technische universität dortmund

SFB 876

Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

December 2019

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

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

Contents

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Jian Jia Chen

Machine Learning on Embedded Systems

Sebastian Buschjäger
Artificial Intelligence Group, Chair 8
Technical University 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. To meet this demand, two approaches are possible: First, machine learning algorithms can be tailored specifically for the hardware at hand. Second, instead of changing the algorithm we can change the hardware to suit the machine learning algorithms better. This report briefly discusses my last years' work which focused largely on the first approach and quickly outlines some ideas for future research.

1 Introduction

To make machine learning universally applicable, we need to bring its algorithms to small and embedded devices including both - the training and the application of models. From a computer architectural point of view, we may optimize these two aspects separately. In model application we rapidly apply an already trained model for predictions and thus focus on the optimization of inference. In model training however, we would like to train models on small devices directly, so that these devices dynamically adjust their prediction rules for new data.

2 Machine learning for Embedded Devices

In the previous year I continued my study of decision tree (DT) ensembles and found, that DT ensembles are memory hungry in practical applications due to the amount and size of DTs required for satisfactory performance. Thus, I explored various ways to reduce amount of memory required by an ensemble of DTs. I noticed that Deep learning methods also suffer from this problem and thus tried to adopt various techniques from the Deep Learning (DL) community for DT ensembles.

Arguably the simplest method to reduce memory requirements is to reduce the number if bits required to store a single model's parameter, e.g. by using Fixed-Point arithmetics. Various approaches have been proposed in the DL community ranging from fixed weight quantization to weight clustering. DTs do not use weights, but axis-aligned splits on features which makes clustering not directly applicable. I explored clustering of trees by introducing different distances measure for trees based on their output distribution. This way, I was able to assign trees to clusters and use cluster centers as a representative for the ensemble. This approach considerable reduces the amount of memory required by the final ensemble with only marginally impact on the classification performance. Unfortunately, the process seems to be somewhat unstable. Most of the time, it is equally good (or sometimes even better) to randomly sample a small set of trees without prior clustering. I suspect, that clustering inherently assumes some continuity in the distances of data points, which does not hold for trees since their predictions can vary a lot. I hypothesize that sampling those trees from the original ensemble which are closest to the predictions of the original ensemble should work well and was able to verify this hypothesis in some preliminary experiments. I continued my efforts to bring pre-trained ML models to embedded devices by the means of code generation. Previously we have been focusing on code generation for DT ensembles [2] which was expanded to Deep Learning methods as well. I specifically implemented the training of Binarised Neural Networks following [3], which only use constraint weights from $\{-1, +1\}$. These types of networks are ideal candidates for the fast execution on small devices, as these networks can be implemented only using XOR, summation and popcount operations. Together with various students we expanded the previously written code generator to include Neural Networks and specifically BNNs. We intend to use this generator for the upcoming summerschool in 2020, as well as various publications. Moreover, we aim to expand the code generator by adding FPGAs a potential computation backend. Due to our ongoing discussions in the A1 project we noticed that BNNs might be ideal candidates for non-volitale memory (NVM). NVM have the interstring property, that the probability of random bitflips is proportional to the energy spend for write operations.We found, that BNNs are ideal in this situation as they can be trained to include redundancies which makes them more resilient towards random bit-fips. We are currently working on a publication in A1.

3 Machine learning on Embedded Devices

Going from model application to model training, I focused small devices and especially smart sensors. A smart sensor produces vast amounts of data but only has limited memory, processing and communication capabilities. Thus, I decided to investigate data aggregation techniques on data streams. More specifically, I looked into data summarization by the means of submodular function maximization. A submodular function is a set function $f : 2^V \to \mathbb{R}^+$ which incorporates a notion of diminishing returns: Adding a new element to a smaller set will increase the utility value of the set more than adding the same element to a larger set. This intuitively captures one aspect of data summarization in which adding new elements to a summary always make it more informative, but less so with each new item. In general, set maximization is NP-hard, which makes sub-set selection a very challenging problem. However, for submodular functions a simple greedy algorithm which always selects that element which maximizes the current function value the most achieves a $1 - 1/e$ approximation guarantee in linear run time [4]. This algorithm assumes, that the data is already stored which is infeasible for small devices. The SieveStreaming algorithm is a streaming version which "sieves-out" un-informative items on-the-fly by maintaining a set of different novelty thresholds [1]. For each threshold a summary is extracted, in which each item exceeds the chosen threshold. Since the optimal threshold is not known beforehand, the authors propose to carefully "sample" different thresholds and show that their approach maintains a $1/2 - \varepsilon$ approximation where ε controls the total amount of sieves. During experiments with this algorithm I noticed, that the amount of sieves (and therefore summaries) continuously maintained by the algorithm quickly grows even for moderate choices of ε. Moreover, many thresholds are either too small so that summaries quickly fill up or are too large so that they rarely add any items to the summaries. Thus, I designed a simpler version of SieveStreaming called ThreeSieves which employs the "Rule of Three" to estimate the probability that given a threshold there will be any item in the remaining stream exceeding it. This way, the algorithm starts with the highest possible threshold and rejects points as long as sufficient evidence is gathered that the threshold can be safely lower without hurting the maximization performance. The resulting algorithm is much faster than SieveStreaming while having similar or even better maximization performance. The corresponding paper is currently under review.

4 Future research

It is well-known that Random Forest - which combines data and feature bagging - produces excellent results in practice which cannot be explained by the classic model-based complexity measures such as VC-Dimension or Rademacher complexity. Model-based complexity measures predict that large models should overfit, whereas Random Forest

tend to become very large but rarely overfit. One theoretical explanations for this behavior can be expressed in terms of the Bias-Variance Decomposition which states that ensemble with enough variance in their individual predictions are less prone to overfitting. As stated above, we were able to derive a Generalized Bias-Co-Variance Decomposition for ensembles. This decomposition shows two things: First, we can understand boosting and bagging in the same framework in which both minimize different sides of the same equality which shows a novel connection between both algorithm. Moreover, this decomposition shows that there are many equally good models from a losses point of view. Using the Bias-Variance decomposition we derived an algorithm which allows us to precisely control the trade-off between the bias and variance term. However, this algorithm does not yield a better result than existing methods and therefore we may pose the question: If neither the Bias-Variance Decomposition, nor the complexity-based measure accurately predict the behavior of an ensemble, what does?

I want to explore this question into two directions: First, maybe the excellent performance of Random Forest is due to the combination of DTs with Bagging. Thus, I will continue to study DTs in general and specifically methods which enable gradient-based learning of DTs. This will also enable us to train DTs on small devices which do not have enough memory to store sufficient data. Second, we will evaluate the Bias-Variance-tradeoff in more settings including Deep Learning to gain a more complete picture.

References

[1] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 671–680. ACM, 2014.

[2] Sebastian Buschjäger, Kuan-hsun Chen, Jian-jia Chen, and Katharina Morik. Realization of Random Forest for Real-Time Evaluation through Tree Framing. *The IEEE International Conference on Data Mining series (ICDM)*, 2018.

[3] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pages 4107–4115, 2016.

[4] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Deep Representations for Searching Related Mathematical Expressions

Lukas Pfahler
Artificial Intelligence Group
TU Dortmund University, Germany
lukas.pfahler@tu-dortmund.de

We learn to search for mathematical expressions in a large collection of scientific documents. We view mathematical expressions as trees and use graph convolutional networks to learn a representation that allows nearest-neighbour retrieval of related formulas. Our learning tasks are inspired by embedding learning, similarity learning and more recent ideas from self-supervised learning. Empirical Evaluations show advantages over classical information retrieval approaches based on partial exact matches between expressions.

Motivation

Given the increasing number of new scientific publications, searching for relevant articles becomes tedious. In many sciences, the most efficient way to judge relevance of scientific manuscripts is by looking at the mathematical expressions. They are compact and comprehensive and a trained reader can derive lots of information from the particular composition of symbols, letters and numbers [1, 2]. We show how we can apply graph convolutional neural networks to judge the relevance of equations for the application in a search engine for mathematical expressions. To do so, we design a self-supervised learning task and train our model using millions of formulas expressed as trees [3].

Mathematical Expressions as Graph Data

We work on all publications on arXiv.org up to April, 2019 and extract all mathematical expressions from the LaTeX sources. We convert these formulas to Presentation

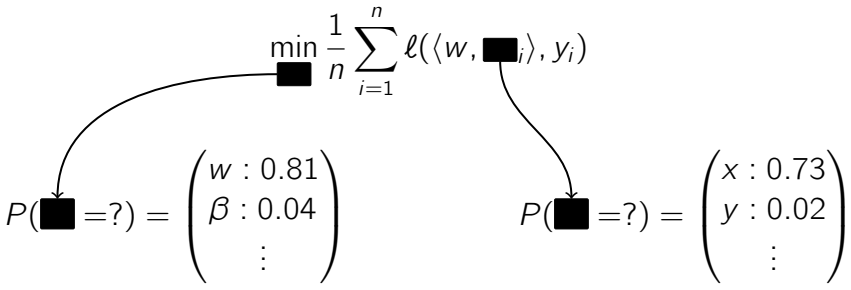


Figure 1: Example of the Masking Task with Fictional Values

MathML, a XML dialect for displaying maths on the web. This way, we obtain 28,973,591 mathematical expressions from 760,041 manuscripts. Viewing this XML representation as a tree-structure, we can represent each expression as a graph where each vertex is associated with a feature vector constructed by one-hot-encoding the XML-nodes. More specifically, we separately encode the tag-name (like `<mi>` for math identifiers or `<mn>` for numbers), any optional attributes (like font-changes) and the text in leaf-nodes (numbers, parenthesis, Latin and Greek letters, etc.) in a total of 256 binary features.

Self-Supervised Representation Learning with Graph Convolutional Neural Networks

We want to learn similarities between formulas or rather graph structures x, x' of the form

$$\text{sim}(x, x') = \langle \phi(x), \phi(x') \rangle \text{ with } \phi(x) \mapsto \mathbb{R}^d$$

where ϕ is a vectorial embedding produced by a graph CNN with average pooling [4]. Learning similarity measures traditionally requires labeled data [5, 6]. It is however infeasible to obtain large quantities of labeled pairs of formulas. To mitigate this, we consider an unsupervised approach that simultaneously optimizes two objectives. First, we use a contextual similarity approach where we label two formulas as related when they appear in the same context, in our case in the same paper [7]. We minimize the histogram loss using these context labels [8]. Second we use a self-supervised masking task originally used in language modeling [9]: We randomly set the feature maps of 15% of the vertices to zero and train our model to reconstruct the hidden inputs using the remaining vertices, as depitcted in Figure 1.

A Search Engine Study Using Semi-Manual Annotations

In order to evaluate our vectorial representations in a search engine use-case, we first curate a set of evaluation queries. To this end, we have asked fellow researchers from different disciplines to provide mathematical expressions annotated with a set of keywords. Checking for these keywords in the surrounding sections of our search results allows us to automatize the relevance-scoring of our results. We show that our machine-learning based search beats a traditional bag-of-words approach by a large margin. In particular we see improvements when query and results use different notations or conventions. In the future we hope to host a demo of our system and collect actual user queries. To further improve the generalization, we will investigate additional sources of supervision.

Outlook

In the future we want to learn embeddings using many labeling heuristics. Currently we use context, but we want to extend this to other sources of supervision. Most promising is the idea to split expressions at equality or in-equality signs and learn these identities using our graph convolutional network. New ideas in self-supervised learning can be included, like instead of predicting masked tokens we can randomly exchange some tokens and predict which tokens are real.

References

[1] Wei Zhong and Richard Zanibbi. Structural Similarity Search for Formulas using Leaf-Root Paths in Operator Subtrees. In *European Conference on Information Retrieval*, pages 116–129. Springer, 2019.

[2] Ferruccio Guidi and Claudio Sacerdoti Coen. A survey on retrieval of mathematical knowledge. *Mathematics in Computer Science*, 10(4):409–427, 2016.

[3] Lukas Pfahler and Katharina Morik. Semantic Search in Millions of Equations. In *under review at KDD2020*, 2020.

[4] Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[5] Andreas Maurer. Learning similarity with operator-valued large-margin classifiers. *Journal of Machine Learning Research*, 9:1049–1082, 2008.

[6] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–119. BMVA Press, 2016.

[7] Lukas Pfahler, Jonathan Schill, and Katharina Morik. The Search for Equations - Learning to Identify Similarities between Mathematical Expressions. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019*, 2019.

[8] Evgeniya Ustinova and Victor Lempitsky. Learning Deep Embeddings with Histogram Loss. In *Advances In Neural Information Precessing Systems 2016*, 2016.

[9] Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A2

Algorithmic aspect of learning methods in embedded systems

Christian Sohler Jens Teubner

Probabilistic smallest enclosing ball in high dimensions

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

We study a variant of the median problem for a collection of point sets in high dimensions. This generalizes the geometric median as well as the (probabilistic) smallest enclosing ball (pSEB) problems. Our main objective is to improve the previously best algorithm for the pSEB problem by reducing its exponential dependence on the dimension to linear. This is achieved via a novel combination of sampling techniques for clustering problems in metric spaces with the framework of stochastic subgradient descent. As a result, the algorithm becomes applicable to shape fitting problems in Hilbert spaces of unbounded dimension via kernel functions. We present an exemplary application by extending the support vector data description (SVDD) shape fitting method to the probabilistic case. These results were published in [2]

Introduction

The (probabilistic) smallest enclosing ball (pSEB) problem in $\mathbb{R}^d$ is to find a center that minimizes the (expected) maximum distance to the input points. It occurs often as a building block for complex data analysis and machine learning tasks like estimating the support of high dimensional distributions, outlier detection, novelty detection, classification and robot gathering. Our goals are reducing the number of points but also keeping the dependence on the dimension as low as possible. We focus on a small dependence on the dimension. Kernel methods are a common technique in machine learning. These methods implicitly project the d-dimensional input data into much larger dimension D where simple linear classifiers or spherical data fitting methods can be applied to obtain

a non-linear separation or non-convex shapes in the original d-dimensional space. The efficiency of kernel methods is usually not harmed since inner products and thus distances in the D-dimensional space can be evaluated in $O(d)$ time.

In some cases a proper approximation relying on sampling and discretizing the ambient solution space may require a polynomial or exponential dependence on D. The algorithm of Munteanu *et al.* [4] is the only FPTAS and the fastest algorithm to date for the pSEB problem in fixed dimension, but it suffers from the stated problems. To make the pSEB algorithm viable in the context of kernel methods and generally in high dimensions, it is desirable to reduce the dependence on the dimension to a small polynomial occurring only in evaluations of inner products and distances of low dimensional vectors.

General notation: Let $[n] = \{1, \ldots, n\}$. For any convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ let $\partial f(x) = \left\{g \in \mathbb{R}^d \mid \forall y \in \mathbb{R}^d\colon f(x) - f(y) \leq \langle g, x - y\rangle\right\}$ be the set of subgradients of f at x. We assume $0 < \varepsilon < 1/9$.

A generalized median problem

The pSEB problem can be reduced to two different types of 1-median problems [4]. One of them is defined on the set of all non-empty locations in $\mathbb{R}^d$ where probabilistic points may appear, equipped with the Euclidean distance. The other is defined on the collection of all possible realizations of probabilistic point sets, and the distance measure between a center $c \in \mathbb{R}^d$ and a realization $P_i \subset \mathbb{R}^d$ is $m(c, P_i) = \max_{p \in P_i} \|c - p\|$. We state a generalized problem that covers both of them, and call it the *set median problem*. In case of singleton sets, the set median problem is equivalent to the 1-median problem. For $N = 1$ it coincides with the smallest enclosing ball problem. For both of these problems there are known algorithms based on the subgradient method from convex optimization. Here we adapt the deterministic subgradient method from [5].

Definition 1 (set median problem)**.** *Let $\mathcal{P} = \{P_1, \ldots, P_N\}$ be a family of finite non-empty sets where $\forall i \in [N]\colon P_i \subset \mathbb{R}^d$ and $n = \max\{|P_i| \mid i \in [N]\}$. The set median problem on $\mathcal{P}$ consists in finding a center $c \in \mathbb{R}^d$ that minimizes the cost function $f(c) = \sum_{i=1}^{N} m(c, P_i)$.*

To minimize f we need to compute a subgradient $g(c_i) \in \partial f(c_i)$ at the current center c_i. The subgradient computation takes $O(dnN)$ time, thus we replace the exact subgradient $g(c_i)$ by a uniform sample of only one nonzero term which points into the right direction in expectation. The result is in expectation a $(1+\varepsilon)$-approximation to the optimal solution. We choose an initial center c_0 to be an arbitrary point in a randomly chosen input set from $\mathcal{P}$, it suffices to have $\ell \in O(1/\varepsilon^2)$ iterations, and a fixed step size s is bounded by average cost on a sample of size $1/\varepsilon$ [3]. Our algorithm iteratively picks a set $P_j \in \mathcal{P}$ uniformly at random and chooses a point $p_j \in P_j$ that attains the maximum distance

to the current center. This point is used to compute an approximate subgradient. The algorithm outputs the best center found in all iterations. To find the best center out of all iterations of our algorithm efficiently we cannot do it exactly since evaluating the cost even for one single center takes time $O(dnN)$. We adapt a sampling technique [7] to find a point that is a $(1+\varepsilon)$-approximation of the best center in a finite set of candidate centers. The main difference to [7] is that the collection of input sets and the set of candidate solutions here may be completely distinct. Putting all together we have:

Theorem 2. *Consider an input $\mathcal{P} = \{P_1, \ldots, P_N\}$, where for every $i \in [N]$ we have $P_i \subset \mathbb{R}^d$ and $n = \max\{|P_i| \mid i \in [N]\}$. There exists an algorithm that computes a center $\tilde{c}$ that is with constant probability a $(1+\varepsilon)$-approximation to the optimal solution c^* of the set median problem (see Definition 1). Its running time is $O\big(dn/\varepsilon^4 \cdot \log^2 1/\varepsilon\big)$.*

The removal of the linear dependence on n for the maximum distance computations was done in [4] via a grid based strong coreset of size $1/\varepsilon^{\Theta(d)}$. This is not an option if we want to work in high dimensions. We show that without an exponential dependence on d, we have to lose an approximation factor of roughly $\sqrt{2}$, adapting the techniques of [1].

Theorem 3. *Any data structure that, with probability at least $2/3$, α-approximates maximum distance queries on a set $S \subset \mathbb{R}^d$ of size $|S| = n$, for $\alpha < \sqrt{2}\left(1 - 2/d^{1/3}\right)$, requires $\Omega\left(\min\{n, \exp\left(d^{1/3}\right)\}\right)$ bits of storage.*

Applications

We apply our result to the pSEB problem, as given in [4]. In such a setting, the input is a set $\mathcal{D} = \{D_1, \ldots, D_n\}$ of n discrete and independent probability distributions. The i-th distribution D_i is defined over a set of z possible locations $q_{i,j} \in \mathbb{R}^d \cup \{\perp\}$, for $j \in [z]$, where $\perp$ indicates that the i-th point is not present in a sampled set, i.e., $q_{i,j} = \perp \Leftrightarrow \{q_{i,j}\} = \emptyset$. We call these points *probabilistic points*. Each location $q_{i,j}$ is associated with the probability $p_{i,j}$, such that $\sum_{j=1}^{z} p_{i,j} = 1$, for every $i \in [n]$. Thus the probabilistic points can be considered as independent random variables X_i. A probabilistic set X of probabilistic points is also a random variable.

Definition 4. *([4]) Let $\mathcal{D}$ be a set of n discrete distributions, where each distribution is defined over z locations in $\mathbb{R}^d \cup \{\perp\}$. The pSEB problem is to find a center $c^* \in \mathbb{R}^d$ that minimizes the expected smallest enclosing ball cost: $c^* \in \operatorname{argmin}_{c \in \mathbb{R}^d} \mathbb{E}\left[m(c, X)\right]$, where the expectation is taken over the randomness of $X \sim \mathcal{D}$.*

Our pSEB algorithm adapts the framework of [4], but plugging in Theorem 2 it differs mainly in three points. First, the number of samples had a dependence on d hidden in the O-notation. This is not the case any more. Second, the sampled realizations are not sketched via coresets of size $1/\varepsilon^{\Theta(d)}$ any more. Third, the running time of the actual optimization task is reduced instead of an exhaustive grid search.

Theorem 5. *Let $\mathcal{D}$ be a set of n discrete distributions, each defined over z locations in $\mathbb{R}^d \cup \{\perp\}$. Let $\tilde{c} \in \mathbb{R}^d$ be the output of our pSEB algorithm on input $\mathcal{D}$. Let $\varepsilon < 1/9$. With constant probability $\tilde{c}$ is a $(1+\varepsilon)$-approximation for the pSEB problem: $\mathbb{E}_X[m(\tilde{c}, X)] \leq (1+\varepsilon)\min_{c\in\mathbb{R}^d} \mathbb{E}_X[m(c, X)]$. The running time is $O(dnz/\varepsilon^4 \cdot \log 1/\varepsilon)$.*

Comparing to the result of [4], the running time is reduced from $O(dnz/\varepsilon^{O(1)} + 1/\varepsilon^{O(d)})$ to $O(dnz/\varepsilon^{O(1)})$. The factor of d plays a role only in computations of distances between two points in $\mathbb{R}^d$. The sample size and the number of centers that need to be evaluated do not depend on the dimension d any more. This is crucial in the application to the SVDD problem (the SEB problem in kernel spaces). We extend the SVDD to its probabilistic version. Let $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive semidefinite kernel function with feature map $\varphi\colon \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is a high dimensional Hilbert space, say $\mathbb{R}^D$, where $D \gg d$ [6].

Theorem 6. *Let $\mathcal{D}$ be a set of n discrete distributions, each defined over z locations in $\mathbb{R}^d \cup \{\perp\}$. There is an algorithm that implicitly computes $\tilde{c} \in \mathcal{H}$ that with constant probability is a $(1+\varepsilon)$-approximation for the probabilistic SVDD problem. It is $\mathbb{E}_X[m(\tilde{c}, \varphi(X))] \leq (1+\varepsilon)\min_{c\in\mathcal{H}} \mathbb{E}_X[m(c, \varphi(X))]$, where the expectation is taken over the randomness of $X \sim \mathcal{D}$, and $\varphi(X) = \{\varphi(x_i) \mid x_i \in X\}$. The running time is $O\big(dn \cdot \big(z/\varepsilon^3 \cdot \log 1/\varepsilon + 1/\varepsilon^8 \cdot \log^2 1/\varepsilon\big)\big)$.*

References

[1] P. K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, 72(1):83–98, 2015.

[2] A. Krivošija and A. Munteanu. Probabilistic smallest enclosing ball in high dimensions via subgradient sampling. In *SoCG*, volume 129 of *LIPIcs*, pages 47:1–47:14, 2019.

[3] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32, 2010.

[4] A. Munteanu, C. Sohler, and D. Feldman. Smallest enclosing ball for probabilistic data. In *SoCG*, pages 214–223, 2014.

[5] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, New York, 2004.

[6] B. Schölkopf and A. J. Smola. *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[7] M. Thorup. Quick k-median, k-center, and facility location for sparse graphs. *SIAM J. Comput.*, 34(2):405–432, 2005.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A3

Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jian Jia Chen Jörg Rahnenführer

Feature Selection for Data Sets with Similar Features

Andrea Bommert
Faculty of Statistics
TU Dortmund University
andrea.bommert@tu-dortmund.de

The task of feature selection is choosing a subset of features such that all relevant information for target prediction is captured while avoiding the selection of irrelevant or redundant features. Highly similar (for example highly correlated) features in a data set make this task particularly challenging. We propose a new strategy for feature selection and model fitting on such data sets. The proposed approach is using L_0-regularized regression and tuning the hyperparameter in a multi-criteria fashion with respect to both predictive accuracy and feature selection stability. We suggest assessing the stability with an adjusted stability measure, that is, a measure that takes into account similarities between features.

We evaluate our approach based on both simulated and real data sets of different sizes and with different similarity structures between the features. Especially in situations with many similar features, our approach outperforms competing approaches. Much fewer irrelevant features are selected at almost no loss of predictive accuracy. In situations with very few similar features, it is still beneficial to consider the stability during tuning, but an unadjusted stability measure is sufficient.

1 Proposed Approach

For data sets with similar features, feature selection is very challenging. Most established methods are not able to select only one feature out of a group of similar features. They select either all or none of the similar features. One method that is able to perform such

a selection among similar features is L_0-regularized regression [1]. Therefore, we propose the following approach: Use L_0-regularized regression as predictive method and tune its hyperparameter considering both the predictive accuracy and the feature selection stability. Assess the stability of the feature selection with an adjusted stability measure. For choosing the best configuration, that is, the best hyperparameter value, employ ϵ-constraint selection, an algorithm introduced by us that focuses on predictive accuracy and employs the stability as a secondary criterion.

2 Evaluation

The proposed approach is evaluated based on both simulated and real classification data sets. Therefore, L_0-regularized logistic regression is employed. The proposed approach will be denoted by "adj" and it is compared to three competing approaches:

- "unadj": Proceed as in the proposed approach but use an unadjusted stability measure instead of an adjusted measure.
- "acc": L_0-regularized logistic regression with single-criteria hyperparameter tuning only with respect to predictive accuracy.
- "stabs": Select features with stability selection [2] using L_0-regularized logistic regression as feature selection method. Then fit an unregularized logistic regression model with the chosen features.
- "truth": Fit an unregularized logistic regression model with the features that were used for target generation. This is only possible for simulated data and gives an upper bound for the predicitve accuracy that can be achieved.

For simulated data, it is investigated whether the proposed approach is beneficial for selecting the correct features when fitting a model with high predictive accuracy on independent data. The correct features, that is, the features that have been used for generating the values of the target variable, are known for simulated data. Also, if instead of a correct feature, a highly similar feature is chosen, this other feature is accepted as correct as well. Based on simulated data, it can be observed (see Figure 1) that especially in situations with many similar features, the proposed approach is beneficial. With the proposed approach, much fewer irrelevant or redundant features are selected than with the other approaches and almost no predictive accuracy is lost compared to the best method. The predictive accuracy on independent test data is high for the methods "acc" and "unadj" as well. But with these methods, much more irrelevant or redundant features are chosen. "stabs" fails at selecting enough relevant features for an acceptable predictive accuracy. If there are no similar features, "stabs" performs best.

On real data sets, only the predictive performance and the sparsity of the models can be assessed. It can be observed (see Figure 2) that models fitted with the proposed

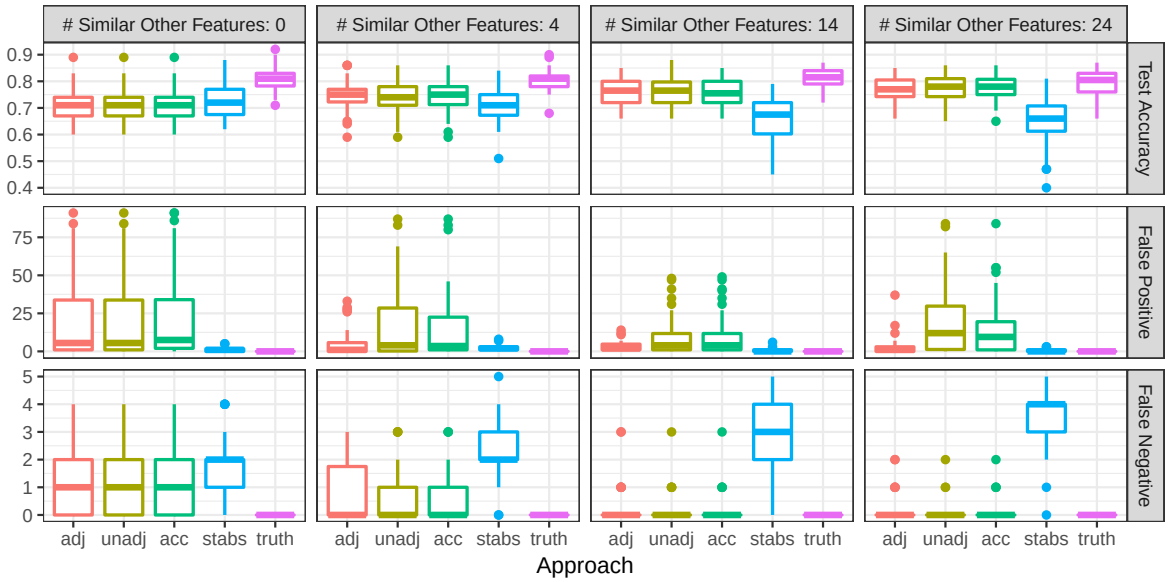


Figure 1: Results for simulated data with 100 instances, 200 features. "# Similar Other Features": Number of features in the data set that each feature is similar to (other than itself). "Test Accuracy": Predictive accuracy on independent test data, "False Positive": Number of selected features that should not have been selected, "False Negative": Number of not-selected features that should have been selected.

approach are of comparable predictive quality as the models of the other approaches and for many data sets, they are more sparse. So, with the proposed approach, the selection of irrelevant or redundant features can be avoided for many data sets. For most data sets with many similar features, the proposed approach outperforms the other approaches. For these data sets, it is necessary to consider an adjusted stability measure. For data sets with few similar features, it is still beneficial to consider the stability during hyperparameter tuning. But for these data sets, an unadjusted stability measure suffices. Also, by additionally considering the stability, almost no predictive accuracy is lost. When writing this report, the computations for "stabs" on data set "arcene" were not finished.

References

[1] Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.

[2] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

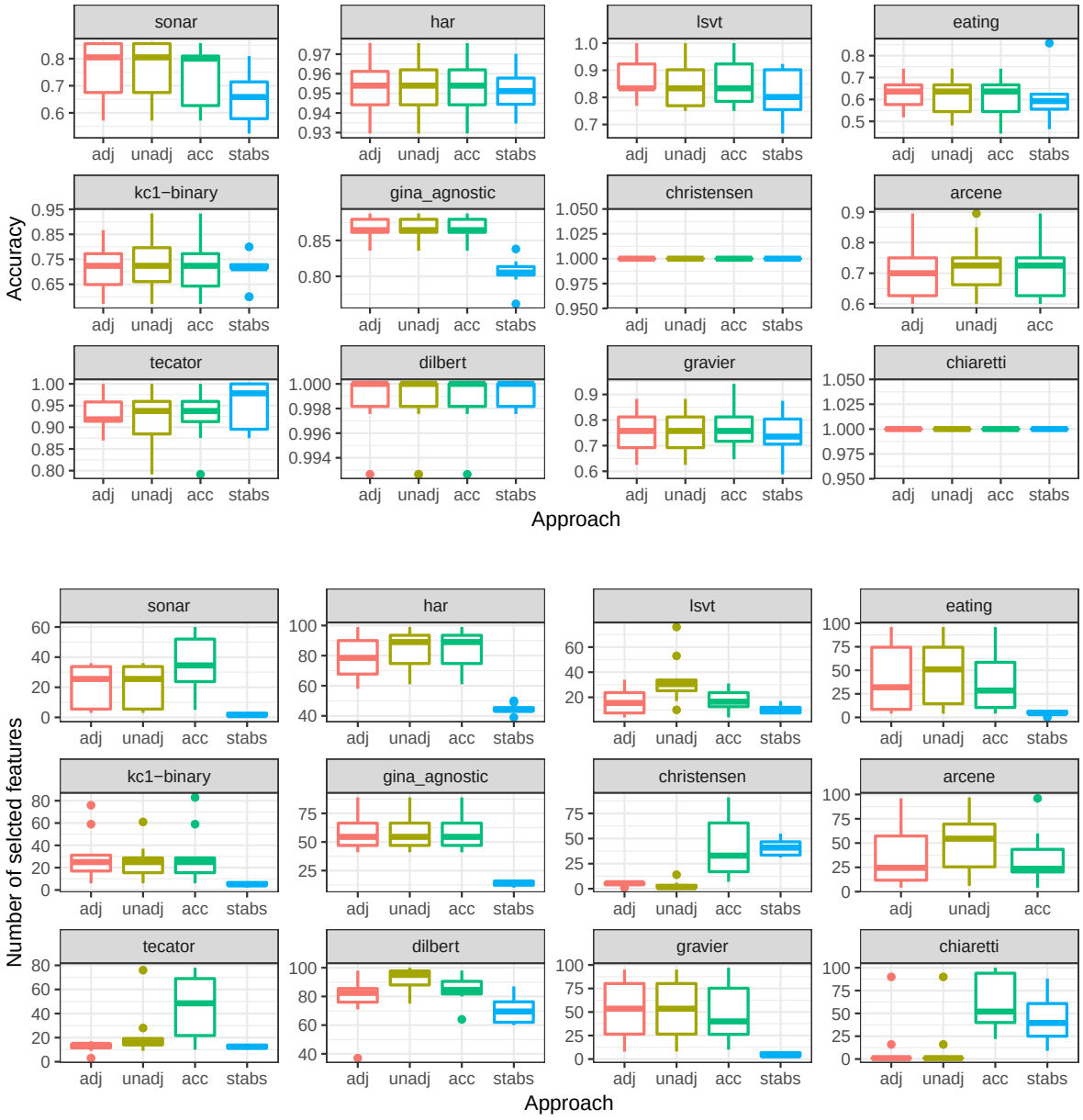


Figure 2: Results for real data sets: predictive accuracy on test data (top) and number of selected features (bottom).

Multiprocessor Synchronization of Periodic Real-Time Tasks Using Dependency Graphs

Junjie Shi
Lehrstuhl für Informatik XII
Technische Universität Dortmund
junjie.shi@tu-dortmund.de

With the increasing computation demand in real-time systems, e.g., for machine learning algorithms in autonomous driving systems, when accelerators, like GPUs, are adopted, they behave like classical shared resources with relative high utilization. Which presents a new challenge for resource synchronization in multiprocessor real-time systems. Although many resource synchronization protocols have been developed in the past decades, their performance highly depends on the task partition and prioritization. The newly proposed Dependency Graph Approach recently attracted a lot of interests due to its excellent performance. However, the original DGA can only handle frame-based tasksets, where all the tasks have the same period and deadline. We further extended the original DGA for periodic real-time tasks and developed two partitioning methods for different potential applications where tasks are tied on different processors. The evaluation results show our extension outperforms all the other existed protocols and the new developed partitioning methods performs comparable with global scheduling methods.

When considering recurrent real-time tasks in multiprocessor systems, access to shared resources, via so-called critical sections, can jeopardize the schedulability of the system. The reason is that resource access is mutual exclusive and a task must finish its execution of the critical section before another task can access the same resource. Therefore, the problem of multiprocessor synchronization has been extensively studied since the 1990s, and a large number of multiprocessor resource sharing protocols have been developed and analyzed. Most protocols assume work-conserving scheduling algorithms which make it

impossible to schedule task sets where a critical section of one task is longer than the relative deadline of another task that accesses the same resource. However, with the increasing computation demand in real-time systems, e.g., autonomous driving systems, when accelerators like GPUs are adopted, they perform like shared resource but have relative high utilizations. Which introduced a new challenging for resource synchronization in real-time multiprocessor systems.

Recently, a new non work-conserving method is presented by Chen et al. in [1], named Dependency Graph Approach where the order in which tasks access a shared resource is not determined online, but based on a pre-computed dependency graph. The initial work only considers frame-based task systems, i.e., all tasks have the same period and release their jobs always at the same time. Which limited the applicability of the approach in real applications.

Towards this, we extend the Dependency Graph Approach to periodic task systems in [2]. We point out the connection to the uniprocessor non-preemptive scheduling problem and exploit the related algorithms to construct dependency graphs for each resource. We consider a set $\mathbf{T}$ of n recurrent tasks to be scheduled on M identical (homogeneous) processors. All tasks have exactly one (non-nested) critical section where they access exactly one of the z shared resources in the system, and are each described by $\tau_i = ((C_{i,1},\ A_{i,1},\ C_{i,2}),\ T_i,\ D_i)$, where:

- $C_{i,1}$ is the worst-case execution time (WCET) of the first non-critical section of the job.
- $A_{i,1}$ is the WCET of the critical section of the job, accessing a dedicated shared resource.
- $C_{i,2}$ is the WCET of the second non-critical section.
- T_i is the period of τ_i.
- D_i is the relative deadline. We consider a constrained deadline task system, i.e., $\forall \tau_i \in \mathbf{T},\ D_i \leq T_i$.

