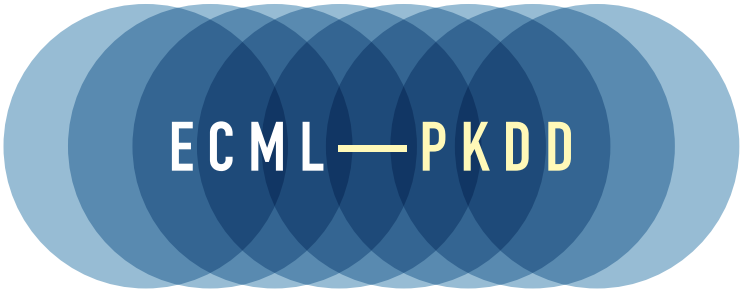


2016
SEPTEMBER 19 — 23
RIVA DEL GARDA,
ITALY

EUROPEAN CONFERENCE ON
MACHINE LEARNING AND PRINCIPLES
AND PRACTICE OF KNOWLEDGE DISCOVERY
IN DATABASES

SCHEDULE
SEPTEMBER 19 – 23 2016

MONDAY 19

08:30 – 8:30 - 9:00 Registration

09:00 – 9:00 - 10:40 Workshops | Tutorials | PhD forum

10:30 – 10:40 - 11:00 Coffe break

11:00 – 11:00 - 12:40 Workshops | Tutorials | PhD forum

12:30 – 12:40 - 14:20 Lunch

14:30 – 14:20 - 16:00 Workshops | Tutorials | PhD forum | Discovery Challenge

16:00 – 16:00 - 16:20 Coffe break

16:30 – 16:20 - 18:00 Workshops | Tutorials | PhD forum | Discovery Challenge

18:30 – 18:30 - 18:45 Opening

18:45 - 19:35 Invited talk: Alex: "Sandy" Pentland (1000A)

19:35 - 24:00 Welcome reception

TUESDAY 20

8:30 - 9:00 Registration

9:10 - 10:00 Invited talk: Susan Athey (1000A)

10:00 - 10:30 Best DM paper

10:30 - 11:00 Coffe break

11:00 - 13:00 Tue1A: Deep Learning and Neural Networks 1(1000A)
Tue1B: Graphs (1000B)
Tue1C: Clustering 1 (300A)
Tue1D: Learning (300B)
Nectar 1 (Belvedere)

13:00 - 14:50 Lunch

14:50 - 16:10 Tue2A: Classification 1 (1000A)
Tue2B: Optimization (1000B)
Tue2C: Topic Modelling (300A)
Tue2D: Patterns (300B)
Nectar 2 (Belvedere)

16:10 - 16:40 Coffe break

16:40 - 18:20 Tue3A: Kernels (1000A)
Tue3B: Probabilistic Learning (1000B)
Tue3C: Clustering 2 (300A)
Tue3D: Data Science (300B)
16:40 - 18:00 Nectar 3 (Belvedere)
18:00 - 18:16 Demo Spotlights

18:20 - 20:00 Demo and Poster session

WEDNESDAY 21

8:30 - 9:00
Registration

9:00 - 9:50
Invited talk: Zoubin Ghahramani
(1000A)

9:50 - 10:20
Test of time presentation

10:20 - 10:50
Coffe break

10:50 - 12:50
Plenary 1 (1000A)

12:50 - 14:40
Lunch

14:40 - 15:30
Invited talk: Rasmus Pagh
(1000A)

15:30 - 16:10
Plenary 2 (1000A)

16:10 - 16:40
Coffe break

16:40 - 18:00
Plenary 3 (1000A)

18:15 - 19:15
Community Meeting (1000A)

20:00 - 23:30
Gala dinner

THURSDAY 22

8:30 - 9:00
Registration

9:10 - 10:00
Invited talk: Thore Graepel
(1000A)

10:00 - 10:30
Best ML paper

10:30 - 11:00
Coffe break

11:00 - 13:00
Thu1A: Graphs and Social Networks 1 (1000A)
Thu1B: Reinforcement Learning (1000B)
Thu1C: Factorization (300A)
Thu1D: Streams and Time Series (300B)

13:00 - 14:50
Lunch

14:50 - 16:10
Thu2A: Classification 2 (1000A)
Thu2B: (Semi-)Supervised Learning (1000B)
Thu2C: Dimensionality (300A)
Thu2D: Patterns in Sequences (300B)

16:10 - 16:40
Coffe break

16:40 - 18:00
Thu3A: Graphs and Social Networks 2 (1000A)
Thu3B: Deep Learning and Neural Networks 2 (1000B)
Thu3C: Bandits & Transfer Learning (300A)
Thu3D: Recommendation (300B)
Thu3E: Mixed Grill (Belvedere)

18:00 - 18:15
Demo spotlights (Belvedere)

18:20 - 20:00
Demo and Poster session

FRIDAY 23

8:30 - 9:00
Registration

9:00 - 9:50
Invited talk: Ravi Kumar
(1000A)

10:00 - 11:20
Workshops | Tutorials | Industrial Track | Discovery Challenge

11:20 - 11:40
Coffe break

11:40 - 13:40
Workshops | Tutorials | Industrial Track | Discovery Challenge

13:40 - 14:40
Lunch

14:40 - 16:20
Workshops | Tutorials | Industrial Track

16:20 - 16:40
Coffe break

16:40 - 18:20
Workshops | Tutorials

16:40 - 17:55
Industrial Track

20:00 - ...:...
Farewell Party

Room	1000B	300A	300B	Belvedere	Stampa	100A	100B	Meeting	Presidenza
09:00	AALTD	NFMcp	Tutorial T3	Tutorial T1	MIDAS	MLSA	Tutorial T2	PhD forum	SSDM
12:40	Lunch								
14:20	AALTD	NFMcp	Tutorial T4	SoGood	MIDAS	MLSA	Discovery Challenges C1/C2	PhD forum	Tutorial T5
18:00									
18:45	Invited talk Alex 'Sandy' Pentland								
19:45	Welcome reception								
21:30									

Workshops

MLSA Machine Learning and Data Mining for Sports Analytics
AALTD 2nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data
MIDAS The First Workshop on MIning DAta for financial applicationS
NFmcp 5th International Workshop on New Frontiers in Mining Complex Patterns
SSDM The 2nd ECML/PKDD 2016 workshop on Statistically Sound Data Mining
SoGood Data Science for Social Good

Tutorials

T1 Preference-based pattern mining
T2 Machine Learning Meets Philosophy: From Epistemology to Ethics
T3 Comparing competing algorithms: Bayesian versus frequentist hypothesis testing
T4 Context Aware Knowledge Discovery: Opportunities, Techniques and Applications
T5 Learning from Hospital Data and Learning from Cohorts Discovery challenges

Discovery challenges

C1 Bank Card Usage Analysis
C2 SPHERE Challenge: Activity Recognition with Multimodal Sensor Data

FRIDAY SEPTEMBER 23 2016

Room	1000B	300A	300B	Belvedere	Stampa	100A	100B	Meeting	Presidenza
09:00	Invited talk Ravi Kumar								
09:50									
10:00	Industrial Track	Tutorial T6	Tutorial T7	DLPM	DMNLP	MML	DADA	MLLS	DARE
13:40	Lunch								
14:40									
14:40	Industrial Track	StreamEvolv	Tutorial T10	DLPM	Tutorial T9	MML	Discovery Challenges C3/C4	Tutorial T8	
18:20									
20:00	Farewall party								
...:...									

Workshops

DLPM	Deep Learning for Precision Medicine
MML	9th International Workshop on Music and Machine Learning
DMNLP	3rd Edition of the Workshop on Interactions between Data Mining and Natural Language Processing
DARE	4th International Workshop on Data Analytics for Renewable Energy Integration
MLLS	4th Workshop on Machine Learning in Life Sciences
DADA	1st International Workshop on Domain Adaptation for Dialog Agents
StreamEvolv	Learning from Data Streams in Large-Scale Evolving Environments: Challenges, Methods and Applications

Tutorials

T6	Large-scale Learning from Data Streams in Evolving Environments
T7	Data scientist's guide for writing papers
T8	An Introduction to Redescription Mining
T9	Probabilistic logics
T10	Learning Bayesian Networks for Complex Relational Data

Discovery challenges

C3	NetCla: The ECML-PKDD Network Classification Challenge
C4	cQA Challenge: Learning to Re-Rank Questions for Community Question Answering

GENERAL INFORMATION

INSURANCE

The organizers cannot be held liable for any injury, loss or damage occurring during the conference.

LUNCHES

Lunches are not provided. Close to the conference venue there is a wide variety of differently-priced bars and restaurants offering all kind of meals.

MOBILE PHONES

Please make sure to to switch off your mobile phones during the sessions.

ELECTRICITY

Italy adopts a 230 volt 50 Hz system.

SHOPPING

Most shops in Riva del Garda are open from 9:00 to 19:00 from Monday to Saturday.

SOCIAL EVENTS

WELCOME RECEPTION

Where: Congress Centre Garden
When: Monday, September 19th 2016, 19:30-24:00
The event is open to all participants.

CONFERENCE DINNER

Where: Arco Casino
When: Wednesday, September 21st, 20:00-23:30
Buses will leave from the parking lot in front of the Conference Centre at 19:30.
Please make sure to wear your badge to access the Gala Dinner..

CATERED EVENING POSTER SESSIONS

Where: Congress Center Palameeting
When: Tuesday, September 20th and Thursday, September 22nd, 18:20 20:00.
Poster sessions are open to all participants.

FAREWELL DINNER&PARTY

Where: Riva del Garda, spiaggia degli ulivi
When: Friday, September 23rd, from 20:00
The Farewell dinner&party ticket is mandatory to access the event.

INSTRUCTIONS FOR PARTICIPANTS

GENERAL INFORMATION

— For general enquiries or projector/computer/internet problems please contact the registration desk.
— Internet access: ***RivaFiereFree*** wireless network (no need for password)
— Conference app: we created an event for the conference on EventBase with all details about the program. You can browse it by downloading the EventBase app from the app store and searching for "ECML-PKDD 2016". Please contact paolo.dragone@unitn.it for further requests.
— Conference Matcher: To help you decide which sessions to attend you can use the ECML-PKDD 2016 Matcher tool at https://www.simsift.com/ecmlpkdd2016/. The tool lets you rank the papers at ECML-PKDD 2016 by their similarity to your own research by simply entering the URL of your homepage or online bibliography author page (DBLP recommended) or by entering some text or keywords representing your current interests.

INSTRUCTIONS FOR SPEAKERS

— Every room will have a digital projector and computer.
— Speakers can bring their own laptop.
— A technician will be available in every main session room. A volunteer will be available to help in every room.
— Please make sure to show up 10 minutes before the session starts to check the equipment works, and, if you will not be using your own laptop, to upload your slides from your USB drive.
— The time allocated to each speaker, including time to set up and time for questions, is:
  - research and nectar tracks: 20 minutes.
  - demo and phd forum tracks: 2 minutes.
  - industry track: 25 minutes.
  - workshops and discovery challenges: per workshop/challenge instructions.
— Should the session chair not be present, the last speaker of the session shall take over the session chair role.

INSTRUCTIONS FOR SESSION CHAIRS

— Please make sure to show up 10 minutes before the session starts to check the equipment works.
— Please stick to the schedule. If a speaker fails to show up, please announce a short break.
— Please moderate questions.
— Please do not start sessions or talks early as attendees may need some time to move between rooms.

INSTRUCTIONS FOR POSTER PRESENTERS

— Two poster sessions are scheduled:
  - one on Tuesday at 18:20-20:00 (for talks held on Tuesday)
  - one on Thursday at 18:20-20:00 (for talks held on Wednesday and Thursday)
— Both poster sessions will take place in the Palameeting of the conference center.
— Finger food will be served during the sessions
— Presenters should set up their on the day of the poster session, no later than the last coffee break , and remove it at the end of the session.
— Each presenter will be provided with a 1x2.4 mt poster board (width times height, portrait orientation recommended) and drawing pins.
— Poster boards will be grouped according to sessions.
— A volunteer will be available to assist you.

INSTRUCTIONS FOR SOFTWARE DEMOS

— Demo sessions will be held together with the poster sessions in the palameeting.
— Demonstration desks will be set up next to the poster area.
— Presenters should show up in advance to set up their demos, and remove them at the end of the session.
— Each demo will be provided with a desk, two chairs, and a (4 or 5 way) power socket.
— Wireless internet will be available.

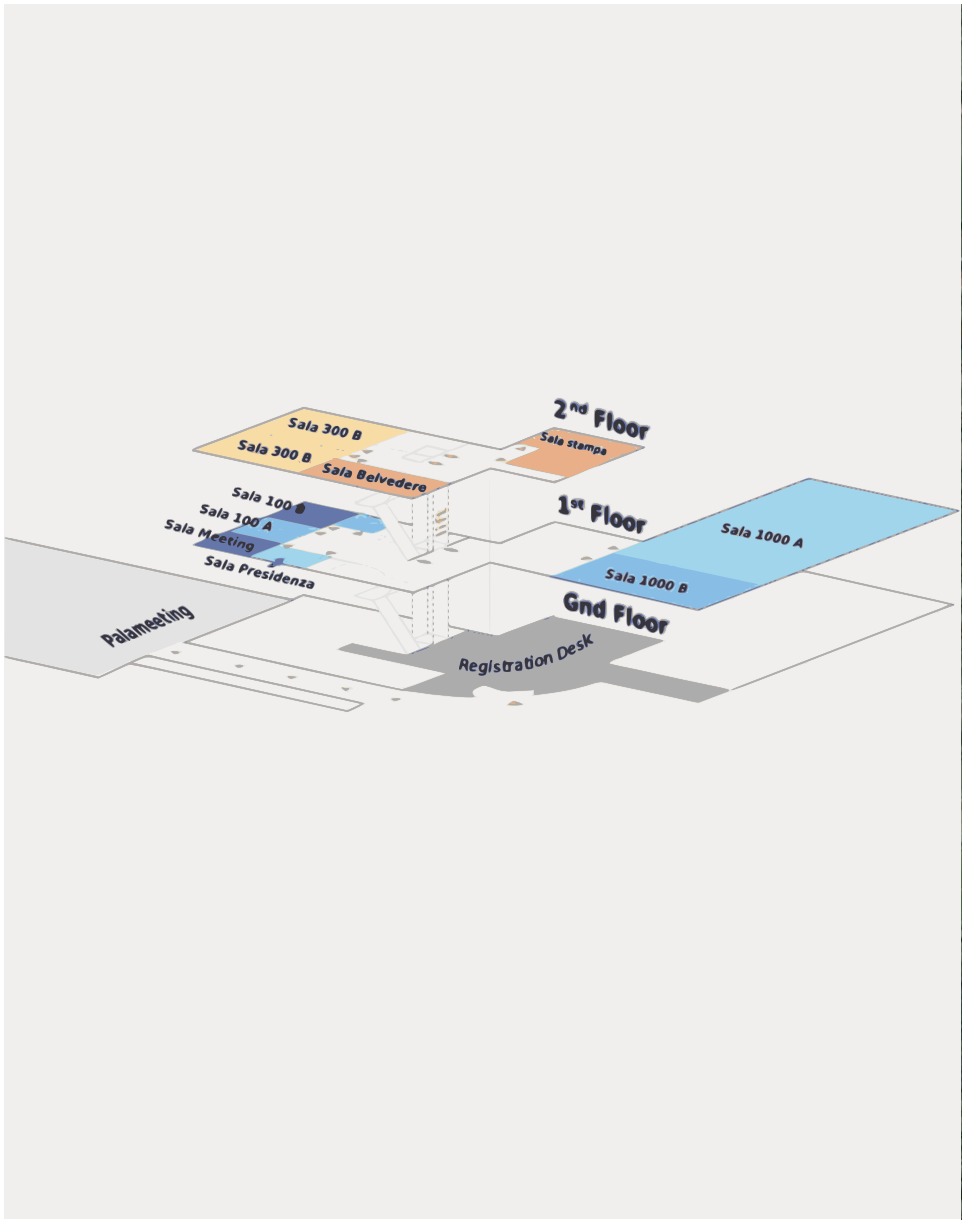
WELCOME

Dear Colleagues,
it is with great pleasure that we welcome you to ECML-PKDD 2016 in the beautiful setting of Riva del Garda.
It is a great honor to have organised this event, which is the result of the joint effort of a valuable team of chairs, volunteers and collaborators – we would like to thank each and all of them for the precious help, support and advice they have tirelessly provided over the last months.
ECML-PKDD has attracted nearly 600 participants this year, exceeding our own expectations. While we have tried our best to offer a memorable event, combining a wide and diversified range of scientific contributions, the presence of high-profile keynotes and an exciting social program, the numerous and valuable contributions we have received, as well as the high attendance rate, have made our work easier, interesting, and very enjoyable.
We are glad to inform you that this edition of ECML PKDD adheres to the Food for Good initiative, which promotes the redistribution of any food which is not consumed during coffee breaks and social events among charities, refugee centers and associations which provide support to people in need.
We would like to thank each of you for being here and hope you enjoy the conference and your stay in Riva!

General Chairs
Andrea Passerini
Fosca Giannotti

Program Chairs
Paolo Frasconi
Niels Landwehr
Giuseppe Manco
Jilles Vreeken

FLOOR PLANS


2nd Floor
Sala 300 B
Sala 300 B
Sala Belvedere
Sala stampa
1st Floor
Sala 100 B
Sala 100 A
Sala Meeting
Sala Presidenza
Sala 1000 A
Sala 1000 B
Palameeting
Gnd Floor
Registration Desk


Gala Dinner (Ar
Riva del Garda


co Casinò, ab. 6 km)
Welcome Reception (Gardens)
Congress Center
Farewell Dinner&Party (Spiagga degli Olivi)
Welcome Reception - DJ Set

KEYNOTE SPEAKERS

SUSAN ATHEY

The Economics of Technology Professor

Causal Inference and Machine Learning: Estimating and Evaluating Policies

Abstract:

In many contexts, a decision-making can choose to assign one of a number of "treatments" to individuals. The treatments may be drugs, offers, advertisements, algorithms, or government programs. One setting for evaluating such treatments involves randomized controlled trials, for example A/B testing platforms or clinical trials. In such settings, we show how to optimize supervised machine learning methods for the problem of estimating heterogeneous treatment effects, while preserving a key desiderata of randomized trials, which is providing valid confidence intervals for estimates. We also discuss approaches for estimating optimal policies and online learning. In environments with observational (non-experimental) data, different methods are required to separate correlation from causality. We show how supervised machine learning methods can be adapted to this problem.

About the speaker...

Susan Athey is The Economics of Technology Professor at Stanford Graduate School of Business. She received her bachelor's degree from Duke University and her Ph.D. from Stanford, and she holds an honorary doctorate from Duke University. She previously taught at the economics departments at MIT, Stanford and Harvard. In 2007, Professor Athey received the John Bates Clark Medal, awarded by the American Economic Association to "that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." She was elected to the National Academy of Science in 2012 and to the American Academy of Arts and Sciences in 2008. Professor Athey's research focuses on the economics of the internet, online advertising, the news media, marketplace design, virtual currencies and the intersection of computer science, machine learning and economics. She advises governments and businesses on marketplace design and platform economics, notably serving since 2007 as a long-term consultant to Microsoft Corporation in a variety of roles, including consulting chief economist.

ZOUBIN GHAHRAMANI

University of Cambridge, Alan Turing Institute

Automating Machine Learning

Abstract:

I will describe the "Automatic Statistician" (http://www.automaticstatistician.com/), a project which aims to automate the exploratory analysis and modelling of data. Our approach starts by defining a large space of related probabilistic models via a grammar over models, and then uses Bayesian marginal likelihood computations to search over this space for one or a few good models of the data. The aim is to find models which have both good predictive performance, and are somewhat interpretable. The Automatic Statistician generates a natural language summary of the analysis, producing a 10-15 page report with plots and tables describing the analysis. I will also link this to recent work we have been doing in the area of Probabilistic Programming (including an new system in Julia) to automate inference, and on the rational allocation of computational resources (and our entry in the AutoML conference).

About the speaker...

Zoubin Ghahramani FRS is Professor of Information Engineering at the University of Cambridge, where he leads the Machine Learning Group, and the Cambridge Liaison Director of the Alan Turing Institute, the UK's national institute for Data Science. He studied computer science and cognitive science at the University of Pennsylvania, obtained his PhD from MIT in 1995, and was a postdoctoral fellow at the University of Toronto. His academic career includes concurrent appointments as one of the founding members of the Gatsby Computational Neuroscience Unit in London, and as a faculty member of CMU's Machine Learning Department for over 10 years. His current research interests include statistical machine learning, Bayesian nonparametrics, scalable inference, probabilistic programming, and building an automatic statistician. He has published over 250 research papers, and has held a number of leadership roles as programme and general chair of the leading international conferences in machine learning including: AISTATS (2005), ICML (2007, 2011), and NIPS (2013, 2014). In 2015 he was elected a Fellow of the Royal Society.

THORE GRAEPEL

Google DeepMind and University College London

AlphaGo – Mastering the Game of Go with Deep Neural Networks and Tree Search

Abstract:

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play.
Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs and beat the human European Go champion Fan Hui by 5 games to 0, a feat thought to be at least a decade away by Go and AI experts alike. Finally, in a dramatic and widely publicised match, AlphaGo defeated Lee Sedol, the top player of the past decade, 4 games to 1. In this talk, I will explain how AlphaGo works, describe our process of evaluation and improvement, and discuss what we can learn about computational intuition and creativity from the way AlphaGo plays.

About the speaker...

Thore Graepel is a research group lead at Google DeepMind and holds a part-time position as Chair of Machine Learning at University College London. He studied physics at the University of Hamburg, Imperial College London, and Technical University of Berlin, where he also obtained his PhD in machine learning in 2001. He spent time as a postdoctoral researcher at ETH Zurich and Royal Holloway College, University of London, before joining Microsoft Research in Cambridge in 2003, where he co-founded the Online Services and Advertising group. Major applications of Thore's work include Xbox Live's TrueSkill system for ranking and matchmaking, the AdPredictor framework for click-through rate prediction in Bing, and the Matchbox recommender system which inspired the recommendation engine of Xbox Live Marketplace. More recently, Thore's work on the predictability of private attributes from digital records of human behaviour has been the subject of intense discussion among privacy experts and the general public. Thore's current research interests include probabilistic graphical models and inference, reinforcement learning, games, and multi-agent systems. He has published over one hundred peer-reviewed papers, is a named co-inventor on dozens of patents, serves on the editorial boards of JMLR and MLJ, and is a founding editor of the book series Machine Learning & Pattern Recognition at Chapman & Hall/CRC. At DeepMind, Thore has returned to his original passion of understanding and creating intelligence, and recently contributed to creating AlphaGo, the first computer program to defeat a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.

RAVI KUMAR

Google

Sequences, Choices, and their Dynamics

Abstract:

Sequences arise in many online and offline settings: urls to visit, songs to listen to, videos to watch, restaurants to dine at, and so on. User-generated sequences are tightly related to mechanisms of choice, where a user must select one from a finite set of alternatives. In this talk, we will discuss a class of problems arising from studying such sequences and the role discrete choice theory plays in these problems. We will present modeling and algorithmic approaches to some of these problems and illustrate them in the context of large-scale data analysis.

About the speaker...

Ravi Kumar has been a senior staff research scientist at Google since 2012. Prior to this, he was a research staff member at the IBM Almaden Research Center and a principal research scientist at Yahoo! Research. His research interests include Web search and data mining, algorithms for massive data, and the theory of computation.

RASMUS PAGH

IT University of Copenhagen

Dimensionality reduction with certainty

Abstract:

Tool such as Johnson-Lindenstrauss dimensionality reduction and 1-bit minwise hashing have been successfully used to transform problems involving very high-dimensional real vectors into lower-dimensional equivalents, at the cost of introducing a random distortion of distances/similarities among vectors. While this can alleviate the computational cost associated with high dimensionality, the effect on the outcome of the computation (compared to working on the original vectors) can be hard to analyze and interpret. For example, the behavior of a basic kNN classifier is easy to describe and interpret, but if the algorithm is run on dimension-reduced vectors with distorted distances it is much less transparent what is happening.

The talk starts with an introduction to randomized (data-independent) dimensionality reduction methods and gives some example applications in machine learning. Based on recent work in the theoretical computer science community we describe tools for dimension reduction that give stronger guarantees on approximation, replacing probabilistic bounds on distance/similarity with bounds that hold with certainty. For example, we describe a "distance sensitive Bloom filter": a succinct representation of high-dimensional boolean vectors that can identify vectors within distance r with certainty, while far vectors are only thought to be close with a small "false positive" probability. We also discuss work towards a deterministic alternative to random feature maps (i.e., dimension-reduced vectors from a high-dimensional feature space), and settings in which a pair of dimension-reducing mappings outperform single-mapping methods. While there are limits to what performance can be achieved with certainty, such techniques may be part of the toolbox for designing transparent and scalable machine learning and knowledge discovery methods.

About the speaker...

Rasmus Pagh graduated from Aarhus University in 2002, and is now a full professor at the IT University of Copenhagen. His work is centered around efficient algorithms for big data, with an emphasis on randomized techniques. His publications span theoretical computer science, databases, information retrieval, knowledge discovery, and parallel computing. His most well-known work is the cuckoo hashing algorithm (2001), which has led to new developments in several fields. In 2014 he received the best paper award at the WWW Conference for a paper with Pham and Mitzenmacher on similarity estimation, and started a 5-year research project funded by the European Research Council on scalable similarity search.

ALEX `SANDY' PENTLAND

MIT

Social Learning

Abstract:

Human decisions are heavily influenced by social interaction, so that predicting or influencing individual behavior requires modeling these interaction effects. In addition the distributed learning strategies exhibited by human communities suggest methods of improving both machine learning and human-machine systems. Several practical examples will be described.

About the speaker...

Professor Alex "Sandy" Pentland directs the MIT Connection Science and Human Dynamics labs and previously helped create and direct the MIT Media Lab and the Media Lab Asia in India. He is one of the most-cited scientists in the world, and Forbes recently declared him one of the "7 most powerful data scientists in the world" along with Google founders and the Chief Technical Officer of the United States. He has received numerous awards and prizes such as the McKinsey Award from Harvard Business Review, the 40th Anniversary of the Internet from DARPA, and the Brandeis Award for work in privacy.

He is a founding member of advisory boards for Google, AT&T, Nissan, and the UN Secretary General, a serial entrepreneur who has co-founded more than a dozen companies including social enterprises such as the Data Transparency Lab, the Harvard-ODI-MIT DataPop Alliance and the Institute for Data Driven Design. He is a member of the U.S. National Academy of Engineering and leader within the World Economic Forum.

INDUSTRY INVITED TALKS

MICHAEL MAY

Siemens Corporate Technology, Munich

Towards Industrial Machine Intelligence

Abstract:

The next decade will see a deep transformation of industrial applications by big data analytics, machine learning and the internet of things. Industrial applications have a number of unique features, setting them apart from other domains. Central for many industrial applications in the internet of things is time series data generated by often hundreds or thousands of sensors at a high rate, e.g. by a turbine or a smart grid. In a first wave of applications this data is centrally collected and analyzed in Map-Reduce or streaming systems for condition monitoring, root cause analysis, or predictive maintenance. The next step is to shift from centralized analysis to distributed in-field or in situ analytics, e.g in smart cities or smart grids. The final step will be a distributed, partially autonomous decision making and learning in massively distributed environments.

In this talk I will give an overview on Siemens' journey through this transformation, highlight early successes, products and prototypes and point out future challenges on the way towards machine intelligence. I will also discuss architectural challenges for such systems from a Big Data point of view.

About the speaker...

Michael May is Head of the Technology Field Business Analytics & Monitoring at Siemens Corporate Technology, Munich, and responsible for eleven research groups in Europe, US, and Asia. Michael is driving research at Siemens in data analytics, machine learning and big data architectures. In the last two years he was responsible for creating the Sinalytics platform for Big Data applications across Siemens' business.

Before joining Siemens in 2013, Michael was Head of the Knowledge Discovery Department at the Fraunhofer Institute for Intelligent Analysis and Information Systems in Bonn, Germany. In cooperation with industry he developed Big Data Analytics applications in sectors ranging from telecommunication, automotive, and retail to finance and advertising.

Between 2002 and 2009 Michael coordinated two Europe-wide Data Mining Research Networks (KDNet, KDubiq). He was local chair of ICML 2005, ILP 2005 and program chair of the ECML/PKDD Industrial Track 2015. Michael did his PhD on machine discovery of causal relationships at the Graduate Programme for Cognitive Science at the University of Hamburg.

MATTHIAS SEEGER

Amazon, Berlin

Machine Learning Challenges at Amazon

Abstract:

At Amazon, some of the world's largest and most diverse problems in e-commerce, logistics, digital content management, and cloud computing services are being addressed by machine learning on behalf of our customers. In this talk, I will give an overview of a number of key areas and associated machine learning challenges.

About the speaker...

Matthias Seeger got his PhD from Edinburgh. He had academic appointments at UC Berkeley, MPI Tuebingen, Saarbruecken, and EPF Lausanne. Currently, he is a principal applied scientist at Amazon in Berlin. His interests are in Bayesian methods, large scale probabilistic learning, active decision making and forecasting.

SCHEDULE SEPTEMBER 19 – 23 2016

MONDAY 19

TUESDAY 20

WEDNESDAY 21

THURSDAY 22

FRIDAY 23

MONDAY 19 SEPTEMBER 2016

TUTORIALS - MORNING

COMPARING COMPETING ALGORITHMS: BAYESIAN VERSUS FREQUENTIST HYPOTHESIS TESTING

Organizers: Giorgio Corani, Alessio Benavoli, Janez Demsar
Time: 09:00 - 12:40
Room: 300B

Hypothesis testing in machine learning – for instance to establish whether the performance of two algorithms is significantly different – is usually performed using null hypothesis significance tests (nhst). Yet the nhst methodology has well-known drawbacks. For instance, the claimed statistical significances do not necessarily imply practical significance. Moreover, nhst cannot verify the null hypothesis and thus cannot recognize equivalent classifiers. Most important, nhst does not answer the question of the researcher: which is the probability of the null and of the alternative hypothesis, given the observed data? Bayesian hypothesis tests overcome such problems. They compute the posterior probability of the null and the alternative hypothesis. This allows to detect equivalent classifiers and to claim statistical significances which have a practical impact. We will review Bayesian counterpart of the most commonly test adopted in machine learning, such as the correlated t-test and the signed-rank test. We will also show software implementing such test for the most common platforms (R, Python, etc.)

MACHINE LEARNING MEETS PHILOSOPHY: FROM EPISTEMOLOGY TO ETHICS

Organizers: Marcello Pelillo, Teresa Scantamburlo
Time: 09:00 - 12:40
Room: 100B

In recent years there has been a revival of interest around the philosophical issues underpinning machine learning research, from both the computer scientist's and the philosopher's camps. This suggests that the time is ripe to attempt establishing a long-term dialogue between the two communities with a view to foster cross-fertilization of ideas. The goal of this tutorial is to provide a timely and coherent picture of the state of the art in the field and to stimulate a discussion and a debate within our community. This could be an opportunity for reflection, reassessment and eventually some synthesis, with the aim of providing the field a self-portrait of where it currently stands and where it is going as a whole. Our discussion will be focused on both the epistemological and the ethical perspectives. On the one hand, we will explore the problem of induction and causality around which machine learning and philosophy have built some of the most intriguing and passionate discussions, offering a number of insights that are still inspiring today's research (e.g., inductivism vs. falsificationism, Bayesianism, model selection, etc.). On the other hand, we will consider how machine learning is raising profound philosophical questions concerning the ethical implications of data-driven decision making, which reminds us of Wiener's visionary lesson. As we shall see, specific causes for concern include the opacity of learning algorithms (transparency and accountability), the potential discriminative effects of algorithmic inference (fairness) and the disclosure of personal data (privacy). We shall assume no pre-existing knowledge of philosophy by the audience, thereby making the tutorial self-contained and understandable by a non-expert.

