaxNC`, the `MinNC`, and `StdNC` are correlated with the shower pixels. At first glance these features seems to be promising candidates to destinguish shower and non-shower pixels.

Conclusion and Outlook The presented method allows to compute six additional features to parameterize the content of a camera pixel. Four of these features seem to be good indicators weather a pixel contains Cherenkov photons or not. Thus, they offer potential for a future machine learning based image cleaning.

Nevertheless, a more sophisticated evaluation of these features' performance has yet to be done. Also the training of learning a algorithm for the image cleaning is planned to be done with these features in the near future.

References

[1] H. Anderhub et al. Design and operation of FACT the first g-apd cherenkov telescope. *JINST*, 8(06):P06008, 2013.

[2] V Blobel and E Lohrmann. *Statistische und numerische Methoden der Datenanalyse*. B. G. Teubner Stuttgart, 1998.

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

Sample Generation for the Measurement of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data

Mathis Börner
Experimentelle Physik 5b
Technische Universität Dortmund
mathis.boerner@tu-dortmund.de

The goal of my work is to measure the atmospheric muon neutrino energy spectrum with IceCube data. The covered energy ranges from ∼100 GeV to multiple PeV. A good knowledge of the atmospheric neutrino flux at those energies is essential on the way of identifying the first astrophysical neutrino source, because it shows the transition from an atmospheric dominated to an astrophysical dominated spectrum. In my work I choose an approach successfully applied to previous years of IceCube data [1, 2]. It will be the first time this approach is applied on multiple years. A more detailed physical motivation and a broader overview of the analysis can be found in [3]. This report focuses on new developments and results achieved in the year 2016.

1 Introduction

The full analysis can be separated into two parts: the generation of a pure muon neutrino sample and the subsequent unfolding.

For the unfolding a pure sample muon neutrinos has to be created. Detected muon neutrino events are the *signal* for the analysis. Whereas, the major *background* are atmospheric muons events. At the beginning of the analysis the signal to background ratio is 1:1 000 000. The final samples of the previous analyses achieved between ∼30 000

and 90 000 signal events per year with purities around 99 %. The approach to obtain such a sample consists of six steps: data selection, precuts, feature generation, feature selection, multivariate classification and the final cut on the classifier score to generate the sample. In the following sections some renewals for the feature selection and the cut on the classification score are shown.

2 Data/Simulation Mismatch

One of the big challenges of the analysis is to ensure a good agreement between the used simulations and the actual data. Moreover with the use of multivariate algorithms like Random Forest (RF) [4] the agreement has to be checked multivariate. In the previous analyses the distributions of all features for simulated and measured events were compared and checked to detect potential mismatches. Another challenge is to judge whether between negligible and problematic mismatches.

For the multiyear unfolding analysis a novel approach is pursued. A Random Forest is trained to differ between simulated and measured events [7]. Without any mismatches the RF should not be capable to separate the events. In case of a clear separation the "feature importance" provided by the RF can help to identify features with the most significant mismatches.

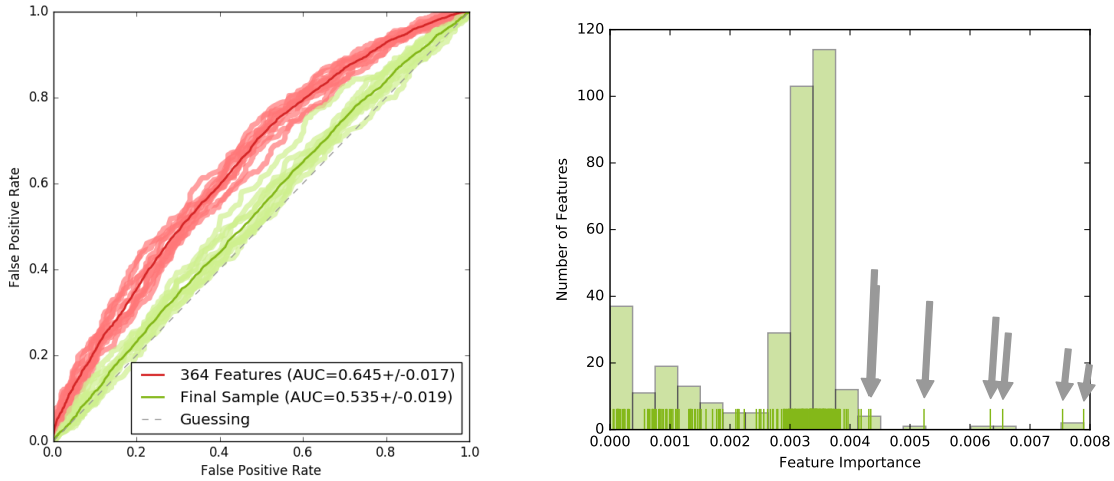
Before the RF is trained a preselection of the features is done. As a first step all constant features are removed. For all features the correlation to the rest of the features is calculated. And features with a *Pearson correlation* ≥ 0.96 only one is kept. This leads to a set of 359 features. Those features are used for the training of the RF. The corresponding ROC curve is shown in Figure 1a.

The ROC curve indicates significant mismatches in at least some of the features. The features importances are shown in Figure 1b and show seven features with a clearly higher importance. Those features are removed from the set.

The final 40 out of the remaining 352 features are selected with a "mRMR" features selection [5]. This features set is again used to train the RF. The retrieved ROC curve (see Figure 1a) with an AUC of 0.535 ± 0.019 indicates a significant reduce of mismatches in the final features.

3 Classification Score Cut

Another change in the separation chain is the way the cut on the classification score is done. After the features selection a Random Forest is trained to separate between atmospheric and neutrino-induced muons. The output of the RF is a score between 0.0



(a) Random Forest score for the classification with the full feature set.

(b) ROC curves for training with the full and the final features set.

Figure 1: Random Forest score and ROC curves for the classification of measured and simulated events to detect mismatches in the simulation.

and 1.0. A score close to 1.0 indicates a neutrino-induced moun and a score close to 0 an atmospheric muon. To retrieve the final sample a cut on this score has to be done. The cut must be tuned to fulfill certain levels of efficiency and purity. In the predecessor analyses either a single threshold or a threshold with a single step in the estimated muon energy were used. A more sophisticated way to define the cut was tested in [6] and is the basis for the approach presented here.

In this analysis the cut is a function $RF_{\text{cut}}\left(E_{\mu}^{\text{est}}\right)$ depending on the estimated energy of the muon. To obtain the curve individual thresholds are determined for small energy regions. The thresholds in each region have to achieve a purity of 99 %. To get the cut curve all thresholds are stringed together. To have a sufficient number of events in each energy region and to get a smooth curve the regions are overlapping. For this analysis 100 regions between 320 MeV and 1 PeV with a width of $0.3/\log_{10}(\text{GeV})$ are used. To prevent the curve from sticking too heavily to individual events of the samples the curve is determined on 100 bootstrapped samples and the final curve is the average of those curves.

The resulting cut curve and the energy dependent classification score is shown in Figure 2. The settings for the whole separation process are not yet fully optimized. Nevertheless, the preliminary results achieve between 700 000 and 100 000 neutrino induced events per year with purity around 99 % for each year.

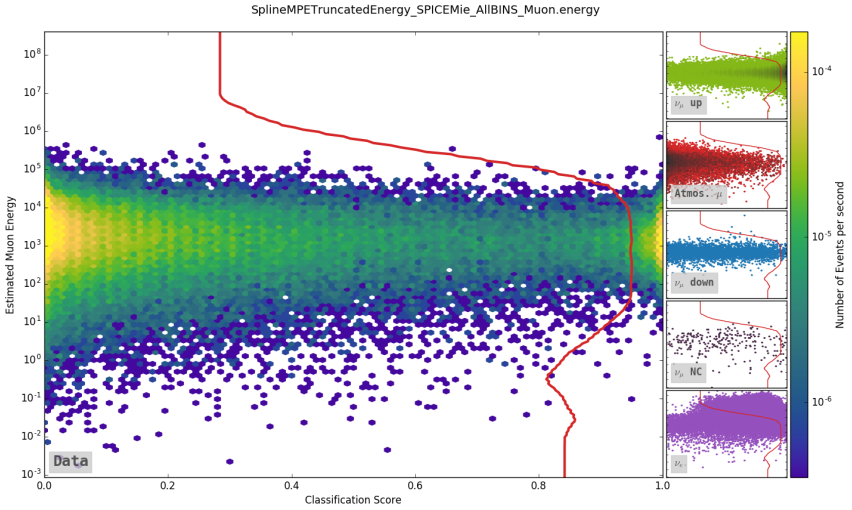


Figure 2: The histogram on the left shows the energy dependent distribution of the classification score for the measured events. On the left side the contribution of the signal component ("ν_μ up") and all the different background components are shown. The red curves illustrates the used cut.

References

[1] M. G. Aartsen et al. "Development of a general analysis and unfolding scheme and its application to measure the energy spectrum of atmospheric neutrinos with IceCube". In: *The European Physical Journal C* 75.3 (2015).

[2] M. G. Aartsen et al. "Unfolding measurement of the Atmospheric Muon Neutrino Spectrum using IceCube-79/86". In: *Proceedings, 34th International Cosmic Ray Conference (ICRC 2015)*. 2015.

[3] M. Börner. *Measurement of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data*. Tech. rep. Technischer Report für das SFB 876 - Graduiertenkolleg. TU Dortmund University, Dec. 2015.

[4] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001).

[5] C. Ding and H. Peng. "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of bioinformatics and computational biology* 3.02 (2005).

[6] J. Luckey. "Optimierung der Separation des IC86-I Myon-Neutrino-Datensatzes". Bachelor Thesis. Technische Universität Dortmund, 2015.

[7] Alex Rogozhnikov. "Reweighting with Boosted Decision Trees". In: *arXiv preprint* (2016).

Search for high-energy astrophysical Tau-Neutrinos in IceCube Data

Maximilian Meier
Lehrstuhl für Experimentelle Physik V
Technische Universität Dortmund
maximilian.meier@tu-dortmund.de

In this work a potential astrophysical flux of tau neutrinos is investigated. The measurement of tau neutrinos would be a clear astrophysical signal since no tau neutrino flux from the atmosphere is expected. Tau neutrinos can be detected with the IceCube detector by their unique double-cascade signature at high-energies. The tau neutrino signal is buried in a large amount of background events with a signal-to-background ratio of about $\mathcal{O}(1 : 10^{10})$ at trigger level. To remove these background events methods from the field of machine learning are applied. A detection of tau neutrinos would require a very pure sample to achieve a high significance and a very efficient analysis to obtain sufficient statistics in limited time of measurement.

1 Introduction

The IceCube detector is a neutrino telescope with an instrumented volume of one cubic kilometer at the south pole. It consists of 5160 digital optical modules (DOMs) deployed on 86 strings at depths between 1450 m and 2450 m in the glacial ice. On top of the surface 81 tanks are installed to detect air showers or work as a veto region for the in-ice array. [3]

In 2013 the IceCube Neutrino Observatory observed a diffuse flux of astrophysical neutrinos with a significance of 5.7σ in an all-flavour analysis [1]. The direct observation of tau neutrinos as a component of this flux has yet to be performed.

The signature of a charged current tau neutrino interaction consists of two parts. The hadronic cascade at the neutrino interaction vertex producing a cascade-like signature

and a tau lepton in the detector and the subsequent decay of the tau lepton, which also produces a cascade in the detector for all non-muonic decay modes. This results in a so called double-cascade signature.

At PeV neutrino energies the tau lepton can travel distances on the order of the DOM-to-DOM distance in IceCube and the event can be properly reconstructed as a "double bang" event with two separate cascades. The reconstruction algorithms can not resolve double-cascade events at lower energies. This analysis aims to find two separate pulses in at least one of the IceCube optical sensors, one from the first and one from the second cascade (as suggested in [4]). This type of event is called a "double pulse" event. Reconstructing events with this technique lowers the detection threshold to $\sim$100 TeV.

The double pulse events have two different kinds of backgrounds, which have to be removed in separate processes. Cascade-like background events, like charged current electron neutrino, as well as all flavor neutral current, interactions, can not produce a substantial double pulse feature in their waveforms. Track-like background events, like cosmic-ray induced muons from the atmosphere and charged current muon neutrino events, can produce double pulse waveforms via a huge stochastic loss near a DOM after having already deposited enough energy to produce a significant pulse in the same DOM.

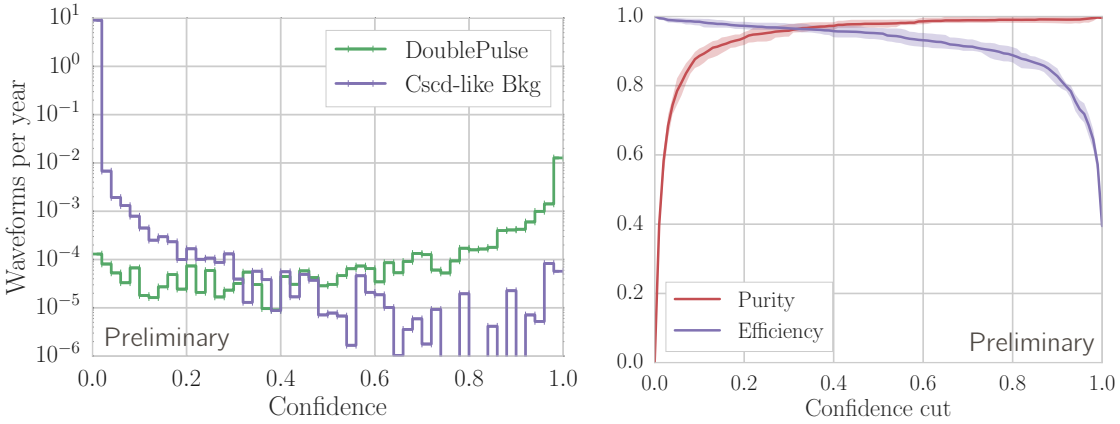
2 Double Pulse Selection

To be able to produce a double pulse waveform, both cascades have to be both near a specific DOM but also sufficiently separated in time. Since this condition does not get fulfilled by every simulated tau neutrino event a subset of signal waveforms has to be identified. This selection depends on Monte Carlo truth information, so it can not be applied on data.

Each waveform gets characterized by features describing its derivate (see [2]), its summary statistics, the number of local maxima and the goodness of fit to an exponential function describing the usual singe pulse waveform.

The separation between double pulse waveforms and single cascade waveforms is performed via a random forest with 200 trees. For training 200 000 background waveforms and only $\mathcal{O}(6500)$ signal waveforms were used due to limited simulation statistics. The random forest was validated in a 10-fold cross validation.

The results of the random forest are shown in figure 1. Since the background events this random forest was trained on are not capable to produce sufficient double peak structures only weak second pulses from e.g. late scattered photons have to be removed. As depicted in figure 1a both confidence distributions can be separated quite well. This also leads to very high purities, even for efficiencies well above 90 %.



(a) Confidence distributions for the double pulse selection. The green distribution contains all preselected signal waveforms. The purple distribution contains all components, that produce cascade-like background events, namely waveforms from all flavor neutral current interactions and electron neutrino charged current interactions.