The hyper-period H of the task set $\mathbf{T}$ is defined as the least common multiple (LCM) of the periods of the tasks in $\mathbf{T}$. In our approach, we unroll the jobs in one hyper-period and design a schedule for all of them. To make sure that the time and space complexity is affordable, we assume that the task set has *Semi-Harmonic Periods*: $T_i \cdot n_i = H\ \forall \tau_i \in \mathbf{T}$ where n_i is a small integer value.

To schedule the derived dependency graphs, one common approach is to apply list scheduling to a given task graph with precedence constraints. However, in some application scenarios, such as the OpenMP task model and multiprocessor partitioned scheduling for resource synchronization using binary semaphores, several operations can be forced to be tied to the same processor, which invalidates the list scheduling. Here, we applied two different kinds of scheduling algorithms: a) List scheduling is combined with earliest deadline first (LIST-EDF); b) Partitioned earliest deadline first (P-EDF) with two different partitioning algorithms

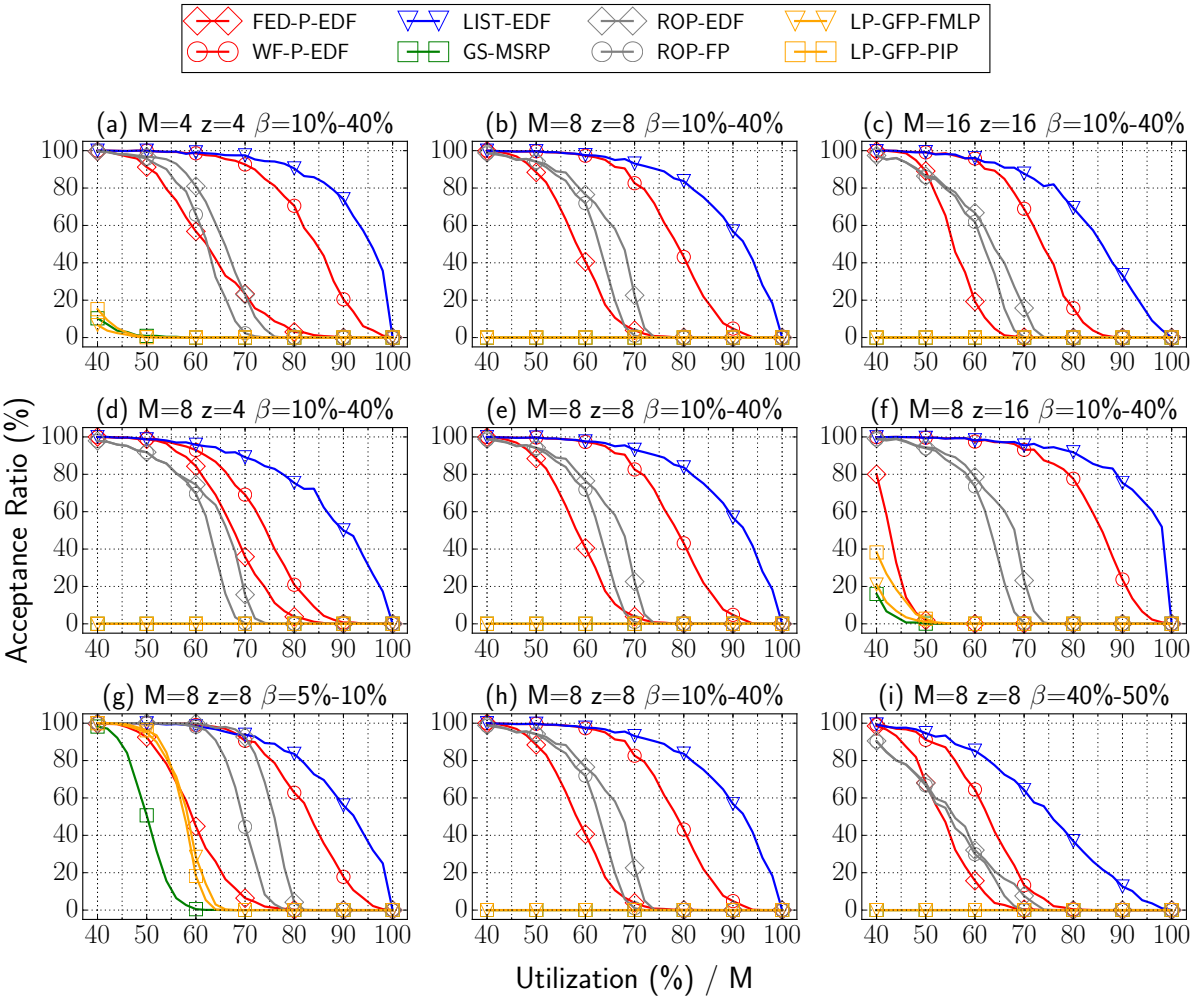


Figure 1: Schedulability of different approaches for periodic task sets.

Two different partitioning algorithms are developed in [3]. First one is based on Federated scheduling. The idea is to assign DAGs, in our case the DAGs resulting from the dependency graph construction, that need to utilize more than one processor, so called *heavy* graphs, to those processors exclusively. Analogously, the graphs that can be feasibly scheduled on a single processor are denoted as *light* graphs and are scheduled jointly on the remaining processors, i.e., non-exclusively allocated processors. After this initial partition, the actual scheduling is done by a work-conserving scheduler on the assigned processors. Second one is a worst-fit heuristic, where the tasks are partitioned one by one. The tasks are first sorted according to a sorting strategy. After that, they are partitioned to the available processors using a worst-fit strategy, i.e., each task is assigned to the processor with the currently lowest utilization. Again, P-EDF is applied to verify whether the resulting partition on M processors is feasible.

We evaluated the performance considering synthesized task sets under different configu-

Max. (Avg.) in μs	Partitioned FMLP	Partitioned List-EDF	Global FMLP	(Global) List-EDF
CXS	34.45 (0.71)	29.93 (0.81)	42.25 (1.14)	30.87 (1.79)
RELEASE	30.01 (0.63)	25.37 (0.75)	65.44 (2.98)	61.63 (12.06)
SCHED	63.91 (0.92)	32.3 (1.03)	80.81 (1.77)	59.05 (4.46)
SCHED2	33.23 (0.13)	25.24 (0.15)	31.43 (0.19)	27.17 (0.25)
SEND-RESCHED	65.81 (11.38)	20.71 (1.44)	92.78 (17.2)	72.09 (20.77)

Table 1: Overheads of different protocols in LITMUS^{RT}.

rations in Figure 1. The results show a significant improvement of LIST-EDF with respect to the acceptance ratio compared to other resource sharing protocols is observed. P-EDF using worst-fit heuristic partitioning algorithms (WF-P-EDF) outperforms all the existing partitioned approaches and perform reasonably compared to global List-EDF. Furthermore, to show the applicability in real-world systems, we implemented our extended method in LITMUS^{RT} and report the resulting scheduling overheads in Table 1.

References

[1] Jian-Jia Chen, Georg von der Bruggen, Junjie Shi, and Niklas Ueter. Dependency graph approach for multiprocessor real-time synchronization. In *IEEE Real-Time Systems Symposium, RTSS*, pages 434–446, 2018.

[2] Junjie Shi, Niklas Ueter, Georg von der Brüggen, and Jian-jia Chen. Multiprocessor synchronization of periodic real-time tasks using dependency graphs. In *2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 279–292, 2019.

[3] Junjie Shi, Niklas Ueter, Georg von der Brüggen, and Jian-Jia Chen. Partitioned scheduling for dependency graphs in multiprocessor real-time systems. In *Proceedings of the 25th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, RTCSA*, 2019.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A4

Resource efficient and distributed platforms for integrative data analysis

Michael ten Hompel Christian Wietfeld

Limits of LPWAN Technologies for Mission-Critical Machine Type Communications

Stefan Böcker
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
stefan.boecker@tu-dortmund.de

The ongoing digitalization and the associated steadily increasing number of distributed sensor devices and Internet-of-Things (IoT) systems implies a massive increase of subscribers. At the same time, the amount of available frequency spectrum resources remains static. In this respect, current 5G networks are already aiming for large-scale connectivity with an ambitious node density of 1,000,000 devices per square kilometer in the area of massive 5G machine-type communications (mMTC). Current cellular mobile radio solutions operating in licensed frequency ranges, like Narrowband-IoT (NB-IoT) or enhanced Machine Type Communications (eMTC), can barely meet these strict requirements. In this context, this report aims at summarizing the potentials of LPWAN technologies, as an additional technology option in unlicensed frequency bands, to contribute to these tight 5G requirement profiles. It can be illustrated that LoRaWAN technology can contribute significantly to the achievement of these objectives. Still, especially against the background of mission-critical MTC (mcMTC), technology limitations are also shown here. Therefore, this work proposes an approach for dynamic spectrum management in unlicensed IoT environments to further enhance capabilities with specific respect to mcMTC applications.

1 Modelling of LoRaWAN performance limits

Performance evaluation of common key performance indicators (KPIs) underlines that LoRaWAN is a sufficient MTC technology solution due to high communication ranges up to multiple kilometers, enabling a high coverage even with a small number of cells.

However, due to a simple aloha channel access mechanism in combination with strict regulatory requirements of a 1% duty cycle defined for the 868MHz short-range device (SRD) frequency band [1], this is accompanied by low data rates and high delays of several seconds in large-scale scenarios. To evaluate the maximum capability of LoRaWAN to contribute to the 5G mMTC connection density requirement, the maximum scalability of LoRaWAN is identified by analyzing the aloha channel access performance in large scale scenarios (see Fig. 1). The underlying parameter set of the depicted evaluation is defined for a generic traffic model of 32 bytes payload with a maximum transmission interval, equal to the maximum duty cycle. It can be seen that the maximum cell density for this worst-case setup is approx. 900 nodes per square kilometer combined with a system throughput of up to 3.3 kbps. Additionally, taking the capture effect into account, this cell density, as well as maximum system throughput, can be increased up to 1400, respectively 4.6 kbps. However, it can also be seen that the latency bottleneck (DR0) leads to mean delays of up to 400 s.

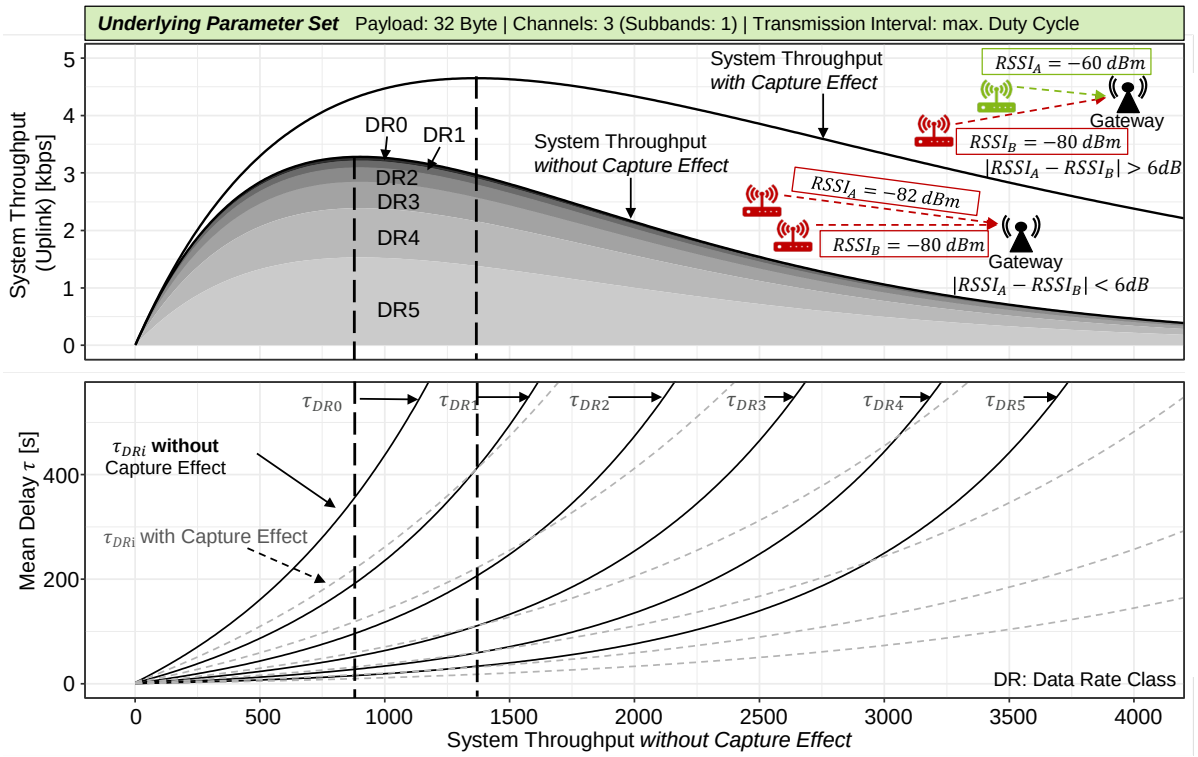


Figure 1: Modeling of LoRaWAN system throughput (uplink) and latency bounds

2 Sensitivity Analysis for Various Application Characteristics

Based on the established LoRaWAN system boundaries, a sensitivity analysis is presented to determine the capabilities of LoRaWAN to contribute to the 5G mMTC objectives. Therefore,the impact of various LoRaWAN parameter configurations on the maximum scalability is illustrated in Fig. 2. The analysis is based on the 5G mMTC traffic model of 32 bytes payload and an inter-arrival time (IAT) of two hours. A significant contribution of LoRaWAN to 5G mMTC targets can be evaluated without the consideration of a limiting latency requirement. In that case, LoRaWAN can contribute to 5G mMTC

targets of 10% (for 3 x 125 kHz channels) up to 25% (for 8 x 125 kHz channels). However, taking additional latency requirements into account, the scalability is reduced by about 25 to 70%, depending on the targeted service quality. Depending on further

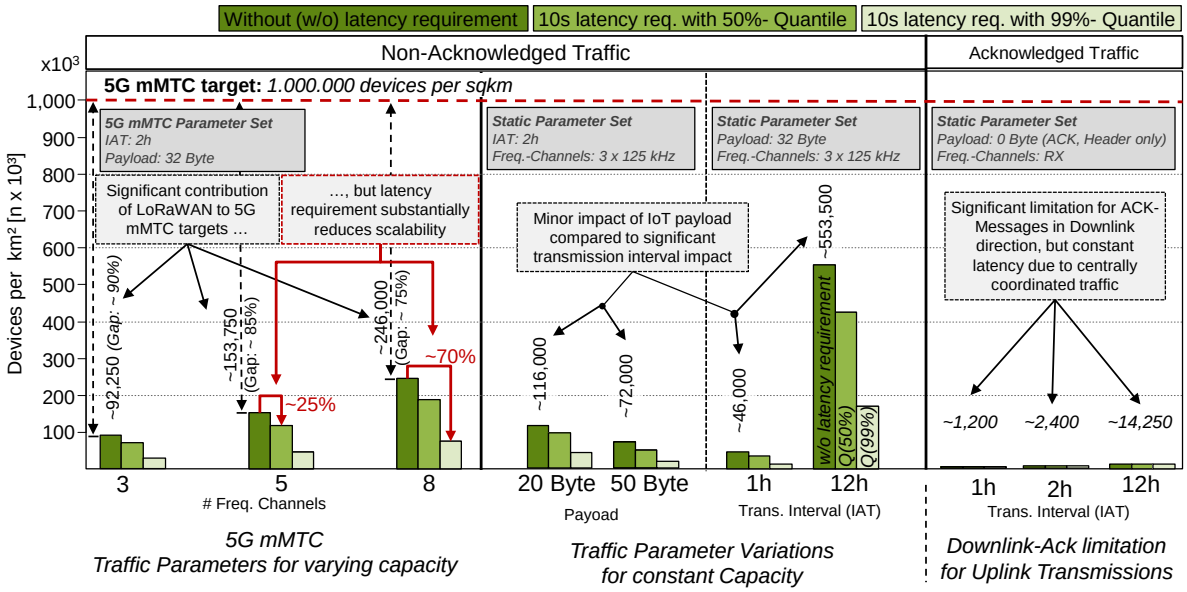


Figure 2: Impact of various LoRaWAN parameter configurations on maximum scalability considering 5G mMTC connection density and latency requirements [2]

parameter variations (payload and IAT), while considering a constant frequency capacity (3x 125 kHz), it can be seen the payload variations have only a minor impact on scalability compared to the significant impact of varying transmission intervals. In contrast, the maximum scalability is significantly limited for acknowledged traffic patterns in downlink direction. In that case, even for an IAT of 12h, the maximum scalability is reduced to approx. 14250 nodes per square kilometers, compared to more than half a million for unacknowledged traffic and no additional latency requirement.

3 Optimization Measures Toward Mission-Critical MTC

To further increase the reliability of LPWAN systems, especially for critical services, different approaches to increase spectral efficiency are discussed. In addition to pure scheduling based approaches [3], an AI-based analysis of the spectral power density to predict and avoid technology-independent interferences (see Fig. 3) is developed in ongoing work. For this purpose, the Energy Detection sensing method is implemented on the basis of a software-defined radio for the SRD frequency band. Based on the identified information of interfered frequencies in the SRD band, the ARIMA model enables a data-driven channel quality and interference prediction. The idea is to increase the robustness of LPWAN systems by centrally deriving communication profiles that address and bypass the predicted interference characteristics. The updated communication behavior are transmitted in beacons to all decentralized cell participants, which in turn adapt their

own communication decisions in favor of lower interference potentials.

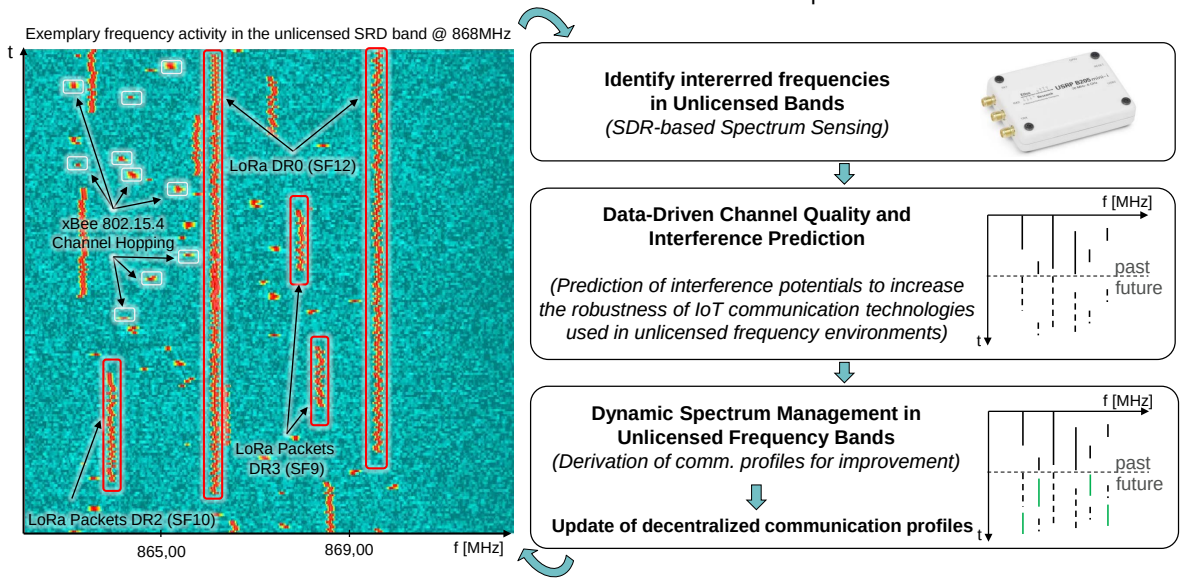


Figure 3: Dynamic Spectrum-Management to improve Scalability of mission-critical MTC Applications

4 Conclusion and Further Research

In future work, the dynamic spectrum-management approach to improve the scalability of mission-critical sensor applications will be further enhanced and evaluated in the lab and real-world environments. In addition, the reliability of mcMTC applications will be increased through ongoing improvements in measures to increase availability [4], as well as the consideration of network slicing functionality in unlicensed IoT environments [5].

References

[1] ETSI, *Short Range Devices (SRD) operating in the frequency range 25 MHz to 1 000 MHz; Part 2: Harmonised Standard covering the essential requirements of article 3.2 of Directive 2014/53/EU for non specific radio equipment*, European Telecommunications Standards Institute Std. EN 300 220-2, 2016.

[2] S. Böcker, C. Arendt, P. Jörke, and C. Wietfeld, "LPWAN in the context of 5G: Capability of LoRaWAN to contribute to mMTC," in *2019 IEEE World Forum on Internet of Things (WF-IoT 2019)*, Apr. 2019.

[3] B. Reynders, Q. Wang, P. Tuset-Peiro, X. Vilajosana, and S. Pollin, "Improving reliability and scalability of lorawans through lightweight scheduling," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1830–1842, June 2018.

[4] P. Jörke, J. Güldenring, S. Böcker, and C. Wietfeld, "Coverage and link quality improvement of cellular IoT networks with multi-operator and multi-link strategies," in *2019 IEEE 89th Vehicular Technology Conference (VTC-Spring)*, Kuala Lumpur, Malaysia, Apr. 2019.

[5] C. Bektas, S. Böcker, F. Kurtz, and C. Wietfeld, "Reliable software-defined RAN network slicing for mission-critical 5G communication networks," in *2019 IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, Hawaii, USA, Dec. 2019.

FALCON: Accurate Real-time Monitoring for Client-based Mobile Network Data Analysis

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

The development of methods to increase the quality of mobile network services requires precise data of the instant network load. However, generally only the network operator has direct access to network-side data. Instead, the entire activity of the mobile network cell can be monitored in terms of resource allocations by analyzing the control channels. Previous open-source approaches lack reliability and accuracy so that the recorded data is largely spoiled by misleading information. In this work, Fast Analysis of LTE Control channels (FALCON) is presented as a reliable open-source instrument that combines previous attempts with a novel shortcut-decoding approach for fast and accurate data extraction from public mobile networks. It is compatible with standard computers and supports numerous Software-Defined Radios (SDRs). Field measurements show that FALCON lowers the fraction of false data by three orders of magnitude, compared to the best previous approach.

1 Introduction

Recently, the 3rd Generation Partnership Project (3GPP) has begun to standardize the interface for a Network Data Analytics Function (NWDAF) [1] within a fifth generation (5G) mobile communication system to allow network components, such as the Policy Control Function (PCF), to initiate countermeasures in case of slice congestion events. The development and evaluation of such functions that mainly depend on real activity in the Radio Access Network (RAN) can only be done with representative data of the production networks. However, to obtain reliable and accurate information about the instant network load and spectral utilization of a mobile network cell is a particular

Table 1: Key validation techniques used by open-source PDCCH decoders.

	Decoder			
Technique	**LTEye** [6]	**OWL** [2]	**C³ACE** [4]	**FALCON** [5]
Signal power	●	●	○	●
Re-encoding	●	●	○	○
RAR tracking	○	●	○	●
RNTI histograms	○	○	●	●
Short-cut (**new**)	○	○	○	●

● Included, ○ Not included

challenge. Although the radio resource management is almost completely governed by the base station, generally only the operator has access to that information from the network side. A regular User Equipment (UE) by design only decodes information about its own resource allocations. Conversely, an observation of the radio spectrum only reveals the total utilization of resources and does not provide information about the number of competing network participants.

In Long Term Evolution (LTE) networks, resource assignments are not encrypted and are transmitted as Downlink Control Information (DCI) over the Physical Downlink Control Channel (PDCCH). They also include information about the direction (uplink or downlink), the Modulation and Coding Scheme (MCS) to be used, further decoding information, and finally Cyclic Redundancy Check (CRC) checksum. Generally, any LTE receiver in cell-range can decode all DCI data structures. However, the base station scrambles the CRC checksum with the 16-bit Radio Network Temporary Identifier (RNTI) of the addressed UE, so a proper validation can only be performed with prior knowledge of RNTIs of all currently active UEs.

Therefore, the major challenges are a fast and accurate discovery of active RNTIs, efficient decoding order, and accurate validation of numerous DCI candidates. This report provides a brief overview of FALCON, which is a novel reliable open-source instrument to accomplish that task by the use of general-purpose computers and SDRs. A detailed description of FALCON is provided in [5].

2 The FALCON Approach

In LTE the PDCCH carries the encoded DCI data structures which contain individual resource assignments to respective UEs. The standard defines plenty of DCI formats which differ in their decoded length and which serve specific purposes, e.g. for uplink allocations, SISO or MIMO transmissions. However, the convolutionally encoded sequence that is visible on PDCCH provides no information about the actual format of the content. As the channel lacks a table of contents, any DCI format must be considered for a message on the PDCCH, but only one of them may be correct. Due to interleaving and rate matching, the consideration of each format (i.e. different output lengths) involves

Table 2: Avg. fraction of subframes with contradictory resource assignments (downlink).

	Network 1	**Network 2**	**Network 3**
RSRP (average)	−91.23 dBm	−99.23 dBm	−107.69 dBm
RSRQ (average)	−7.11 dB	−10.05 dB	−14.39 dB
Signal Quality	Good	Fair	Poor
OWL	0.284 %	0.516 %	**2.527 %**
FALCON	0.000 %	0.001 %	**0.005 %**

a separate Viterbi decoding and CRC calculation. Valid candidates can be identified by their checksum, provided that the RNTIs are known. Since this is usually not the case, several researchers came up with different solutions. An overview is given in Tab. 1.

By considering the signal power, vacant sections of the PDCCH can be skipped, reducing the number of decoding attempts. Re-encoding the decoded sequences and comparing them to the channel bits was introduced with LTEye [6] and included in OWL [2] as a fallback. However, our analysis show, that this approach is sensitive to interference and noise and leads to numerous false positives at the end. With OWL [2] the authors introduce the capturing of initial RNTI assignments during Random Access Response (RAR) to obtain a list of active RNTIs. But RNTIs of UEs that were already active before the observation period are not captured and the re-encoding fallback is applied. Our previous approaches C^{3}ACE [3, 4] are based on the assumption that true RNTIs have a significantly higher activity than false and random RNTIs resulting from incorrect decoding. However, a trade-off between reliability and detection speed is necessary, i.e. the frequency of individual RNTIs in a time-windowed histogram exceeds a threshold value.

FALCON includes the most reliable techniques from previous attempts and introduces the new short-cut decoding approach. The data areas of the PDCCH are examined in the form of a recursive descent. Areas in which no known RNTI was found are divided into two halves and examined further. If the investigation of a shortened sequence yields the same RNTI as the complete sequence, the RNTI is marked as valid and the DCI is accepted immediately. This enables a fast discovery of RNTIs at their first occurrence in the observation period even if the initial RAR was missed. For a more detailed insight into this approach, including corner cases, the interested reader is referred to [5].

3 Evaluation and Results

In order to validate the superiority of FALCON over the previously best approach OWL, a measurement campaign over several days was carried out in three mobile networks at one location. At intervals of 5 min, 5 s of raw LTE signal were recorded and then decoded by OWL and FALCON. Since the ground-truth is unknown in the live network, the accuracy is evaluated by the fraction of contradictory and mutually exclusive resource allocations. Tab. 2 shows, that in the downlink the error rate of FALCON is significantly below that of OWL especially at poor radio conditions. The characteristics of the uplink are omitted here for space reasons and are available in [5]. Fig. 1 underlines the accuracy of FALCON. The histogram plot shows the activity of RNTIs classified as valid by the

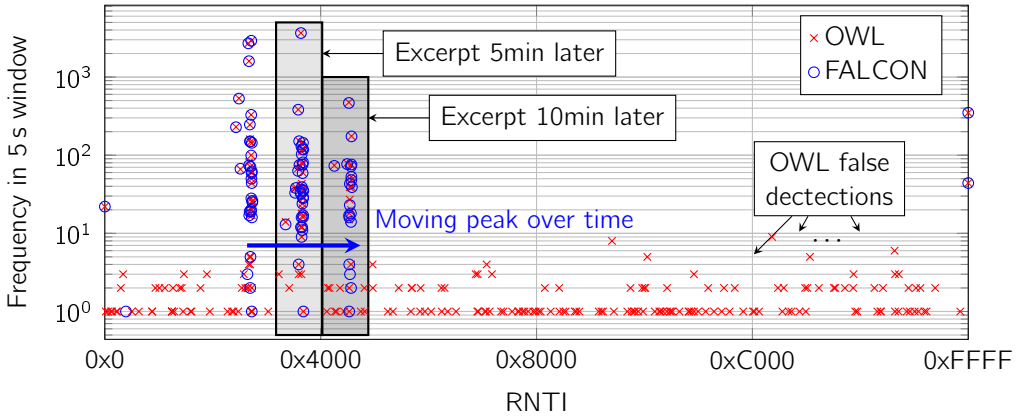


Figure 1: Distribution and Frequency of RNTIs from detected DCI by OWL and FALCON for three subsequent records in network 3 (poor signal). True RNTIs concentrate in a dense peak region that moves rightwards over time.

respective decoder. Both approaches detect the main activity in a narrow range of values that gradually moves to the right. However, OWL accepts additional false assignments that are equally distributed over the entire RNTI value range.

4 Conclusion and Perspective

This report introduced FALCON, a fast and accurate open-source instrument for real-time monitoring of the radio resource utilization in public LTE networks. Compared to the previous best approach, FALCON lowers the amount of false detection in a significant manner and paves the way for reliable data acquisition from live production networks. It provides researchers a foundation for the development and the evaluation of load-dependent functions, such as NWDAF and UE-based network selection. It also enables the derivation of realistic traffic models and an analysis of the network load over time.

References

[1] 3GPP TS 29.520 - network data analytics services (release 15), December 2018.

[2] Nicola Bui and Joerg Widmer. OWL: A reliable online watcher for LTE control channel measurements. In *All Things Cellular: Operations, Applications and Challenges*, ATC '16, New York, NY, USA, October 2016. ACM.

[3] Robert Falkenberg, Karsten Heimann, and Christian Wietfeld. Discover your competition in LTE: Client-based passive data rate prediction by machine learning. In *IEEE Globecom*, Singapore, December 2017.

[4] 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.

[5] Robert Falkenberg and Christian Wietfeld. FALCON: An accurate real-time monitor for client-based mobile network data analytics. In *GLOBECOM 2019 - 2019 IEEE Global Communications Conference*, Waikoloa, Hawaii, USA, Dec 2019. IEEE.

[6] Swarun Kumar, Ezzeldin Hamed, Dina Katabi, and Li Erran Li. LTE radio analytics made easy and accessible. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, New York, NY, USA, August 2014. ACM.

Coverage and Link Quality Improvement of Cellular IoT Networks with Multi-Operator and Multi-Link Strategies

Pascal Jörke
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
pascal.joerke@tu-dortmund.de

Mobile Network Operators (MNOs) all over the world are rolling out the fifth generation of mobile communication networks. But while these networks are rolled out, existing networks like LTE are still suffering from major coverage gaps even years after they have been introduced. However, with the upcoming Internet of Things a mobile network with comprehensive coverage is a mandatory prerequisite. This work addresses the challenge of coverage gaps and investigates the potential for improvement when using multi-operator strategies for LTE and NB-IoT networks in the urban area of the Smart City Dortmund. Besides closing coverage gaps, the potential of data rate improvement through multi-link strategies is evaluated, providing sufficient data rates of cellular IoT technologies even in extreme coupling loss conditions. Facing the overall IoT challenge of low power consumption for long battery lifetimes, an improved data rate reduces the time on air and therefore extends the battery lifetime [2]. The results show, that both LTE and cellular IoT can provide 100% coverage in outdoor scenarios, but in indoor and deep indoor scenarios, none of the Mobile Network Operators can provide full network coverage. With multi-operator strategies, LTE coverage is increased by up to 40% in deep indoor scenarios. With a MCL of 164 dB, NB-IoT can provide full coverage in deep indoor scenarios using a single operator, but can only provide very low data rates. When using multi-link strategies, NB-IoT data rates can be increased by a factor of 2.8.

1 Proposed Methods for Connectivity Enhancements

In order to analyze the potential of coverage and link quality improvement through multi-operator and multi-link strategies, the area of the Smart City Dortmund is considered as a scenario with single and multi-operator mobile networks. Fig. 1 shows a scenario with an exemplary constellation of 2 MNOs.

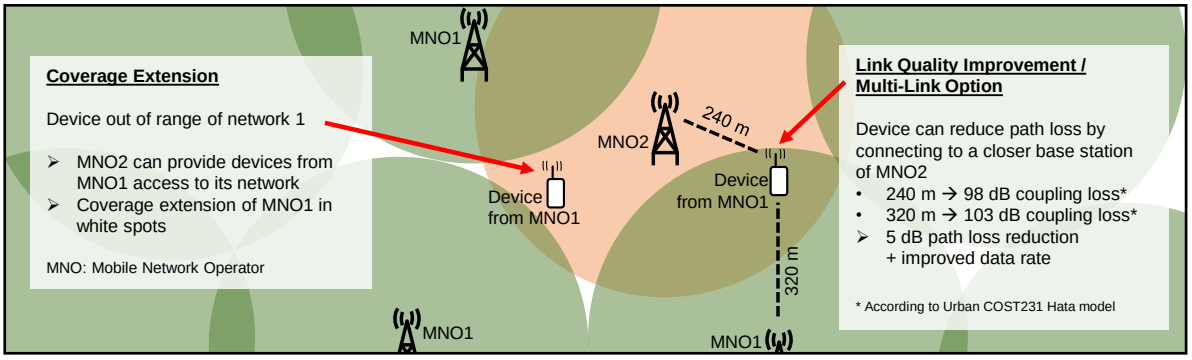


Figure 1: Nearest server analysis based on proposed grid search method for different Mobile Network Operators (MNOs) enables coverage extension as well as link quality improvement and multi-link operations.

The left device is out of range of network 1. In this case MNO2 can provide access to its network to close the coverage gap. While the device on the right has sufficient coverage by MNO1, it can increase its link quality by connecting to a closer base station of MNO2. The increased link quality can help improving the data rate of the device.

For our analysis, Dortmund is evaluated by dividing the area into a grid with a grid spacing of 25 m. Since the position and frequency of the base stations are known, the nearest base station for each MNO and grid point is searched as well as the path loss to these base stations is calculated using the COST Hata radio propagation model [1]. Additional building entry losses are derived from [3] (Table 1).

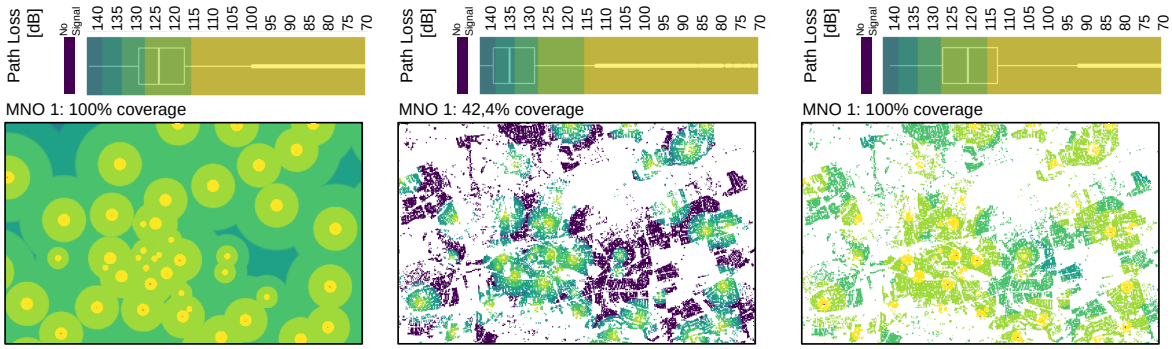
Table 1: Additional Coupling Loss for Indoor and Deep Indoor scenarios [3]

Frequency / Scenario	800 MHz / 900 MHz	1800 MHz
Indoor	15.4 dB	15.8 dB
Deep Indoor	20.9 dB	25.0 dB

After gathering all information including areal characteristics of Dortmund, such as borders and building areas, base station locations, installed technologies and frequencies, a detailed analysis on the coverage extension, link quality improvement and transport layer multilink aggregation is performed.