PREFERENCE-BASED PATTERN MINING

Organizers: Bruno Cremilleux, Marc Plantevit, Arnaud Soulet
Time: 09:00 - 12:40
Room: Belvedere

This tutorial focuses on the recent shift from constraint-based pattern mining to preference-based pattern mining. Constraint-based pattern mining is now a mature domain of data mining that makes it possible to handle various different pattern domains (e.g., itemsets, sequences, graphs) with a large variety of constraints thanks to solid theoretical foundations and an efficient algorithmic machinery. Even though, it has been realized for a long time that it is difficult for the end-user to model her interest in term of constraints and above to overcome the well-known thresholding issue, researchers have only recently intensified their study of methods for finding high-quality patterns according to the user's preferences. In this tutorial, we discuss the need of preferences in pattern mining, the principles and methods of the use of preferences in pattern mining. Many methods are derived from constraint-based pattern mining by integrating utility functions or interestingness measures as quantitative preference model. This approach transforms pattern mining in an optimization problem guided by user specified preferences. However, in practice, the user has only a vague idea of what useful patterns could be. The recent research field of interactive pattern mining relies on the automatic acquisition of these preferences.

TUTORIALS - AFTERNOON

CONTEXT-AWARE KNOWLEDGE DISCOVERY: OPPORTUNITIES, TECHNIQUES AND APPLICATIONS

Organizers: Cesar Ferri, Peter Flach, Meelis Kull, Nicolas Lachiche
Time: 14:20 - 18:00
Room: 300B

Traditionally, knowledge discovery aims to learn patterns from given data that are expected to apply to future data. Often there is relevant contextual information that, although it can have a considerable effect on the quality of the results, is rarely taken into account as it is not directly represented in the training data. This tutorial aims to elucidate the role of context and context changes in the knowledge discovery process, and to demonstrate how recent research advances in contextaware machine learning and data mining can be put to practical use. The tutorial will cover the main types of context changes, including changes in costs, data distribution and others. Participants will develop basic skills in choosing the appropriate modelling techniques and visualisation tools for the construction, selection, adaptation and understanding of versatile and context aware models. We will discuss how data mining methodologies such as CRISPDM can be extended to take context change into account. The tutorial will not only equip the attendees with new technical and methodological knowledge, but also encourage an anticipatory attitude towards context change.

LEARNING FROM HOSPITAL DATA AND LEARNING FROM COHORTS

Organizers: Panagiotis Papapetrou, Myra Spiliopoulou
Time: 14:20 - 18:00
Room: Presidenza

Data mining is intensively used in medicine and healthcare. Electronic Health Records (EHRs) are perceived as big medical data. On them, scientists strive to perform predictions on patients' progress while in the hospital, to detect adverse drug effects, and to identify phenotypes of correlated diseases (as they occur in a hospital), among other learning tasks. Next to EHRs, medical research is no less interested in learning from cohort data, i.e., from a carefully selected set of persons with and without the outcome under observation. From these data, which are small in numbers but have a big number of dimensions, scientists want, e.g., to predict how people with and without a disease evolve, to assess how they respond to a treatment, and to identify phenotypes of a disease as it occurs in the population. In this tutorial, we discuss learning on hospital data and learning on cohorts. We begin by introducing key terms and then discuss example objectives for mining on hospital data and on cohort data. Then, we focus on specific application areas. For the cohort data, we present examples of exploratory analysis on population-based and clinical studies, with emphasis on the role of time in these studies. For the hospital data, we present examples of learning from time-stamped data, heterogeneous data, and then focus on the problem of discovering adverse drug effects.

MONDAY 19 SEPTEMBER 2016

WORKSHOPS – FULL DAY

MLSA: MACHINE LEARNING AND DATA MINING FOR SPORTS ANALYTICS

Organizers: Jan Van Haaren, Mehdi Kaytoue, Jesse Davis
Time: 09:00 - 18:00
Room: 100A
Website: https://dtai.cs.kuleuven.be/events/MLSA16/index.php

Sports Analytics has been a steadily growing and rapidly evolving area over the last decade, both in US professional sports leagues and in European football leagues. The recent implementation of strict financial fair-play regulations in European football will definitely increase the importance of Sports Analytics in the coming years. In addition, there is the popularity of sports betting. The developed techniques are being used for decision support in all aspects of professional sports, including:

- Match strategy, tactics, and analysis
- Player acquisition, player valuation, and team spending
- Training regimens and focus
- Injury prediction and prevention
- Performance management and prediction
- Match outcome and league table prediction
- Tournament design and scheduling
- Betting odds calculation

The interest in the topic has grown so much that there is now an annual conference on Sports Analytics at the MIT Sloan School of Management, which has been attended by representatives from over 70 professional sports teams in eminent leagues such as the Major League Baseball, National Basketball Association, National Football League, National Hockey League, Major League Soccer, English Premier League, and the German Bundesliga. Furthermore, sports data providers such as OPTA have started making performance data publicly available to stimulate researchers who have the skills and vision to make a difference in the sports analytics community. Traditionally, the definition of sports has also included certain non-physical activities, such as chess – in other words, games. Especially in the last decade, so-called e-sports, based on a number of computer games, have become very relevant commercially. Professional teams have been formed for games such as Starcraft 2, Defense of the Ancients (DOTA) 2, and League of Legends. Moreover, tournaments offer large sums of prize money and are important broadcast events. Given that topics such as strategy analysis and match forecasting apply in equal measure to these new sports (and other topics might apply as well but are not very well explored so far), and data collection is in fact somewhat easier than for off-line sports. Therefore, we have chosen to broaden the scope of the workshop and solicit e-sports submissions as well. The majority of techniques used in the field so far are statistical. While there has been some interest in the Machine Learning and Data Mining community, it has been somewhat muted so far. Building off our successful workshops on Sports Analytics at ECML/PKDD 2013 and ECML/PKDD 2015, we intend to change this by hosting a third edition at ECML/PKDD 2016.

PROGRAM

09:00 - 9:20	Introduction to the workshop and prediction challenge
09:20 - 9:40	Antoine Adam - Generalised Linear Model for Predicting Football Matches
09:40 - 10:00	Maryam Tavakol, Hamid Zafartavanaelmi and Ulf Brefeld - Feature Extraction and Aggregation for Predicting the EURO 2016
10:00 - 10:20	Jan Lasek - EURO 2016 Predictions Using Team Rating Systems
10:20 - 10:40	Lucas Maystre, Victor Kristof, Antonio J. Gonzalez Ferrer and Matthias Grossglauser - The Player Kernel: Learning Team Strengths Based on Implicit Player Contributions
16:00 - 16:20	Coffee break
11:00 - 11:20	Aleksandr Semenov, Peter Romov, Kirill Neklyudov, Daniil Yashkov and Daniil Kireev - Applications of Machine Learning in DOTA 2: Literature Review and Practical Knowledge Sharing
11:20 - 11:40	Madan Gopal Jhanwar and Vikram Pudi - Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach
11:40 - 12:00	Harm Eggels, Ruud van Elk and Mykola Pechenizkiy - Explaining Soccer Match Outcomes With Goal Scoring Opportunities Predictive Analytics
12:00 - 12:20	Vincent Vercruyssen, Luc De Raedt and Jesse Davis - Qualitative Spatial Reasoning for Soccer Pass Prediction
12:20 - 12:40	Konstantinos Pelechrinis - Decision Making in American Football: Evidence from 7 Years of NFL Data
12:40 - 14:20	Lunch break
14:20 - 15:20	Invited talk: Dr. Daniel Link - Data Mining and Sports Analytics - Bridging the gap between science and practice
15:20 - 15:40	Boris Doux, Clement Gautrais and Benjamin Negrevergne - Detecting Key Strategic Events in HearthStone Matches
15:40 - 16:00	Albrecht Zimmermann - Wages of Wins: Could an Amateur Make Money From Match Outcome Predictions?
16:00 - 16:20	Coffee break
16:20 - 16:40	Juan Alfonso Lara Torralbo, José María Barreiro, David de La Peña and David Lizcano - Data Mining in Stabilometry: Application to Patient Balance Study for Sports Talent Mapping
16:40 - 17:00	Daniel Ruiz-Mayo, Estrella Pulido and Gonzalo Martínez-Muñoz - Marathon Performance Prediction of Amateur Runners based on Training Session Data
17:00 - 17:20	Javier Fernández, Daniel Medina, Antonio Gómez, Marta Arias and Ricard Gavaldà - Does Training Affect Match Performance? A Study Using Data Mining And Tracking Devices
17:20 - 17:40	Dimitri de Smet, Marc Francaux, Julien M. Hendrickx and Michel Verleysen - Cardiac Parameters Identification for Fitness Assessment
17:40 - 18:00	Wrap up and discussion

MONDAY 19 SEPTEMBER 2016

WORKSHOPS – FULL DAY

AALTD: 2ND ECML/PKDD WORKSHOP ON ADVANCED ANALYTICS AND LEARNING ON TEMPORAL DATA

Organizers: Andrés M. Alonso, Benjamin Bustos, Ahlame Douzal-Chouakria, Simon Malinowski, Pierre-François Marteau, Edoardo Otranto, Romain Tavenard, José Antonio Vilar Fernández
Time: 09:00 - 18:00
Room: 1000B
Webpage: https://aaltd16.irisa.fr/

Temporal data are frequently encountered in a wide range of domains such as bio-informatics, medicine, finance and engineering, among many others. They are naturally present in applications covering language, motion and vision analysis, or more emerging ones as energy efficient building, smart cities, dynamic social media or sensor networks. Contrary to static data, temporal data are of complex nature, they are generally noisy, of high dimensionality, they may be non-stationary (i.e. first order statistics vary with time) and irregular (involving several time granularities), they may have several invariant domain-dependent factors as time delay, translation, scale or tendency effects. These temporal peculiarities make limited the majority of standard statistical models and machine learning approaches, that mainly assume i.i.d data, homoscedasticity, normality of residuals, etc. To tackle such challenging temporal data, one appeals for new advanced approaches at the bridge of statistics, time series analysis, signal processing and machine learning. Defining new approaches that transcend boundaries between several domains to extract valuable information from temporal data is undeniably a hot topic in the near future, that has been yet the subject of active research this last decade. The aim of this workshop is to bring together researchers and experts in machine learning, data mining, pattern analysis and statistics to share their challenging issues and advance researches on temporal data analysis. Analysis and learning from temporal data cover a wide scope of tasks including learning metrics, learning representations, unsupervised feature extraction, clustering and classification. The proposed workshop welcomes papers that cover, but not limited to, one or several of the following topics:

— Temporal data clustering
— Semi-supervised and supervised classification on temporal data
— Early classification of temporal data
— Deep learning and learning representations for temporal data
— Metric and kernel learning for temporal data
— Modeling temporal dependencies
— Advanced forecasting and prediction models
— Space-temporal statistical analysis
— Functional data analysis methods
— Temporal data streams
— Dimensionality reduction, sparsity, algorithmic complexity and big data challenge
— Bio-informatics, medical, energy consumption, multimedia and other applications on temporal data
— Benchmarking and assessment methods for temporal data

PROGRAM

SESSION I (9:00 - 10:40)

09:00 - 09:20 Welcome speech by workshop chairs

09:20 - 09:40 Stephan Spiegel - Transfer Learning for Time Series Classification in Dissimilarity Spaces

09:40 - 10:00 Weiwei Shi, Yongxin Zhu, Xiao Pan, Philip Yu and Yufeng Chen - Missing Data Prediction in Multi-source Time Series with Sensor Network Regularization

10:00 - 10:20 Pablo Montero-Manso and Jose A. Vilar - A time series two-sample test based on comparing distributions of pairwise distances

10:20 - 10:40 Romain Brault, Néhémy Lim and Florence d'Alché-buc - Scaling up Vector Autoregressive Models with Operator-Valued Random Fourier Features

10:40 - 11:00 Coffee break

SESSION II (11:00 - 12:40)

11:00 - 11:40 Invited talk: Tony Bagnall

11:40 - 12:00 Xavier Renard, Maria Rifqi, Gabriel Fricout and Marcin Detyniecki - EAST representation: fast discovery of discriminant temporal patterns from time series

12:00 - 12:20 Christina Papagiannopoulou, Diego Miralles, Mathieu Depoorter, Niko Verhoest, Wouter Dorigo and Willem Waegeman - Discovering relationships in climate-vegetation dynamics using satellite data

12:20 - 12:40 David Tolpin - Progressive Temporal Window Widening

12:40 - 14:20 Lunch break

SESSION III (14:20 - 16:00)

14:20 - 15:00 Invited talk: Marco Cuturi

15:00 - 15:20 Katsiaryna Mirylenka, Christoph Miksovic and Paolo Scotton - Recurrent Neural Networks for Modeling Company-Product Time Series

15:20 - 15:40 Arthur Le Guennec, Simon Malinowski and Romain Tavenard - Data Augmentation for Time Series Classification using Convolutional Neural Networks

15:40 - 16:00 Yulong Pei, Jianpeng Zhang, George H. L. Fletcher and Mykola Pechenizkiy - Node Classification in Dynamic Social Networks

16:00 - 16:20 Coffee break

CHALLENGE SESSION (16:20 - 18:00)

16:20 - 16:40 Presentation of the challenge by challenge chairs

16:40 - 18:00 Presentation of competing methods by their authors

MONDAY 19 SEPTEMBER 2016

WORKSHOPS – FULL DAY

MIDAS: THE FIRST WORKSHOP ON MINING DATA FOR FINANCIAL APPLICATIONS

Organizers: Ilaria Bordino, Guido Caldarelli, Fabio Fumarola, Francesco Gullo, Tiziano Squartini
Time: 09:00 - 18:00
Room: Stampa
Website: http://networks.imtlucca.it/conferences/midas

Like the famous King Midas, popularly remembered in Greek mythology for his ability to turn everything he touched with his hand into gold, we believe that the wealth of data generated by modern technologies, with widespread presence of computers, users and media connected by Internet, is a goldmine for tackling a variety of problems in the financial domain. Nowadays, people's interactions with technological systems provide us with gargantuan amounts of data documenting collective behaviour in a previously unimaginable fashion. Recent research has shown that by properly modeling and analyzing these massive datasets, for instance representing them as network structures, it is possible to gain useful insights into the evolution of the systems considered (i.e., trading, disease spreading, political elections). Investigating the impact of data arising from today's application domains on financial decisions may be of paramount importance. Knowledge extracted from data can help gather critical information for trading decisions, reveal early signs of impactful events (such as stock market moves), or anticipate catastrophic events (e.g., financial crises) that result from a combination of actions, and affect humans worldwide. Despite its well recognized relevance and some recent related efforts, data mining in finance is still not stably part of the main stream of datamining conferences. This makes the topic particularly appealing for a workshop, whose small, interactive, and possibly interdisciplinary context provides a unique opportunity to advance research in a stimulating but still quite unexplored field. The MIDAS workshop aims at discussing challenges, potentialities, and applications of leveraging datamining tasks to tackle problems in the financial domain. The workshop will provide a premier forum for sharing findings, knowledge, insights, experience and lessons learned from mining data generated in various domains. The intrinsic interdisciplinary nature of the workshop will promote the interaction between computer scientists, physicists, mathematicians, economists and financial analysts, thus paving the way for an exciting and stimulating environment involving researchers and practitioners from different areas.

PROGRAM

SESSION I (09:00 - 10:40)

09:00 - 09:10 Opening

09:10 - 10:10 Invited Talk: Fabrizio Lillo

10:10 - 10:40 Marco Bianchetti, Davide Emilio Galli, Camilla Ricci, Angelo Salvatori and Marco Scaringi - Brexit or Bremain? Evidence from bubble analysis

10:40 - 11:10 Coffee break

SESSION II (11:10 - 12:40)

11:10 - 11:40 Andrea Pazienza, Sabrina Francesca Pellegrino, Stefano Ferilli and Floriana Esposito - Clustering underlying stock trends via non-negative matrix factorization

11:40 - 12:10 Argimiro Arratia and Marti Renedo - Clustering of exchange rates and their dynamics under different dependence measures

12:10 - 12:40 Huisu Jang and Jaewook Lee - A general framework for building machine learning models for pricing american index options with no-arbitrage

12:40 - 14:20 Lunch break

SESSION III (14:20 - 15:50)

14:20 - 15:20 Invited Talk: Marcello Paris

15:20 - 15:50 Argimiro Arratia, Alejandra Cabaña and Àlex Serès - Towards a sharp estimation of transfer entropy for identifying causality in financial time series

15:50 - 16:20 Coffee break

SESSION IV (16:20 - 18:00)

16:20 - 16:50 Salvatore Cuomo, Pasquale De Michele, Vittorio Di Somma and Giovanni Ponti - A probabilistic approach for financial IoT data

16:50 - 17:20 Alya Al Nasseri, Faek Menla Ali and Allan Tucker - Good news and bad news: using machine learning to identify investor sentiment reaction to return news

17:20 - 17:50 Ali Caner Türkmen - Sentiment extraction from financial public disclosure documents

17:50 - 18:00 Concluding Remarks

WORKSHOPS – FULL DAY

NFMCP: 5TH INTERNATIONAL WORKSHOP ON NEW FRONTIERS IN MINING COMPLEX PATTERNS

Organizers: Annalisa Appice, Michelangelo Ceci, Corrado Loglisci, Elio Masciari, Zbigniew W. Ras
Time: 09:00 - 18:00
Room: 300A
Website: http://www.di.uniba.it/~loglisci/NFmcp2016/index.html

Modern automatic systems are able to collect huge volumes of data, often with a complex structure (e.g. multi-table data, XML data, web data, time series and sequences, graphs and trees). This fact poses new challenges for current information systems with respect to storing, managing and mining these big sets of complex data. The purpose of this workshop is to bring together researchers and practitioners of data mining who are interested in the advances and latest developments in the area of extracting patterns from big and complex data sources like blogs, event or log data, biological data, spatio-temporal data, social networks, mobility data, sensor data and streams, and so on. The workshop aims at integrating recent results from existing fields such as data mining, statistics, machine learning and relational databases to discuss and introduce new algorithmic foundations and representation formalisms in pattern discovery. We are interested in advanced techniques which preserve the informative richness of data and allow us to efficiently and efficaciously identify complex information units present in such data. A non-exclusive list of topics for the complex pattern mining research is reported in the following: Foundations on pattern mining, pattern usage, and pattern understanding

— Mining stream, time-series and sequence data
— Mining networks and graphs
— Mining biological data
— Mining dynamic and evolving data
— Mining environmental and scientific data
— Mining heterogeneous and ubiquitous data
— Mining multimedia data
— Mining multi-relational data
— Mining semi-structured and unstructured data
— Mining spatio-temporal data
— Mining Big Data
— Social Media Analytics
— Ontology and metadata
— Privacy preserving mining
— Semantic Web and Knowledge Databases

PROGRAM

09:00 - 09:05 Opening
09:05 - 10:00 Invited Talk: Jaakko Hollmén

10:00-10:40 SESSION I: FEATURE SELECTION AND INDUCTION
10:00 - 10:20 Giorgio Roffo and Simone Melzi - Features Selection via Eigenvector Centrality
10:20 - 10:40 Konstantinos Pliakos and Celine Vens - Feature Induction based on Extremely Randomized Tree Paths

10:40 - 11:00 Coffee Break

11:00 - 12:40 SESSION II: CLASSIFICATION AND PREDICTION
11:00 - 11:20 Elzbieta Kubera, Alicja Wieczorkowska, Tomasz Slowik, Andrzej Kuranc and Krzysztof Skrzypiec - Speed Change Classification for Engines in Vehicles
11:20 - 11:40 Dariusz Brzezinski, Zbigniew Grudziński and Izabela Szczęch - Bayesian Confirmation Measures in Rule-based Classification
11:40 - 12:00 Jerzy Stefanowski, Krystyna Napierala and Izabela Szczech - Increasing the Interpretability of Rules Induced from Imbalanced Data by Using Bayesian Confirmation Measures
12:00 - 12:20 Fabio Leuzzi, Giovanni Tessitore, Stefano Delfino, Claudio Fusco, Massimo Gneo, Gianpaolo Zambonini and Stefano Ferilli - A statistical approach to speaker identification in forensic phonetics field
12:20 - 12:40 Hiroki Takahashi and Masahiro Kimura - Analyzing Time-decay Effects of Mediating-objects in Creating Trust-links

12:40 - 14:20 Lunch break

14:20 - 15:00 SESSION III: SEQUENCES AND TIME-SERIES
14:20 - 14:40 Martin Atzmueller, Andreas Schmidt, Benjamin Kloepper and David Arnu - HypGraphs: An Approach for Modeling and Comparing Graph-Based and Sequential Hypotheses
14:40 - 15:00 Nita Valmarska, Dragana Miljkovic, Nada Lavrac and Marko Robnik-Šikonja - Multi-view Approach to Parkinson's Disease Quality of Life Data Analysis

15:00 - 16:00 SESSION IV: CLUSTERING
15:00 - 15:20 Maël Chiapino and Anne Sabourin - Feature clustering for extreme events analysis, with application to extreme stream-flow data
15:20 - 15:40 Massimo Guarascio, Francesco Sergio Pisani, Ettore Ritacco and Pietro Sabatino - Human Behavior Discovery via Multidimensional Latent Factor Modeling
15:40 - 16:00 Mamoun Almardini, Ayman Hajja, Zbigniew Ras, Lina Clover and David Olaleye - Predicting the primary medical procedure through personalization

16:00 - 16:20 Coffee Break

16:20 - 17:00 SESSION V: RULE MINING
16:20 - 16:40 Bart Goethals, Emmanuel Mueller and Thomas Van Brussel - Randomized Quantitative Association Rule Mining
16:40 - 17:00 Stefano Ferilli, Domenico Redavid and Sergio Angelastro - Mining Chess Playing as a Complex Process

17:00 - 18:00 SESSION VI: PATTERN DISCOVERY
17:00 - 17:20 Laura Genga, Domenico Potena, Orazio Martino, Mahdi Alizadeh, Claudia Diamantini and Nicola Zannone - Subgraph Mining for Anomalous Pattern Discovery in Event Logs
17:20 - 17:40 Saket Maheshwary - Mining Keystroke Timing Pattern for User Authentication
17:40 - 18:00 Corrado Loglisci, Michelangelo Ceci, Angelo Impedovo and Donato Malerba - Mining Spatio-Temporal Patterns of Periodic Changes in Climate Data

MONDAY 19 SEPTEMBER 2016

WORKSHOPS – MORNING

SDDM: 2ND ECML/PKDD 2016 WORKSHOP ON STATISTICALLY SOUND DATA MINING

Organizers: Wilhelmiina Hämäläinen, Geoff Webb
Time: 09:00 - 12:40
Room: Presidenza
Website: https://sites.google.com/site/whamalaipages/ssdm2016

The field of statistics has developed sophisticated well-founded methods for inference from data. While some of these place computational or practical limits that made them infeasible to apply directly to many data mining problems, the field of data mining has much to gain from a more sophisticated understanding of the strengths and limitations of these techniques and from greater utilization of them where they are appropriate. The workshop topic is extremely important and topical, because a clear trend towards statistically sound data mining is currently emerging. It seems that a paradigm shift from the so-called Method-first to the Patterns-first approach is currently occurring. So, instead of defining easily searchable but possibly low-value or spurious patterns, one should rather design algorithms for statistically meaningful and valid patterns. This paradigm shift is strongly supported by the application fields, where the data mining methods are all the time gaining in popularity. Researchers of application fields cannot build their work on arbitrary, false or spurious discoveries, but they require guarantees of statistical significance and validity. The goal of this workshop is to bring together researchers from data mining, machine learning and statistics as well as application fields to discuss the problems and share ideas on statistically sound data mining.

PROGRAM

09:00 - 10:40	SESSION I
09:00 - 09:15	Opening
09:15 - 10:15	Koji Tsuda - Significant Pattern Mining: Efficient Algorithms and Biomedical Applications
10:15 - 10.40	Matthijs van Leeuwen and Antti Ukkonen - Expect the unexpected? On the significance of subgroups
10:40 - 11:00	Coffee Break
11:00 - 12:40	SESSION II
11:00 - 11:40	Francois Petitjean - Scaling log-linear analysis to datasets with thousands of variables
11:40 - 12:10	Jan Ramon - Statistically sound analysis of populations resulting from haplotype evolution
12:10 - 12:40	Closing discussion with open problems
12:40	Lunch break

WORKSHOPS – AFTERNOON

SOGOOD: DATA SCIENCE FOR SOCIAL GOOD

Organizers: Ricard Gavaldà, Indrà Zliobaitè, João Gama
Time: 14:20 - 18:00
Room: Belvedere
Website: https://sites.google.com/site/ecmlpkdd2016sogood/

The "Big Data" term is often perceived by society more as a threat than as a blessing. Commercial entities and governments have taken the lead in actual use of large data technologies in advertising, marketing, surveillance, finances, search and many more. Most people are not aware that data science applications are contributing to creating social good, for example helping people at the bottom of the economic pyramid or with special needs, improving healthcare systems, reinforcing international cooperation, or dealing with environmental problems, disasters, and climate change. Nobody doubts that such applications are important, but as they do not promise a direct financial benefit, it is not obvious in what ways they should be carried out, who should be responsible for their development, and what could be the best ways to make them happen. Social good applications do find their place sporadically in the regular scientific forums on Data Science (conferences and journals). However, not only they are not given any particular prominence, but they are often not advertised under any indicative label; they may appear separated e.g. in sessions on "Social networks" or "Privacy" or "Predictive models" or even under the catch-all term "Applications". Additionally, such forums tend to have a strong bias for papers that are novel in the strictly technical sense (new algorithms, new kinds of data analysis, new technologies) rather than on the novelty or the social impact of the application. This workshop will discuss and (re)define what is data science for social good, how it could be pushed forward, and in what ways the scientific community can contribute. In this workshop we plan to attract papers presenting applications (which may, or may not require new methods) of Data Science to Social Good, or else that take into account social aspects of Data Science methods and techniques. Application domains should be as varied as possible. A non-exclusive list includes:

- Public safety and disaster relief
- Access to food, water, and utilities
- Efficiency and sustainability
- Government transparency
- Data journalism
- Economic development
- Education
- Social services
- Healthcare

Profit-driven projects are not out of the scope of the workshop per se, but the goal here is to find good means and define resources how the needs (that do not potentially promise much profit but are relevant for the society) can be addressed.

MONDAY 19 SEPTEMBER 2016

WORKSHOPS – AFTERNOON

PROGRAM

14:20 - 14:25 Welcome - presentation

14:25 - 16:00 Session 1

Yann-Aël Le Borgne, Adriana Homolova and Gianluca Bontempi
OpenTED Browser: Insights into European Public Spendings

Diego Garcia-Olano, Marta Arias and Josep L Larriba-Pey
Automated Construction and Analysis of Political Networks via open government and media sources

Antonio Rodriguez, Frederic Bartumeus and Ricard Gavaldà
Machine Learning Assists the Classification of Reports by Citizens on Disease-Carrying Mosquitoes

Mohammed Ahmed, Gianni Barlacchi, Stefano Braghin, Francesco Calabrese, Michele Ferretti, Vincent Lonij, Rahul Nair, Rana Novack, Jurij Paraszczak and Andeep Toor
A multi-scale approach to data-driven mass migration analysis

Nayantara Kotoky and Saradhi Vijaya
Right to Information Query Modelling via Graded Response Model

Emanuele Di Buccio, Andrea Lorenzet, Massimo Melucci and Federico Neresini
Unveiling Latent States Behind Social Indicators

16:00 - 16:20 Coffee break

16:20 - 16:50 Session 2

Fernando Martínez-Plumed, Cesar Ferri and Lidia Contreras-Ochando
Cycling network projects: a decision-making aid approach

Pavlos Paraskevopoulos, Giovanni Pellegrini and Themis Palpanas:
When a Tweet Finds its Place - Fine-Grained Tweet Geolocalisation

16:50 - 17:20 Panel session

Panelists: André Carvalho, Mykola Pechenizkiy, Dino Pedreschi, Antti Ukkonen

17:20 - 18:00 Poster session

PhD FORUM

PHD FORUM

Organizers: Leman Akoglu, Tijl de Bie
Time: 09:00 - 18:00
Room: Meeting

The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) includes a PhD Forum on machine learning and knowledge discovery.
The purpose of this forum is to provide an environment specifically for junior PhD students to exchange ideas and experiences with peers in an interactive atmosphere and to get constructive feedback from senior researchers in data mining, machine learning, and related areas. The focus of the discussion at the PhD Forum would be the work in progress of junior PhD students, with 1-3 years of research experience, towards their dissertation. During the forum, senior researchers with experience in supervising and examining PhD students will be able to participate and provide feedback and advice to the participants. It will be an excellent opportunity for developing person-to-person networks to the benefit of the PhD students in their future careers.

PROGRAM

08:45 - 09:00	Welcome & Remarks
09:00 - 09:35	Jilles Vreeken - Keynote Presentation: "How to Survive and Enjoy PhD"
09:35 - 10:15	Mini-Panel 1 (Q&A mentors and students)
10:15 - 10:40	Spotlight talks (2 minutes each)
10:40 - 11:00	Coffee break
11:00 - 12:45	Poster session (morning spotlights)
12:40 - 14:20	Lunch break
14:20 - 14:55	Dafna Shahaf - Keynote Presentation: "Experience Sharing" (abstract)
14:55 - 15:35	Mini-Panel 2 (Q&A mentors and students)
15:35 - 16:00	Spotlight talks (2 minutes each)
16:00 - 16:20	Coffee break
16:20 - 18:00	Poster session (afternoon spotlights)
18:00	Closing

MONDAY 19 SEPTEMBER 2016

DISCOVERY CHALLENGES – AFTERNOON

BANK CARD USAGE ANALYSIS

Discovery Challenge Chairs: Elio Masciari, Alessandro Moschitti
Challenge Organizers: Illés Gozlán, Gábor Káposztási, Róbert Pálovics, Frederick Ayala Gomez, András Benczúr
Time: 14:20 – 16:00
Room: 100B
Website: https://dms.sztaki.hu/ecml-pkkd-2016

The ECML/PKDD Discovery Challenge 2016 on Bank Card Usage Analysis asks you to predict the user behavior of the OTP Bank Hungary, a key bank in CEE Region. We give you one year list of card payment events with geolocation information.
The Bank wants to know wich branch will be visited by each customer to be able to optimize proactive contact list and plan distribution. The customer will be proactively called in campaigns from the branch that will be visited with the highest probability. The bank expects higher conversion rates in branch campaigns if the call is made in the branch mostly preferred by the customer.

PROGRAM

14:20 - 14:50	Introduction of the task, announcement of the winners, prize ceremony
14:50 - 15:05	Task 1 winner presentation
15:05 - 15:20	Task 2 winner presentation
15:20 - 15:30	Task 1 second place presentation
15:30 - 15:40	Task 2 second place presentation
15:40 - 15:50	Task 1 third place presentation
15:50 - 16:00	Task 2 third place presentation

SPHERE CHALLENGE: ACTIVITY RECOGNITION WITH MULTIMODAL SENSOR DATA

Discovery Challenge Chairs: Elio Masciari, Alessandro Moschitti
Challenge Organizers: Niall Twomey, Tom Diethe, Meelis Kull, Peter Flach, Ian Craddock
Time: 16:20 – 18:00
Room: 100B
Website: https://www.drivendata.org/competitions/sphere

Obesity, depression, stroke, falls, cardiovascular and musculoskeletal disease are some of the biggest health issues and fastest-rising categories of health-care costs. The financial expenditure associated with these is widely regarded as unsustainable and the impact on quality of life is felt by millions of people in the UK each day. Smart technologies can unobtrusively quantify activities of daily living, and these can provide long-term behavioural patterns that are objective, insightful measures for clinical professionals and caregivers.
To this end the EPSRC-funded "Sensor Platform for HEalthcare in Residential Environment (SPHERE)" Interdisciplinary Research Collaboration (IRC) has designed a multi-modal sensor system driven by data analytics requirements. The system is under test in a single house, and will be deployed in a general population of 100 homes in Bristol (UK). The data sets collected will be made available to researchers in a variety of communities.
Data is collected from the following three sensing modalities:

— wrist-worn accelerometer;
— RGB-D cameras (i.e. video with depth information); and passive environmental sensors.