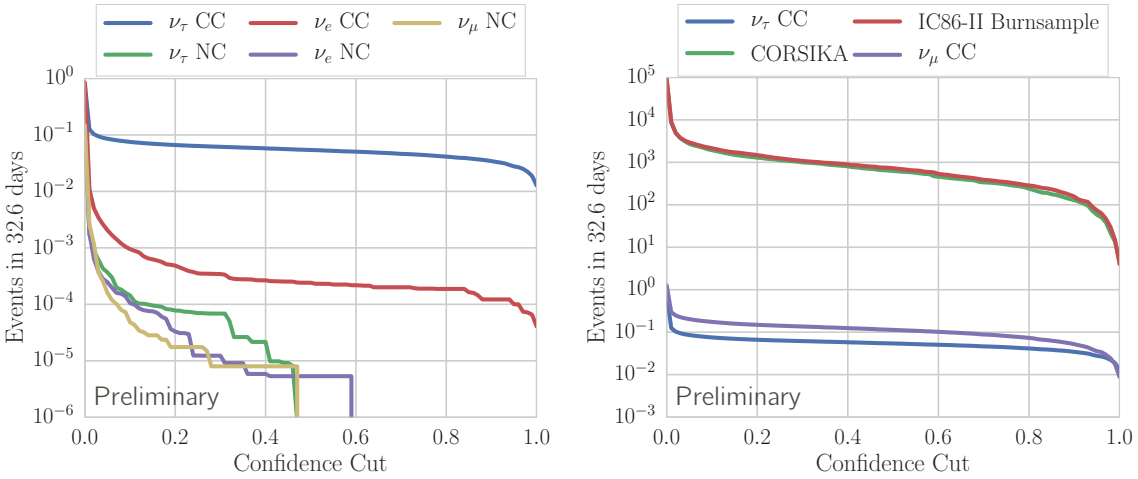
(b) Quality parameters as a function of the confidence cut. The shaded areas cover the central 68 % interval determined from the distribution in a 10-fold cross validation.

Figure 1: Results of the double pulse selection.

Since preselecting signal waveforms requires Monte Carlo truth information, the estimation of event rates resulting from the double pulse selection has to be carried out on the whole simulation dataset. For this estimation only waveforms, that the random forest had no access to during the training process, are used. Figure 2 shows the rate estimates after the double pulse selection. As depicted the signal dominates the cascade-like background but is still buried under four orders of magnitude of track-like background events for a confidence cut of 0.2. These track-like components have to be discriminated in a separate process.

3 Outlook

In this work a search for high-energy astrophysical tau neutrinos is performed. The discrimination between cascade-like background waveforms and double pulse waveforms already shows promising results. The remaining signal events are still dominated by track-like background events, which have to be removed to reach a high sensitivity to charged current tau neutrino events.



(a) Passing signal events (ν_τ CC) with all cascade-like background components, charged current electron neutrino events (ν_e CC) and all flavor neutral current events (ν_X NC).

(b) Passing signal events (ν_τ CC) with all track-like background components, charged current muon neutrino events (ν_μ CC) and atmospheric muons (CORSIKA). In addition the passing events of the IC86-II burnsample are shown as a comparison.

Figure 2: Events with at least one waveform passing a confidence cut in one year of burnsample livetime (32.6 d) as a function of the confidence cut.

References

[1] M. G. Aartsen et al. "Observation of High-Energy Astrophysical Neutrinos in Three Years of IceCube Data". In: *Phys. Rev. Lett.* 113 (2014), p. 101101.

[2] M. G. Aartsen et al. "Search for Astrophysical Tau Neutrinos in Three Years of IceCube Data". In: *Physical Review D* 93 (2016), p. 022001.

[3] A. Achterberg et al. "First year performance of the IceCube neutrino telescope". In: *Astroparticle Physics* 26.3 (2006), pp. 155–173.

[4] D. F. Cowen and the IceCube Collaboration. "Tau Neutrinos in IceCube". In: *Journal of Physics: Conference Series* 60.1 (2007), p. 227.

Unfolding of the Neutrino Energy Spectrum for Different Source Types

Thorben Menne
Experimentelle Physik 5b
Technische Universität Dortmund
thorben.menne@tu-dortmund.de

In this work three topics are addressed. The first being related to the recent discovery of neutrinos of astrophysical origin by the IceCube neutrino observatory [2]. In a follow-up analysis no significant source of these high energy events could be found. This analysis aims to find a signal from lower energy neutrinos originating from these positions. In a stacked and single catalog search all positions are tested against this hypothesis.

The second topic is to investigate the unfolding method described in [6] in detail and compare it to other methods for obtaining a true distribution from a measured one. Emphasis is put on the comparison of this method to the often referenced *forward folding* technique. Furthermore applicability and implications of the handling of systematic errors within the unfolding framework are investigated.

The third topic is still devoted to an all sky stacking search of known source positions followed by an unfolding of the neutrino spectrum. This analysis aims to improve the flux limits originating from an earlier analysis based on data taken with the IceCube detector in the 59 string configuration. The goal is to achieve an improved unfolding procedure by combining three years of IC86 data with better statistics and smaller systematic errors. The results from the unfolding studies can be used to achieve correct and transparent incorporation of systematic errors in addition to the already understood handling of statistical errors.

1 Introduction

The IceCube detector is a cubic kilometer sized neutrino detector located at the south pole. It consists of 5160 digital optical modules (DOMs) mounted on 86 strings in depths between 1455 m to 2450 m directly in the antarctic ice. Additionally 81 instrumented ice tanks, the IceTop array, are installed on the surface to detect air showers [3]. Neutrinos can be measured indirectly via secondary particles produced in charged or neutral current interactions of the neutrinos with the surrounding matter. Those secondary particles then create tertiary particles in further interactions with the ice. All produced charged particles that are faster than the speed of light in the surrounding medium emit Cherenkov light which is measured and used to reconstruct the original neutrino properties.

With large neutrino detectors like IceCube the physics of astrophysical sources can be studied by measuring neutrinos originating from them. Currently the neutrino signal for single source positions on the sky is too low to be seen against the large background of atmospheric neutrinos. With the stacking method multiple sources of the same type are be bundled into one catalog. The combined signals have a better signal over background ratio than a single source so the time needed to measure a significant signal from a class of sources can be reduced. [5]

However, the measured energy distribution is not the true sought after neutrino energy. It is convolved with the charged current cross section and effects from limited detector resolution and acceptance. The used unfolding procedure provides a model independent estimation of the true neutrino energy. [1, 6]

2 General Point Source Likelihood Approach and Stacked HESE Search

The general likelihood approach in IceCube for a point source search is described in [7]. The most basic likelihood function $\mathcal{L}$ can be expressed as

$$\mathcal{L}(n_S) = \prod_{i=1}^{N} \left[\frac{n_S}{N} \, S_i + \left(1 - \frac{n_S}{N}\right) B_i \right] \tag{1}$$

where n_S is the number of assumed signal events, S_i, B_i the expected signal and background probabilities per event i and N the number of total events.

To search for a priori unknown point source in the sky the likelihood can be evaluated on each point on a grid finer than the spatial detector resolution. For a given point on the sky the likelihood value is compared to the value obtained with a background only hypotheses. The background is usually estimated by scrambling real data events in the

azimuth which is justified by the assumption of only having a small amount of signal events.

For a stacked analysis only a small modification to the likelihood function (1) is needed [4]. The signal hypotheses is replaced by a sum over M weighted hypotheses for each source in the catalog

$$S_i \rightarrow S_i^{\text{tot}} = \frac{\sum_{j=1}^{M} W^j R^j S_i^j}{\sum_{j=1}^{M} W^j R^j} \tag{2}$$

where W^j is the relative theoretical weight, R^j the relative detector acceptance for each source and S_i^j the signal probability for each event i regarding source j.

In this work the predefined source positions are given by the positions of measured high energy starting events, which are known to be of astrophysical origin [2]. It is then searched for spatial clustering of lower energy events in a larger sample at those source locations. The likelihood approach has to be modified, as the stacked sources are not point-like, but have a finite angular size. To take also irregular shaped structures into account, a whole sky map of the reconstructed likelihood space from the directional fit of the high energy event is used as the spatial signal PDF for each source.

3 Unfolding

For the unfolding of the neutrino energy spectrum the algorithm described in [6] and implemented for example in the software TRUEE [8] is used. The advantage of this method is, that a model independent energy spectrum can be obtained. The unfolding step is done by building a so called *response matrix* on Monte Carlo data. This matrix incorporates all measurement effects mapping the true sought after observable to the measured variables correlating with the true variable. Several tests exist to ensure that the unfolding procedure yields reliable and stable results.

In this work effort is made to understand the underlying algorithm in more depth. For this systematic studies and comparisons to other unfolding methods are made. The focus is set on comparing the unfolding method to *forward folding* methods, which are binned poissonian likelihood fits of a given hypothesis directly to the measured data.

Another major approach is to incorporate correct handling of systematic errors in the unfolding method. Especially neutrino detectors or Imaging Atmospheric Cherenkov Telescopes (IACTs) are heavily relying on Monte Carlo simulation data due to a lack of a proper calibration method. The uncertainties in many details of those simulations arise from insufficient or unavailable measurements of the cross sections or properties of surrounding matter thus resulting in great systematic errors. Hence it is very important to correctly estimate and incorporate these systematic errors in the analysis process to give a reliable and statistically correct result.

4 Outlook

In this work three topics are covered. The stacking search with a catalog of high energy starting neutrinos event positions is a new idea to find a long awaited source or source class in the neutrino sky. The in depth study of the unfolding method will provide valuable insight also for many other experiments and analysis. It is also related to the catalog search with a follow up unfolding of the energy spectrum of source classes. A correct incorporation of systematic errors in the result is crucial for a statistically reliable result.

References

[1] M. G. Aartsen et al. The IceCube Neutrino Observatory Part I: Point Source Searches. In *Proceedings, 33rd International Cosmic Ray Conference (ICRC2013)*, pages 24–27, 2013.

[2] M. G. Aartsen et al. The IceCube Neutrino Observatory - Contributions to ICRC 2015 Part II: Atmospheric and Astrophysical Diffuse Neutrino Searches of All Flavors. In *Proceedings, 34th International Cosmic Ray Conference (ICRC 2015): The Hague, The Netherlands, July 30-August 6, 2015*, 2015.

[3] M. G. Aartsen, N. van Eijndhoven, J. C. Groh, F. Huang, H. P. Bretz, L. Classen, J. Daughhetee, G. Yodh, P. Berghaus, G. W. Sullivan, et al. Letter of intent: The precision icecube next generation upgrade (pingu). Technical report, Inst, 2014.

[4] 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. Time-integrated Searches for Point-like Sources of Neutrinos with the 40-string IceCube Detector. *Astrophys. Journal*, 732:18, May 2011.

[5] A. Achterberg et al. On the selection of AGN neutrino source candidates for a source stacking analysis with neutrino telescopes. *Astropart. Phys.*, 26:282–300, 2006.

[6] Volker Blobel. An Unfolding method for high-energy physics experiments. In *Advanced Statistical Techniques in Particle Physics. Proceedings, Conference, Durham, UK, March 18-22, 2002*, pages 258–267, 2002.

[7] Jim Braun, Jon Dumm, Francesco De Palma, Chad Finley, Albrecht Karle, and Teresa Montaruli. Methods for point source analysis in high energy neutrino telescopes. *Astropart. Phys.*, 29:299–305, 2008.

[8] 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. *Nuclear Instruments and Methods in Physics Research A*, 697:133–147, January 2013.

Optical and Temporal Resolution of the FACT Telescope

Maximilian Nöthe
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
maximilian.noethte@tu-dortmund.de

In Imaging Atmospheric Cherenkov Astronomy, Muon ring images provide a valuable event class to estimate detector properties.

After selecting those events using supervised machine learning techniques, the data can be used to estimate the optical resolution from the width of the rings and the temporal resolution from the standard deviation of the mean arrival times in the pixels.

The results show a mismatch between simulated and measured data. Two strategies are pursued to tackle this problem: first the measured data is calibrated with a newly developed technique then the remaining mismatch is resolved by adjusting the simulation.

1 Muons in gamma-ray Cerenkov Astronomy

Muons, albeit background for the analysis of gamma-ray induced extensive air showers, provide a valuable event class for the estimation of detector properties. As visualized in Figure 1, they produce ring images in the camera plane. The radial width of these rings can be used to estimate the optical resolution of the telescope. The time in which photons of a muon ring image arrive in the camera plane is also short compared to the temporal resolution of the telescope. Thus, muon ring images can be used to estimate also the temporal resolution of the entire detection system up to the analysis software.

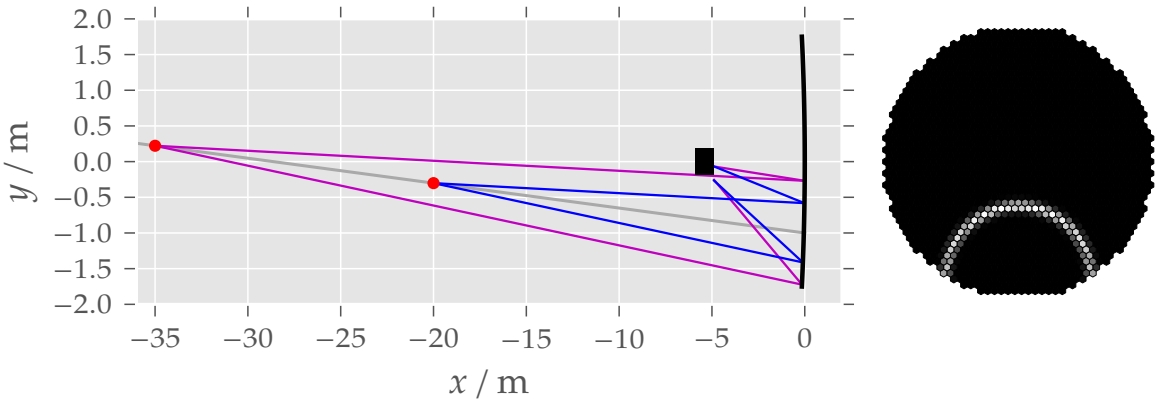


Figure 1: A muon emitting Cherenkov radiation moves towards a telescope, producing a ring image. Because of the offset to the optical axis of the telescope, the ring is not lit evenly and because of the angle to the optical axis, the ring center is not in the camera center. [3]

2 Selecting muon ring images from FACT data

To get a muon ring dataset, supervised machine learning techniques are used. In this case, a Random Forest classifier is trained with simulated Muon ring images as signal class and Proton induced air showers as background class. As proton induced showers can also produce muons and thus muon ring images, all events where more than 80 % of the photons were emitted by muons are removed from the background training data.

The ROC curve for classifiers trained on different feature sets is shown in Figure 2.

3 Optical and Temporal resolution

The optical resolution of the telescope can be estimated by the width of the muon rings. This is done by performing a maximum likelihood fit with a gaussian profile on the ring images. In Figure 3 the standard deviation of the gaussian profile is hown for two time periods: one before a mirror alignment in May 2014 and one for the time after.

By calculating the standard deviation of the mean arrival time of the photons in the pixels of the FACT image sensor, the overall temporal resolution of the FACT telescope and the analysis chain is estimated.

Figure 4 shows the result for measured data with and without correction for offsets between the pixels.

