SweDS 2017

The 5th Swedish Workshop on Data Science

Data Science Division
Department of Computer Science and Engineering
University of Gothenburg | Chalmers

at the

Wallenberg Conference Center
University of Gothenburg
December 12–13, 2017

Abstracts

Alexander Schliep (Editor)

INFORMATION AND COMMUNICATION TECHNOLOGY
A CHALMERS AREA OF ADVANCE

Overview

The SweDS 2017 workshop brought together researchers, practitioners, and opinion leaders with interest in data science. The goal was to further establish this important area of research and application in Sweden, foster the exchange of ideas, and to promote collaboration. It was the fifth workshop after very successful previous meetings held at the University of Borås, Stockholm University, Blekinge Institute of Technology and the University of Skövde. In 2017 there were four keynotes, two invited industry presentations and nine contributed plenary presentations selected from 28 submissions as well as 19 poster presentations.

Organizers

SweDS 2017 was organized by Devdatt Dubhashi, Richard Johansson, Graham Kemp and Alexander Schliep, who are faculty members of the Data Science Division at the Department of Computer Science and Engineering (CSE), which is a joint department of the University of Gothenburg and Chalmers University of Technology.

The workshop took place at the Wallenberg Conference Center of the University of Gothenburg.

Funding

The workshop was made possible due to generous funding from the ICT Area of Advance of Chalmers, the IT University of Gothenburg University and Chalmers, and the Applied Data Science Masters Program at Gothenburg University.

Panel discussion

The panel discussion on ”Data Science Education from Industry and Academia Perspectives” with Josephine Sullivan (KTH), Niklas Lavesson (BTH), Daniel Langkilde, Jan Wassén (Volvo), Richard Johansson (GU), Ellinor Rånge (Ericsson) was moderated by Peter Ljunglöf (CSE, Chalmers and GU).

Acknowledgements

The organizers would like to thank the invited speakers, plenary speakers, panel members, session chairs and poster presenters as well as all participants of the meeting. The organization has been made much easier due to support by Karin Nyström and Sanna Erling from the CSE department and Marcelo Valderrama and all other staff at the Wallenberg for administrative support.

Copyright

Invited Keynote and Industry Presentations

Cloud Control

Tue 12/12
13:15–14:05

Erik Elmroth
Umeå University and Elastisys

Semi-autonomous resource management systems are required to ensure robust, responsive, and cost-efficient cloud applications. This presentation will provide an overview of research challenges and efforts for transforming today's static and energy consuming cloud data centers into self-managed, dynamic, and dependable infrastructures. The ambition is to constantly deliver expected quality of service with acceptable operation costs and carbon footprint for large-scale services despite highly varying capacity demands. The presentation will outline the field of research and elaborate (by examples) on interdisciplinary approaches to solving resource allocation problems such as performance anomaly detection and bottleneck identification; workload analysis and prediction; vertical and horizontal autoscaling for throughput and (tail) latency control; replication optimization for availability control; decentralized scheduling; multi-dimensional server consolidation; optimizing live VM migration; dynamic resource rationing, to name a few. Target infrastructures span from individual datacenters and disaggregated systems to federated clouds and highly distributed edge clouds. In common to them all is the extreme scale and load variations in combination with the need for rapid management actions, to be determined based on immense amounts of monitoring data.

Decoding brain cancer drug responses

Tue 12/12
16:45–17:35

Sven Nelander
Uppsala University

In the last 10 years, large scale genetic investigations have revealed that human cancers are genetically complex. Application of algorithms to high dimensional cancer data can help reveal new disease subtypes, prognostic subgroups and help us gain insight into how cancer arises. But so far, it has been hard to connect cancer genetics data to actionable therapies: how to we know which molecular therapies to apply in which patients? To address this problem, our team has investigated a large collection of patient-derived brain tumor stem cells from Swedish patients. Unlike standard cancer genetics studies, our cell models provide us with a model to explore - on a large scale - functional differences in patient-specific biology. By combining multiple genomics platforms, drug screening and integrative data analysis, have thus found new functional subtypes of tumor cells and

more than 1500 new biomarkers of cancer drug response. As cancer research is turning into a data science, new integrative approaches of this type will be important, if not essential.

Data intensive computing research at SICS and KTH: towards scalable continuous deep analytics

Wed 12/13
9:00–9:50

Seif Haridi
KTH/SICS

We will present the research done at KTH/SICS on scalable platforms for data intensive computing platforms for deep analytics that requires processing large amount of data (Big Data) and scalable computation resources (Big Compute). In order to perform advanced analytics and machine learning at a scale there is a hierarchy of needs that should be fulfilled before doing advanced machine learning that is deployable in applications. These comprises storages and streams, resource management, data processing frameworks and tools for developing and deploying applications and sharing data-sets. In this context we present HOPS a complete platform for big data analytics, Apache Flink a stream processing framework for stream analytics. We also touch on our new project "Continuous Deep Analytics"

Working with data analytics in the telecommunication industry – my experience as a master thesis student and being part of a data analytics team

Wed 12/13
11:00–11:25

Ellinor Rånge
Ericsson

Around 40% of the world's mobile traffic runs through networks that are built by Ericsson. Mobile networks are constantly generating a large amount of data and Ericsson wants to use machine learning and artificial intelligence on this data for automation, insight and increased serviceability. A radio access network contains large numbers of radio base station configuration topologies. To explore methods on how to gain insight of configuration topologies present in radio access networks Ericsson had a master thesis project in spring 2017. In this thesis, we used graph-based machine learning methods with unsupervised learning, such as the Weisfeiler-Lehman subtree kernel. This kernel showed promising results on the given data set of configuration topologies as it managed to capture important differences between topological attributes. This presentation will cover parts of the master thesis, which led to a position at a data analytics team at Ericsson and what the team is working on today. We will also talk about the lessons learned transitioning from a master thesis student to a developer in a data analytics team and the challenges working with real-world data in one of the largest telecommunication companies in the world, including preparation of data. We will also mention other machine learning projects that we have at Ericsson and how we work at the Lindholmen site to share knowledge and insights between our teams.

Data intensive computing research at SICS and KTH: towards scalable continuous deep analytics

Wed 12/13
11:25–11:50

Nasser Mohammadiha
Zenuity

Autonomous Driving (AD) is currently one of the most challenging real-word problems with an expected high impact on our everyday lives. Self-driving cars will play a revolutionary role in traffic safety and sustainable mobility with major societal and individual benefits. Machine Learning (ML) is an important enabler to solve many AD-related problems; ML, and particularly deep learning (DL), is primarily used to get a description of the surrounding by detecting objects and road information, which is then utilized to develop AD functionalities. More recently, ML has been also used to develop methods for decision and control as well as tools for AD verification. In this presentation, we will give an overview of the main problems related to AD and discuss some ML-based solutions for some of these problems. In a broad classification, ML can be used to design modular systems, holistic and end-to-end systems, or systems with more affordable abstraction levels. In modular system designs, data flow is controlled explicitly so that the sensor signals are first processed, so far DL has mostly contributed here, and then fused to describe the surrounding environments in a similar way that we humans perceive. The obtained information is then used to implement required functionalities to control the vehicle, which itself could be based on ML or DL. End-to-end decision-making systems, on the other hand, rely on a holistic module that takes the unprocessed sensor signals as input and outputs the required signals to control the vehicle. In this case, the intermediate levels are abstracted and may not correspond to how humans perceive the environment. There are approaches for a third alternative also, where the level of the abstract representations is reduced to something that can be interpreted better by humans, but could still differ from how we primarily perceive the environment. The first two alternatives are the more dominant lines of research currently, and in this presentation, some recent works in these topics are reviewed. Clearly in a complete AD design, one could combine all these approaches in a unified system design to get the best out of all. Considering the verification aspects, data analysis and ML play an important role to develop efficient and effective verification strategies to (partially) overcome the need to drive hundreds of millions of miles needed for AD verification. This can be done by, e.g., developing methods and frameworks that enable the analysis of large amounts of data for active safety and AD. Such methods should for example describe and predict the sensor behavior in real traffic situations that are relevant to active safety and AD applications.