With these sensor data, we can learn patterns of behaviour, and can track the deterioration/progress of persons that suffer or recover from various medical conditions. To achieve this, we focus activity recognition over multiple tiers, with the two main prediction tasks of SPHERE including:

— prediction of Activities of Daily Living (ADL) (e.g. tasks such as meal preparation, watching television); and
— prediction of posture/ambulation (e.g. walking, sitting, transitioning).

Reliable predictions of ADL allows us to model behaviour and of residents over time, e.g. what does a typical day consist of, what times are particular activities performed etc. Prediction of posture and ambulation will complement ADL predictions, and can inform us about the physical well-being of the participant, how mobile/responsive is the participant, how activie/sedintary, etc.

PROGRAM

16:20 - 16:40	Challenge Presentation
16:40 - 17:55	Challenger Presentations
17:55 - 18:00	Award Ceremony

INVITED TALK

SOCIAL LEARNING

Speaker: Alex "Sandy" Pentland
Time: 18:45 – 19:35
Room: 1000A

Abstract:

Human decisions are heavily influenced by social interaction, so that predicting or influencing individual behavior requires modeling these interaction effects. In addition the distributed learning strategies exhibited by human communities suggest methods of improving both machine learning and human-machine systems. Several practical examples will be described.

Bio:

Professor Alex "Sandy" Pentland directs the MIT Connection Science and Human Dynamics labs and previously helped create and direct the MIT Media Lab and the Media Lab Asia in India. He is one of the most-cited scientists in the world, and Forbes recently declared him one of the "7 most powerful data scientists in the world" along with Google founders and the Chief Technical Officer of the United States. He has received numerous awards and prizes such as the McKinsey Award from Harvard Business Review, the 40th Anniversary of the Internet from DARPA, and the Brandeis Award for work in privacy.

He is a founding member of advisory boards for Google, AT&T, Nissan, and the UN Secretary General, a serial entrepreneur who has co-founded more than a dozen companies including social enterprises such as the Data Transparency Lab, the Harvard-ODI-MIT DataPop Alliance and the Institute for Data Driven Design. He is a member of the U.S. National Academy of Engineering and leader within the World Economic Forum.

TUESDAY 20 SEPTEMBER 2016

INVITED TALK

CAUSAL INFERENCE AND MACHINE LEARNING: ESTIMATING AND EVALUATING POLICIES

Speaker: Susan Athey
Time: 09:10 - 10:00
Room: 1000A
Abstract:

In many contexts, a decision-making can choose to assign one of a number of "treatments" to individuals. The treatments may be drugs, offers, advertisements, algorithms, or government programs. One setting for evaluating such treatments involves randomized controlled trials, for example A/B testing platforms or clinical trials. In such settings, we show how to optimize supervised machine learning methods for the problem of estimating heterogeneous treatment effects, while preserving a key desiderata of randomized trials, which is providing valid confidence intervals for estimates. We also discuss approaches for estimating optimal policies and online learning. In environments with observational (non-experimental) data, different methods are required to separate correlation from causality. We show how supervised machine learning methods can be adapted to this problem.

Bio:

Susan Athey is The Economics of Technology Professor at Stanford Graduate School of Business. She received her bachelor's degree from Duke University and her Ph.D. from Stanford, and she holds an honorary doctorate from Duke University. She previously taught at the economics departments at MIT, Stanford and Harvard. In 2007, Professor Athey received the John Bates Clark Medal, awarded by the American Economic Association to "that American economist under the age of forty who is adjudged to have made the most significant contribution to economic thought and knowledge." She was elected to the National Academy of Science in 2012 and to the American Academy of Arts and Sciences in 2008. Professor Athey's research focuses on the economics of the internet, online advertising, the news media, marketplace design, virtual currencies and the intersection of computer science, machine learning and economics. She advises governments and businesses on marketplace design and platform economics, notably serving since 2007 as a long-term consultant to Microsoft Corporation in a variety of roles, including consulting chief economist.

BEST DM PAPER

A00: CANCER: ANOTHER ALGORITHM FOR SUBTROPICAL MATRIX FACTORIZATION

Authors: Sanjar Karaev, Pauli Miettinen
Time: 10:00 - 10:30
Room: 1000A
Abstract:

Subtropical algebra is a semi-ring over the nonnegative real numbers with standard multiplication and the addition defined as the maximum operator. Factorizing a matrix over the subtropical algebra gives us a representation of the original matrix with element-wise maximum over a collection of nonnegative rank-1 matrices. Such structure can be compared to the well-known Nonnegative Matrix Factorization (NMF) that gives an element-wise sum over a collection of nonnegative rank-1 matrices. Using the maximum instead of sum changes the `parts-of-whole' interpretation of NMF to `winner-takes-it-all' interpretation. We recently introduced an algorithm for subtropical matrix factorization, called Capricorn, that was designed to work on discrete-valued data with discrete noise [Karaev & Miettinen, SDM '16]. In this paper we present another algorithm, called Cancer, that is designed to work over continuous-valued data with continuous noise - arguably, the more common case. We show that Cancer is capable of finding sparse factors with excellent reconstruction error, being better than either Capricorn, NMF, or SVD in continuous subtropical data. We also show that the winner-takes-it-all interpretation is usable in many real-world scenarios and lets us find structure that is different, and often easier to interpret, than what is found by NMF.

TUESDAY 20 SEPTEMBER 2016

TUESDAY SESSIONS AT A GLANCE

PARALLEL SESSIONS 11:00 - 13:00

Tue1A: Deep Learning and Neural Networks 1 — Room 1000A

11:00 - 11:20 A001: AUC-Maximized Deep Convolutional Neural Fields for Protein Sequence Labeling
Labeling Sheng Wang, Siqi Sun, Jinbo Xu

11:20 - 11:40 A002: adaQN: An Adaptive Quasi-Newton Algorithm for Training RNNs
Nitish Shirish Keskar, Albert Berahas

11:40 - 12:00 A003: Multilabel classification on heterogeneous graphs with gaussian embeddings
Ludovic DOS SANTOS, Benjamin Piwowarski, Patrick Gallinari

12:00 - 12:20 A004: Composite denoising autoencoders
Krzysztof Geras, Charles Sutton

12:20 - 12:40 A005: Sequential Labeling with Online Deep Learning: Exploring Model Initialization
Gang Chen, Ran Xu, Sargur Srihari

12:40 - 13:00 A006: Deep Metric Learning with Data Summarization
Wenlin Wang, Changyou Chen, Wenlin Chen, Lawrence Carin

Tue1B: Graphs — Room 1000B

11:00 - 11:20 A007: Asynchronous Distributed Incremental Computation on Evolving Graphs
Jiangtao Yin, Lixin Gao

11:20 - 11:40 A008: Optimizing Network Robustness by Edge Rewiring: A General Framework
Hau Chan, Leman Akoglu

11:40 - 12:00 A009: Locating the Contagion Source in Networks with Partial Timestamps
Kai Zhu, Zhen Chen, Lei Ying

12:00 - 12:20 A010: An Efficient Exact Algorithm for Triangle Listing in Large Graphs
Sofiane Lagraa, Hamida Seba

12:20 - 12:40 A011: A Distributed Approach for Graph Mining in Massive Networks.
Nilothpal Talukder, Mohammed J. Zaki.

12:40 - 13:00 A012: M-Flash: Fast Billion-Scale Graph Computation Using a Bimodal Block Processing Model
Hugo Gualdron Colmenares, Robson Leonardo Ferreira Cordeiro, José Fernando Rodrigues Jr., Duen Horng Chau, Minsuk Kahng, U Kang

Tue1C: Clustering 1 — Room 300A

11:00 - 11:20 A013: Infection hot spot mining from social media trajectories
Roberto Souza, Renato Assunção, Derick Oliveira, Denise Brito, Meira Wagner

11:20 - 11:40 A014: A Novel Incremental Covariance-guided One-class Support Vector Machine
Takoua Kefi, Riadh Ksantini, Mohamed Kaâniche, Adel Bouhoula

11:40 - 12:00 A015: Incremental Commute Time Using Random Walks and Online Anomaly Detection
Khoa Nguyen, Sanjay Chawla

12:00 - 12:20 A016: Node Re-Ordering as a Means of Anomaly Detection in Time-Evolving Graphs
Lida Rashidi, Christopher Leckie, Sutharshan Rajasegarar, Andrey Kan, Jeffrey Chan, James Bailey

12:20 - 12:40 A017: Multi-Graph Clustering based on Interior-Node Topology with Applications to Brain Networks
Guixiang Ma, Lifang He, Bokai Cao, Jiawei Zhang, Philip Yu

12:40 - 13:00 A018: A A Split-Merge DP-means Algorithm to Avoid Local Minima
Shigeyuki Odashima, Miwa Ueki, Naoyuki Sawasaki

Tue1D: Learning — Room 300B

11:00 - 11:20 A019: Efficient Distributed Decision Trees for Robust Regression
Tian Guo, Konstantin Kutzkov, mohamed Ahmed, jean-paul Calbimonte, Karl Aberer

11:20 - 11:40 A020: On the Convergence of A Family of Robust Losses for Stochastic Gradient Descent
Bo Han, Ivor Tsang, Ling Chen

11:40 - 12:00 A021: Actively Interacting with Experts: A Probabilistic Logic Approach
Phillip Odom, Sriraam Natarajan

12:00 - 12:20 A022: Stochastic CoSaMP: Randomizing Greedy Pursuit for Sparse Signal Recovery
Dipan Pal, OLE MENGSHOEL

12:20 - 12:40 A023: OSLa: Online Structure Learning using Background Knowledge Axiomatization
Evangelos Michelioudakis, Anastasios Skarlatidis, George Paliouras, Alexander Artikis

12:40 - 13:00 A024: Lifted generative learning of Markov logic networks
Jan Van Haaren, Guy Van den Broeck, Wannes Meert, Jesse Davis

Tue1E: Nectar 1 — Room Belvedere

ML/DM in engineering and natural and behavioural sciences/applications

11:00 - 11:20 A025: Practical Bayesian Inverse Reinforcement Learning for Robot Navigation
Billy Okal, Kai Arras

11:20 - 11:40 A026: Learning from Software Project Histories
Verena Honsel, Steffen Herbold, Jens Grabowski

11:40 - 12:00 A027: Resource-Aware Steel Production Through Data Mining
Hendrik Blom, Katharina Morik

12:00 - 12:20 A028: Machine learning challenges for single cell data
Sofie Van Gassen, Tom Dhaene, Yvan Saeys

12:20 - 12:40 A029: Query Log Mining for Inferring User
Rishabh Mehrotra, Emine Yilmaz

12:40 - 13:00 A030: Data Mining Meets HCI: Data and Visual Analytics of Frequent PatternsTasks and Needs
C Leung, Christopher Carmichael , Yaroslav Hayduk , Fan Jiang

TUESDAY 20 SEPTEMBER 2016

TUESDAY SESSIONS AT A GLANCE

PARALLEL SESSIONS 14:50 - 16:10

Tue2A: Classification 1 — Room 1000A

14:50 - 15:10 A031: F-Measure Maximization in Multi-Label Classification with Conditionally Independent Label Subsets
Maxime Gasse, Alex Aussem

15:10 - 15:30 A032: ALRn: Accelerated Higher-order Logistic Regression
Nayyar Zaidi, Geoffrey Webb, Francois Petitjean, Mark Carman, Jesus Crequides

15:30 - 15:50 A033: Beyond the Boundaries of SMOTE: A Framework for Manifold-Based Synthetically Oversampling
Colin Bellinger, Chris Drummond, Nathalie Japkowicz

15:50 - 16:10 A034: CHADE: Metalearning with Classifier Chains for Dynamic Combination of Classifiers
Fábio Pinto, Carlos Soares, Joao Moreira

Tue2B: Optimization — Room 1000B

14:50 - 15:10 A035: Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition
Hamed Karimi, Julie Nutini, Mark Schmidt

15:10 - 15:30 A036: Fast and Scalable Lasso via Stochastic Frank-Wolfe Methods with a Convergence Guarantee
Emanuele Frandi, Ricardo Ñanculef, Stefano Lodi, Claudio Sartori, Johan A. K. Suykens

15:30 - 15:50 A037: Two-Stage Transfer Surrogate Model for Automatic Hyperparameter Optimization
Martin Wistuba, Nicolas Schilling, Lars Schmidt-Thieme

15:50 - 16:10 A038: Scalable Hyperparameter Optimization with Products of Gaussian Process Experts
Nicolas Schilling, Martin Wistuba, Larst Schmidt-Thieme

Tue2C: Topic Modelling — Room 300A

14:50 - 15:10 A039: Detecting Public Influence on News Using Topic-aware Dynamic Granger Test
Lei Hou, Juanzi Li, Xiao-Li Li, Jianbin Jin

15:10 - 15:30 A040: C-BiLDA: Extracting Cross-lingual Topics from Non-Parallel Texts by Distinguishing Shared from Unshared Content
Geert Heyman, Ivan Vulic, Marie-Francine Moens

15:30 - 15:50 A041: Learning beyond Predefined Label Space via Bayesian Nonparametric Topic Modelling
Changying Du, Fuzhen Zhuang, Jia He, Qing He, Guoping Long

15:50 - 16:10 A042: Aspect Mining with Rating Bias
Yitong Li, Chuan Shi, Huidong Zhao, Fuzhen Zhuang, Bin Wu

Tue2D: Patterns — Room 300B

14:50 - 15:10 A043: Subgroup Discovery with Proper Scoring Rules
Hao Song, Meelis Kull, Peter Flach, Georgios Kalogridis

15:10 - 15:30 A044: A Bayesian Network Model for Interesting Itemsets
Jari Fowkes, Charles Sutton

15:30 - 15:50 A045: Transactional Tree Mining
Mostafa Haghir Chehreghani, Morteza Haghir Chehreghani

15:50 - 16:10 A046: Mining Rooted Ordered Trees under Subtree Homeomorphism
Mostafa Haghir Chehreghani, Maurice Bruynooghe

Tue2E: Nectar 2 Room Belvedere

ML/DM in systems and social sciences/applications

14:50 - 15:10 A047: Time and Again: Time Series Mining via Recurrence Quantification Analysis
Stephan Spiegel, Norbert Marwan

15:10 - 15:30 A048: Personality-Based User Modeling for Music Recommender Systems
Bruce Ferwerda, Markus Schedl

15:30 - 15:50 A049: Local Exceptionality Detection on Social Interaction Networks
Martin Atzmueller

15:50 - 16:10 A050: Machine Learning for Crowdsourced Spatial Data
Musfira Jilani, Padraig Corcoran, Michela Bertolotto

TUESDAY 20 SEPTEMBER 2016

TUESDAY SESSIONS AT A GLANCE

PARALLEL SESSIONS 16:40 - 18:20

Tue3A: Kernels — Room 1000A

16:40 - 17:00 A051: Improving Locality Sensitive Hashing Based Similarity Search and Estimation for Kernels
Aniket Chakrabarti, Bortik Bandyopadhyay, Srinivasan Parthasarathy

17:00 - 17:20 A052: Continuous Kernel Learning
John Moeller, Vivek Srikumar, Sarathkrishna Swaminathan, Suresh Venkatasubramanian, Dustin Webb

17:20 - 17:40 A053: Robust Dictionary Learning on the Hilbert Sphere in Kernel Feature Space
Nishanth Koushik, Suyash Awate

17:40 - 18:00 A054: Huber-Norm Regularization for Linear Prediction Models
Oleksandr Zadorozhnyi, Gundhart Benecke, Stephan Mandt, Tobias Scheffer, Marius Kloft

18:00 - 18:20 A055: Efficient Bayesian Maximum Margin Multiple Kernel Learning
Changying Du, Changde Du, Guoping Long, Xin Jin, Yucheng Li

Tue3B: Probabilistic Learning — Room 1000B

16:40 - 17:00 A056: An Online Gibbs Sampler Algorithm for Hierarchical Dirichlet Processes Prior
Yongdai Kim, Minwoo Chae, Kuhwan Jeong, Byungyup Kang, Hyoju Chung

17:00 - 17:20 A057: Gaussian Process Pseudo-Likelihood Models for Sequence Labeling
Srijith P.K., Balamurugan P, Shirish Shevade

17:20 - 17:40 A058: Copula PC Algorithm for Causal Discovery from Mixed Data
Ruifei Cui, Perry Groot, Tom Heskes

17:40 - 18:00 A059: Irrevocable-Choice Algorithms for Sampling from a Stream.
Yan Zhu, Eamonn Keogh

Tue3C: Clustering 2 — Room 300A

16:40 - 17:00 A060: ClusPath: A Temporal-driven Clustering to Infer Typical Evolution Paths
Marian-Andrei Rizoiu, Julien Velcin, Stéphane Bonnevay, Stéphane Lallich

17:00 - 17:20 A061: Renyi Divergence Minimization based Co-regularized Multiview Clustering
Shalmali Joshi, Joydeep Ghosh, Mark Reid, Oluwasanmi Koyejo

17:20 - 17:40 A062: Adaptive Trajectory Analysis of Replicator Dynamics for Data Clustering
Morteza Haghir Chehreghani

17:40 - 18:00 A063: Aggregating Crowdsourced Ordinal Labels via Bayesian Clustering
Xiawei Guo, James Kwok

Tue3D: Data Science — Room 300B

16:40 - 17:00 A064: A Hybrid Knowledge Discovery Approach for Mining Predictive Biomarkers in Metabolomic Data
Dhouha Grissa, Blandine Comte, Estelle Pujos Guillot, Amedeo Napoli

17:00 - 17:20 A065: Functional Bid Landscape Forecasting for Display Advertising
Yuchen Wang, Kan Ren, Weinan Zhang, Jun Wang, Yong Yu

17:20 - 17:40 A066: Enhancing Traffic Congestion Estimation with Social Media by Coupled Hidden Markov Model
Senzhang Wang, Fengxiang Li, Leon Stenneth, Philip Yu

17:40 - 18:00 A067: Joint Learning of Entity Semantics and Relation Pattern for Relation Extraction
Zheng Suncong, Xu Jiaming, Bao Hongyun, Qi Zhenyu, Zhang Jie, Hao Hongwei, Xu bo

Tue3E: Nectar 3 — Room Belvedere

ML/DM across domains and integrated processes

16:40 - 17:00 A068: From Plagiarism Detection to Bible Analysis: The Potential of Machine Learning for Grammar-Based Text Analysis
Michael Tschuggnall, Günther Specht

17:00 - 17:20 A069: Multi-Target Classification: Methodology and Practical Case Studies
Mark Last

17:20 - 17:40 A070: A KDD Process for Discrimination Discovery
Salvatore Ruggieri, Franco Turini

17:40 - 18:00 A071: Discussion

PARALLEL SESSIONS 18:00 - 18:16

Tue4A: Demo Spotlights 1 — Room Belvedere

18:00 - 18:02 A072: Bipeline: a Web-based Visualization Tool for Biclustering of Multivariate Time Series
Ricardo Cachucho, Kaihua Liu, Siegfried Nijssen, Arno Knobbe

18:02 - 18:04 A073: Coordinate transformations for characterization and cluster analysis of spatial configurations in football
Gennady Andrienko, Natalia Andrienko, Guido Budziak, Tatiana von Landesberger, Hendrik Weber

18:04 - 18:06 A074: Exploratory Analysis of Text Collections Through Visualization and Hybrid Biclustering
Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif

18:06 - 18:08 A075: GMMbuilder - User-Driven Discovery of Clustering Structure for Bioarchaeology
Markus Mauder, Eirini Ntoutsi, Yulia Bobkova

18:08 - 18:10 A076: INSIGHT: Dynamic Traffic Management Using Heterogeneous Urban Data
Nikolaos Zygouras, Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Nikos Zacheilas, Ioannis Mpoutsis, Vana Kalogeraki, stephen Lynch, Brendan O'Brien, Dermot Kinane, Jakub Marecek, Jia Yuan Yu, Rudi Verargo, Elizabeth Daly, Nico Piatkowksi, Thomas Liebig, Christian Bockermann, Katharina Morik, Matthias Weidlich, Francois Schnitzler, Avigdor Gal, Shie Mannor, Hendrik Stange, Werner Halft, Gennady Andrienko

18:10 - 18:12 A077: Learning Language Models from Images with ReGLL
Leonor Becerra-Bonache, Hendrik Blockeel, María Galván, Francois Jacquenet

18:12 - 18:14 A078: Leveraging Spatial Abstraction in Traffic Analysis and Forecasting with Visual Analytics
Natalia Andrienko, Gennady Andrienko, Salvatore Rinzivillo

18:14 - 18:16 A079: The SPMF Open-Source Data Mining Library Version 2
Philippe Fournier-Viger, Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, Thanh Lam Hoang

TUESDAY SESSIONS, WITH ABSTRACTS

TUE1A: DEEP LEARNING AND NEURAL NETWORKS 1 — ROOM 1000A

11:00 - 11:20 **A001: AUC-Maximized Deep Convolutional Neural Fields for Protein Sequence Labeling**
Sheng Wang, Siqi Sun, Jinbo Xu

Deep Convolutional Neural Networks (DCNN) has shown excellent performance in a variety of machine learning tasks. This paper presents Deep Convolutional Neural Fields (DeepCNF), an integration of DCNN with Conditional Random Field (CRF), for sequence labeling with an imbalanced label distribution. The widely-used training methods, such as maximum-likelihood and maximum labelwise accuracy, do not work well on imbalanced data. To handle this, we present a new training algorithm called maximum-AUC for DeepCNF. That is, we train DeepCNF by directly maximizing the empirical Area Under the ROC Curve (AUC), which is an unbiased measurement for imbalanced data. To fulfill this, we formulate AUC in a pairwise ranking framework, approximate it by a polynomial function and then apply a gradient-based procedure to optimize it. We then test our AUC-maximized DeepCNF on three very different protein sequence labeling tasks: solvent accessibility prediction, 8-state secondary structure prediction, and disorder prediction. Our experimental results confirm that maximum-AUC greatly outperforms the other two training methods on 8-state secondary structure prediction and disorder prediction since their label distributions are highly imbalanced and also has similar performance as the other two training methods on solvent accessibility prediction, which has three equally-distributed labels. Furthermore, our experimental results show that our AUC-trained DeepCNF models greatly outperform existing popular predictors of these three tasks.

11:20 - 11:40 **A002: adaQN: An Adaptive Quasi-Newton Algorithm for Training RNNs**
Nitish Shirish Keskar, Albert Berahas

Recurrent Neural Networks (RNNs) are powerful models that achieve exceptional performance on a plethora pattern recognition problems. However, the training of RNNs is a computationally difficult task owing to the well-known"vanishing/exploding" gradient problem. Algorithms proposed for training RNNs either exploit no (or limited) curvature information and have cheap per-iteration complexity, or attempt to gain significant curvature information at the cost of increased per-iteration cost. The former set includes diagonally-scaled first-order methods such as ADAGRAD and ADAM, while the latter consists of second-order algorithms like Hessian-Free Newton and K-FAC. In this paper, we present adaQN, a stochastic quasi-Newton algorithm for training RNNs. Our approach retains a low per-iteration cost while allowing for non-diagonal scaling through a stochastic L-BFGS updating scheme. The method uses a novel L-BFGS scaling initialization scheme and is judicious in storing and retaining L-BFGS curvature pairs. We present numerical experiments on two language modeling tasks and show that adaQN is competitive with popular RNN training algorithms.

11:40 - 12:00 **A003: Multilabel classification on heterogeneous graphs with gaussian embeddings**
Ludovic DOS SANTOS, Benjamin Piwowarski, Patrick Gallinari

We consider the problem of node classification in heterogeneous graphs where both nodes and relations may be of different types and a different set of categories is associated to each node type. When graph node classification has mainly been addressed for homogeneous graphs, heterogeneous classification is a recent problem which has been motivated by applications in fields such as social networks where the graphs are intrinsically heterogeneous. We propose a transductive approach to this problem based on learning graph embeddings and model the uncertainty associated to the node representations using Gaussian embeddings. A comparison with representative baselines is provided on three heterogeneous datasets.

12:00 - 12:20 **A004: Composite denoising autoencoders**
Krzysztof Geras, Charles Sutton

In representation learning, it is often desirable to learn features at different levels of scale. For example, in image data, some edges will span only a few pixels, whereas others will span a large portion of the image. We introduce an unsupervised representation learning method called a composite denoising autoencoder (CDA) to address this. We exploit the observation from previous work that in a denoising autoencoder, training with lower levels of noise results in more specific, fine-grained features. In a CDA, different parts of the network are trained with different versions of the same input, corrupted at different noise levels. We introduce a novel cascaded training procedure which is designed to avoid types of bad solutions that are specific to CDAs. We show that CDAs learn effective representations on two different image data sets.

12:20 - 12:40 **A005: Sequential Labeling with Online Deep Learning: Exploring Model Initialization**
Gang Chen, Ran Xu, Sargur Srihari

In this paper, we leverage both deep learning and conditional random fields (CRFs) for sequential labeling. More specifically, we explore parameter initialization and randomization in deep CRFs and train the whole model in a simple but effective way. In particular, we pretrain the deep structure with greedy layer-wise restricted Boltzmann machines (RBMs), followed with an independent label learning step. Finally, we re-randomize the top layer weight and update the whole model with an online learning algorithm -- a mixture of perceptron training and stochastic gradient descent to estimate model parameters. We test our model on different challenge tasks, and show that this simple learning algorithm yields the state of the art results.

12:40 - 13:00 A006: Deep Metric Learning with Data Summarization
Wenlin Wang, Changyou Chen, Wenlin Chen, Lawrence Carin

We present Deep Stochastic Neighbor Compression (DSNC), a framework to compress training data for instance-based methods (such as k-nearest neighbors). We accomplish this by inferring a smaller set of pseudo-inputs in a new feature space learned by a deep neural network. Our framework can equivalently be seen as jointly learning a nonlinear distance metric (induced by the deep feature space) and learning a compressed version of the training data. In particular, compressing the data in a deep feature space makes DSNC robust against label noise and issues such as within-class multi-modal distributions. This leads to DSNC yielding better accuracies and faster predictions at test time, as compared to other competing methods. We conduct comprehensive empirical evaluations, on both quantitative and qualitative tasks, and on several benchmark datasets, to show its effectiveness as compared to several baselines.

TUE1B: GRAPHS

ROOM 1000B

11:00 - 11:20 A007: Asynchronous Distributed Incremental Computation on Evolving Graphs
Jiangtao Yin, Lixin Gao

Graph algorithms have become an essential component in many real-world applications. An essential property of graphs is that they are often dynamic. Many applications must update the computation result periodically on the new graph so as to keep it up-to-date. Incremental computation is a promising technique for this purpose. Traditionally, incremental computation is typically performed synchronously, since it is easy to implement. In this paper, we illustrate that incremental computation can be performed asynchronously as well. Asynchronous incremental computation can bypass synchronization barriers and always utilize the most recent values, and thus it is more efficient than its synchronous counterpart. Furthermore, we develop a distributed framework, GraphIn, to facilitate implementations of incremental computation on massive evolving graphs. We evaluate our asynchronous incremental computation approach via extensive experiments on a local cluster as well as the Amazon EC2 cloud. The evaluation results show that it can accelerate the convergence speed by as much as 14x when compared to recomputation from scratch.

11:20 - 11:40 A008: Optimizing Network Robustness by Edge Rewiring: A General Framework
Hau Chan, Leman Akoglu

Spectral measures have long been used to quantify the robustness of real-world graphs. For example, spectral radius (or the principal eigenvalue) is related to the effective spreading rate of dynamic processes (e.g., rumor, disease, information propagation) on graphs. Algebraic connectivity (or the Fiedler value), which is a lower bound on the node and edge connectivity of a graph, captures the"partitionability" of a graph into disjoint components. In this work we address the problem of modifying a given graph's structure under a given budget so as to maximally improve its robustness, as quantified by spectral measures. We focus on modifications based on degree-preserving edge rewiring, such that the expected load (e.g., airport flight capacity) or physical/hardware requirement (e.g., count of ISP router traffic switches) of nodes remain unchanged. Different from a vast literature of measure-independent heuristic approaches, we propose an algorithm, called EdgeRewire, which optimizes a specific measure of interest directly. Notably, EdgeRewire is general to accommodate six different spectral measures. Experiments on real-world datasets from three different domains (Internet AS-level, P2P, and airport flights graphs) show the effectiveness of our approach, where EdgeRewire produces graphs with both (i) higher robustness, and (ii) higher attack-tolerance over several state-of-the-art methods.

11:40 - 12:00 A009: Locating the Contagion Source in Networks with Partial Timestamps
Kai Zhu, Zhen Chen, Lei Ying

This paper studies the problem of identifying a single contagion source when partial timestamps of a contagion process are available. We formulate the source localization problem as a ranking problem on graphs, where infected nodes are ranked according to their likelihood of being the source. Two ranking algorithms, cost-based ranking (CR) and tree-based ranking (TR), are proposed in this paper. Experimental evaluations with synthetic and real-world data show that our algorithms significantly improve the ranking accuracy compared with four existing algorithms.

12:00 - 12:20 A010: An Efficient Exact Algorithm for Triangle Listing in Large Graphs
Sofiane Lagraa, Hamida Seba

This paper presents a new efficient exact algorithm for listing triangles in a large graph. While the problem of listing triangles in a graph has been considered before, dealing with large graphs continues to be a challenge. Although previous research has attempted to tackle the challenge, this is the first contribution that addresses this problem on a compressed copy of the input graph. In fact, the proposed solution lists the triangles without decompressing the graph. This yields interesting improvements in both storage requirement of the graphs and their time processing.

TUESDAY SESSIONS, WITH ABSTRACTS

12:20 - 12:40 A011: A Distributed Approach for Graph Mining in Massive Networks.
Nilothpal Talukder, Mohammed J. Zaki.

We propose a novel distributed algorithm for mining frequent subgraphs from a single, very large, labeled network. Our approach is the first distributed method to mine a massive input graph that is too large to fit in the memory of any individual compute node. The input graph thus has to be partitioned among the nodes, which can lead to potential false negatives. Furthermore, for scalable performance it is crucial to minimize the communication among the compute nodes. Our algorithm, DistGraph, ensures that there are no false negatives, and uses a set of optimizations and efficient collective communication operations to minimize information exchange. To our knowledge DistGraph is the first approach demonstrated to scale to graphs with over a billion vertices and edges. Scalability results on up to 2048 IBM Blue Gene/Q compute nodes, with 16 cores each, show very good speedup.

12:40 - 13:00 A012: M-Flash: Fast Billion-Scale Graph Computation Using a Bimodal Block Processing Model
Hugo Gualdron Colmenares, Robson Leonardo Ferreira Cordeiro, José Fernando Rodrigues Jr., Duen Horng Chau, Minsuk Kahng, U Kang

Recent graph computation approaches have demonstrated that a single PC can perform efficiently on billion-scale graphs. While these approaches achieve scalability by optimizing I/O operations, they do not fully exploit the capabilities of modern hard drives and processors. To overcome their performance, in this work, we introduce the Bimodal Block Processing (BBP), an innovation that is able to boost the graph computation by minimizing the I/O cost even further. With this strategy, we achieved the following contributions: (1) M-Flash, the fastest graph computation framework to date; (2) a flexible and simple programming model to easily implement popular and essential graph algorithms, including the first single-machine billion-scale eigensolver; and (3) extensive experiments on real graphs with up to 6.6 billion edges, demonstrating M-Flash's consistent and significant speedup.