2 Evaluation of Coverage and Link Quality Improvement

Fig. 2 shows an extract from the coverage analysis. While LTE can provide 100% coverage in outdoor scenarios, only 42% of deep indoor areas can be reached with LTE. With a 22 dB higher Maximum Coupling Loss, NB-IoT is able to close the gaps of a single LTE MNO and provides 100% coverage in all deep indoor areas.



(a) LTE Outdoor Area (including building areas)

(b) LTE Deep Indoor (buildings only)

(c) NB-IoT Deep Indoor (buildings only)

Figure 2: Extract from coverage maps of Dortmund for Outdoor and Deep Indoor scenarios

Fig. 3 gives an overview of the results of all analyzed scenarios. Unlike in outdoor scenarios, where all MNOs can provide 100% coverage by their own, the coverage in indoor scenarios can be increased by up to 26% for two-operator scenarios. An additional third MNO further extends the coverage by at-most 9%. Therefore, depending on the coverage requirements, two collaborating MNOs may be sufficient.

With Narrowband IoT (NB-IoT), an IoT technology for long ranges and long battery lifetime has been developed by 3GPP. With up to 2048 repetitions of each transmission, the Maximum Coupling Loss (MCL) is extended to up to 164 dB. For better comparison, the coverage analysis for NB-IoT is performed for three different coverage classes, which are based on 144 dB, 154 dB and 164 dB MCL.

Since it can provide 100% coverage in all scenarios including 164 dB MCL, NB-IoT benefits less from multi operator strategies. Though it can gain its link quality up to 13.6 dB, which is a significant improvement and therefore recommended (ref. Table 2).

Average Signal Power Gain for LTE and NB-IoT	MNO 1&2		MNO 1&3		MNO 2&3		MNO 1-3 (National Roaming)		
	MNO 1	MNO 2	MNO 1	MNO 3	MNO 2	MNO 3	MNO 1	MNO 2	MNO 3
Outdoor	9.5 dB	8.0 dB	10.2 dB	8.3 dB	9.0 dB	7.8 dB	11.4 dB	10.0 dB	9.0 dB
Indoor	11.2 dB	9.1 dB	12.3 dB	10.1 dB	10.2 dB	9.0 dB	13.6 dB	11.0 dB	10.6 dB
Deep Indoor	10.4 dB	8.9 dB	11.0 dB	9.1 dB	9.6 dB	8.4 dB	12.6 dB	10.9 dB	10.0 dB

Table 2: Results of the coupling loss reduction potential for different coupling loss scenarios and cellular communication technologies

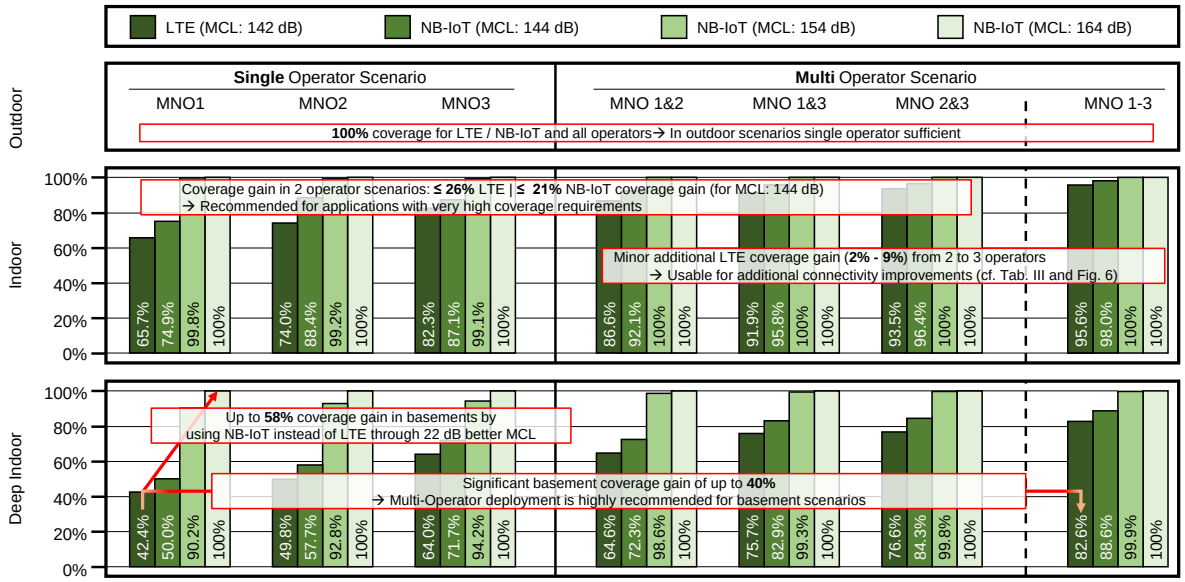


Figure 3: Results of the coverage analysis for outdoor, indoor and deep indoor scenarios and different cellular communication technologies in urban environment

Taking multi operator strategies into account, NB-IoT devices can improve their data rate from 3.8 kbit/s - 5.0 kbit/s to 12.2 kbit/s, reducing time on air for e.g. firmware update transmissions and therefore extending the battery lifetime of the devices.

3 Conclusion

Enabling 100% coverage for LTE devices even in deep indoor scenarios, multi-operator strategies are highly recommended for LTE networks. With 164 dB MCL, NB-IoT can provide full coverage for every MNO without the usage of multi-operator strategies. Although with multi-link strategies the overall data rate can be increased by a factor of 2.8, making the usage of multiple operators recommended as well.

References

[1] E. Damosso. *Digital mobile radio towards future generation systems: COST action 231*. European Commission, 1999.

[2] P. Jörke, R. Falkenberg, and C. Wietfeld. Power Consumption Analysis of NB-IoT and eMTC in Challenging Smart City Environments. In *IEEE Global Communications Conference (GLOBECOM) Workshops, Workshop on Green and Sustainable 5G Wireless Networks*, Abu Dhabi, United Arab Emirates, Dec 2018.

[3] S. Monhof, S. Bocker, J. Tiemann, and C. Wietfeld. Cellular Network Coverage Analysis and Optimization in Challenging Smart Grid Environments. In *2018 IEEE SmartGridComm*, pages 1–6, Oct 2018.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A6

Resource-efficient Graph Mining

Nils Kriege Petra Mutzel
Frank Weichert

Maximum Matching Problems: All Cavity Bipartite Matching and a Recursive Matching Approach

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

The all-cavity maximum weight matching problem is defined as follows. Given a weighted bipartite graph $G = (U \uplus V, E)$, determine a maximum weight matching on $G \setminus \{v\}$ for each vertex $v \in U \uplus V$. This problem has been introduced in the computation of unrooted evolutionary trees and was later used as a backbone to compute maximum agreement subtrees. Another application is in the computation of a subgraph homeomorphism between trees and when computing a subtree isomorphism. We study the all-cavity maximum weight matching problem for balanced graphs followed by known and new results on unbalanced graphs.

For engineering maximum matching algorithms we suggest new concepts on general graphs which can be used as alternatives in augmenting path approaches such as, e.g., Edmonds or Micali and Vazirani. Our newly introduced *alternating rooted sets (ARSs)* for finding augmenting paths generalize the state-of-the-art *alternating trees*. We experimentally evaluate our new recursive metagraph approach on a wide set of benchmark instances including a comparison to publically available state-of-the-art software.

1 All Cavity Maximum Weight Bipartite Matching

Let $G = (U \uplus V, E, w)$ be a bipartite graph with n vertices and m edges and no isolated vertices. Let $w : E \to \mathbb{N}$ be upper bounded by N. Kao, Lam, Sung, and Ting [6]

Author	Restriction	Running time
Kao et al. (1997)	integral weights	$\mathcal{O}(\sqrt{n}m \log N)$
D.; based on Kao et al. (1997)		$\mathcal{O}(nm + n^2 \log n)$
D.; based on Chung (1987)	unweighted	$\mathcal{O}(m\sqrt{s})$
Milo et al. (2013)		$\mathcal{O}(s^3 + st)$
Droschinsky et al. (2016)		$\mathcal{O}(s^2 t + st \log t)$
– considering m		$\mathcal{O}(ms + st \log t)$
Droschinsky et al. (2018)		$\mathcal{O}(s^2 t)$
Droschinsky, to be published		$\mathcal{O}(\min\{s^2, m\}s + s^2 \log s + m)$
	integral weights	$\mathcal{O}(\min\{s^2, m\}\sqrt{s} \log N + m)$

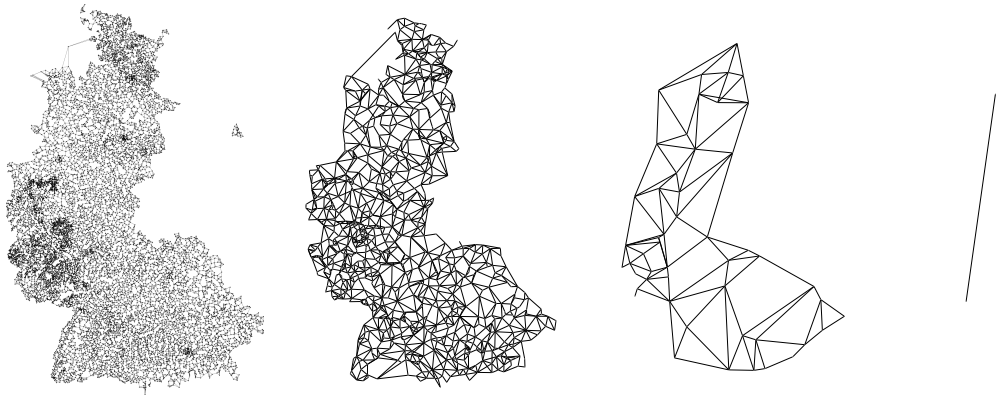
Table 1: Worst-case running times for the all-cavity maximum weight matching problem on a bipartite graph $(U \uplus V, E)$ without isolated vertices, where $s := |U| \leq |V| =: t$, $m := |E|$, $n := s + t$, and N as maximum edge weight.

solved the all-cavity maximum weight matching problem on G by reducing it to the single-destination longest paths problem. When the work of Kao et al. was published they showed an upper time bound of $\mathcal{O}(\sqrt{n}m \log(nN))$. However, with a result from Duan and Su for computing a maximum weight matching this bound may be improved to $\mathcal{O}(\sqrt{n}m \log N)$ [5].

The result from Kao et al. transfers directly to arbitrary edge weights, i.e., each edge has a real-valued weight ($\mathbb{R}$). The single-destination longest paths problem without positive weight cycles may be solved with the well known Bellman-Ford algorithm in time $\mathcal{O}(nm)$. Using the Hungarian method for the initial maximum weight matching allows a total time bound of $\mathcal{O}(nm + n^2 \log n)$.

For unbalanced graphs, i.e., graphs where one vertex set is much smaller than the other, we want so seek running times based on the smaller set. Here, we need maximum weight matching algorithms that consider unbalanced bipartite graphs, e. g. the algorithm from Ramshaw and Tarjan [2012]. They proved a running time of $\mathcal{O}(ms + s^2 \log s)$, where s is the size of the smaller vertex set. Applying the Bellman-Ford algorithm on the reduced graph from Kao et al. exceeds that time bound. We researched a reduction on a smaller graph with only $\mathcal{O}(s)$ vertices, which allows solving the single-source shortest paths problem in time $\mathcal{O}(ms)$. We also studied the approach on graphs with integral edge weights of at most N. Table 1 summarizes the results for the different all-cavity maximum weight matching problems.

Using our new results for the all-cavity maximum weight matching problem allows an improved time bound for the maximum common subtree problem on integral weights



(a) The base graph (b) The first metagraph (c) The second metagraph (d) The third metagraph

Figure 1: Different layers of the recursive metagraph approach on the 4 nearest neighbor graph of the brd14051 TSPLib instance. Nodes in the metagraph have the same coordinates as the root of their represented tree.

compared to real-valued weights, which was a remaining open question in [3]. Let T and T' be trees and $\Delta := \min\{\Delta(T), \Delta(T')\}$, where $\Delta(T)$ is the maximum degree of T. We showed, that is is possible to compute a maximum common subtree isomorphism between T and T' under a weight function ω between the vertices and edges in time $\mathcal{O}(|T||T'|\sqrt{\Delta}\log(N\min\{|T|,|T'|\})$, if ω is integral and bounded by N.

2 A Recursive Maximum Matching Approach

Computing matchings in bipartite graphs is easy and can be solved using alternating trees. However, they cannot be applied to general graphs, as those graphs may contain odd cycles. Let G be an undirected graph and $\mathcal{M}$ a matching on G. An alternating rooted set (ARS) is a set of vertices $V \subseteq V(G)$ with exactly one $\mathcal{M}$-free node $r \in V$ called root, such that for every node $u \in V$ there exists an alternating path from r to u in the subgraph of G induced by V. We showed, that these ARSs generalize the state-of-the-art alternating trees. A data structure, which we called cherry trees, is an implementation of ARSs [4]. This data structure allows us to recursively compute a maximum matching as shown in Figure 1. We denote the graphs in this recursive approach metagraphs. We experimentally compared this new approach with state-of-the-art algorithms for maximum matching; cf. Figure 2 for average running times.

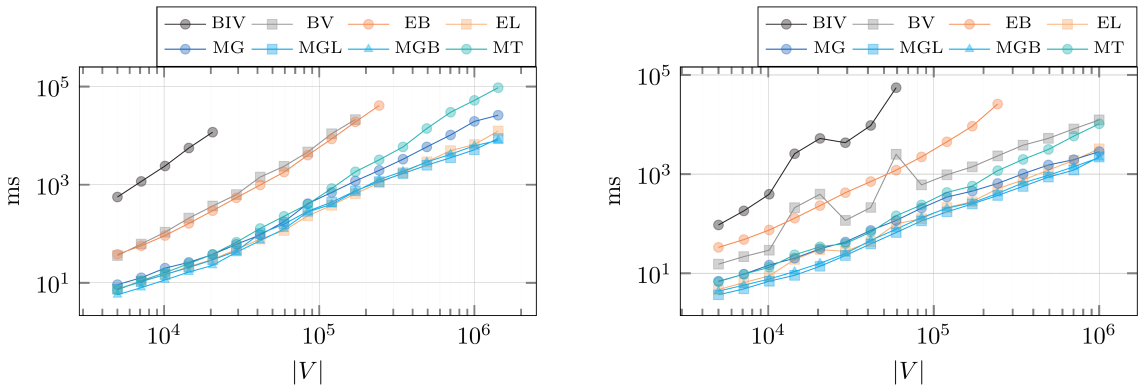


Figure 2: Average running times for different maximum matching algorithms. MG is the Metagraph approach; MGL and MGB are variants thereof. BIV and BV are Blossom IV and V. EB and EL are implementations of Edmonds' matching algorithm in the Boost and the LEMON framework.

References

[1] M. J. Chung. "$O(n^{2.5})$ time algorithms for the subgraph homeomorphism problem on trees". In: *Journal of Algorithms* 8.1 (1987), pp. 106–112. DOI: 10.1016/0196-6774(87)90030-7.

[2] A. Droschinsky, N. M. Kriege, and P. Mutzel. "Faster Algorithms for the Maximum Common Subtree Isomorphism Problem". In: *41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016)*. 2016.

[3] A. Droschinsky, N. M. Kriege, and P. Mutzel. "Largest Weight Common Subtree Embeddings with Distance Penalties". In: *43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018)*. 2018. DOI: 10.4230/LIPIcs.MFCS.2018.54.

[4] A. Droschinsky, P. Mutzel, and E. Thordsen. "Shrinking Trees not Blossoms: A Recursive Maximum Matching Approach". In: *2020 Proceedings of the Symposium on Algorithm Engineering and Experiments (ALENEX)*, pp. 146–160. DOI: 10.1137/1.9781611976007.12.

[5] R. Duan and H.-H. Su. "A Scaling Algorithm for Maximum Weight Matching in Bipartite Graphs". In: *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '12. Kyoto, Japan: SIAM, 2012, pp. 1413–1424.

[6] M. Kao, T. W. Lam, W. Sung, and H. Ting. "All-Cavity Maximum Matchings". In: *Algorithms and Computation, 8th International Symposium, ISAAC '97, Singapore, December 17-19, 1997, Proceedings*. 1997, pp. 364–373.

[7] N. Milo, S. Zakov, E. Katzenelson, E. Bachmat, Y. Dinitz, and M. Ziv-Ukelson. "Unrooted unordered homeomorphic subtree alignment of RNA trees". In: *Algorithms for Molecular Biology* 8 (2013), p. 13. DOI: 10.1186/1748-7188-8-13.

Towards Effective Graph Representation Learning

Matthias Fey
Department of Computer Graphics
TU Dortmund University
matthias.fey@tu-dortmund.de

Graph neural networks (GNNs) have been emerged as one of the most successful approaches to tackle the representation learning on graphs and point clouds. Here, we introduced a deep learning library for fast graph representation learning on graphs, and introduced a novel neighborhood aggregation formulation that tackles the problem of deeply stacking GNN layers. Furthermore, we proposed a two-stage neural architecture for the deep graph matching problem which is able to establish meaningful structural correspondences of nodes between graphs. Lastly, we tackled the problem of shape generation via an adversarial formulation based on implicit shapes.

1 Fast Graph Representation Learning

We introduced *PyTorch Geometric* [2], a library for deep learning on irregularly structured input data such as graphs, point clouds and manifolds, built upon PyTorch. In addition to general graph data structures and processing methods, it contains a variety of recently published methods from the domains of relational learning and 3D data processing. PyTorch Geometric achieves high data throughput by leveraging sparse GPU acceleration, by providing dedicated CUDA kernels and by introducing efficient mini-batch handling for input examples of different size. PyTorch Geometric is released under the MIT license and is available on GitHub.[1]

[1] `https://github.com/rusty1s/pytorch\_geometric`

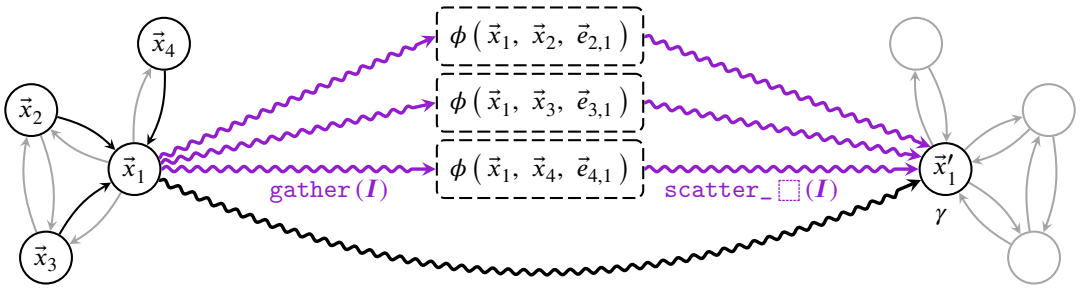


Figure 1: Computation scheme of a GNN layer by leveraging gather and scatter methods, hence alternating between node parallel space and edge parallel space.

Neighborhood Aggregation. Generalizing the convolutional operator to irregular domains is typically expressed as a *neighborhood aggregation* or *message passing* scheme

$$\vec{x}'_i = \gamma\left(\vec{x}_i,\ \square_{j\in\mathcal{N}(i)}\ \phi\left(\vec{x}_i,\ \vec{x}_j,\ \vec{e}_{j,i}\right)\right) \tag{1}$$

where $\square$ denotes a differentiable, permutation invariant function, *e.g.*, sum, mean or max, and γ and ϕ denote differentiable functions, *e.g.*, MLPs. In practice, this can be achieved by gathering and scattering of node features and vectorized element-wise computation of γ and ϕ, as visualized in Figure 1. Although working on irregularly structured input, this scheme can be heavily accelerated by the GPU. In contrast to implementations via sparse matrix multiplications, the usage of gather/scatter proves to be advantageous for low-degree graphs and non-sorted input, and allows for the integration of central node and multi-dimensional edge information while aggregating. We provide the user with a general `MessagePassing` interface to allow for rapid and clean prototyping of new research ideas. In addition, we already implemented over 20 convolutional operators often found in literature.

2 Dynamic Neighborhood Aggregation

Most of proposed message passing operators result in in gradually decreasing performance when deeply stacking those layers, despite having, in principal, access to a wider range of information. This phenomenon is caused by locally differeing graph structures leading to strongly varying speed of expansion, and is tackled, *e.g.*, by node-adaptively *jumping back* to earlier representations if those fit the task at hand more precisely. Inspired by this, we explored a highly *dynamic neighborhood aggregation (DNA)* procedure based [1]. However, in contrast, we proposed to allow jumps to earlier knowledge immediately *while* aggregating information from neighboring nodes. This results in a highly-dynamic receptive-field in which neighborhood information is potentially gathered from representations of differing locality. Therefore, each node's representation controls

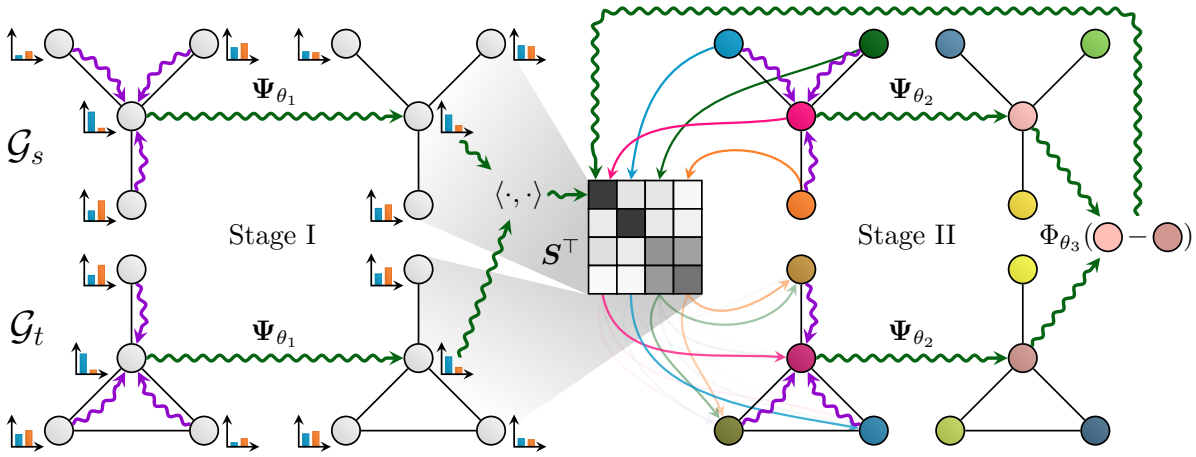


Figure 2: High-level illustration of our two-stage matching consensus architecture.

its own spread-out, possibly aggregating more global information in one branch, and falling back to more local information in others. We showed that this approach, when additionaly combined with grouped linear projections, outperforms traditional stacking of GNN layers, even when those are enhanced by traditional jumping.

3 Deep Graph Matching Consensus

Graph matching refers to the problem of establishing meaningful *structural correspondences* of nodes between two or more graphs by taking both node similarities and pairwise edge similarities into account. We also presented a two-stage neural architecture for learning and refining structural correspondences between two graphs $\mathcal{G}_s$ and $\mathcal{G}_t$ [3], *cf.* 2. First, we use localized node embeddings computed by a graph neural network $\mathbf{\Psi}_{\theta_1}$ to obtain an initial ranking of soft correspondences between nodes. Secondly, we employ synchronous message passing networks $\mathbf{\Psi}_{\theta_2}$ based on random node colorings to iteratively re-rank the soft correspondences to reach a matching consensus in local neighborhoods between graphs. We showed, theoretically and empirically, that our message passing scheme computes a well-founded measure of consensus for corresponding neighborhoods, which is then used to guide the iterative re-ranking process. Our purely local and sparsity-aware architecture scales well to large, real-world inputs while still being able to recover global correspondences consistently. We demonstrated the practical effectiveness of our method on real-world tasks from the fields of computer vision and entity alignment between knowledge graphs, on which we improve upon the current state-of-the-art.

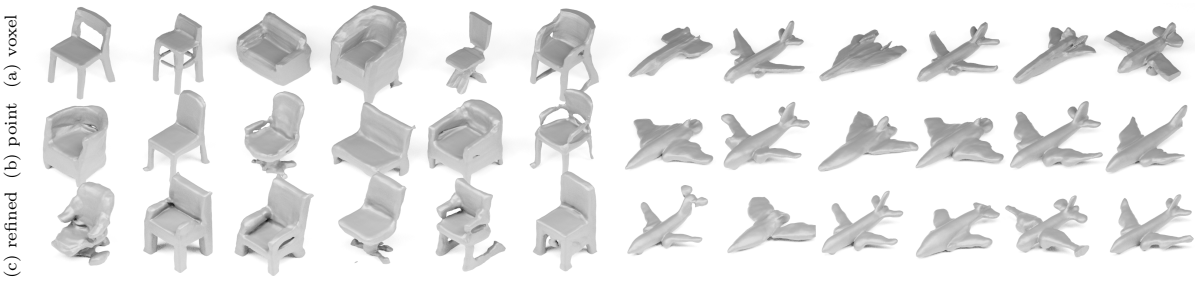


Figure 3: Qualitative examples of generated shapes.

4 Adversarial Generation of Implicit Shapes

We presented a generative adversarial architecture for generating three-dimensional shapes based on signed distance representations [4]. Here, our generator learns to approximate the signed distance for any point in space given prior latent information. Although structurally similar to generative point cloud approaches, this formulation can be evaluated with arbitrary point density during inference, leading to fine-grained details in generated outputs. Furthermore, we studied the effects of using either progressively growing voxel- or point-processing networks as discriminators, and proposed a refinement scheme to strengthen the generator's capabilities in modeling the zero iso-surface decision boundary of shapes. We trained our approach on the ShapeNet benchmark dataset and validated, both quantitatively and qualitatively, its performance in generating realistic 3D shapes. Qualitative results of our models are shown in Figure 3.

References

[1] Matthias Fey. Just jump: Dynamic neighborhood aggregation in graph neural networks. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[2] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[3] Matthias Fey, Jan E. Lenssen, Christopher Morris, Jonathan Masci, and Nils M. Kriege. Deep graph matching consensus. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] Marian Kleineberg, Matthias Fey, and Frank Weichert. Adversarial generation of continuous implicit shape representations. In *Under review at Eurographics 2020*, 2020.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B3

Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik
Petra Wiederkehr

Synchronization of measured and simulated force signals of milling processes

Felix Finkeldey
Chair 14 for Software Engineering
Department of Computer Science
TU Dortmund University
felix.finkeldey@tu-dortmund.de

In order to analyze milling processes, measurement techniques and simulation approaches can be used to evaluate certain process characteristics, e.g., process forces or tool vibrations. Various challenges arise if the matching of time series of these characteristics, acquired by different data sources, is necessary for, e.g., performing a supervised learning procedure. These challenges comprise different sample frequencies, deviating amplitudes of the signals or measurement noise. This work presents different methods to synchronize time series, acquired during milling, sample by sample as a processing operation for further learning tasks.

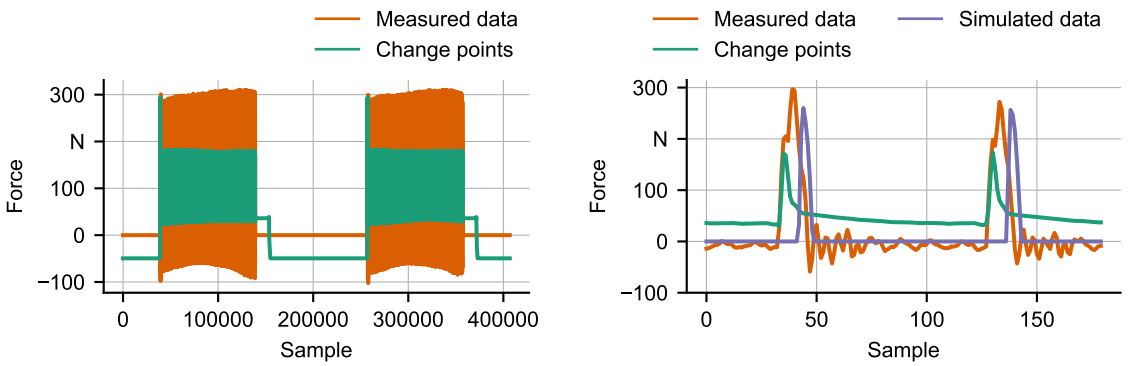
1 Introduction

Simulation techniques offer numerous optimization potentials for various manufacturing processes. For milling operations, geometric physically-based process simulation approaches can be used to generate location errors on the workpiece surface and time series of process forces and tool vibrations, which are crucial process characteristics and can be utilized for an analysis of the resulting workpiece quality for different process parameter values [1]. Different empirical models, which can be used to represent the relationship between geometric features of cutting operations and the process characteristics of interest, comprise a set of parameters, which have to be calibrated on the combination of the

tool geometry and workpiece material of the considered milling process. These relationships tend to be highly complex and challenging to model analytically, which can result in high deviations between measured and simulated characteristics. The incorporation of machine learning (ML)-based methods is a promising approach to enable predictions with a suitable accuracy on unseen data. Using ML models to represent process characteristics of milling operations inside process simulation systems, each sample of the time series of the simulated features has to be aligned with the corresponding measured sample of the considered target characteristic. This report presents different approaches to perform the matching between the samples of measured and simulated time series of process forces.

2 Synchronization of time series

Usually, there is a fixed value for the tooth feed f_z for each process configuration for milling processes. Given the sample frequency f_s of either measurements or simulation data, the number of samples s_z of the time series corresponding to a single tooth feed can be calculated as $s_z = 60 \cdot \frac{z \cdot f_s}{n}$, where z is the number of cutting edges and n is the spindle speed. For calibration experiments, often simple slot milling operations are utilized to generate a suitable data set of measurements. In this case, the total number of tooth engagements per slot and, therefore, samples of the time series is fixed and can be estimated by incorporating the width of cut. Consider a constant and equaling sample frequency for the measured and simulated time series for a slot milling process using a fixed value for the width of cut. In addition, consider the negligence of the acceleration and deceleration properties of the machine drives, which can result in a deviating amount of performed tooth engagements in the experiments compared to the analytically calculated, idealized amount of tooth engagements. In this case, only a time-related delay between the two regarded time series has to be identified. To achieve this, the absolute difference $\left| i_{1,e}^{(p)} - i_{1,e}^{(s)} \right|$ between the indices $i_{1,e}^{(p)}$ and $i_{1,e}^{(s)}$, which denote the indices of the first engagements between the tool and the workpiece for measured $\Xi^{(p)} = \left\{ \xi_1^{(p)}, \ldots, \xi_N^{(p)} \right\}$ and simulated time series $\Xi^{(s)} = \left\{ \xi_1^{(s)}, \ldots, \xi_M^{(s)} \right\}$, respectively, can be utilized. Consider the measured and simulated time series of process forces of two milled slots as well as a magnified illustration of two tooth engagements, which are depicted in Figure 1. Identifying $i_{1,e}^{(s)}$ for $\Xi^{(s)}$ is trivial. Since no noise is present in simulated data, $i_{1,e}^{(s)}$ is the index of the first sample whose value is greater than zero. For measured forces, often noise and vibrations can superimpose the signals. A promising approach to find $i_{1,e}^{(p)}$ is to extract features for each sample of $\Xi^{(p)}$, which indicate the probability of the sample corresponding to an engagement between the tool and the workpiece. The sequentially discounting auto-regression model learning algorithm [2] can be used to determine change points for a given time series which quantify deviations in the semantic behavior covered by the time series. For the measured time series shown



(a) (b)

Figure 1: (a) Measured time series with the corresponding set of change points of two milled slots and (b) a magnified illustration of simulated and measured time series and the time series of change points

in Figure 1, a clear increase of the values of the change points can be observed as soon as there is a shift from no engagement to a cutting operation, enabling the use of simple thresholds to classify the samples.

The major limitations of the presented approach are the numerous necessary preconditions, e.g., the process being a slot milling process and neglecting deviations of the length of each engagement between the simulation and the real process. Utilizing the wavelet transform [3], the spectrum of the time series can be analyzed without the necessity to deal with the trade-off between the time and frequency resolution. Figure 2 shows measured and simulated time series of six tooth engagements and their corresponding wavelet transform using the mexican hat mother wavelet, since it resembles the shape of a forces of a tooth engagement [4]. The intensity of the wavelet transform is the convolution between a scaled mother wavelet, representing a certain frequency, and the signal, indicating the correlation between the amplitude of a sample with a frequency. The classification of a sample to the state of engagement or periods of no engagement can be performed by evaluating the wavelet transform of the considered signal at the tooth engagement frequency. If the intensity is greater than zero, the sample corresponds to an engagement. Using this approach, each sample of time series can be classified, assuming the tooth engagement frequency is known. These time series can be acquired from processes of arbitrary parameter values.

3 Further work

As future work, the approach of using the wavelet transform to classify the samples of time series will be utilized to perform the matching between simulated features and

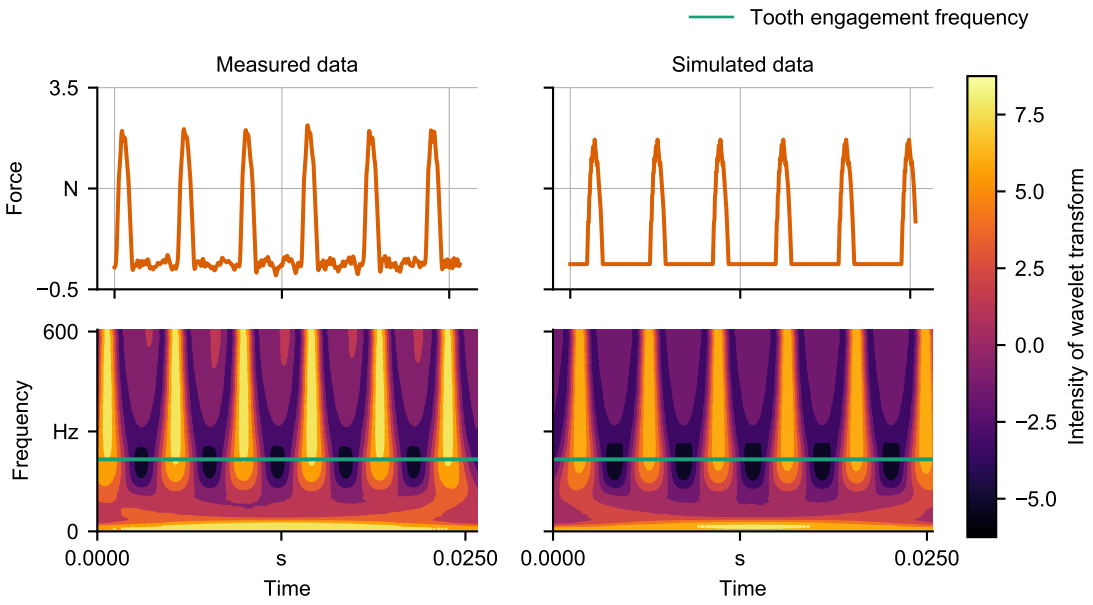


Figure 2: Wavelet transform of measured and simulated data

measured targets for supervised learning tasks to model process characteristics of milling operations, e.g., process forces or tool vibrations.

References

[1] P. Wiederkehr and T. Siebrecht, "Virtual machining: Capabilities and challenges of process simulations in the aerospace industry," *Procedia Manufacturing*, vol. 6, no. Supplement C, pp. 80–87, 2016. 16th Machining Innovations Conference for Aerospace Industry – MIC 2016.

[2] K. Yamanishi and J. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 676–681, 2002.

[3] A. Grossmann and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

[4] W. K. Ngui, M. S. Leong, L. M. Hee, and A. M. Abdelrhman, "Wavelet analysis: Mother wavelet selection methods," in *Advances in Manufacturing and Mechanical Engineering*, vol. 393, pp. 953–958, 11 2013.