Transfer learning and some attempts to illuminate the black-boxes of deep convolutional networks

Wed 12/13
14:10–15:00

Josephine Sullivan
KTH

Deep convolutional networks (ConvNets) have been responsible for tremendous performance gains on many recognition tasks in computer vision. Perhaps even more significant is that these deep networks, trained on huge labelled image collections, learn a generic image representation. Thus one can take advantage of deep ConvNets, for many varied recognition tasks without large scale and expensive supervised training. But deep

ConvNets remain mainly regarded as impenetrable black-boxes and the community still does not have a compelling mathematical explanation of why they work so well or what the network has learnt. In this talk I will review, after an introduction to the latest developments in classification ConvNets and transfer learning, how the community has probed ConvNets to illuminate what they have learnt and also introduce some of our work at KTH to try and give a mathematical characterization of the pre-images of rectifier ConvNets.

Contributed Plenary Presentations

A Marine Spatial Data Infrastructure for Sweden

Tue 12/12
14:05–14:25

David Dodd
University of Gothenburg and IIT Technologies

A Marine Spatial Data Infrastructure (MSDI) is a framework consisting of marine geographic data, associated metadata, and the information technology infrastructure to enable the discovery and use of this data by the wider community. Nascent MSDIs traditionally consist of hydrographic data, as this data is immediately available and easily organized through hydrographic offices in standardized formats. However, an MSDI can and should include all data related to the marine environment, including maritime boundaries.

In Sweden the Swedish Meteorological and Hydrological Institute is hosting national marine monitoring data, except for marine geological data including sediment pollutants, which is hosted by the Swedish Geological Survey, and hydrographic data that is collected for nautical chart production by the Swedish Maritime Administration. From a management perspective, the Swedish Agency for Water and Marine Management (SWaM), is the primary stakeholder of marine data and responsible for implementing data in environmental analyses according to national and EU standards. This could concern evaluation of environmental status of different marine departments to monitor ecosystem health and ensure that measures are taken if ecosystem degeneration is discovered. It could also concern marine spatial planning of activities such as off shore installations, shipping, aquaculture and sea mining. To ensure long term resource efficiency and to facilitate future usage of marine data, both for management authorities and the scientific community, there is an urgent need for the development of a MSDI in Sweden.

A comprehensive MSDI provides easy, and controlled, access to all marine related georeferenced information. Data resides with the responsible agency, and accessed by others through an MSDI portal. “Others” includes government agencies as well as the general public. Users are assigned privileges that dictate the level of access and permissions for use of the information. With a functional MSDI, coastal zone managers, researchers, offshore developers and educators will have easy access to the ALL the information needed for sustainable use of the oceans, and provide the tools necessary for education and public outreach.

Key components of an MDSI include; a management system, internet access portal, and spatial data bases with metadata. The management system controls user access, monitors changes in the data by monitoring the metadata, and provides the link between users and the data. Vital to the success of a comprehensive MSDI management system is metadata. The management system can only access information that has up-to-date metadata in a standard format. An MSDI is not a data storage facility. Rather, it provides access to marine geospatial information stored by the proper authority.

This presentation provides an overview of an ideal MSDI, how it can be applied in Sweden, and examines some of challenges in its creation, use and management.

Joint work with: Ida-Maja Hassellöv (Chalmers University)

Tue 12/12
14:25–14:45

Recent developments on integral privacy

Navoda Senavirathne
Högskolan i Skövde

Development of data science and its related disciplines have provided us with an outstanding insight into data. Despite all the tempting benefits, it has also raised an important concern for data privacy. In the era of ”big data” this concern has become a major problem thus induce the requirement for new legislation; e.g., General Data Protection Regulation (GDPR) for EU.

Data privacy comprises the concepts, methods and tools that can be used to avoid the disclosure risk of sensitive information thus ensuring confidentiality. For the date, quite a few data privacy models have been introduced in the literature. A privacy model defines when a data set can be considered as protected and/or offer degrees of privacy. The definition of privacy models is a first step towards the definition of data protection mechanisms that are compliant with these models. Examples of privacy models include re-identification, k-anonymity and differential privacy. Nowadays there exist a plethora of data protection methods for each of these models. Different data protection methods compete on the type of data to be considered (e.g., databases, streaming data), the quality of the protected data (e.g., low information loss), the level of privacy achieved.

In a recent paper, we introduced the concept of integral privacy, which is based on the databases that are updated frequently. This privacy model becomes significant when data controllers have to adopt the new privacy regulations introduced. In this data-driven age, the new EU General Data Protection Regulation aims towards enhancing privacy and minimizing data breaches. This regulation improves rights of data subjects. One such vital introduction is ”the right to be forgotten”. This means a data subject has the right to request the data controller to erase his or her personal data, consequently ceasing further dissemination and halt processing of data by any third parties. This does not limit to deleting data from the source, rather we need to investigate the implications on aggregated data and inferences extracted from the original data. In data science point of view, this entails changes to the machine learning models built on the original data.

The definition of integral privacy is based on the idea that models inferred from a dataset should not allow disclosure of the training data or on how data has been updated (records deleted, records modified, etc.). Simply an adversary should not be able to compare different machine learning models built on different versions of a data set to extract knowledge on the modifications or underlying training data set. In this paper, we will present the privacy model and our latest results in this area especially with regarding analyzing the probability space of machine learning models.

Joint work with: Torra, V., and Navarro-Arribas, G.

Entity-based Data Federation with Privacy Concern

Tue 12/12
15:45–16:05

Lili Jiang
Umeå University

Motivated by the need for cross-database analysis on heterogeneous data and facilitating distributed data usage, we launched an academic project about privacy-aware data federation, which is supported by Umeå University on federated database research. The goal is exploring the scientific solutions for heterogeneous data federation and data privacy preservation. An infrastructure is thus being built to provide privacy-guarantee federated data for practical usage and research testbed.

Regarding data federation, the main challenges lie in processing federated database queries originate from the data distribution, heterogeneity and autonomy. We develop a data federation engine that provides users a unified interface to access multiple data sources and especially work on developing algorithms of the following three topics: (1) natural language query semantic parsing, (2) data heterogeneity resolution, and (3) personality prediction based differential privacy. Entity-based knowledge graphs are proposed to reveal the relations between different attributes and entities. Meanwhile, entity linking techniques are used for data heterogeneity resolution. To avoid privacy leakage, we develop a module of privacy preservation, which focuses on balancing the needs of the researchers to pursue scientific research as well as the privacy of individuals in the dataset. The datasets we are focusing on are registry data and social network data. The outputs of our work are planned to solve real challenges in disease prevention and analysis (e.g. cardiovascular) and health care (e.g. psychological).

In presentation, I would like to introduce the entire privacy-aware data federation infrastructure we are building. Especially I will describe the challenges and our correspondingly proposed solutions in heterogeneous data federation and privacy guarantee for data analysis.