TUE1C: CLUSTERING 1 — ROOM 300A

11:00 - 11:20 A013: Infection hot spot mining from social media trajectories
Roberto Souza, Renato Assunção, Derick Oliveira, Denise Brito, Meira Wagner

Traditionally, in health surveillance, high risk zones are identified based only on the residence address or the working place of diseased individuals. This provides little information about the places where people are infected, the truly important information for disease control. The recent availability of spatial data generated by geotagged social media posts offers a unique opportunity: by identifying and following diseased individuals, we obtain a collection of sequential geo-located events, each sequence being issued by a social media user. The sequence of map positions implicitly provides an estimation of the users' social trajectories as they drift on the map. The existing data mining techniques for spatial cluster detection fail to address this new setting as they require a single location to each individual under analysis. In this paper we present two stochastic models with their associated algorithms to mine this new type of data. The Visit Model finds the most likely zones that a diseased person visits, while the Infection Model finds the most likely zones where a person gets infected while visiting. We demonstrate the applicability and effectiveness of our proposed models by applying them to more than 100 million geotagged tweets from Brazil in 2015. In particular, we target the identification of infection hot spots associated with dengue, a mosquito-transmitted disease that affects millions of people in Brazil annually, and billions worldwide. We applied our algorithms to data from 11 large cities in Brazil and found infection hot spots, showing the usefulness of our methods for disease surveillance.

11:20 - 11:40 A014: A Novel Incremental Covariance-guided One-class Support Vector Machine
Takoua Kefi, Riadh Ksantini, Mohamed Kaâniche, Adel Bouhoula

Covariance-guided One-Class Support Vector Machine (COSVM) is a very competitive kernel classifier, as it emphasizes the low variance projectional directions of the training data, which results in high accuracy. However, COSVM training involves solving a constrained convex optimization problem, which requires large memory and enormous amount of training time, especially for large scale datasets. Moreover, it has difficulties in classifying sequentially obtained data. For these reasons, this paper introduces an incremental COSVM method by controlling the possible changes of support vectors after the addition of new data points. The control procedure is based on the relationship between the Karush-Kuhn-Tuker conditions of COSVM and the distribution of the training set. Comparative experiments have been carried out to show the effectiveness of our proposed method, both in terms of execution time and classification accuracy. Incremental COSVM results in better classification performance when compared to canonical COSVM and contemporary incremental one-class classifiers.

11:40 - 12:00 A015: Incremental Commute Time Using Random Walks and Online Anomaly Detection
Khoa Nguyen, Sanjay Chawla

Commute time is a random walk based metric on graphs and has found widespread successful applications in many application domains. However, the computation the commute time is expensive, involving the eigen decomposition of the graph Laplacian matrix. There has

been effort to approximate the commute time in offline mode. Our interest is inspired by the use of commute time in online mode. We propose an accurate and efficient approximation for computing the commute time in an incremental fashion in order to facilitate real-time applications. An online anomaly detection technique is designed where the commute time of each new arriving data point to any point in the current graph can be estimated in constant time ensuring a real-time response. The proposed approach shows its high accuracy and efficiency in many synthetic and real datasets and takes only 8 milliseconds on average to detect anomalies online on the DBLP graph which has more than 600,000 nodes and 2 millions edges.

12:00 - 12:20 A016: Node Re-Ordering as a Means of Anomaly Detection in Time-Evolving Graphs

Lida Rashidi, Christopher Leckie, Sutharshan Rajasegarar, Andrey Kan, Jeffrey Chan, James Bailey

Anomaly detection is a vital task for maintaining and improving any dynamic system. In this paper, we address the problem of anomaly detection in time-evolving graphs, where graphs are a natural representation for data in many types of applications. A key challenge in this context is how to process large volumes of streaming graphs. We propose a pre-processing step before running any further analysis on the data, where we permute the rows and columns of the adjacency matrix. This pre-processing step expedites graph mining techniques such as anomaly detection, PageRank, graph coloring and visualization. We can then detect graph anomalies based on rank correlations of the reordered nodes. The merits of our approach lie in its simplicity and resilience to challenges such as unsupervised input, large volumes and high velocities of data. We evaluate the scalability and accuracy of our method on real graphs, where our method facilitates graph processing while producing more deterministic orderings. We show that the proposed approach is capable of revealing anomalies in a more efficient manner based on node rankings. Furthermore, our method can produce visual representations of graphs that are useful for graph compression.

12:20 - 12:40 A017: Clustering based on Interior-Node Topology with Applications to Brain Networks

Guixiang Ma, Lifang He, Bokai Cao, Jiawei Zhang, Philip Yu

Learning from graph data has been attracting much attention recently due to its importance in many scientific applications, where objects are represented as graphs. In this paper, we study the problem of multi-graph clustering i.e., clustering multiple graphs).We propose a multi-graph clustering approach (MGCT) based on the interior-node topology of graph.Specifically, we extract the interior-node topological structure of each graph through a sparsity-inducing interior-node clustering. We merge the interior-node clustering stage and the multi-graph clustering stage into a unified iterative framework, where the multi-graph clustering will influence the interior-node clustering and the updated interior-node clustering results will be further exerted on multi-graph clustering. We apply MGCT on two real brain network data sets i.e.,, ADHD and HIV). Experimental results demonstrate the superior performance of the proposed model on multi-graph clustering.

12:40 - 13:00 A018: A Split-Merge DP-means Algorithm to Avoid Local Minima

Shigeyuki Odashima, Miwa Ueki, Naoyuki Sawasaki

We present an extension of the DP-means algorithm, a hard-clustering approximation of nonparametric Bayesian models. Although a recent work reports that the DP-means can converge to a local minimum, the condition for the DP-means to converge to a local minimum is still unknown. This paper demonstrates one reason the DP-means converges to a local minimum: the DP-means cannot assign the optimal number of clusters when many data points exist within small distances. As a First attempt to avoid the local minimum, we propose an extension of the DP-means by the split-merge technique. The proposed algorithm splits clusters when a cluster has many data points to assign the number of clusters near to optimal. The experimental results with multiple datasets show the robustness of the proposed algorithm.

TUE1D: LEARNING

ROOM 300B

11:00 - 11:20 A019: Efficient Distributed Decision Trees for Robust Regression

Tian Guo, Konstantin Kutzkov, mohamed Ahmed, jean-paul Calbimonte, Karl Aberer

The availability of massive volumes of data and recent advances in data collection and processing platforms have motivated the development of distributed machine learning algorithms. In numerous real-world applications large datasets are inevitably noisy and contain outliers. These outliers can dramatically degrade the performance of standard machine learning approaches such as regression trees. To this end, we present a novel distributed regression tree approach that utilizes robust regression statistics, statistics that are more robust to outliers, for handling large and noisy data. We propose to integrate robust statistics based error criteria into the regression tree. A data summarization method is developed and used to improve the efficiency of learning regression trees in the distributed setting. We implemented the proposed approach and baselines based on Apache Spark, a popular distributed data processing platform. Extensive experiments on both synthetic and real datasets verify the effectiveness and efficiency of our approach.

TUESDAY SESSIONS, WITH ABSTRACTS

11:20 - 11:40 **A020: On the Convergence of A Family of Robust Losses for Stochastic Gradient Descent**
Bo Han, Ivor Tsang, Ling Chen

The convergence of Stochastic Gradient Descent (SGD) using convex loss functions has been widely studied. However, vanilla SGD methods using convex losses cannot perform well with noisy labels, which adversely affect the update of the primal variable in SGD methods. Unfortunately, noisy labels are ubiquitous in real world applications such as crowdsourcing. To handle noisy labels, in this paper, we present a family of robust losses for SGD methods. By employing our robust losses, SGD methods successfully reduce negative effects caused by noisy labels on each update of the primal variable. We not only reveal the convergence rate of SGD methods using robust losses, but also provide the robustness analysis on two representative robust losses. Comprehensive experimental results on six real-world datasets show that SGD methods using robust losses are obviously more robust than other baseline methods in most situations with fast convergence.

11:40 - 12:00 **A021: Actively Interacting with Experts: A Probabilistic Logic Approach**
Phillip Odom, Sriraam Natarajan

Machine learning approaches that utilize human experts combine domain experience with data to generate novel knowledge. Unfortunately, most methods either provide only a limited form of communication with the human expert and/or are overly reliant on the human expert to specify their knowledge upfront. Thus, the expert is unable to understand what the system could learn without their involvement. Allowing the learning algorithm to query the human expert in the most useful areas of the feature space takes full advantage of the data as well as the expert. We introduce active advice-seeking for relational domains. Relational logic allows for compact, but expressive interaction between the human expert and the learning algorithm. We demonstrate our algorithm empirically on several standard relational datasets.

12:00 - 12:20 **A022: Stochastic CoSaMP: Randomizing Greedy Pursuit for Sparse Signal Recovery**
Dipan Pal, OLE MENGSHOEL

In this paper, we formulate the K-sparse compressed signal recovery problem with the L0 norm within a Stochastic Local Search (SLS) framework. Using this randomized framework, we generalize the popular sparse recovery algorithm CoSaMP, creating Stochastic CoSaMP (StoCoSaMP). Interestingly, our deterministic worst case analysis shows that under the Restricted Isometric Property (RIP), even a purely random version of StoCoSaMP is guaranteed to recover a notion of strong components of a sparse signal, thereby leading to support convergence. Empirically, we find that StoCoSaMP outperforms CoSaMP, both in terms of signal recoverability and computational cost, on different problems with up to 1 million dimensions. Further, StoCoSaMP outperforms several other popular recovery algorithms (including StoGradMP and StoIHT) on large scale real-world gene-expression datasets.

12:20 - 12:40 **A023: OSLa: Online Structure Learning using Background Knowledge Axiomatization**
Evangelos Michelioudakis, Anastasios Skarlatidis, George Paliouras, Alexander Artikis

We present OSLa --- an online structure learner for Markov Logic Networks (MLNs) that exploits background knowledge axiomatization in order to constrain the space of possible structures. Many domains of interest are characterized by uncertainty and complex relational structure. MLNs is a state-of-the-art Statistical Relational Learning framework that can naturally be applied to domains governed by these characteristics. Learning MLNs from data is challenging, as their relational structure increases the complexity of the learning process. In addition, due to the dynamic nature of many real-world applications, it is desirable to incrementally learn or revise the model's structure and parameters. Experimental results are presented in activity recognition using a probabilistic variant of the Event Calculus (MLN-EC) as background knowledge and a benchmark dataset for video surveillance.

12:40 - 13:00 **A024: Lifted generative learning of Markov logic networks**
Jan Van Haaren, Guy Van den Broeck, Wannes Meert, Jesse Davis

Markov logic networks (MLNs) are a well-known statistical relational learning formalism that combines Markov networks with first-order logic. MLNs attach weights to formulas in first-order logic. Learning MLNs from data is a chal- lenging task as it requires searching through the huge space of possible theories. Additionally, evaluating a theory's likelihood requires learning the weight of all formulas in the theory. This in turn requires performing probabilistic inference, which, in general, is intractable in MLNs. Lifted inference speeds up probabilistic inference by exploiting symmetries in a model. We explore how to use lifted infer- ence when learning MLNs. Specifically, we investigate generative learning where the goal is to maximize the likelihood of the model given the data. First, we pro- vide a generic algorithm for learning maximum likelihood weights that works with any exact lifted inference approach. In contrast, most existing approaches opti- mize approximate measures such as the pseudo-likelihood. Second, we provide a concrete parameter learning algorithm based on first-order knowledge compila- tion. Third, we propose a structure learning algorithm that learns liftable MLNs, which is the first MLN structure learning algorithm that exactly optimizes the likelihood of the model. Finally, we perform an empirical evaluation on three

real- world datasets. Our parameter learning algorithm results in more accurate models than several competing approximate approaches. It learns more accurate models in terms of test-set log-likelihood as well as prediction tasks. Furthermore, our tractable learner outperforms intractable models on prediction tasks suggesting that liftable models are a powerful hypothesis space, which may be sufficient for many standard learning problems.

TUE1E: NECTAR 1

ROOM BELVEDERE

ML/DM in engineering and natural and behavioural sciences/applications

11:00 - 11:20 A025: Practical Bayesian Inverse Reinforcement Learning for Robot Navigation

Billy Okal, Kai Arras

IRL provides a concise framework for learning behaviors from human demonstrations; and is highly desired in practical and difficult to specify tasks such as normative robot navigation. However, most existing IRL algorithms are often ladened with practical challenges such as representation mismatch and poor scalability when deployed in real world tasks. Moreover, standard RL representations often do not allow for incorporation of task constraints common for example in robot navigation. In this paper, we present an approach that tackles these challenges in a unified manner and delivers a learning setup that is both practical and scalable. We develop a graph-based spare representation for RL and a scalable IRL algorithm based on sampled trajectories. Experimental evaluation in simulation and from a real deployment in a busy airport demonstrate the strengths of the learning setup over existing approaches.

11:20 - 11:40 A026: Learning from Software Project Histories

Verena Honsel, Steffen Herbold, Jens Grabowski

In software project planning project managers have to keep track of several things simultaneously including the estimation of the consequences of decisions about, e.g., the team constellation. The application of machine learning techniques to predict possible outcomes is a widespread research topic in software engineering. In this paper, we summarize our work in the field of learning from project history.

11:40 - 12:00 A027: Resource-Aware Steel Production Through Data Mining

Hendrik Blom, Katharina Morik

Today's steel industry is characterized by overcapacity and increasing competitive pressure. There is a need for continuously improving processes, with a focus on consistent enhancement of efficiency, improvement of quality and thereby better competitiveness. About 70% of steel is produced using the BF-BOF (Blast Furnace - Blow Oxygen Furnace) route worldwide. The BOF is the first step of controlling the composition of the steel and has an impact on all further processing steps and the overall quality of the end product. Multiple sources of process-related variance and overall harsh conditions for sensors and automation systems in general lead to a process complexity that is not easy to model with thermodynamic or metallurgical approaches. In this paper we want to give an insight how to improve the output quality with machine learning based modeling and which constraints and requirements are necessary for an online application in real-time.

12:00 - 12:20 A028: Machine learning challenges for single cell data

Sofie Van Gassen, Tom Dhaene, Yvan Saeys

Recent technological advances in the fields of biology and medicine allow measuring single cells into unprecedented depth. This results in new types of high-throughput datasets that shed new lights on cell development, both in healthy as well as diseased tissues. However, studying these biological processes into greater detail crucially depends on novel computational techniques that efficiently mine single cell data sets. In this paper, we introduce machine learning techniques for single cell data analysis: we summarize the main developments in the field, and highlight a number of interesting new avenues that will likely stimulate the design of new types of machine learning algorithms.

12:20 - 12:40 A029: Query Log Mining for Inferring User Tasks and Needs

Rishabh Mehrotra, Emine Yilmaz

Search behavior, and information seeking behavior more generally, is often motivated by tasks that prompt search processes that are often lengthy, iterative, and intermittent, and are characterized by distinct stages, shifting goals and multitasking. Current search systems do not provide adequate support for users tackling complex tasks due to which the cognitive burden of keeping track of such tasks is placed on the searcher. In this note, we summarize our recent efforts towards extracting search tasks from search logs. Based on recent advancements in Bayesian Nonparametrics and distributional semantics, we propose novel algorithms to extract task and subtasks from a query collection. The models discussed can inform the design of the next generation of task-based search systems that leverage user's task behavior for better support and personalization.

12:40 - 13:00 A030: Data Mining Meets HCI: Data and Visual Analytics of Frequent Patterns

C Leung, Christopher Carmichael , Yaroslav Hayduk , Fan Jiang

As a popular data mining tasks, frequent pattern mining discovers implicit, previously unknown and potentially useful knowledge in the form of sets of frequently co-occurring items or events. Many existing data mining algorithms return to users with long textual lists of

TUESDAY SESSIONS, WITH ABSTRACTS

frequent patterns, which may not be easily comprehensible. As a picture is worth a thousand words, having a visual means for humans to interact with computers would be beneficial. This is when human-computer interaction (HCI) research meets data mining research. In particular, the popular HCI task of data and result visualization could help data miners to visualize the original data and to analyze the mined results (in the form of frequent patterns). In this paper, we present a few systems for data and visual analytics of frequent patterns, which integrate (i) data analytics and mining with (ii) data and result visualization.

TUE2A: CLASSIFICATION 1 — ROOM 1000A

14:50 - 15:10 **A031: F-Measure Maximization in Multi-Label Classification with Conditionally Independent Label Subsets**
Maxime Gasse, Alex Aussem

We discuss a method to improve the exact F-measure maximization algorithm called GFM, proposed in [DembczynskiWCH11] for multi-label classification, assuming the label set can be partitioned into conditionally independent subsets given the input features. If the labels were all independent, the estimation of only m parameters (m denoting the number of labels) would suffice to derive Bayes-optimal predictions in $O(m^2)$ operations [NanCLC12]. In the general case, $m^2 + 1$ parameters are required by GFM, to solve the problem in $O(m^3)$ operations. In this work, we show that the number of parameters can be reduced further to m^2/n, in the best case, assuming the label set can be partitioned into n conditionally independent subsets. As this label partition needs to be estimated from the data beforehand, we use first the procedure proposed in [GasseAE15] that finds such partition and then infer the required parameters locally in each label subset. The latter are aggregated and serve as input to GFM to form the Bayes-optimal prediction. We show on a synthetic experiment that the reduction in the number of parameters brings about significant benefits in terms of performance.

15:10 - 15:30 **A032: ALRn: Accelerated Higher-order Logistic Regression**
Nayyar Zaidi, Geoffrey Webb, Francois Petitjean, Mark Carman, Jesus Crequides

This paper introduces Accelerated Logistic Regression: a hybrid generative- discriminative approach to training Logistic Regression with high-order features. We present two main results: (1) that our combined generative-discriminative approach significantly improves the efficiency of Logistic Regression and (2) that incorporating higher order features (i.e. features that are the Cartesian products of the original features) reduces the bias of Logistic Regression, which in turn significantly reduces its error on large datasets. We assess the efficacy of Accelerated Logistic Regression by conducting an extensive set of experiments on 75 standard datasets. We demonstrate its competitiveness, particularly on large datasets, by comparing against state-of-the-art classifiers including Random Forest and Averaged n-Dependence Estimators.

15:30 - 15:50 **A033: Beyond the Boundaries of SMOTE: A Framework for Manifold-Based Synthetically Oversampling**
Colin Bellinger, Chris Drummond, Nathalie Japkowicz

Problems of class imbalance appear in diverse domains, ranging from gene function annotation to spectra and medical classification. On such problems, the classifier becomes biased in favour of the majority class. This leads to inaccuracy on the important minority classes, such as specific diseases and gene functions. Synthetic oversampling mitigates this by balancing the training set, whilst avoiding the pitfalls of random under and oversampling. The existing methods are primarily based on the SMOTE algorithm, which employs a bias of randomly generating points between nearest neighbours. The relationship between the generative bias and the latent distribution has a significant impact on the performance of the induced classifier. Our research into gamma-ray spectra classification has shown that the generative bias applied by SMOTE is inappropriate for domains that conform to the manifold property, such as spectra, text, image and climate change classification. To this end, we propose a framework for manifold-based synthetic oversampling, and demonstrate its superiority in terms of robustness to the manifold with respect to the AUC on three spectra classification tasks and 16 UCI datasets.

15:50 - 16:10 **A034: CHADE: Metalearning with Classifier Chains for Dynamic Combination of Classifiers**
Fábio Pinto, Carlos Soares, Joao Moreira

Dynamic selection or combination (DSC) methods allow to select one or more classifiers from an ensemble according to the characteristics of a given test instance x. Most methods proposed for this purpose are based on the nearest neighbors algorithm: it is assumed that if a classifier performed well on a set of instances similar to x, it will also perform well on x. We address the problem of dynamically combining a pool of classifiers by combining two approaches: metalearning and multi-label classification. Taking into account that diversity is a fundamental concept in ensemble learning and the interdependencies between the classifiers cannot be ignored, we solve the multi-label classification problem by using a widely known technique: Classifier Chains (CC). Additionally, we extend a typical metalearning approach by combining metafeatures characterizing the interdependencies between the classifiers with the base-level features. We executed experiments on 42 classification datasets and compared our method with several state-of-the-art DSC techniques, including another metalearning approach. Results show that our method allows an improvement over the other metalearning approach and is very competitive with other four DSC methods.

TUE2B: OPTIMIZATION

ROOM 1000B

14:50 - 15:10 **A035: Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition**

Hamed Karimi, Julie Nutini, Mark Schmidt

In 1963, Polyak proposed a simple condition that is sufficient to show a global linear convergence rate for gradient descent. This condition is a special case of the Lojasiewicz inequality proposed in the same year, and it does not require strong convexity (or even convexity). In this work, we show that this much-older Polyak-Lojasiewicz (PL) inequality is actually weaker than the main conditions that have been explored to show linear convergence rates without strong convexity over the last 25 years. We also use the PL inequality to give new analyses of coordinate descent and stochastic gradient for many non-strongly-convex (and some non-convex) functions. We further propose a generalization that applies to proximal-gradient methods for non-smooth optimization, leading to simple proofs of linear convergence for support vector machines and L1-regularized least squares without additional assumptions.

15:10 - 15:30 **A036: Fast and Scalable Lasso via Stochastic Frank-Wolfe Methods with a Convergence Guarantee**

Emanuele Frandi, Ricardo Ñanculef, Stefano Lodi, Claudio Sartori, Johan A. K. Suykens

Frank-Wolfe (FW) algorithms have been often proposed over the last few years as efficient solvers for a variety of optimization problems arising in the field of Machine Learning. The ability to work with cheap projection-free iterations and the incremental nature of the method make FW a very effective choice for many large-scale problems where computing a sparse model is desirable. In this paper, we present a high-performance implementation of the FW method tailored to solve large-scale Lasso regression problems, based on a randomized iteration, and prove that the convergence guarantees of the standard FW method are preserved in the stochastic setting. We show experimentally that our algorithm outperforms several existing state of the art methods, including the Coordinate Descent algorithm by Friedman et al. (one of the fastest known Lasso solvers), on several benchmark datasets with a very large number of features, without sacrificing the accuracy of the model. Our results illustrate that the algorithm is able to generate the complete regularization path on problems of size up to four million variables in less than one minute.

15:30 - 15:50 **A037: Two-Stage Transfer Surrogate Model for Automatic Hyperparameter Optimization**

Martin Wistuba, Nicolas Schilling, Lars Schmidt-Thieme

The choice of hyperparameters and the selection of algorithms is a crucial part in machine learning. Bayesian optimization methods have been used very successfully to tune hyperparameters automatically, in many cases even being able to outperform the human expert. Recently, these techniques have been massively improved by using meta-knowledge. The idea is to use knowledge of the performance of an algorithm on given other data sets to automatically accelerate the hyperparameter optimization for a new data set. In this work we present a model that transfers this knowledge in two stages. At the first stage, the function that maps hyperparameter configurations to hold-out validation performances is approximated for previously seen data sets. At the second stage, these approximations are combined to rank the hyperparameter configurations for a new data set. In extensive experiments on the problem of hyperparameter optimization as well as the problem of combined algorithm selection and hyperparameter optimization, we are outperforming the state of the art methods.

15:50 - 16:10 **A038: Scalable Hyperparameter Optimization with Products of Gaussian Process Experts**

Nicolas Schilling, Martin Wistuba, Larst Schmidt-Thieme

In machine learning, hyperparameter optimization is a challenging but necessary task that is usually approached in a computationally expensive manner such as grid-search. Out of this reason, surrogate based black-box optimization techniques such as sequential model-based optimization have been proposed which allow for a faster hyperparameter optimization. Recent research proposes to also integrate hyperparameter performances on past data sets to allow for a faster and more efficient hyperparameter optimization. In this paper, we use products of Gaussian process experts as surrogate models for hyperparameter optimization. Naturally, Gaussian processes are a decent choice as they offer good prediction accuracy as well as estimations about their uncertainty. Additionally, their hyperparameters can be tuned very effectively. However, in the light of large meta data sets, learning a single Gaussian process is not feasible as it involves inversion of a large kernel matrix. This directly limits their usefulness for hyperparameter optimization if large scale hyperparameter performances on past data sets are given. By using products of Gaussian process experts the scalability issues can be circumvened, however, this usually comes with the price of having less predictive accuracy. In our experiments, we show empirically that products of experts nevertheless perform very well compared to a variety of published surrogate models. Thus, we propose a surrogate model that performs as well as the current state of the art, is scalable to large scale meta knowledge, does not include hyperparameters itself and finally is even very easy to parallelize.

TUESDAY SESSIONS, WITH ABSTRACTS

TUE2C: TOPIC MODELLING

ROOM 300A

14:50 - 15:10 A039: Detecting Public Influence on News Using Topic-aware Dynamic Granger Test

Lei Hou, Juanzi Li, Xiao-Li Li, Jianbin Jin

With the rapid proliferation of Web 2.0, user-generated content (UGC), which is formed by the public to reflect their views and voice, presents rich and timely feedback on news events. Existing research either studies the common and private features between news and UGC, or describes the ability of news media to influence the public opinion. However, in the current highly media-user interactive environment, investigating the public influence on news is of great significance to risk and credible management for government and enterprises. In this paper, we propose a novel topic-aware dynamic Granger test framework to quantify and characterize the public influence on news. In particular, we represent words and documents as distributed low-dimensional vectors which facilitates the subsequent topic extraction. Then, a topic-aware dynamic strategy is proposed to transfer news and UGC streams into topic series, and finally we apply Granger causality test to investigate the public influence on news. Extensive experiments on 45 diverse real-world events demonstrate the effectiveness of the proposed method, and the results show promising prospects on predicting whether an event will be properly handled at its early stage.

15:10 - 15:30 A040: C-BiLDA: Extracting Cross-lingual Topics from Non-Parallel Texts by Distinguishing Shared from Unshared Content

Geert Heyman, Ivan Vulic, Marie-Francine Moens

We study the problem of extracting cross-lingual topics from non-parallel multilingual text datasets with partially overlapping thematic content (e.g., aligned Wikipedia articles in two different languages). To this end, we develop a new bilingual probabilistic topic model called comparable bilingual latent Dirichlet allocation (C- BiLDA), which is able to deal with such comparable data, and, unlike the standard bilingual LDA model (BiLDA), does not assume the availability of document pairs with identical topic distributions. We present a full overview of C-BiLDA, and show its utility in the task of cross-lingual knowledge transfer for multi-class document classification on two benchmarking datasets for three language pairs. The proposed model outperforms the baseline LDA model, as well as the standard BiLDA model and two standard low-rank approximation methods (CL-LSI and CL-KCCA) used in previous work on this task.

15:30 - 15:50 A041: Learning beyond Predefined Label Space via Bayesian Nonparametric Topic Modelling

Changying Du, Fuzhen Zhuang, Jia He, Qing He, Guoping Long

In real world machine learning applications, testing data may contain some meaningful new categories that have not been seen in labeled training data. To simultaneously recognize new data categories and assign most appropriate category labels to the data actually from known categories, existing models assume the number of unknown new categories is pre-specified, though it is difficult to determine in advance. In this paper, we propose a Bayesian nonparametric topic model to automatically infer this number, based on the hierarchical Dirichlet process and the notion of latent Dirichlet allocation. Exact inference in our model is intractable, so we provide an efficient collapsed Gibbs sampling algorithm for approximate posterior inference. Extensive experiments on various text data sets show that: (a) compared with parametric approaches that use pre-specified true number of new categories, the proposed nonparametric approach can yield comparable performance; and (b) when the exact number of new categories is unavailable, i.e. the parametric approaches only have a rough idea about the new categories, our approach has evident performance advantages.

15:50 - 16:10 A042: Aspect Mining with Rating Bias

Yitong Li, Chuan Shi, Huidong Zhao, Fuzhen Zhuang, Bin Wu

Due to the personalized needs for specific aspect evaluation on product quality, these years have witnessed a boom of researches on aspect rating prediction, whose goal is to extract ad hoc aspects from online reviews and predict rating or opinion on each aspect. Most of the existing works on aspect rating prediction have a basic assumption that the overall rating is the average score of aspect ratings or the overall rating is very close to aspect ratings. However, after analyzing real datasets, we have an insightful observation: there is an obvious rating bias between overall rating and aspect ratings. Motivated by this observation, we study the problem of aspect mining with rating bias, and design a novel RAting-center model with BIas (RABI). Different from the widely used review-center models, RABI adopts the overall rating as the center of the probabilistic model, which generates reviews and topics. In addition, a novel aspect rating variable in RABI is designed to effectively integrate the rating bias priori information. Experiments on two real datasets (Dianping and TripAdvisor) validate that RABI significantly improves the prediction accuracy over existing state-of-the-art methods.

TUE2D: PATTERNS

ROOM 300B

14:50 - 15:10 A043: Subgroup Discovery with Proper Scoring Rules

Hao Song, Meelis Kull, Peter Flach, Georgios Kalogridis

Subgroup Discovery is the process of finding and describing sufficiently large subsets of a given population that have unusual distributional characteristics with regard to some target attribute. Such subgroups can be used as a statistical summary which improves on the default summary of stating the overall distribution in the population. A natural way to evaluate such summaries is to quantify the difference between predicted and empirical distribution of the target. In this paper we propose to use proper scoring rules, a well-known family of evaluation measures for assessing the goodness of probability estimators, to obtain theoretically well-founded evaluation measures for subgroup discovery. From this perspective, one subgroup is better than another if it has lower divergence of target probability estimates from the actual labels on average. We demonstrate empirically on both synthetic and real-world data that this leads to higher quality statistical summaries than the existing methods based on measures such as Weighted Relative Accuracy.

15:10 - 15:30 A044: A Bayesian Network Model for Interesting Itemsets

Jari Fowkes, Charles Sutton

Mining itemsets that are the most interesting under a statistical model of the underlying data is a commonly used and well-studied technique for exploratory data analysis, with the most recent interestingness models exhibiting state of the art performance. Continuing this highly promising line of work, we propose the first, to the best of our knowledge, generative model over itemsets, in the form of a Bayesian network, and an associated novel measure of interestingness. Our model is able to efficiently infer interesting itemsets directly from the transaction database using structural EM, in which the E-step employs the greedy approximation to weighted set cover. Our approach is theoretically simple, straightforward to implement, trivially parallelizable and retrieves itemsets whose quality is comparable to, if not better than, existing state of the art algorithms as we demonstrate on several real-world datasets.

15:30 - 15:50 A045: Mining Rooted Ordered Trees under Subtree Homeomorphism

Mostafa Haghir Chehreghani, Maurice Bruynooghe

Mining frequent tree patterns has many applications in different areas such as XML data, bioinformatics and World Wide Web. The crucial step in frequent pattern mining is frequency counting, which involves a matching operator to find occurrences (instances) of a tree pattern in a given collection of trees. A widely used matching operator for tree-structured data is subtree homeomorphism, where an edge.

15:50 - 16:10 A046: Transactional Tree Mining

Mostafa Haghir Chehreghani, Morteza Haghir Chehreghani

In the transactional setting of finding frequent embedded patterns from a large collection of tree-structured data, the crucial step is to decide whether a tree pattern is subtree homeomorphic to a database tree. Our extensive study on the properties of real-world tree-structured datasets reveals that while many vertices in a database tree may have the same label, no two vertices on the same path are identically labeled. In this paper, we exploit this property and propose a novel and efficient method for deciding whether a tree pattern is subtree homeomorphic to a database tree. Our algorithm is based on a compact data-structure, called EMET, that stores all information required for subtree homeomorphism. We propose an efficient algorithm to generate EMETs of larger patterns using EMETs of the smaller ones. Based on the proposed subtree homeomorphism method, we introduce TTM, an effective algorithm for finding frequent tree patterns from rooted ordered trees. We evaluate the efficiency of TTM on several real-world and synthetic datasets and show that it outperforms well-known existing algorithms by an order of magnitude.

TUE2E: NECTAR 2

ROOM BELVEDERE

ML/DM in systems and social sciences/applications

14:50 - 15:10 A047: Time and Again: Time Series Mining via Recurrence Quantification Analysis

Stephan Spiegel, Norbert Marwan

Recurrence quantification analysis (RQA) was developed in order to quantify differently appearing recurrence plots (RPs) based on their small-scale structures, which generally indicate the number and duration of recurrences in a dynamical system. Although RQA measures are traditionally employed in analyzing complex systems and identifying transitions, recent work has shown that they can also be used for pairwise dissimilarity comparisons of time series. We explain why RQA is not only a modern method for nonlinear data analysis but also is a very promising technique for various time series mining tasks.