A Drift-based Dynamic Ensemble Members Selection using Clustering

Amal SAADALLAH
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
amal.saadallah@tu-dortmund.de

This report sums up our recent research to use dynamic model selection for ensemble pruning for time series forecasting. This work is devised in the context of the work package "Management of many models". A drift detection mechanism for model-based performance is devised to select best performing ensemble members. In addition, model clustering is proposed to enhance ensemble diversity. We evaluate our framework using many real-world data sets. Experimental results are detailed in this report, followed up by our future work within the same topic.

1 Introduction

Both complex and evolving nature of time series structure make forecasting among one of the most important and challenging tasks in time series analysis. Typical methods for forecasting are designed to model time-evolving dependencies between data observations. However, it is generally accepted that none of them is universally valid for every application. Therefore, methods for learning heterogeneous ensembles by combining a diverse set of forecasts together appear as a promising solution to tackle this task. Hitherto, in classical ML literature, ensemble techniques such as stacking, cascading and voting are mostly restricted to operate in a static manner. To deal with changes in the relative performance of models as well as changes in the data distribution, we propose a drift-aware meta-learning approach for adaptively selecting and combining forecasting models. Our assumption is that different forecasting models have different areas of expertise and a varying relative performance. Our method ensures dynamic selection of initial

ensemble base models candidates through a performance drift detection mechanism. Since diversity is a fundamental component in ensemble methods, we propose a second stage selection with clustering that is computed after each drift detection. Predictions of final selected models are combined into a single prediction. An exhaustive empirical testing of the method was performed, evaluating both generalization error and scalability of the approach using time series from several real world domains. Empirical results show the competitiveness of the method in comparison to state-of-the-art approaches for combining forecasters.

2 Methodology

2.1 Problem Formulation

A time series X is a temporal sequence of values $X = \{x_1, x_2, \cdots, x_t\}$, where x_i is the value of X at time i. Typical solution for time series forecasting include traditional univariate time series analysis models, such as the popular Box-Jenkins ARIMA family of methods [1] or exponential smoothing methods [1]. Typical regression models can be applied in the context of forecasting by using a time delay embedding which maps a set of target observations to a K-dimensional feature space corresponding to the K past lagged values of each observation [1].

Denote with $P_M = \{M_1, M_2, \cdots, M_N\}$ the pool of trained base forecasting models. Let $\hat{x} = (\hat{x}_{M_1}, \hat{x}_{M_2}, \cdots, \hat{x}_{M_N})$ be the vector of forecast values of X at time instant $t+1$ (i.e. x_{t+1}) by each of the base model in P_M. The goal of the dynamic selection is identifying which $\hat{x}_{M_i}$ values should be integrated in the weighted average.

2.2 A drift-based model pre-selection

Suppose we want to compute the prediction for time instant $t+1$, the validation sliding-window of size W over X is defined by the sequence $X_{W,t} = \{x_{t-W+1}, x_{t-W+2}, \cdots, x_t\}$. Let $\hat{X}^{M_i}_{W,t} = \{\hat{x}^{M_i}_{t-W+1}, \hat{x}^{M_i}_{t-W+2}, \cdots, \hat{x}^{M_i}_{t}\}$ be the predicted sequence of values by the model M_i on $X_{W,t}$, where $M_i \in P_M$. A subset K of highly correlated models with the target, denoted "top-base" models, are selected using a sliding-window similarity measure computed on $X_{W,t}$. K is a user-defined hyperparameter. Hereby, we propose to use a custom measure based on the Pearson's correlation - commonly used to deal with time series data-denoted as *SRC*-Scaled Root Correlation and defined. Naturally, with time-evolving data, dependencies change over time and follow non-stationary concepts. We can assume that the distance between the two most dissimilar random processes within the same pool of models sets its boundary under a form of a logical *diameter*. If this boundary diverges in a significant way over time, a drift is assumed to take place. We propose to detect the validity of

such assumption using the well-known Hoeffding Bound [3], which states that after W independent observations of a real-value random variable with range R, its true mean has not diverged if the sample mean is contained within $\pm\epsilon_F$:

$$\epsilon_F = \sqrt{\frac{R^2 \ln(1/\delta)}{2W}} \tag{1}$$

with a probability of $1 - \delta$ (a user-defined hyperparameter). Once a drift is detected, an alarm is triggered, the top base models using C_t are updated. Afterwards, the dependency monitoring process is continued by sliding the time window for the next prediction and the reference *diameter* μ is reset by setting $t_i = t$.

2.3 Model Clustering

One of the most important aspects for successful ensembles is diversity [1]. Typically, this diversity is initially reflected in the distinctive patterns of each base learner's inductive bias derived from the different hypothesis on which each base learner is built to model the input data and its dependence structure. Surprisingly, the enforcement and evaluation of diversity on ensembles for time series data is still a quite unexplored topic-especially for forecasting problems [1]. However, the expected error decomposition for ensemble schemata [1] in general helps to get an intuition about the importance of diversity. More precisely, the expected error can be decomposed into *bias*, *variance* and *covariance*. In DEMSC, we propose a second-stage selection that tries to ensure such diversity through *clustering*. Predictions of K top-base models on the time sequence $X_{W,t}$, are considered as W-dimensional vector features to cluster the models. The final step in the selection consists of selecting one representative model for each cluster. We simply select the closest model to each cluster center. The final selected base-models are integrated using a sliding-window weighted average [1].

3 Experiments

The results are evaluated using the RMSE and ranks of our method compared to s.o.a methods for ensemble learning are computed. A rank of 1 means the model was the best performing on all time series.

DEMSC has advantages over the compared methods except for ADE. The approaches for combining individual forecasters, which are SE, SWE, OGD, FS, EWA and MLPOL, show a big difference in the average rank compared to DEMSC. Common ensemble methods like RF, GBM, OGD and Stacking, compare poorly to all methods specialized for combining forecasters. ADE is competitive to DEMSC and have a higher average rank, but it is comparable to DEMSC in terms of wins and losses. DEMSC is within the range

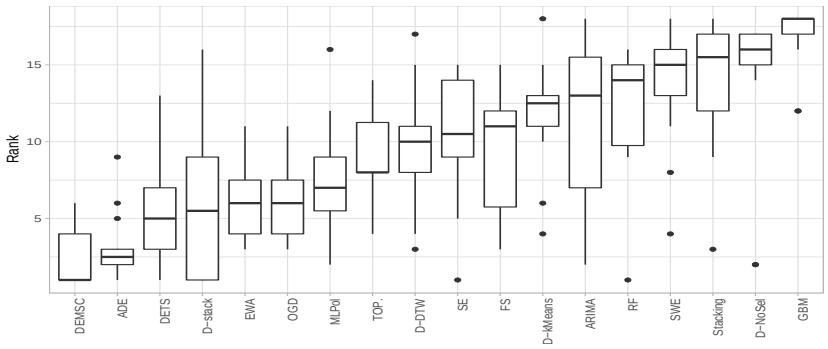


Figure 1: Distribution of the ranks of ensemble methods across the different time series, *D* is used as abbreviation for DEMSC

of the first 4 ranks and has a median of 1 with no clear outliers. More details about the method and the evaluation can be found in [2].

4 Conclusion and Future work

DEMSC: a novel, practically useful dynamic ensemble members selection framework for time series forecasting. DEMSC uses a two-staged selection procedure which on the one hand enhances accuracy by performing informed selection of base learners at test time based on a base models performance drift detection mechanism and diversity on the other hand through an online clustering approach. An exhaustive empirical evaluation, including 16 real-world datasets and multiple comparison algorithms shows the advantages of DEMSC. As a future work, we aim to add a drift-informed procedure for retraining the base-learners.

References

[1] Saadallah, A., Moreira-Matias, L., Sousa, R., Khiari, J., Jenelius, E., Gama, J.: Bright-drift-aware demand predictions for taxi networks. IEEE TKDE (2018)

[2] Saadallah, Amal et al. “A Drift-based Dynamic Ensemble Members Selection using Clustering for Time Series Forecasting.”. ECML PKDD. Springer (2019)

[3] Hoeffding, W.: Probability inequalities for sums of bounded random variables. In: The Collected Works of Wassily Hoeffding, pp. 409–426. Springer (1994)

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B4

Analysis and Communication for Dynamic Traffic Prognosis

Thomas Liebig Michael Schreckenberg
Christian Wietfeld

Performance Analysis of C-V2X Mode 4 Communication

Fabian Eckermann
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
fabian.eckermann@tu-dortmund.de

Communication systems for upcoming Intelligent Transportation Systems (ITSs) are an essential part of automated vehicles and intelligent traffic control. To analyze the performance capabilities of those communication systems simulations are indispensable, especially for scalability studies. In this report, the first available open source Cellular Vehicle-to-Everything (C-V2X) mode 4 simulator based on the discrete-event network simulator ns-3 is presented and the performance of C-V2X mode 4 is analyzed.

1 C-V2X Communication

For V2X communication two competing standards are proposed. The WLAN based IEEE (Institute of Electrical and Electronics Engineers) 802.11p and 3GPPs (3rd Generation Partnership Project) LTE(Long Term Evolution)/5G based C-V2X standard, that this report is focused on. V2X communication is mainly based on periodic Cooperative Awareness Messages (CAMs) and event-based Decentralized Environmental Notification Messages (DENMs). CAM contains information regarding the vehicle status such as position, speed and direction while DENMs are send if for example a car accident is detected by the vehicles sensors. For C-V2X two modes are defined: Mode 3 an infrastructure based mode like common cellular services and mode 4, that is decentralized. To take advantage of the periodic and predictable nature of V2X communication services, sensing based Semi-Persistent Scheduling (SPS) is introduced as distributed scheduling protocol of C-V2X's mode 4.
With sensing-based SPS V-UEs (Vehicular User Equipment) reserve the for the transmission size necessary number of subchannels in the frequency domain for a random

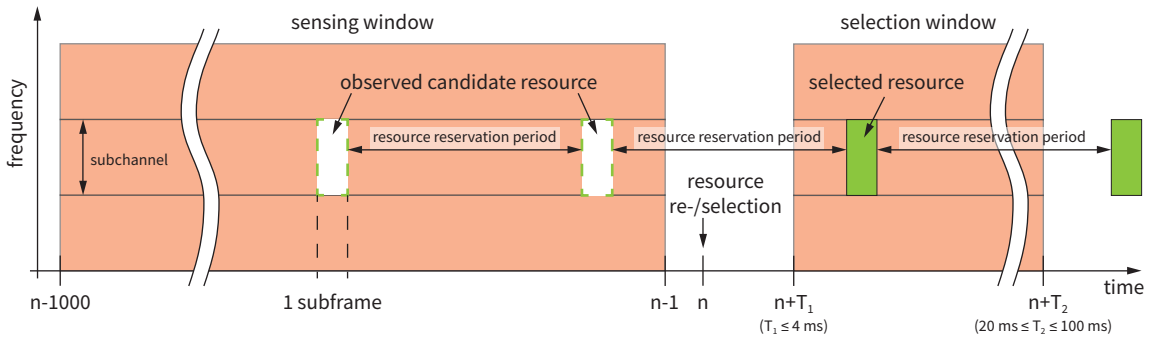


Figure 1: C-V2X distributed channel access: Sensing-based Semi-Persistent Scheduling (SPS).

number of consecutive periodic transmissions. Figure 1 shows a timeline of the resource allocation mechanism. Based on the received data from other vehicles within the past 1000 ms, the sensing window, a resource at time n is selected or re-selected. The received data and the periodicity of the messages is used to create a list of potentially free resources and to pick one of these resources within the selection window. The selection window is defined by the V-UEs configuration (lower bound, T_1) and the maximum Packet Inter-Reception time allowed for the type of transmission (Upper bound, T_2). If no free resources are available the V-UE uses a occupied resource with low receive power, since vehicular communication services are only important for receivers nearby.

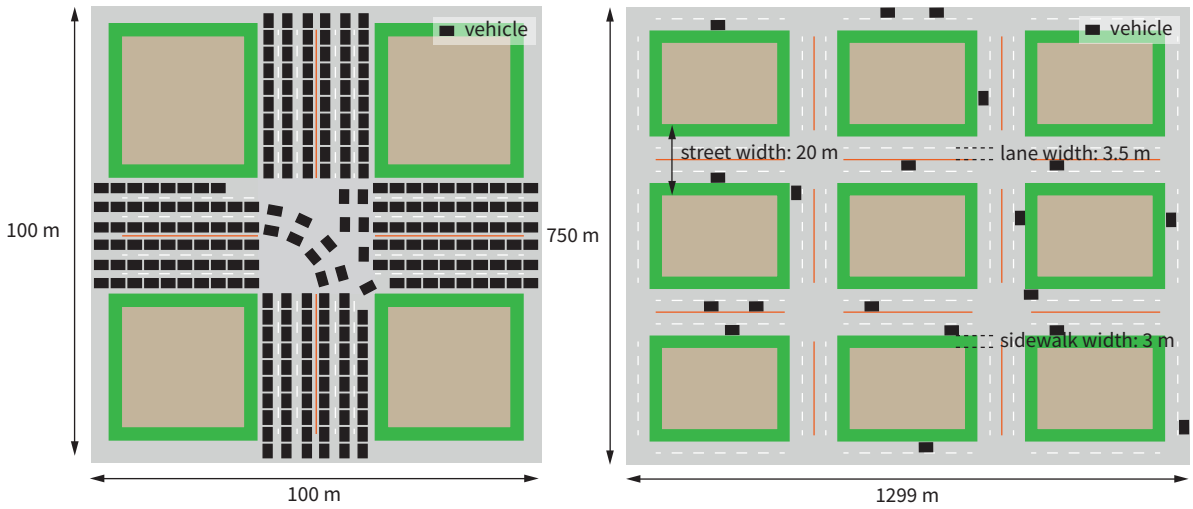
2 Simulative Performance Analysis

For the validation of the simulation results the Packet Reception Ratio (PRR) as specified in [2] is used:

- The PRR is calculated by X/Y, where Y is the number of V-UEs that are located in the baseline distance (a, b) from the transmitter, and X is the number of V-UEs with successful packet reception among Y.

Two simulation scenarios are defined for the performance analysis. A static worst case scenario with high vehicle density and a more realistic urban Manhattan grid scenario as used by 3GPP [2] including a mobility simulation based on the microscopic traffic simulator SUMO [4]. The scenarios are illustrated in Figure 2. The static scenarios dimensions are such that all transmitted packets will be received by all other vehicles. Beside the two scenarios, two cellular bandwidths are simulated, 10 MHz and 20 MHz, where higher bandwidths offer more resources. To align with the 3GPP studies the WINNER+ B1 channel model [5] has been added to the ns-3 simulator. For more details on the simulation parameters see [3].

The results shown in this report slightly differ from the original results presented in [3]



(a) Static. (b) Urban (Manhattan grid).

Figure 2: Simulation scenarios.

as the number of simulation runs and therefore the statistical relevance is higher.

The scalability of vehicular communication is mandatory to guarantee road safety even for crowded inner city scenarios with a very large number of traffic participants.

The results of the scalability analysis is shown by Figure 3. Even in the static worst case scenario with 10 MHz cellular bandwidth the LTE Rel. 14 requirements [1] for up to

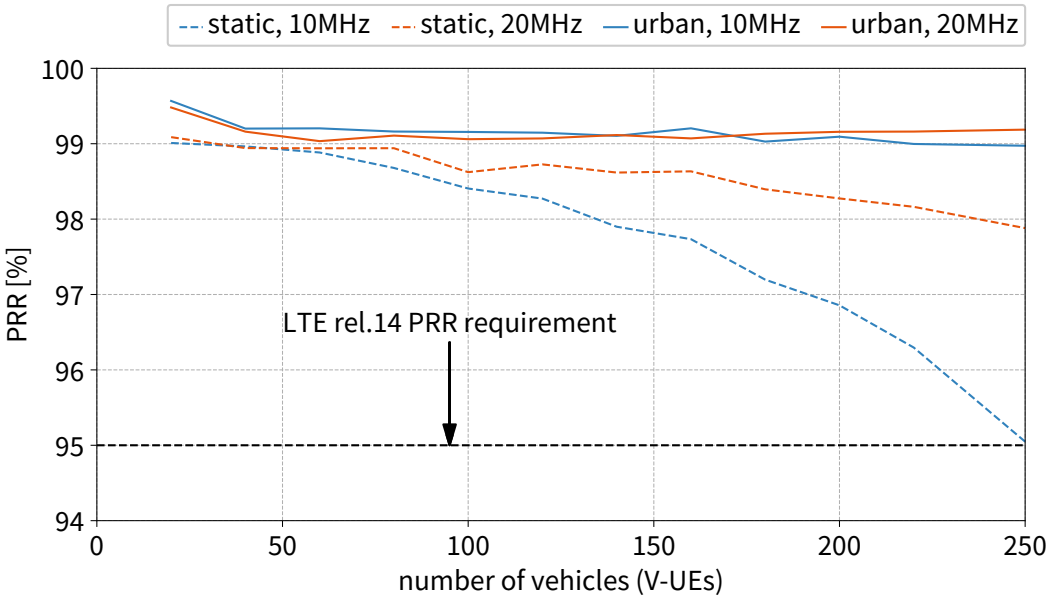


Figure 3: Packet Reception Ratio for an increasing number of V-UEs for the static and urban scenario.

250 vehicles can be fulfilled. For 20 MHz and 250 vehicles a packet reception ratio of ca. 98 % is achieved. In the more realistic urban Manhattan grid scenario, due to the mobility of the vehicles and the bigger playground only several vehicles are within the communication range of each other. Here a packet reception ratio at around 99 % is achieved.

3 Conclusion

In this report an open source C-V2X mode 4 simulator implemented in the ns-3 network simulator is presented and the performance of C-V2X is analyzed. The performance analysis depicts that C-V2X mode 4 is highly scalable even for worst case scenarios. For more realistic, dynamic scenarios including vehicular mobility the scalability is even better and above the 3GPP rel. 14 V2X requirements. In contrast to other V2X simulators the presented simulator is open source and available for open research (`https://github.com/FabianEckermann/ns-3_c-v2x`) to gain not only comparable simulation results between different researchers but also to have other researchers verify the functionality of the simulator and thereby the presented simulation results.

References

[1] 3GPP. Study on LTE support for vehicle-to-everything (V2X) services. TR 22.885 V14.0.0, 3GPP, December 2015.

[2] 3GPP. Study on LTE-based V2X services. TR 36.885 V14.0.0, 3GPP, July 2016.

[3] Fabian Eckermann, Moritz Kahlert, and Christian Wietfeld. Performance analysis of C-V2X mode 4 communication introducing an open-source C-V2X simulator. In *2019 IEEE 90th Vehicular Technology Conference (VTC-Fall)*, Honolulu, Hawaii, USA, September 2019.

[4] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.

[5] J. Meinilä, P. Kyösti, L. Hentilä, T. Jämsä, E. Suikkanen, E. Kunnari, and M. Narandžić. WINNER+ final channel models. Tr, CELTIC, June 2010.

Towards 5G: A Software–Defined End–to–End Mobile Network System for the Evaluation of 5G's Non–Standalone Mode

Karsten Heimann
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
karsten.heimann@tu-dortmund.de

In the course of the deployment of public networks of the fifth generation of mobile communication (5G), a transition process is needed to firstly rely on current long term evolution (LTE) infrastructure while launching novel services and performance improvements enabled by 5G. With our end–to–end evaluation platform containing a software–defined radio millimeter wave (mmWave) transceiver system, such a non–standalone 5G system approach is assessed. The setup utilizes a mmWave radio link at 28 GHz in addition to an LTE network. Initial results prove the system's suitability for ensuring quality of service for prioritized applications.

1 Motivation

Upcoming 5G networks are believed to enhance quality of service (QoS) support especially of applications with contrasting requirements. Thus, the application oriented QoS level needs to differently prioritize several traffic flows as the case may even be on one single device. For example, a vehicle may offer some real–time service parallel to basic telemetry transmissions. However, due to the supposed lengthy transition period from LTE to 5G, LTE's infrastructure is initially supposed to be reused for 5G's non–standalone mode. As depicted in Figure 1, our vision is to design an end–to–end system–of–systems suitable for 5G–like QoS during the transition phase as presented in [1].

For this, previous works form the building blocks:

- The characteristics of mmWave communications focusing on beam alignment and propagation peculiarities by means of phased array antennas are studied in [2].
- In [3], *tinyLTE* as a lean mobile network is presented running on commercial off–the–shelf (COTS) hardware with an open–source LTE implementation.
- From the core network viewpoint, our software–defined networking (SDN) management and orchestration (MANO) controller is introduced in [4] as part of our network slicing platform for managing data traffic flow rules.

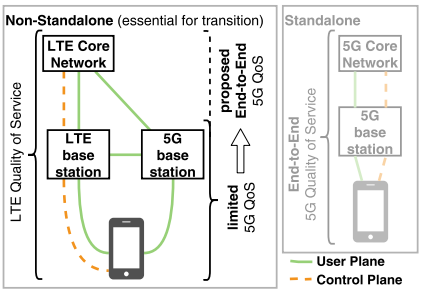


Figure 1: Non–standalone operation mode relying on LTE is supposed to already introduce 5G–like End–to–End QoS.

With [1], aforementioned contributions are merged to allow for the evaluation of these key concepts and their interworking within an end–to–end mobile network development platform.

2 Architecture Concept

In the following, the confluence of the named contributions is expounded based on Figure 2. Due to the space limitation, only key aspects are subsequently highlighted, whereas further details can be found in [1].

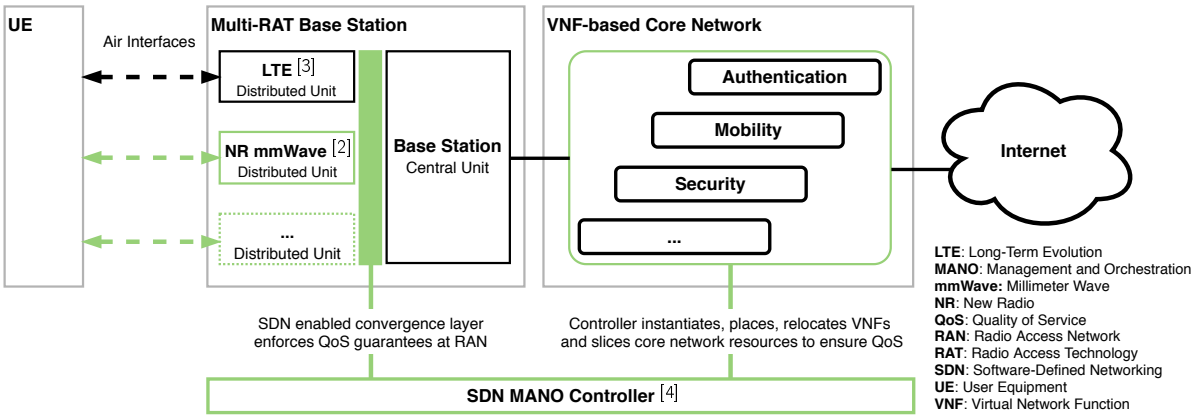


Figure 2: In the proposed architecture concept, the base station logic is split into a central unit, which distributes the traffic flow to the LTE as well as a mmWave link according to the EN–DC approach. Additionally, the SDN MANO controller centrally coordinates flow rules for ensuring end–to–end QoS.

The basis is formed by *tinyLTE* of [3], which hosts the core as well as the radio access network and user equipment (UE) components by using network function virtualization (NFV) concepts. With the supposed dense deployment of base stations in 5G, the centralization of some of their functionalities may lead to a higher resource efficiency. For this reason, the split into a central and distributed units allows for concentrating higher protocol layer routines at the central unit while at the same time open up the lower protocol layers down to the physical transmission for other technologies like a mmWave radio link. Similar to the established dual connectivity (DC) operation mode, the split is implemented at the packet data convergence protocol (PDCP) layer to relax requirements compared to a split approach comparable with carrier aggregation, which would demand a precise synchronization of distributed units at a shared MAC layer.

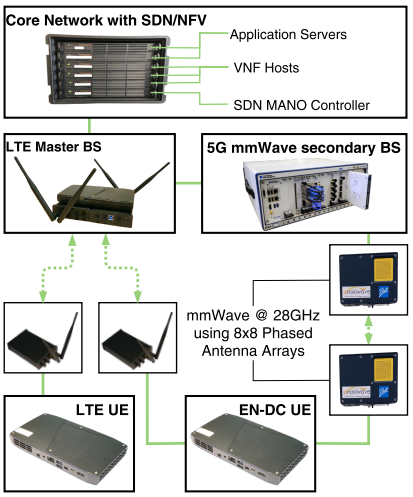


Figure 3: Hardware setup.

5G NR introduces a radio access network (RAN) with improvements for the distinct application scenarios. The novel frequency spectrum in the mmWave domain constitutes more challenging radio propagation conditions on the one hand, but on the other hand, a vast amount of bandwidth is available for an increased data throughput. To overcome the RF peculiarities, directional communications by means of electronically steerable pencil beams are studied in [2]. With antenna arrays, the severe path loss is compensated by their beamforming gains in case of a proper beam alignment.

At the core network, network slicing enables the logical separation of end–to–end networks, which may still share a common physical infrastructure. Depending on QoS constraints, fractions of available physical resources are dynamically assigned to each isolated network slice. The SDN MANO controller of [4] facilitates the dynamical rearrangement of resources as programmable control plane of the entire network.

3 Experimental Setup and Performance Evaluation

In Figure 3, the hardware setup used for the subsequent evaluation is shown. As one advantage of virtualization, COTS hardware could be used for the core network as well as the LTE RAN part. Beside the software–defined radio (SDR) based LTE links, our mmWave transceiver system provides the 5G link by means of SDR hardware and phased antenna arrays using analog beamforming at 28.5 GHz.

The measurements are conducted by emulating data traffic in three phases. After firstly only data of a prioritized control application is transmitted, two additional data links, a

telemetry link at the same device and a conventional application on a separate device, are started in phases two and three, respectively.

The statistical evaluation of the end–to–end delay of the three applications during the consecutive phases is depicted in Figure 4. Since the control application uses the dedicated mmWave link, its delay is offset due to current interface constraints of the development platform. While the telemetry link is impaired by the conventional device competing for the same resources in phase three, the prioritized control application stays on the same level during all phases. This proves the logical separation of this network slice against the other two applications within the end–to–end mobile network architecture.

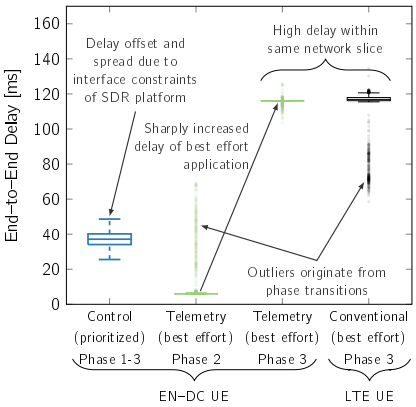


Figure 4: Statistical analysis of the delay distribution.

4 Conclusion

As the deployment of 5G is believed to be a lengthy transition period relying on LTE infrastructure, the presented work illustrates an end–to–end network architecture, where some 5G technology aspects are already incorporated to the predecessor mobile network. The results of lab measurements show, that virtualization, network slicing and the usage of mmWave spectrum can be exploited to build an end–to–end mobile network development platform base on LTE with 5G features like handling different application types on different, isolated network slices.

References

[1] K. Heimann, P. Gorczak, C. Bektas, F. Girke, and C. Wietfeld, "Software–defined end–to–end evaluation platform for quality of service in non–standalone 5G systems," in *Annual IEEE International Systems Conference (SysCon)*, 2019.

[2] K. Heimann, J. Tiemann, S. Boecker, and C. Wietfeld, "On the potential of 5G mmWave pencil beam antennas for UAV communications: An experimental evaluation," in *22nd International ITG Workshop on Smart Antennas (WSA)*, 2018.

[3] F. Eckermann, P. Gorczak, and C. Wietfeld, "tinyLTE: Lightweight, ad hoc deployable cellular network for vehicular communication," in *IEEE Vehicular Technology Conference (VTC-Spring)*, 2018.

[4] F. Kurtz, C. Bektas, N. Dorsch, and C. Wietfeld, "Network slicing for critical communications in shared 5G infrastructures - an empirical evaluation," in *4th IEEE International Conference on Network Softwarization (NetSoft 2018)*, June 2018.

A smartphone-based GNSS real-time tracker system for efficient traffic data acquisition of participants in carnival parades

Petros Polichronidis
Physics of Transport and Traffic
University of Duisburg-Essen
petros.polichronidis@uni-due.de

Carnival parades are little studied so far. Recent studies show interesting and surprising physical properties [1]. In order to find out whether this described phenomenon is universal in nature and whether it occurs in other types of processions as well, movement data from other processions are needed. Especially in the case of carnival processions, this is a very difficult task, as almost all carnival processions take place on same day, and same time. This requires a corresponding amount of effort and resources. Within this contribution the development and application of a smart-phone based real-time tracking system for the special requirements in carnival parades is described.

1 Introduction

In a traffic system, the purpose of movement determines the "natural" speeds of road users, up to the point where the traffic density is so high that the interaction of the participants determines the traffic dynamics and thus the individual travel times. As a transport system that has so far been little described in literature, processions with a festive character (such as the Rhenish carnival processions, rifleman and costume parades, etc.) have interesting and surprising physical characteristics. Above a certain traffic density, a collapse of speeds and corresponding congestion is to be expected. Based

on an analysis of movement data of the Cologne Rose Monday Parade from 2014, a contrary effect is actually observed: Congestion or stop & go traffic, which dominates the kinematics of parades, causes no delay at all, but in fact shortens the traveltime, the later a participant starts to parade [1]. To verify the generic nature of this phenomenon for carnival parades, more measurements of different carnival parades are essential. A major difficulty in the anyway complex measurement of this transport system is the fact that carnival parades take place simultaneously on a single day, namely Rose Monday, in a wide variety of cities. In order to be able to carry out measurements on several parades on this one day and at the same time be virtually unlimited in terms of the number of measuring stations (participants in the parades), the measuring system was developed, the conception and application of which is explained in the following.

2 Identification of Requirements

Starting point of the data analyzed in [1] was the first use of conventional, GPS-capable handheld radio devices in the described parade. These should transmit the local coordinates and the time point of the current position to the central event control center at predefined time intervals. There, the data is displayed in real time on a digital map, which provides information about the current state of movement and location of the parade, and archived for post-processing. In the following years the number of handheld radio devices and at the same time the temporal resolution of the measurement was increased on our advice. In this actually more accurate measurement, a considerable part of the data was lost during transmission to the event control center. The temporal and spatial structure (see fig. 1) of the data loss allows conclusions to be drawn about the original cause. The data loss takes place within a period of time for almost the entire spatial area along the parade route. The transmission interference is proportional to the number of visitors, with large data losses mainly between 12:30 pm and 16:30 pm. Nevertheless, spatial sections along the parade path are shown in fig. 1, which exhibit almost no data loss. On closer examination, it can be seen that these sections are exactly those where the pageant itself is closest to the event control center, and data transmission is subject to considerably less interference. In the area between 3000m and 4000m the pageant follows the most remote sections of the parade pathway. From there to the event control center, the transmitted signals pass through the entire city center of Cologne with its almost 1 million visitors, an area with a high electromagnetic load at that time. The smart-phone-based measuring system is intended to use the internet access of a smart-phone via the mobile phone connection as a transmission channel. Normally (with sufficient supply via transmission and reception masts) no significant transmission error is to be expected with this choice of method of measurement. A positioning of the measuring points at the parade's leader and the last participant, as well as at preferably equal distances within the parade is also no challenge. In addition, the temporal reso-

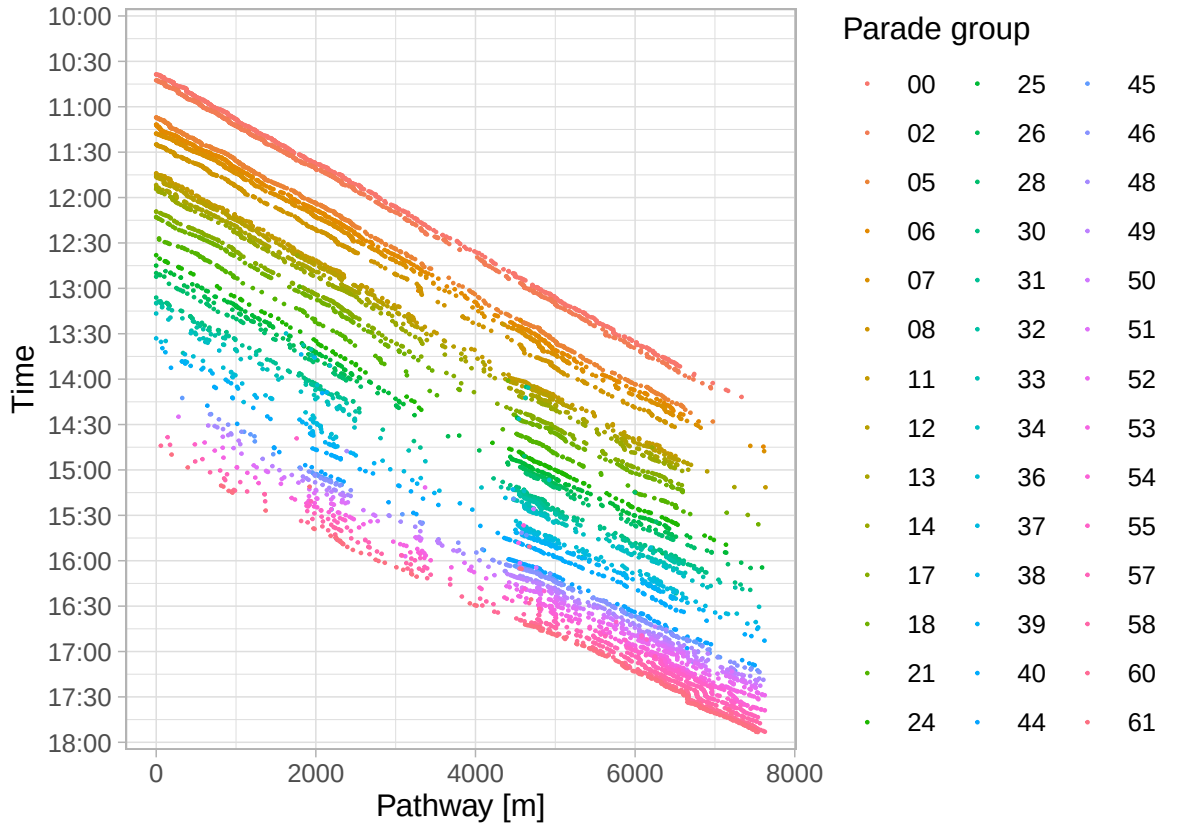
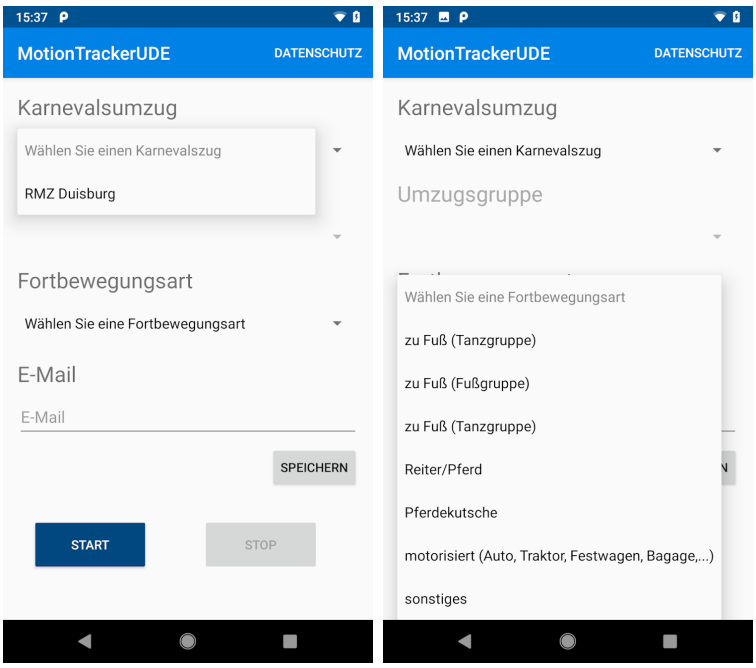


Figure 1: The distance-time diagram of the Cologne Rose Monday Parade 2015. The particularly high number of visitors and the specific design of the radio concept led to a spatially and temporally structured data loss.

lution of the measurement can be freely adjusted on the server side. In the event of a connection failure but continued GPS localization, the data to be transmitted is pre-stored locally on the smart-phone until an Internet connection is re-established in order to send the data stored in the cache. Besides the disturbance of the data processing the presented measurement by means of loggers is limited on the one hand by the number of available devices and on the other hand by the time needed to install the devices and to cover the long distances between the measuring stations. The use of the smart-phones of the parade participants virtually removes restrictions of the measurement regarding these two points. If the application is marketed accordingly, it is possible to measure tens of thousands of participants simultaneously - i.e. a complete carnival parade - which would be almost impossible with loggers. The system has been designed in such a way that, in addition to position data, some other metadata are collected, which the user can select himself (e.g. type of locomotion: on foot/on horseback/on car/etc.).