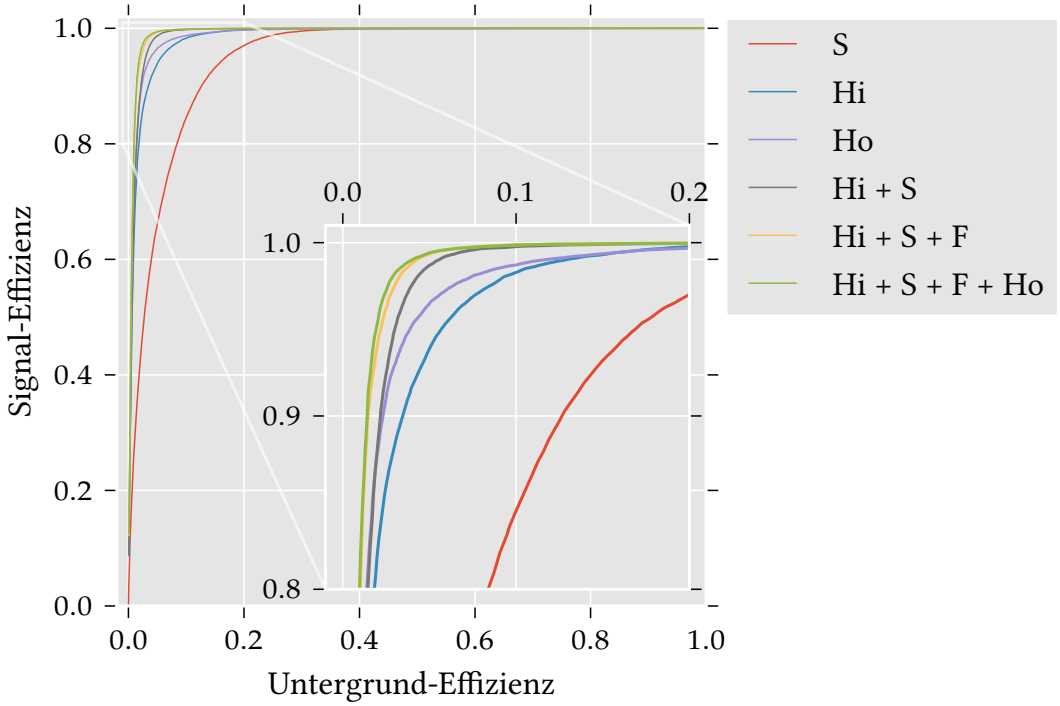


Figure 2: ROC curves for a Random Forest classifier using different feature sets. The best performance is reached with calculating additional features from a rather time consuming Hough-transform on the camera images [3]

.

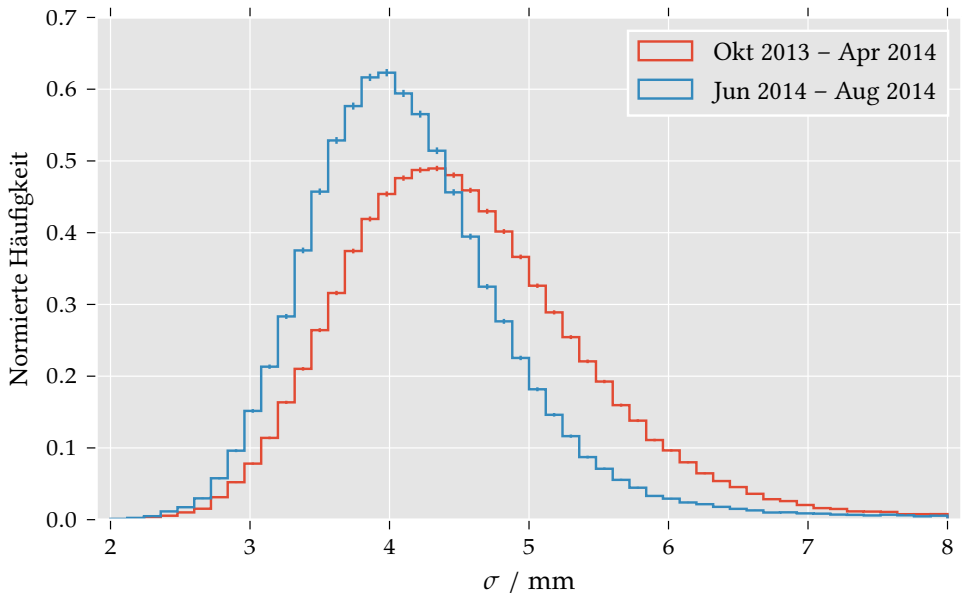


Figure 3: Width of the muon ring images estimated by a gaussian likelihood fit for the time before and after a mirror alignment in 2014. [3]

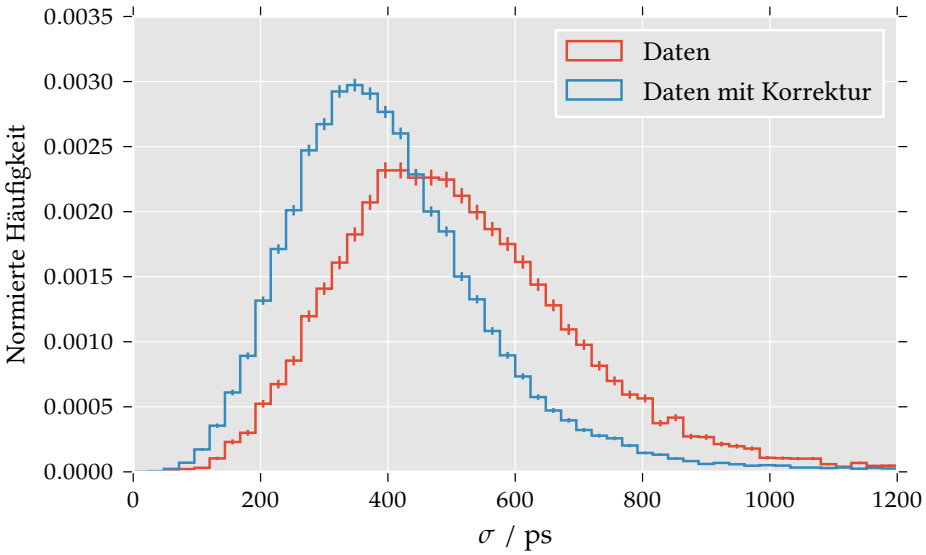


Figure 4: Standard deviation of the arrival time of Muon ring photons with and without correction for inter-pixel offsets measured with lighpulser calibration runs.

References

[1] H. Anderhub et al. Design and operation of fact – the first g-apd cherenkov telescope. *JINST*, 8(06):P06008, 2013.

[2] A. Biland et al. Calibration and performance of the photon sensor response of fact — the first g-apd cherenkov telescope. *JINST*, 9(10):P10012, 2014.

[3] Maximilian Nöthe. Ermittlung des zeitlichen und optischen auflösungsvermögens des first g-apd cherenkov telescope mithilfe von myon-ringen, 2016.

Reconstructing the Atmospheric Muon Neutrino Energy Spectrum with IceCube DeepCore

Philipp Schlunder
Experimentelle Physik 5b
Technische Universität Dortmund
philipp.schlunder@tu-dortmund.de

In this work the reconstruction of the energy spectrum of atmospheric muon neutrinos within an energy range of around 10 GeV to 1000 GeV is targeted. Using simulations for creating an event-selection chain enabling the tagging of muon neutrino candidates within the data taken by IceCube DeepCore. Inferring a neutrinos energy can only be done by measuring the secondary particles from neutrino interactions. Energy estimates of these secondary particles are used to reconstruct the initial neutrino energy by accounting for effects introduced by the measuring process. This analysis aims to bridge the gap within the energy spectrum of the atmospheric neutrino flux between current measurements from Super-Kamiokande [Ric+15] and previous IceCube [Aar+15] results. Therefore a new method combining event selection and energy reconstruction is to be developed. This report gives an overview of the initial event selection.

1 Introduction

IceCube is a neutrino observatory currently consisting of three sub detectors: the in-ice-array, DeepCore and IceTop. The in-ice-array and DeepCore are constituted of 5160 digital optical modules (DOMs), deployed on 86 strings in a depth between 1455 m to 2450 m in the antarctic ice. The instrumented volume covers around one cubic kilometer

of ice. The the in-ice-array has a string spacing of around 125 m and an inter-dom-distance of 17 m. DeepCore is a more dense instrumented array with an inter-dom-distance of 7 m to 10 m and a string-to-string spacing of around 60 m, thus allowing to measure neutrinos with energies as low as 10 GeV. IceTop is an ensemble of 81 tanks placed on the surface each containing two DOMs to measure air showers. [Ach+06]

Neutrinos only interact with matter via weak force, hence only indirect measurements via secondary leptons produced in weak interactions are possible. These leptons are able to excite the surrounding ice, which emits Cerenkov light during relaxation. Measurements of these photons allows to reconstruct a leptons energy. Using the reconstructed lepton energies it is possible to infer the initial neutrino energies on a statistical basis.

Reconstructing the true neutrino energy is complicated by the fact, that its distribution is folded with effects like limited resolution and acceptance introduced by the detector as well as uncertainties raised by the weak interaction. One method to gain the true distribution of the measured lepton flux is called unfolding [Mil+13].

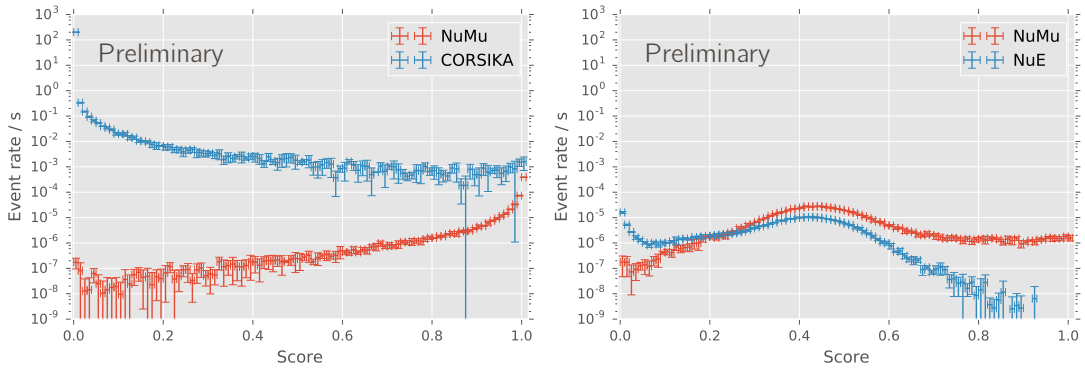
2 Muon neutrino event selection

Before reconstructing the neutrino energy from lepton energy estimates a pure muon neutrino sample has to be obtained. Atmospheric muon neutrinos are produced via interactions of cosmic rays with nuclei in the earths atmosphere. During these interactions not only muon neutrinos but also electron neutrinos and muons are created, besides other particles not relevant for this selection.

Muons are able to reach the detector if their energy is high enough. Their event topology within IceCube is the same as for muons from muon neutrinos interactions. The signal to background ratio of these two particles is around $1{:}10^6$. This ratio can be reduced with pre-selection cuts to a ratio of $1{:}10^3$ within DeepCore.

Following a previous analysis [Sch15] the separation of neutrino induced muons from atmospheric muons in the given energy range is done in multiple stages. First neutrino induced events are separated from atmospheric muons in order to reduce the vast amount of atmospheric muons. Since the topology of electron and muon neutrinos is very similar for energies between 10 GeV and 100 GeV a second separation step to differentiate between these types is needed.

A first look into these two separation steps is shown in figure 1a and figure 1b. A classifier is trained on IceCube Monte Carlo data sets. After applying standard reconstructions to the data, attributes with constants and high correlations are removed. For both separation steps a random forest with 100 trees able to sample 5 attributes per node was trained within a five-fold-cross-validation. The distributions display event rates averaged over the five validation steps, using the standard deviation as an error.



(a) Initial separation step to separate muon neutrinos (red) from atmospheric muons (blue).

(b) Potential follow up separation to diffentiate between muon (red) and electron (blue) neutrinos. Assuming noch remaining atmospheric muon background.

Figure 1: Mean expected event rates for given random forest scores. Averaged are the rates from a five-fold-cross-validation. Shown are the first two separation stages.

Considering the event rates of the initial separation step (figure 1a) after a given cut in the random forest score it can be seen, that improvements to the atmospheric muon suppression are still necessary. Also the results already on pair with similar separations within DeepCore. Whereas it has to be noted, that the standard pre-selection cuts mentioned before, weren't applied for this selection. It turns out, that the human chosen pre-selection cuts can be dropped if a cut in the random forest score above ~ 0.05 is applied.

The similarity between the topologies of electron and muon neutrinos are reflected in the overlap of the maxima within the shown distributions in figure 1b. For random forest scores above 0.6 a purity of above 0.9 is achieved. Since the efficiency is quite low at this point, with values around 0.16, further techniques to increase the separability have to be applied.

3 Energy reconstruction

An initial energy reconstruction for the muon neutrino spectrum will be performed on the selected muon neutrino events with the software package TRUEE [Mil+13]. TRUEE uses a model independent unfolding approach to infer the true energy distribution from three given attributes. This method builds a response matrix on Monte Carlo data to characterize information about named detector effects. Folding the response matrix with

up to three input attributes of the measured data results in a reconstruction of the true energy distribution.

Uncertainties of the resulting spectrum are dependent on the amount of muon neutrino events available. Hence a quality measure for the event selection can be developed with regards to the energy distribution selected.

4 Outlook

The presented results are a first glance at the different event selection stages needed to create a muon neutrino sample that will be used to reconstruct the atmospheric muon neutrino energy spectrum in the energy range of 10 GeV to 1000 GeV. Further investigations have to be made in order to understand the energy distributions of selected events and the needed amount of muon neutrinos necessary for an unfolding with reasonable uncertainties. Enhancements to the event selection are expected due to feature engineering and optimization of used classifiers.

References

[Aar+15] M. G. Aartsen et al. "Unfolding measurement of the Atmospheric Muon Neutrino Spectrum using IceCube-79/86". In: *PoS(ICRC2015)* 1098 (2015).

[Ach+06] A. Achterberg et al. "First year performance of the IceCube neutrino telescope". In: *Astroparticle Physics* 26.3 (2006), pp. 155–173. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2006.06.007. URL: http://www.sciencedirect.com/science/article/pii/S0927650506000855.

[Mil+13] N Milke et al. "Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 697 (2013), pp. 133–147.

[Ric+15] E. Richard et al. "Measurements of the atmospheric neutrino flux by Super-Kamiokande: energy spectra, geomagnetic effects, and solar modulation". In: November (Oct. 2015), pp. 1–30. arXiv: 1510.08127. URL: http://arxiv.org/abs/1510.08127.

[Sch15] Philipp Schlunder. "Multivariate Selektion mit DeepCore gemessener atmosphärischer Myon-Neutrinos". Masterarbeit. TU Dortmund, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C4

Regression approaches for large-scale high-dimensional data

Katja Ickstadt Christian Sohler

On choosing observations in the case of linear regression with missing y-values

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

Setting

Linear regression is a method that is frequently employed in statistics and computer science. In the usual setting, a regression model is given by Equation 1:

$$Y = x\beta + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^n$ is a random variable containing the values of the response or dependent variable and $x \in \mathbb{R}^{n \times d}$ is a matrix containing the values of the d independent variables. $\beta \in \mathbb{R}^d$ is the parameter vector one wishes to estimate, β is assumed to be unknown, but fixed. The last component, $\varepsilon \in \mathbb{R}^n$, is a random vector that is unobservable and introduces unsystematic measurement error. ε is assumed to follow a normal distribution, $\varepsilon \sim N(0, \sigma^2 I_n)$, which means that Y in turn is also normally distributed with $Y \sim N(x\beta, \sigma^2 I_n)$.

In the classical setting, β is estimated using realisations y of Y and values of x, often employing the ordinary least squares estimator. We now deviate from the classical setting by assuming that the values of y are unknown, but observable. I.e., every y-value can be obtained, but this is associated with a cost per observation. For that reason, it is our aim to build a model with good fit while using only $k < n$ of the y-values.

Leverage Scores

Leverage scores are one possibility to assess the importance of observations. They are solely based on the x-values. Leverage scores can be calculated via the hat matrix, H,

$$H = x(x'x)^{-1}x'. \tag{2}$$

The leverage score of observation i is defined as the ith diagonal element of H, h_{ii} ($i = 1, \ldots, n$). Roughly speaking, the leverage scores are low for observations that lie near the centre of the data and high for observations that are far away from the centre, thus giving higher weight to observations that are striking in some way.

While leverage scores are defined as given in Equation 2, the calculation is done via a QR decomposition for numerical reasons, see Hoaglin and Welsch (1978) [3].

Other importance measures also employ the y-values. For that reason, these cannot in general by applied in our setting. However, some measures like Cook's distance can be used to decide which areas are not covered well enough by the current model.

Simulation Study

We conduct a simulation study including different methods which were adapted from existing methods in computer science and statistics. To accentuate differences between the methods, we include both cases where the assumptions of the linear regression model are met and cases where some assumptions were violated. The latter ones are the more interesting cases as a simple random sample leads to a solution close to the optimal one if all assumptions are met.