TUESDAY SESSIONS, WITH ABSTRACTS

15:10 - 15:30 A048: Personality-Based User Modeling for Music Recommender Systems
Bruce Ferwerda, Markus Schedl

Applications are getting increasingly interconnected. Although the interconnectedness provide new ways to gather information about the user, not all user information is ready to be directly implemented in order to provide a personalized experience to the user. Therefore, a general model is needed to which users' behavior, preferences, and needs can be connected to. In this paper we present our works on a personality-based music recommender system in which we use users' personality traits as a general model. We identified relationships between users' personality and their behavior, preferences, and needs, and also investigated different ways to infer users' personality traits from user-generated data of social networking sites (i.e., Facebook, Twitter, and Instagram). Our work contributes to new ways to mine and infer personality-based user models, and show how these models can be implemented in a music recommender system to positively contribute to the user experience.

15:30 - 15:50 A049: Local Exceptionality Detection on Social Interaction Networks
Martin Atzmueller

Local exceptionality detection on social interaction networks includes the analysis of resources created by humans (e.g. social media) as well as those generated by sensor devices in the context of (complex) interactions. This paper provides a structured overview on a line of work comprising a set of papers that focus on data-driven exploration and modeling in the context of social network analysis, community detection and pattern mining.

15:50 - 16:10 A050: Machine Learning for Crowdsourced Spatial Data
Musfira Jilani, Padraig Corcoran, Michela Bertolotto

Recent years have seen a significant increase in the number of applications requiring accurate and up-to-date spatial data. In this context crowdsourced maps such as OpenStreetMap (OSM) have the potential to provide a free and timely representation of our world. However, one factor that negatively influences the proliferation of these maps is the uncertainty about their data quality. This paper presents structured and unstructured machine learning methods to automatically assess and improve the semantic quality of streets in the OSM database.

TUE3A: KERNELS — ROOM 1000A

16:40 - 17:00 A051: Improving Locality Sensitive Hashing Based Similarity Search and Estimation for Kernels
Aniket Chakrabarti, Bortik Bandyopadhyay, Srinivasan Parthasarathy

We present a novel data embedding that significantly reduces the estimation error of locality sensitive hashing (LSH) technique when used in reproducing kernel Hilbert space (RKHS). Efficient and accurate kernel approximation techniques either involve the kernel principal component analysis (KPCA) approach or the Nyström approximation method. In this work we show that extant LSH methods in this space suffer from a bias problem, that moreover is difficult to estimate apriori. Consequently, the LSH estimate of a kernel is different from that of the KPCA/Nyström approximation. We provide theoretical rationale for this bias, which is also confirmed empirically. We propose an LSH algorithm that can reduce this bias and consequently our approach can match the KPCA or the Nyström methods' estimation accuracy while retaining the traditional benefits of LSH. We evaluate our algorithm on a wide range of realworld image datasets (for which kernels are known to perform well) and show the efficacy of our algorithm using a variety of principled evaluations including mean estimation error, KL divergence and the Kolmogorov-Smirnov test.

17:00 - 17:20 A052: Continuous Kernel Learning
John Moeller, Vivek Srikumar, Sarathkrishna Swaminathan, Suresh Venkatasubramanian, Dustin Webb

Kernel learning is the problem of determining the best kernel (either from a dictionary of fixed kernels, or from a smooth space of kernel representations) for a given task. In this paper, we describe a new approach to kernel learning that establishes connections between the Fourier-analytic representation of kernels arising out of Bochner's theorem and a specific kind of feed-forward network using cosine activations. We analyze the complexity of this space of hypotheses and demonstrate empirically that our approach provides scalable kernel learning superior in quality to prior approaches.

17:20 - 17:40 A053: Robust Dictionary Learning on the Hilbert Sphere in Kernel Feature Space
Nishanth Koushik, Suyash Awate

This paper presents a novel dictionary learning method in kernel feature space that is part of a reproducing kernel Hilbert space (RKHS). Our method focuses on several popular kernels, e.g., radial basis function kernels like the Gaussian, that implicitly map data to a Hilbert sphere, a Riemannian manifold, in RKHS. Our method exploits this manifold structure of the mapped data in RKHS, unlike typical

methods for kernel dictionary learning that use linear methods in RKHS. We show that dictionary learning on a Hilbert sphere in RKHS is possible without the need of the explicit lifting map underlying the kernel, but using solely the Gram matrix. Unlike the typical L1 norm sparsity prior, we incorporate the non-convex Lp quasi-norm based penalty, with $p < 1$, on coefficients to enforce a stronger sparsity prior and achieve more robust dictionary learning in the presence of corrupted training data. We evaluate our method for image classification on two large publicly available datasets and demonstrate the improved performance of our method over the state of the art dictionary learning methods.

17:40 - 18:00 **A054: Huber-Norm Support Vector Machines**
Oleksandr Zadorozhnyi, Gundhart Benecke, Stephan Mandt, Tobias Scheffer, Marius Kloft

In order to avoid overfitting, it is a common practice to regularize linear prediction models using squared or absolute-value norms of the model parameters. In our article we consider a new method of regularization: Huber-norm regularization imposes a combination of L1 and L2-norm regularization on the model parameters. We derive the dual optimization problem, prove an upper bound on the statistical risk of the model class by means of the Rademacher complexity and establish a simple type of oracle inequality on the optimality of the decision rule. Empirically, we observe that logistic regression with Huber-norm regularizer outperforms L1-norm, L2-norm, and elastic-net regularization for a wide range of benchmark data sets.

18:00 - 18:20 **A055: Efficient Bayesian Maximum Margin Multiple Kernel Learning**
Changying Du, Changde Du, Guoping Long, Xin Jin, Yucheng Li

Multiple Kernel Learning (MKL) suffers from slow learning speed and poor generalization ability. Existing methods seldom address these problems well simultaneously. In this paper, by defining a multiclass (psudo-) likelihood function that accounts for the margin loss for kernelized classification, we develop a robust Bayesian maximum margin MKL framework with Dirichlet and the three parameter Beta normal priors imposed on the kernel and sample combination weights respectively. For inference, we exploit the data augmentation idea and devise an efficient MCMC algorithm in the augmented variable space, employing the Riemann manifold Hamiltonian Monte Carlo technique to sample from the conditional posterior of kernel weights, and making use of local conjugacy for all other variables. Such geometry and conjugacy based posterior sampling leads to very fast mixing rate and scales linearly with the number of kernels used. Extensive experiments on classification tasks validate the superiority of the proposed method in both efficacy and efficiency.

TUE3B: PROBABILISTIC LEARNING — ROOM 1000B

16:40 - 17:00 **A056: An Online Gibbs Sampler Algorithm for Hierarchical Dirichlet Processes Prior**
Yongdai Kim, Minwoo Chae, Kuhwan Jeong, Byungyup Kang, Hyoju Chung

The hierarchical Dirichlet processes (HDP) is a Bayesian nonparametric model that provides a flexible mixed-membership to documents. In this paper, we develop a novel mini-batch online Gibbs sampler algorithm for the HDP which can be easily applied to massive and streaming data. For this purpose, a new prior process so called the generalized hierarchical Dirichlet processes (gHDP) is proposed. The gHDP is an extension of the standard HDP where some prespecified topics can be included in the top-level Dirichlet process. By analyzing various datasets, we show that the proposed mini-batch online Gibbs sampler algorithm performs significantly better than the online variational algorithm for the HDP.

17:00 - 17:20 **A057: Gaussian Process Pseudo-Likelihood Models for Sequence Labeling**
Srijith P.K., Balamurugan P, Shirish Shevade

Several machine learning problems arising in natural language processing can be modelled as a sequence labelling problem. Gaussian processes (GPs) provide a Bayesian approach to learning such problems in a kernel based framework. We develop Gaussian process models based on a pseudo-likelihood to solve sequence labelling problems. The pseudo-likelihood model enables one to capture multiple dependencies among the output components of the sequence without becoming computationally intractable. We use an efficient variational Gaussian approximation method to perform inference in the proposed model. We also provide an iterative algorithm which can effectively make use of the information from the neighbouring labels to perform prediction. The ability to capture multiple dependencies makes the proposed approach useful for a wide range of sequence labelling problems. Numerical experiments on some sequence labelling problems in natural language processing demonstrate the usefulness of the proposed approach.

17:20 - 17:40 **A058: Copula PC Algorithm for Causal Discovery from Mixed Data**
Ruifei Cui, Perry Groot, Tom Heskes

We propose the 'Copula PC' algorithm for causal discovery from a combination of continuous and discrete data, assumed to be drawn from a Gaussian copula model. It is based on a two-step approach. The first step applies Gibbs sampling on rank-based data to obtain samples of correlation matrices. These are then translated into an average correlation matrix and an effective number of data points,

TUESDAY SESSIONS, WITH ABSTRACTS

which in the second step are input to the standard PC algorithm for causal discovery. A stable version naturally arises when rerunning the PC algorithm on different Gibbs samples. Our 'Copula PC' algorithm extends the 'Rank PC' algorithm, which has been designed for Gaussian copula models for purely continuous data. In simulations, 'Copula PC' indeed outperforms 'Rank PC' in cases with mixed variables, in particular for larger numbers of data points, at the expense of a slight increase in computation time.

17:40 - 18:00 A059: Irrevocable-Choice Algorithms for Sampling from a Stream.
Yan Zhu, Eamonn Keogh

The problem of sampling from data streams has attracted significant interest in the last decade. Whichever sampling criteria is considered (uniform sample, maximally diverse sample, etc.), the challenges stem from the relatively small amount of memory available in the face of unbounded streams. In this work we consider an interesting extension of this problem, the framework of which is stimulated by recent improvements in sensing technologies and robotics. In some situations it is not only possible to digitally sense some aspects of the world, but to physically capture a tangible aspect of that world. Currently deployed examples include devices that can capture water/air samples, and devices that capture individual insects or fish. Such devices create an interesting twist on the stream sampling problem, because in most cases, the decision to take a physical sample is irrevocable. In this work we show how to generalize diversification sampling strategies to the irrevocable-choice setting, demonstrating our ideas on several real world domains.

TUE3C: CLUSTERING 2 ROOM 300A

16:40 - 17:00 A060: ClusPath: A Temporal-driven Clustering to Infer Typical Evolution Paths
Marian-Andrei Rizoiu, Julien Velcin, Stéphane Bonnevay, Stéphane Lallich

We propose ClusPath, a novel algorithm for detecting general evolution tendencies in a population of entities. We show how abstract notions, such as the Swedish socio- economical model (in a political dataset) or the companies fiscal optimization (in an economical dataset) can be inferred from low-level descriptive features. Such high- level regularities in the evolution of entities are detected by combining spatial and temporal features into a spatio-temporal dissimilarity measure and using semi- supervised clustering techniques. The relations between the evolution phases are modeled using a graph structure, inferred simultaneously with the partition, by using a"slow changing world" assumption. The idea is to ensure a smooth passage for entities along their evolution paths, which catches the long-term trends in the dataset. Additionally, we also provide a method, based on an evolutionary algorithm, to tune the parameters of ClusPath to new, unseen datasets. This method assesses the fitness of a solution using four opposed quality measures and proposes a balanced compromise.

17:00 - 17:20 A061: Renyi Divergence Minimization based Co-regularized Multiview Clustering
Shalmali Joshi, Joydeep Ghosh, Mark Reid, Oluwasanmi Koyejo

Multiview clustering is a framework for grouping objects given multiple views, e.g. text and image views describing the same set of entities. This paper introduces co- regularization techniques for multiview clustering that explicitly minimize a weighted sum of divergences to impose coherence between per-view learned models. Specifically, we iteratively minimize a weighted sum of divergences between posterior memberships of clusterings, thus learning view-specific parameters that produce similar clusterings across views. We explore a flexible family of divergences, namely Rènyi divergences for co-regularization. An existing method of probabilistic multiview clustering is recovered as a special case of the proposed method. Extensive empirical evaluation suggests improved performance over a variety of existing multiview clustering techniques as well as related methods developed for information fusion with multiview data.

17:20 - 17:40 A062: Adaptive Trajectory Analysis of Replicator Dynamics for Data Clustering
Morteza Haghir Chehreghani

We study the use of replicator dynamics for data clustering and structure identification. We investigate that replicator dynamics, while running, reveals informative transitions that correspond to the significant cuts over data. Occurrence of such transitions is significantly faster than the convergence of replicator dynamics. We exploit this observation to design an efficient clustering algorithm in two steps: i) Cut Identification, and ii) Cluster Pruning. We propose an appropriate regularization to accelerate the appearance of transitions which leads to an adaptive replicator dynamics. A main computational advantage of this regularization is that the optimal solution of the corresponding objective function can be still computed via performing a replicator dynamics. Our experiments on synthetic and real-world datasets show the effectiveness of our algorithm compared to the alternatives.

17:40 - 18:00 A063: Aggregating Crowdsourced Ordinal Labels via Bayesian Clustering
Xiawei Guo, James Kwok

Crowdsourcing allows the collection of labels from a crowd of workers at low cost. In this paper, we focus on ordinal labels, whose underlying order is important. However, the labels can be noisy as there may be amateur workers, spammers and/or even malicious

workers. Moreover, some workers/items may have very few labels, making the estimation of their behavior difficult. To alleviate these problems, we propose a novel Bayesian model that clusters workers and items together using the nonparametric Dirichlet process priors. This allows workers/items in the same cluster to borrow strength from each other. Instead of directly computing the posterior of this complex model, which is infeasible, we propose a new variational inference procedure. Experimental results on a number of real-world data sets show that the proposed algorithm are more accurate than the state-of-the-art.

TUE3D: DATA SCIENCE

ROOM 300B

16:40 - 17:00 **A064: A Hybrid Knowledge Discovery Approach for Mining Predictive Biomarkers in Metabolomic Data**

Dhouha Grissa, Blandine Comte, Estelle Pujos Guillot, Amedeo Napoli

The analysis of complex and massive biological data issued from metabolomic analytical platforms is a challenge of high importance. The analyzed datasets are constituted of a limited set of individuals and a large set of features where predictive biomarkers of clinical outcomes should be mined. Accordingly, in this paper, we propose a new hybrid knowledge discovery approach for discovering meaningful predictive biological patterns. This hybrid approach combines numerical classifiers such as SVM, Random Forests (RF) and ANOVA, with a symbolic method, namely Formal Concept Analysis (FCA). The related experiments show how we can discover among the best potential predictive biomarkers of metabolic diseases thanks to specific combinations of classifiers mainly involving RF and ANOVA. The visualization of predictive biomarkers is based on heatmaps while FCA is mainly used for visualization and interpretation purposes, complementing the computational power of numerical methods.

17:00 - 17:20 **A065: Functional Bid Landscape Forecasting for Display Advertising**

Yuchen Wang, Kan Ren, Weinan Zhang, Jun Wang, Yong Yu

Real-time auction has become an important online advertising trading mechanism. A crucial issue for advertisers is to model the market competition, i.e., bid landscape forecasting. It is formulated as predicting the market price distribution for each ad auction provided by its side information. Existing solutions mainly focus on parameterized heuristic forms of the market price distribution and learn the parameters to fit the data. In this paper, we present a functional bid landscape forecasting method to automatically learning the function mapping from each ad auction features to the market price distribution without any assumption about the functional form. Specifically, to deal with the categorical feature input, we propose a novel decision tree model with a node splitting scheme by attribute value clustering. Furthermore, to deal with the problem of right-censored market price observations, we propose to incorporate a survival model into tree learning and prediction, which largely reduces the model bias. The experiments on real-world data demonstrate that our models achieve substantial performance gains over previous work in various metrics.

17:20 - 17:40 **A066: Enhancing Traffic Congestion Estimation with Social Media by Coupled Hidden Markov Model**

Senzhang Wang, Fengxiang Li, Leon Stenneth, Philip Yu

Estimating traffic conditions in arterial networks with GPS probe data is a practically important while substantially challenging problem. With the increasing availability of GPS equipments installed in various vehicles, GPS probe data is currently becoming a significant data source for traffic monitoring. However, limited by the lack of reliability and low sampling frequency of GPS probes, probe data are usually not sufficient for fully estimating traffic conditions of a large arterial network. For the first time this paper studies how to explore social media as an auxiliary data source and incorporate it with GPS probe data to enhance traffic congestion estimation. Motivated by the increasing amount of traffic information available in Twitter, we first extensively collect tweets that report various traffic events such as congestion, accident, and road construction. Next we propose an extended Coupled Hidden Markov Model which can effectively integrate GPS probe readings and traffic related tweets to more accurately estimate traffic conditions of an arterial network. To address the computational challenge, a sequential importance sampling based EM algorithm is also introduced. We evaluate the proposed model on the arterial network of downtown Chicago. The experimental results demonstrate the superior performance of the model by comparison with previous methods.

17:40 - 18:00 **A067: Joint Learning of Entity Semantics and Relation Pattern for Relation Extraction**

Zheng Suncong, Xu Jiaming, Bao Hongyun, Qi Zhenyu, Zhang Jie, Hao Hongwei, Xu bo

Relation extraction is identifying the relationship of two given entities in the text. It is an important step in the task of knowledge extraction, which plays a vital role in automatic construction of knowledge base. When extracting entities' relations from sentences, some keywords can reflect the relation pattern, besides, the semantic properties of given entities can also help to distinguish some confusing relations. Based on the above observations, we propose a mixture convolutional neural network for the task of relation classification, which can simultaneously learn the semantic properties of entities and the keyword information related to the relation. We conduct experiments on the SemEval-2010 Task8 dataset. The method we propose achieves the state-of-the-art result without using any external information. Additionally, the experimental results also show that our approach can represent the semantic relationship of the given entities effectively.

TUESDAY SESSIONS, WITH ABSTRACTS

TUE3E: NECTAR 3

ROOM BELVEDERE

ML/DM across domains and integrated processes

16:40 - 17:00 A068: From Plagiarism Detection to Bible Analysis: The Potential of Machine Learning for Grammar-Based Text Analysis

Michael Tschuggnall, Günther Specht

The amount of textual data available from digitalized sources such as free online libraries or social media posts has increased drastically in the last decade. In this paper, the main idea to analyze authors by their grammatical writing style is presented. In particular, tasks like authorship attribution, plagiarism detection or author profiling are tackled using the presented algorithm, revealing promising results. Thereby all of the presented approaches are ultimately solved by machine learning algorithms.

17:00 - 17:20 A069: Multi-Target Classification: Methodology and Practical Case Studies

Mark Last

Most classification algorithms are aimed at predicting the value or values of a single target (class) attribute. However, some real-world classification tasks involve several targets that need to be predicted simultaneously. The Multi-Objective Info-Fuzzy Network (M-IFN) algorithm builds an ordered (oblivious) decision-tree model for a multi-target classification task. After summarizing the principles and the properties of the M-IFN algorithm, this paper reviews three case studies of applying M-IFN to practical problems in industry and science.

17:20 - 17:40 A070: A KDD Process for Discrimination Discovery

Salvatore Ruggieri, Franco Turini

The acceptance of analytical methods for discrimination discovery by practitioners and legal scholars can be only achieved if the data mining and machine learning communities will be able to provide case studies, methodological refinements, and the consolidation of a KDD process. We summarize here an approach along these directions.

TUE4A: DEMO SPOTLIGHTS 1

ROOM BELVEDERE

18:00 - 18:02 A071: Bipeline: a Web-based Visualization Tool for Biclustering of Multivariate Time Series

Ricardo Cachucho, Kaihua Liu, Siegfried Nijssen, Arno Knobbe

Large amounts of multivariate time series data are being generated every day. Understanding this data and finding patterns in it is a contemporary task. To find prominent patterns present in multivariate time series, one can use biclustering, that is looking for patterns both in subsets of variables that show coherent behavior and in a number of time periods. For this, an experimental tool is needed. Here, we present Bipeline, a web-based visualization tool that provides both experts and non-experts with a pipeline for experimenting with multivariate time series biclustering. With Bipeline, it is straightforward to save experiments and try different biclustering algorithms, enabling users to intuitively go from pre-processing to visual analysis of biclusters.

18:02 - 18:04 A072: Coordinate transformations for characterization and cluster analysis of spatial configurations in football

Gennady Andrienko, Natalia Andrienko, Guido Budziak, Tatiana von Landesberger, Hendrik Weber

Current technologies allow movements of the players and the ball in football matches to be tracked and recorded with high accuracy and temporal frequency. We demonstrate an approach to analyzing football data with the aim to find typical patterns of spatial arrangement of the field players. It involves transformation of original coordinates to relative positions of the players and the ball with respect to the center and attack vector of each team. From these relative positions, we derive features for characterizing spatial configurations in different time steps during a football game. We apply clustering to these features, which groups the spatial configurations by similarity. By summarizing groups of similar configurations, we obtain representation of spatial arrangement patterns practiced by each team. The patterns are represented visually by density maps built in the teams' relative coordinate systems. Using additional displays, we can investigate under what conditions each pattern was applied.

18:04 - 18:06 A073: Exploratory Analysis of Text Collections Through Visualization and Hybrid Biclustering

Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif

We propose a visual analytics tool to support analytic journalists in the exploration of large text corpora. Our tool combines graph modularity-based diagonal biclustering to extract high-level topics with overlapping bi-clustering to elicit fine-grained topic variants. A hybrid topic treemap visualization gives the analyst an overview of all topics. Coordinated sunburst and heatmap visualizations let the analyst inspect and compare topic variants and access document content on demand.

18:06 - 18:08 A074: GMMbuilder - User-Driven Discovery of Clustering Structure for Bioarchaeology
Markus Mauder, Eirini Ntoutsi, Yulia Bobkova

We present the GMMbuilder tool that allows domain scientists to build Gaussian Mixture Models (GMM) that adhere to domain specific knowledge. Domain experts use the tool to generate different models, extract stable object communities across these models and use these communities to interactively design a final clustering model that explains the data but also considers prior beliefs and expectations of the domain experts.

18:08 - 18:10 A075: INSIGHT: Dynamic Traffic Management Using Heterogeneous Urban Data
Nikolaos Zygouras, Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Nikos Zacheilas, Ioannis Mpoutsis, Vana Kalogeraki, stephen Lynch, Brendan O'Brien, Dermot Kinane, Jakub Marecek, Jia Yuan Yu, Rudi Verargo, Elizabeth Daly, Nico Piatkowksi, Thomas Liebig, Christian Bockermann, Katharina Morik, Matthias Weidlich, Francois Schnitzler, Avigdor Gal, Shie Mannor, Hendrik Stange, Werner Halft, Gennady Andrienko

In this demo we present INSIGHT, a system that provides traffic event detection in Dublin by exploiting Big Data and Crowdsourcing techniques. Our system is able to process and analyze input from multiple heterogeneous urban data sources.

18:10 - 18:12 A076: Learning Language Models from Images with ReGLL
Leonor Becerra-Bonache, Hendrik Blockeel, María Galván, Francois Jacquenet

In this demonstration, we present ReGLL, a system that is able to learn language models taking into account the perceptual context in which the sentences of the model are produced. Thus, ReGLL learns from pairs (Context, Sentence) where: Context is given in the form of an image whose objects have been identified, and Sentence gives a (partial) description of the image. ReGLL uses Inductive Logic Programming Techniques and learns some mappings between n-grams and first order representations of their meanings. The demonstration shows some applications of the language models learned, such as generating relevant sentences describing new images given by the user and translating some sentences from one language to another without the need of any parallel corpus.

18:12 - 18:14 A077: Leveraging Spatial Abstraction in Traffic Analysis and Forecasting with Visual Analytics
Natalia Andrienko, Gennady Andrienko, Salvatore Rinzivillo

By applying spatio-temporal aggregation to traffic data consisting of vehicle trajectories, we generate a spatially abstracted transportation network, which is a directed graph where nodes stand for territory compartments (areas in geographic space) and links (edges) are abstractions of the possible paths between neighboring areas. From time series of traffic characteristics obtained for the links, we reconstruct mathematical models of the interdependencies between the traffic intensity (a.k.a. traffic flow or flux) and mean velocity. Graphical representations of these interdependencies have the same shape as the fundamental diagram of traffic flow through a physical street segment, which is known in transportation science. This key finding substantiates our approach to traffic analysis, forecasting, and simulation leveraging spatial abstraction. We present the process of data-driven generation of traffic forecasting and simulation models, in which each step is supported by visual analytics techniques.

18:14 - 18:16 A078: The SPMF Open-Source Data Mining Library Version 2
Philippe Fournier-Viger, Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, Thanh Lam Hoang

SPMF is an open-source data mining library, specialized in pattern mining, offering implementations of more than 120 data mining algorithms. It has been used in more than 310 research papers to solve applied problems in a wide range of domains from authorship attribution to restaurant recommendation. Its implementations are also commonly used as benchmarks in research papers, and it has also been integrated in several data analysis software programs. After three years of development, this paper introduces the second major revision of the library, named SPMF 2, which provides (1) more than 60 new algorithm implementations (including novel algorithms for sequence prediction), (2) an improved user interface with pattern visualization (3) a novel plug-in system, (4) improved performance, and (5) support for text mining.

WEDNESDAY 21 SEPTEMBER 2016

INVITED TALK

AUTOMATING MACHINE LEARNING

Speaker: Zoubin Ghahramani
Time: 09:00 - 09:50
Room: 1000A
Abstract:

I will describe the "Automatic Statistician" (http://www.automaticstatistician.com/), a project which aims to automate the exploratory analysis and modelling of data. Our approach starts by defining a large space of related probabilistic models via a grammar over models, and then uses Bayesian marginal likelihood computations to search over this space for one or a few good models of the data. The aim is to find models which have both good predictive performance, and are somewhat interpretable. The Automatic Statistician generates a natural language summary of the analysis, producing a 10-15 page report with plots and tables describing the analysis. I will also link this to recent work we have been doing in the area of Probabilistic Programming (including an new system in Julia) to automate inference, and on the rational allocation of computational resources (and our entry in the AutoML conference).

Bio:

Zoubin Ghahramani FRS is Professor of Information Engineering at the University of Cambridge, where he leads the Machine Learning Group, and the Cambridge Liaison Director of the Alan Turing Institute, the UK's national institute for Data Science. He studied computer science and cognitive science at the University of Pennsylvania, obtained his PhD from MIT in 1995, and was a postdoctoral fellow at the University of Toronto. His academic career includes concurrent appointments as one of the founding members of the Gatsby Computational Neuroscience Unit in London, and as a faculty member of CMU's Machine Learning Department for over 10 years. His current research interests include statistical machine learning, Bayesian nonparametrics, scalable inference, probabilistic programming, and building an automatic statistician. He has published over 250 research papers, and has held a number of leadership roles as programme and general chair of the leading international conferences in machine learning including: AISTATS (2005), ICML (2007, 2011), and NIPS (2013, 2014). In 2015 he was elected a Fellow of the Royal Society.

TEST-OF-TIME PRESENTATION

BANDIT BASED MONTE-CARLO PLANNING

Authors: Levente Kocsis, Csaba Szepesvari
Time: 09:50 - 10:20
Room: 1000A
Abstract:

For large state-space Markovian Decision Problems Monte-Carlo planning is one of the few viable approaches to find near-optimal solutions. In this paper we introduce a new algorithm, UCT, that applies bandit ideas to guide Monte-Carlo planning. In finite-horizon or discounted MDPs the algorithm is shown to be consistent and finite sample bounds are derived on the estimation error due to sampling. Experimental results show that in several domains, UCT is significantly more efficient than its alternatives.

INVITED TALK

DIMENSIONALITY REDUCTION WITH CERTAINTY

Speaker: Rasmus Pagh
Time: 14:40 - 15:30
Room: 1000A

Abstract:

Tool such as Johnson-Lindenstrauss dimensionality reduction and 1-bit minwise hashing have been successfully used to transform problems involving very high-dimensional real vectors into lower-dimensional equivalents, at the cost of introducing a random distortion of distances/similarities among vectors. While this can alleviate the computational cost associated with high dimensionality, the effect on the outcome of the computation (compared to working on the original vectors) can be hard to analyze and interpret. For example, the behavior of a basic kNN classifier is easy to describe and interpret, but if the algorithm is run on dimension-reduced vectors with distorted distances it is much less transparent what is happening. The talk starts with an introduction to randomized (data-independent) dimensionality reduction methods and gives some example applications in machine learning. Based on recent work in the theoretical computer science community we describe tools for dimension reduction that give stronger guarantees on approximation, replacing probabilistic bounds on distance/similarity with bounds that hold with certainty. For example, we describe a "distance sensitive Bloom filter": a succinct representation of high-dimensional boolean vectors that can identify vectors within distance r with certainty, while far vectors are only thought to be close with a small "false positive" probability. We also discuss work towards a deterministic alternative to random feature maps (i.e., dimension-reduced vectors from a high-dimensional feature space), and settings in which a pair of dimension-reducing mappings outperform single-mapping methods. While there are limits to what performance can be achieved with certainty, such techniques may be part of the toolbox for designing transparent and scalable machine learning and knowledge discovery methods.

Bio:

Rasmus Pagh graduated from Aarhus University in 2002, and is now a full professor at the IT University of Copenhagen. His work is centered around efficient algorithms for big data, with an emphasis on randomized techniques. His publications span theoretical computer science, databases, information retrieval, knowledge discovery, and parallel computing. His most well-known work is the cuckoo hashing algorithm (2001), which has led to new developments in several fields. In 2014 he received the best paper award at the WWW Conference for a paper with Pham and Mitzenmacher on similarity estimation, and started a 5-year research project funded by the European Research Council on scalable similarity search.

WEDNESDAY 21 SEPTEMBER 2016

WEDNESDAY SESSIONS AT A GLANCE

PLENARY 1 10:50 - 12:50 — Room 1000A

10:50 - 11:10 B001: Interactive Visual Data Exploration with Subjective Feedback
Kai Puolamaki, Bo Kang, Jefrey Lijffijt, Tijl De Bie

11:10 - 11:30 B002: On the Need for Structure Modelling in Sequence Prediction
Niall Twomey, Tom Diethe, Peter Flach

11:30 - 11:50 B003: Multi-Objective Group Discovery on the Social Web
Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Pierre-Francois Dutot, Denis Trystram

11:50 - 12:10 B004: The Matrix Generalized Inverse Gaussian Distribution: Properties and Applications
Farideh Fazayeli, Arindam Banerjee

12:10 - 12:30 B005: Top-k overlapping densest subgraphs.
Esther Galbrun, Aristides Gionis, Nikolaj Tatti.

12:30 - 12:50 B006: Coupled Hierarchical Dirichlet Process Mixtures for Simultaneous Clustering and Topic Modeling
Masamichi Shimosaka, Takeshi Tsukiji, Shoji Tominaga, Kota Tsubouchi

PLENARY 2 15:30 - 16:10 — Room 1000A

15:30 - 15:50 B007: Cost-sensitive boosting algorithms: Do we really need them?
Nikolaos Nikolaou, Narayanan Unny Edakunni, Meelis Kull, Peter Flach, Gavin Brown

15:50 - 16:10 B008: Warped Matrix Factorisation for Multi-View Data Integration
Naruemon Pratanwanich, Pietro Lio', Oliver Stegle

PLENARY 3 16:40 - 18:00 — Room 1000A

16:40 - 17:00 B009: Graphical Model Sketch
Branislav Kveton, Hung Bui, Mohammad Ghavamzadeh, Georgios Theocharous, S Muthukrishnan, Siqi Sun

17:00 - 17:20 B010: Ensembles of label noise filters: a ranking approach.
Luìs P. F. Garcia, Ana C. Lorena, Stan Matwin, André C. P. L. F. de Carvalho.

17:20 - 17:40 B011: Asynchronous Feature Extraction for Large-scale Linear Predictors Shin Matsushima

17:40 - 18:00 B012: Native Advertisement Selection and Allocation in Social Media Post Feeds
Iordanis Koutsopoulos, Panagiotis Spentzouris

WEDNESDAY SESSIONS, WITH ABSTRACTS

PLENARY 1

ROOM 1000A

10:50 - 11:10 **B001: Interactive Visual Data Exploration with Subjective Feedback**

Kai Puolamaki, Bo Kang, Jefrey Lijffijt, Tijl De Bie

Data visualization and iterative/interactive data mining are growing rapidly in attention, both in research as well as in industry. However, integrated methods and tools that combine advanced visualization and data mining techniques are rare, and those that exist are often specialized to a single problem or domain. In this paper, we introduce a novel generic method for interactive visual exploration of high-dimensional data. In contrast to most visualization tools, it is not based on the traditional dogma of manually zooming and rotating data. Instead, the tool initially presents the user with an `interesting' projection of the data and then employs data randomization with constraints to allow users to flexibly and intuitively express their interests or beliefs using visual interactions that correspond to exactly defined constraints. These constraints expressed by the user are then taken into account by a projection-finding algorithm to compute a new `interesting' projection, a process that can be iterated until the user runs out of time or finds that constraints explain everything she needs to find from the data. We present the tool by means of two case studies, one controlled study on synthetic data and another on real census data.