(a) (b)

Figure 2: A smartphone-based GNSS real-time tracker application system for position data acquisition of participants in carnival parades: MotionTrackerUDE. a) Firstly the user selects the carnival parade and b) secondly chooses metadata information about his participation in the parade (on foot, on horseback, in the car, etc.)

3 Outlook

The measurement system presented here is to be used in the coming carnival season to collect traffic data from several parades (Rose Monday Parade Duisburg, Rose Monday Parade Mainz, Carnival parade in Duisburg Wehofen) simultaneously for the first time. Independent of the applications presented here, this measuring system can be easily adapted to other transport modes. A survey of individual traffic participants in order to record the corresponding individual dynamics within the framework of the systematic measurement inaccuracies is also possible without further ado.

References

[1] Petros Polichronidis, Dominik Wegerle, Alexander Dieper, and Michael Schreckenberg. Traffic dynamics of carnival processions. *EPL (Europhysics Letters)*, 121(6):68003, 2018.

Data-driven Network Simulation for Anticipatory Communications

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

In this report, Data-driven Network Simulation (DDNS) is presented as a novel methodological approach for analyzing the performance of vehicular anticipatory communication networks. In contrast to existing solutions such as Discrete Event Simulation (DES), DDNS allows to derive close to reality representations of concrete real world networks based on machine learning.

1 Data-driven Network Simulation

System-level network simulation based on DES has a emerged as the de-facto standard method for evaluating mobile communication networks. However, due to the highly complex parameter initialization – related to unknown and confidential value definitions – as well as other impact factors such as missing features, simulating models are often not able to mimic the behavior of a novel method in a concrete real world scenario. In order to overcome these issues and to provide a highly accurate and controllable simulation environment, the proposed DDNS [1] method relies on machine learning-based modeling of end-to-end performance indicators such as the end-to-end date rate. Real world traces of network quality indicators are replayed and fed into the machine learning pipeline which leverages the results of previous work focusing on data rate prediction in vehicular networks [2, 3]. A comparison of the system architectures of the two considered simulation approaches is shown in Fig. 1. Since DES relies on modeling actual communicating entities and their respective protocol stacks, the resulting system complexity is much higher than for the proposed DDNS method.

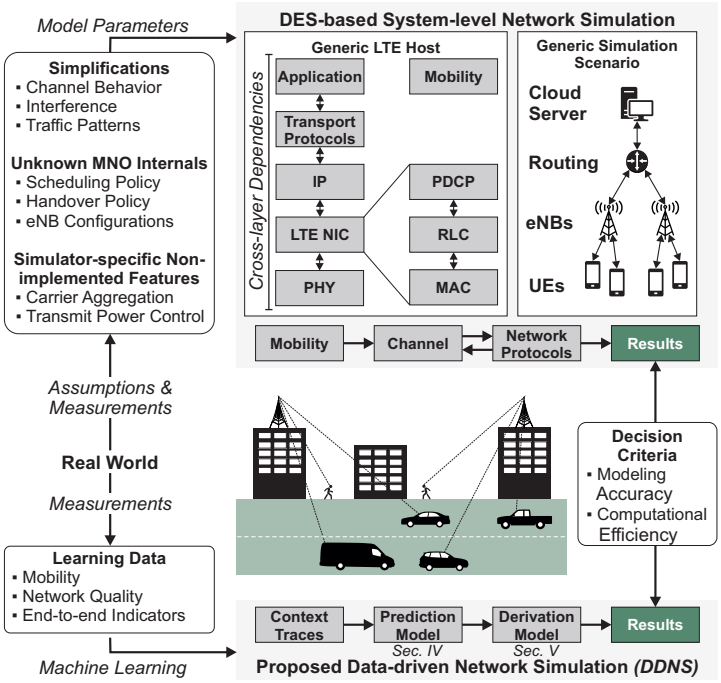


Figure 1: Comparison of classical DES and the proposed DDNS approach.

The actual simulation process relies on a two-stage approach that sequentially executes specialized machine learning models:

- **Prediction**: Based on the replayed context traces, data rate prediction is performed according to [2]. For each time step, a Random Forest model is applied to forecast the currently achievable data rate $\tilde{\mathbf{Y}}_{\mathrm{RF}}$ based on the measured network context features.

- **Derivation Modeling**: Since the predictions are imperfect, a *virtual measurement* is required which pays attention to the derivations between prediction model and real world observations. For this purpose, a Gaussian Process Regression (GPR) model is used to learn the statistical derivations the predictions $\tilde{\mathbf{Y}}_{\mathrm{RF}}$ and measurements $\mathbf{Y}$.

  With the assumption of a gaussian error distribution, each prediction $\tilde{y}_{\mathrm{RF}}$ is converted into a virtual measurement $\tilde{y}_{\mathrm{GPR}}$ as shown in Fig. 2 by sampling from the confidence interval of the GPR model as

$$\tilde{y}_{\mathrm{GPR}}(\tilde{y}_{\mathrm{RF}}) = \mathcal{N}\left(\tilde{\mathbf{Y}}_{\mathrm{GPR}}(\tilde{y}_{\mathrm{RF}}), \boldsymbol{\sigma}^2_{\mathrm{GPR}}(\tilde{y}_{\mathrm{RF}})\right) \tag{1}$$

  However, due to the statistical properties of the sampling process, impossible values (such as negative data rates) might occur. Therefore, a final correction step is

applied which aligns the sample with the value range of the indicator as

$$\hat{y} = \begin{cases} \min(\tilde{\mathbf{Y}}_{\mathrm{RF}}) & \tilde{y}_{\mathrm{GPR}} < \min(\tilde{\mathbf{Y}}_{\mathrm{RF}}) \\ \max(\tilde{\mathbf{Y}}_{\mathrm{RF}}) & \tilde{y}_{\mathrm{GPR}} > \max(\tilde{\mathbf{Y}}_{\mathrm{RF}}) \\ \tilde{y}_{\mathrm{GPR}} & \text{else} \end{cases} \tag{2}$$

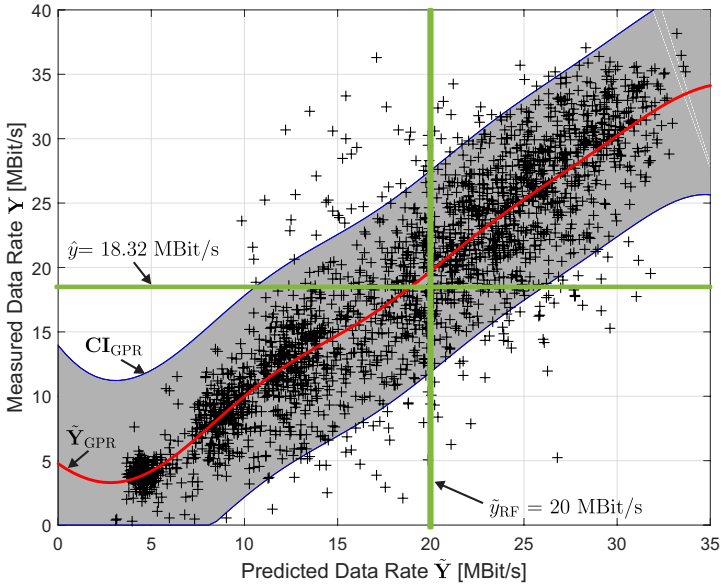


Figure 2: Application of GPR for consideration of the derivations of the prediction model from the real world measurements.

2 Real World Validation

In order to validate the proposed method, DDNS is compared to DES (based on the SimuLTE framework of Objective Modular Network Testbed in C++ (OMNeT++)) as well as to real world measurements. Different variants of the anticipatory data transfer method Channel-aware Transmission (CAT) are analyzed. As a reference, periodic data transmissions are considered. For the DES, the network topology is modeled according to the real world locations of the Long Term Evolution (LTE) base stations. During the simulations, the trajectory of the vehicle is replayed based on the real world location measurements. DDNS furthermore, considers the real world network indicator measurements. For both simulative approaches, the goal is to achieve a good match with the real world measurements. An overview about the behavior of the different data transfer methods and analysis methods is shown in Fig. 3.

It can be seen that the proposed DDNS achieves a significantly more realistic behavior or the considered methods and is able to provide a good representation of the statistical

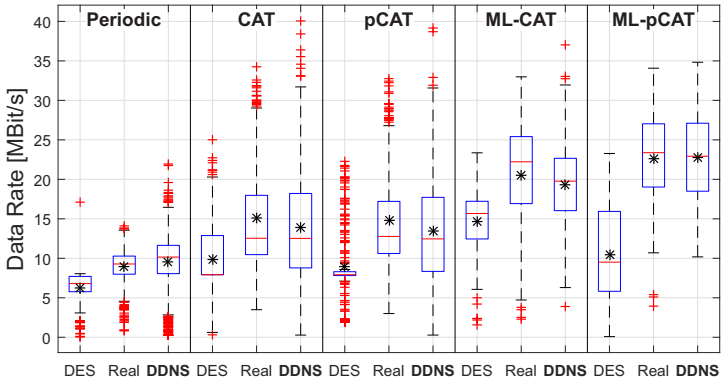


Figure 3: Comparison of modeling accuracies achieved with DES and DDNS for different anticipatory data transfer methods.

properties which are observed in the real world. In contrast, the DES approach is not able to model the observed benefits of the CAT schemes as the channel dynamics and the interplay of the network participants and the network infrastructure show significant derivations from the real world behavior.

3 Conclusion and Further Research

The proof-of-concept results show that DDNS can be utilized as an accurate method for simulating the end-to-end behavior of mobile communication systems. In future work, the developed methodological approach will be utilized for performing the exploration phase of reinforcement learning-based data transfer methods.

References

[1] Benjamin Sliwa and Christian Wietfeld. Data-driven network simulation for performance analysis of anticipatory vehicular communication systems. *IEEE Access*, Nov 2019.

[2] Benjamin Sliwa and Christian Wietfeld. Empirical analysis of client-based network quality prediction in vehicular multi-MNO networks. In *2019 IEEE 90th Vehicular Technology Conference (VTC-Fall)*, Honolulu, Hawaii, USA, Sep 2019.

[3] Benjamin Sliwa and Christian Wietfeld. Towards data-driven simulation of end-to-end network performance indicators. In *2019 IEEE 90th Vehicular Technology Conference (VTC-Fall)*, Honolulu, Hawaii, USA, Sep 2019.

Adaptation of the Lee-model for urban and automated traffic in different models

Tim Vranken
Physik von Transport und Verkehr
Universität Duisburg-Essen
tim.vranken@uni-due.de

This report sums up our recent research on adapting the Lee-model [4] in such a way, that it can be used to simulate urban traffic in front of and passing through a intersection, as well as to simulate automated and heterogeneous traffic. To this end, a new method to design intersections in a cellular automata model was introduced, the time step length of the model was changed and additional rules were added to the Lee model.

1 Introduction

The simulation of urban traffic with the help of cellular automatons results in many problems which restrict the number of lanes before a intersection or the number of directions vehicle can drive to from each lane like in [1], or [6]. The Lee problem furthermore implemented a maximal deceleration possibility which complicates the interaction at intersections since agents have to decide multiple time steps ahead how to behave. For these reasons, changes to the Lee model and a new way so simulate intersections in cellular automaton models have been researched and will be summarised in section 2. Afterwards, section 3 will show how automated vehicle can be simulated in a cellular automaton model, even through their reaction times vastly differ from humans. To this end, the Lee model will again be adapted into a different model, with a time step length below one second and a few changes to the rules.

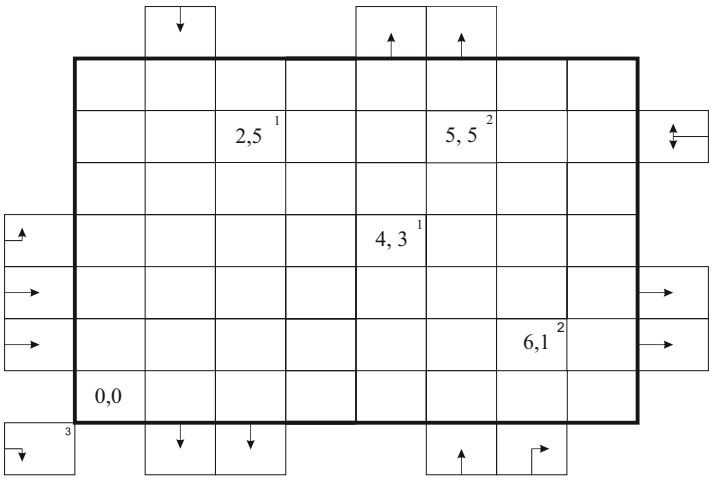


Figure 1: A example intersection for the cellular automaton model. The arrows show the directions a agent is allow to take, based on the lane he passes into the intersection from.

2 Intersection simulation

To simulate intersection traffic in Lee's cellular automaton model [4], first Potmeiers accident free version [5] was implemented. Then a general rule to create intersections out of cells (see fig. 1 for a example intersection) was introduced and it was defined how agent interact with traffic lights.

As fig. 2 shows, the car following time —which descripes how long a agent takes to enter the intersection after its leading vehicle— is steadily above 4 s. Experimental values in [3] show that this value should be around 2 s instead. To adapt these car following times into the Lee-model the acceleration was changed from 1 to:

$$a_n^{t+1} = \begin{cases} k \cdot a, & \text{if } v_{n+1}^t - v_n^t + a_{n+1}^t \tau \geq \Delta v_a \\ a, & \text{else} \end{cases}$$

With this, Agent could travel through intersections in a realistic way. Afterwards, additional rules were implemented to prevent turning vehicle from stopping or reducing crossing flows and lastly rules which prevent deadlocks were included.

3 Automated vehicle

After analysing intersection traffic, the current work is about automated vehicles and heterogeneous traffic in a cellular automaton model. To this end, the Lee-model (more

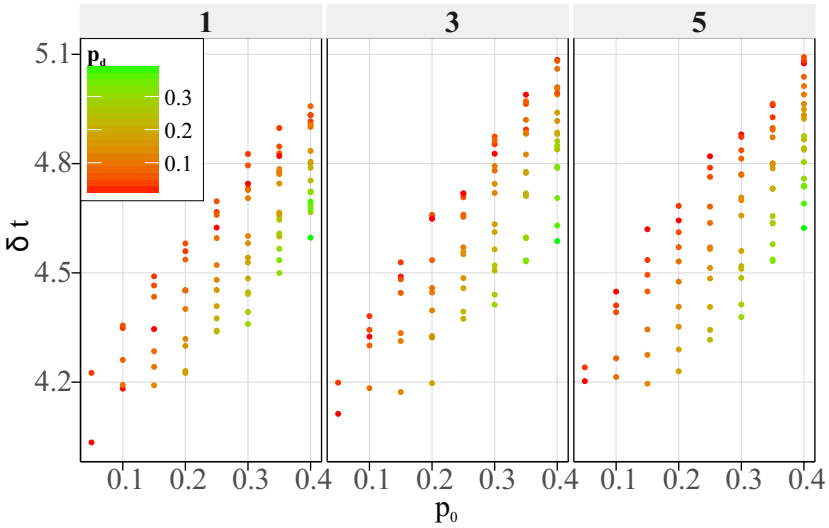


Figure 2: Car following times for different sets of system parameter.

precisely Pottmeiers accident free version of it) are once again adapted. This time, the time step length was changed to 0.1 s, which is the current communication time required for 4G, while still retaining the reaction time of 1 s for human driven vehicles. As one can see in fig. 3, the 0.1 s time step length model is able to reproduce the original results with only a few small differences due to of differences in statistics.

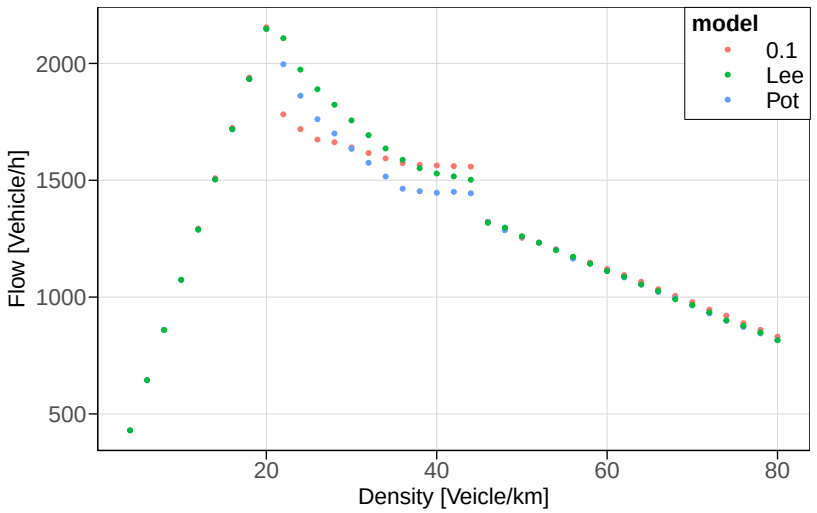


Figure 3: Fundamental diagram for the original version of the Lee-model as well as for Pottmeiers and the 0.1 time step length version.

Automated vehicle were then introduced in the model. They have no dawdle probability and a reduced reaction as well as car following time of 0.5 s. As shown in fig. 4 the

heterogeneous traffic with automated vehicle improves the road capacity in a non linear way up to $\approx 4800\ \frac{\text{vehicle}}{\text{hour}}$ and is a well fit with the theory presented in [2].

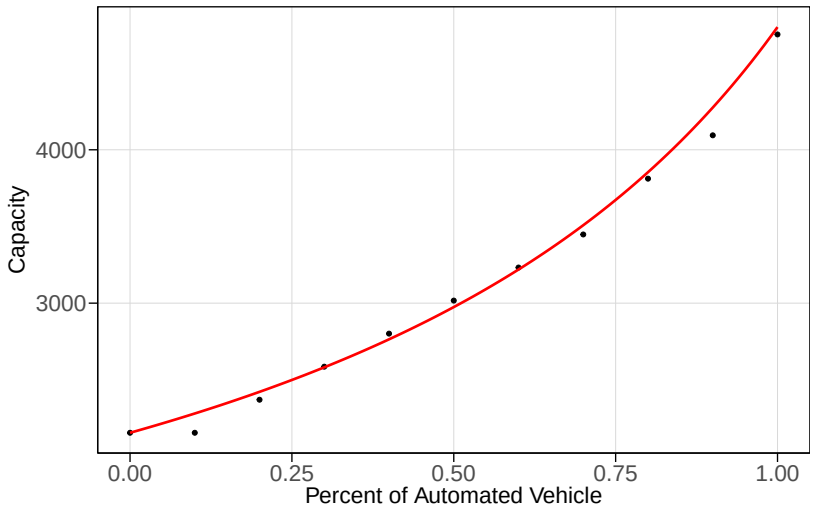


Figure 4: Road capacity in dependence of the percentage of automated vehicle. The red line represents the theoretical obtainable capacities.

References

[1] Esser Jörg and Michael Schreckenberg. Microscopic simulation of urban traffic based on cellular automata. *Journal of Modern Physics C*, 8(5):1025–1036, 1997.

[2] Bernhard Friedrich. The effect of autonomous vehicles on traffic. In Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, editors, *Autonomous driving*, pages 317–334. Springer Open, Berlin, 2016.

[3] Bruce D. Greenshields, Donald Schapiro, and Ericksen Elroy L. Traffic performace at urban street intersections.

[4] Hyun Keun Lee, Robert Barlovic, Michael Schreckenberg, and Doochul Kim. Mechanical restriction versus human overreaction triggering congested traffic states. *Phys. Rev. Lett.*, 92:238702, Jun 2004.

[5] Andreas Pottmeier, Christian Thiemann, Andreas Schadschneider, and Michael Schreckenberg. Mechanical restriction versus human overreaction: Accident avoidance and two-lane traffic simulations. In Springer, editor, *Traffic And Granular Flow*, pages 503–508. 2005.

[6] O. K. Tonguz, W. Viriyasitavat, and Fan Bai. Modeling urban traffic: A cellular automata approach. *IEEE Communications Magazine*, 47(5):142–150, 2009.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C1

Feature selection in high dimensional data for risk prognosis in oncology

Sven Rahmann Alexander Schramm

Strategies for handling errors in genomic data

Till Hartmann
Lehrstuhl für Genominformatik
Universität Duisburg-Essen
till.hartmann@uni-due.de

Genomic data obtained from experiments is never free of errors. Either errors in the data have to be corrected or existing methods have to be adapted to take certain error profiles into account, such that results can be stated with respect to these profiles.

1 Error correction and barcode design in MERFISH data

Multiplexed Error Robust Fluorescence In Situ Hybridisation (MERFISH) [5] is an approach for single-cell transcriptomics with spatial information, based on the repeated application of single molecule FISH (smFISH). For each species of RNA (e.g. mRNA transcript under investigation), a barcode sequence $\in \{0,1\}^n$ and corresponding fluorescent probes are designed such that in the i-th round of smFISH imaging a fluorescent signal is emitted iff the barcode's i-th entry is 1 (and no signal if the entry is 0). By observing the bit pattern at each imaged pixel, it can be inferred which transcript is present at the corresponding location. Due to background noise and biological as well as chemical issues [1, 5], real-world data does not exhibit a one-to-one correspondence between designed barcode and observed bit pattern: Some true signals may not be recorded (false negatives), and similarly some spurious signals may be observed (false positives). Experience shows that false negatives happen more frequently than false positives. Because these errors are hard to avoid technologically, a decoding strategy is part of MERFISH protocols, which estimates true RNA transcript abundances from the observed barcode frequencies. Due to observational error rates and design decisions, the number of transcripts that can be targeted simultaneously using this technology has been limited so

far. We designed an iterative algorithm that alternates between estimating error rates and decoding barcodes (i.e. correcting erroneous barcodes), which allows to increase the number of targets by one to two orders of magnitude while achieving reasonable transcript quantification accuracy and lifting restrictions on protocol design.
For each pair $(b_j, b_k) \in \{0,1\}^n \times \{0,1\}^n$ of barcodes, we can model a probability $a_{jk} = \mathbb{P}(b_j \leftarrow b_k)$ to transition from b_k to b_j using $2n$ positional error variables $\varepsilon_i^{\{0\leftarrow 1, 1\leftarrow 0\}}$. Since each b_i has an associated observed frequency $y_i \in [0,1]$, we can model the *true* frequencies x_i as the solution of

$$\mathbb{E}[y_j] = \sum_k a_{jk}(\boldsymbol{\varepsilon}) \cdot x_k\,, \quad \text{or} \quad \mathbb{E}[y] = A(\boldsymbol{\varepsilon}) \cdot x \text{ in matrix-vector notation.}$$

In MERFISH protocol MHD4.2, $n = 16$ bits are used for the barcodes; hence, error transition matrix A has shape $2^{16} \times 2^{16}$ and – assuming double-precision entries – would require about 32GiB of RAM. Instead of constructing the full matrix and inverting that[1], we assume that transition probabilities below a certain threshold to be zero instead, hence increasing sparsity, allowing us to employ successive over-relaxation for solving the system instead. At this point the error probabilities $\boldsymbol{\varepsilon} = (\varepsilon_i^{\{0\leftarrow 1, 1\leftarrow 0\}})_{i=0,\dots,n}$ are not known yet. We use gradient descent to solve

$$\min_{\boldsymbol{\varepsilon}} \frac{1}{2n^2} \|y - A(\boldsymbol{\varepsilon})\, x\|_2^2\,.$$

However, x needs to be known in order to solve for $\boldsymbol{\varepsilon}$. So we initially either need to estimate x or $\boldsymbol{\varepsilon}$ before we can start alternating between solving for $\boldsymbol{\varepsilon}$ and x.

2 Fusion detection in Oxford nanopore sequencing data

In contrast to Illumina sequencers, which produce reads of length 150bp or 300bp, Oxford nanopore sequencing reads can span tens to hundreds of kilobases [3], thus allowing analysis of larger structural variations, such as large deletions or gene fusions. The latter can for example be the result of interstitial deletions, causing two genes that were far apart to "fuse" and be recognized as a new gene comprised of a prefix of the first gene and a suffix of the second one. Fusion genes are often associated with disease; it is therefore of clinical relevance to be able to identify fusion genes in patients' samples.

However, the way nanopore sequencing works makes reads error-prone to so called *homopolymer errors*: during base-calling, the length of runs of identical bases in a sequence are misjudged to be either larger or smaller than the actual run-lengths. To cope with these errors, we propose to collapse any homopolymer run in both nanopore reads and target sequence. While this increases ambiguity for possible alignments of the reads (or

[1] if at all invertible

kmers thereof) to the target sequence, it eliminates the need to address homopolymer errors specifically.

Building a kmer index of the collapsed target sequence allows quick lookup of the positions of each kmer of a collapsed read in the target. Then, finding possible candidates for gene fusions involves identifying runs of kmers in a read, such that

1. the read is made of at least two consecutive runs of kmers,
2. the gap between these runs is large enough to be considered relevant (i.e. can be assumed to be drawn from the distribution of distances between neighbouring genes),
3. the runs stem from different genes.

Item 1 is not as trivial to satisfy as it may seem, since even though homopolymer errors have been ignored, nanopore reads may still contain substitutions and insertions/deletions, resulting in read kmers which either do not have a corresponding target kmer at all or way too many target kmers. However, a single substitution in a window of length k in a read sequence will lead to exactly (as long as it does not occur within a homopolymer run) k kmers having either different target positions than the preceding and following kmers or no target positions at all, i.e. results in a pattern that can easily be recognized and corrected.

The specifics of this approach still need to be validated, and fusion candidates obtained using this approach need to be verified experimentally.

3 Extending PairHMMs to consider homopolymer errors

Pair Hidden Markov Models are a way of comparing two sequences, enabling a probabilistic assessment of the relatedness of these two sequences (with respect to the underlying model). The standard PairHMM [2] for sequence alignment has 3 relevant states[2], *Match*, Gap_x and Gap_y. However, as seen above, nanopore reads have a higher rate of homopolymer errors, which, as long as these are not close to the probabilities to open and extend gaps, are not part of the standard model. We are working on extending the standard model to be able to give proper probabilities for the relatedness of two sequences obtained by nanopore sequencing, in order to improve postprocessing variant callers such as *varlociraptor* [4]. Error profiles including substitutions, insertions/deletions as well as certain motif dependent errors for Illumina platforms and Oxford Nanopore MinION have been studied extensively [3, 6]. These studies do not adress homopolymer errors specifically. So to obtain transition probabilities for use in the extended model, we use Oxford nanopore data publicly available for the NA12878 sample.

[2] omitting start and end states

References

[1] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, Apr 2015.

[2] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[3] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):239, 2016.

[4] Johannes Köster, Louis J Dijkstra, Tobias Marschall, and Alexander Schönhuth. Enhancing sensitivity and controlling false discovery rate in somatic indel discovery. *BioRxiv*, page 741256, 2019.

[5] J. R. Moffitt, J. Hao, G. Wang, K. H. Chen, H. P. Babcock, and X. Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U.S.A.*, 113(39):11046–11051, 09 2016.

[6] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):125, 2016.

Detection of Lung Cancer Fusion Genes by Nanopore Sequencing

Alicia Isabell Tüns
Molekulare Onkologie
Innere Klinik/Tumorforschung
Universitätsklinikum Essen
Alicia.Tuens@uk-essen.de

Lung cancer (LC) is the leading cause of cancer-related death and five-year survival rates are below 20% due to advanced stage at diagnosis, intratumoral heterogeneity and therapy resistance. Early detection of tumor progression and therapy resistance is therefore an unmet medical need. Fusion genes and transcripts such as the EML4-ALK variant that occur in 3-5% of all LCs are good candidates for monitoring tumor cells. Nanopore sequencing is an emerging technology potentially allowing detection of tumor-specific alterations including fusion genes with high sensitivity and specificity. In the current project we establish a workflow for detecting fusion transcripts in the EML4-ALK variant 3a positive lung cancer cell line NCI-H2228.

1 Introduction

Lung cancer (LC) is the most common cancer type and the leading cause of cancer-related death. LC can be categorized into small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) that account for 15% and 85%, respectively [1]. Molecular profiling of advanced NSCLC revealed various genetic alterations such as anaplastic lymphoma receptor tyrosine kinase (ALK) rearrangements, resulting in the development of specific targeted therapies. Despite advances in detection and treatment, LC patients still have a poor long-term survival. Therefore, a better understanding of the molecular tumor evolution would improve clinical outcome. Conventional whole genome and transcriptome sequencing technologies are long-established methods for uncovering the

genetic diversity of many cancers. In contrast, nanopore sequencing is a relatively new method that allows real-time sequencing of long reads by measuring ion current changes when a DNA/RNA strand passes through a nanopore (Figure 2)[3]. The extended read length enables simpler detection of structural variants like deletions, duplications, insertions and fusion genes such as EML4-ALK.

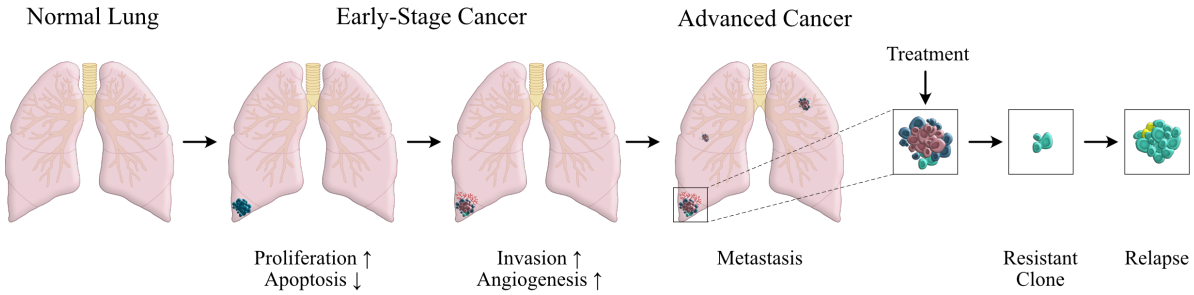


Figure 1: **Lung Cancer Evolution** Environmental or genetic factors can influence carcinogenesis of normal epithelial cells by aberrant activation of tumor promoting pathways (proliferation, invasion, angiogenesis) or downregulation of tumor suppressing pathways (apoptosis). During the course of the disease, most tumors become more heterogeneous, leading to diverse subpopulations. Even after LC patients are treated, many suffer from relapses due to resistant clones [3,4].

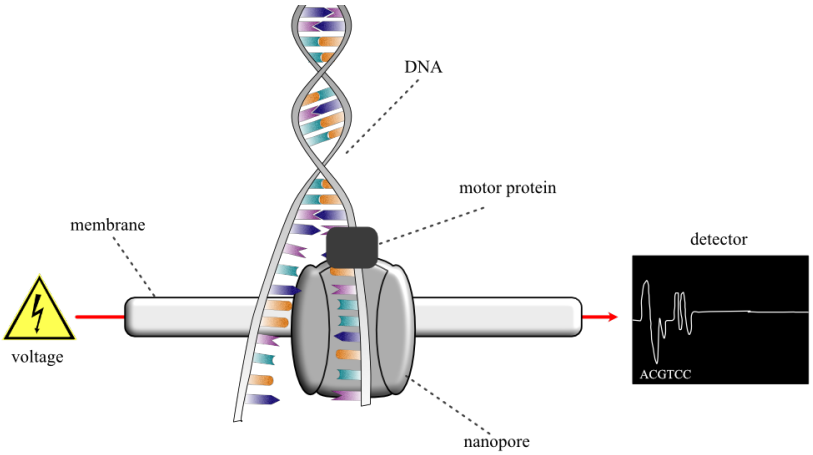


Figure 2: **Nanopore Sequencing Principle** When a DNA strand approaches the nanopore, a motor protein unzips the DNA and pulls the single-strand through the nanopore. The nucleotides entering the nanopore cause characteristic changes in the ion current that can be measured and characterized.

2 Detecting ALK fusions in an EML4-ALK expressing cell line using the nanopore sequencer MinION

Preliminary experiments revealed positive EML4-ALK variant 3a gene- and protein expression in the LC cell line NCI-H2228 (Figure 3a,b). Based on these results, a pilot sequencing run was performed. Using the nanopore sequencer MinION and the direct RNA kit (both *Oxford Nanopore Technologies*), the transcriptome of NCI-H2228 was determined. Reads that passed quality control were aligned against the human Ensembl cDNA database (GRCh38) and the EML4-ALK transcript variant 3a (AB374361.1) using the alignment algorithm Sublong (Rsubread). The mapping of 898,284 reads against the cDNA database led to 727,838 positively aligned reads (81%) including eleven reads and two reads for EML4 and ALK, respectively. The mapping against the EML4-ALK variant 3a reference revealed two aligned reads however only one is covering the fusion breakpoint (Figure 3c).

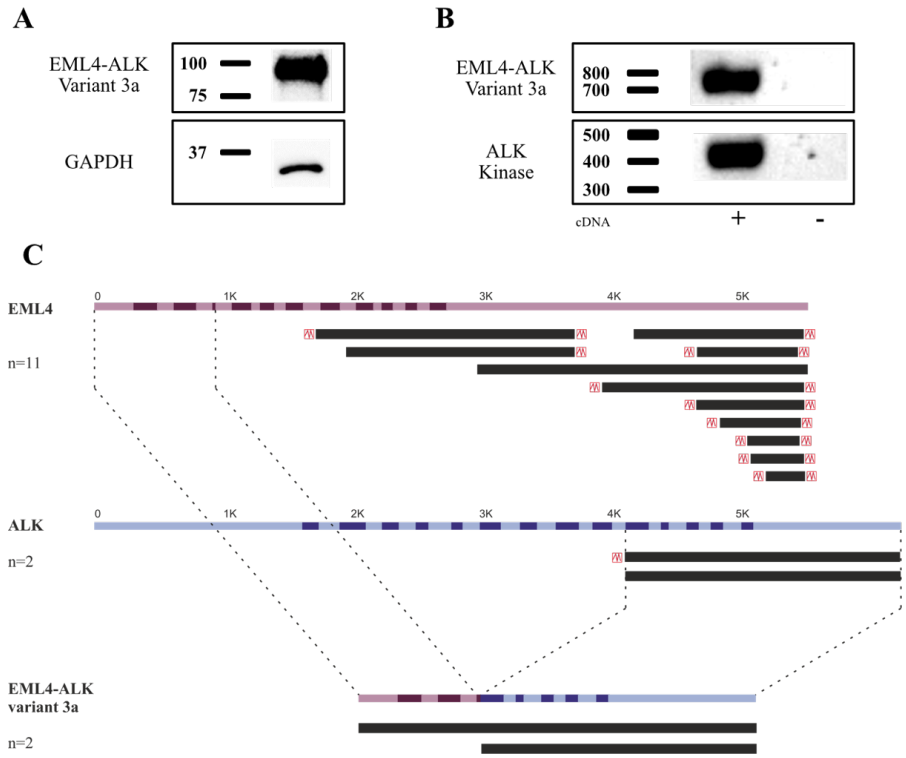


Figure 3: **Detection of EML4-AK Variant 3a** Gene and protein expression in NCI-H2228 cells were verified by PCR (A) and Western Blot (B) analyses, respectively. The alignment of Nanopore reads revealed 11 reads against EML4 and 2 reads against ALK. Both ALK reads mapped to EML4-ALK variant 3a, however only one covered the breakpoint, thereby confirming the fusion transcript (C).