Integrative analysis of genomic data: finding joint signal across subsets of groups of observations and variables

Tue 12/12
16:05–16:25

Jonatan Kallus
University of Gothenburg and Chalmers University of Technology

Integrative analysis of several related high-dimensional data sets is increasingly relevant in molecular biology (Richardson et al., 2016). This presentation will focus on genomic data where each subject is characterized by several high-dimensional features (gene expression, copy number aberration, methylation etc.) and subjects are divided among several classes according to disease subtype. Data sets of similar size and structure exist in e.g. image analysis and natural language processing. This structure calls for simultaneous integrative analysis both vertically (among groups of observations) and horizontally (among groups of variables). Two existing methods address general simultaneous vertical and horizontal integration: bi-modal OnPLS (Löfstedt et al., 2012) and linked matrix factorization (O'Connel and Lock, 2017). These methods aim at separating a signal that is global across all variable/observation groups from individual signals for each group. Already when the number of groups is greater than two, there may be signal that is common to subsets of groups but not common to all groups. Existing methods are not able to find such subset-common signal. This drawback becomes increasingly

relevant as the number of groups grows. We address the problem of finding which subsets of groups that have joint signal. Thus we address the harder problem of finding the jointness structure, as opposed to merely extracting the globally joint signal.

We define an objective function based on the singular value decomposition for each group. The optimum corresponds to low rank orthonormal bases for each group's row and column space. A reparametrization to angles reduces the number of parameters and avoids optimization constraints. We define a distance between orthonormal bases which is used in a fusing penalty term. A sparsity penalty is also added to increase interpretability of results. Global optimization is avoided since singular value decompositions give the unpenalized optimum. Small increments of the fusing and sparsity penalties yield a regularization path. Less computationally demanding approximative algorithms are also proposed and evaluated. The proposed method can be used for explorative analysis, e.g. integrative biclustering and driving combinations of variables across groups. Furthermore, the uncovered jointness structure can be used for improving separation of signal from noise. The application to genomic data suggests variables that may be of clinical importance and highlights similarities and differences between disease genotypes. Joint signal between different types of genomic data corresponds to connections between mutations, epigenetics and phenotype that may be mechanistic, and thus have relevance for development of new drugs.

Tue 12/12
16:25–16:45

Challenges in face expression recognition from video

Sergey Redyuk
Interaction Lab, University of Skövde

Identification of emotion from face expressions is a relatively well understood problem where state-of-the-art solutions perform almost as well as humans. However, in many practical applications, disrupting factors still make identification of face expression a very challenging problem. Within the project DREAM [1] - Development of Robot Enhanced Therapy for Children with Autism Spectrum Disorder (ASD), we are identifying face expressions from children with ASD, during therapy. Identified face expressions are used both in the online system, to guide the behavior of the robot, and off-line, to automatically annotate video for measurements of clinical outcomes.

This setup puts several new challenges on the face expression technology. First of all, in contrast to most open databases of face expressions comprising adult faces, we are recognizing emotions from children between the age of 4 to 7 years. Secondly, children with ASD may show emotions differently, compared to typically developed children. Thirdly, the children move freely during the intervention and, despite the use of several cameras tracking the face of the child from different angles, we rarely have a full frontal view of the face. Fourthly, and finally, the amount of native data is very limited.

Although we have access to extensive video recorded material from therapy sessions with ASD children, potentially constituting a very valuable dataset for both training and testing of face expression implementations, this data proved to be difficult to use. A session of 10 minutes of video may comprise only a few instances of expressions e.g. smiling. As such, although we have many hours of video in total, the data is very sparse and the number of clear face expressions is still rather small for it to be used as training data in most machine learning (ML) techniques.

We therefore focused on the use of synthetic datasets for transfer learning, trying to overcome the challenges mentioned above. Three techniques were evaluated: (1) convolutional neural networks for image classification by analyzing separate video frames,

(2) recurrent neural networks for sequence classification to capture facial dynamics, and (3) ML algorithms classifying pre-extracted facial landmarks.

The performance of all three models are unsatisfactory. Although the proposed models were of high accuracy, approximately 98%, while classifying a test set, they performed poorly on the real-world data. This was due to the usage of a synthetic dataset which had mostly a frontal view of faces. The models which "have not seen" similar examples before failed to classify them correctly. The accuracy decreased drastically when the child rotated her head or covered a part of her face. Even if the frame clearly captured a facial expression, ML algorithms were not able to provide a stable positive classification rate. Thus, elaboration on training datasets and designing robust ML models are required. Another option is to incorporate voice and gestures of the child into the model to classify emotional state as a complex concept.

Joint work with: Erik A. Billing (University of Skövde)

Detecting abnormal behavior of elderlies by analyzing energy consumption of individual households

Wed 12/13
9:50–10:10

Christian Nordahl
Blekinge Institute of Technology (BTH)

The elderly population is growing as years go by. Projections state that by year 2020, approximately 20% of the worlds population will be of age 60 or older. Providing assistance to the elderly population that wants to continue living at home has become an interesting area of research. Much of the focus has been towards ambient sensors in smart homes to identify abnormal behavior of its resident. However, installing a number of sensors in the households is tedious and could be seen as intrusive by the resident. We instead aim to detect abnormal behavior in a non-intrusive way, by only collecting the household's energy consumption data. We believe this is a valid approach because a major part of daily activities include the use of electrical equipment. Today's energy consumption meters in households are being upgraded to smart meters, and in Sweden we already have a 100% coverage. These smart energy meters allow for remote monitoring of a household's energy consumption with a high granularity, up to once per minute.

In our preliminary study, we acquired real world energy consumption data from 17 anonymous households in Sweden. The data was collected with a one minute interval and spanned over 2 months. We conducted an experiment in two parts, to investigate the accuracy when modeling the households normal behavior and the detection of abnormal behavior. In the first part of the experiment, we evaluated three different regression methods, six feature sets, and four data sets with different collection intervals. In the second part, we synthetically created two abnormal behaviors which we tried to detect using the regression models, feature sets, and data set intervals that were the most accurate in the first part of the experiment.

Results show that modeling normal behavior using regression on a household level is a viable approach. We end up with an average error rate between 18-25%. The results for the detection of abnormal behavior does, however, show that further study is required. The models misclassify many of the normal data points, up towards 40% in some cases.

This approach allows for an initial step of remote monitoring that is non-intrusive, of low cost, and easily deployable as smart meters are already underway. Currently, we are collaborating with the local eldercare and municipality to conduct a pilot study.

In this pilot study, we aim to collect energy consumption data from elderly households over six months and further study the possibility of this approach. We also also in the works of another study where we, instead of forecasting, cluster the consumers' energy consumption to identify consumption signatures and deviations from normal behavior.

Joint work with: Marie Persson, and Håkan Grahn (BTH)

Wed 12/13
10:10—10:30

Knowledge Structures for Explainable Machine Learning

Slawomir Nowaczyk
Halmstad University

It is a challenge to develop algorithms, methods and, ultimately, practical tools that can explain, in a clinical setting, the complex reasoning of a decision support system, based on data mining information from several sources of varying uncertainty. In order to create models that provide explainable decisions, one needs first to build a well-structured knowledge representation for the medical domain and then exploit it when learning predictive models and generating explanations.

Our claim is that the creation of Machine Learning (ML) models whose decisions are explainable can be facilitated by a Knowledge Structure (KS). This requires methods that: 1) automatically build a knowledge structure for a given domain, by fusing existing expert knowledge with several data sources; 2) generate understandable explanations for the decisions made by data-driven prediction models, based on such a knowledge structure; and 3) take advantage of the knowledge structure when building multiple decision models, for several different but related tasks.