The cases we considered were: all assumptions met with small unexplained variance, all assumptions met with large unexplained variance, heteroscedasticity, outliers in the values of x, outliers in the values of y and some model misspecifications such as a mixture of two models, interaction effects that are not taken into account and systematic variation around the regression line which follows a sine curve.

We varied the number of observations n and the number of variables d, where n was taken from the set $\{1\,000, 2\,000, 5\,000\}$ and d was chosen from $d \in \{5, 10, 20\}$. The number of observations was reduced to $k \in \{50, 100, 200, 500\}$ and in the case of $n = 5\,000$ also to $k = 1\,000$. The simulation study was conducted using R, version 3.2.1 [5].

Methods

In our analysis, we found three methods that generally work well in the setting and the specifiations outlined above. All methods have the common idea of selecting observations that represent observations around them. Consequently, they all avoid selecting observations that are close to each other in terms of x-values as these observations carry less additional information than others in different areas of the data cloud.

The first method, Clusterscore, is a combination of k-means clustering and selection according to the leverage scores. In this approach, the standardised data set is split into k clusters. In each cluster, the observation with the highest leverage score is selected and used to build the linear regression model. The selected observations are weighted proportional to the number of observations they represent. Clusterscore outperforms both selection based on k-means clustering alone and selection based on leverage scores (random as well as deterministic). A possible modification is to obtain only $k^* = \lfloor 0.95k \rfloor$ clusters in the first step. This allows for choosing the remaining $k - k^*$ observations after building the linear model and checking which areas need refinement and additional information. In this case, importance measures that also take into account the y-values such as Cook's distance can be employed. This is especially useful to reduce the influence of single outliers on the model.

The second method, Deterministic Epsilon Method (DEM), is based on an algorithm by Feldman et al. (2011) [2] which is employed to find coresets to train models in the presence of mixture models. In our case, we modified the algorithm as follows. In every step, select the observation with the highest leverage score among all eligible observations. Observations that have been selected in an earlier step become ineligible. As this observation is assumed to represent close-by observations, after each selection we exclude the $\left\lceil \frac{n-k}{k} \right\rceil$ observations closest to the one currently selected. This process is repeated k times. As each of the k observations represents the same amount of observations by definition, weighting them is not necessary.

The third method, Maximal Distance Algorithm (MDA), is similar to an algorithm by Agarwal et al. (2005) [1]. The first observation is selected at random. From the second step on, the distance of all remaining observations to the ones already selected is calculated. The observation which maximises the minimal distance to the selected observations is selected next. This is repeated until k observations are selected. MDA does not take the number of observations around the selected ones into account. For that reason, the selected observations are again weighted proportional to the number of observations they represent.

In the simulation study, three more methods were evaluated in addition to the three methods introduced in this report. For more information on the additional methods, please refer to [4]. A simple random sample of k observations was also included as a baseline comparison.

Results

Our aim is to find a model that is close to the model that would be obtained on the full data set. To assess this, we consider two cost criteria, $C_1 = \sqrt{(\hat{\beta} - \hat{\beta}_k)^T(\hat{\beta} - \hat{\beta}_k)}$ and $C_2 = \frac{\|(X\hat{\beta}_k - y)\|^2}{\|(X\hat{\beta} - y)\|^2}$, where $\hat{\beta}$ is the optimal solution estimated from the full data set and $\hat{\beta}_k$ is estimated from the reduced data set with k observations.

In cases were the assumptions are met, all methods including simple random sample show equally good results. For the more demanding cases where assumptions are violated, differences are more distinct. Generalised over all cases, Clusterscore and DEM exhibit the lowest values in both cost criteria with MDA being a close third. If it is unknown beforehand which assumptions might not be valid, Clusterscore or DEM should be considered to select the observations. As expected, both cost criteria decrease as the number of observations increase.

In our simulation study, we also consider the importance of finding the "right" observations, i.e. finding observations that are influential for the values of the parameter estimates. In general, the importance is hardly influenced by the number of observations in the original data set, while the importance decreases if the number of variables increases. In situations where all assumptions are met, the importance is lower than in situations were some assumptions are violated. The most demanding case is the case of outliers in the y-values, especially if they are spread out along the range of x-values.

References

[1] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

[2] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150, 2011.

[3] David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

[4] Niels Lategahn. Vergleich von methoden zur auswahl von beobachtungen bei regression mit fehlenden y-werten. Master's thesis, TU Dortmund, 2016.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

Sampling and Coresets for efficient Poisson regression

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

Our current research aims at extending methods for efficient large-scale linear ℓ_p-regression to *generalized linear regression* models such as Poisson regression. We follow the approach of obtaining small size coresets by subsampling and reweighting the data. The subsequent calculations can be carried out more efficiently on the coreset leading to fast approximate solutions. We show that in general it is not possible to have any randomized sublinear coreset construction. However, in well-suited situations where the data follows the assumptions of the generalized linear model up to some outliers, it seems that ℓ_2 leverage score sampling can help identifying the outliers and thus yield an heuristic approach to constructing coresets of small size.

Introduction and preliminaries

A *generalized linear regression* model takes the form $\mathbf{E}[Y|x] = h(\beta^T x)$ for a non-linear link function h. For example, the Poisson regression model is frequently used for predicting non-negative integer valued count variables Y from independent real variables x. The model is defined using the exponential function as its link function $h(\cdot) = \exp(\cdot)$. Note that we can assume $Y > 0$. Otherwise, by linearity we have $\mathbf{E}[Y+1|x] = \mathbf{E}[Y|x] + 1$ and thus we have that $\mathbf{E}[Y+1|x] = \exp(\beta^T x)$ is equivalent to $\mathbf{E}[Y|x] = \exp(\beta^T x) - 1$. This means we can still predict Y using the transformed variable $Y' = Y + 1$ by applying the inverse transform to the usual prediction.

Given an observed data set $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{N}^n$ we would like to estimate the underlying true parameter β^*. This can be done by maximizing the likelihood of observing the

fixed data over the parameter space $\beta \in \mathbb{R}^d$. From an optimization point of view, this is equivalent to the analytically and computationally more convenient task of minimizing the convex negative log-likelihood $\ell(\beta|X,Y) = \sum_{i=1}^{n} \exp(\beta^T x_i) - y_i \cdot \beta^T x_i$, which is presented here omitting the regularization constants which do not depend on $beta$. For the sake of brevity we will write $\ell(\beta)$ omitting the data from the notation.

A *coreset* for Poisson regression is a smaller sized substitute data set that mimics the behavior of the original log-likelihood in the following sense.

Definition 1. Given a data set $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{N}^n$, an η-coreset for Poisson regression is a succinct representation $S(X) \in \mathbb{R}^{k \times d}, S(Y) \in \mathbb{N}^k$ for $k \ll n$ of the data such that

$$\forall \beta \in \mathbb{R}^d : \eta^{-1} \cdot \ell(\beta|X,Y) \leq \ell(\beta|S(X),S(Y)) \leq \eta \cdot \ell(\beta|X,Y)$$

Coresets for regression problems are often constructed by subsampling and reweighting the original data [2, 3]. In principle it can be any other succinct summary of the data satisfying the above upper and lower bound guarantees for some small η and all $\beta \in \mathbb{R}^d$.

Unfortunately it turns out that in general there cannot exist coresets of sublinear size with reasonable approximation guarantee.

Theorem 2. *Any η-coreset for Poisson regression with $\eta < \frac{\exp(\frac{n}{4})}{\sqrt{2}n^2}$ uses $\Omega(n)$ bits of space.*

Proof. We reduce from the indexing problem which is known to have $\Omega(n)$ one-way randomized communication complexity. Alice is given a vector $b \in \{0,1\}^n$. She produces for every i with $b_i = 1$ the points $x_i = (r \cdot \omega^i, 1) \in \mathbb{R}^3$, where $\omega^i, i \in \{0, \ldots, n-1\}$ denote the n^{th} unit roots in the plane, i.e., the vertices of a regular n-polygon of radius $r = n^3$ in canonical order. The corresponding counts are set to $y_i = 1$. She builds and sends an η-coreset of size $s(n)$ to Bob, whose task is to guess the bit b_j. He chooses β in such a way that it points away from the origin, has unit norm and is normal to a hyperplane passing exactly through $r \cdot \omega^{j-1}$ and $r \cdot \omega^{j+1}$. Now we know that if $b_j = 0$, the cost is at most $\ell(\beta) \leq n \cdot 2r = 2n^4$. If $b_j = 1$ then x_j is in the expensive halfspace and at distance at least $r \cdot (1 - cos(\frac{2\pi}{n})) \geq r \cdot \frac{1}{n^2} = n$. So the cost is bounded below by $\ell(\beta) \geq \exp(n) - n \geq exp(\frac{n}{2})$. Given $\eta < \frac{\exp(\frac{n}{4})}{\sqrt{2}n^2}$ Bob can distinguish these two cases based on the coreset only and consequently $s(n) = \Omega(n)$. □

Note that the bound is given in bit complexity, but assuming we have a sampling based coreset and every data point can be expressed in $O(d \log n)$ bits, this means we still have a lower bound of $k = \Omega(\frac{n}{d \log n})$ samples. At this point it seems very likely that a similar argument can be used to rule out any $o(n)$ space constant approximation algorithm. This remains an open problem for now.

Leverage score sampling for generalized linear models

Despite the above negative result, there is still hope for constructing coresets if the data follows the assumptions of the underlying linear model. Consider as an extreme case that all data points of a linear model lie exactly in a linear subspace. Then any uniform sample that is large enough to preserve the rank will do the job. Such behavior has been observed in [4] for linear regression. On the other hand, the Poisson model will be very sensitive to outliers in the exponential half space and behave less sensitive on outliers in the other half space causing only linear cost. Nevertheless, these will also have a crucial function, because if there are outliers only on one side of the hyperplane then the model will behave like a robust ℓ_1-regression model. So outliers in both directions will be interesting to us. Our approach for Poisson regression is given in the following algorithm which is inspired by the similar approach of [1] for ℓ_p-regression. In a first step the data is linearized in the following sense. Recall that by definition $Y' = \log Y$ follows a linear model. So we apply the logarithm to the variable Y given in the input data. Now we can compute or approximate the ℓ_2-leverage scores on the transformed data X, Y' from its QR-decomposition. See [2] for computational details on leverage scores and their use for outlier detection. In Line 3 we take k_1 samples proportional to the distribution given by the leverage scores. This sample is our first heuristic coreset, together with properly chosen weights to make up an unbiased estimator for the likelihood function.

Algorithm for Poisson regression:

1. Reduce data to linear model $\log(\mathbf{E}[Y \mid X]) = X\beta$
2. Compute ℓ_2-leverage scores $\lambda_i, i \in [n]$
3. Take k_1 samples proportional to the leverage scores by assigning weight $w_i = \frac{1}{p_i}$ with probability $p_i = \frac{\lambda_i}{\sum \lambda_j}$ and $w_i = 0$ otherwise
4. Solve *weighted* convex optimization problem on the sample to obtain $\tilde{\beta} = \operatorname{argmin}_\beta \sum_{i=1}^n w_i \left(\exp(\beta' x_i) - y_i \beta' x_i\right)$
5. Take k_2 samples proportional to the log-likelihood by assigning weight $w_i = \frac{1}{q_i}$ with probability $q_i \propto \max\left\{\frac{f_i(\tilde{\beta})}{\sum f_j(\tilde{\beta})}, p_i\right\}$ and $w_i = 0$ otherwise
6. Solve *weighted* convex optimization problem on sample and output $\hat{\beta} = \operatorname{argmin}_\beta \sum_{i=1}^n w_i \left(\exp(\beta' x_i) - y_i \beta' x_i\right)$

We can perform approximate tasks like maximum likelihood estimation or even take a Bayesian regression sample from the distribution on the parameter space. However, the predictions will still have some bias from the original ones and there will be a high variance around the true predictions. This is due to the leverage scores, which by symmetry, treat outliers on both sides of the hyperplane as equally important. Contrarily the log-likelihood puts only linear cost on one side and exponential cost on the other. To reduce

the remaining bias and variance, we obtain a maximum likelihood estimator $\tilde{\beta}$ on the weighted coreset in Line 4. Then we take another sample of size k_2 and this time, in addition to the leverage scores, we use the contribution of single observations to the log-likelihood in the sampling distribution to get all of the outliers and to put special emphasis on the ones that cause large cost. This yields another coreset which is more accurate than the previous one. In step 6, again we can solve the weighted convex optimization problem or draw MCMC-samples from the parameter distribution approximately using the small coreset only. We have performed preliminary experiments with artificial data that inhibits the model assumptions up to a small constant number of influential outliers. Due to their importance to the full model there is no hope to get any reasonable approximation by a uniform sublinear sample. The leverage scores are capable of identifying the relevant points. We have empirically quantified the sample sizes needed to get $(1+\epsilon)$-approximations of the predictions made approximate parameters. The bounds are $k_1 = O(\frac{d}{\epsilon^{1.5}})$ and $k_2 = O(\frac{d^{0.6}}{\epsilon^{1.5}})$. Running the whole algorithm is mainly beneficial when dealing with higher dimensions. For practical purposes, running only the first three lines already yields the desired accuracy with high probability.

Conclusion and future research

We will have to find out more on the limitations of sublinear algorithms for Poisson regression. Our lower bound indicates that such algorithms will most likely not exist in general. However, due to the good practical results, more theoretical work should be done under the natural model assumptions. Moreover our methods can be used in practice for efficiently estimating count data in cooperation with other projects.

References

[1] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

[2] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[3] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of SODA*, pages 1127–1136, 2006.

[4] Ping Ma, Michael W. Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C5

Real-Time Analysis and Storage of High-Volume Data in Particle Physics

Bernhard Spaan Jens Teubner

Measurement of *CP* observables in $B_s^0 \to D_s^{\mp} K^{\pm}$ decays with the LHCb experiment

Ulrich Eitschberger
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ulrich.eitschberger@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally, particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

A new physics analysis is being performed that studies time-dependent charge-parity (CP) violation in $B_s^0 \to D_s^{\mp} K^{\pm}$ decays. CP violation is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. The analysis can be used to demonstrate the need for an improved processing of the data collected by the LHCb detector.

The time-dependent analyses of neutral B^0 and B_s^0 tree-level decays [3–6] provide sensitivity to the measurement of the CKM angle $\gamma = \arg\left[-\frac{V_{ud}\,V_{ub}^*}{V_{cd}\,V_{cb}^*}\right]$, without theoretical complications of penguin contributions or other hadronic uncertainties. The time-dependent analysis of $B_s^0 \to D_s^{\mp} K^{\pm}$ decays provides a complementary measurement of γ with respect to the time-integrated analyses using the $B \to DK$ decays, where the charged and neutral B modes need different auxiliary inputs in the final determination of γ.

Sensitivity to the CKM angle γ arises from the interference of $b \to u$ and $b \to c$ amplitudes (see Fig. 1a). In the case of the time-dependent γ measurements with neutral B-mesons, the B^0 (B_s^0) mixing provides the interfering amplitudes. In practice, the sum of γ and the mixing phase are thus measured, namely $\gamma + 2\beta$ and $\gamma - 2\beta_s$ in the B^0 and B_s^0 systems, respectively (see Fig. 1b). The CP-parameters $\sin 2\beta$ and $2\beta_s$ are measured accurately, and are used as external input to the determination of the weak phase γ. The measurement uses a data set corresponding to 1.0 (2.0) fb^{-1} of pp collisions recorded with the LHCb detector at $\sqrt{s} = 7\,(8)$ TeV in 2011 (2012). It is expected to yield the precise determination of $\gamma - 2\beta_s$ in these decays.