3 Conclusion and Outlook

Although Western Blot and PCR analyses revealed positive EML4-ALK variant 3a expression in NCI-H2228, RNA sequencing identified only one read covering the predicted breakpoint. This in turn suggests that also EML4 and ALK wild-type alleles are expressed in these cells. The paucity of EML4-ALK reads could be explained by the excessive background of ribosomal protein coding- and housekeeping genes that are frequently observed in nanopore RNA sequencing. These technical issues are currently being addressed by e.g. depletion of ribosomal genes and enrichment of coding mRNA.

References

[1] HERBST, R.; HEYMACH, John V.; LIPPMAN, S. Molecular origins of cancer. N Engl J Med, 2008, 359. Jg., Nr. 13, S. 1367-80.

[2] ZAPPA, Cecilia; MOUSA, Shaker A. Non-small cell lung cancer: current treatment and future advances. Translational lung cancer research, 2016, 5. Jg., Nr. 3, S. 288.

[3] JAIN, Miten, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology, 2016, 17. Jg., Nr. 1, S. 239.

[4] IBIAYI, Dagogo-Jack, and SHAW, Alice T. Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology 15.2 (2018): 81.

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

Active Class Selection for Simulation Control

Mirko Bunse
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
mirko.bunse@tu-dortmund.de

Active class selection can make resource-aware decisions on what training data to simulate. In the last year, we have proposed a distinction between this learning scheme and a different one also related to learning from simulations. Moreover, we have surveyed the existing publications on active class selection to develop expectations for future work on the topic, particularly in the context of controllable simulations.

1 Introduction

The high computational cost associated with simulation requires smart decisions on which scenarios to simulate. Here, we consider simulations which produce training data for classification problems. Moreover, we assume that the production of new data is conditioned on the class, i.e. one can decide for a class for which an observation is produced. The simulations employed in astro-particle physics depict a prime use case for this learning scheme; they produce data for arbitrarily chosen particle classes, which is then used to train classification systems. The high amount of energy required to run these simulations demands for a trade-off between classification accuracy and simulation cost.

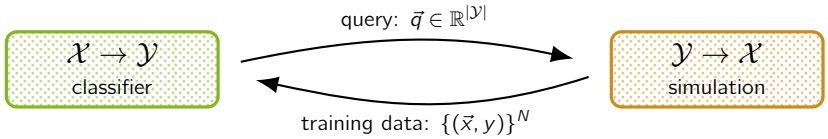


Figure 1: N additional samples are produced from desired label proportions $\vec{q} \in \mathbb{R}^{|\mathcal{Y}|}$.

2 Forward and Backward Learning

There is a recent attention towards the employment of simulated data in machine learning, an integration sometimes termed *simulation data mining* [5, 6]. The benefit of this paradigm is that less or even no data is required from the actual system, e.g. if data acquisition is costly or if this system is not yet deployed.

We argue that every simulation is based on some kind of generative model which evolves the system's state over time. Specifically, we observe that simulations produce an outcome $\vec{s} \in \mathcal{S}$ from an initial state $\vec{s}_0 \in \mathcal{S}_0$. If we intend to make resource-aware decisions on what data to simulate for training, we must clarify in which causal direction the learning task is defined. Most commonly, the learning task is defined in a "*forward*" fashion, i.e. the goal is to predict the outcome $\vec{s}$ of an initial state $\vec{s}_0$. In other use cases like astro-particle physics, the learning task is "*backward*", i.e. the goal is to predict $\vec{s}_0$ from $\vec{s}$. This causal direction has an immediate implication on how we can reduce the number of simulation runs. While the forward learning scheme can be optimized by active learning (selecting objects to label), the backward scheme can be optimized by active class selection (selecting labels to generate objects for) [2].

3 Active Class Selection

The promise of active class selection is that the proportions of classes can be optimized in newly acquired data. We surveyed the results on existing methods for this task [3, 4] and found that random sampling, the baseline method, performs highly competitive with respect to resource efficiency [1]. Namely, it is amongst the best strategies in five out of eight data sets. To date, it is not yet clear if this observation stems from the difficulty of active class selection itself or from the methods proposed.

Another noteworthy observation is that all data sets used in active class selection consist of at least three classes of varying difficulty [1]. In fact, when selecting class labels to generate data for, we can only expect to improve performance for pairs of classes in between which the decision boundary is not yet based on a sufficient number of samples. Moreover, we cannot reasonably expect the existing methods to optimize data acquisition in binary classification problems or in problems with homogeneous class difficulty.

4 Recent Progress and Future Work

We have recently considered the proportions of classes in the training data as a meta-parameter of the learning method. From this perspective, meta-parameter tuning becomes a vehicle for active class selection. Where the existing heuristics for active class

selection [4] evaluate the performance of a classifier in the last acquisition round, we designed our approach to learn from all past rounds and to extrapolate into the next acquisition. What we found, however, is that the level of noise in this setting is too high for a straight-forward application of meta-parameter tuning techniques. Namely, one fixed three-class configuration (N_1, N_2, N_3) consisting of the number of samples in each class can have a wide range of performance values. The quality of a configuration strongly depends on whether the samples are close to a decision boundary or not. Therefore, we cannot learn from label configurations (N_1, N_2, N_3) alone. We must take the feature space into account, as indicated by Fig. 2.

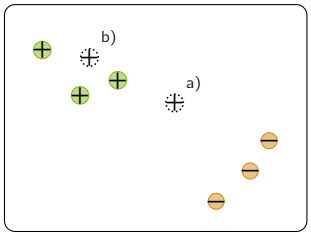


Figure 2: Generating one additional observation of the positive class might be a) informative or b) not informative with respect to the previous decision boundary.

In the future, we also intend to study relaxations of the "pure" active class selection problem. Indeed, the simulations used in astro-particle physics (and also other data generators) are not only controlled by class proportions, but also from additional parameters. Optimizing simulations only with respect to class proportions therefore means to limit the actual task artificially—and maybe even detrimentally.

5 Conclusion

The idea of actively selecting classes to generate data for promises that simulations can be optimized for resource efficiency. We are constantly trying to push the limits of what can be achieved with this learning scheme, also gaining a better understanding of why the task appears to be so difficult. As one particular difficulty, we have identified the randomness of the data-generating process.

Since simulated data is the basis for every analysis in astro-particle physics, optimizing simulations for resource efficiency can have an impact on the resource footprint of all work packages of the C3 project. We are hosting the latest information on our efforts at `https://sfb876.tu-dortmund.de/simulation-data-mining`.

References

[1] Mirko Bunse and Katharina Morik. What can we expect from active class selection? In *Proc. of the Conf. on "Lernen, Wissen, Daten, Analysen"*, volume 2454 of *CEUR Workshop Proceedings*, pages 79–83, 2019.

[2] Mirko Bunse, Amal Saadallah, and Katharina Morik. Towards active simulation data mining. In *Proc. of the 3rd Int. Workshop and Tutorial on Interactive Adaptive Learning at ECML-PKDD 2019*, volume 2444 of *CEUR Workshop Proceedings*, pages 104–107, 2019.

[3] Daniel Kottke, Georg Krempl, Marianne Stecklina, Cornelius Styp von Rekowski, Tim Sabsch, Tuan Pham Minh, Matthias Deliano, Myra Spiliopoulou, and Bernhard Sick. Probabilistic active learning for active class selection. In *Proc. of the NIPS Workshop on the Future of Interactive Learning Machines*, 2016.

[4] Rachel Lomasky, Carla E. Brodley, M. Aernecke, D. Walt, and Mark A. Friedl. Active class selection. In *Proc. of the ECML 2007*, volume 4701 of *LNCS*, pages 640–647. Springer, 2007.

[5] Yanli Shao, Yusheng Liu, Xiaoping Ye, and Shuting Zhang. A machine learning based global simulation data mining approach for efficient design changes. *Advances in Engineering Software*, 124:22–41, 2018.

[6] Holger Trittenbach, Martin Gauch, Klemens Böhm, and Katrin Schulz. Towards simulation-data science – a case study on material failures. In *Proc. of the 5th Int. Conf. on Data Science and Advanced Analytics (DSAA)*, pages 450–459. IEEE, 2018.

Energy Reconstruction of Cosmic Rays with a Machine Learning Approach using IACT Data

Alicia Fattorini
Experimentelle Physik 5
Technische Universität Dortmund
alicia.fattorini@tu-dortmund.de

The MAGIC telescopes, two Air Imaging Cherenkov Telescopes at La Palma are sensitive in the energy regime from the GeV to the TeV range and are extremely capable instruments for studying gamma-ray sources in the Universe. With a ratio of photons to hadrons up to 1:10000 for the detected showers, the background offers large statistics. Energy reconstruction for charged particles can yield interesting results both for background rejection and scientific studies of the Cosmic Ray spectrum. Two methods for the energy regression are presented and compared. While the first method uses look-up tables to estimate the particles energy and performs the standard analysis method of the MAGIC collaboration, the second approach is based on a random forest algorithm. Using simulated proton showers, we perform the energy estimation of energetic charged particles.

The MAGIC Telescopes

MAGIC [1] is a system of two Imaging Air Cherenkov Telescopes at the Roque de los Muchachos, a mountain of 2200 m height at La Palma, Canary Islands. Originally it was built for the detection and investigation of gamma-ray sources. Besides high energetic photons, cosmic rays also cause showers entering the atmosphere. The ratio of photons to hadrons is up to 1:10000 for the showers detected in the MAGIC's energy detection range. Hence the measurements of IACTs offer besides the analysis of gamma-ray sources a great opportunity to investigate the cosmic ray spectrum in the certain energy regime.

The Cosmic Ray Spectrum

Unlike photons or neutrinos, hadrons pose the challenge that their origin is difficult to reconstruct because they can be deflected by intergalactic magnetic fields on the way from their source to earth. This leads to the fact, that a diffuse cosmic ray flux is measured on earth. Although the cosmic ray spectrum is now being investigated in many experiments, the energy range of MAGIC is relatively unexplored. Therefore it is reasonable to create hadronic simulations and build models to reconstruct the energy of the cosmic rays detected by MAGIC.

Energy Reconstruction with Look-up Tables

The currently used tool for the reconstruction of the primary gamma particle is the so-called lookup table. It is quite evident that this method can also be applied to proton energy reconstruction and check the performance. The lookup table makes use of the fact, that the energy of a primary gamma particle is approximately proportional to the number of Cherenkov photons in the shower and therefore also to the amount of photo-electrons measured by the photo multiplier tubes in the camera. The amount of detected shower photo-electrons is called 'size' and one of the most important parameter for the energy reconstruction.
The principle of a lookup table is straight forward: the simulated data is divided into bins for important parameters like the size. The average of the simulated energy in every bin will be the prediction for the data to be analyzed. In the standard analysis of MAGIC, the simulations are binned in two dimensions, the parameter 'sqrt(log(size))' and the ratio between 'impact' and 'cherenkov radius'. The entries in the table are the 'energy'*'cherenkov density'/'size' for every simulated event. The energy is reconstructed separately for the telescopes and then calculated by the average of both estimations.

Energy Reconstruction with a Random Forest

The energy of the primary particle can also be estimated by a random forest. For this work, two applications of a random forest are tested: the implemented tool of the MAGIC standard analysis software, which is tested for the energy estimation of gamma rays, in the following called 'random forest' and the framework named 'aict-tools' [3], especially written for the analysis of CTA and FACT data.
Both random forests are trained on the same data set with the same features and hyper parameter tuning.

Results and Outlook

The available simulations contain protons and helium particles. The lookup tables and random forests were trained on proton simulations. The resulting migration matrices are

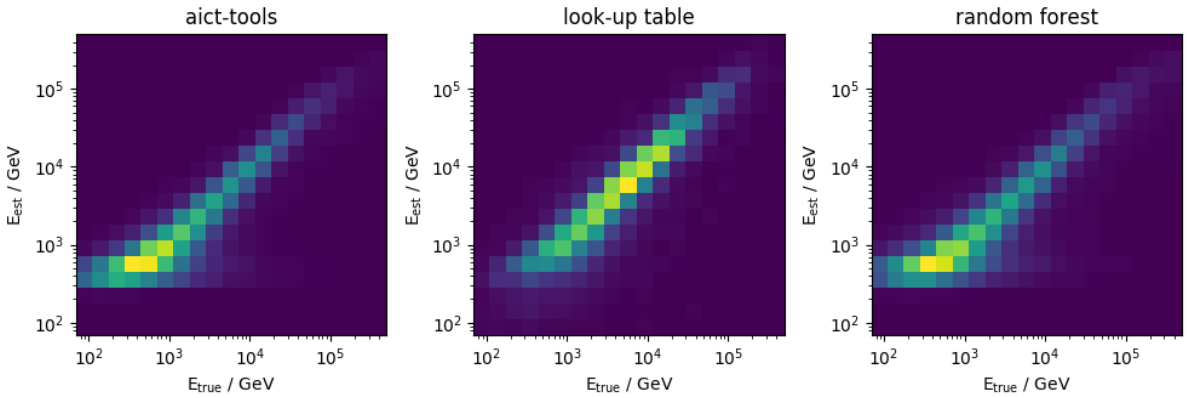


Figure 1: Migration matrices for the energy estimation with three different methods. Aict-tools and the so-called random forest are based on a machine learning approch while the lookup table is constructed with the binning of simulations.

shown in figure 1. The bias and the resolution of the energy estimation with the different methods are shown in figure 2. According to the graphs, the energy reconstruction

proportion

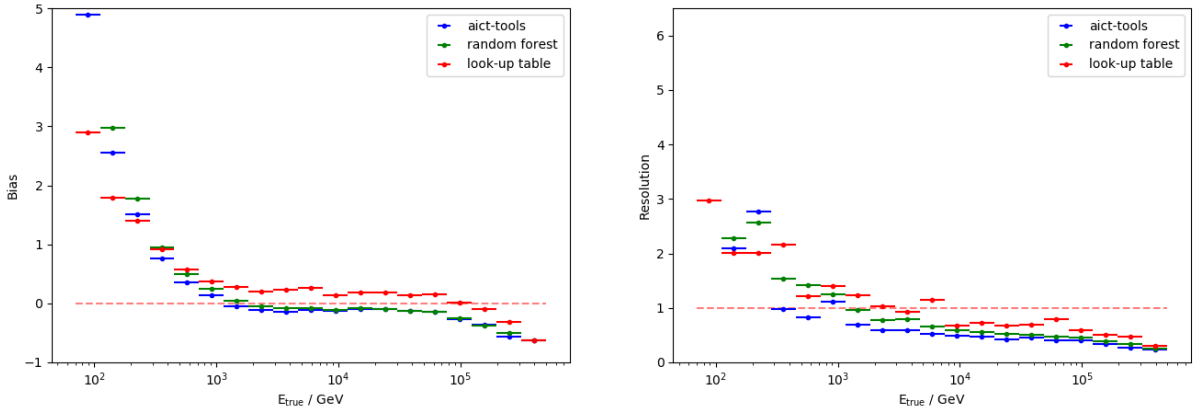


Figure 2: Bias and resolution of the different methods. Aict-tools and the so-called random forest are based on a machine learning approch while the lookup table is constructed with the binning of simulations.

works with all three methods, whereby the lookup table performs better at lower energies than the random forests. At higher energies the random forests have a smaller bias

then the lookup table. In order to test the methods in the application on real data, the reconstructed energy of both real data and simulations is displayed and compared in a histogram, see figure 3. The combined data set consists of proton and helium simulations

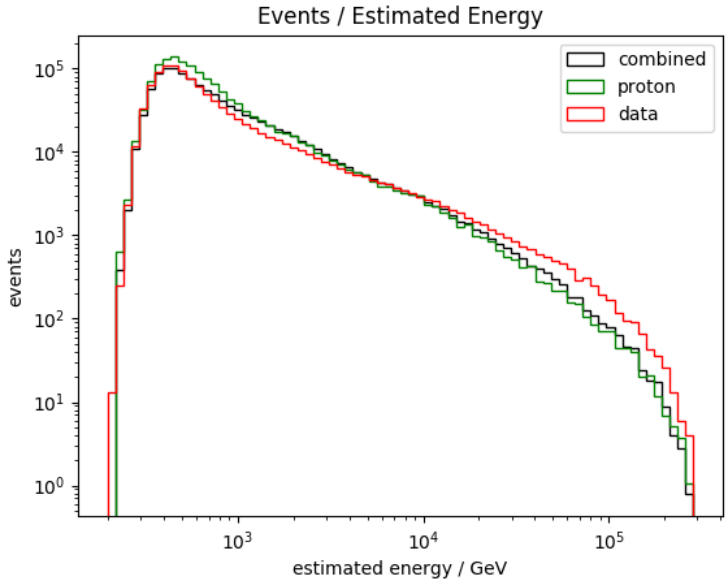


Figure 3: The distribution of the energy reconstructed with aict-tools. The random forest was trained on proton simulations and then applied to a second set of proton simulations, as well as on a proton-helium set and real observation data.

to represent the real cosmic ray spectrum. The simulation matches the data at lower energies, a mismatch is clearly seen in higher energies. This mismatch can be explained by the fact that at higher energies the fraction of heavier nuclei in the cosmic ray spectrum increases.
The next steps in this project are the application of a larger training set with heavier nuclei like iron, oxygen and carbon and the particle identification using machine learning.

References

[1] Aleksic, J, et al., The major upgrade of the MAGIC telescopes, Part II: A performance study using observations of the Crab Nebula. *Astroparticle Physics 72*, pp. 76-94.

[2] Nöthe, M., Brügge, K. and Buß, J., *aict-tools* - Reproducible Artificial Intelligence for Cherenkov Telescopes. *https://github.com/fact-project/aict-tools*.

[3] Particle Data Group, Patrignani, C. et al., Review of Particle Physics - 29. Cosmic Rays. Chin. Phys. C, 40, 100001 (2016) and 2017 update. *http://pdg.lbl.gov*.

Event Reconstruction for the Cascade Real-Time Alert Stream in IceCube

Mirco Hünnefeld
Lehrstuhl für Experimentelle Physik E5b
Technische Universität Dortmund
mirco.huennefeld@tu-dortmund.de

IceCube is a neutrino detector located at the geographic South Pole, instrumenting a cubic kilometer of glacial ice. A major goal of IceCube is the detection of astrophysical neutrino sources. Therefore, a real-time alert system was implemented to enable multi-messenger astronomy. Events that pass certain selection criteria are reconstructed in real-time on-site at the South Pole. If these events meet the requirements, they are sent out as alerts to telescopes around the world, enabling follow-up observations. Precise and fast reconstruction methods are necessary to abide the harsh resource constraints given on-site. The inclusion of cascade-like events harbors further challenges due to their inherently difficult angular reconstruction. In this paper, a deep learning-based reconstruction method is presented, which enables the accurate and fast reconstruction of cascade-like events and thus the implementation of a cascade alert stream.

1 CNN-based Real-Time Cascade Reconstruction

In order to implement a cascade real-time alert stream in IceCube, reliable and efficient cascade reconstruction methods are required. We use a Convolutional Neural Network (CNN)-based reconstruction method [1] and develop a technique to obtain accurate uncertainty contours. The CNN reconstructs the neutrino direction by predicting the components of the unit direction vector. In addition, uncertainty estimates on each direction vector component are obtained. This is accomplished through the use of a loss function that is composed of three independent 1D Gaussian Likelihoods for each

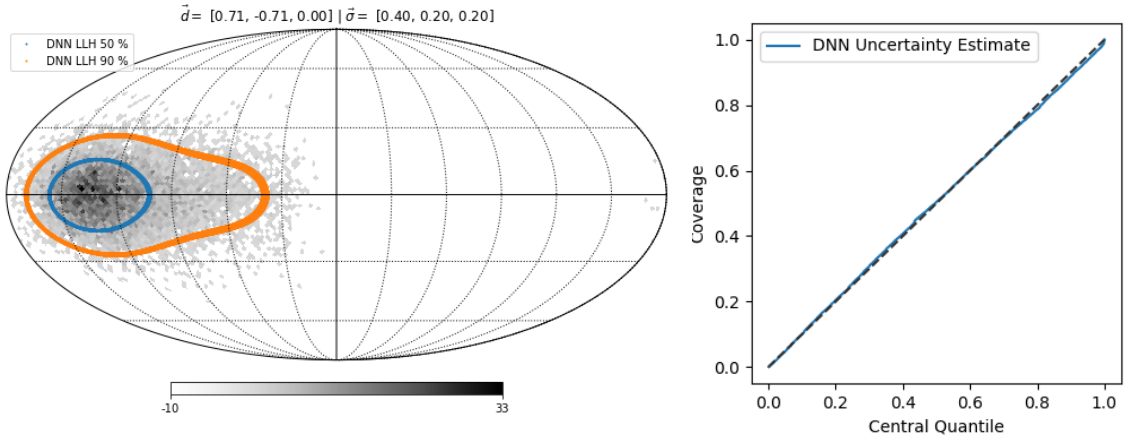


Figure 1: The 50% and 90% containment contours are shown on the left for an example event. The quality of the uncertainty estimate can be quantified by calculating the coverage, e.g. how many events actually fall within the estimated contour, for a certain central quantile.

direction vector component. The output of the CNN therefore consists of the best fit direction as well as the estimated standard deviation of the Gaussian residuals.

The obtained uncertainty estimates on each of the direction vectors can then be utilized to construct uncertainty contours around the best fit direction. An example contour is shown on the left of Fig. 1 and the coverage obtained on a sample of simulated events is illustrated on the right.

The CNN is trained on a baseline simulation set as well as simulation sets with varied systematic parameters. This enables the CNN to include known systematic uncertainties and to extend the contours appropriately. As shown on the right of Fig. 1, an excellent coverage is obtained on the test set.

Events that pass certain selection criteria undergo initial and simplified reconstructions in real-time on-site at the South Pole and are then sent north via satellite [2]. Once the events are received per satellite, alerts are sent out to telescopes around the world and more advanced reconstruction methods are triggered. When the advanced reconstruction methods are completed, refined directions and uncertainty contours are distributed.

The developed CNN-based method is fast and efficient enough to be run directly at the South Pole. However, in order to test the developed method prior to deployment on-site, a real-time bot is implemented that automatically reconstructs the events which are sent north via satellite and then distributes the results internally via the messenger platforms Slack and Telegram. Figure 2 shows the result of the reconstruction for the HESE [3] alert IceCube-200110A.

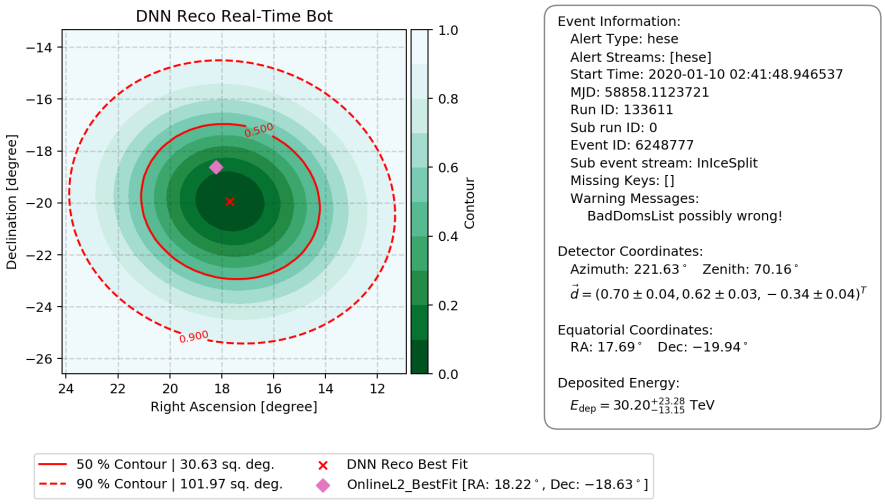


Figure 2: The real-time bot result for HESE [3] alert IceCube-200110A is shown.

2 Cascade-Generator: Overcoming Limitations of CNNs

Despite its success, the CNN-based reconstruction has its limitations. The employed convolutional layers do not optimally use available timing information and cannot naturally handle the geometry of the detector, which limits the full exploitation of translational invariance. This could potentially be solved with the help of geometric deep learning. However, translational invariance in the measured data in IceCube is only approximately given due to inhomogeneities in the detector medium, and more symmetries and prior knowledge exist, which remain unused.

An alternative approach to include further symmetries and prior knowledge is developed. This approach combines the strengths of both maximum-likelihood-estimation and deep learning [4]. A generative network, the cascade-generator, is trained to estimate the expected pulses in the detector based on a cascade hypothesis. Once the generator is trained, it can be used in reverse to fit the cascade hypothesis for a given set of measured pulses in the detector. In this approach, symmetries and prior knowledge can easily be utilized, similarly to maximum-likelihood methods. As a result, the reconstruction performance can be further improved compared to the CNN-based (DNN reco) and standard reconstruction method (Monopod [5]) in IceCube as shown in fig. 3.

3 Conclusion and Outlook

A real-time bot is implemented that utilizes the CNN-based cascade reconstruction method to automate the reconstruction of cascade events for a new cascade alert-stream

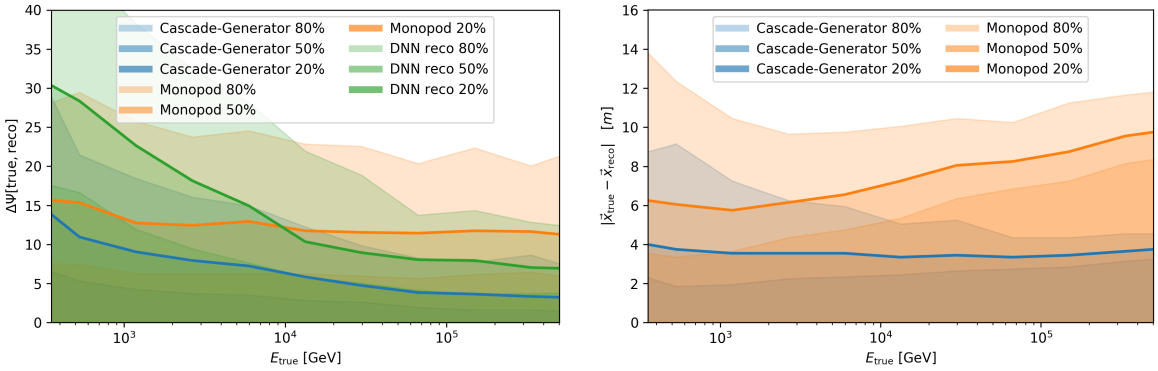


Figure 3: On the left, the angular resolution for cascade events is shown for the current state-of-the-art reconstruction method (Monopod [5]) and for the newly developed CNN-based reconstruction (DNN reco) and the cascade-generator. On the right, the cascade vertex resolution is shown for the current standard method and the cascade-generator.

in IceCube. The CNN-based reconstruction method can provide fast and accurate reconstructions combined with reliable uncertainty contours. Currently, the real-time bot is being tested internally in IceCube. In parallel, a novel approach is developed which combines the benefits of maximum-likelihood-estimation and deep learning. This approach can exploit symmetries and prior knowledge and therefore has the capability to surpass current state-of-the-art reconstructions in IceCube.

References

[1] M. Huennefeld (IceCube Collaboration), *Deep Learning in Physics exemplified by the Reconstruction of Muon-Neutrino Events in IceCube*, PoS **ICRC2017**, 1057 (2017)

[2] M. Aartsen et al. (IceCube Collaboration), *The IceCube realtime alert system*, Astropart. Phys. **92**, 30–41 (2017)

[3] M. Aartsen et al. (IceCube Collaboration), *Observation of High-Energy Astrophysical Neutrinos in Three Years of IceCube Data*, Phys. Rev. Lett. **113**, 101101 (2014)

[4] M. Huennefeld (IceCube Collaboration), *Reconstruction Techniques in IceCube using Convolutional and Generative Neural Networks*, EPJ Web of Conferences **207**, 05005 (2019)

[5] M. Aartsen et al. (IceCube Collaboration), *Energy reconstruction methods in the IceCube neutrino telescope*, JINST **9**, P03009 (2014)

Multiwavelength Analysis of 3C 84/NGC 1275

Lena Linhoff
Experimentelle Physik 5
Technische Universität Dortmund
lena.linhoff@tu-dortmund.de

The elliptical galaxy 3C 84 ($z = 0.017$) is located at the Perseus cluster and one of the closest and brightest radio galaxies. Because of its proximity it has been observed and studied quite well over years with different ground- and space-based detectors and telescopes. The fact, that we measure both gamma and radio emission and are able to distinguish different radio emission regions within the source opens quite unique opportunities. Since the acceleration mechanisms are quite unclear, TeV radio galaxies like 3C 84 are perfect candidates to study these mechanisms. As a first approach I segmented the different radio emission regions, investigate the correlation of multiwavelength lightcurves and undertake some calculations to restrict the possible gamma-ray emission region, by using estimations of the optical depth of the broad line region.

Data

3C 84 is monitored by multiple gamma-ray detectors, e.g. the Major Atmospheric Gamma-Ray Imaging Cherenkov (MAGIC) telescopes, that measure the source at very high energies since 2010 in an irregular order. MAGIC detected a strong gamma-ray flare at the beginning of 2017 [2]. FermiLAT also observes at MeV range since 2008 very regularly and reported a continuous increase in flux from 2008 to 2018, since then, the flux decreased quite abrupt. Radio data from the Very Long Baseline Array (VLBA) reveal the structure of the source as follows: the core component C1, which is assumed to host the black hole, is rather stable in size and flux; C2 an old component south east from the

core is very faint in these days and a newer component C3 that emerges from the core in 2007, reaches its peak intensity in 2016 and is now nearly vanished [7]. Additionally to that, optical data are available.

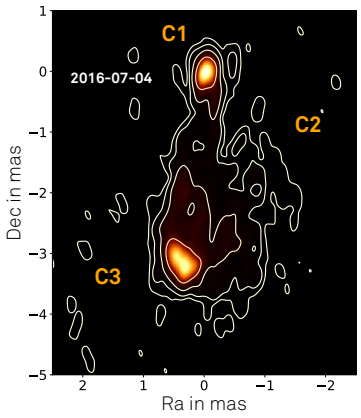


Figure 1: Radio Map of 3C 84

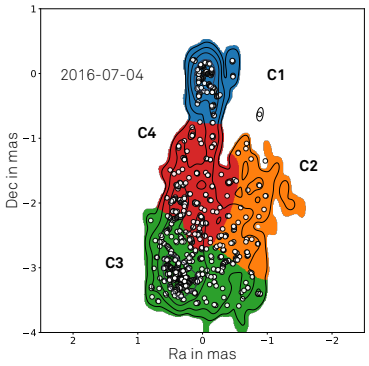


Figure 2: Segmented emission regions found by the random walker algorithm.

Total Intensity Emission Regions

To study the total intensity of the separate emission regions, it was necessary to develope and use new techniques, because the standard approach (fitting Gaussian components manually in Difmap [8]) is not suitable for the amount of data (85 maps), which is available these days. For this reason I, used a random walker algorithm [6] to separate the different emission regions in all maps automatically 2. Some adjustments still have to be done with this approach, but far less than with Difmap. The segmentation reveals that the increase in total flux clearly comes from the variable and moving C3 region. C1 is stable as expected 3.

Unbinned Discrete Correlation Function

To further investigate the behavior of the multiwavelength lightcurves and search for correlations, an UDCF [3] was adapted to the data. For this purpose, 1000 artificial lightcurves were produced and correlated using DelcGen [4] to estimate the significance of the result. The C1 lightcurve show a 3σ correlation with the FermiLAT gamma-ray lightcurve, as well as the optical R-band data correlated with VLBA 43 GHz C1.

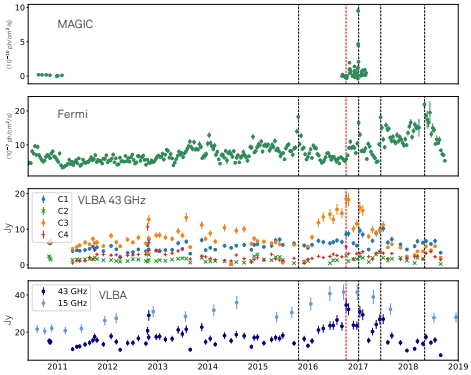


Figure 3: Multiwavelength lightcuves of NGC 1275 (aka. 3C 84)

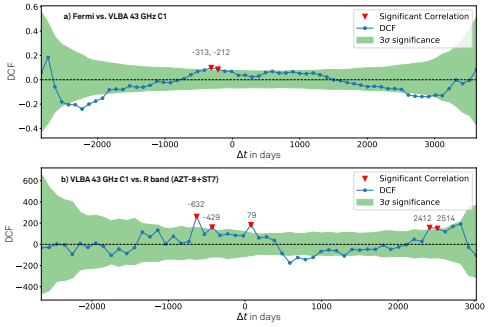


Figure 4: Unbinned discrete correlation function.

Restriction to the Gamma-Ray Emission Region

The fact, that we see gamma-rays as well as radio emission from a single source challenge some models, which try to explain the acceleration mechanisms within these sources. It is still unclear, if the emission is produced in the core of a source near the central black hole or further downstream the jet. Following calculations of Finke [5], it is possible to give at least a minimal distance of the emission region from the black hole, by calculating the absorption of photons in the broad line region of a source, depending on its energy and distance to the black hole. For this purpose, I fitted a logparabola spectrum modified by a photo-absorption term to the data measured by MAGIC and FermiLAT. The fitted values clearly states, that (under the assumption the model used by Finke is correct and only one emission region exists) the gamma-ray emission takes place outside the broad line region and therefore not near the central black hole. With this findings, some models, e.g. the magnetospheric model [1], can be excluded.

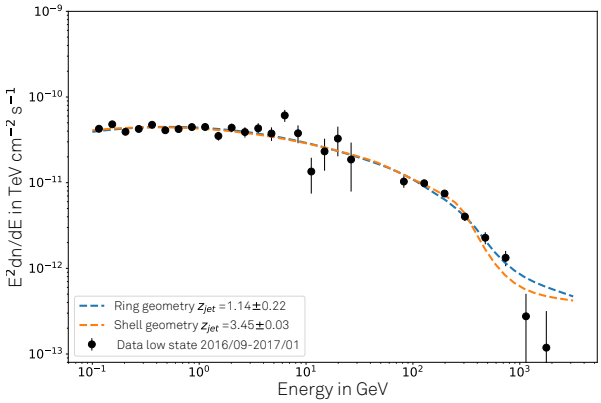


Figure 5: Photo-absorbed logparabola-fit on FermiLAT and MAGIC data, using two different shapes of the BLR, following [5].

Outlook

As a next step the longterm lightcurve of the VHE emission is needed from the MAGIC observations. For this reason all the data, MAGIc has ever taken from this source, has to be analysed. This longterm analysis is a quite crucial task, due to the current state of the MAGIC analysis software, which is not suitable scalable for such longterm analyses. The next step is therefore to automatize these analysis, so that longterm analyses will be possible and produce a very valuable scientific outcome.

References

[1] FA Aharonian, MV Barkov, and D Khangulyan. Scenarios for ultrafast gamma-ray variability in agn. *The Astrophysical Journal*, 841(1):61, 2017.

[2] S Ansoldi, LA Antonelli, C Arcaro, D Baack, A Babić, B Banerjee, P Bangale, U Barres de Almeida, JA Barrio, J Becerra González, et al. Gamma-ray flaring activity of ngc 1275 in 2016-2017 measured by magic. *arXiv preprint arXiv:1806.01559*, 2018.

[3] RA Edelson and JH Krolik. The discrete correlation function-a new method for analyzing unevenly sampled variability data. *The Astrophysical Journal*, 333:646–659, 1988.

[4] D Emmanoulopoulos, IM McHardy, and IE Papadakis. Generating artificial light curves: revisited and updated. *Monthly Notices of the Royal Astronomical Society*, 433(2):907–927, 2013.