In particular, one possible demonstration can focus on Congestive Heart Failure (CHF) patients, who contribute to approximately 20% of all of admissions and 12% of all emergency visits at Halland hospitals. The challenge in avoiding re-hospitalisations comes mainly from the difficulty in ensuring, through consistent patient follow-up, the maintenance of health through proper behaviours and medication adherence. The proposed solution will support clinicians and improve patient recovery by providing risk profiles for medical personnel, as well as personalised care coaching for patients, by combining three diverse data types: 1) information from Electronic Health Records (EHR), 2) clinical knowledge including regulations and recommendations, 3) surveys and wearable sensors readings.

Improved health outcomes can only be achieved by developing a decision support system for clinical applications which is able to communicate not only the indicated course of action, but also the reasoning behind suggestions and recommendations offered. This implies tracing the advice back to the information sources that led to a particular conclusion, together with a trail of the inference and how did the level of certainty evolve. In order to design systems capable of providing a justification, we need to determine what first principles, and relation to their data-driven counterparts, were considered when making a particular decision.

Doing this directly over a model which was learned purely from the data is not possible; an intermediate representation, connecting the expert knowledge and the data available, needs to be created as a foundation for the final ML predictor. We propose that such representations should be based on a KS that formally structures the data and defines relations between different influencing concepts. The main research questions are 1) to develop methods and tools for automatic or semi-automatic construction of a KS, and to

enrich it over time; 2) to combine data measurements with domain knowledge in a way that facilitates explanations, in particular in the face of data and model uncertainty; and 3) to demonstrate the generality of KS in supporting the creation of multiple models, for different but related task, that can share similar explainable decisions.

AVFDT: Adaptive Very Fast Decision Tree. Preliminary Results

Wed 12/13
13:30–13:50

Eva García-Martín
Blekinge Institute of Technology

Recent advancements in hardware together with the availability of large volumes of data has lead to an increase on the development of machine learning algorithms that build predictive models on those datasets. This trend is present in companies such as Google and Facebook, creating a challenging situation since machine learning algorithms account for a significant amount of the energy consumed in their data centers. For example, Google has created the Tensor Processing Unit (TPU) to speed up the computations of neural networks, by obtaining a 30X-80X gain in performance/Watt in comparison to CPUs and GPUs.

Machine learning algorithms are usually optimized towards scalability and predictive performance. This is also the case for state-of-the-art streaming algorithms. The algorithm investigated in this study is the Very Fast Decision Tree Algorithm, an online learning decision tree. The VFDT achieves competitive predictive performance results and scales well w.r.t. the number of instances. However, when profiling its energy consumption, we discover energy hotspots that could be addressed to reduce the overall energy consumption.

VFDT builds a tree incrementally as the data arrives. After nmin instances are observed at a node, the algorithm calculates if there is a clear attribute to make a confident split. Calculating the best attributes consumes high levels of energy. Since nmin is a fixed value and the same for each node, there are some nodes where nmin instances are not enough to create a split, thus computing those functions unnecessarily.

To address the mentioned hotspot, we present the Adaptive Very Fast Decision Tree algorithm (AVFDT); an energy efficient extension of the VFDT that adapts nmin based on the incoming data, independently for each node, to guarantee a split when the best attributes are computed. We evaluated the predictive performance and energy consumption of AVFDT against VFDT on seven datasets, with three different setups of nmin. The results show that AVFDT consumes up to 89% less energy than VFDT, trading off up to only 3% percent of accuracy. Averaging over all setups and all datasets, AVFDT consumes 23.5% less energy than AVFDT, sacrificing less than 1% of accuracy.

Our approach can be used to trade off energy consumption with predictive and computational performance in the strive towards resource-aware machine learning.

Wed 12/13
13:50–14:10

Exact inference for graphical models

Petter Mostad
Chalmers University of Technology

Graphical networks are central tools in traditional stochastic modelling. Such networks may often contain a mixture of discrete distributions and continuous distributions from exponential families. Given such a network, a common inference strategy is to use simulation, e.g., Gibbs sampling. We work on extending exact (i.e., non-stochastic) inference strategies to such networks. Preliminary results show that our algorithms may be faster and more accurate than, e.g., programs like Stan.

Poster Presentations

Functional Federated Learning in Erlang

Poster 1

Gregor Ulm
Fraunhofer-Chalmers Research Centre for Industrial Mathematics

A modern connected car produces gigabytes to terabytes of data per day. Collecting data generated by an entire fleet of cars, and processing it centrally on a server farm, is thus not feasible. The problem is that the total amount of data generated by cars, i.e. on edge devices, is too large to be efficiently transmitted to a central server. However, CPUs used in edge devices such as connected cars but also regular smart phones that connect to the cloud, have been getting more and more powerful in recent years. Tapping into this computational resource is one way of addressing the problem of processing big data that is generated by large numbers of edge devices.

One such approach consists of distributed data processing. Using the example of training an Artificial Neural Network, we introduce a framework for distributed data processing. A particular focus is on the implementation language Erlang. Arguably the biggest strength of the functional programming language Erlang is how straightforward it is to implement concurrent and distributed programs with it. Numerical computing, on the other hand, is not necessarily seen as one of its strengths.

The recent introduction of Federated Learning, a concept according to which edge devices are leveraged for decentralized machine learning tasks, while a central server only updates and distributes a global model, provides the motivation for exploring how well Erlang is suited to such a use case. We present a framework for Federated Learning in Erlang, written in a purely functional style. Erlang is used for coordinating data processing tasks but also for performing numerical computations. Initial results show that Erlang is well-suited for that kind of task. We provide an overview of the general framework and also discuss an existing and fully realized in-house prototypical implementation that performs distributed machine learning tasks according to the Federated Learning paradigm. While we focus on Artificial Neural Networks, our Federated Learning framework is of a more general nature and could also be used with other machine learning algorithms.

The novelty of our work is that we present the first publicly available implementation of a Federated Learning framework; our work is also the first implementation of Federated Learning in a functional programming language, with the added benefit of being purely functional. In addition, we demonstrate that Erlang can not only be leveraged for message passing but that it also performs adequately for practical machine learning tasks.

Our presentation is based on our work-in-progress paper “Purely Functional Federated Learning in Erlang”, which we presented at IFL 2017. The context of this research is our ongoing involvement in the Vinnova-funded project ”On-board/off-board distributed data analysis” (OODIDA), which is a joint-project between the Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Chalmers University of Technology, Volvo Car Corporation, Volvo Trucks, and Alkit Communications.