The decay-time-dependent decay rates of the initially produced flavour eigenstates

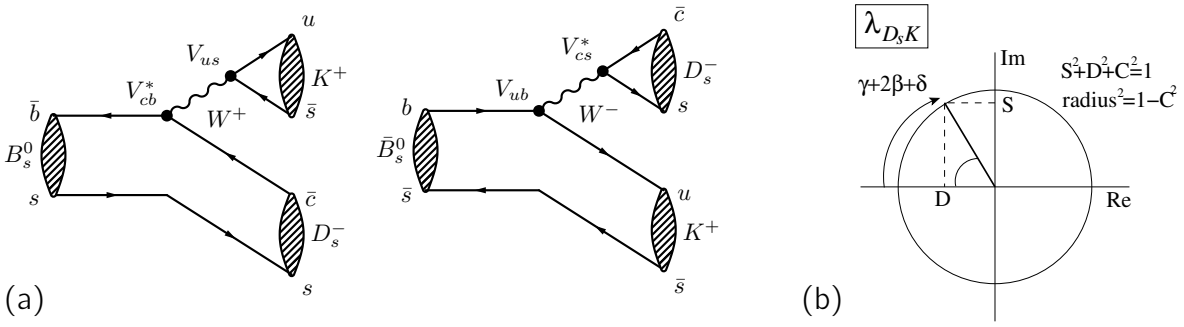


Figure 1: (a) Both the B_s^0 and $\overline{B}_s^0$ meson can decay to the same final state $D_s^- K^+$, which allows for interference effects when the B_s^0 oscillates. The relative phase between the two amplitudes leads to sensitivity to $\gamma = \arg\left[-\frac{V_{ud}\,V_{ub}^*}{V_{cd}\,V_{cb}^*}\right]$. (b) Illustration of the parameter λ_{D_sK} in the complex plane, together with the CP-asymmetry observables S, C and D mentioned in the text.

$|B_s^0(t=0)\rangle$ and $|\overline{B}_s^0(t=0)\rangle$ are proportional to

$$\frac{\mathrm{d}\Gamma_{B_s^0\to f}(t)}{\mathrm{d}t} \propto e^{-\Gamma_s t}\left[\cosh\left(\frac{\Delta\Gamma_s t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma_s t}{2}\right) + C\cos\left(\Delta m_s t\right) - S_f \sin\left(\Delta m_s t\right)\right], \quad (1)$$

$$\frac{\mathrm{d}\Gamma_{\overline{B}_s^0\to f}(t)}{\mathrm{d}t} \propto e^{-\Gamma_s t}\left[\cosh\left(\frac{\Delta\Gamma_s t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma_s t}{2}\right) - C\cos\left(\Delta m_s t\right) + S_f \sin\left(\Delta m_s t\right)\right], \quad (2)$$

where Γ_s is the average B_s^0 decay width, and $\Delta\Gamma_s$ is the positive [1] decay-width difference between the heavy and light mass eigenstates in the B_s^0 system. Similar equations can be written for the *CP* conjugate decays replacing S_f by $S_{\bar{f}}$, and D_f by $D_{\bar{f}}$, while $C = -C_{\bar{f}}$ under the assumption of no *CP* violation in either the decay or mixing amplitudes. The *CP* observables are related to the magnitude of the amplitude ratio $r_{D_sK} \equiv |\lambda_{D_sK}| = |A(\overline{B}_s^0 \to D_s^- K^+)/A(B_s^0 \to D_s^- K^+)|$, the strong phase difference between the favoured and suppressed amplitudes δ, and the weak phase difference $\gamma - 2\beta_s$ by the following equations:

$$\begin{aligned} C &= \frac{1 - r_{D_sK}^2}{1 + r_{D_sK}^2}, \\ D_f = \frac{-2r_{D_sK}\cos(\delta - (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}, &\quad D_{\bar{f}} = \frac{-2r_{D_sK}\cos(\delta + (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}, \\ S_f = \frac{2r_{D_sK}\sin(\delta - (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}, &\quad S_{\bar{f}} = \frac{-2r_{D_sK}\sin(\delta + (\gamma - 2\beta_s))}{1 + r_{D_sK}^2}. \end{aligned} \quad (3)$$

The sinusoidal *CP* observables can only be measured using signal candidates in which the initial flavour of the B_s^0 meson is determined, a process known as "flavour tagging". By contrast, all signal candidates provide sensitivity to the hyperbolic *CP* observables. The analysis strategy largely follows that described in Ref. [2]. The kinematically similar mode $B_s^0 \to D_s^-\pi^+$ is used as a control channel which helps in the determination of the decay-time-dependent efficiency and flavour tagging performance. A two-stage likelihood fit is performed: first the signal and background are separated and the so called *sWeights* [7] obtained; second, these *sWeights* are used to statistically subtract the background, and the decay-time-dependent *CP* observables defined in Eq. 3 are measured by fitting the signal Probability Density Function (PDF) to the decay-time distribution. The analysis is far advanced and close to its publication.

The LHCb experiment will multiply the amount of collected data during the upcoming LHC Runs. To be able to fully exploit the potential of these large datasets analysis tools need to be scalable in terms of both storing and processing. A measurement of

CP violation as described above could profit in several ways from making use of data handling technologies in the considered open source frameworks Apache Hadoop and Apache Flink. Improvements in the handling of large amounts of data would make looser preselections possible and thus retain more potential signal candidates. Necessary to achieve this is the transformation of the experiments data from existing data formats into modern versatile data formats like HDF5. A cut-based offline selection could then be replaced by a selection based on more sophisticated machine learning techniques. This makes it necessary to perform multiple computationally intensive trainings of Neural Nets or Boosted Decision Trees on large data samples. Network bottlenecks due to heavy data access can be overcome by exploiting the individual storage of cluster nodes. Making use of the MapReduce programming model or even more advanced ideas at the same time should lead to reasonable processing times.

References

[1] R. Aaij et al. Determination of the sign of the decay width difference in the B_s^0 system. *Phys. Rev. Lett.*, 108:241801, 2012.

[2] R. Aaij et al. Measurement of *CP* asymmetry in $B_s^0 \to D_s^{\mp} K^{\pm}$ decays. *JHEP*, 11:060, 2014.

[3] R. Aleksan, I. Dunietz, and B. Kayser. Determining the CP violating phase γ. *Z. Phys.*, C54:653–660, 1992.

[4] Kristof De Bruyn, Robert Fleischer, Robert Knegjens, Marcel Merk, Manuel Schiller, and Niels Tuning. Exploring $B_s \to D_s^{(*)\pm} K^{\mp}$ Decays in the Presence of a Sizable Width Difference $\Delta\Gamma_s$. *Nucl. Phys.*, B868:351–367, 2013.

[5] Isard Dunietz and Robert G. Sachs. Asymmetry Between Inclusive Charmed and Anticharmed Modes in B^0, Anti-B^0 Decay as a Measure of CP Violation. *Phys.Rev.*, D37:3186, 1988.

[6] Robert Fleischer. New strategies to obtain insights into CP violation through $B_{(s)} \to D_{(s)}^{\pm} K^{\mp}$, $D_{(s)}^{*\pm} K^{\mp}$, ... and $B_{(d)} \to D^{\pm}\pi^{\mp}$, $D^{*\pm}\pi^{\mp}$, ... decays. *Nucl.Phys.*, B671:459–482, 2003.

[7] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Realtime Analysis and Storage for High-Volume Data in Particle Physics

Michael Kußmann
Lehrstuhl für Datenbanken und Informationssysteme
Technische Universität Dortmund
michael.kussmann@cs.tu-dortmund.de

This report will introduce *DeLorean*, a new system architecture for high-volume data processing in the domain of particle physics. *DeLorean* combines the simplicity and performance of relational database technology with the massive scalability of modern cloud execution platforms. Experiments show a four-fold performance improvement over state-of-the-art solutions.

Introduction

As part of the "Big Data" revolution, the way how "data" is being used in applications has seen a dramatic shift over the past years. Increasingly, relevant information is no longer carried in single pieces of data (*e.g.*, data points or records), but comes from the statistical relevance within very large data volumes.

Particle physics is a prime example of this trend that by today has reached virtually any application domain, from science to engineering to business. Analyses over massive sets of experimental data take the role that the close inspection of a single experiment had just a few years ago. To illustrate, the LHCb experiment at CERN's Large Hadron Collider (LHC) produces about 4 TB of raw data every second – year-round. Existing systems, even the largest ones, are still overwhelmed by data volumes of this scale.

In this report, we advocate the use of database technology to accelerate DaVinci (the current processing software). And we report on *DeLorean*, our new, intelligent storage back-end to accelerate data analyses at CERN.

- *DeLorean* builds on *Apache Drill* [1], the open source counterpart to Google's Dremel system [3]. Drill enables relational-style data processing at massive scale, leveraging technologies such as the *Hadoop Distributed File System (HDFS)* or *Apache Zookeeper* for coordination.
- We show how real-world analysis tasks can be broken up into a *data-intensive part* – leveraging Drill's potential to *scan* massive data volumes in parallel – and into a *compute-intensive part* which covers most of the analysis's complexity but needs to run only on a fraction of the original data set.
- An important ingredient to *DeLorean* is an aggressive reduction of the data volumes that must be *scanned* during the analysis. We achieve this by applying *column-store technology* to a *synopsis* of the original data set, heavily optimized for scanning. In addition, we leverage *lightweight compression* to save bandwidth at the storage layer.
- We illustrate the potential of *DeLorean* using a *reference analysis*, on which we achieve performance improvements of up to a factor 4.6.

Characteristics of Physical Analyses

Analysis programs typically filter the full, petabyte-scale data set according to complex criteria, strongly dependent on the particular physical question being asked. Thereby, (partial) criteria can be as simple as "return all events that produced a muon particle with an energy of at least ...," but also as complex as graph conditions on the 3d tracks that can be inferred – through compute-intensive algorithms – from the data points recorded. Figure 1 here on the right illustrates a strongly simplified example of an analysis that searches for $A^0 \to b^+b^-$ events. As can be seen in the pseudo code, simple predicates (*e.g.* on charge and mass) are interspersed with compute-intensive calculations. This type of complexity and diversity essentially rules out access

```
for all evt in events do
    for all particle in evt.particles do
        if conditions on tracks and states then          ▷ first cut
            calculate information e.g. charge
            if charge < 0 then
                neg ← neg ∪ {particle}
            else if charge > 0 then
                pos ← pos ∪ {particle}
            end if
        end if
    end for
    for all pp in pos, for all np in neg do
        calculate combined mass np.mom. + pp.mom.
        if mass in A^0 mass window then                  ▷ second cut
            emit (np, pp)
        end if
    end for
end for
```

Figure 1: Simplified $A^0 \to b^+b^-$ analysis task (stripping line cut).

structures like (multi-dimensional) indexes, leaving *scans* as the only viable search mechanism. Analysis performance is, therefore, heavily influenced by the *volume* of the data that is being scanned. To reduce this volume, the existing platform at CERN uses a mechanism that physicists refer to as *stripping*: a *preprocessing stage* segregates all events into *stripping lines* on the storage cluster; each stripping line corresponds to pre-defined search criteria.[1]

The stripping concept is both, a blessing and a curse. While it reduces the scan cost for common analysis (types), stripping *(a)* occupies scarce disk resources and *(b)* is limited to classes of analyses that have been pre-declared to the stripping process. In fact, most physicists would like to get rid of stripping rather sooner than later.

Experiments

Using an out-of-the-box embedded Drill instance we were able to achieve significant event throughput increase for *DeLorean*: The single-threaded throughput gain is of factor 4.6. Unfortunately, a parallel configuration of DaVinci is currently not available in our laboratory setting. For the sake of fairness we assume a linear scalability for DaVinci (best case). Figure 2 shows the scalability for different compression algorithms compared to the DaVinci projection. One noteable result of the experiment is that *DeLorean* even outperforms the linear projection of DaVinci.

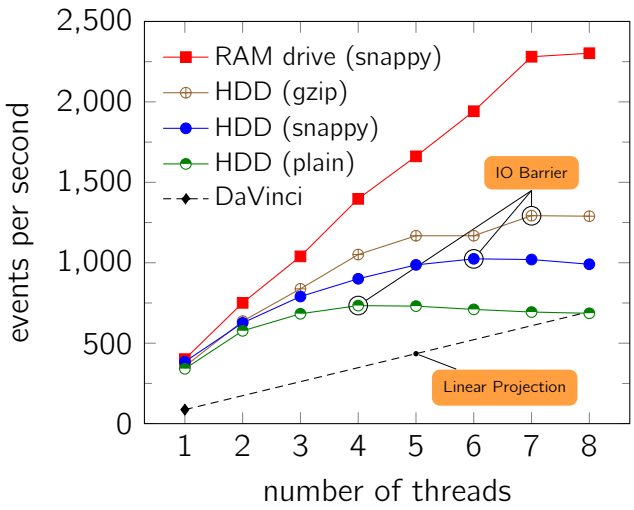


Figure 2: Single node scalability of *DeLorean* compared to DaVinci (estimation).

Figure 2 clearly shows that the HDD bound experiments hit the IO barrier.[2] Here we can see that compression can leverage the IO bottleneck significantly. Gzip, the "heaviest" of the compression algorithms, performs best in this scenario, so it might be worthwhile to invest in higher compression ratios using domain specific compression algorithms (like run-length encoding or delta encoding).

Talking about IO bottlenecks one should also look at the IO behaviour. Figure 3 shows the total amount of data read by the different approaches and the corresponding throughput at the IO barrier. Although DaVinci uses a very heavy weight compression algorithm, total reading IO is higher than for *DeLorean* using lighter weight algorithms.

[1] A stripping line compares best to a *materialized view* in a relational database engine.

[2] The RAM drive experiment is there to verify the IO bottleneck: Regarding the massive amount of data, it is unrealistic to process the whole data set in memory.

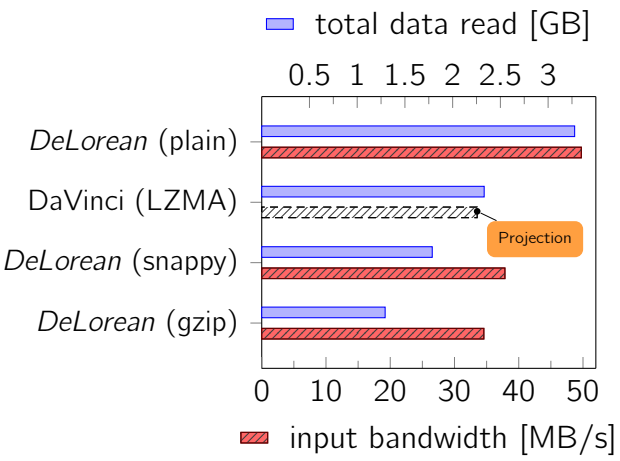


Figure 3: Comparison of HDD traffic.

Due to the columnar storage design, *DeLorean* avoids reading irrelevant data from disk. Additionally, DaVinci spends about 50 % of its total runtime decompressing data that will never be touched later. Looking at the IO throughput in Figure 3, it becomes clear that the limiting factor cannot be the bandwidth but must be a bad access pattern.[3] We are currently working on strategies to enhance the access pattern by using intelligent caches and exploiting data inherent sortedness (our extracted tables are inherently sorted by ID). Additionally, we expect a lesser impact of access patterns when further scaling *DeLorean*, allowing for much bigger block sizes per thread.

Conclusions and Future Work

There is still potential in improving the IO behaviour of *DeLorean*. Therefore we will examine other Big Data platforms such as Apache Pig [2]. Pig's LZO support will allow us to evaluate LZO compressed Parquet storage in *DeLorean*. Additionally, exploiting the natural ordering of the data (by *event_id*, *protoparticle_id*, *track_id* etc.) can aid in generating more favourable access patterns.

A more detailed presentation of *DeLorean* is soon to be published as a full-size paper.

References

[1] Apache drill - schema-free sql for hadoop, nosql and cloud storage, 2016. `https://drill.apache.org`, 15.09.16.