[5] Justin D Finke. External compton scattering in blazar jets and the location of the gamma-ray emitting region. *The Astrophysical Journal*, 830(2):94, 2016.

[6] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1768–1783, 2006.

[7] Hiroshi Nagai, Monica Orienti, Motoki Kino, Kenta Suzuki, Gabriele Giovannini, A Doi, K Asada, M Giroletti, J Kataoka, F D'Ammando, et al. Vlbi and single-dish monitoring of 3c 84 for the period 2009–2011. *Monthly Notices of the Royal Astronomical Society: Letters*, 423(1):L122–L126, 2012.

[8] MC Shepherd, TJ Pearson, and GB Taylor. Difmap: an interactive program for synthesis imaging. In *Bulletin of the American Astronomical Society*, volume 26, pages 987–989, 1994.

Automatic Analysis for the MAGIC Telescopes

Simone Mender
Experimentelle Physik 5
Technische Universität Dortmund
simone.mender@tu-dortmund.de

The MAGIC telescopes, two imaging atmospheric Cherenkov telescopes, are used to monitore very-high gamma-ray emission of active galactic nuclei. Especially in the case of variable sources, long-term observations can provide information on emission and acceleration mechanisms. In the analysis of MAGIC data, the main challenges are the separation of gamma-ray events from background and the reconstruction of energy and direction. Since data of a long-term observation are collected under various environmental conditions, it is a challenge to analyze these data in a consistent way. In order to ensure an uniform analysis, automated analysis is a suitable solution.

Gamma-Ray Astronomy with the MAGIC Telescopes

MAGIC is a system of two imaging atmospheric Cherenkov telescopes located at the Roque de los Muchachos on La Palma, Canary Islands, at about 2200 m above sea level. MAGIC detects Cherenkov light from air showers induced by gamma rays with energies from 50 GeV to 50 TeV [4]. In 2011 and 2012, the performance of MAGIC has been significantly improved by series of upgrades [5]. MAGIC achieves its best performance under dark conditions. But to extend the duty cycle, observations are performed with different background level, e.g. under moonlight [1].

The analysis of data is performed with *MARS*, the MAGIC analysis and reconstruction software [6] following the standard analysis chain described in [5]. One main challenge

in gamma-ray astronomy is the separation of gamma-ray-induced events from the background, consisting of hadron-induced events. This separation as well as the reconstruction of origin and energy of the primary gamma ray is estimated with look-up tables and random forests. In *MARS* Monte Carlo (MC) simulations of gamma-ray events are used for traFining of models. As several hardware modifications took place after 2012 and the conditions (e.g. dust on the mirrors of the telescopes) are variable, different MC sets are available for different periods. For the analysis of data taken under moonlight, the standard analysis has to be adapted with proper cleaning levels [1].

Long Term Analysis of Active Galactic Nuclei

The emission range of active galactic nuclei (AGN) cover the entire electromagnetic spectrum. To study emission and acceleration models it is essential to observe AGNs at all wavelengths from radio to gamma rays. As some AGNs show a huge variability, it is reasonable to monitor those sources over a longer period. Individual flux outbursts, in particular, can provide insights into emission processes.

To investigate the gamma-ray variability of AGNs, MAGIC is monitoring some sources over several years. For example, the radio galaxy IC 310 is monitored since its gamma-ray activity has been detected with the MAGIC telescopes (*ATel* #2510) in 2012. A large flare was observed [2] during a multi-wavelength campaign between 2012 November and 2013 January.

Since then, the IC 310 is deeply monitored with the MAGIC telescopes under various conditions. IC 310 is located next to the galaxy NGC 1275, which are both part of the Perseus cluster. MAGIC is either observing Perseus-MA or NGC 1257. Therefore, the offset from the camera center to IC 310 is dependent on the dataset. In the different nights, shown in Figure 1, data was taken under different conditions. To analyze the data, various moon conditions, zenith ranges, MC data and weather conditions must be taken into account.

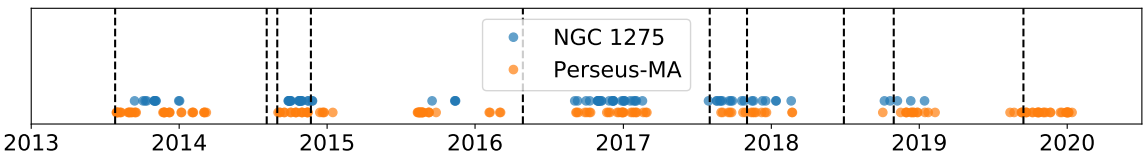
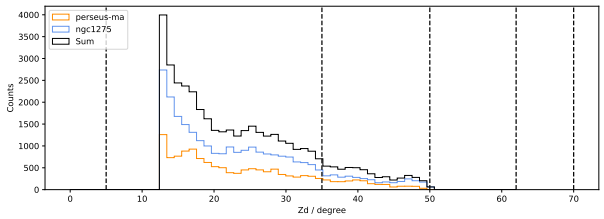


Figure 1: MAGIC observations of NGC 1275 and Perseus-MA from 2013 to now. The dashed lines indicates different MC periods.

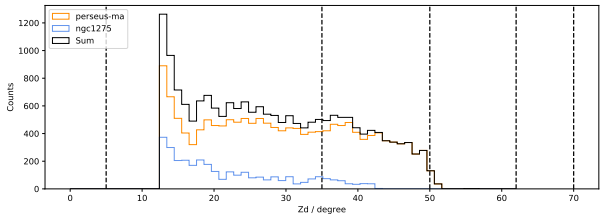
Automatic Analysis of MAGIC data

To ensure a consistent analysis of longterm observations with MAGIC, an automated analysis is a suitable solution. This is realized by a `Makefile`, in which all analysis steps are defined. Individual settings can be made in a .makerc file. To call *MARS* the python package `subprocess` is used.

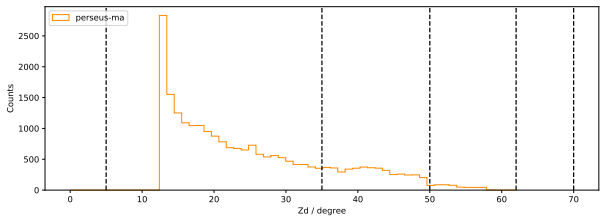
The first step is the investigation of the data. Zenith, weather condition and moon condition during the observations are important parameters for this investigation. In Figure 2, the zenith distribution is exemplary shown for three MC periods. The data is divided into different sets depending on the environmental conditions, so that suitable off data can be created for each set. With these off data and suitable MCs, models are trained for separation and reconstruction of the gamma events. The models can then be applied to the data.



(a) Zenith distribution in the MC period ST.03.07.



(b) Zenith distribution in the MC period ST.03.11.



(c) Zenith distribution in the MC period ST.03.12.

Figure 2: Zenith distribution of Perseus-MA and NGC 1275 for three different MC periods. Dashed lines are indicating the limits of MC sets.

Outlook

The next step is to include high-level analysis, like production of lightcurve and spectrum, in the automatic analysis. For this purpose, all different data sets must be merged together. To make sure that the automatic analysis corresponds to the MAGIC standard analysis, the results of the automatic analysis will be compared with the manual analysis.

Afterward light curve and spectrum can be used for long-term variability or multi-wavelength studies.

References

[1] M.L. Ahnen et al. Performance of the magic telescopes under moonlight. *Astroparticle Physics*, 94:29 – 41, 2017.

[2] J. Aleksić et al. Black hole lightning due to particle acceleration at subhorizon scales. *Science*, 346(6213):1080–1084, Nov 2014.

[3] J. Aleksi et al. Detection of very high energy γ-ray emission from the perseus cluster head-tail galaxy ic 310 by the magic telescopes. *The Astrophysical Journal Letters*, 723(2), 2010.

[4] J. Aleksi et al. The major upgrade of the magic telescopes, part i: The hardware improvements and the commissioning of the system. *Astroparticle Physics*, 72:61–75, 2016.

[5] J. Aleksi et al. The major upgrade of the magic telescopes, part ii: A performance study using observations of the crab nebula. *Astroparticle Physics*, 72:76–94, 2016.

[6] R. Zanin. Mars, the magic analysis and reconstruction software. In *Proceedings, 33rd International Cosmic Ray Conference (ICRC2013): Rio de Janeiro, Brazil, July 2-9, 2013*, page 0773, 2013.

Image Reconstruction of Radio Interferometric Data Using Neural Networks

Kevin Schmidt
Experimentelle Physik 5
Technische Universität Dortmund
kevin3.schmidt@tu-dortmund.de

Very long baseline radio interferometry allows the observation of distant astronomical objects with the highest resolution. This technique combines the data of several radio telescopes to achieve an effective diameter equal to the greatest baseline. Radio interferometers measure visibilities depending on the baselines between the individual telescopes. Based on their sparse distribution, much visibility space remains uncovered. This lack of information causes noise artifacts in the recorded data, which must be removed in a time-consuming analysis to receive a clean image.

With increasing data rates of modern radio interferometers, fast solutions are necessary to analyze observations in a reasonable time. One approach is the usage of machine learning techniques like neural networks. In this report, the feasibility study of reconstructing sparse radio interferometric data using convolutional neural networks is presented based on a toy dataset.

Imaging Radio Interferometric Data

Radio interferometers consist of several radio antennas connected to one big telescope. This technique achieves very high resolutions, as the effective diameter of the interferometer corresponds to its largest baseline. In this way, it is possible to observe the finest structures of astrophysical sources.

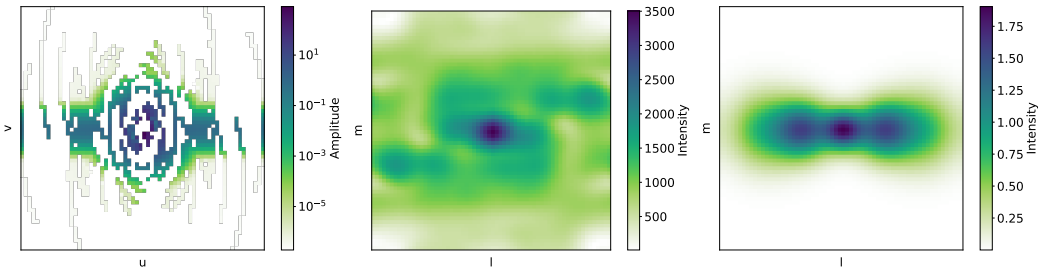


Figure 1: Sampled visibility space of an interferometric observation (left). Reconstructed image by taking the inverse Fourier transformation of the incomplete visibility space (middle) and real image obtained with complete visibility space (right).

The data recorded by radio interferometers contain no direct information about the intensity distribution in the sky. Radio interferometers collect information in Fourier space. Every telescope pair can measure one so-called visibility per timestep, depending on the baseline between the two telescopes. By sampling data points, the visibility space of the observation is filled. The inverse Fourier transformation of this visibility space leads to an image of the observed source.

The visibility coverage of a radio interferometer depends on its layout and the duration of observation. As there is only a limited number of antennas available, visibility space always remains incomplete. This leads to a lack of information for the inverse Fourier transformation causing noisy artifacts in the image of the astrophysical source. Figure 1 visualizes this reconstruction problem.

For decades different implementations of the CLEAN algorithm [1] are used to remove these noisy artifacts. With this method, the flux of the source is cleaned iteratively from the data starting with the brightest component. Continuing until all parts of the source are cleaned, a complete model is built, which is the basis to create a clean image of the observed source. Nevertheless, the CLEAN algorithm performs slowly. Thus, its usage on big datasets is unhandy. Increasing data rates of modern radio interferometers [2, 4] require faster solutions to analyze the observations in a reasonable time. One approach is the reconstruction of radio interferometric data with neural networks.

Image Reconstruction with Neural Networks

A big advantage of neural networks is their ability to process images very fast. In recent years, their application on image completion tasks became more and more prominent [3, 6]. Our idea is to reconstruct the incomplete visibility space and to create the inverse Fourier transformation using a convolutional neural network. This will make it possible to obtain a cleaned image of the astrophysical source from the calibrated dataset measured by the radio interferometer.

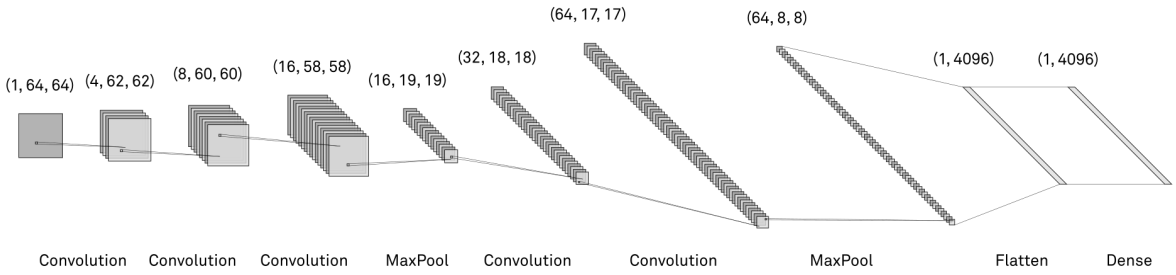


Figure 2: Architecture used to reconstruct incomplete visibility spaces.

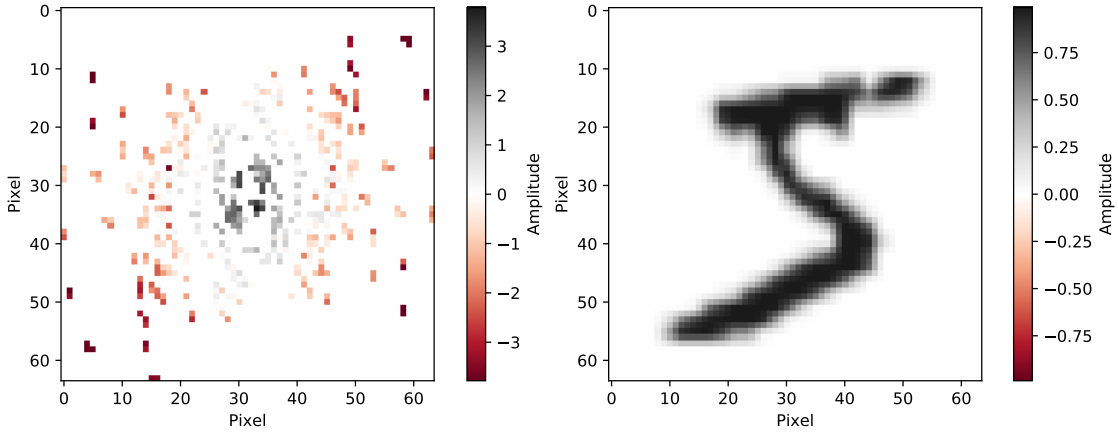


Figure 3: Input images (left) and target images (right) to train the neural network.

For a first feasibility study, we use a convolutional neural network with two convolutional blocks, each followed by a max-pooling. All convolutional layers have a batch normalization and a ReLU function attached. The network ends with a dense layer to fit the image size at the output. Figure 2 visualizes the complete architecture of the neural network.

As we had no simulations of radio data by hand at the start of the analysis, we started with the MNIST dataset. Initially, it had to be prepared to become a radio interferometric like dataset. Therefore, we applied two tasks on the images of handwritten digits: Firstly, we calculated the Fourier transform of the pictures. Secondly, we sampled frequencies in Fourier space to get an incomplete visibility space using a tool to simulate radio interferometric observations. By doing this, we generate a sample of input images to train the neural network. Figure 3 specifies the input and target images. We used this toy dataset to train the model for 350 epochs with mean squared error (MSE) loss. Afterward, we utilize a test set to evaluate the training. Figure 4 illustrates the reconstruction of the incomplete visibility space for one example.

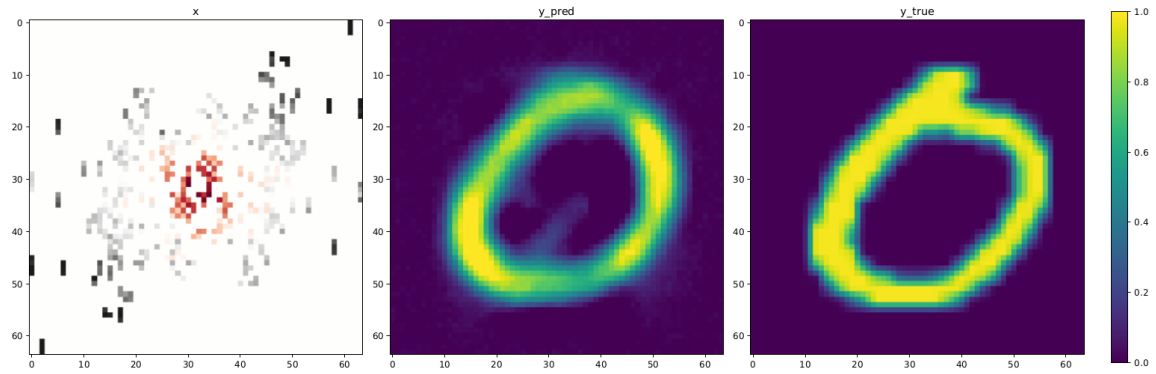


Figure 4: Visualization of the reconstruction results of the convolutional neural network. Input image (left), reconstructed image (middle) and true image (right).

Conclusion and Further Work

The results shown in the previous section indicate the feasibility of reconstructing incomplete data samples with convolutional neural networks. However, the analysis has to be adapted to the case of radio interferometric data more adequately. We plan to generate an improved toy dataset consisting of radio galaxies built of Gaussian components. This will lead to a more realistic description of astrophysical sources. Furthermore, we plan to use prior information like the PSF of the observation to improve the reconstruction. Another approach is improving the loss function to fit the problem in a better way [5].

References

[1] B. G. Clark. An efficient implementation of the algorithm 'CLEAN'. *Astronomy and Astrophysics*, 89(3):377, Sep 1980.

[2] P. E. Dewdney et al. The square kilometre array. *Proceedings of the IEEE*, 97(8):1482–1496, 2009.

[3] Jicong Fan and Tommy Chow. Deep learning based matrix completion. *Neurocomputing*, 266:540 – 549, 2017.

[4] van Haarlem, M. P. et al. Lofar: The low-frequency array. *A&A*, 556:A2, 2013.

[5] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

[6] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C4

Regression approaches for large-scale high-dimensional data

Katja Ickstadt Christian Sohler

An R package for linear regression on very large data sets

Esther Denecke
Lehrstuhl für Mathematische Statistik und
biometrische Anwendungen
Technische Universität Dortmund
esther.denecke@tu-dortmund.de

The overall aim of project C4 is the development of regression approaches for very large and high-dimensional data. In the course of this, the concept of *Merge & Reduce* – mainly known from data structures and coresets – has been transferred to statistical models. This makes the execution of both frequentist and Bayesian regression possible even when the data are too large to fit into main memory. This technical report introduces the `R` package `mrregression` containing implementations for both frequentist and Bayesian linear regression within the *Merge & Reduce* scheme.

1 Merge and Reduce

Merge & Reduce enables the execution of frequentist and Bayesian regression on very large data sets, e.g. when the data set does not fit into main memory. The methodology with respect to statistical models is introduced in chapter 4 of [2]. The basic idea is to read in a subset (block) of the data, to perform statistical analysis on this block of data, storing certain summary statistics of this model. Following this, models are merged (and reduced) following a tree structure. Overall, [2] proposes three merging approaches, where approaches 1 and 3 can be used in the frequentist and approach 2 in the Bayesian case. The first two approaches (1 and 2) are based on weighted means of the summary statistics (and some corrections factors). The third approach makes use of products of Gaussians recovering the OLS-estimator.

For frequentist linear regression the summary statistics include the regression coefficients and their standard errors and for Bayesian linear regression the distributions of the coefficients are characterized by the mean, median, quartiles, 2.5% and 97.5% quantiles.

2 Implementation

The `R` package `mrregression` – implemented in the statistical programming language `R` [3] – contains functionalities for conducting frequentist and Bayesian linear regression according to the *Merge & Reduce* principle as introduced in chapter 4 of [2]. The package contains two major functions, `mrfrequentist` and `mrbayes`. The functions return objects of classes `mrfrequentist` and `mrbayes`, respectively.

We start with a toy example where we are given simulated data with 50 000 observations and 50 variables. The data set is stored as the `R` object `exampleData`. The dependent variable has the column name `"y"`. An exemplary call of `mrfrequentist` using this data set can be specified as follows:

```
R> fit1 = mrfrequentist(dataMr = exampleData, approach = "1",
                        obsPerBlock = 10000, formula = y ~ .)
```

To fit the model, we specified four arguments where `dataMr` is a `data.frame` containing the data, `approach` is specified as `"1"` referring to a weighted mean procedure for merging the individual models, `obsPerBlock` gives the number of observations used per block and `formula` the formula as known from the standard function for linear regression, `lm`.

By default, calling the fitted object invokes the S3 method `print` for class `mrfrequentist`. We receive the following output (here truncated):

```
R> fit1
 Approach:
 1

 Model:
 y ~ .

 Coefficients:
 (Intercept)       x1        x2         x3         x4         x5
    -2.01286  1.29284  -9.17557  -12.37795  -15.81465  -12.52735
```

The output contains information on the approach used, the specified formula (named model) and the regression coefficients. The style of the coefficients returned follows the style the user knows from the `print` method for class `lm`.

Additionally, a summary method is available. The output contains more detailed information such as the level of the final model in the tree structure, the number of observations

the final model is based on as well as the regression coefficients and their standard errors. Again, the output is similar to that of the method `summary` for class `lm`. Note that for approach 3, the output of the `print` method has the same structure. Differences arise with the `summary` method and the internal structure.

In the example above, the data `exampleData` is stored within `R` as an object. However, one aim of the *Merge & Reduce* method is to provide a method for fitting regression models when the data are too large to fit into memory at once. To that end, the package offers the alternative argument `fileMr` where the data are read in blockwise from, e.g. a text-file. Internally, a connection to a file is opened, lines are read with the function `readLines` and formatted using function `fread` of the package `data.table` [1]. This results in chunks of size `obsPerBlock` being read in sequentially. After creating the model and saving the summary statistics, these chunks are deleted to set free memory. Upon using the argument `fileMr`, more arguments that can be specified when reading in data with `fread` become available. These include `sep, header, col.names, dec` and `na.strings`. Note that defaults can differ from package `data.table`. For the sake of consistency within the package `mrregression`, arguments within `mrfrequentist` and `mrbayes` are always written in camelCase, i.e. `colNames` and `naStrings`.

For fitting Bayesian linear models using the *Merge & Reduce* scheme the package `rstan` [4] is employed. A basic call to fit a Bayesian linear regression with intercept to data stored as a text-file could be

```
R> fit2 = mrbayes(fileMr = "exampleData2.txt", obsPerBlock = 500,
                  y = "y"),
```

where `y` is a character string with the column name of the dependent variable. Users familiar with `rstan` may have noticed that the arguments `data` and `model_code` from function `stan` are missing in the call. These are set to default values. It is, however, possible to specify ones own data within an argument called `dataStan` using `mrbayes` (this also explains the naming of `dataMr`). Currently, it is not possible to specify the `model_code` as known from `rstan`. Instead, the user has the choice between a default linear regression model with or without intercept.

As noted above, objects returned are of classes `mrfrequentist` and `mrbayes`, respectively. The internal structure is always a list. All internal structures contain list entries of the `level` of the final model in the tree structure, the number of observations (`numberObs`), the summary statistics, and `dataHead` containing the first six rows of the data of the first block. The latter entry serves as a sanity check when using the argument `fileMr`. Thus, the overall number of observations the final model is based on can be accessed in the following way:

```
R> fit1$numberObs
 [1] 50000
```

In the Bayesian case diagnostic measures for Bayesian methods are available. These can be accessed via the list entry `diagnostics` of the resulting object, i.e.

```
R> fit2$diagnostics
```

in the above example. The diagnostic measures contain the minimal observed effective sample size on level 1 and the potential scale reduction factor on split chains – here the maximum observed value on level 1. Level 1 contains the models built directly on the blocks of data.

3 Summary and outlook

The `R` package `mrregression` allows performing frequentist and Bayesian linear regression using the *Merge & Reduce* technique as described in chapter 4 of [2]. The package offers functionality for sequentially reading in blocks of data from a file when the data set is too large to fit into main memory. Moreover, for frequentist linear regression the user can choose out of two different merging approaches. Sensible summary statistics are saved for both frequentist and Bayesian linear regression.

The package's calculations are currently done sequentially, one block after the other. To speed up calculations it would be beneficial to parallelize computations in the future. Additionally, the class of regression models could be extended to e.g. Poisson regression as described in chapter 4 of [2]. Moreover, it could also be possible to allow the user to specify their own merge functions.

References

[1] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2019. R package version 1.12.6.

[2] Leo N. Geppert. *Bayesian and Frequentist Regression Approaches for Very Large Data Sets*, 2018. Doctoral Thesis, TU Dortmund University.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[4] Stan Development Team. *RStan: the R interface to Stan*, 2019. R package version 2.19.2.

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

Studies and application of newly developed LHCb Flavour Tagging algorithms

Kevin Heinicke
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
kevin.heinicke@tu-dortmund.de

The Flavour Tagging software package, used to deduce the initial flavour of neutral and oscillating B mesons has been re-optimised until the end of 2017. A study of the reliability of a newly developed Flavour Tagging algorihm within the framework has been performed since then. Furthermore the new framework is used in a measurement of the B_s oscillation frequency Δm_s with $B_s \to D_s^- \pi^+$ decays.

1 Introduction to the LHCb Experiment

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 New Physics effects in CP-violating and rare 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).

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 sub detectors 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.

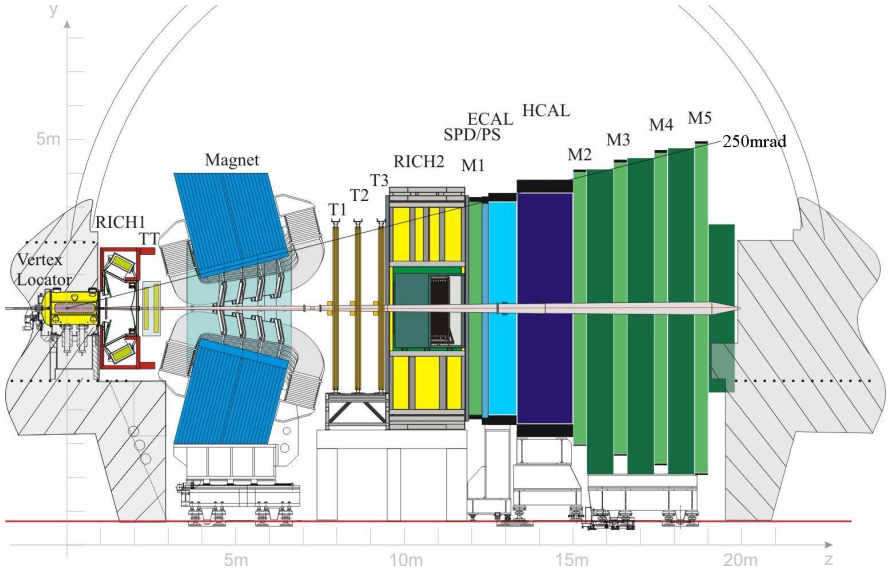


Figure 1: The LHCb detector with the various sub detectors for the identification of particles and reconstruction of their tracks [1].

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 50 ns / 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.

2 Study of the Inclusive Tagging Algorithm

A major subset of the LHCb analyses involve studies of *CP* violation, which can be used to test the Standard Model of particles physics (SM). These measurements include time-dependent decay rate measurements, many of which are subject to mixing of neutral B meson states. The knowledge of the initial flavour of these mesons is therefore crucial. It is being extracted with several Flavour Tagging algorithms, which are executed after the particle-decays have been fully reconstructed. The algorithms are designed to reconstruct particles on the same-side (SS) and opposite-side (OS) of the signal B candidate. The charge information of these particles is correlated with the initial flavour of the B candidate via different weak transitions. The different types of algorithms are depicted in figure 2.

In general each algorithm aims to identify a decay product either from the OS non-signal B meson, or from a SS hadron which hadronised with the signal B meson by applying different selection criteria. The selected particles are referred to as tagging particles.

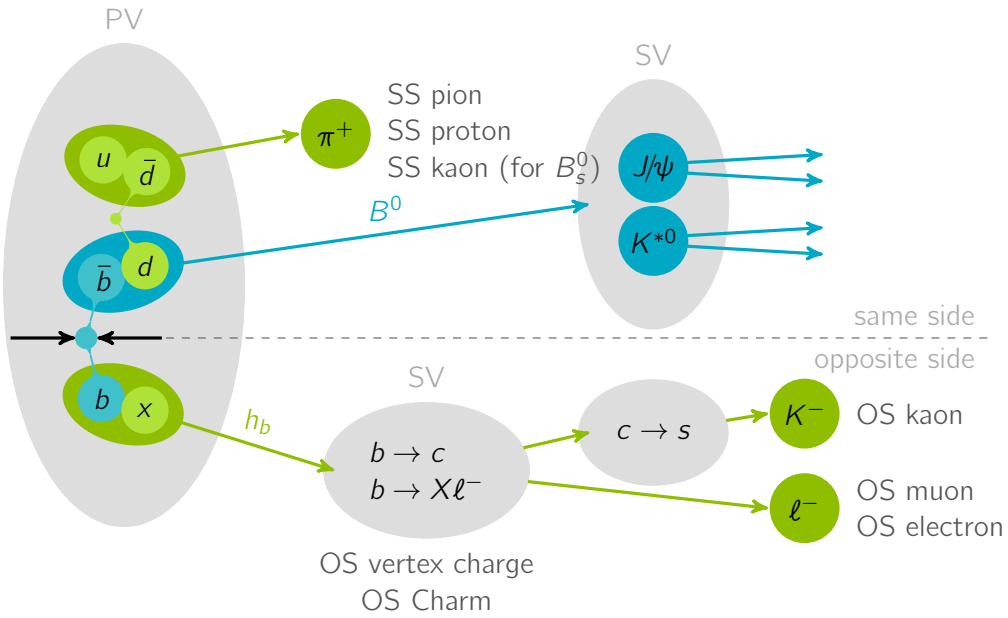


Figure 2: Schematic description of different Flavour Tagging algorithms. The same side algorithms infer the initial b flavour from pions and protons that hadronise alongside with the B meson. The opposite side algorithms infer this information from the non-signal B partner and its decay into leptons or $b \to c \to s$ transitions.

The algorithms differ in their specific decay product and selection strategies. Due to the large number of tracks which do not necessarily originate from the primary vertex, this process is error prone. The quality of the identification is therefore evaluated by calculating an overall tagging efficiency

$$\varepsilon_{\text{tag}} = \frac{N_{\text{tagged}}}{N_{\text{tagged}} + N_{\text{untagged}}}\,, \tag{1}$$

describing the ratio of B candidates for which a tagging particle has been found within all B candidates, and the mistag rate

$$\omega = \frac{N_{\text{incorrect}}}{N_{\text{tagged}}}\,, \tag{2}$$

which is the rate of incorrectly tagged events or false-positive rate. The overall performance of these algorithms is usually described in terms of the tagging power

$$\bar{\varepsilon}_{\text{eff}} = \varepsilon_{\text{tag}}\,(1-2\omega)^2\,. \tag{3}$$

To estimate the tagging power on untagged data, the mistag rate ω_i is predicted on a per-event basis by different multivariate analysis tools for each individual tagging algorithm, leading to the final figure of merit for Flavour Tagging, the per-event tagging power

$$\varepsilon_{\text{eff}} = \frac{1}{N_{\text{tagged}} + N_{\text{untagged}}} \sum (1-2\omega_i)^2\,. \tag{4}$$

A newly proposed strategy to infer the initial B flavour has been implemented together with Vukan Jevtic and Quentin Führing, using the previously reported, re-factored Flavour

Tagging software package. The Inclusive Tagger aims to defer the flavour information via a Recurrent Neural Network, trained on simulated event samples, using `keras` and the `tensorflow` backend [3, 4]. It will replace the individual flavour tagging algorithms depicted in Figure 2 with a single algorithm, allowing to deduce flavour information from the correlation between the different physical processes. After a study of the interpretability of the new algorithm has been performed and reported in the previous period, the validation of the algorithm is prepared. Part of this is the calibration of the established flavour tagging algorithms on the full LHCb Run 2 data sample from 2015 to 2018 which is described in the upcoming section for a sample of $B_s \to D_s^- \pi^+$ decays.

3 Measurement of the B_s oscillation frequency Δm_s

One of the key parameters of aforementioned measurements of *CP* violation in the B_s^0 system is the oscillation frequency Δm_s of neutral B_s^0 mesons. To measure Δm_s, a data sample of a flavour-specific process is needed for which the decay $B_s^0 \to D_s^- \pi^+$ is a perfect candidate: The charge of the final state $\pi^\pm$ determines the final flavour of the B_s meson. Combined with the flavour tagging information from previously mentioned flavour tagging algorithms, and the decay time, which can be measured with high precision in the VELO detector, the oscillation frequency Δm_s can be measured.

A first study of the expected signal yield has been performed, which is determined to be $N_{\text{sig}} = 395\,500 \pm 700$. Furthermore a pre-calibration of the OS flavour tagging algorithms is performed. The expected combined tagging power of OS combination and SS Kaon tagger is estimated to be $\varepsilon_{\text{eff}} = (7.03 \pm 0.03_{\text{stat.}} \pm 0.32_{\text{syst.}})\,\%$.

Together with a calibration of the decay time resolution, the expected systematic uncertainty on Δm_s can be determined. A blind fit to data yields $\sigma_{\Delta m_s,\text{stat}} = 0.0053\,\text{fs}^-1$, which indicates a statistical limitation of the measurement, if similar systematic uncertainties as the previous measurement are assumed [2].

References

[1] A. A. Alves, Jr. et al. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[2] R Aaij et al. Precision measurement of the B_s^0-$\bar{B}_s^0$ oscillation frequency with the decay $B_s^0 \to D_s^- \pi^+$. *New J. Phys.*, 15:053021, 2013.

[3] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[4] François Chollet et al. Keras. `https://keras.io`, 2015.