Joint work with: Emil Gustavsson and Mats Jirstrand

Poster 2

Anomaly Detection in Video Session Data

Shahrooz Abghari
Blekinge Institute of Technology

Online video service providers (OVSPs) continuously improve their services to satisfy the subscribers' expectation. This requires analysing massive amount of log files and different video event types. We use sequential pattern mining to analyse video data sequences to detect unexpected issues that can highly affect the subscribers' experience. The video session data has temporal order and contains detailed information regarding which video is requested, what type of device is used for watching the video, and the list of occurrences of all event types. The initial assumption with using sequential pattern mining is that most frequent sequential patterns (MFSPs) can be considered as normal system behaviour, while the others, non-most frequent sequential patterns (NMFSPs), can be potential anomalies. By performing clustering analysis, the MFSPs can be grouped based on their similarities. Finally, NMFSPs can be evaluated by the created model. The goodness-of-fit of the NMFSPs can be identified by applying an internal cluster validation measure such as Silhouette Index (SI). The proposed method has six steps as follows: 1) The video sessions are divided into equal-sized segments, e.g., daily. 2) The PrefixSpan algorithm is used to extract frequent sequential patterns. Such sequential patterns can lead us to detect collective anomalies, i.e., a collection of related data points (event types) assumed to be anomalous based on their occurrences together. 3) The extracted frequent sequential patterns are mapped with the video sessions and extra information related to date and time such as workday or weekend for finding contextual anomalies will be added to them. 4) The frequent sequential patterns are divided into two groups based on how frequent they are. Those patterns that occurred in more than one segment are named MFSPs with initial assumption that they are normal. The NMFSPs, on the other hand can be assumed as potentially anomalies. 5) MFSPs are clustered into partitions based on their similarities. 6) The clustering model built in the previous step is used to analyse the NMFSPs by matching each pattern into a cluster. To evaluate the goodness-of-fit of each NMFSP, SI is used. The SI has a range of [-1, 1]. A score 1 shows the NMFSP is assigned to a correct cluster. When score is about zero, this indicates that the NMFSP is on the decision boundary between two neighbouring clusters. Finally, a score close to -1 indicates the pattern is misclassified and assigned to an erroneous cluster, i.e., such NMFSP can be identified as anomaly. The proposed approach is applied on two months (October-November 2016) of data for a large OVSP company. The results show an increase in the number of quality adaptation events for many video sessions in both months. Such surge in the number of video streaming performance events during video sessions can be related to the fact that many viewers simultaneously try to watch the same video (e.g., a special live show) or an issue at the system level. In both cases, additional analysis by the company experts is needed for better understanding and interpretation of the results. *Joint work with: Veselka Boeva, Niklas Lavesson, Jörgen Gustafsson, Junaid Shaikh and Håkan Grahn*

Outlier Mining in Supervised Classification Problems

Poster 3

Veselka Boeva

Blekinge Institute of Technology

When the data are modeled using machine learning algorithms, the presence of noise and outliers can affect the model that is generated. Improving how learning algorithms handle noise and outliers can produce better models. In this study, we propose an outlier mining technique, entitled Cluster Validation Index (CVI)-based Outlier Mining, that is close to class outlier detection approaches which find suspicious instances taking into account the class label. The proposed approach identify and eliminate outliers from the training set, and a classification hypothesis is then built from the set of remaining instances.

Cluster validation measures are usually used for evaluating and interpreting clustering solutions in unsupervised learning. However, we apply these well-known and scientifically proven measures in a different context; we use them for filtering outliers in training sets in supervised learning scenarios. In supervised learning the clusters (in the form of classes) are known, and if there exists a strong relation among the instances of these clusters the classes of new instances can be accurately predicted. The intuition behind our approach is that instances in the training set that are not strongly connected to their clusters are outliers and should be removed prior to training to improve the classification performance of the classifier. Our approach assigns each instance in the training set several cluster validation scores representing its potential of being an outlier with respect to the clustering properties the used validation measures assess. In this respect, the proposed approach may be referred to a multi-criteria outlier filtering measure. Namely, it uses a combination of different cluster validation indices in order to reflect different aspects of the clustering model determined by the labeled instances of the training set.

We examine the effects of mining outliers for five commonly used learning algorithms on ten data sets from the UCI data repository using two different cluster validation indices (Silhouette Index and Connectivity). In addition, we study two approaches for filtering outliers: local and global. In case of local filtering we remove x percent from each class; in global filtering it is enough that we filter x percent from the entire training set. Our results show that for most learning algorithms and data sets, using the union of the two cluster validation indices and global filtering of outliers produces the greatest increase in classification accuracy.

Joint work with Lars Lundberg, Milena Angelova

Poster 4

Unbiased and efficient estimation with batch mode sampling in importance weighted active learning

Henrik Imberg

Matematiska vetenskaper - Chalmers tekniska högskola och Göteborgs universitet

Active learning is a family of prediction algorithms that iterate between data collection and model fitting, where the learner itself chooses the data from which it learns. These algorithms operate in a semi-supervised setting, where both labelled and unlabelled instances are available, by querying an oracle about the class membership of unlabelled instances. Active learning systems provide a cost efficient alternative to traditional passive learners as equal performance may be achieved with fewer labelled instances, and consequently at lower cost.

New instances have typically been selected by deterministic rules, resulting in a selection bias, and the reliability of predictions derived from such a procedure could be questioned. This selection bias may sometimes be neglected and sometimes adjusted for, but it is in general desirable to use methods that provide unbiased inferences.

A recent approach is to select new instances according to a random mechanism, where each element is assigned a unique probability of selection, so that the idea of active learning is utilized. Estimation can then be carried out by minimization of an unbiased estimator of the targeted loss function, by weighting the contribution of the labelled instances by the reciprocal of their corresponding sampling probabilities. Such procedures have been shown to yield approximately unbiased and consistent estimators, and label complexity bounds that outperform passive learners.

In our research, we investigate importance weighted batch mode active learning algorithms, where multiple elements are queried and labelled simultaneously. The model is updated after each new batch is drawn, and the updated model is used to construct a new sampling scheme in the next iteration.

Batch querying may be preferable to single element querying in settings where the cost of labelling not only depends on the number of labelled instances but also on the number of batches. It is also the case that single element querying tends to give highly variable estimates due to large importance weights, which is attenuated by batch querying. When batches have different sizes, we show how to optimally weight the contribution from each batch so that the variance is minimized. This may also be useful in single query algorithms for weighing the contributions from an initial sample and the actively queried instances, whenever such an initial sample is available.

We illustrate our methods on simulated and real traffic safety data, where the risk of a rear end collision is modelled as a function of glancing and speed and other covariates related to driver style and behaviour. Our aim is not only to make predictions of safety critical events, but also to interpret and make inference about the model parameters. The inferential framework we propose extends the traditional active learning framework in two ways. First, it allows for some of the covariates to be unknown and require annotation. Second, we allow for auxiliary information, not necessarily included in the target model, to be utilized in the sampling procedure.

Joint work with Olle Nerman, Marina Axelson-Fisk

Document Image Binarization Using Recurrent Neural Networks

Poster 5

Florian Westphal
Blekinge Institute of Technology

The readability of historical document images is a key issue for people wishing to read those documents, as well as for algorithms aiming to extract information from these images. Therefore, document image binarization, the separation of text foreground from page background is an important step to aid these attempts. However, this classification problem is a challenging task, due to common image degradations, such as faded ink, stains covering written text or text bleeding through from the other side of the page, as well as due to the large variability in appearance of those document images. In our work, we propose a binarization algorithm based on Grid Long Short-Term Memory (Grid LSTM) cells, which uses the capability of Grid LSTMs to perform multidimensional processing to take more context information about the currently binarized pixels into account. For the binarization, an input image is cut into square blocks and each of these blocks is read from all four corners by the first Grid LSTM layer. At each time step, a small block of the configured footprint size is read by the LSTM cells of the first input dimension. Simultaneously, the LSTM cells of the second input dimension read a footprint sized block of pixels above those of the first input dimension, while the LSTM cells of a third input dimension read another footprint sized block of a scaled down version of the surroundings of the pixels of the first input dimension. The output of this first Grid LSTM layer is then further processed by a second bi-directional Grid LSTM layer, whose output is ultimately converted into the binarized output by a fully connected output layer. Together with the binarization algorithm, we propose a dynamically weighted binary cross-entropy loss function for training. The dynamic weights take into consideration the impact certain labelling errors have on the readability of the binarized document image, penalizing those labelling errors more, which have a higher negative impact on readability. This follows the idea of pseudo F-Measure, a common measure for binarization quality of document images, and thus uses the same method as pseudo F-Measure to compute those weights for each training image. In this study, we analyze the impact of different footprint sizes, different scale factors for the third input dimension and different loss functions on the binarization quality. Additionally, we compare different footprint sizes with respect to their respective training and binarization time. Through our analysis, we show that the proposed binarization algorithm produces the best binarization results for a footprint size of 4 times 4 pixels, a scale factor of 2 and when trained using our proposed loss function. This configuration achieves the best average performance over all reported competition results of the 2016 binarization contest in 2 out of 4 evaluation measures used in this contest.