[2] Apache pig, 2016. `https://pig.apache.org`.

[3] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: Interactive analysis of web-scale datasets. *Proc. VLDB Endow.*, 3(1-2):330–339, September 2010.

[3] Common HDDs exceed 150 MB/s reading throughput for sequential reads.

Improving Efficiency of the LHCb High Level Trigger using Distributed Heterogeneous Low Power Hardware Compute Units

Thomas Lindemann
Lehrstuhl für Datenbanken und Informationssysteme (DBIS)
Technische Universität Dortmund
thomas.lindemann@cs.tu-dortmund.de

In the LHCb Project, a continuous stream of events is produced by the aggregates of the detector, which have to be processed in real time, since there are no capabilities to store all collision events permanently with the current storage technology. The High Level Trigger (HLT) needs to pick out the events, which have to be stored for further analysis. In our research, we are evaluating different techniques to handle with these restrictions. Since the last report, we went on testing state of the art database techniques using heterogeneous hardware of modern server CPUs and accelerating Co-Processors. A drawback of these system configurations is the big amount of energy consumption of modern server hardware, especially if it is extended by Co-Processors like Intel's Xeon Phi or General Purpose Graphics Processing Units. For that reason, we further introduce the idea for making use of new hardware from the low-power segment. This would allow a much bigger count of low-power processing units under the same constraints of energy consumption and related thermal discharge.

1 Introduction

The LHCb project is a large and complex research project grown over the last decades. Named after the b-quark, LHCb is one of the four big experiments at CERN. The general scope is to explain the matter/anti-matter asymmetry. Its main focus is the study of particle decays involving beauty and charm quarks. [1]

At the time of writing this report, we working on this project for over a year. Our specific research topic is to improve the High Level Trigger (HLT). The HLT has to process all the experiment data in hard time constraints and sample it down to the maximum load which the storage can handle. (Figure 1) To achieve this goal, the current configuration of the trigger is divided in HLT1 and HLT2. The HLT1 is the first software trigger in this process after the hardware filters and has to reduce the data frequency to a level which can be locally stored on the trigger machines. After this, the HLT2 software trigger can make further calculations on a smaller residual of events. The local storage and later processing is only possible by the fact that the detector has some phases of an unstable beam where no data is recorded, thus there is a time slot for the HTL2 to make up the locally stored events.

Figure 1: Trigger system setup

The challenge is to find new solutions for processing this big amounts of data with limited resources much faster than it has been performed in the first run of the LHCb project and allow the physicists to make experiments with more precise decisions. An approach to achieve this objective is using modern hardware components such as GPUs (Graphics Processing Units) or APUs (Accelerated Processing Units) and place the operations to the best device as described in the last report. A drawback is a much bigger energy consumption and high programming effort for optimizing the algorithms for the event processing to these new kinds of hardware architectures. Another issue of using high performance acceleration Co-Processors in SIMD architecture is the heterogeneous variety of the events data. Because of the drawbacks of high performance accelerators, we decided to investigate another field of the modern hardware section, which is low power hardware.

2 Decay Event Processing on Low-Power Compute Units

At the time of writing this report, the High Level Triggers HLT1 and HLT2 concept used by the LHCb Project is a large computing grid of standard server machines. (Figure 2) Our leading idea is the same is the same as described in the last report, the workload is not just computed by high performance server machines, instead it is separated to a much larger number of small compute units.

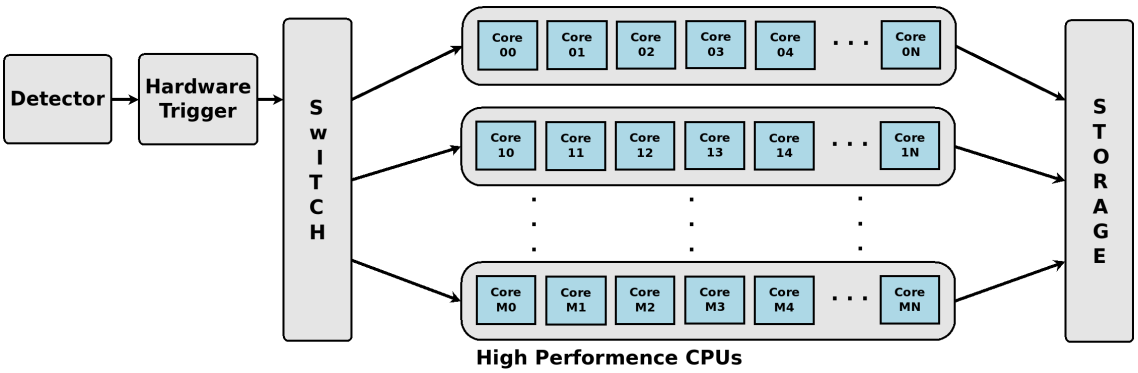


Figure 2: LHCb Data Processing Model

In contrast to the previous report one year ago, the concept was refined by using new types of heterogeneous energy efficient hardware solutions. New evolutions in low power System on Chip (SOC) technologies may reveal a better trigger solution, which is able to achieve the same overall throughput rate while less power consuming. This saved resources of energy and thermal constraints may be used to expand the computing grid and use the new compute capabilities for more advanced trigger decisions. Our approach for including heterogeneous low power hardware is based on our experiments with common used high performance CPUs and GPUs compared to modern low power hardware. We tested different key values as the compute capability, the memory bandwidth in linear scans and random address accesses and the network transfer rate.

The tested SOCs are available as homogeneous hardware, which have only one type of low power processing unit, but also as heterogeneous variants consisting of big low power cores, small ultra low power cores and integrated low power GPUs with general purpose capabilities with OpenCL. (Figure 3) This heterogeneous SOCs provide the best variety to choose the best hardware for a specific task and place the operation there.

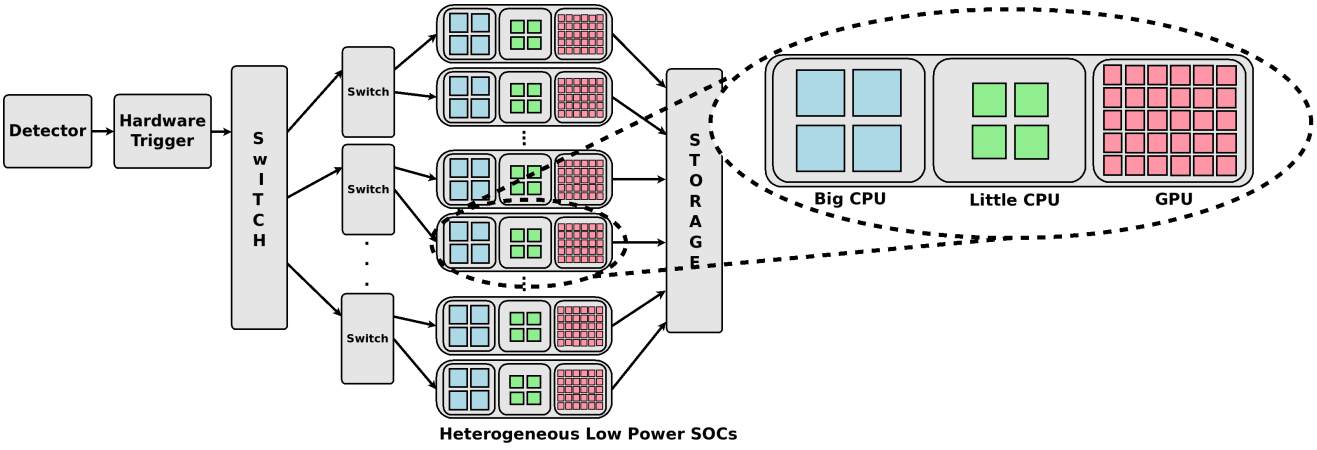


Figure 3: Newly Proposed Data Processing Model

The ideas presented in the last report were based on high performance GPUs and APUs. But the usage of a GPU and similar devices assumes that the problem is highly parallelizable, which is essential for the utilization with an SIMD-architecture. A disadvantage is that there is a high number of events with a relatively small amount of data. A solution might be to process many events simultaneously with the same algorithm, but there is the adverse fact that every event is different in the recorded data, because of varieties in the number of tracks to be reconstructed and the number and kind of vertices, due to different decay channels. Using smaller computing units as low power CPUs and especially smaller low power GPUs obtain the advantage that there lasts a much smaller workload per device.

Another advantage of the SOC integrated low power GPUs is that data transfers between devices over an PCI-e bus are avoided when all processor types on the SOC are using the same main memory.

One consequence of a shared main memory is that the bottleneck may be caused by the main memory bandwidth, especially on scan/stream computations. Additionally the memory bandwidth of only one low power device much lower than in a high performance Server. This drawback might be handled by the fact, that there is not only a few high performance processor or GPU co-processors, instead there are many highly distributes low power devices, all of them with their own main memory bandwidth. This distributed memory provides a high overall memory bandwidth and theoretically supports compute-bound operations on the processors.

3 Conclusion and Future Work

Since the last report we are working on a refinement of the new concepts already presented in the last report including the integration of further ideas as described in this report. Besides another essential step to the intended improvement of the trigger is to proof the concepts on real algorithms used by the LHCb Trigger system. Thus, a main objective for the future is to reimplement selected trigger algorithms as used in the real LHCb Trigger software framework and extract real detector data on which the algorithms should run on.

References

[1] C. Langenbruch et. al., The LHCb collaboration. Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay. *LHCb-CONF-2015-002*, 2015.

Measurement of *CP* violation using $B^0 \to D^+D^-$ decays with the LHCb experiment

Frank Meier
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
frank.meier@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer. In the collisions of protons hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction. The interaction rate of up to 25 ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

Physics analyses have been performed to demonstrate the need for an improved processing of the data collected by the LHCb detector [5]. These analyses study decay-time-dependent charge-parity (*CP*) violation which is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. The data exploited in these analyses is collected at the Large Hadron Collider (LHC), where proton bunches are accelerated and collided at centre-of-mass energies of up to 13 TeV. In 2011 and 2012, referred to as Run I, a data sample corresponding to an integrated luminosity of 3 fb^{-1} has been recorded with the LHCb detector. This data sample, collected at centre-of-mass energies of 7 TeV and 8 TeV, is the world's largest sample of B^0 mesons.

From a theoretical point of view the decay mode $B^0 \to J/\psi\, K^0_S$ offers a very clean determination of $\sin 2\beta$, as the dominating contributions can be calculated perturbatively. Experimentally, the quite large branching fractions and the clear signature in the detector allow a measurement with very high precision. In contrast, the main purpose of the study of $B^0 \to D^+D^-$ decays is to constrain higher-order Standard Model corrections occurring in measurements of the *CP*-violating phase. These contributions need to be controlled to distinguish them from effects caused by physics beyond the Standard Model of particle physics, often referred to as "New Physics".

In the B^0 meson system, *CP* violation in the mixing is negligible, as is the decay width difference $\Delta\Gamma$ of the mass eigenstates [3]. In contrast, sizeable *CP* violation from the interference between the direct (unmixed) decay into the *CP*-even final state D^+D^- and the decay to the same final state after B^0–$\overline{B}^0$ mixing, or from the interference of different decay processes, leads to a decay-time-dependent decay rate of

$$\frac{\mathrm{d}\Gamma(t,d)}{\mathrm{d}t} \propto \mathrm{e}^{-t/\tau}\Big(1 - d\,S\sin(\Delta m t) + d\,C\cos(\Delta m t)\Big)\,, \tag{1}$$

where t is the proper decay time, d represents the B^0 flavour at production and takes a value of $+1$ for mesons whose initial flavour is B^0 and -1 for $\overline{B}^0$, τ is the mean lifetime and Δm is the mass difference between the physical B^0 meson eigenstates. The *CP* observables S and C are related to the B^0 mixing phase ϕ_d and a phase shift $\Delta\phi$ from the decay amplitudes via $S/\sqrt{1-C^2} = -\sin(\phi_d + \Delta\phi)$ [4].

The subsequent decays $D^+ \to K^-\pi^+\pi^+$ and $D^+ \to K^-K^+\pi^+$ are used to reconstruct $B^0 \to D^+D^-$ decays, with combinations of two $D \to KK\pi$ candidates omitted due to the low branching fraction. A selection is required to suppress the combinatorial background. The first step is a centrally organized preselection performed by the LHCb collaboration which takes several weeks. Afterwards, the individual tuple of the size of several hundred GB is produced. The amount of background in $B^0 \to D^+D^-$ is still too high to perform a significant measurement of *CP* violation without any further selection. Therefore, a set of requirements with high signal efficiency is applied, followed by a special treatment of misidentified backgrounds and a multivariate analysis which further reduces combinatorial background. After all selection steps about 6000 candidates remain in the data sample, of which about 25 % are signal candidates. The exact number of signal candidates is determined in a mass fit which also discriminates between signal and background candidates by calculating signal weights via the ${}_s\mathcal{P}lot$ method [7].

The *CP* violation observables S and C are determined from a multidimensional fit to the background-subtracted tag and decay time distributions of the tagged $B^0 \to D^+D^-$ candidates. The conditional PDF describing the reconstructed decay time t' and tag decisions $\vec{d}' = (d'_{\mathrm{OS}}, d'_{\mathrm{SS}})$, given a per-event decay time resolution $\sigma_{t'}$ and per-event mistag probability estimates $\vec{\eta} = (\eta_{\mathrm{OS}}, \eta_{\mathrm{SS}})$, is

$$P\left(t', \vec{d}' \mid \sigma_{t'}, \vec{\eta}\right) \propto \epsilon(t')\left(\mathcal{P}(t, \vec{d}' \mid \vec{\eta}) \otimes \mathcal{R}(t'-t \mid \sigma_{t'})\right)\,, \tag{2}$$

where

$$\mathcal{P}(t, \vec{d}' \,|\, \vec{\eta}) \propto \sum_d \mathcal{P}(\vec{d}' \,|\, d, \vec{\eta})[1 - d\, A_{\mathrm{P}}]\, e^{-t/\tau} \left\{1 - d\, S \sin(\Delta m t) + d\, C \cos(\Delta m t)\right\} , \quad (3)$$

and where t is the true decay time, d is the true production flavor, A_{P} is the production asymmetry, and $\mathcal{P}(\vec{d}' \,|\, d, \vec{\eta})$ is a two-dimensional binomial PDF describing the distribution of tagging decisions given $\vec{\eta}$ and d. Normalization factors are omitted for brevity. In the fit, the mass difference Δm and the lifetime τ are constrained to their known values within uncertainties [6]. The production asymmetry A_{P} is constrained separately for the 7 TeV and 8 TeV samples to the values obtained from weighting the results from the measurements in Ref. [1] according to the kinematic distribution of the B^0 signal candidates. The parameters of the decay time resolution model $\mathcal{R}$ are determined from simulation. The average decay time resolution in data is 49 fs. The function $\epsilon(t')$ describes the efficiency for all reconstruction and selection steps as a function of the reconstructed decay time. It is represented by cubic splines, with the spline coefficients left unconstrained in the fit.

The *CP* observables are measured to be $S = -0.54\,^{+0.17}_{-0.16}$ and $C = 0.26\,^{+0.18}_{-0.17}$ with a correlation coefficient of $\rho = 0.48$. The decay-time-dependent signal yield asymmetry $(N_{\overline{B}^0} - N_{B^0})/(N_{\overline{B}^0} + N_{B^0})$, where N_{B^0} is the number of $B^0 \to D^+D^-$ decays with a B^0 flavor tag, and $N_{\overline{B}^0}$ the number with a $\overline{B}^0$ tag, is shown in Fig. 1.