Measurement of CP violation in $B^0 \to \psi(\to \ell\ell) K^0_S(\to \pi^+\pi^-)$ and $B^0_s \to J/\psi\,(\to \mu\mu) K^0_S(\to \pi^+\pi^-)$ decays

Vukan Jevtić
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
vukan.jevtic@tu-dortmund.de

The LHCb detector [1] is one of the major particle detectors at the LHC. The main research subjects at LHCb are measurements of CP violation and the search for rare decays in the decays of beauty -and charm mesons. Since B mesons are mainly created in the forward region (along the beam axis), the LHCb detector is constructed as a one-arm forward-spectrometer. The LHCb detector that recorded the Run 2 dataset (2015 to 2018) that is used in this analysis is shown in figure 1. While the LHC is running, proton bunches are brought to collision in the Vertex Locator (VELO) at a rate of 40 MHz. The Vertex Locator measures the primary and secondary vertices of a proton-proton collision. The particles that are created in these collisions then traverse the rest of the detector. For physics analyses, a measurement of the particles mass and energy and charge is crucial. For this reason the LHCb detector is equipped with multiple subdetectors that measure the track hits created by charged particles traversing the detector as well as the momentum of the tracks. The tracking components of the LHCb detector are the VELO, the Tracker Turicensis (TT), which is a silicon strip detector located upstream of the LHCb magnet, the tracking stations T1-T3 consisting of an inner silicon tracker and a straw tube outer tracker as well as muon chambers. If charged particles reach these detector components, electromagnetic interactions occur and the detectors produce signals. That way, tracks can be reconstructed from multiple space-time points inside the detector. The warm LHCb dipole magnet bends tracks of charged particles and therefore allows for a measurement of the track curvature from which the track charge can be inferred. By measuring the track curvature and the track

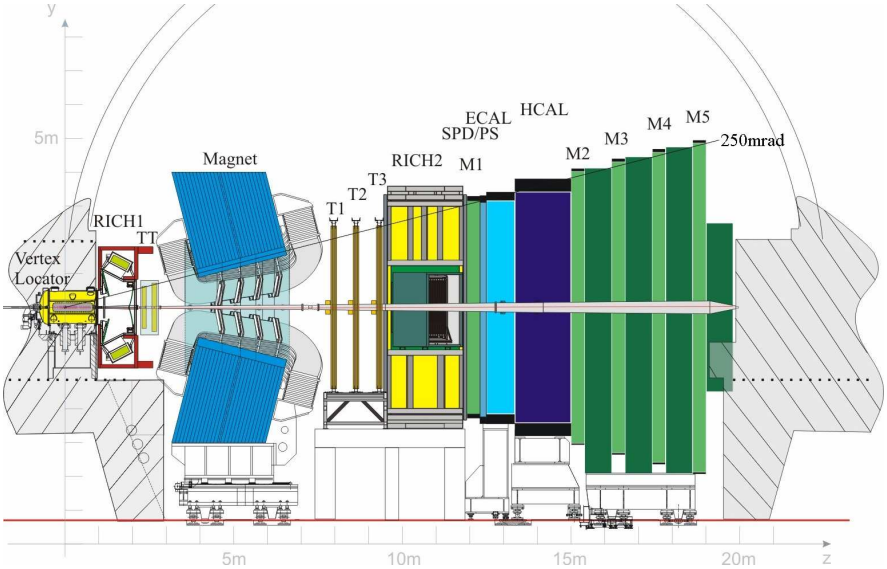


Figure 1: The LHCb detector with the subdetectors for the identification of particles and reconstruction of their tracks [1].

momentum, the particles mass can be computed which is a major indicator of a particles type. The track momentum is measured by the two Ring Imaging Cherenkov detectors RICH 1 and RICH 2, one is located upstream and the other is located downstream of the magnet, and they measure in two different momentum ranges. Inside the RICH detectors, particles are travelling faster than the speed of light of the RICH medium, and therefore a Cherenkov light cone is created at an angle that is directly related to the particles momentum. In the electromagnetic calorimeter (ECAL), the energy of interacting particles is measured. Particles like electrons and photons produce showers in the ECAL medium and lose all of their initial energy in this process. This energy loss is estimated by the amount of photons that are created in the ECALs scintillating layers. Muons and charged hadrons, like pions or protons traverse the ECAL without losing a significant fraction of their energy. The hadronic calorimeter (HCAL) is located after the ECAL and measures the energy of hadrons that produce showers of photons and gluons. Muons generally traverse the whole detector without losing a significant fraction of their initial energy. In the muon chambers M1 to M5 the tracks hits produced by muons are measured and allow in combination with the remaining LHCb tracking components, a very precise measurement of the muon momentum.

In this analysis, the *CP*-violation parameters S and C are measured in $B^0 \to J/\psi(\to \mu\mu)K^0_S$, $B^0 \to J/\psi(\to ee)K^0_S$ and $B^0 \to \psi(2S)(\to \mu\mu)K^0_S$ decays which are also referred to as *golden modes* for the measurement of indirect *CP* violation. This is due to the fact, that higher-order decay processes like penguin decays with non-*CP* violating phases only add small corrections to the tree-level decay [6]. This allows for an exceptionally clean

measurement of the *CP*-violation parameter S, a constant parameterizing, in part, the *CP*-violating nature of weak interactions, according to the Standard Model of particle physics. In these decays, *CP* violation occurs in the interference of mixing, i.e. the time dependent oscillation of B mesons between their matter and antimatter states B^0 and $\overline{B}^0$ and their decay into a *CP*-invariant final state. The *CP*-invariant final state analysed in the scope of this analysis is a charmonium state ψ consisting of two charm quarks and a strange meson K^0_S consisting of a strange quark and a down quark. The charmonium state J/ψ is reconstructed from either two electrons or two muons, while the charmonium state $\psi(2S)$ is only reconstructed from decays into two muons. The K^0_S meson is in all channels reconstructed from decays into two oppositely charged pions.

The *CP*-violation parameter S is determined by measuring the time dependent *CP*-asymmetry which is defined as

$$\mathcal{A}(t) \equiv \frac{\Gamma(\overline{B}^0(t) \to \psi K^0_S) - \Gamma(B^0(t) \to \psi K^0_S)}{\Gamma(\overline{B}^0(t) \to \psi K^0_S) + \Gamma(B^0(t) \to \psi K^0_S)} = \frac{S \sin(\Delta m\, t) - C \cos(\Delta m\, t)}{\cosh(\frac{\Delta\Gamma\, t}{2}) + A_{\Delta\Gamma} \sinh(\frac{\Delta\Gamma\, t}{2})}.$$

In this specific decay, $\Delta\Gamma \approx 0$ and the *CP*-violation parameter C is close to zero, and therefore measuring the parameter S is almost synonymous to measuring the amplitude of the time dependent *CP*-asymmetry. Additionally, *CP* violation is measured in the U-spin related decay $B^0_s \to J/\psi\, K^0_S$. This measurement can be used to estimate penguin contributions in the golden modes [5].

The challenge of this analysis is therefore to select as many signal decays as possible while achieving a good background rejection, and identifying the B flavour at production. Background is rejected, with the help of machine learning algorithms like boosted decisions trees. Boosted decision trees are supervised learning algorithms that are trained to separate a dataset into classes, for example signal and background events. In addition, the B flavour at production needs to be determined. This is achieved by measuring the charges of particles that are created in the hadronisation of either the signal B meson or the signal B's partner-beauty hadron. This is possible since b quarks are created in pairs. This is a especially challenging task which generally significantly decreases the effective size of the data sample. Hence, a huge effort is made to utilize LHCb existing flavour tagging algorithms to their fullest potential as well as developing new kinds of flavour tagging algorithms.

The final goal of this analysis is to measure the *CP* violation parameter S in a simultaneous fit to all discussed golden modes. This is likely going to result in the most precise single measurement of the *CP*-violation parameter S to date and the measurement precision will be comparable to the current world average. Previous measurements by LHCb that have been performed using he Run 1 dataset [2–4] are likely going to be superseded in precision.

References

[1] A. A. Alves, Jr. et al. The LHCb detector at the LHC. *JINST*, 3:S08005, 2008.

[2] R. Aaij et al. Measurement of *CP* violation in $B^0 \to J/\psi K^0_S$ decays. *Phys. Rev. Lett.*, 115:031601, 2015.

[3] R. Aaij et al. Measurement of the time-dependent *CP* asymmetries in $B^0_s \to J/\psi K^0_S$. *JHEP*, 06:131, 2015.

[4] R. Aaij et al. Measurement of *CP* violation in $B^0 \to J/\psi K^0_S$ and $B^0 \to \psi(2S)K^0_S$ decays. *JHEP*, 11:170, 2017.

[5] Kristof De Bruyn and Robert Fleischer. A Roadmap to Control Penguin Effects in $B^0_d \to J/\psi K^0_S$ and $B^0_s \to J/\psi\phi$. *JHEP*, 03:145, 2015.

[6] Philipp Frings, Ulrich Nierste, and Martin Wiebusch. Penguin contributions to *cp* phases in $B_{d,s}$ decays to charmonium. *Phys. Rev. Lett.*, 115:061802, Aug 2015.

Resource-aware Scientific Data Processing

Thomas Lindemann
Databases and Information Systems Group (DBIS)
TU Dortmund University
thomas.lindemann@cs.tu-dortmund.de

The main objective of our research is to get over hardware restrictions in processing capabilities due to electrical power consumption or thermal discharge constraints. Our current work aims to handle high volume scientific data on energy-efficient modern hardware clusters.

To validate our approaches, we are processing real world scientific use cases on our self-developed experimental systems. The use cases get provided by our collaboration partners of the particle physics department in the SFB876 collaboration. We aim to improve the processing efficiency from different approaches by adjusting the hardware configuration, the algorithms and the data storage patterns, while metering the execution time and energy consumption compared to existing solutions.

We are working in particular on handling large continuous event data streams, which are characterized by a enormous heterogeneity of the event complexity caused by a large variety of particles in every event. One of our approaches is to use modern hardware of different architectures to aim the event heterogeneity by placing and get the best results in performance and energy consumption at run time.

Since the last report, we worked in the SFB876-C5 on analysis on the efficiency of Tracking Algorithms used in the LHCb software framework. In our ongoing work, we made further analysis of the event structure and we are using the results to improve the energy consumption by Event-complexity-driven Machine Placement Algorithm, briefly referred as *EcoMap*.

Furthermore, we also cooperate in the SFB876-C3 project and analyzed the efficient processing of simulations for high energy particle interactions in the atmosphere. With a similar approach, we analyzed the process parameters and investigate parallel heterogeneous processing possibilities of this very different use case and achieved an MPI distributed approach.

1 SFB876-C5 - Particle Track Reconstruction

The LHCb project at CERN is a large and complex research project grown over the last decades. Its general scope is to explain the matter/anti-matter asymmetry. Our main topic in the SFB876 is the C5 Sub-Project, where a continuous stream of hits inside the LHC (Large Hadron Collider) is produced by the several stages of the LHCb 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 goal is to find new solutions for processing this big amounts of data with limited resources 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.

For our evaluation, we are using very different modern efficient ARM Cortex-A53 core hardware, from which we've constructed a 160-core cluster and we compare our execution results to a dual socket Intel Xeon E5-2695 Server with 48 virtual cores, that represents state of the art hardware. The single core performance of the Cortes-A53 cores is much lower than the Xeon cores, but these architecture provides more efficiency in energy, package size and thermal discharge. As a result, we can operate with a much higher core count under similar power constraints and process event data in a high distributed way.

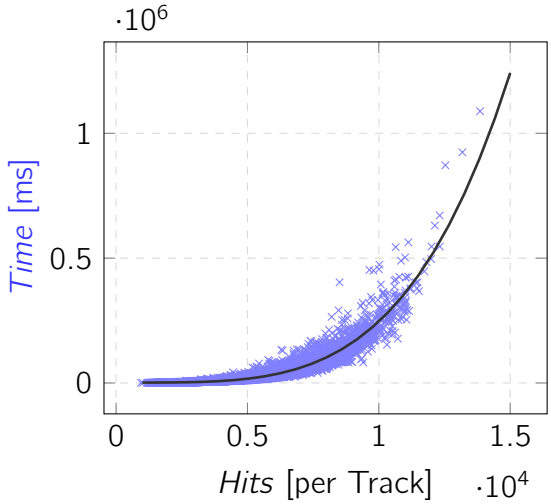


Figure 1: Single Core Event Processing Time over Hits on Intel Xeon CPU.

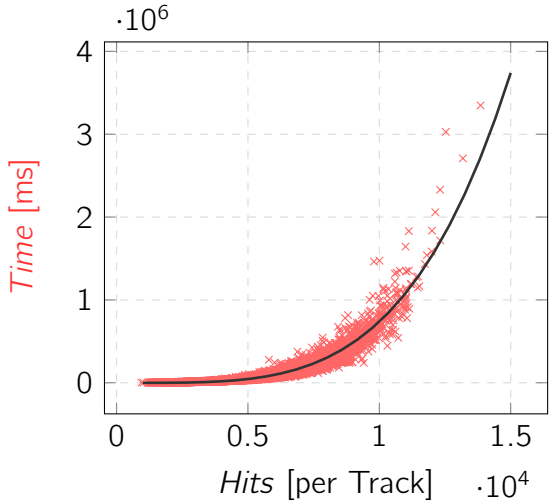


Figure 2: Single Core Event Processing Time over Hits on ARM Cortex-A53 Low Power CPU.

Our experiments confirm our approach. One challenge of the data is the wide spread of heterogeneity of the events, which is caused by different interactions of the particles after the collision. As seen in Figure 1 and 2, the execution time per grows exponential over the number of hits in a track, tested one a single core, due to lower single core performance. On contrast, Figure 3 and 4 show the opposite effect in energy mileage, the energy on a low power core grows much lower over the event size despite the higher execution time.

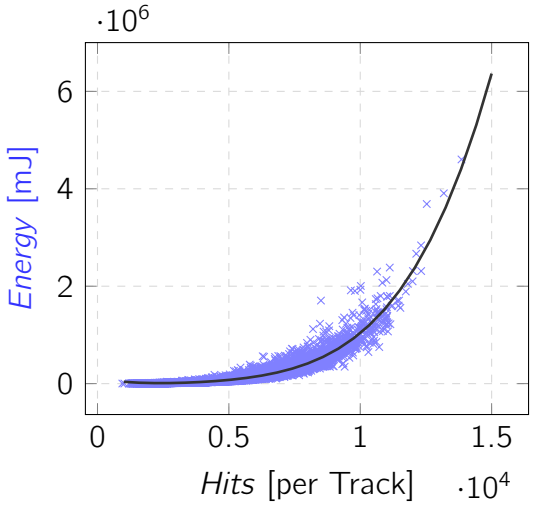


Figure 3: Single Core Event Energy Mileage over Hits on Intel Xeon CPU.

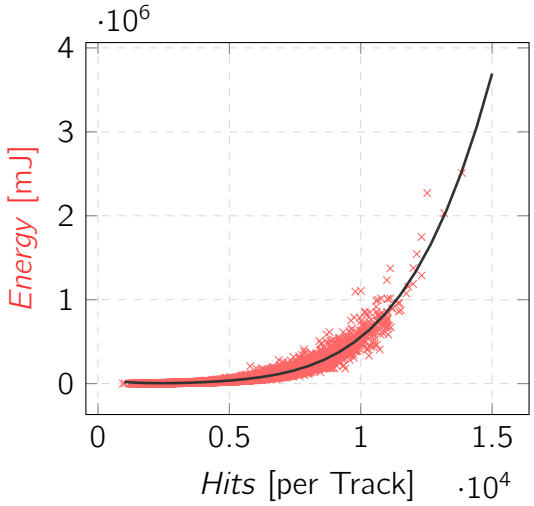


Figure 4: Single Core Event Energy Mileage over Hits on ARM Cortex-A53 Low Power CPU.

2 SFB876-C3/C5 Cooperation - Atmospheric Particle Shower Simulation with CORSIKA

CORSIKA (COsmic Ray SImulations for KAscade) is a program for detailed simulation of extensive air showers initiated by high energy cosmic ray particles. [1]
We did several experiments and gathered process data which is statistically evaluated to find correlation between simulation parameters and the execution time of an event on our reference machine to use the perceptions for placement decisions on the most suitable hardware.
Furthermore, we initially implemented a MPI supported variant of CORSIKA to the community that provides the opportunity to take sub particles of a single event from the particle stack and place them and their interactions to different heterogeneous machines in a network cluster.

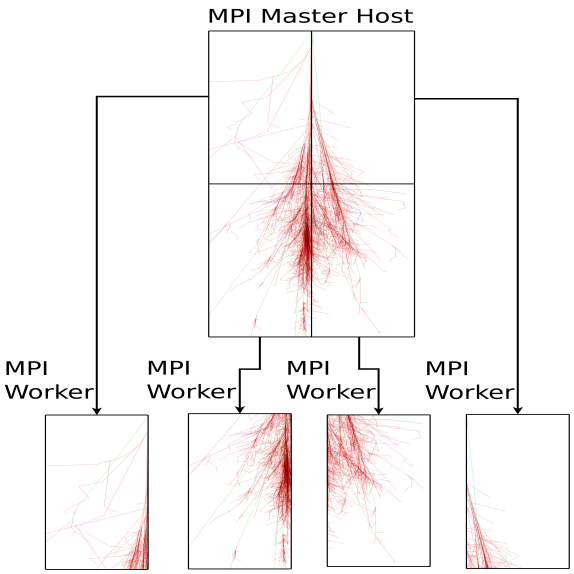


Figure 5: CORSIKA Proton Shower split into Sub-Showers. Particle Shower Picture from [1].

The approach is to place either single particles or whole sub-showers on different hardware architectures, which are the most suitable for the job at the moment of processing to minimize the processing time and the energy. Single particles distribution shows a huge amount of network traffic, so the overall execution time is worse than by placing a complete sub-shower. We are working on finding the right sub-shower size together with a hardware placement decision in order to minimize the energy amount.

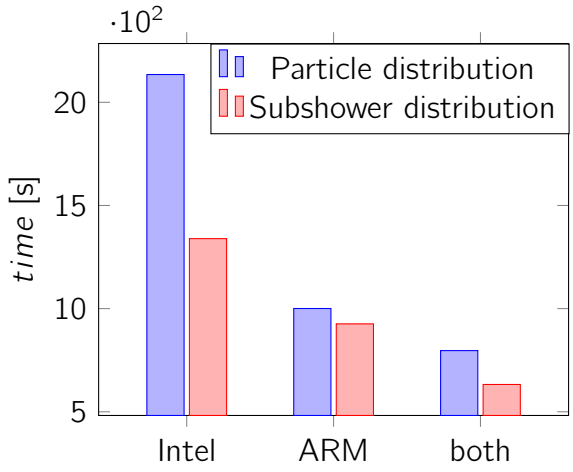


Figure 6: CORSIKA 8 Execution Time of MPI Distributed Events for Different Distribution strategies.

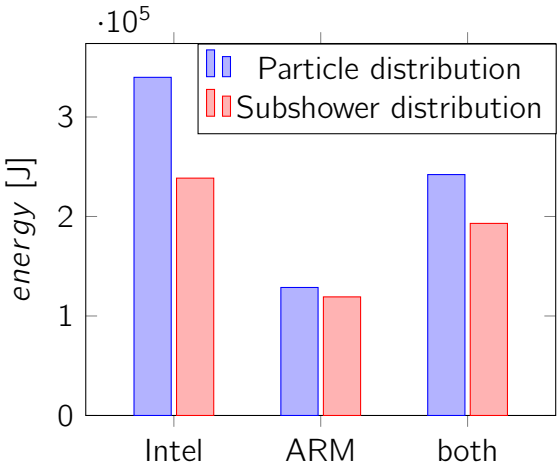


Figure 7: CORSIKA 8 Energy Consumption of MPI Distributed Events for Different Distribution strategies.

3 Conclusion and Future Work

Our experiments have shown that the ARM Cortex-A53 low power cluster has got a better energy efficiency than a state-of-the-art Xeon dual socket server system, while the Xeon Server has a better single-thread performance, the bigger number of low-power cores has an overall better execution time in all our experiments, depending on the assumption of parallelization. Moreover, there is still optimization potential of energy and time consumption by placing work at run time to the most suitable hardware. So we try to improve the EcoMaP approach to find the best trade-off between execution time and energy consumption.

References

[1] Karlsruhe Institute of Technology. CORSIKA-COsmic Ray SImulations for KAscade. `https://www.ikp.kit.edu/corsika/`.

[2] The LHCb collaboration. Technical Report LHCb Tracker Upgrade Technical Design Report. *CERN-LHCC-2014-001. LHCB-TDR-015*, Feb 2014.

Measurement of CP violation in $B^0 \to J/\psi\, K^0_S$ decay with $J/\psi \to e^+e^-$ and $K^0_S \to \pi^+\pi^-$

Gerwin Meier
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
gerwin.meier@tu-dortmund.de

The European Organization for Nuclear Research (CERN) is located near Geneva at the border between Switzerland and France. There, the largest machine of the world, the Large Hadron Collider (LHC) is operated. At the particle detection of the LHC, the aim is to study and constrain the Standard Model (SM), the best model for particle physics at the moment, and search for physics beyond the SM, so called "New Physics". The LHCb experiment is one of the four main experiments at the LHC and study high precision measurements with b- and c-Quarks. That includes charge and parity (CP) violating decays, one of three necessary conditions to explain the matter-antimatter asymmetry [4], and rare decays.

In proton-proton collisions, b quarks are produced in pairs mainly in a small cone of the beam line, and is therefore a single-arm forward spectrometer, as is shown in Fig. 1.

Closest to the beam is the Vertex Locator (VELO), which is used to identify the proton-proton collision point and the decay vertices of the produced particles with a small life time. The tracking turicensis (TT), the tracking stations (T1-T3) and the muon chambers (M1-M5) are used to reconstruct the tracks of the particles and determine the momentum with the bending of the magnet. The type of the particles is determined with the help of the ring-imaging cherenkov detectors (RICH1-RICH2) and the muon chambers. Finally the energy is measured with the calorimeter system, including the scintillating pad detector (SPD), preshower (PS), electromagnetic and hadronic calorimeter (ECAL, HCAL).

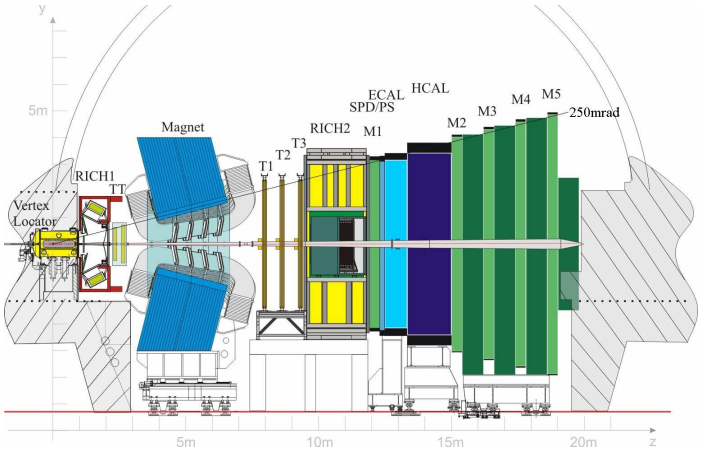


Figure 1: The LHCb detector with the various subdetectors for the identification of particles and reconstruction of their tracks.

There are several challenges for the reconstruction of all particles. In proton-proton collisions, many particles are produced and therefore there are thousands of hits in the detector, which make it very challenging to reconstruct the hundreds of tracks, because the number of possible combinations of the hits to form a track increase quadratically with the number of hits. Another challenge is the operation time of the LHC. The proton bunches collide every 25 ns and therefore a decision to save or reject the event, a so called trigger decision, has to be done very fast. The amount of data that needs to be recorded is very larg and a huge effort is made in finding ways to efficiently and safely store huge amount of data.

A measurement of time-dependent *CP* violation in the golden mode $B^0 \to J/\psi\, K^0_S$ is performed, to show the necessity of handling the data in an efficient way. The J/ψ is reconstructed with two electrons and the K^0_S with two oppositely charged pions. The analysis used the data collected with the LHCb detector in the second data period from 2015 to 2018, which doubles the number of signal candidates from the first period in 2011 and 2012 and increase the center-of-mass energy to $\sqrt{s} = 13\,\text{TeV}$. The analysis from the first period is already published in [1].

Neutral mesons like B^0 can oscillate into their antiparticle and the analysed finalstate $J/\psi\, K^0_S$ is common for both mesons. Therefore a decay-time-dependent *CP* asymmetry in the interference between the amplitudes of the direct decay and the decay after B^0-$\bar{B}^0$ mixing can be measured:

$$\mathcal{A}(t) \equiv \frac{\Gamma(\bar{B}^0(t) \to J/\psi\, K^0_S) - \Gamma(B^0(t) \to J/\psi\, K^0_S)}{\Gamma(\bar{B}^0(t) \to J/\psi\, K^0_S) + \Gamma(B^0(t) \to J/\psi\, K^0_S)} = \frac{S \sin(\Delta m\, t) - C \cos(\Delta m\, t)}{\cosh(\frac{\Delta\Gamma\, t}{2}) + A_{\Delta\Gamma} \sinh(\frac{\Delta\Gamma\, t}{2})}. \quad (1)$$

Here, $B^0(t)$ and $\bar{B}^0(t)$ indicate the flavour of the B meson at production, while t indicates the decay time. The parameters Δm and $\Delta\Gamma$ are the mass and the decay width differences

between the heavy and light mass eigenstates of the B^0–$\overline{B}^0$ system, and S, C, and $A_{\Delta\Gamma}$ are CP observables. As $\Delta\Gamma$ is negligible for the B^0–$\overline{B}^0$ system [2], the time-dependent asymmetry simplifies to $\mathcal{A}(t) = S\sin(\Delta m\, t) - C\cos(\Delta m\, t)$.

The first step in the analysis is a signal and background selection. One background component comes from wrong reconstructed particles, e.g. the J/ψ is reconstructed from two random electrons and not the two electrons, the J/ψ was decaying into. Another type of background are misidentified particles, e.g. one pion is in fact a kaon and built with the other pion a K^* instead of the K^0_S, or partially reconstructed backgrounds, physical decays where one or more particles are not reconstructed, like $B^0 \to J/\psi\, K^*(\to K^0_S(\to \pi^+\pi^-)\pi^0)$ where the neutral pion is lost. However, these backgrounds have different kinematical or topological properties. Therefore a cut based method on these variables is used at the beginning. Afterwards some special backgrounds are reduced as $\Lambda^0_b \to J/\psi\, \Lambda(\to p\pi)$, where the decay is refitted while one pion has a proton mass. Then near the Λ mass the probability of the pion candidate to be a proton has to be under 5%. Another background is $B^0 \to J/\psi\, K^*(892)^0(\to K^\pm\pi^\mp)$, where the kaon is misidentified as an pion. These background are reduced with an decay time requirement for the K^0_S to be more than 0.5 ps due to the fact, that the K^* is decaying almost immediately. Furthermore a background of $B^\pm \to J/\psi\, K^\pm$ is reduced, where one pion is randomly added to mimic the signal. These background are effectively reduced with a loose requirement that the pion candidate is not a kaon. The last step in the selection is a multivariate approach. A supervised learner is used, where a signal simulation is used for the signal proxy and wrong reconstructed data as a background proxy. The data is used to train a boosted decision tree (BDT) [3], which uses a couple of decision trees to decide for a candidate its signal likeliness. After the BDT is applied to the full dataset and determine the signal likeliness, a cut on this prediction is performed in a way, that as much signal candidates as possible survives and as much background candidates as possible are rejected.

After this huge effort to reduce the background there are still some background candidates remaining. Therefore the sFit method [5] is used to unfold the signal shape. That is done with a fit to the invariant mass of the B^0 meson. The signal and each background component is modelled with a function and then a fit to the data is performed. Afterwards the signal function parameters described the signal component very well, which can be seen in Fig. 2, and one obtains about 50000 signal candidates. The signal parameters are then used to calculate signal weights, which can be used to extract the signal out of all data in all variables, which are uncorrelated to the mass of the B^0 meson like the decay time, one of the parameters, which are needed to calculate the CP-parameters.

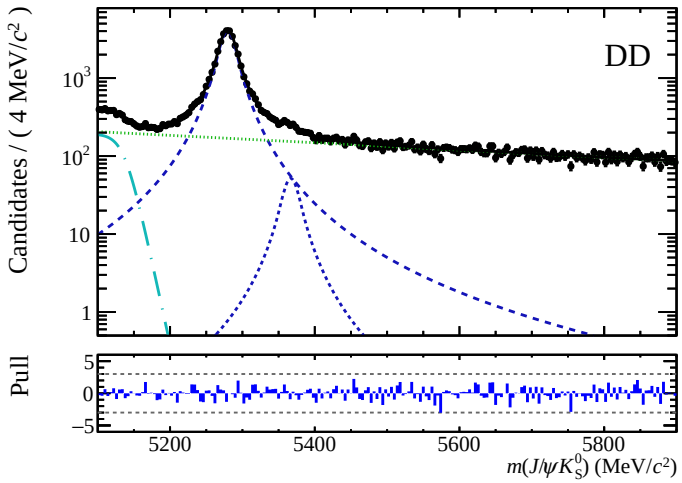


Figure 2: Massfit on selected data with the B^0 signal component in blue and long-dashed, the B^0_s background in blue dashed, the wrong reconstructed background in green and dashed and the partial background in cyan and long-dashed-dotted. The DD indicate, that the K^0_S decay after the vertex locator.

References

[1] R. Aaij et al. Measurement of *CP* violation in $B^0 \to J/\psi K^0_S$ and $B^0 \to \psi(2S)K^0_S$ decays. *JHEP*, 11:170, 2017.

[2] Y. Amhis et al. Averages of *b*-hadron, *c*-hadron, and τ-lepton properties as of summer 2016. *Eur. Phys. J.*, C77:895, 2017. updated results and plots available at `https://hflav.web.cern.ch`.

[3] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[4] A. D. Sakharov. Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe. *Pisma Zh. Eksp. Teor. Fiz.*, 5:32–35, 1967. [JETP Lett.5,24(1967); Sov. Phys. Usp.34,no.5,392(1991); Usp. Fiz. Nauk161,no.5,61(1991)].

[5] Yuehong Xie. sFit: A method for background subtraction in maximum likelihood fit. 2009.

GPU accelerators in the future LHCb-Triggersystem

Holger Stevens
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
holger.stevens@tu-dortmund.de

The LHCb experiment entered an upgrade phase in Dezember 2018, which will last 2 years. Not only the detector but also the computing farm will be changed. The planning for this started 10 years ago. A certain amount of computing power was planed to be bought by a fixed budget, but meanwhile "Moores Law died" so the plan cannot be converted. There a different solution approaches in the LHCb collaboration to reach the necessary computing power, one is the usage of GPUs.
Computing tasks can only benefit from the multiprocessor architecture of a GPU, if they are massively parallelizable. Different algorithms are running in the online farm to reconstruct the events. Most of them are independent and parallelizable. On the one hand it is the combination of hits to reconstruct the tracks on the other hand the processing of RAW information.
This report will describe the decoding of the so-called RAW banks of the SciFi tracker. These contain all information of the channels which detect a signal. The known channel IDs need to be translated to physical positions in the detector.

1 Introduction

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, as mentioned before an upgrade of the experiment is foreseen [6]. Most of the

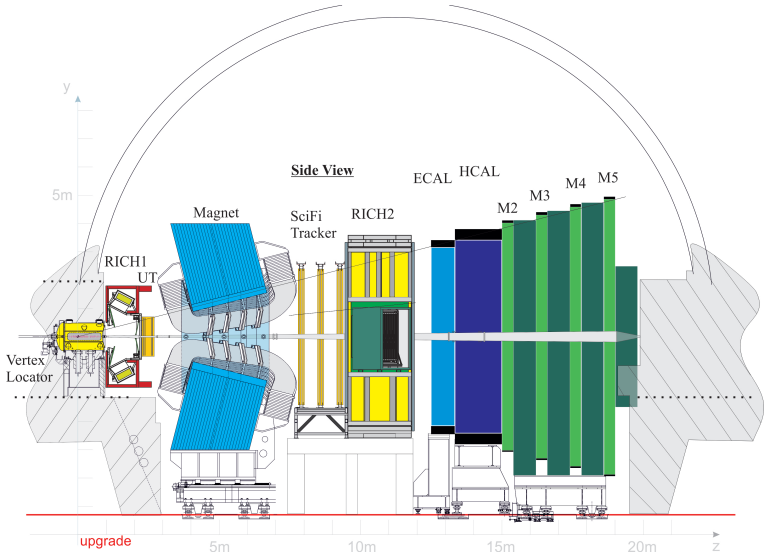


Figure 1: The LHCb upgrade detector with the various subdetectors for the identification of particles and reconstruction of their tracks [3].

existing detector will be replaced, but the general structure of the detector 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.
A major change is the new triggerless readout of the full detector with 40 MHz. This leads to a data rate of 40 Tb/s which is the input of the high level trigger(HLT) [7].

2 Details of the SciFi Tracker

Before the actual decoding part is described, some technical details of the SciFi Tracker are explained to understand the data format.
In total the Tracker consists of 3 stations with 4 layers each. These layers are build out of so-called modules, which again are composed of 8 fibre mats. Each mat is connected to 4 SiPM arrays. These are connected to the so-called FrontEnd electronics (FE). An FPGA based algorithm is looking for clusters of hits. A real signal has a different

width and height than a background hit, caused for example by thermal noise. Only the information from found clusters, so-called zero suppressed data, is sent from the FE to the BE (BackEnd electronics). The BE has to pack the data from several FE boards and send it to the computing farm where the actual decoding will take place. It is important to mention that the stations have a different width, T1 and T2 consist of 10 modules, and T3 is larger with its 12 modules. This difference leads to a branching of the decoding code, which causes a significant longer runtime.

3 Status and outlook

The aim of one working group of the LHCb collaboration is to run the first Trigger stage (HLT1) on GPU [5]. The project, Allen, can be found at `https://gitlab.cern.ch/lhcb/Allen`. Multiple streams can send data to the GPU. Then a visitor service will load the necessary Algorithms. One of these is the SciFi hit decoding. Every data-package has a global header which contains information about the associated proton-proton collision, like a timestamp. This information is followed by detector component-specific informations, for the SciFi this are the station number, MatID etc. with these information a position in the detector is unambiguously defined. The actual numbering scheme is constant for a specific time and can be a loud as a constant in the code [8]. During the development of the decoder, the data-format was changed to reduce the needed bandwidth. The amount of Hits is no longer in the header [4]. To allocate the correct size of memory for the result vector of decoded Hits, a pre count is necessary.

Successfully a GPU version of the SciFi decoder was implemented. Also, the other parts of the HLT1 sequence are ported to a GPU based version. The sequence can now be used to compare the performance to the baseline solution, which are standard x64 CPU server. The general idea and the status are presented in an article in *Computing and Software for Big Science* [2]. Different GPUs were tested to optimize the throughput of data, respectively Events. Figure 2 shows these performance studies. As the decision of using GPUs in LHCb or not wasn't taken yet, there is also no need to choose a specific GPU. It is possible to wait for the next generation ,which might have an even better cost/performance ratio. This can be expected as the performance of the sequence scales almost linear with the computing power (FLOPS) of the different GPUs.

References

[1] Letter of Intent for the LHCb Upgrade. Technical Report CERN-LHCC-2011-001. LHCC-I-018, CERN, Geneva, Mar 2011.

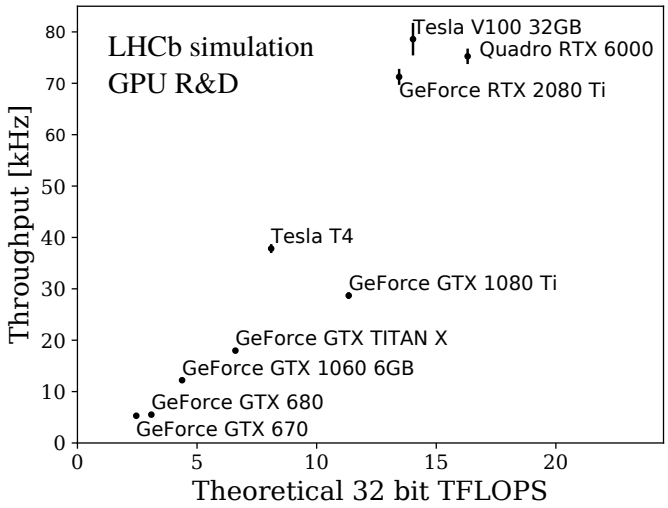


Figure 2: Allen throughput on various GPUs with respect to their reported peak 32-bit FLOPS performance.

[2] Roel Aaij et al. Allen: A high level trigger on GPUs for LHCb. 2019.

[3] LHCb Collaboration. LHCb Tracker Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-001. LHCB-TDR-015, Feb 2014.

[4] Sevda Esen, Jeroen Van Tilburg, Luigi Del Buono, Pierre Billoir, and Louis Henry. Clustering and rawbank decoding for the SciFi detector. Technical Report LHCb-INT-2018-024. CERN-LHCb-INT-2018-024, CERN, Geneva, Jul 2018.

[5] Stefano Gallorini, Donatella Lucchesi, Alessio Gianelle, Silvia Amerio, and Marco Corvo. First experiences with a parallel architecture testbed in the LHCb trigger system. Technical Report LHCb-PUB-2017-015. CERN-LHCb-PUB-2017-015. 3, CERN, Geneva, Apr 2017.

[6] Christian Joram. 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.

[7] CERN (Meyrin) LHCb Collaboration. Computing Model of the Upgrade LHCb experiment. Technical Report CERN-LHCC-2018-014. LHCB-TDR-018, CERN, Geneva, May 2018.

[8] Jeroen Van Tilburg. SciFi readout numbering scheme. Technical Report LHCb-INT-2016-044. CERN-LHCb-INT-2016-044, CERN, Geneva, Nov 2016.