11:10 - 11:30 **B002: On the Need for Structure Modelling in Sequence Prediction**

Niall Twomey, Tom Diethe, Peter Flach

There is no uniform approach in the literature for modelling sequential correlations in sequence classification problems. It is easy to find examples of unstructured models (e.g. logistic regression) where correlations are not taken into account at all, but there are also many examples where the correlations are explicitly incorporated into a -- potentially computationally expensive -- structured classification model (e.g. conditional random fields). In this paper we lay theoretical and empirical foundations for clarifying the types of problem which necessitate direct modelling of correlations in sequences, and the types of problem where unstructured models that capture sequential aspects solely through features are sufficient. The theoretical work in this paper shows that the rate of decay of auto-correlations within a sequence is related to the excess classification risk that is incurred by ignoring the structural aspect of the data. This is an intuitively appealing result, demonstrating the intimate link between the auto- correlations and excess classification risk. Drawing directly on this theory, we develop well-founded visual analytics tools that can be applied a priori on data sequences and we demonstrate how these tools can guide practitioners in specifying feature representations based on auto-correlation profiles. Empirical analysis is performed on three sequential datasets. With baseline feature templates, structured and unstructured models achieve similar performance, indicating no initial preference for either model. We then apply the visual analytics tools to the datasets, and show that classification performance in all cases is improved over baseline results when our tools are involved in defining feature representations.

11:30 - 11:50 **B003: Multi-Objective Group Discovery on the Social Web**

Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Pierre-Francois Dutot, Denis Trystram

We are interested in discovering user groups from collaborative rating datasets. Each user has a set of attributes that help find labeled groups such as young computer scientists in France and American female designers. We formalize the problem of finding user groups whose quality is optimized in multiple dimensions and show that it is NP-Complete. We develop alpha-MOMRI, an alpha-approximation algorithm, and h-MOMRI, a heuristic-based algorithm, for multi-objective optimization to find high quality groups. Our extensive experiments on real datasets from the social Web examine the performance of our algorithms and report cases where alpha-MOMRI and h-MOMRI are useful.

11:50 - 12:10 **B004: The Matrix Generalized Inverse Gaussian Distribution: Properties and Applications**

Farideh Fazayeli, Arindam Banerjee

While the Matrix Generalized Inverse Gaussian (MGIG) distribution arises naturally in some settings as a distribution over symmetric positive semi-definite matrices, certain key properties of the distribution and effective ways of sampling from the distribution have not been carefully studied. In this paper, we show that the MGIG is unimodal, and the mode can be obtained by solving an Algebraic Riccati Equation (ARE) equation [7]. Based on the property, we propose an importance sampling method for the MGIG where the mode of the proposal distribution matches that of the target. The proposed sampling method is more efficient than existing approaches [32,33], which use proposal distributions that may have the mode far from the MGIG's mode. Further, we illustrate that the posterior distribution in latent factor models, such as probabilistic matrix factorization (PMF) [24], when marginalized over one latent factor has the MGIG distribution. The characterization leads to a novel Collapsed Monte Carlo (CMC) inference algorithm for such latent factor models. We illustrate that CMC has a lower log loss or perplexity than MCMC, and needs fewer samples.

WEDNESDAY SESSIONS, WITH ABSTRACTS

12:10 - 12:30 B005: Top-k overlapping densest subgraphs.
Esther Galbrun, Aristides Gionis, Nikolaj Tatti.

Finding dense subgraphs is an important problem in graph mining and has many practical applications. At the same time, while large real-world networks are known to have many communities that are not well-separated, the majority of the existing work focuses on the problem of finding a single densest subgraph. Hence, it is natural to consider the question of finding the top-k densest subgraphs. One major challenge in addressing this question is how to handle overlaps: eliminating overlaps completely is one option, but this may lead to extracting subgraphs not as dense as it would be possible by allowing a limited amount of overlap. Furthermore, overlaps are desirable as in most real-world graphs there are vertices that belong to more than one community, and thus, to more than one densest subgraph. In this paper we study the problem of finding top-k overlapping densest subgraphs, and we present a new approach that improves over the existing techniques, both in theory and practice. First, we reformulate the problem definition in a way that we are able to obtain an algorithm with constant-factor approximation guarantee. Our approach relies on using techniques for solving the max-sum diversification problem, which however, we need to extend in order to make them applicable to our setting. Second, we evaluate our algorithm on a collection of benchmark datasets and show that it convincingly outperforms the previous methods,

12:30 - 12:50 B006: Coupled Hierarchical Dirichlet Process Mixtures for Simultaneous Clustering and Topic Modeling
Masamichi Shimosaka, Takeshi Tsukiji, Shoji Tominaga, Kota Tsubouchi

In this paper, we propose a nonparametric Bayesian mixture model that simultaneously optimizes the topic extraction and group clustering while allowing all topics to be shared by all clusters for grouped data. In addition, in order to enhance the computational efficiency on par with today's large-scale data, we formulate our model so that it can use a closed-form variational Bayesian method to approximately calculate the posterior distribution. Experimental results with corpus data show that our model has a better performance than existing models, achieving 22% improvement against state-of-the-art model. Moreover, an experiment with location data from mobile phones shows that our model performs well in the field of big data analysis.

PLENARY 2 ROOM 1000A

15:30 - 15:50 B007: Cost-sensitive boosting algorithms: Do we really need them?
Nikolaos Nikolaou, Narayanan Unny Edakunni, Meelis Kull, Peter Flach, Gavin Brown

We provide a unifying perspective for two decades of work on cost-sensitive Boosting algorithms. When analyzing the literature 1997-2016, we find 15 distinct cost-sensitive variants of the original algorithm; each of these has its own motivation and claims to superiority -- so who should we believe? In this work we critique the Boosting literature using four theoretical frameworks: Bayesian decision theory, the functional gradient descent view, margin theory, and probabilistic modelling. Our finding is that only three algorithms are fully supported -- and the probabilistic model view suggests that all require their outputs to be calibrated for best performance. Experiments on 18 datasets across 21 degrees of imbalance support the hypothesis -- showing that once calibrated, they perform equivalently, and outperform all others. Our final recommendation -- based on simplicity, flexibility and performance -- is to use the original Adaboost algorithm with a shifted decision threshold and calibrated probability estimates.

15:50 - 16:10 B008: Warped Matrix Factorisation for Multi-View Data Integration
Naruemon Pratanwanich, Pietro Lio', Oliver Stegle

Matrix factorisation is a widely used tool and has found applications in collaborative filtering, image analysis and in genomics. Several extensions of the classical model have been proposed, such as modelling of multiple related ``data views" or accounting for side information on the latent factors. However, as the complexity of these models increases even subtle mismatches of the distributional assumptions on the input data can severely affect model performance. Here, we propose a simple yet effective solution to address this problem by modelling the observed data in a transformed or warped space. We derive a joint model of a multi-view matrix factorisation model that infers view-specific data transformations and provide a computationally efficient variational approach for parameter inference, which exploits low-rank of side information. We validate the model on synthetic data before applying it to a matrix completion problem in genomics, finding that our model improves the imputation of missing values in gene-disease association analysis and allows to discover enhanced consensus structures across multiple data views.

PLENARY 3

ROOM 1000A

16:40 - 17:00 B009: Graphical Model Sketch

Branislav Kveton, Hung Bui, Mohammad Ghavamzadeh, Georgios Theocharous, S Muthukrishnan, Siqi Sun

Structured high-cardinality data arises in many domains, and poses a major challenge for both modeling and inference. Graphical models are a popular approach to modeling structured data but they are unsuitable for high-cardinality variables. The count-min (CM) sketch is a popular approach to estimating probabilities in high-cardinality data but it does not scale well beyond a few variables. In this work, we bring together the ideas of graphical models and count sketches; and propose and analyze several approaches to estimating probabilities in structured high-cardinality streams of data. The key idea of our approximations is to use the structure of a graphical model and approximately estimate its factors by"sketches", which hash high-cardinality variables using random projections. Our approximations are computationally efficient and their space complexity is independent of the cardinality of variables. Our error bounds are multiplicative and significantly improve upon those of the CM sketch, a state-of-the-art approach to estimating probabilities in streams. We evaluate our approximations on synthetic and real-world problems, and report an order of magnitude improvements over the CM sketch.

17:00 - 17:20 B010: Ensembles of label noise filters: a ranking approach.

Luìs P. F. Garcia, Ana C. Lorena, Stan Matwin, André C. P. L. F. de Carvalho.

Label noise can be a major problem in classification tasks, since most machine learning algorithms rely on data labels in their inductive process. Thereupon, various techniques for label noise identification have been investigated in the literature. The bias of each technique defines how suitable it is for each dataset. Besides, while some techniques identify a large number of examples as noisy and have a high false positive rate, others are very restrictive and therefore not able to identify all noisy examples. This paper investigates how label noise detection can be improved by using an ensemble of noise filtering techniques. These filters, individual and ensembles, are experimentally compared. Another concern in this paper is the computational cost of ensembles, once, for a particular dataset, an individual technique can have the same predictive performance as an ensemble. In this case the individual technique should be preferred. To deal with this situation, this study also proposes the use of meta-learning to recommend, for a new dataset, the best filter. An extensive experimental evaluation of the use of individual filters, ensemble filters and meta-learning was performed using public datasets with imputed label noise. The results show that ensembles of noise filters can improve noise filtering performance and that a recommendation system based on meta-learning can successfully recommend the best filtering technique for new datasets. A case study using a real dataset from the ecological niche modeling domain is also presented and evaluated, with the results validated by an expert.

17:20 - 17:40 B011: Asynchronous Feature Extraction for Large-scale Linear Predictors

Shin Matsushima

Learning from datasets with a massive number of possible features to obtain more accurate predictors is being intensively studied. In this paper, we aim to perform effective learning by using the L1 regularized risk minimization problems regarding both time and space computational resources. This is accomplished by concentrating on the effective features from among a large number of unnecessary features. To achieve this, we propose a multithreaded scheme that simultaneously runs processes for developing seemingly important features in the main memory and updating parameters regarding only the important features. We verified our method through computational experiments, showing that our proposed scheme can handle terabyte-scale optimization problems with one machine.

17:40 - 18:00 B012: Native Advertisement Selection and Allocation in Social Media Post Feeds

Iordanis Koutsopoulos, Panagiotis Spentzouris

We study native advertisement selection and placement in social media post feeds. In the prevalent pay-per-click model, each ad click leads to certain amount of revenue for the platform. The probability of click for an ad depends on attributes that are either inherent to the ad (e.g ad quality) or related to user profile and activity or related to the post feed. While the first two types of attributes are also encountered in web-search advertising, the third one fundamentally differentiates native from web-search advertising, and it is the one we model and study in this paper. Evidence from online platforms suggests that the main attributes of the third type that affect ad clicks are the relevance of ads to preceding posts, and the distance between consecutively projected ads; e.g the fewer the intervening posts between ads, the smaller the click probability is, due to user saturation. We model the events of ad clicks as Bernoulli random variables. We seek the ad selection and allocation policy that optimizes a metric which is a combination of (i) the platform expected revenue, and (ii) uncertainty in revenue, captured by the variance of provisionally consumed budget of selected ads. Uncertainty in revenue should be minimum, since this translates into reduced profit or wasted advertising opportunities for the platform. On the other hand, the expected revenue from ad clicking should be maximum. The constraint is that the expected revenue attained for each selected ad should not exceed its apriori set budget. We show that the optimization problem above reduces to an instance of a resource-constrained minimum-cost path problem on a weighted directed acyclic graph. Through numerical evaluation, we assess the impact of various parameters on the objective, and the way they shape the tradeoff between revenue and uncertainty.

THURSDAY 22 SEPTEMBER 2016

INVITED TALK

ALPHAGO – MASTERING THE GAME OF GO WITH DEEP NEURAL NETWORKS AND TREE SEARCH

Speaker: Thore Graepel
Time: 09:10 - 10:00
Room: 1000A
Abstract:

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play. Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs and beat the human European Go champion Fan Hui by 5 games to 0, a feat thought to be at least a decade away by Go and AI experts alike. Finally, in a dramatic and widely publicised match, AlphaGo defeated Lee Sedol, the top player of the past decade, 4 games to 1. In this talk, I will explain how AlphaGo works, describe our process of evaluation and improvement, and discuss what we can learn about computational intuition and creativity from the way AlphaGo plays.

Bio:

Thore Graepel is a research group lead at Google DeepMind and holds a part-time position as Chair of Machine Learning at University College London. He studied physics at the University of Hamburg, Imperial College London, and Technical University of Berlin, where he also obtained his PhD in machine learning in 2001. He spent time as a postdoctoral researcher at ETH Zurich and Royal Holloway College, University of London, before joining Microsoft Research in Cambridge in 2003, where he co-founded the Online Services and Advertising group. Major applications of Thore's work include Xbox Live's TrueSkill system for ranking and matchmaking, the AdPredictor framework for click-through rate prediction in Bing, and the Matchbox recommender system which inspired the recommendation engine of Xbox Live Marketplace. More recently, Thore's work on the predictability of private attributes from digital records of human behaviour has been the subject of intense discussion among privacy experts and the general public. Thore's current research interests include probabilistic graphical models and inference, reinforcement learning, games, and multi-agent systems. He has published over one hundred peer-reviewed papers, is a named co-inventor on dozens of patents, serves on the editorial boards of JMLR and MLJ, and is a founding editor of the book series Machine Learning & Pattern Recognition at Chapman & Hall/CRC. At DeepMind, Thore has returned to his original passion of understanding and creating intelligence, and recently contributed to creating AlphaGo, the first computer program to defeat a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.

BEST ML PAPER

BOO: PROBABILISTIC INFERENCE FOR DETERMINING OPTIONS IN REINFORCEMENT LEARNING

Authors: Christian Daniel, Herke van Hoof, Jan Peters, Gerhard Neumann

Time: 10:00 - 10:30

Room: 1000A

Abstract:

Tasks that require many sequential decisions or complex solutions are hard to solve using conventional reinforcement learning algorithms. Based on the semi Markov decision process setting (SMDP) and the option framework, we propose a model which aims to alleviate these concerns. Instead of learning a single monolithic policy, the agent learns a set of simpler sub-policies as well as the initiation and termination probabilities for each of those sub-policies. While existing option learning algorithms frequently require manual specification of components such as the sub-policies, we present an algorithm which infers all relevant components of the option framework from data. Furthermore, the proposed approach is based on parametric option representations and works well in combination with current policy search methods, which are particularly well suited for continuous real-world tasks. We present results on SMDPs with discrete as well as continuous state-action spaces. The results show that the presented algorithm can combine simple sub-policies to solve complex tasks and can improve learning performance on simpler tasks.

THURSDAY 22 SEPTEMBER 2016

THURSDAY SESSIONS AT A GLANCE

PARALLEL SESSIONS 11:00 - 13:00

Thu1A: Graphs and Social Networks 1 — Room 1000A

11:00 - 11:20 B013: Temporal PageRank
Polina Rozenshtein, Aristides Gionis

11:20 - 11:40 B014: Discovering Topically- and Temporally-Coherent Events in Interaction Networks
Han Xiao, Polina Rozenshtein, Aristides Gionis

11:40 - 12:00 B015: BASS: A Bootstrapping Approach for Aligning Heterogenous Social Networks
Xuezhi Cao, Yong Yu

12:00 - 12:20 B016: Structure Pattern Analysis and Cascade Prediction in Social Networks
Bolei Zhang, Zhuzhong Qian, Sanglu Lu

12:20 - 12:40 B017: Link Prediction in Dynamic Networks Using Graphlet
Mahmudur Rahman, Mohammad Al Hasan

12:40 - 13:00 B018: Persistent Roles in Online Social Networks
Matt Revelle, Carlotta Domeniconi, Aditya Johri

Thu1B: Reinforcement Learning — Room 1000B

11:00 - 11:20 B019: Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes
Jordi Grau-Moya, Felix Leibfried, Tim Genewein, Daniel Braun

11:20 - 11:40 B020: Learning to Control a Structured-Prediction Decoder for Detection of HTTP-Layer DDoS Attackers
Uwe Dick, Tobias Scheffer

11:40 - 12:00 B021: Local Roots: A Tree-based Subgoal Discovery Method to Accelerate Reinforcement Learning
Erkin Cilden, Alper Demir, Faruk Polat

12:00 - 12:20 B022: Anti Imitation-based Policy Learning
Michelle Sebag, Marc Schoenauer, Riad Akrour, Basile Mayeur

12:20 - 12:40 B023: Pure Exploration for Max-Quantile Bandits
Yahel David, Nahum Shimkin

Thu1C: Factorization — Room 300A

11:00 - 11:20 B024: M-Zoom: Fast Dense-Block Detection in Tensors with Quality Guarantees
Kijung Shin, Bryan Hooi, Christos Faloutsos

11:20 - 11:40 B025: Semi-supervised Tensor Factorization for Brain Network Analysis
Bokai Cao, Chun-Ta Lu, Xiaokai Wei, Philip Yu, Alex Leow

11:40 - 12:00 B026: Bayesian Wishart Matrix Factorization.
Cheng Luo, Xiongcai Cai.

12:00 - 12:20 B027: Sparse Topical Analysis of Dyadic Data using Matrix Tri-factorization
Ranganath B. N.

12:20 - 12:40 B028: Factorizing LambdaMART for cold start recommendations
Phong Nguyen, Jun Wang, A. Kalousis

Thu1D: Streams And Time Series — Room 300B

11:00 - 11:20 B029: Online Density Estimation of Heterogeneous Data Streams in Higher Dimensions
Michael Geilke, Andreas Karwath, Stefan Kramer

11:20 - 11:40 B030: Fast Hoeffding Drift Detection Method for Evolving Data Streams
Ali Pesaranghader, Herna Viktor

11:40 - 12:00 B031: On Dynamic Feature Weighting for Feature Drifting Data Streams
Jean Barddal, Heitor Gomes, Fabrício Enembreck, Bernhard Pfahringer, Albert Bifet

12:00 - 12:20 B032: Cost-aware early classification of time series
Romain Tavenard, Simon Malinowski

12:20 - 12:40 B033: Scalable Time Series Classification
Patrick Schäfer

12:40 - 13:00 B034: Generalized Random Shapelet Forests.
Isak Karlsson, Panagiotis Papapetrou, Henrik Bostrom.

PARALLEL SESSIONS 14:50 - 16:10

Thu2A: Classification 2 — Room 1000A

14:50 - 15:10 B035: Learning to Aggregate using Uninorms
Vitalik Melnikov, Eyke Huellermeier

15:10 - 15:30 B036: Consistency of Probabilistic Classifier Trees
Krzysztof Dembczynski, Wojciech Kotlowski, Willem Waegeman, Robert Busa-Fekete, Eyke Huellermeier

15:30 - 15:50 B037: Is Attribute-Based Zero-Shot Learning an Ill-Posed Strategy?
Ibrahim Alabdulmohsin, Moustapha Cisse, Xiangliang Zhang

15:50 - 16:10 B038: Building Ensembles of Adaptive Nested Dichotomies with Random-Pair Selection
Tim Leathart, Bernhard Pfahringer, Eibe Frank

Thu2B: (Semi-)Supervised Learning — Room 1000B

14:50 - 15:10 B039: Learning Efficiently in Semantic Based Regularization
Vincenzo Scoca, Michelangelo Diligenti, Marco Gori

15:10 - 15:30 B040: Uncovering Locally Discriminative Structure for Feature Analysis
Sen Wang, Feiping Nie, Xiaojun Chang, Xue Li, Quan Z. Sheng, Lina Yao

15:30 - 15:50 B041: Ballpark Learning: Estimating Labels from Rough Group Comparisons
Tom Hope, Dafna Shahaf

15:50 - 16:10 B042: Interactive Learning from Multiple Noisy Labels
Shankar Vembu, Sandra Zilles

Thu2C: Dimensionality — Room 300A

14:50 - 15:10 B043: Using regression makes extraction of shared variation in multiple datasets easy.
Jussi Korpela, Andreas Henelius, Lauri Ahonen, Arto Klami, Kai Puolamäki.

15:10 - 15:30 B044: Graph-Margin Based Multi-Label Feature Selection
Peng Yan, Yun Li

15:30 - 15:50 B045: Robust Principal Component Analysis by Reverse Iterative Linear Programming
Andrea Visentin, Steven Prestwich, Armagan Tarim

15:50 - 16:10 B046: Measuring the Stability of Feature Selection
Sarah Nogueira, Gavin Brown

Thu2D: Patterns in Sequences — Room 300B

14:50 - 15:10 B047: An Efficient Algorithm for Mining Frequent Sequence with Constraint Programming
John AOGA, Pierre Schaus, Tias Guns

15:10 - 15:30 B048: SkOPUS: Mining top-k sequential patterns under leverage.
Francois Petitjean, Tao Li, Nikolaj Tatti, Geoffrey I. Webb.

15:30 - 15:50 B049: Efficient Discovery of Sets of Co-occurring Items in Event Sequences
Boris Cule, Len Feremans, Bart Goethals

15:50 - 16:10 B050: Semigeometric Tiling of Event Sequences
Andreas Henelius, Isak Karlsson, Panagiotis Papapetrou, Antti Ukkonen, Kai Puolamaki

THURSDAY 22 SEPTEMBER 2016

THURSDAY SESSIONS AT A GLANCE

PARALLEL SESSIONS 16:40 - 18:00

Thu3A: Graphs and Social Networks 2 — Room 1000A

16:40 - 17:00 B051: Credible Review Detection with Limited Information using Consistency Features
Subhabrata Mukherjee, Sourav Dutta, Gerhard Weikum

17:00 - 17:20 B052: Maximizing Time-decaying Influence in Social Networks
Naoto Ohsaka, Yutaro Yamaguchi, Naonori Kakimura, Ken-ichi Kawarabayashi

17:20 - 17:40 B053: Trust Hole Identification in Signed Networks
Jiawei Zhang, Qianyi Zhan, Lifang He, Charu Aggarwal, Philip Yu

17:40 - 18:00 B054: Learning Distributed Representations of Users for Source Detection in Online Social Networks
Simon Bourigault, Sylvain Lamprier, Patrick Gallinari

Thu3B: Deep Learning and Neural Networks 2 — Room 1000B

16:40 - 17:00 B055: A topological insight into restricted Boltzmann machines
Decebal Constantin Mocanu, Elena Mocanu, Phuong H. Nguyen, Madeleine Gibescu, Antonio Liotta

17:00 - 17:20 B056: Attribute Conjunction Learning with Recurrent Neural Network
Kongming Liang, Hong Chang, Shiguang Shan, Xilin Chen

17:20 - 17:40 B057: Exploring a mixed representation for encoding Temporal Coherence
Jon Parkinson, Ubai Sandouk, Ke Chen

17:40 - 18:00 B058: Sequential Data Classification in the Space of Liquid State Machines
Yang Li, Junyuan Hong, Huanhuan Chen

Thu3C: Bandits & Transfer Learning — Room 300A

16:40 - 17:00 B059: Linear Bandits in Unknown Environments
Thibault Gisselbrecht, Sylvain Lamprier, Patrick Gallinari

17:00 - 17:20 B060: Interpretable Domain Adaptation via Optimization over the Stiefel Manifold
Christian Poelitz, Wouter Duivesteijn, Katharina Morik

17:20 - 17:40 B061: Communication-Efficient Distributed Online Learning with Kernels
Michael Kamp, Sebastian Bothe, Mario Boley, Michael Mock

17:40 - 18:00 B062: Proactive Transfer Learning for Heterogeneous Feature and Label Spaces
Seungwhan Moon, Jaime Carbonell

Thu3D: Recommendation — Room 300B

16:40 - 17:00 B063: Top-N Recommendation via Joint Cross-Domain User Clustering and Similarity Learning
Dimitrios Rafailidis, Fabio Crestani

17:00 - 17:20 B064: Collaborative Expert Recommendation for Community-Based Question Answering
Congfu Xu, Xin Wang, Yunhui Guo

17:20 - 17:40 B065: Selecting Collaborative Filtering algorithms using Metalearning
Tiago Cunha, Carlos Soares, André Carvalho

17:40 - 18:00 B066: Modeling Sequential Preferences with Dynamic User and Context Factors
Duc-Trong Le, Yuan Fang, Hady Lauw

Thu3E: Mixed Grill — Room Belvedere

16:40 - 17:00 B067: Laplacian Hamiltonian Monte Carlo
Yizhe Zhang, Changyou Chen, Ricardo Henao, Lawrence Carin

17:00 - 17:20 B068: Augmented leverage score sampling with bounds
Daniel Perry, Ross Whitaker

17:20 - 17:40 B069: Exact and efficient top-K inference for multi-target prediction by querying separable linear relational models
Michiel Stock, Krzysztof Dembczynski, Bernard De Baets, Willem Waegeman

17:40 - 18:00 B070: Differentially Private User Data Perturbation with Multi-Level Privacy Controls
Yilin Shen, Rui Chen, Hongxia Jin, Ninghui Li

PARALLEL SESSIONS 18:00 - 18:16

Thu4A: Demo Spotlights 2 — Room Belvedere

18:00 - 18:02 B071: A Tool for Subjective and Interactive Visual Data Exploration
Bo Kang, Kai Puolamaki, Jefrey Lijffijt, Tijl De Bie

18:02 - 18:04 B072: DANCer: dynamic attributed network with community structure generator
Christine Largeron, Oualid Benyahia, Baptiste Jeudy, Osmar Zaiane

18:04 - 18:06 B073: Finding incident-related social media messages for emergency awareness
Alexander Nieuwenhuijse, Jorn Bakker, Mykola Pechenizkiy

18:06 - 18:08 B074: h(odor): Interactive Discovery of Hypotheses on the Structure-Odor Relationship in Neuroscience
Guillaume Bosc, Marc Plantevit, Moustafa Bensafi, Jean-Francois Boulicaut, Mehdi Kaytoue

18:08 - 18:10 B075: Ranking Researchers through Collaboration Pattern Analysis
Mario Cataldi, Luigi Di Caro, Claudio Schifanella

18:10 - 18:12 B076: SITS-P2miner: Pattern-Based Mining of Satellite Image Time Series
Tuan Nguyen, Nicolas MEGER, Christophe Rigotti, Catherine Pothier, Remi Andreoli

18:12 - 18:14 B077: Topy: Real-time Story Tracking via Social Tags
Gevorg Poghosyan, M. Atif Qureshi, Georgiana Ifrim

18:14 - 18:16 B078: TwitterCracy: Exploratory Monitoring of Twitter Streams for the 2016 U.S. Presidential Election Cycle
Muhammad Atif Qureshi, Arjumand Younus, Derek Greene

THURSDAY SESSIONS, WITH ABSTRACTS

THU1A: GRAPHS AND SOCIAL NETWORKS 1

ROOM 1000A

11:00 - 11:20 B013: Temporal PageRank

Polina Rozenshtein, Aristides Gionis

PageRank is one of the most popular measures for ranking the nodes of a network according to their importance. However, PageRank is defined as a steady state of a random walk, which implies that the underlying network needs to be fixed and static. Thus, to extend PageRank to networks with a temporal dimension, the available temporal information has to be judiciously incorporated into the model. Although numerous recent works study the problem of computing PageRank on dynamic graphs, most of them consider the case of updating static PageRank under node/edge insertions/deletions. In other words, PageRank is always defined as the static PageRank of the current instance of the graph. In this paper we introduce temporal PageRank, a generalization of PageRank for temporal networks, where activity is represented as a sequence of time-stamped edges. Our model uses the random-walk interpretation of static PageRank, generalized by the concept of temporal random walk. By highlighting the actual information flow in the network, temporal PageRank captures more accurately the network dynamics. A main feature of temporal PageRank is that it adapts to concept drifts: the importance of nodes may change during the lifetime of the network, according to changes in the distribution of edges. On the other hand, if the distribution of edges remains constant, temporal PageRank is equivalent to static PageRank. We present temporal PageRank along with an efficient algorithm, suitable for online streaming scenarios. We conduct experiments on various real and semi-real datasets, and provide empirical evidence that temporal PageRank is a flexible measure that adjusts to changes in the network dynamics.

11:20 - 11:40 B014: Discovering Topically- and Temporally-Coherent Events in Interaction Networks

Han Xiao, Polina Rozenshtein, Aristides Gionis

With the increasing use of online communication platforms, such as email, Twitter, and messaging applications, we are faced with a growing amount of data that combine content (what is said), time (when), and user (by whom) information. Discovering meaningful patterns and understand what is happening in this data is an important challenge. We consider the problem of mining online communication data and finding top-k temporal events. A temporal event is a coherent topic that is dis- cussed frequently in a relatively short time span, while its information flow respects the underlying network. Our method consists of two steps. We first introduce the notion of in- teraction meta-graph, which connects associated interactions. Using this notion, we define a temporal event to be a subset of interactions that (i) are topically and temporally close and (ii) correspond to a tree that captures the information flow. Finding the best temporal event leads to a budget version of the prize-collecting Steiner-tree (PCST) prob- lem, which we solve using three different methods: a greedy approach, a dynamic-programming algorithm, and an adaptation to an existing ap- proximation algorithm. Finding the top-k events maps to a maximum set-cover problem, and thus, solved by greedy algorithm. We compare and analyze our algorithms in both synthetic and real datasets, such as Twitter and email communication. The results show that our methods are able to detect meaningful temporal events.

11:40 - 12:00 B015: BASS: A Bootstrapping Approach for Aligning Heterogenous Social Networks

Xuezhi Cao, Yong Yu

Most people now participate in more than one online social network (OSN). However, the alignment indicating which accounts belong to same natural person is not revealed. Aligning these isolated networks can provide united environment for users and help to improve online personalization services. In this paper, we propose a bootstrapping approach BASS to recover the alignment. It is an unsupervised general-purposed approach with minimum limitation on target networks and users, and is scalable for real OSNs. Specifically, we jointly model user consistencies of usernames, social ties, and user generated contents, and then employ EM algorithm for the parameter learning. For analysis and evaluation, We collect and publish large-scale data sets covering various types of OSNs and multi-lingual scenarios. We conduct extensive experiments to demonstrate the performance of BASS, concluding that our approach significantly outperform state-of-the-art approaches.

12:00 - 12:20 B016: Structure Pattern Analysis and Cascade Prediction in Social Networks

Bolei Zhang, Zhuzhong Qian, Sanglu Lu

As information spreads across social links, it may reach different people and become cascades in social networks. However, the elusive micro-foundations of social behaviors and the complex underlying social networks make it very difficult to model and predict the information diffusion process precisely. From a different perspective, we can often observe the interplay between information diffusion and the cascade structures. On one hand, information driven by different mechanics may evolve into diverse structures; On the other hand, different cascade structures will reach different groups people and thus affect the diffusion process. In this paper, we explore the relationships between information diffusion and the cascade structures in social networks. By embedding the cascades in a lower dimensional space and employing spectral clustering algorithm, we find that the cascades generally evolve into five typical structure

patterns with distinguishable characteristics. In addition, these patterns can be identified by observing the initial footprints of the cascades. Based on this observation, we propose to predict cascade growth with the structure patterns. The experiment results show that the accuracy of predicting both the structure and virality of cascades can be improved significantly.