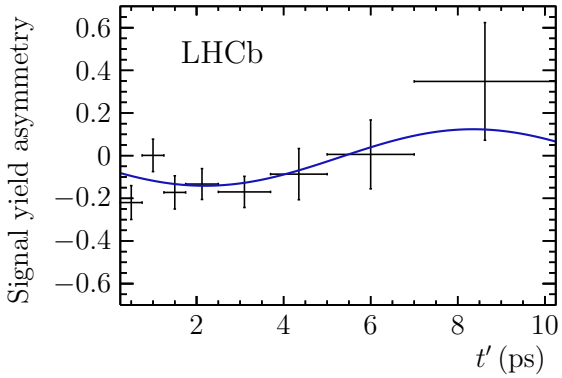


Figure 1: Decay-time-dependent signal yield asymmetry [2].

Several sources of systematic uncertainties on the *CP* observables are studied with pseudoexperiments. The largest systematic uncertainty arises from neglecting backgrounds in which the final state contains only one charm meson, such as $B^0 \to D^- K^- K^+ \pi^+$. The yield of these backgrounds is estimated to be about 2 % of the signal yield and their impact is assessed by assuming that they maximally violate *CP* symmetry and have the eigenvalue opposite to the signal mode. This leads to a systematic uncertainty of ± 0.05 on S and ± 0.013 on C. Further systematic uncertainties on S are related to the assumption $\Delta\Gamma = 0$ (± 0.014), and to the modeling of the dependence of the efficiency on

decay time (± 0.007). For C the second largest systematic uncertainty of ± 0.007 is due to neglecting the correlation between the invariant mass and the decay time. Additional systematic uncertainties arise from the decay time resolution, the uncertainty on the knowledge of the length scale, the parametrization of the mass model, and from uncertainties on the B^0 production asymmetry and mass difference Δm. The total systematic uncertainty, calculated as the sum in quadrature of all contributions, is ± 0.05 for S and ± 0.02 for C, with a correlation coefficient of $\rho = -0.69$.

In conclusion, the *CP* observables S and C are determined to be

$$\begin{aligned} S &= -0.54\,^{+0.17}_{-0.16}\,(\text{stat}) \pm 0.05\,(\text{syst})\,, \\ C &= \quad 0.26\,^{+0.18}_{-0.17}\,(\text{stat}) \pm 0.02\,(\text{syst})\,. \end{aligned}$$

This result excludes the conservation of *CP* symmetry by 4.0 standard deviations. It is compatible with the previous measurement by the BaBar experiment while being significantly more precise. The result presented here corresponds to $\sin(\phi_d + \Delta\phi) = 0.56\,^{+0.16}_{-0.17}$, which constrains the phase shift to the world's most precise value of $\Delta\phi = -0.16\,^{+0.19}_{-0.21}$ rad, and thus implies only a small contribution from higher-order Standard Model corrections [2].

References

[1] R. Aaij et al. Measurement of the $\overline{B}^0$–B^0 and $\overline{B}^0_s$–B^0_s production asymmetries in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Lett.*, B739:218, 2014.

[2] R. Aaij et al. Measurement of *CP* violation in $B \to D^+D^-$ decays. *Phys. Rev. Lett.*, 117:261801, 2016.

[3] Y. Amhis et al. Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014. 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[4] Kristof De Bruyn and Robert Fleischer. A roadmap to control penguin effects in $B^0_d \to J/\psi\, K^0_S$ and $B^0_s \to J/\psi\, \phi$. 2014.

[5] Frank Meier and Bernhard Spaan. *Measurements of* $\sin 2\beta$ *using charmonium and open charm decays at LHCb*. PhD thesis, Tech. U., Dortmund (main), Nov 2016.

[6] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[7] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Measurement of *CP* violation in $B^0 \to J/\psi(e^+e^-)K^0_S$ and $B^0 \to \psi(2S)K^0_S$ decays with the LHCb experiment

Ramon Niet
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ramon.niet@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP* violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer (see Figure 1), whereas the other three experiments ATLAS, ALICE and CMS are so called General Purpose Detectors, covering a symmetrical region around the proton-proton interaction point.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally, particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the

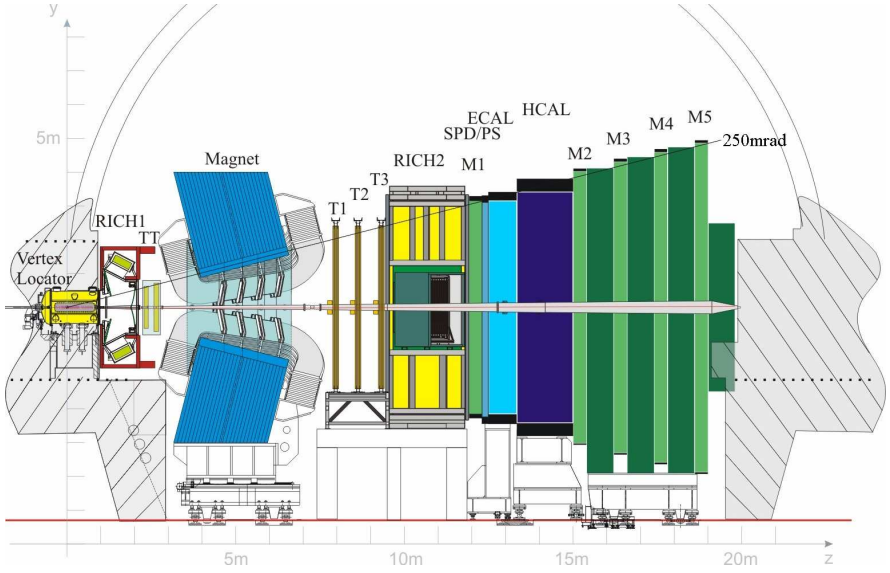


Figure 1: The LHCb detector with the various subdetectors for the identification of particles and reconstruction of their tracks.

rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

Decays of neutral B mesons involving $b \to c\overline{c}s$ transitions are referred to as "golden modes" for measuring CP violation in the interference of direct decay and decay after B–$\overline{B}$ mixing. These decay channels are theoretically clean, as higher-order contributions that could introduce additional strong and weak phases in the decay amplitudes are expected to be small [3–5].

In the B^0 system, the decay modes $B^0 \to c\overline{c}K^0_S$ and $B^0 \to c\overline{c}K^0_L$ belong to this class of decays, where $c\overline{c}$ denotes a charmonium resonance like $J/\psi, \psi(2S), \eta_c, etc.$ Here, 'K^0_S' and 'K^0_L' do not denote the undecayed K^0 mass eigenstates, but rather their $\pi\pi$ (CP even) and $\pi\pi\pi$ (CP odd) final states, respectively. At LHCb, only the K^0_S with the $\pi^+\pi^-$ final state is considered, as it does not contain neutrals. As the decay width difference $\Delta\Gamma$ is very small, $|\Delta\Gamma_d/\Gamma| = (0.1 \pm 1.0) \cdot 10^{-2}$ [1], and CP violation in the mixing is negligible, $|q/p| = 1$, the time-dependent decay rate asymmetry can be written as

$$\begin{aligned} \mathcal{A}_{J/\psi K^0_S}(t) &\equiv \frac{\Gamma(\overline{B}^0(t) \to c\overline{c}K^0_S) - \Gamma(B^0(t) \to c\overline{c}K^0_S)}{\Gamma(\overline{B}^0(t) \to c\overline{c}K^0_S) + \Gamma(B^0(t) \to c\overline{c}K^0_S)} \\ &= S\sin(\Delta m\, t) - C\cos(\Delta m\, t). \end{aligned} \tag{1}$$

The states $\overline{B}^0(t)$ and $B^0(t)$ represent evolving B^0 mesons decaying at decay time t after being produced as $\overline{B}^0$ and B^0 at $t = 0$, respectively. The parameter Δm represents the

mass difference between the two B^0 mass eigenstates. While the sine term accounts for *CP* violation in the interference of decay and mixing, a non-vanishing cosine term results from *CP* violation in the decay. As higher order penguins are suppressed, direct *CP* violation is expected to be negligible at the current level of experimental precision, hence $C \approx 0$ and $S \approx \sin 2\beta$, where β is one of the angles of the unitarity triangle of the CKM-matrix.

The present work is on the LHCb measurements of S and C using $B^0 \to J/\psi\, K^0_S$ candidates reconstructed in the di-electron final state of the J/ψ and the $B^0 \to \psi(2S)K^0_S$ mode with the charmonium state $\psi(2S)$ reconstructed in the di-muon final state. The former will be referred to as the J/ψ mode, the latter as the $\psi(2S)$ mode. In both decays, only the charged $\pi^+\pi^-$ final state of the K^0_S meson is considered. The analysed data sample corresponds to the same 3 fb^{-1} of pp collisions used in the analysis of $B^0 \to J/\psi(\mu^+\mu^-)K^0_S$ decays. While the analysis of the $\psi(2S)$ mode will allow to increase the precision of the $\sin 2\beta$ measurements by LHCb while being very similiar to the $B^0 \to J/\psi(\mu^+\mu^-)K^0_S$ mode, the J/ψ mode with the di-electron final state additionally represents a benchmark measurement for flavour tagged, decay time dependent *CP* analyses in di-electron modes at LHCb.

The selection strategy for both decay channels $B^0 \to J/\psi(e^+e^-)K^0_S$ and $B^0 \to \psi(2S)K^0_S$ is basically similar. However, the selection of the ψ(2S) channel can be performed more efficiently than in the J/ψ channel, where losses through Bremsstrahlung and relatively low identification rates of the electrons lead to efficiency losses at various stages of the selection chain. Data samples are produced from the LEPTONIC stream for the J/ψ channel and from the DIMUON stream for the ψ(2S) channel. Several selection steps are applied, where the first consists of the trigger system. Then a loose and very general selection, the so called stripping, is applied to the triggered data written to tape. The final step is the offline selection of the remaining candidates to further extract the specific decays and reconstructed final states. In the offline selection Boosted Decision Trees (BDT) [2] are used. In case of the J/ψ channel an additional preselection is applied to further reduce the sample size in order to be able to handle the amount of data and reject obvious background before applying the BDT. Fig. 2 shows the reconstructed mass distribution of candidates in both channels, which is used to extract ${}_s\mathcal{W}eights$ for the final simultaneous decay time fit to both modes that is currently being finalized.

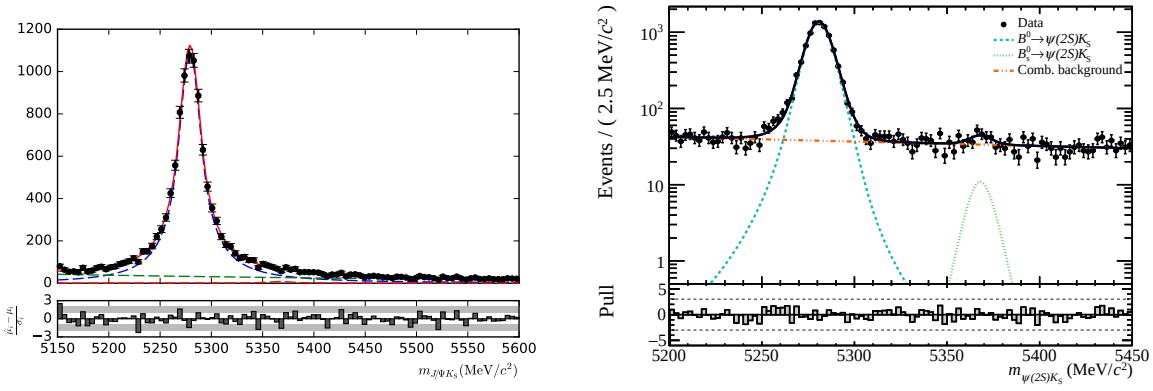


Figure 2: Nominal massfit to extract sWeights for the *CP* fit to the decay time to $B^0 \to J/\psi(e^+e^-)K^0_S$ (left) and $B^0 \to \psi(2S)K^0_S$ (right) candidates.

References

[1] Y. Amhis et al. Averages of *b*-hadron, *c*-hadron, and τ-lepton properties as of summer 2014. 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth international group, Belmont, California, USA, 1984.

[3] Sven Faller, Robert Fleischer, Martin Jung, and Thomas Mannel. The golden modes $B^0 \to J/\psi K^0_{S,L}$ in the era of precision flavour physics. *arXiv.org*, D79:014030. 4 p, September 2008.

[4] Robert Fleischer. Penguin effects in $\phi_{d,s}$ determinations. 2012.

[5] Martin Jung. Determining weak phases from $B \to J/\psi P$ decays. *Phys.Rev.*, D86:053008, 2012.

Improved precision measurement of *CP* violation

Margarete Schellenberg
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
margarete.schellenberg@tu-dortmund.de

One of the four big experiments at the Large Hadron Collider (LHC) near Geneva is the LHCb experiment [1]. The main focus lies on the research of the asymmetry of matter and anti-matter in the observable universe. During the Big Bang matter and anti matter should have been produced in equal parts. Today we observe a large asymmetry between the two, so that it is assumable, that physical laws influence matter and antimatter in different ways. Physicist at the LHCb experiment are investigating the charge-parity (CP) violation in decays of beauty and charm hadrons as one possible cause for this asymmetry. Due to the flight behaviour of b and c quarks after the collision of two protons the LHCb detector is designed as a single-arm forward spectrometer as illustrated in Figure 1.

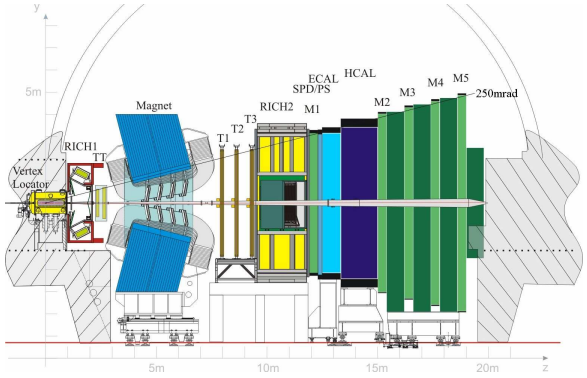


Figure 1: The LHCb detector designed as a single-arm forward spectrometer with the various subdetectors for identification and reconstruction of particles and their tracks. [2]

At the interaction point of the proton beams the Vertex Locater is situated. Due to several physical processes many particles are produced at the interaction point. They decay further into new particles that fly through the detector and interact with the detector material. For the reconstruction of these particles different subdetectors are necessary. The flight path reconstruction is done by the tracking system consisting of the VELO, TT, IT and OT. The particle identification system composed of the RICH1 and RICH2, ECAL, HCAL and the Muon chambers determines the type of detected particles. Information from both systems are used to combine the particles to heavier particles until the primary beauty and charm hadron and their complete decay chain is fully reconstructed.

At LHC, collisions are produced at a rate of 40 million collisions per second. Considering down times of the collider the experiment has to handle nearly 40×10^{14} collisions per year. If every collision has the size of 100 kB the result is a massive data volume, that is filtered on the first stage by an online and offline trigger system. Only a few per mille of the data passes the trigger and is saved on a large storage cluster.

After a centrally organized loose preselection of the complete dataset, this data is used by physicist, who store the data that correspond to their analysis conditions, in form of ROOT nTuple structures. These nTuples are much smaller than the full data before. Nevertheless they can reach a size of several hundreds of Gigabytes. Therefore it is even more important to be able to process this data efficiently.

To show the necessity of handling the data with a good performance, a physics analysis has been made. In this analysis a measurement of the time-dependent CP violation in the decay $B^0 \to D^{*\pm} D^{\mp}$ has been performed with the final states $D^{*+} \to D^0 \pi^+$, where the D^0 decays into K^- π^+ and $D^- \to K^+ \pi^- \pi^-$. The decay chain is illustrated in Figure 2.