Joint work with Niklas Lavesson, Håkan Grahn

Poster 6

Architecture for Sustainable Enterprise Alignment of Model Driven Information Systems

Tomas Jonsson
Genicore AB

Model driven IT systems development with code generation is today used in a variety of, mostly technical, application areas. However, for large IT based Information Systems (ITbIS) it is still a rare an approach. An innovative case of low cost and high quality ITbIS is presented along with the model driven framework applied.

For over 20 years and still, the FMV (Swedish Defence Material Administration) ERP system [1] has been designed, extended and modified in pace with changes in the as well as in response to disruptive changes in information technology.

This innovative level of sustained enterprise alignment for FMV ERP is made possible by a coherent model driven approach with integrated method and tool support for the complete development and maintenance cycles of an ITbIS.

The purpose of an ITbIS is to provide meaningful information to people within or related to an enterprise. For information to be meaningful it must be relevant and understandable and thus related to peoples perception of enterprise “reality” which in turn is related to knowledge about the enterprise.

Core Enterprise Architecture Framework (CoreEAF) includes a multi perspective modelling method and framework, comprising three integrated perspective models. The perspectives are Phenomena Model (PM), View Model (VM) and Organization Model (OM).

PM is a phenomena model, based on a declarative object oriented expressional language with a combination of graphical and textual syntax. PM represents collective knowledge of enterprise phenomena, their relationships, their value attributes, the rules and calculations relating to these and if applicable, their value chains. The value chain concept is a state model guiding a goal driven rather than a process driven work system. Each node in a value chain represents a specific value state. E.g. the value states of an order could be tender state, confirmed order, delivered and paid.

When actors of an organization perform different activities, only specific fragments of the collective information is relevant. Therefore each actor needs their own perspective with a relevant choice of information content. Thus the VM defines perspectives of the PM relevant for the various actors and activities of the enterprise.

OM describes the organization structure, roles, groups, individuals and their relationships. OM is connected to VM by defining the relationship between roles and views, and to PM defining role responsibilities in relation to value chains.

Finally, model management tools and execution environment (CorePro) for CoreEAF models, support development and life cycle management of small and large information systems, sustainably aligned to the enterprise.

There is also CoreWeb, an easy to use web based tool and execution environment based on CoreEAF, intended for teaching and prototyping purposes, available to educational and research institutions.

Joint work with Håkan Enquist

Automatic blood glucose prediction with confidence using recurrent neural networks

Poster 7

Olof Mogren
CSE, Chalmers

Low-cost sensors and mobile platforms combined with machine-learning (ML) solutions enable personalized precision health and disease management. The large number of possible sensors, analysis tasks and adaptation to individuals raises the issue of scale, with respect to the number of ML systems which need to be developed. We present an approach for partially automated ML using deep learning in the context of managing diabetes. Continuous glucose monitor (CGM) systems help diabetics to closely follow their blood glucose values by storing a value every 5-15 minutes. They also provide an excellent source of data for predictions. This paper presents a model that can predict blood glucose levels for diabetics for up to two hours into the future. The solution uses a long short-term memory (LSTM) model which works on the raw data from a CGM device. The approach needs no feature engineering or data pre-processing, it obtains accuracy matching or exceeding the state-of-the-art, is computationally inexpensive, and provides along with the predictions an estimate of their variance, helping users to interpret the predicted levels.

The increasingly wide adoption of continuous blood glucose monitoring sys- tems (CGM) has given diabetics a valuable tool for closely monitoring and acting upon their current blood glucose levels and trends. While CGM devices typically store the glucose level once every 5-15 minutes, it can still be difficult for pa- tients to estimate how the glucose levels will develop in the nearest hours. Blood glucose levels adhere to complex dynamics that depend on many different vari- ables (such as carbohydrate intake, recent insulin injections, physical activity In this work, we present a novel approach using recurrent neural networks to predict blood glucose values based on the history of a patient. The system needs only the blood glucose history as input, and does not require the user to input insulin injections or meal intakes. In fact, our experiments show that such information did not help make better predictions with the proposed model. The model is trained on a patient's raw CGM data, with no pre-processing or feature-extraction needed, to predict future glucose values. The output from the system is modeled using a univariate Gaussian distribution, which means that along with the prediction, a user can get a sense of confidence in the predicted values. The proposed model outperforms state-of-the-art solutions in the CEGA metric, having 98% of all predictions in zone A+B (compared to 92% for the baseline). The model has only modest computational requirements, allowing deployment in mobile devices with no requirements of processing in the cloud.

Joint work with Christian Meijner, Simon Persson, Alexander Schliep and Björn Eliasson.

Poster 8

Enhancement of Energy Control Routing Protocol for Mobile Ad hoc Network Based on Hybrid Grey Wolf Optimizer with Ant Colony-based Energy Control Routing

Hasan Shakir
The National University of Malaysia

MANET is an autonomous collection of distributed mobile nodes. Every node in a MANET works as a source and a sink and that relays packets for other nodes. The key features of a MANET include dynamic network topology, distributed network nature, multi-hop communication, limited bandwidth, and limited energy constraints. Given that the battery of the nodes is limited, the energy of the nodes and the lifetime of network is a critical problem in MANETs. Moreover, nodes maintain static or less movement after being deployed. The energy of the MANET nodes cannot be recharged, which leads to dead nodes. This study improves the energy cost for the ACECR and boosts advancement through its contributions. Areas in the ad hoc network where much work is needed are discussed. This study only explored the impact of GWO on ACECR. Results indicate that ACECR-GWO performed better than the other protocols in terms of balanced energy consumption and extended network lifetime.

Poster 9

Locating CNV candidates in WGS data using wavelet-compressed Bayesian HMM

John Wiedenhoeft
CSE, Chalmers

The avalanche of NGS data and the growing demand for Bayesian methods pose huge algorithmic challenges in the case of whole-genome CNV inference. At the same time, fast, accurate and efficient computation is crucial in both clinical and fundamental research settings.

Recently, Wiedenhoeft, Brugel, and Schliep (2016) presented a method to drastically improve computation of full latent state marginals of Bayesian HMM in terms of speed and convergence behavior, but handling the memory requirements due to the sheer size of the input remained challenging.

We present an improved implementation of HaMMLET, a wavelet-compressed Forward-Backward Gibbs sampler for Bayesian HMM. We present a new data structure for dynamic compression, which can be constructed in-place and in linear time. We demonstrate its application for CNV inference on rat populations divergently selected for tame and aggressive behavior.

Joint work with Alexander Schliep.