12:20 - 12:40 B017: Link Prediction in Dynamic Networks Using Graphlet
Mahmudur Rahman, Mohammad Al Hasan

Predicting the link state of a network at a future time given a collection of link states at earlier time is an important task with many real-life applications. In existing literature this task is known as link prediction in dynamic networks. Solving this task is more difficult than its counterpart in static networks because an effective feature representation of node-pair instances for the case of dynamic network is hard to obtain. In this work we solve this problem by designing a novel graphlet transition based feature representation of the node-pair instances of a dynamic network. We propose a method GraTFEL which uses unsupervised feature learning methodologies on graphlet transition based features to give a low-dimensional feature representation of the node-pair instances. GraTFEL models the feature learning task as an optimal coding task where the objective is to minimize the reconstruction error, and it solves this optimization task by using a gradient descent method. We validate the effectiveness of the learned feature representations by utilizing it for link prediction in real-life dynamic networks. Specifically, we show that GraTFEL, which use the extracted feature representation of graphlet transition events, outperforms existing methods that use well-known link prediction features.

12:40 - 13:00 B018: Persistent Roles in Online Social Networks
Matt Revelle, Carlotta Domeniconi, Aditya Johri

Users in online social networks often have very different structural positions which may be attributed to a latent factor: roles. In this paper, we analyze dynamic networks from two datasets (Facebook and Scratch) to find roles which define users' structural positions. Each dynamic network is partitioned into snapshots and we independently find roles for each network snapshot. We present our role discovery methodology and investigate how roles differ between snapshots and datasets. Six persistent roles are found and we investigate user role membership, transitions between roles, and interaction preferences.

THU1B: REINFORCEMENT LEARNING — ROOM 1000B

11:00 - 11:20 B019: Planning with Information-Processing Constraints and Model Uncertainty in Markov Decision Processes
Jordi Grau-Moya, Felix Leibfried, Tim Genewein, Daniel Braun

Recently, information-theoretic principles for learning and acting have been proposed to solve particular classes of Markov Decision Problems. Mathematically, such approaches are governed by a variational free energy principle and allow solving MDP planning problems with information-processing constraints expressed in terms of a Kullback-Leibler divergence with respect to a reference distribution. Here we consider a generalization of such MDP planners by taking model uncertainty into account. As model uncertainty can also be formalized as an information-processing constraint, we can derive a unified solution from a single generalized variational principle. We provide a generalized value iteration scheme together with a convergence proof. As limit cases, this generalized scheme includes standard value iteration with a known model, Bayes Adaptive MDP planning, and robust planning. We demonstrate the benefits of this approach in a grid world simulation.

11:20 - 11:40 B020: Learning to Control a Structured-Prediction Decoder for Detection of HTTP-Layer DDoS Attackers
Uwe Dick, Tobias Scheffer

We focus on the problem of detecting clients that attempt to exhaust server resources by flooding a service with protocol-compliant HTTP requests. Attacks are usually coordinated by an entity that controls many clients. Modeling the application as a structured-prediction problem allows the prediction model to jointly classify a multitude of clients based on their cohesion of otherwise inconspicuous features. Since the resulting output space is too vast to search exhaustively, we employ greedy search and techniques in which a parametric controller guides the search. We apply a known method that sequentially learns the controller and the structured-prediction model. We then derive an online policy-gradient method that finds the parameters of the controller and of the structured-prediction model in a joint optimization problem; we obtain a convergence guarantee for the latter method. We evaluate and compare the various methods based on a large collection of traffic data of a web-hosting service.

11:40 - 12:00 B021: Local Roots: A Tree-based Subgoal Discovery Method to Accelerate Reinforcement Learning
Erkin Cilden, Alper Demir, Faruk Polat

Subgoal discovery in reinforcement learning is an effective way of partitioning a problem domain with large state space. Recent research

THURSDAY SESSIONS, WITH ABSTRACTS

mainly focuses on automatic identification of such subgoals during learning, making use of state transition information gathered during exploration. Mostly based on the options framework, an identified subgoal leads the learning agent to an intermediate region which is known to be useful on the way to goal. In this paper, we propose a novel automatic subgoal discovery method which is based on analysis of predicted shortcut history segments derived from experience, which are then used to generate useful options to speed up learning. Compared to similar existing methods, it performs significantly better in terms of time complexity and usefulness of the subgoals identified, without sacrificing solution quality. The effectiveness of the method is empirically shown via experimentation on various benchmark problems compared to well known subgoal identification methods.

12:00 - 12:20 B022: Anti Imitation-based Policy Learning

Michelle Sebag, Marc Schoenauer, Riad Akrour, Basile Mayeur

The Anti Imitation-based Policy Learning (AIPoL) approach, taking inspiration from the Energy-based learning framework (LeCun et al. 2006), aims at a pseudo-value function such that it induces the same order on the state space as a (nearly optimal) value function. By construction, the greedification of such a pseudo-value induces the same policy as the value function itself. The approach assumes that, thanks to prior knowledge, not-to-be-imitated demonstrations can easily be generated. For instance, applying a random policy on a good initial state (e.g., a bicycle in equilibrium) will on average lead to visit states with decreasing values (the bicycle ultimately falls down). Such a demonstration, that is, a sequence of states with decreasing values, is used along a standard learning-to-rank approach to define a pseudo-value function. If the model of the environment is known, this pseudo-value directly induces a policy by greedification. Otherwise, the bad demonstrations are exploited together with off-policy learning to learn a pseudo-Q-value function and likewise thence derive a policy by greedification. To our best knowledge the use of bad demonstrations to achieve policy learning is original. The theoretical analysis shows that the loss of optimality of the pseudo value-based policy is bounded under mild assumptions, and the empirical validation of AIPoL on the mountain car, the bicycle and the swing-up pendulum problems demonstrates the simplicity and the merits of the approach.

12:20 - 12:40 B023: Pure Exploration for Max-Quantile Bandits

Yahel David, Nahum Shimkin

We consider a variant of the pure exploration problem in Multi-Armed Bandits, where the goal is to find the arm for which the λ-quantile is maximal. Within the PAC framework, we provide a lower bound on the sample complexity of any (ε, δ)-correct algorithm, and propose algorithms with matching upper bounds. Our bounds sharpen existing ones by explicitly incorporating the quantile factor λ.
We further provide experiments that compare the sample complexity of our algorithms with that of previous works.

THU1C: FACTORIZATION — ROOM 300A

11:00 - 11:20 B024: M-Zoom: Fast Dense-Block Detection in Tensors with Quality Guarantees

Kijung Shin, Bryan Hooi, Christos Faloutsos

Given a large-scale and high-order tensor, how can we find dense blocks in it? Can we find them in near-linear time but with a quality guarantee? Extensive previous work has shown that dense blocks in tensors as well as graphs indicate anomalous or fraudulent behavior (e.g., lockstep behavior in social networks). However, available methods for detecting such dense blocks are not satisfactory in terms of speed, accuracy, or flexibility. In this work, we propose M-Zoom, a flexible framework for finding dense blocks in tensors, which works with a broad class of density measures. M-Zoom has the following properties: (1) Scalable: M-Zoom scales linearly with all aspects of tensors and is up to 114X faster than state-of-the-art methods with similar accuracy. (2) Provably accurate: M-Zoom provides a guarantee on the lowest density of the blocks it finds. (3) Flexible: M-Zoom supports multi-block detection and size bounds as well as diverse density measures. (4) Effective: M-Zoom successfully detected edit wars and bot activities in Wikipedia, and spotted network attacks from a TCP dump with near-perfect accuracy (AUC=0.98).

11:20 - 11:40 B025: Semi-supervised Tensor Factorization for Brain Network Analysis

Bokai Cao, Chun-Ta Lu, Xiaokai Wei, Philip Yu, Alex Leow

Brain networks characterize the temporal and/or spectral connections between brain regions and are inherently represented by multi-way arrays (tensors). In order to discover the underlying factors driving such connections, we need to derive compact representations from brain network data. Such representations should be discriminative so as to facilitate the identification of subjects performing different cognitive tasks or with different neurological disorders. In this paper, we propose semiBAT, a novel semi-supervised Brain network Analysis approach based on constrained Tensor factorization. semiBAT 1) leverages unlabeled resting-state brain networks for task recognition, 2) explores the temporal dimension to capture the progress, 3) incorporates classifier learning procedure to introduce supervision from labeled data, and 4) selects discriminative latent factors for different tasks. The Alternating Direction Method of Multipliers (ADMM)

framework is utilized to solve the optimization objective. Experimental results on EEG brain networks illustrate the superior performance of the proposed semiBAT model on graph classification with a significant improvement 31.60% over plain vanilla tensor factorization. Moreover, the data-driven factors can be readily visualized which should be informative for investigating cognitive mechanisms.

11:40 - 12:00 B026: Bayesian Wishart Matrix Factorization.
Cheng Luo, Xiongcai Cai.

User tastes are constantly drifting over time as users are exposed to different types of products. The ability to model the tendency of both user preferences and product attractiveness is vital to the success of recommender systems (RSs). We propose a Bayesian Wishart matrix factorization method to model the temporal dynamics of variations among user preferences and item attractiveness in a novel algorithmic perspective. The proposed method is able to well model and properly control diverse rating behaviors across time frames and related temporal effects within time frames in the tendency of user preferences and item attractiveness. We evaluate the proposed method on two synthetic and three real-world benchmark datasets for RSs. Experimental results demonstrate that our proposed method significantly outperforms a variety of state-of-the-art methods in RSs.

12:00 - 12:20 B027: Sparse Topical Analysis of Dyadic Data using Matrix Tri-factorization
Ranganath B. N.

Many applications involve dyadic data, where associations between one pair of domain entities, such as (documents, words) and associations between another pair, such as (documents, users) are completely observed. We motivate the analysis of such dyadic data introducing an additional discrete dimension, which we call topics, and explore sparse relationships between the domain entities and the topic, such as user-topic and document-topic relationships. For this problem of sparse topical analysis of dyadic data, we propose a formulation using sparse matrix tri-factorization. This formulation requires sparsity constraints, not only on the individual factor matrices, but also on the product of two of the factors. To the best of our knowledge, this problem of sparse matrix tri-factorization has not been studied before. We propose a solution that introduces a surrogate for the product of factors and enforce sparsity on this surrogate as well as on the individual factors through L1- regularization. The resulting optimization problem is efficiently solvable in an alternating minimization framework over sub-problems involving individual factors using the well known FISTA algorithm. For the sub-problems that are constrained, we use a projected variant of the FISTA algorithm. We also show that our formulation leads to independent sub-problems towards solving a factor matrix, thereby supporting parallel implementation leading to scalable solution. We perform experiments over bibliographic and product review data to show that the proposed framework based on sparse tri-factorization formulation results in better generalization ability and factorization accuracy compared to baselines that use sparse bi-factorization.

12:20 - 12:40 B028: Factorizing LambdaMART for cold start recommendations
Phong Nguyen, Jun Wang, A. Kalousis

Recommendation systems often rely on point-wise loss metrics such as the mean squared error. However, in real recommendation settings only few items are presented to a user. This observation has recently encouraged the use of rank-based metrics. LambdaMART is the state-of-the-art algorithm in learning to rank which relies on such a metric. Motivated by the fact that very often the users' and items' descriptions as well as the preference behavior can be well summarized by a small number of hidden factors, we propose a novel algorithm, LambdaMART Matrix Factorization (LambdaMART-MF), that learns latent representations of users and items using gradient boosted trees. The algorithm factorizes lambdaMART by defining relevance scores as the inner product of the learned representations of the users and items. We regularise the learned latent representations so that they reflect the user and item manifolds as these are defined by their original feature based descriptors and the preference behavior. Finally we also propose to use a weighted variant of NDCG to reduce the penalty for similar items with large rating discrepancy. We experiment on two very different recommendation datasets, meta-mining and movies-users, and evaluate the performance of LambdaMART-MF, with and without regularization, in the cold start setting as well as in the simpler matrix completion setting. The experiments show that the factorization of LambdaMart brings significant performance improvements both in the cold start and the matrix completion settings. The incorporation of regularisation seems to have a smaller performance impact.

THU1D: STREAMS AND TIME SERIES ROOM 300B

11:00 - 11:20 B029: Online Density Estimation of Heterogeneous Data Streams in Higher Dimensions
Michael Geilke, Andreas Karwath, Stefan Kramer

The joint density of a data stream is suitable for performing data mining tasks without having access to the original data. However, the methods proposed so far only target a small to medium number of variables, since their estimates rely on representing all the interdependencies between the variables of the data. High-dimensional data streams, which are becoming more and more frequent due to increasing numbers of interconnected devices, are, therefore, pushing these methods to their limits. To mitigate these limitations, we

THURSDAY SESSIONS, WITH ABSTRACTS

present an approach that projects the original data stream into a vector space and uses a set of representatives to provide an estimate. Due to the structure of the estimates, it enables the density estimation of higher-dimensional data and approaches the true density with increasing dimensionality of the vector space. Moreover, it is not only designed to estimate homogeneous data, i.e., where all variables are nominal or all variables are numeric, but it can also estimate heterogeneous data. The evaluation is conducted on synthetic and real-world data.

11:20 - 11:40 **B030: Fast Hoeffding Drift Detection Method for Evolving Data Streams**
Ali Pesaranghader, Herna Viktor

Decision makers increasingly require near-instant models to make sense of fast evolving data streams. Learning from such evolving environments is, however, a challenging task. This challenge is partially due to the fact that the distribution of data often changes over time, thus potentially leading to degradation in the overall performance. In particular, classification algorithms need to adapt their models after facing such distributional changes (also referred to as concept drifts). Usually, drift detection methods are utilized in order to accomplish this task. It follows that detecting concept drifts as soon as possible, while resulting in fewer false positives and false negatives, is a major objective of drift detectors. To this end, we introduce the Fast Hoeffding Drift Detection Method (FHDDM) which detects the drift points using a sliding window and Hoeffding's inequality. FHDDM detects a drift when a significant difference between the maximum probability of correct predictions and the most recent probability of correct predictions is observed. Experimental results confirm that FHDDM detects drifts with less detection delay, less false positive and less false negative, when compared to the state-of-the-art.

11:40 - 12:00 **B031: On Dynamic Feature Weighting for Feature Drifting Data Streams**
Jean Barddal, Heitor Gomes, Fabrício Enembreck, Bernhard Pfahringer, Albert Bifet

The ubiquity of data streams has been encouraging the development of new incremental and adaptive learning algorithms. Data stream learners must be fast, memory-bounded, but mainly, tailored to adapt to possible changes in the data distribution, a phenomenon named concept drift. Recently, several works have shown the impact of a so far nearly neglected type of drift: feature drifts. Feature drifts occur whenever a subset of features becomes, or ceases to be, relevant to the learning task. In this paper we (i) provide insights into how the relevance of features can be tracked as a stream progresses according to information theoretical Symmetrical Uncertainty; and (ii) how it can be used to boost two learning schemes: Naive Bayesian and k-Nearest Neighbor. Furthermore, we investigate the usage of these two new dynamically weighted learners as prediction models in the leaves of the Hoeffding Adaptive Tree classifier. Results show improvements in accuracy (an average of 10.69% for k-Nearest Neighbor, 6.23% for Naive Bayes and 4.42% for Hoeffding Adaptive Trees) in both synthetic and real-world datasets at the expense of a bounded increase in both memory consumption and processing time.

12:00 - 12:20 **B032: Cost-aware early classification of time series**
Romain Tavenard, Simon Malinowski

In time series classification, two antagonist notions are at stake. On the one hand, in most cases, the sooner the time series is classified, the more rewarding. On the other hand, an early classification is more likely to be erroneous. Most of the early classification methods have been designed to take a decision as soon as sufficient level of reliability is reached. However, in many applications, delaying the decision with no guarantee that the reliability threshold will be met in the future can be costly. Recently, a framework dedicated to optimizing a trade-off between classification accuracy and the cost of delaying the decision was proposed, together with an algorithm that decides online the optimal time instant to classify an incoming time series. On top of this framework, we build in this paper two different early classification algorithms that optimize a trade-off between decision accuracy and the cost of delaying the decision. These algorithms are non-myopic in the sense that, even when classification is delayed, they can provide an estimate of when the optimal classification time is likely to occur. Our experiments on real datasets demonstrate that the proposed approaches are more robust than existing methods.

12:20 - 12:40 **B033: Scalable Time Series Classification**
Patrick Schäfer

Time series classification tries to mimic the human understanding of similarity. When it comes to long or larger time series datasets, state-of-the-art classifiers reach their limits because of unreasonably high training or testing times. One representative example is the 1-nearest-neighbor DTW classifier (1-NN DTW) that is commonly used as the benchmark to compare to. It has several shortcomings: it has a quadratic time complexity in the time series length and its accuracy degenerates in the presence of noise. To reduce the computational complexity, early abandoning techniques, cascading lower bounds, or recently, a nearest centroid classifier have been introduced. Still, classification times on datasets of a few thousand time series are in the order of hours. We present our Bag-Of-SFA-Symbols in Vector Space (BOSS VS) classifier that is accurate, fast and robust to noise. We show that it is significantly more accurate than 1-NN DTW while being multiple orders of magnitude faster. Its low computational complexity combined with its good classification accuracy makes it relevant for use cases like long or large amounts of time series or real-time analytics.

12:40 - 13:00 B034: Generalized Random Shapelet Forests.
Isak Karlsson, Panagiotis Papapetrou, Henrik Bostrom.

Shapelets are discriminative subsequences of time series, usually embedded in shapelet-based decision trees. The enumeration of time series shapelets is, however, computationally costly, which in addition to the inherent difficulty of the decision tree learning algorithm to effectively handle high-dimensional data, severely limits the applicability of shapelet-based decision tree learning from large (multivariate) time series databases. This paper introduces a novel tree-based ensemble method for univariate and multivariate time series classification using shapelets, called the generalized random shapelet forest (gRSF) algorithm. The algorithm generates a set of shapelet-based decision trees, where both the choice of instances used for building a tree and the choice of shapelets are randomized. For univariate time series, it is demonstrated through an extensive empirical investigation that the proposed algorithm yields predictive performance comparable to the current state-of-the-art and significantly outperforms several alternative algorithms, while being at least an order of magnitude faster. Similarly for multivariate time series, it is shown that the algorithm is significantly less computationally costly and more accurate than the current state-of- the-art.

THU2A: CLASSIFICATION 2

ROOM 1000A

14:50 - 15:10 B035: Learning to Aggregate using Uninorms
Vitalik Melnikov, Eyke Huellermeier

In this paper, we propose a framework for a class of learning problems that we refer to as"learning to aggregate". Roughly, learning-to-aggregate problems are supervised machine learning problems, in which instances are represented in the form of a composition of a (variable) number on constituents; such compositions are associated with an evaluation, score, or label, which is the target of the prediction task, and which can presumably be modeled in the form of a suitable aggregation of the properties of its constituents. Our learning-to-aggregate framework establishes a close connection between machine learning and a branch of mathematics devoted to the systematic study of aggregation functions. We specifically focus on a class of functions called uninorms, which combine conjunctive and disjunctive modes of aggregation. Experimental results for a corresponding model are presented for a review data set, for which the aggregation problem consists of combining different reviewer opinions about a paper into an overall decision of acceptance or rejection.

15:10 - 15:30 B036: Consistency of Probabilistic Classifier Trees
Krzysztof Dembczynski, Wojciech Kotlowski, Willem Waegeman,
Robert Busa-Fekete, Eyke Huellermeier

Label tree classifiers are commonly used for efficient multi-class and multi-label classification. They represent a predictive model in the form of a tree-like hierarchy of (internal) classifiers, each of which is trained on a simpler (often binary) subproblem, and predictions are made by (greedily) following these classifiers' decisions from the root to a leaf of the tree. Unfortunately, this approach does normally not assure consistency for different losses on the original prediction task, even if the internal classifiers are consistent for their subtask. In this paper, we thoroughly analyze a class of methods referred to as probabilistic classifier trees (PCTs). Thanks to training probabilistic classifiers at internal nodes of the hierarchy, these methods allow for searching the tree-structure in a more sophisticated manner, thereby producing predictions of a less greedy nature. Our main result is a regret bound for 0/1 loss, which can easily be extended to ranking-based losses. In this regard, PCTs nicely complement a related approach called filter trees (FTs), and can indeed be seen as a natural alternative thereof. We compare the two approaches both theoretically and empirically.

15:30 - 15:50 B037: Is Attribute-Based Zero-Shot Learning an Ill-Posed Strategy?
Ibrahim Alabdulmohsin, Moustapha Cisse, Xiangliang Zhang

One transfer learning approach that has gained a wide popularity lately is attribute-based zero-shot learning. Its goal is to learn novel classes that were never seen during the training stage. The classical route towards realizing this goal is to incorporate a prior knowledge, in the form of a semantic embedding of classes, and to learn to predict classes indirectly via their semantic attributes. Despite the amount of research devoted to this subject lately, no known algorithm has yet reported a predictive accuracy that could exceed the accuracy of supervised learning with very few training examples. For instance, the direct attribute prediction (DAP) algorithm, which forms a standard baseline for the task, is known to be as accurate as supervised learning when as few as two examples from each hidden class are used for training on some popular benchmark datasets! In this paper, we argue that this lack of significant results in the literature is not a coincidence; attribute-based zero-shot learning is fundamentally an ill-posed strategy. The key insight is the observation that the mechanical task of predicting an attribute is, in fact, quite different from the epistemological task of learning the "correct meaning" of the attribute itself. This renders attribute-based zero-shot learning fundamentally ill-posed. In more precise mathematical terms, attribute-based zero-shot learning is equivalent to the mirage goal of learning with respect to one distribution of instances, with the hope of being able to predict with respect to any arbitrary distribution. We demonstrate this overlooked fact on some synthetic and real datasets.

THURSDAY SESSIONS, WITH ABSTRACTS

15:50 - 16:10 B038: Building Ensembles of Adaptive Nested Dichotomies with Random-Pair Selection
Tim Leathart, Bernhard Pfahringer, Eibe Frank

A system of nested dichotomies is a method of decomposing a multi-class problem into a collection of binary problems. Such a system recursively splits the set of classes into two subsets, and trains a binary classifier to distinguish between each subset. Even though ensembles of nested dichotomies with random structure have been shown to perform well in practice, using a more sophisticated class subset selection method can be used to improve classification accuracy. We investigate an approach to this problem called random-pair selection, and evaluate its effectiveness compared to other published methods of subset selection. We show that our method outperforms other methods in many cases when forming ensembles of nested dichotomies, and is at least on par in all other cases.

THU2B: (SEMI-)SUPERVISED LEARNING — ROOM 1000B

14:50 - 15:10 B039: Learning Efficiently in Semantic Based Regularization
Vincenzo Scoca, Michelangelo Diligenti, Marco Gori

Semantic Based Regularization (SBR) is a general framework to integrate semi-supervised learning with the application specific background knowledge, which is assumed to be expressed as a collection of first-order logic (FOL) clauses. While SBR has been proved to be a useful tool in many applications, the underlying learning task often requires to solve an optimization problem that has been empirically observed to be challenging. Heuristics and experience to achieve good results are therefore the key to success in the application of SBR. The main contribution of this paper is to study why and when training in SBR is easy. In particular, this paper shows that exists a large class of prior knowledge that can be expressed as convex constraints, which can be exploited during training in a very efficient and effective way. This class of constraints provides a natural way to break the complexity of learning by building a training plan that uses the convex constraints as an effective initialization step for the final full optimization problem. Whereas previous published results on SBR have employed Kernel Machines to approximate the underlying unknown predicates, this paper employs Neural Networks for the first time, showing the flexibility of the framework. The experimental results show the effectiveness of the training plan on categorization of real world images.

15:10 - 15:30 B040: Uncovering Locally Discriminative Structure for Feature Analysis
Sen Wang, Feiping Nie, Xiaojun Chang, Xue Li, Quan Z. Sheng, Lina Yao

Manifold structure learning is often used to exploit geometric information among data in semi-supervised feature learning algorithms. In this paper, we find that local discriminative information is also of importance for semi-supervised feature learning. We propose a method that utilizes both the manifold structure of data and local discriminant information. Specifically, we define a local clique for each data point. The k-Nearest Neighbors (kNN) is used to determine the structural information within each clique. We then employ a variant of Fisher criterion model to each clique for local discriminant evaluation and sum all cliques as global integration into the framework. In this way, local discriminant information is embedded. Labels are also utilized to minimize distances between data from the same class. In addition, we use the kernel method to extend our proposed model and facilitate feature learning in a high-dimensional space after feature mapping. Experimental results show that our method is superior to all other compared methods over a number of datasets.

15:30 - 15:50 B041: Ballpark Learning: Estimating Labels from Rough Group Comparisons
Tom Hope, Dafna Shahaf

We are interested in estimating individual labels given only coarse, aggregated signal over the data points. In our setting, we receive sets ("bags") of unlabeled instances with constraints on label proportions. We relax the unrealistic assumption of known label proportions, made in previous work; instead, we assume only to have upper and lower bounds, and constraints on bag differences. We motivate the problem, propose an intuitive formulation and algorithm, and apply our methods to real-world scenarios. Across several domains, we show how using only proportion constraints and no labeled examples, we can achieve surprisingly high accuracy. In particular, we demonstrate how to predict income level using rough stereotypes and how to perform sentiment analysis using very little information. We also apply our method to guide exploratory analysis, recovering geographical differences in twitter dialect.

15:50 - 16:10 B042: Interactive Learning from Multiple Noisy Labels
Shankar Vembu, Sandra Zilles

Interactive learning is a process in which a machine learning algorithm is provided with meaningful, well-chosen examples as opposed to randomly chosen examples typical in standard supervised learning. In this paper, we propose a new method for interactive learning from multiple noisy labels where we exploit the disagreement among annotators to quantify the easiness (or meaningfulness) of an example. We demonstrate the usefulness of this method in estimating the parameters of a latent variable classification model, and conduct experimental analyses on a range of synthetic and benchmark data sets. Furthermore, we theoretically analyze the performance of perceptron in this interactive learning framework.

THU2C: DIMENSIONALITY — ROOM 300A

14:50 - 15:10 **B043: Using regression makes extraction of shared variation in multiple datasets easy.**
Jussi Korpela, Andreas Henelius, Lauri Ahonen, Arto Klami, Kai Puolamäki.

In many data analysis tasks it is important to understand the relationships between different datasets. Several methods exist for this task but many of them are limited to two datasets and linear relationships. In this paper, we propose a new efficient algorithm, termed COCOREG, for the extraction of variation common to all datasets in a given collection of arbitrary size. COCOREG extends redundancy analysis to more than two datasets, utilizing chains of regression functions to extract the shared variation in the original data space. The algorithm can be used with any linear or non- linear regression function, which makes it robust, straightforward, fast, and easy to implement and use. We empirically demonstrate the efficacy of shared variation extraction using the COCOREG algorithm on five artificial and three real datasets.

15:10 - 15:30 **B044: Graph-Margin Based Multi-Label Feature Selection**
Peng Yan, Yun Li

Since instances in multi-label problems are associated with several labels simultaneously, most traditional feature selection algorithms for single label problems are inapplicable. Therefore, new criteria to evaluate features and new methods to model label correlations are needed. In this paper, we adopt the graph model to capture the label correlation, and propose a feature selection algorithm for multi-label problems according to the graph combining with the large margin theory. The proposed multi-label feature selection algorithm GMBA can efficiently utilize the high order label correlation. Experiments on real world data sets demonstrate the effectiveness of the proposed method.

15:30 - 15:50 **B045: Robust Principal Component Analysis by Reverse Iterative Linear Programming**
Andrea Visentin, Steven Prestwich, Armagan Tarim

Principal Components Analysis (PCA) is a data analysis technique widely used in dimensionality reduction. It extracts a small number of orthonormal vectors that explain most of the variation in a dataset, which are called the Principal Components. Conventional PCA is sensitive to outliers because it is based on the L2-norm, so to improve robustness several algorithms based on the L1-norm have been introduced in the literature. We present a new algorithm for robust L1-norm PCA that computes components iteratively in reverse, using a new heuristic based on Linear Programming. This solution is focused on finding the projection that minimizes the variance of the projected points. It has only one parameter to tune, making it simple to use. On common benchmarks it performs competitively compared to other methods.

15:50 - 16:10 **B046: Measuring the Stability of Feature Selection**
Sarah Nogueira, Gavin Brown

In feature selection algorithms, ``stability" is the sensitivity of the chosen feature set to variations in the supplied training data. As such it can be seen as an analogous concept to the statistical variance of a predictor. However unlike variance, there is no unique definition of stability, with numerous proposed measures over 15 years of literature. In this paper, instead of defining a new measure, we start from an axiomatic point of view and identify what properties would be desirable. Somewhat surprisingly, we find that the simple Pearson's correlation coefficient has all necessary properties, yet has somehow been overlooked in favour of more complex alternatives. Finally, we illustrate how the use of this measure in practice can provide better interpretability and more confidence in the model selection process.

THU2D: PATTERNS IN SEQUENCES — ROOM 300B

14:50 - 15:10 **B047: An Efficient Algorithm for Mining Frequent Sequence with Constraint Programming**
John AOGA, Pierre Schaus, Tias Guns

The main advantage of Constraint Programming (CP) approaches for sequential pattern mining (SPM) is their modularity, which includes the ability to add new constraints (regular expressions, length restrictions, etc). The current best CP approach for SPM uses a global constraint (module) that computes the projected database and enforces the minimum frequency; it does this with a filtering algorithm similar to the PrefixSpan method. However, the resulting system is not as scalable as some of the most advanced mining systems like Zaki's cSPADE. We show how, using techniques from both data mining and CP, one can use a generic constraint solver and yet outperform existing specialized systems. This is mainly due to two improvements in the module that computes the projected frequencies: first, computing the projected database can be sped up by pre-computing the positions at which a symbol can become unsupported by a sequence, thereby avoiding to scan the full sequence each time; and second by taking inspiration from the trailing used in CP solvers to devise a backtracking-aware data structure that allows fast incremental storing and restoring of the projected database. Detailed experiments show how this approach outperforms existing CP as well as specialized systems for SPM, and that the gain in efficiency translates directly into increased efficiency for other settings such as mining with regular expressions.

THURSDAY SESSIONS, WITH ABSTRACTS

15:10 - 15:30 B048: SkOPUS: Mining top-k sequential patterns under leverage.
Francois Petitjean, Tao Li, Nikolaj Tatti, Geoffrey I. Webb.

This paper presents a framework for exact discovery of the top-k sequential patterns under Leverage. It combines (1) a novel definition of the expected support for a sequential pattern - a concept on which most interestingness measures directly rely - with (2) SkOPUS: a new branch-and-bound algorithm for the exact discovery of top-k sequential patterns under a given measure of interest. Our interestingness measure employs the partition approach. A pattern is interesting to the extent that it is more frequent than can be explained by assuming independence between any of the pairs of patterns from which it can be composed. The larger the support compared to the expectation under independence, the more interesting is the pattern. We build on these two elements to exactly extract the k sequential patterns with highest leverage, consistent with our definition of expected support. We conduct experiments on both synthetic data with known patterns and real-world datasets; both experiments confirm the consistency and relevance of our approach with regard to the state of the art.

15:30 - 15:50 B049: Efficient Discovery of Sets of Co-occurring Items in Event Sequences
Boris Cule, Len Feremans, Bart Goethals

Discovering patterns in long event sequences is an important data mining task. Most existing work focuses on frequency-based quality measures that allow algorithms to use the anti-monotonicity property to prune the search space and efficiently discover the most frequent patterns. In this work, we step away from such measures, and evaluate patterns using cohesion --- a measure of how close to each other the items making up the pattern appear in the sequence on average. We tackle the fact that cohesion is not an anti-monotonic measure by developing a novel pruning technique in order to reduce the search space. By doing so, we are able to efficiently unearth rare, but strongly cohesive, patterns that existing methods often fail to discover.

15:50 - 16:10 B050: Semigeometric Tiling of Event Sequences
Andreas Henelius, Isak Karlsson, Panagiotis Papapetrou, Antti Ukkonen, Kai Puolamaki

Event sequences are ubiquitous, e.g., in finance, medicine, and social media. Often the same underlying phenomenon, such as television advertisements during Superbowl, is reflected in independent event sequences, like different Twitter users. It is hence of interest to find combinations of temporal segments and subsets of sequences where an event of interest, like a particular hashtag, has an increased occurrence probability. Such patterns allow exploration of the event sequences in terms of their evolving temporal dynamics, and provide more fine-grained insights to the data than what for example straightforward clustering can reveal. We formulate the task of finding such patterns as a novel matrix tiling problem, and propose two algorithms for solving it. Our first algorithm is a greedy set cover heuristic, while in the second approach we view the problem as time-series segmentation. We apply the algorithms on real and artificial datasets and obtain promising results.