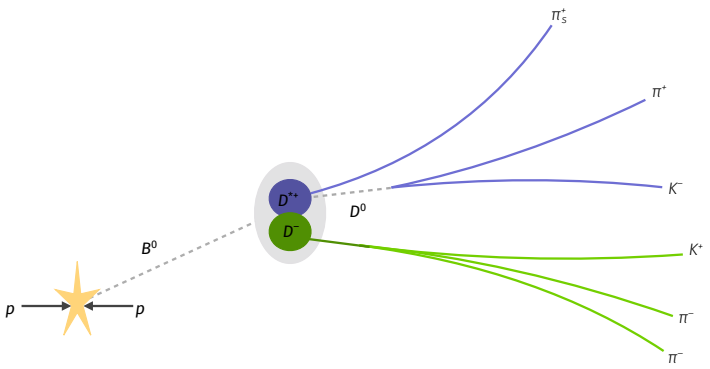


Figure 2: Schematic diagram of the decay chain of the reconstructed decay $B^0 \to D^{*+} D^-$. The D^- meson decays into $D^- \to K^+ \pi^- \pi^-$ and the D^{*+} meson in $D^{*+} \to D^0 \pi^+$, at which the D^0 meson decays further into $D^0 \to K^- \pi^+$

Due to the fact that the free propagating B^0 can mix into their anti particle $\overline{B}^0$ and vice versa and that the charge conjugated final states $D^{*+}\, D^-$ and $D^{*-}\, D^+$ are reachable from

B^0 and $\overline{B}^0$ mesons, one gets the decay-time dependent CP asymmetry, resulting from the interference between the amplitudes for the direct decay and after B^0-$\overline{B}^0$ mixing:

$$A_f(t) = \frac{\Gamma(\overline{B}^0(t) \to f) - \Gamma(B^0(t) \to f)}{\Gamma(\overline{B}^0(t) \to f) + \Gamma(B^0(t) \to f)} = \frac{S_f \sin(\Delta m t) - C_f \cos(\Delta m t)}{\cosh\left(\frac{\Delta\Gamma t}{2}\right) + D_f \sinh\left(\frac{\Delta\Gamma t}{2}\right)}. \quad (1)$$

The decay-time dependent asymmetry is made up of the quotient of the difference of the decay width of a B^0 or $\overline{B}^0$ meson decaying into the final state f and the sum of those. $B^0(t)$ and $\overline{B}^0(t)$ state the flavour of the B meson at the moment of production and the parameters Δm and $\Delta\Gamma$ are the differences of the masses and decay widths between the heavy and light mass eigenstates concerning the B^0-$\overline{B}^0$ system [4]. With some assumptions the time-dependent asymmetry for the charge conjugated final states become

$$A_f(t) = S_f \sin(\Delta m t) - C_f \cos(\Delta m t), \qquad A_{\overline{f}}(t) = S_{\overline{f}} \sin(\Delta m t) - C_{\overline{f}} \cos(\Delta m t), \quad (2)$$

where S_f, $S_{\overline{f}}$, C_f and $C_{\overline{f}}$ are the CP observables.

The first step for measuring CP observables is the selection, the separation of signal and background. Background are particles, that are mistaken with the signal due to wrong reconstruction or misidentification. If one wants to extract information from such large data samples, it is necessary to achieve a good control over the different backgrounds. The selection that is applied consists of different selection steps. Starting with a rectangular cuts on kinematical and geometrical requirements. Figure 3 shows some of the distributions of the variables that are used for the selection. The red areas are selected and the grey areas are rejected. The cut points are not yet optimized and are chosen by eye.

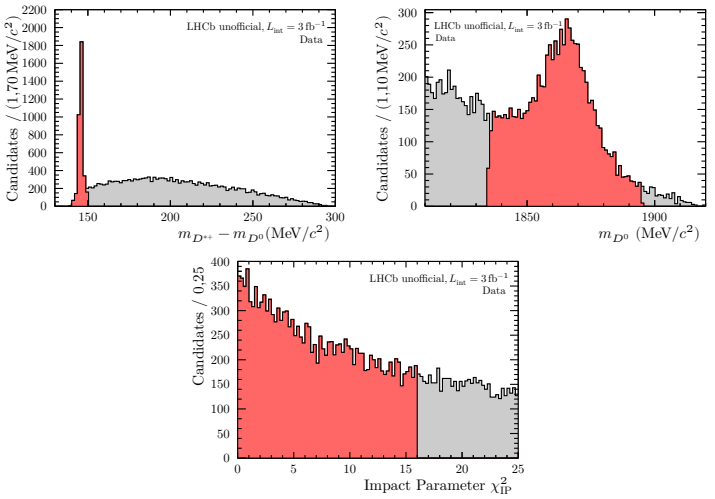


Figure 3: Distributions of variables, that were used for the cut-based selection. The red area is selected, the grey area rejected.

Afterwards a multivariate analysis is performed. As before, the cut point of the classifier is not optimized. The next step for measuring CP observables is a maximization of a likelihood

of a probability density function in different variables. Because the analysis is not finished yet, the fit is performed blinded. The preliminary result is a statistical uncertainty of 0.27 which is a factor of two larger compared to results from the experiments Belle and BaBar [3,5]. To become more compatible with Belle and BaBar, the statistics needs to be increased.

Since the amount of measured data is fixed, the only way to increase the statistics is to improve the selection. One possibility to optimize the selection in the analysis of $B^0 \to D^{*\pm}D^{\mp}$ is to use more elaborated cut points. For that a scan of the cut point on a variable has to be performed and at every step the cut efficiency has to be determined, considering the background rejection and signal efficiency. The optimal cut point is the cut value with the best preservation of the signal while rejecting as much as possible of the background. Executing a cut on a nTuple, that has a size of several hundreds GB is very time consuming. To increase the performance a programming model like MapReduce for parallelizing the cut process which results in a massively reduced processing time, especially compared to nonparallelized processes on a single multi-core machine.

The Apache Hadoop framework [7] fulfils the conditions, with the Hadoop Distributed File System (HDFS) [6] that allows for scalable distributed storing abilities, while the MapReduce programming model can be used for the processing task.

References

[1] A. A. Alves, Jr. et al. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[2] R Aaij et al. Letter of Intent for the LHCb Upgrade. Technical Report CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

[3] Bernard Aubert et al. Measurements of time-dependent CP asymmetries in $B^0 \to D^{(*)+}D^{(*)-}$ decays. *Phys. Rev.*, D79:032002, 2009.

[4] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[5] M. Röhrken et al. Measurements of Branching Fractions and Time-dependent CP Violating Asymmetries in $B^0 \to D^{(*)\pm}D^{\mp}$ Decays. *Phys. Rev.*, D85:091106, 2012.

[6] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[7] Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.

Use of GPUs in the LHCb Online Farm for the track reconstruction in the SciFi tracker of the LHCb upgrade detector

Holger Stevens
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
holger.stevens@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider (LHC) near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP*-violation in decays of beauty and charm hadrons [1]. Due to some physical constraints in the production of b and c quarks through proton proton collisions the LHCb detector is designed as a single-arm forward spectrometer. Over the past years the understanding of the detector and its systematical effects has reached an almost perfect level. At the moment, the most limiting factor for analyses is the statistical uncertainty. The only way to improve this is to massively increase the dataset. For this reason, an upgrade of the experiment is foreseen in 2019 [5]. The upgrade LHCb detector is shown in figure 1. There are smaller changes to the existing detector, but the general structure of the components will remain the same. In the Vertex Locator (Velo) the position of the primary interaction is detected. The Upstream Tracker (UT) and the SciFi Tracker also belong to the tracking system. Other components like the Ring Imaging Cherenkov Detectors (RICH), the Electronic Calorimeter (ECAL), the Hadronic Calorimeter (HCAL) and the Muon Chambers (M2-M5) are used for the particle identification.
Induced by an already implemented upgrade of the LHC the rate of proton-proton collision is doubled from 20 MHz to 40 MHz. There is no major problem for the old detector to handle this higher event rate. But due to the trigger system, which will be explained later, the output of data is not doubled. After the next upgrade of the LHC the performance of the old LHCb detector would be reduced significantly. The center-of-mass energy will be increased to 14 TeV and the luminosity to $2\times10^{33}\,\mathrm{cm}^{-2}s^{-1}$. This causes a lot more hits in the detector. Most components are designed for a single hit resolution and can not

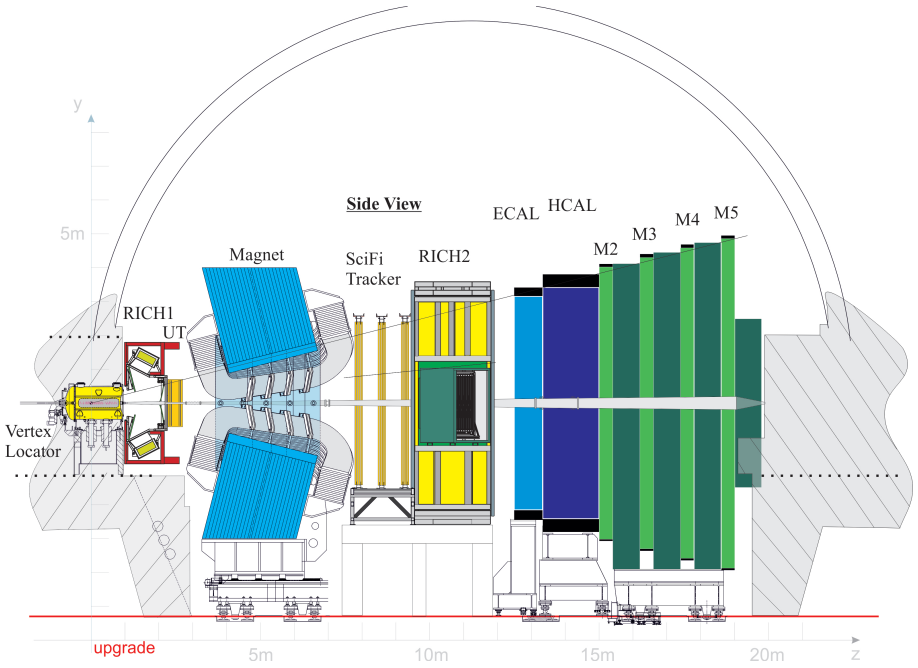


Figure 1: The LHCb upgrade detector with the various subdetectors for the identification of particles and reconstruction of their tracks [4].

distinguish multiple hits in one event. The most meaningful example are the drift tubes in the existing tracking stations. When a particle passes a tube, a signal is produced. This tube is not operative for a certain time the so called dead time, which is in this case more than 25 ns. Because there is no detector technology without dead time, a new tracking station with a higher granularity was developed, the so called SciFi Tracker. In the SciFi Tracker scintillating fibre with a diameter of 250 μm are used. This is tiny compared to the 5 cm thick drift tubes.

As mentioned before, the trigger system is a bottleneck in the LHCb data acquiring system. The trigger is needed, because the amount of data would be too high if every event would be stored. The trigger system of the present LHCb detector consists of two stages: the Level-0 (L0) hardware trigger and a software trigger, the so called High Level Trigger (HLT). The aim of the L0 is to reduce the rate to 1 MHz. The HLT reduces the rate to 12.5 kHz which can be stored. Not all of the dismissed events are uninteresting, so it would be good to be able to store more. Therefore, not only the LHCb detector will be upgraded but also the trigger system.

The new trigger is a full software trigger, as a result the whole electronics has to handle a trigger-less readout frequency of 40 MHz. The Online Farm will receive about 32 Tbit/s [3]. These are more or less just raw information like channel IDs from the tracking system where a hit was and factors for energy deposit in the calorimeter systems. Because of the asynchronous arrival of data from different subdetectors for one event the data needs

to be buffered and then are collated. This takes place in the Eventbuilder farm and the assembled event is sent to the Eventfilter farm (EF). In the EF the raw information is decoded. This requires a detailed model of the detector. Then, the reconstruction is performed. Several algorithms are trying to build tracks from the hits in the detector. There are two types of algorithms, on the one hand the independent and on the other hand the dependent. The independent algorithms only get information from one tracking component to find tracks and the dependent ones get input tracks from other algorithms to extend them.

This project tries to build an independent GPU-based algorithm to reconstruct tracks in the SciFi tracker so called T tracks. The different types of tracks are shown in figure 2. Up next, to motivate the usage of GPUs, a simplified representation of the tracking

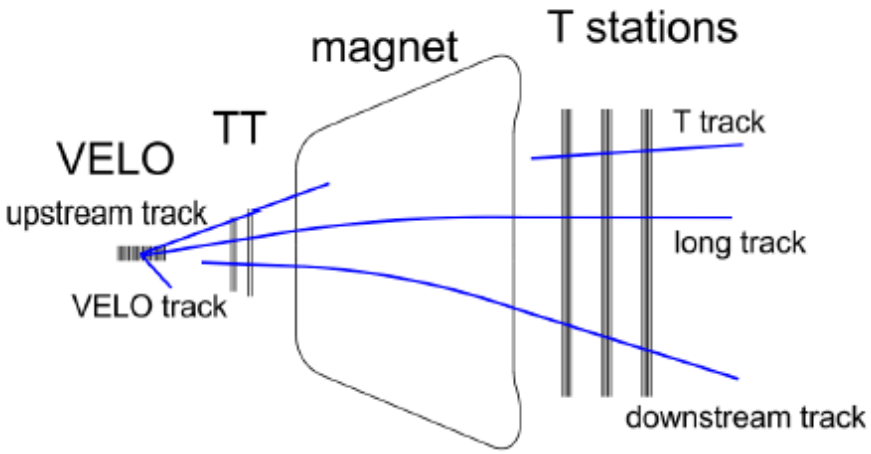


Figure 2: The different track types in the LHCb detector [2].

procedure.
The SciFi tracker consist of 12 layers and two hits from different layers are picked. The hits are connected via a straight line, which is interpolated through all layers. In a certain window around this line another hit is searched. If another hit is found the triplet is stored as a track candidate. This calculation has to be performed for every possible combination. Because the calculations are independent it is possible to parallelise them. As it is well known GPUs are ideal for massive parallel computations. A very important quality standard for every tracking algorithm is the tracking efficiency. Due to the electronic noise and some other physics effects it is possible to build so called ghost tracks. The quality or purity of tracks is directly correlated with the significance of later analyses.
It is not easy to evaluate the quality of a new algorithm, because there are no datasets with truth informations. Only Monte Carlo simulation of events for an also simulated detector are available. With infinite computing power it would be easier to calculate

only good tracks, but there is a hard time constraint for the whole event reconstruction progress. At the moment the time budget for one event is 13 ms [2]. This is estimated by the probable computing power of the Online Farm available in 2019. Of course one limit is the cost, but another factor is the free space in the server room. As a remark the LHCb experiment is an underground experiment with a limited cavern. The power consumption and the cooling capacity are also limited.
At the moment only CPUs are used in the LHCb computing farm. For this reason the algorithm has to be built from scratch. The format of the data structure which will be transferred to the GPU was chosen. The next steps are to find physical constraints for track candidates. These should reduce the workload of the GPU. After a good track quality is reached, the tuning of the runtime can be challenged.

References

[1] Letter of Intent for the LHCb Upgrade. Technical Report CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

[2] LHCb Trigger and Online Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-016. LHCB-TDR-016, May 2014.

[3] D. H. Campora Perez, A. Falabella, D. Galli, F. Giacomini, V. Gligorov, M. Manzali, U. Marconi, N. Neufeld, A. Otto, F. Pisani, and V. M. Vagnoni. The 40 MHz trigger-less DAQ for the LHCb Upgrade. *Nuclear Instruments and Methods in Physics Research A*, 824:280–283, July 2016.

[4] LHCb Collaboration. LHCb Tracker Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-001. LHCB-TDR-015, Feb 2014.

[5] Christian Joram, Ulrich Uwer, Blake Dean Leverington, Thomas Kirn, Sebastian Bachmann, Robert Jan Ekelhof, and Janine Müller. LHCb Scintillating Fibre Tracker Engineering Design Review Report: Fibres, Mats and Modules. Technical Report LHCb-PUB-2015-008. CERN-LHCb-PUB-2015-008, CERN, Geneva, Mar 2015.