Privacy Preserving Data Collection

Poster 10

Hamid Ebadi
Chalmers

Differential privacy assures the privacy of individuals by not revealing too much about them when their data is used in an aggregated statistical analysis. A differentially private data analysis usually works by the controlled addition of noise to the computation, aiming for enough to achieve a given degree of privacy at the same time as still giving statistically useful results.

In this poster we introduce a general framework targeting this local setting in which there is no centralized database or trusted curator, and the differential privacy mechanism must be applied at the data source.

Towards interactive correction of speech recognition errors

Poster 11

Peter Ljunglöf
CSE, University of Gothenburg and Chalmers University of Technology

In this project we explore how to make quick fixes to simple texts using as few interactions as possible. There are several situations where this could be useful, such as when you are driving (and don't have access to a keyboard), if your device is too small for a proper keyboard (such as a mobile phone), or if you have a communicative disability (e.g., cerebral palsy, visual impairment, or something else).

The main contribution of this work is about improving the online interaction for a user who wants to correct speech recognition errors. The actual error correction algorithm that we have used - based on the Levenshtein edit distance - has not been in focus. During spring 2018 the project will focus on improving the error correction algorithm.

Scan-o-matic: High-throughput quantitative phenotyping

Poster 12

Martin Zackrisson
Fraunhofer-Chalmers Research Centre for Industrial Mathematics

Scan-o-matic is an free and open source software suite and a methodology for performing automated high quality quantitative monitoring of growth of fungal or bacterial colonies, with better throughput than conventional methods. This enables us to accurately and precisely monitor the growth responses of microbial populations in various environments at an enormous scale. A standard flatbed film scanner measures the amount of light the colony is absorbing or scattering and through a calibration process this is converted to a precise estimation of the number of cells in the colony. With calibration and normalization to account for environmental differences across a plate, Scan-o-matic offers a inexpensive way to monitor the behaviour of many colonies of most cultivable colony forming microbes in parallel. For instance, the lab setup at Göteborgs Universitet can measure the growth of more than 220,000 colonies at the same time. Due to massive parallelisation of measurements, Scan-o-matic is a tool that can help us understand antibiotic resistance.

Joint work with Joakim Möller, Jonas Warringer, Anders Blomberg, and Andreas Skyman

---

Poster 13

Probabilistic Modelling of Sensors in Autonomous Vehicles

Edvin Listo Zec
Zenuity

The use of advance driver assistance systems (ADAS) is widely increasing today. Autonomous vehicles are equipped with different sensors such as camera, radar and lidar that gather a lot of information regarding the surroundings of the vehicle in order to make the best and safest decision.

Testing the quality of the sensors in autonomous vehicles is crucial for safety verification. This is usually done by collecting a lot of data in many different settings. However, this can be a very time consuming and expensive task and thus one is interested in realistic virtual verification methods that simulate these situations. By doing so, one is able to test many different scenarios without actual hazards more efficiently. Although it is useful to have an ideal model (ground truth), the drawback is that it completely disregards sensor noise and disturbance from the environment, which makes it near to impossible to evaluate and verify simulations. The objective is thus to probabilistically model the sensor behaviors in order to have as realistic simulations as possible.

For this purpose, a generative model has been created for the errors in production sensors used by Volvo Cars. The model is an extension to the hidden Markov model (HMM), called autoregressive input/output hidden Markov model (AIOHMM).

The main differences between the HMM and the AIOHMM are two-fold. The first being that we break the very restricting independence assumption between the observations by conditioning the output probabilities on previous outputs and also on inputs. By doing so, we are able to catch the autoregressiveness of the error and thus we are able to model the error more smoothly and catch long-term dependencies better. Secondly, a drawback of the regular HMM is that the transition matrix is time homogeneous. The AIOHMM breaks this homogeneity by conditioning the transition probabilities on an input vector. Thus, the transition probabilities are affected both by the previous state but also by the input at time t making them time inhomogeneous.

In this work we compare AIOHMMs to HMMs gradually by first only breaking the independence assumption between the observations and show that it is much better at preserving the long-term dependencies. Then we also add the time inhomogeneous transition matrix and show that it is important in order to catch the overall behavior of the sensor. Lastly, we use the Jensen-Shannon distance in order to evaluate how good the generative model is compared to a validation data set.

Joint work with Nasser Mohammadiha and Alexander Schliep

Transfer learning for Models of Human activity patterns

Poster 14

Rebeen Ali Hamad
Halmstad University

Machine learning algorithms classically rely on the assumptions of training data and testing data share the same feature space as well as residing from the same distribution. However, real-world applications often do not satisfy this assumption, such as when data is collected from different homes equipped with sensors in order to model human activity patterns of individual residents. In this aspect, the use of Transfer Learning (TL) could be considered. TL aims to leverage learning tasks for a target domain given learned knowledge from a different but related source domain. In this work we consider the learning of spatio-temporal human activity patterns from smart homes having different layouts, residents, and sensors. We hypothesize that even if these smart homes are unique their data share a common latent manifold which resides in a lower-dimensional subspace. This is explored by using Random Forest and t-SNE algorithms. For this, the use of TL is studied in order to develop a method for sharing knowledge across different physical homes to increase prediction accuracy and learning rate. Typical applications considered are deviation detection and pattern discovery for short-term and long-term planning of home care. Situation Awareness for Ambient Assisted Living (SA3L) is a project aimed at developing robust machine learning algorithms to improve the understanding of resident behavior patterns within smart home environments. One of the challenges relating to the modelling of activity patterns is to determine what knowledge should be transferred across different homes and residents to enhance the method's ability to generalize and to prevent negative transfer. Another challenge is to determine how the knowledge could be transferred from different homes. In this poster, the stability of t-SNE mapping of smart home data is investigated through the analysis of reproducibility of low-dimensional manifolds. The stability investigation is a key-step towards the goal which is to transfer knowledge across home domains. Manifolds are compared by the alignment of different runs of t-SNE using Procustes Analysis and the results from the ongoing work is presented

Joint work with Jens Lundström and Eric Järpe

Advanced Analytics Centre - A data science hub at AstraZeneca

Poster 15

Daniel Dalevi
AstraZeneca

The advanced analytics centre (AAC) at AstraZeneca is a problem-solving hub, working primarily in the late phase of drug development. The main focus is on generating support to data driven decision making in order to bring the right medicines to the right patients. AAC consists of 35 data scientists with different but overlapping skills, divided into five areas: Data science solutions, Decision sciences, Biomedical informatics, Statistical innovation and Health informatics. AAC supports the drug projects with advanced statistical expertise, interactive visualizations, machine learning competence and software solutions. Both internal and externa data sources, e.g. clinical trial data, operational data, competitive data, registry data, claims data, social media and observational data, are used.

In the poster we will show examples of both recent achievements and current challenges. The examples will include Innovative visualizations of safety data from clinical

trials, Central automated quality monitoring of hospitals globally, Predictive modelling of medicine supply, Real-time data analytics from biosensors and Synthetic control-arm models.

Joint work with Mattis Gottlow, Jesper Havsol and Martin Karpefors.

Poster 16

Toward Automatic Data-Driven Short-Term Traffic Prediction

Bin Sun
BTH

Here is a summary of the procedure from my PhD thesis to archive more accurate and more automatic machine learning based short-term traffic prediction. Short-term traffic prediction on freeways has been an active research subject in the past several decades. Various algorithms covering a broad range of topics regarding performance, data requirements and efficiency have been proposed. However, the implementation of machine learning based algorithms in traffic management centres is still limited. Two main reasons for this situation are, the data is messy or missing, and the parameter tuning requires experienced engineers. The main objective of this thesis was to develop a procedure that can improve the performance and automation level of short-term traffic prediction. Missing data is a problem that prevents many prediction algorithms in ITS from working effectively. Much work has been done to impute those missing data. Among different imputation methods, k-nearest neighbours (kNN) has shown excellent accuracy and efficiency. However, the general kNN is designed for matrix instead of time series so it lacks the usage of time series characteristics such as windows and weights that are gap-sensitive. We introduce gap-sensitive windowed kNN (GSW-kNN) imputation for time series. The results show that GSW-kNN is 34% more accurate than benchmarking methods, and it is still robust even if the missing ratio increases to 90Lacking accurate accident information (labels) is another problem that prevents huge amount of traffic data to be fully used. We improve a Mahalanobis distance based algorithm to be able to handle differential data to estimate flow fluctuations and detect accidents and use it to support correcting and complementing accident information. The outlier detection algorithm provides accurate suggestions for accident occurring time, duration and direction. We also develop a system with interactive user interface to realize this procedure. There are three contributions for data handling. Firstly, we propose to use multi-metric traffic data instead of single metric for traffic outlier detection. Secondly, we present a practical method to organise traffic data and to evaluate the organisation for Mahalanobis distance. Thirdly, we describe a general method to modify Mahalanobis distance algorithms to be updatable. For automatic parameter tuning, the experiments show that the flow-aware strategy performs better than the time-aware one. Thus, we use all parameter strategies simultaneously as ensemble strategies especially by including window in flow-aware strategies. Based on the above studies, we have developed online-orientated and offline-orientated algorithms for real-time traffic forecasting. The online automatic tuned version is performing near the optimal manual tuned performance. The offline version gives the performance that cannot be achieved using the manual tuning. It is also 3.05% better than XGB and 11.7% better than traditional SARIMA.

Breath-Sensing Robots: Looking Toward a Person using Change Point Detection on Privacy-Preserving Gas Sensor Data

Poster 17

Martin Cooney
Halmstad University

Loneliness has been described as a ”rising epidemic” which when prolonged can be a higher mortality risk than moderate daily smoking or obesity. To provide some comfort to lonely people, robots capable of interacting in a contingent and enjoyable way can be used; for example, by acknowledging and looking toward a person who is seeking to interact. A problem is that typical robot sensors such as cameras and microphones enable potential misusage of personal data. As an alternative to avoid this problem we propose an approach for using a common inexpensive gas sensor (MQ-135) to detect an interacting person's relative location via their breath. The algorithm combines a fast component using adaptive thresholds to recognize sudden large changes, and a more robust component to detect slight changes over longer times: the latter component applies the reweighted norm minimization version of the TREnd Filtering with EXponentials (rTREFEX) algorithm to detect change points between exponentials fit to the gas sensor readings. Our approach was evaluated by acquiring some initial data from seven participants (3 female, 4 male; age: 30.1 years, SD = 2.5) who interacted with a robot prototype we built. As a result, our assumption that more peace of mind would be felt with a gas sensor than typical sensors such as cameras or microphones was supported. Furthermore, basic feasibility of our approach was confirmed, with our prototype reacting to changes in a person's location on average in 6.5 - 7.7s. We believe that these results suggest the usefulness of considering some unconventional modalities toward designing robots which can be accepted in people's homes and that future work is required to investigate capabilities and limitations and improve speed and accuracy of detections and responses.

Joint work with Sepideh Pashami.

Centre for Automation and Integration Technology - a graduate school and an incubator for the new transport sector

Poster 18

Torsten Linders
University of Gothenburg

We take ocean observing as our starting point and example. This activity is now undergoing dramatic changes. Large manned research vessels are no longer the preferred carriers of observing sensors. Instead a multitude of automated systems have gained importance, including satellites, moorings, floats and self-propelled platforms. Most maturity in this field has the remote sensing by satellites. Simultaneously the output from ocean observing is also changing. The data used to consist of relatively simple sets, with a limited user group, well known by (or identical to) the group collecting the observations. Today the output is typically referred to as “products”, which are automatically integrated and distributed for diverse purposes and diverse user groups. Often it is anticipated that not all purposes or users can be known in advance. Most maturity in this field has the assimilation of observations into operational numerical forecasting models. Analysing ongoing changes of ocean observing we find that it is characterised by three developments:

- Automation - Integration - Big data These three developments are of course interdependent. Even more striking is that they could easily be used to describe the contemporary technological changes in many (most?) sectors of our society. One sector which arguably is greatly affected by the developments above is transportation. The transport industry is a traditionally important in Sweden. Its transformation is a challenge and an opportunity. As an effort to rise to this challenge we suggest the creation of a ”Centre for Automation and Integration Technology”. The Centre should initially consist of: - A graduate school - An incubator for start-up enterprises Fostering a new generation of people for a transformed Swedish transport industry must be done jointly between academia, industry and authorities. As indicated with the example of ocean observing, it can be anticipated that if the Swedish transport sector succeeds to transform, then the positive effect will spread and drive positive change in other sectors of Swedish industry.

Joint work with: Emil Gustavsson and Mats Jirstrand

Poster 19 **Data Integration Analysis for Supporting Decisions in Engineering Design**

Siva Krishna Dasari
Blekinge Institute of Technology

Aircraft designs are large-scale industrial projects involving multidisciplinary studies that address the behaviour of the design from mechanical, aerothermal and producibility aspects. Hence, it is a complex task characterised by large search spaces and high-level of uncertainty in requirements. The recent advancements in automation of CAD (Computer Aided Design) modelling and Finite Element Method simulations open to the possibility of embodying and assessing many design variants amongst a range of requirements early in the design phase. Despite the fact that computational tools and methods have been developed to support design engineers in decision making in early phases, the design of aircraft still remains complex to analyse due to multidisciplinary objectives resulting in long project times from the conceptual phase to the product delivery.

In our study, we aim to support the design of the Turbine Rear Structure (TRS) of an aero engine using machine learning techniques. Design exploration is divided into several studies evaluating different aspects of the proposed design, piece by piece unveiling behaviour and constraints. Each study has a design objective where the design engineers analyse specific parameters, for instance, the analysis of geometric configurations, such as engine mount positions, number of struts etc. These studies share a few common design parameters (for instance, thermal zones), however, they are focussed on different design objective of the TRS. The analysis of each design objective provides insight on how design parameters affect the performance of the engine in terms of quality, cost weight etc. However, it is difficult to explore how various design objectives of the TRS (studies) relate to each other. Therefore, there is a need to combine different studies (design objectives) into one design objective to understand the design space by analysing design parameters. This will be done by integrative analysis of simulation results (datasets: inputs and outputs) from two or more studies of the TRS. However, it is challenging to combine and analyse these datasets because the combined design objective (combined study) is not explored using simulations.

In this study, we plan to apply cluster analysis techniques to analyse design parameters of two design objectives (datasets from two studies of the TRS) separately. It is expected that performing analysis of studies separately may cause bias in conclusions

about design configurations. Thus, in the later phase, we focus on how to integrate the results from the analysis of those studies for more thorough understanding of design parameters. For this, we plan to conduct quantitative experiments to apply and evaluate cluster integrated techniques to analyse data from two design studies (from an aerospace industry) in order to extract valuable insights about design parameters to support decision making in engineering design.

Joint work with Veselka Boeva, Johan Wall, Niklas Lavesson and Petter Andersson.

Index of Presenting Authors