THU3A: GRAPHS AND SOCIAL NETWORKS 2 ROOM 1000A

16:40 - 17:00 B051: Credible Review Detection with Limited Information using Consistency Features
Subhabrata Mukherjee, Sourav Dutta, Gerhard Weikum

Online reviews provide viewpoints on the strengths and shortcomings of products/services, influencing potential customers' purchasing decisions. However, the proliferation of non-credible reviews — either fake (promoting/demoting an item), incompetent (involving irrelevant aspects), or biased — entails the problem of identifying credible reviews. Prior works involve classifiers harnessing rich information about items/users — which might not be readily available in several domains — that provide only limited interpretability as to why a review is deemed non-credible. This paper presents a novel approach to address the above issues. We utilize latent topic models leveraging review texts, item ratings, and timestamps to derive consistency features without relying on item/user histories, unavailable for"long-tail" items/users. We develop models, for computing review credibility scores to provide interpretable evidence for non-credible reviews, that are also transferable to other domains — addressing the scarcity of labeled data. Experiments on real-world datasets demonstrate improvements over state-of-the-art baselines.

17:00 - 17:20 B052: Maximizing Time-decaying Influence in Social Networks
Naoto Ohsaka, Yutaro Yamaguchi, Naonori Kakimura, Ken-ichi Kawarabayashi

Influence maximization is a well-studied problem of finding a small set of highly influential individuals in a social network such that the spread of influence under a certain diffusion model is maximized. We propose new diffusion models that incorporate the time-decaying phenomenon by which the power of influence decreases with elapsed time. In standard diffusion models such as the independent cascade and linear threshold models, each edge in a network has a fixed power of influence over time. However, in practical settings, such as rumor spreading, it is natural for the power of influence to depend on the time influenced. We generalize the independent cascade and linear threshold models with time-decaying effects. Moreover, we show that by using an analysis framework based on submodular functions, a natural greedy strategy obtains a solution that is provably within (1-1/e) of optimal. In addition, we propose theoretically and practically

fast algorithms for the proposed models. Experimental results show that the proposed algorithms are scalable to graphs with millions of edges and outperform baseline algorithms based on a state-of-the-art algorithm.

17:20 - 17:40 B053: Trust Hole Identification in Signed Networks

Jiawei Zhang, Qianyi Zhan, Lifang He, Charu Aggarwal, Philip Yu

In the trust-centric context of signed networks, the social links among users are associated with specific polarities to denote the attitudes (trust vs distrust) among the users. Different from traditional unsigned social networks, the diffusion of information in signed networks can be affected by the link polarities and users' positions significantly. In this paper, a new concept called "trust hole" is introduced to characterize the advantages of specific users positions in signed networks. To uncover the trust holes, a novel trust hole detection framework named "Social Community based tRust hOLe expLoration" (SCROLL) is proposed in this paper. Framework SCROLL is based on the signed community detection technique. By removing the potential trust hole candidates, SCROLL aims at maximizing the community detection cost drop to identify the optimal set of trust holes. Extensive experiments have been done on real-world signed network datasets to show the effectiveness of SCROLL.

17:40 - 18:00 B054: Learning Distributed Representations of Users for Source Detection in Online Social Networks

Simon Bourigault, Sylvain Lamprier, Patrick Gallinari

In this paper, we study the problem of source detection in the context of information diffusion through online social networks. We propose a representation learning approach that leads to a robust model able to deal with the sparsity of the data. From learned continuous projections of the users, our approach is able to efficiently predict the source of any newly observed diffusion episode. Our model does not rely neither on a known diffusion graph nor on a hypothetical probabilistic diffusion law, but directly infers the source from diffusion episodes. It is also less complex than alternative state of the art models. It showed good performances on artificial and real-world datasets, compared with various state of the art baselines.

THU3B: DEEP LEARNING AND NEURAL NETWORKS 2 ROOM 1000B

16:40 - 17:00 B055: A topological insight into restricted Boltzmann machines

Decebal Constantin Mocanu, Elena Mocanu, Phuong H. Nguyen, Madeleine Gibescu, Antonio Liotta

Restricted Boltzmann Machines (RBMs) and models derived from them have been successfully used as basic building blocks in deep artificial neural networks for automatic features extraction, unsupervised weights initialization, but also as density estimators. Thus, their generative and discriminative capabilities, but also their computational time are instrumental to a wide range of applications. Our main contribution is to look at RBMs from a topological perspective, bringing insights from network science. Firstly, here we show that RBMs and Gaussian RBMs (GRBMs) are bipartite graphs which naturally have a small-world topology. Secondly, we demonstrate both on synthetic and real-world datasets that by constraining RBMs and GRBMs to a scale-free topology (while still considering local neighborhoods and data distribution), we reduce the number of weights that need to be computed by a few orders of magnitude, at virtually no loss in generative performance. Thirdly, we show that, for a fixed number of weights, our proposed sparse models (which by design have a higher number of hidden neurons) achieve better generative capabilities than standard fully connected RBMs and GRBMs (which by design have a smaller number of hidden neurons), at no additional computational costs.

17:00 - 17:20 B056: Attribute Conjunction Learning with Recurrent Neural Network

Kongming Liang, Hong Chang, Shiguang Shan, Xilin Chen

Searching images with multi-attribute queries shows practical significance in various real world applications. The key problem in this task is how to effectively and efficiently learn from the conjunction of query attributes. In this paper, we propose Attribute Conjunction Recurrent Neural Network (AC-RNN) to tackle this problem. Attributes involved in a query are mapped into the hidden units and combined in a recurrent way to generate the representation of the attribute conjunction, which is then used to compute a multi-attribute classifier as the output. To mitigate the data imbalance problem of multi-attribute queries, we propose a data weighting procedure in attribute conjunction learning with small positive samples. We also discuss on the influence of attribute order in a query and present two methods based on attention mechanism and ensemble learning respectively to further boost the performance. Experimental results on aPASCAL, ImageNet Attributes and LFWA datasets show that our method consistently and significantly outperforms the other comparison methods on all types of queries.

THURSDAY SESSIONS, WITH ABSTRACTS

17:20 - 17:40 B057: Exploring a mixed representation for encoding Temporal Coherence
Jon Parkinson, Ubai Sandouk, Ke Chen

Guiding representation learning towards temporally stable features improves object identity encoding from video. Existing models have applied temporal coherence uniformly over all features based on the assumption that optimal object identity encoding only requires temporally stable components. We explore the effects of mixing temporally coherent 'invariant' features alongside 'variable' features in a single representation. Applying temporal coherence to different proportions of available features, we introduce a mixed representation autoencoder. Trained on several datasets, model outputs were passed to an object classification task to compare performance. Whilst the inclusion of temporal coherence improved object identity recognition in all cases, the majority of tests favoured a mixed representation.

17:40 - 18:00 B058: Sequential Data Classification in the Space of Liquid State Machines
Yang Li, Junyuan Hong, Huanhuan Chen

This paper proposes a novel classification approach to carrying out sequential data classification. In this approach, each sequence in a data stream is approximated and represented by one state space model -- liquid state machine. Each sequence is mapped into the state space of the approximating model. Instead of carrying out classification on the sequences directly, we discuss measuring the dissimilarity between models under different hypotheses. The classification experiment on binary synthetic data demonstrates robustness using appropriate measurement. The classifications on benchmark univariate and multivariate data confirm the advantages of the proposed approach compared with several common algorithms.

THU3C: BANDITS & TRANSFER LEARNING — ROOM 300A

16:40 - 17:00 B059: Linear Bandits in Unknown Environments
Thibault Gisselbrecht, Sylvain Lamprier, Patrick Gallinari

In contextual bandit problems, an agent has to choose an action among a bigger set of available ones at each decision step, according to features observed on them. The goal is to define a decision strategy that maximizes the cumulative reward of actions over time. We focus on the specific case where the features of each action correspond to some kind of a constant profile, which can be used to determine its intrinsic utility for the task in concern. If there exists an unknown linear application that allows rewards to be mapped from profiles, this can be leveraged to greatly improve the exploitation-exploration trade-off of stationary stochastic methods like UCB. In this paper, we consider the case where action profiles are unknown beforehand. Instead, the agent only observes sample vectors, with mean equal to the true profiles, for a subset of actions at each decision step. We propose a new algorithm, called SampLinUCB, and derive a finite time high probability upper bound on its regret. We also provide numerical experiments on a task of focused data capture from online social networks.

17:00 - 17:20 B060: Interpretable Domain Adaptation via Optimization over the Stiefel Manifold
Christian Poelitz, Wouter Duivesteijn, Katharina Morik

In domain adaptation, the goal is to find common ground between two, potentially differently distributed, data sets. By finding common concepts present in two sets of words pertaining to different domains, one could leverage the performance of a classifier for one domain for use on the other domain. We propose a solution to the domain adaptation task, by efficiently solving an optimization problem through Stochastic Gradient Descent. We provide update rules that allow us to run Stochastic Gradient Descent directly on a matrix manifold: the steps compel the solution to stay on the Stiefel manifold. This manifold encompasses projection matrices of word vectors onto low-dimensional latent feature representations, which allows us to interpret the results: the rotation magnitude of the word vector projection for a given word corresponds to the importance of that word towards making the adaptation. Beyond this interpretability benefit, experiments show that the Stiefel manifold method performs better than state-of-the-art methods.

17:20 - 17:40 B061: Communication-Efficient Distributed Online Learning with Kernels
Michael Kamp, Sebastian Bothe, Mario Boley, Michael Mock

We propose an efficient distributed online learning protocol for low-latency real-time services. It extends a previously presented protocol to kernelized online learners that represent their models by a support vector expansion. While such learners often achieve higher predictive performance than their linear counterparts, communicating the support vector expansions becomes inefficient for large numbers of support vectors. The proposed extension allows for a larger class of online learning algorithms - including those alleviating the problem above through model compression. In addition, we characterize the quality of the proposed protocol by introducing a novel criterion that requires the communication to be bounded by the loss suffered.

17:40 - 18:00 B062: Proactive Transfer Learning for Heterogeneous Feature and Label Spaces
Seungwhan Moon, Jaime Carbonell

We propose a framework for learning new target tasks by leveraging existing heterogeneous knowledge sources. Unlike the traditional transfer learning, we do not require explicit relations between source and target tasks, and instead let the learner actively mine transferable knowledge from a source dataset. To this end, we develop (1) a transfer learning method for source datasets with heterogeneous feature and label spaces, and (2) a proactive learning framework which progressively builds bridges between target and source domains in order to improve transfer accuracy. Experiments on a challenging transfer learning scenario (learning from hetero-lingual datasets with non-overlapping label spaces) show the efficacy of the proposed approach.

THU3D: RECOMMENDATION

ROOM 300B

16:40 - 17:00 B063: Top-N Recommendation via Joint Cross-Domain User Clustering and Similarity Learning
Dimitrios Rafailidis, Fabio Crestani

A cross-domain recommendation algorithm exploits user preferences from multiple domains to solve the data sparsity and cold-start problems, in order to improve the recommendation accuracy. In this study, we propose an efficient Joint cross-domain user Clustering and Similarity Learning recommendation algorithm, namely JCSL. We formulate a joint objective function to perform adaptive user clustering in each domain, when calculating the user-based and cluster-based similarities across the multiple domains. In addition, the objective function uses an $L_{2,1}$ regularization term, to consider the sparsity that occurs in the user-based and cluster-based similarities between multiple domains. The joint problem is solved via an efficient alternating optimization algorithm, which adapts the clustering solutions in each iteration so as to jointly compute the user-based and cluster-based similarities. Our experiments on ten cross-domain recommendation tasks show that JCSL outperforms other state-of-the-art cross-domain strategies.

17:00 - 17:20 B064: Collaborative Expert Recommendation for Community-Based Question Answering
Congfu Xu, Xin Wang, Yunhui Guo

With the development of Internet, users can share knowledge by asking and answering questions on community question answering (CQA) websites. How to find related experts to contribute their answers is hence worthy of studying. In this paper, we propose a recommendation algorithm called collaborative expert recommendation (CER) for this purpose. We take full advantage of the heterogeneous information including question tags, content, answer's votes, which are considered important for identifying experts. Moreover, we combine such information by a causal assumption of questions and answers, and inner connection exploitation among different types of information such as (questioner, question), (answer, question) and (answerer, question, answer) correlations, which are more explicable and reasonable comparing with the existing methods. Experiments carried out on six real-world datasets prove that CER has a better performance.

17:20 - 17:40 B065: Selecting Collaborative Filtering algorithms using Metalearning
Tiago Cunha, Carlos Soares, André Carvalho

Recommender Systems are an important tool in e-business, for both companies and customers. Several algorithms are available to developers, however, there is little guidance concerning which is the best algorithm for a specific recommendation problem. In this study, a metalearning approach is proposed to address this issue. It consists of relating the characteristics of problems (metafeatures) to the performance of recommendation algorithms. We propose a set of metafeatures based on the application of systematic procedure to develop metafeatures and by extending and generalizing the state of the art metafeatures for recommender systems. The approach is tested on a set of Matrix Factorization algorithms and a collection of real-world Collaborative Filtering datasets. The performance of these algorithms in these datasets is evaluated using several standard metrics. The algorithm selection problem is formulated as classification tasks, where the target attributes are the best Matrix Factorization algorithm, according to each metric. The results show that the approach is viable and that the metafeatures used contain information that is useful to predict the best algorithm for a dataset.

17:40 - 18:00 B066: Modeling Sequential Preferences with Dynamic User and Context Factors
Duc-Trong Le, Yuan Fang, Hady Lauw

Users express their preferences for items in diverse forms, through their liking for items, as well as through the sequence in which they consume items. The latter, referred to as"sequential preference", manifests itself in scenarios such as song or video playlists, topics one reads or writes about in social media, etc. The current approach to modeling sequential preferences relies primarily on the sequence information, i.e., which item follows another item. However, there are other important factors, due to either the user or the context, which may dynamically affect the way a sequence unfolds. In this work, we develop generative modeling of sequences, incorporating dynamic user-biased emission and context-biased transition for sequential preference. Experiments on publicly-available real-life datasets as well as synthetic data show significant improvements in accuracy at predicting the next item in a sequence.

THURSDAY SESSIONS, WITH ABSTRACTS

THU3E: MIXED GRILL

ROOM BELVEDERE

16:40 - 17:00 B067: Laplacian Hamiltonian Monte Carlo

Yizhe Zhang, Changyou Chen, Ricardo Henao, Lawrence Carin

We proposed a Hamiltonian Monte Carlo (HMC) method with Laplace kinetic energy, and demonstrate the connection between slice sampling and proposed HMC method in one-dimensional cases. Based on this connection, one can perform slice sampling using a numerical integrator in an HMC fashion. We provide theoretical analysis on the performance of such sampler in several univariate cases. Furthermore, the proposed approach extends the standard HMC by enabling sampling from discrete distributions. We compared our method with standard HMC on both synthetic and real data, and discuss its limitations and potential improvements.

17:00 - 17:20 B068: Augmented leverage score sampling with bounds

Daniel Perry, Ross Whitaker

We present an interesting modification to the traditional leverage score sampling approach by augmenting the scores with information from the data variance, which improves empirical performance on the column subsample selection problem (CSSP). We further present, to our knowledge, the first deterministic bounds for this augmented leverage score, and discuss how it compares to the traditional leverage score. We present some experimental results demonstrating how the augmented leverage score performs better than traditional leverage score sampling on CSSP in both a deterministic and probabilistic setting.

17:20 - 17:40 B069: Exact and efficient top-K inference for multi-target prediction by querying separable linear relational models

Michiel Stock, Krzysztof Dembczynski, Bernard De Baets, Willem Waegeman

Many complex multi-target prediction problems that concern large target spaces are characterised by a need for efficient prediction strategies that avoid the computation of predictions for all targets explicitly. Examples of such problems emerge in several subfields of machine learning, such as collaborative filtering, multi-label classification, dyadic prediction and biological network inference. In this article we analyse efficient and exact algorithms for computing the top-K predictions in the above problem settings, using a general class of models that we refer to as separable linear relational models. We show how to use those inference algorithms, which are modifications of well-known information retrieval methods, in a variety of machine learning settings. Furthermore, we study the possibility of scoring items incompletely, while still retaining an exact top-K retrieval. Experimental results in several application domains reveal that the so-called threshold algorithm is very scalable, performing often many orders of magnitude more efficiently than the naive approach.

17:40 - 18:00 B070: Differentially Private User Data Perturbation with Multi-Level Privacy Controls

Yilin Shen, Rui Chen, Hongxia Jin, Ninghui Li

Service providers typically collect user data for profiling users in order to provide high-quality services, yet this brings up user privacy concerns. One one hand, service providers oftentimes need to analyze multiple user data attributes that usually have different privacy concern levels. On the other hand, users often pose different trusts towards different service providers based on their reputation. However, it is unrealistic to repeatedly ask users to specify privacy levels for each data attribute towards each service provider. To solve this problem, we develop the *first* lightweight and provably framework that not only guarantees differential privacy on both *service provider* and *different data attributes* but also allows configurable *utility functions* based on service needs. Using various large-scale real-world datasets, our solution helps to significantly improve the utility up to 5 times with negligible computational overhead, especially towards numerous low reputed service providers in practice.

THU4A: DEMO SPOTLIGHTS 2

ROOM BELVEDERE

18:00 - 18:02 B071: A Tool for Subjective and Interactive Visual Data Exploration

Bo Kang, Kai Puolamaki, Jefrey Lijffijt, Tijl De Bie

We present SIDE, a tool for Subjective and Interactive Visual Data Exploration, which lets users explore high dimensional data via subjectively informative 2D data visualizations. Many existing visual analytics tools are either restricted to specific problems and domains or they aim to find visualizations that align with user's belief about the data. In contrast, our generic tool computes data visualizations that are surprising given a user's current understanding of the data. The user's belief state is represented as a set of projection tiles. Hence, this user-awareness offers users an efficient way to interactively explore yet-unknown features of complex high dimensional datasets.

18:02 - 18:04 B072: DANCer: dynamic attributed network with community structure generator

Christine Largeron, Oualid Benyahia, Baptiste Jeudy, Osmar Zaiane

We propose a new generator for dynamic attributed networks with community structure which follow the known properties of real-world networks such as preferential attachment, small world and homophily. After the generation, the different graphs forming the dynamic

network as well as its evolution can be displayed in the interface. Several measures are also computed to evaluate the properties verified by each graph. Finally, the generated dynamic network, the parameters and the measures can be saved as a collection of files.

18:04 - 18:06 **B073: Finding incident-related social media messages for emergency awareness**
Alexander Nieuwenhuijse, Jorn Bakker, Mykola Pechenizkiy

An information retrieval framework is proposed which searches for incident-related social media messages in an automated fashion. Using P2000 messages as an input for this framework and by extracting location information from text, using simple natural language processing techniques, a search for incident-related messages is conducted. A machine learned ranker is trained to create an ordering of the retrieved messages, based on their relevance. This provides an easy accessible interface for emergency response managers to aid them in their decision making process.

18:06 - 18:08 **B074: h(odor): Interactive Discovery of Hypotheses on the Structure-Odor Relationship in Neuroscience**
Guillaume Bosc, Marc Plantevit, Moustafa Bensafi, Jean-Francois Boulicaut, Mehdi Kaytoue

From a molecule to the brain perception, olfaction is a complex phenomenon that remains to be fully understood in neuroscience. Latest studies reveal that the physico-chemical properties of volatile molecules can partly explain the odor perception. Neuroscientists are then looking for new hypotheses to guide their research: physico-chemical descriptors distinguishing a subset of perceived odors. To answer this problem, we present the platform h(odor) that implements descriptive rule discovery algorithms suited for this task. Most importantly, the ol- faction experts can interact with the discovery algorithm to guide the search in a huge description space w.r.t their non-formalized background knowledge thanks to an ergonomic user interface.

18:08 - 18:10 **B075: Ranking Researchers through Collaboration Pattern Analysis**
Mario Cataldi, Luigi Di Caro, Claudio Schifanella

The academic world utterly relies on the concept of scientific collaboration. As in every collaborative network, however, the production of research articles follows hidden co-authoring principles as well as temporal dynamics which generate latent, and complex, collaboration patterns. In this paper, we present an online advanced tool for real-time rankings of computer scientists under these perspectives.

18:10 - 18:12 **B076: SITS-P2miner: Pattern-Based Mining of Satellite Image Time Series**
Tuan Nguyen, Nicolas MEGER, Christophe Rigotti, Catherine Pothier, Remi Andreoli

This paper presents a mining system for extracting patterns from Satellite Image Time Series. This system is a fully-fledged tool comprising four main modules for pre-processing, pattern extraction, pattern ranking and pattern visualization. It is based on the extraction of grouped frequent sequential patterns and on swap randomization.

18:12 - 18:14 **B077: Topy: Real-time Story Tracking via Social Tags**
Gevorg Poghosyan, M. Atif Qureshi, Georgiana Ifrim

We present the Topy system, which automates real-time story tracking by utilizing crowd-sourced tagging on social media platforms. The system employs a state-of-the-art Twitter hashtag recommender to continuously annotate news articles with hashtags, a rich meta-data source that allows connecting articles under drastically different timelines than typical keyword based story tracking systems. Employing social tags for story tracking has the following advantages: (1) story tracking via social annotation of news enables the detection of emerging concepts and topic drift; (2) hashtags allow going beyond topics by grouping articles based on connected themes (e.g., #rip, #blacklivesmatter, #icantbreath); (3) hashtags allow linking articles that focus on subplots of the same story (e.g., #palmyra, #isis, #refugeecrisis).

18:14 - 18:16 **B078: TwitterCracy: Exploratory Monitoring of Twitter Streams for the 2016 U.S. Presidential Election Cycle**
Muhammad Atif Qureshi, Arjumand Younus, Derek Greene

We present TwitterCracy, an exploratory search system that allows users to search and monitor across the Twitter streams of political entities. Its exploratory capabilities stem from the application of lightweight time-series based clustering together with biased PageRank to extract facets from tweets and presenting them in a manner that facilitates exploration.

FRIDAY 23
SEPTEMBER 2016

INVITED TALK

SEQUENCES, CHOICES, AND THEIR DYNAMICS

Speaker: Ravi Kumar – Google
Time: 09:00 – 09:50
Room: 1000A
Abstract:

Sequences arise in many online and offline settings: urls to visit, songs to listen to, videos to watch, restaurants to dine at, and so on. User-generated sequences are tightly related to mechanisms of choice, where a user must select one from a finite set of alternatives. In this talk, we will discuss a class of problems arising from studying such sequences and the role discrete choice theory plays in these problems. We will present modeling and algorithmic approaches to some of these problems and illustrate them in the context of large-scale data analysis.

Bio:

Ravi Kumar has been a senior staff research scientist at Google since 2012. Prior to this, he was a research staff member at the IBM Almaden Research Center and a principal research scientist at Yahoo! Research. His research interests include Web search and data mining, algorithms for massive data, and the theory of computation.

INDUSTRIAL KEYNOTE

TOWARDS INDUSTRIAL MACHINE INTELLIGENCE

Speaker: Michael May
Time: 10:00 - 10:45
Room: 1000B

Abstract:

The next decade will see a deep transformation of industrial applications by big data analytics, machine learning and the internet of things. Industrial applications have a number of unique features, setting them apart from other domains. Central for many industrial applications in the internet of things is time series data generated by often hundreds or thousands of sensors at a high rate, e.g. by a turbine or a smart grid. In a first wave of applications this data is centrally collected and analyzed in Map-Reduce or streaming systems for condition monitoring, root cause analysis, or predictive maintenance. The next step is to shift from centralized analysis to distributed in-field or in situ analytics, e.g in smart cities or smart grids. The final step will be a distributed, partially autonomous decision making and learning in massively distributed environments. In this talk I will give an overview on Siemens' journey through this transformation, highlight early successes, products and prototypes and point out future challenges on the way towards machine intelligence. I will also discuss architectural challenges for such systems from a Big Data point of view.

Bio:

Michael May is Head of the Technology Field Business Analytics & Monitoring at Siemens Corporate Technology, Munich, and responsible for eleven research groups in Europe, US, and Asia. Michael is driving research at Siemens in data analytics, machine learning and big data architectures. In the last two years he was responsible for creating the Sinalytics platform for Big Data applications across Siemens' business. Before joining Siemens in 2013, Michael was Head of the Knowledge Discovery Department at the Fraunhofer Institute for Intelligent Analysis and Information Systems in Bonn, Germany. In cooperation with industry he developed Big Data Analytics applications in sectors ranging from telecommunication, automotive, and retail to finance and advertising. Between 2002 and 2009 Michael coordinated two Europe-wide Data Mining Research Networks (KDNet, KDubiq). He was local chair of ICML 2005, ILP 2005 and program chair of the ECML/PKDD Industrial Track 2015. Michael did his PhD on machine discovery of causal relationships at the Graduate Programme for Cognitive Science at the University of Hamburg.

INDUSTRIAL KEYNOTE

MACHINE LEARNING CHALLENGES AT AMAZON

Speaker: Matthias Seeger
Time: 14:40 - 15:25
Room: 1000B

Abstract:

At Amazon, some of the world's largest and most diverse problems in e-commerce, logistics, digital content management, and cloud computing services are being addressed by machine learning on behalf of our customers. In this talk, I will give an overview of a number of key areas and associated machine learning challenges.

Bio:

Matthias Seeger got his PhD from Edinburgh. He had academic appointments at UC Berkeley, MPI Tuebingen, Saarbruecken, and EPF Lausanne. Currently, he is a principal applied scientist at Amazon in Berlin. His interests are in Bayesian methods, large scale probabilistic learning, active decision making and forecasting.

FRIDAY SESSIONS AT A GLANCE

Industrial 1 10:45 - 11:20 — Room 1000B

10:45 - 11:20 C001: Intelligent Urban Data Monitoring for Smart Cities
Nikolaos Zygouras, Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Nikos Zacheilas, Ioannis Mpoutsis, Vana Kalogeraki, stephen Lynch, Brendan O'Brien

11:20 - 11:40 Coffee Break

Industrial 2 11:40 - 13:20 — Room 1000B

11:40 - 12:05 C002: Using Social Media to Promote STEM Education: Matching College Students with Role Models
Ling He, Jiebo Luo, Lee Murphy

12:05 - 12:30 C003: Active Learning with Rationales for Identifying Operationally Significant Anomalies in Aviation
Manali Sharma, Kamalika Das, Mustafa Bilgic, Bryan Matthews, David Nielsen, Nikunj Oza

12:30 - 12:55 C004: Finding Dynamic Co-evolving Zones in Spatial-temporal time series Data
Yun Cheng, Xiucheng Li, Yan Li

12:55 - 13:20 C005: PULSE: A Real Time System for Crowd Flow Prediction at Metropolitan Subway Stations
Ermal Toto, Elke Rundensteiner, Yanhua Li, Kajal Claypool, Richard Jordan, Mariya Ishutkina, Jun Luo, Fan Zhang

13:30 - 14:40 Lunch Break

Industrial 3 15:25 - 16:15 — Room 1000B

15:25 - 15:50 C006: Engine Misfire Detection With Pervasive Mobile Audio
Joshua Siegel, Sumeet Kumar, Isaac Ehrenberg, Sanjay Sarma

15:50 - 16:15 C007: Do Street Fairs Boost Local Businesses? A Quasi-Experimental Analysis Using Social Network Data
Ke Zhang, Konstantinos Pelechrinis

16:20 – 16:40 Coffee Break

Industrial 4 16:40 - 17:55 — Room 1000B

16:40 - 17:05 C008: ECG Monitoring in Wearable Devices by Sparse Models
Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, Daniele Zambon, Giacomo Boracchi

17:05 - 17:30 C009: Concept Neurons - Handling Drift Issues for Real-Time Industrial Data Mining
Luis Matias, Joao Gama, Joao Moreira

17:30 - 17:55 C010: Automatic Detection of Non-Biological Artifacts in ECGs Acquired During Cardiac Computed Tomography
Rustem Bekmukhametov, Sebastian Pölsterl, Nassir Navab, thomas Allmendinger, Minh-duc Doan

FRIDAY SESSIONS, WITH ABSTRACTS

INDUSTRIAL 1

ROOM 1000B

10:45 - 11:20 C001: Intelligent Urban Data Monitoring for Smart Cities

Nikolaos Zygouras, Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Nikos Zacheilas, Ioannis Mpoutsis, Vana Kalogeraki, stephen Lynch, Brendan O'Brien

Traffic monitoring systems have recently become an essential service in smart cities as they enable the authorities to monitor and handle traffic incidents in real-time. However, these applications need to process streaming data from a highly heterogeneous and voluminous data sources in real-time. Moreover, it is necessary to be able to get feedback from the citizens to complement the information retrieved from the rest of the sensors. In this paper we present our system called INSIGHT which provides traffic event detection in Dublin by exploiting Big Data and crowdsourcing techniques. Our system is able to receive and to process input from multiple heterogeneous sources such as the locations of buses moving around the city, traffic volume information, social media, and the human crowd.

INDUSTRIAL 2

ROOM 1000B

11:40 - 12:05 C002: Using Social Media to Promote STEM Education: Matching College Students with Role Models

Ling He, Jiebo Luo, Lee Murphy

STEM (Science, Technology, Engineering, and Mathematics) fields have become increasingly central to U.S. economic competitiveness and growth. The shortage in the STEM workforce has brought promoting STEM education upfront. The rapid growth of social media usage provides a unique opportunity to predict users' real-life identities and interests from online texts and photos. In this paper, we propose an innovative approach by leveraging social media to promote STEM education: matching Twitter college student users with diverse LinkedIn STEM professionals using a ranking algorithm based on the similarities of their demographics and interests. We share the belief that increasing STEM presence in the form of introducing career role models who share similar interests and demographics will inspire students to develop interests in STEM related fields and emulate their models. Our evaluation on 2,000 real college students demonstrated the accuracy of our ranking algorithm. We also design a novel implementation that recommends matched role models to the students.

12:05 - 12:30 C003: Active Learning with Rationales for Identifying Operationally Significant Anomalies in Aviation

Manali Sharma, Kamalika Das, Mustafa Bilgic, Bryan Matthews, David Nielsen, Nikunj Oza

A major focus of the commercial aviation community is discovery of unknown safety events in flight operational data. Data-driven unsupervised anomaly detection methods are better at capturing unknown safety events compared to rule-based methods which only look for known violations. However, not all statistical anomalies that are discovered by these unsupervised anomaly detection methods are operationally significant (e.g., represent a safety concern). Subject Matter Experts (SMEs) have to spend significant time reviewing these statistical anomalies individually to identify a few operationally significant ones. In this paper we propose an active learning algorithm that incorporates SME feedback in the form of rationales to build a classifier that can distinguish between uninteresting and operationally significant anomalies. Experimental evaluation on real aviation data shows that our approach improves detection of operationally significant events by as much as 75% compared to the state-of-the-art. The learnt classifier also generalizes well to additional validation data sets.

12:30 - 12:55 C004: Finding Dynamic Co-evolving Zones in Spatial-temporal time series Data

Yun Cheng, Xiucheng Li, Yan Li

Co-evolving patterns exist in many Spatial-temporal time series Data, which shows invaluable information about evolving patterns of the data. However, due to the sensor readings' spatial and temporal heterogeneity, how to find the stable and dynamic co-evolving zones remains an unsolved issue. In this paper, we proposed a novel divide-and-conquer strategy to find the dynamic co-evolving zones that systematically leverages the heterogeneity challenges. The precision of spatial inference and temporal prediction improved by 7% and 8% respectively by using the found patterns, which shows the effectiveness of the found patterns. The system has also been deployed with the Haidian Ministry of Environmental Protection, Beijing, China, providing accurate spatial-temporal predictions and help the government make more scientific strategies for environment treatment.

12:55 - 13:20 C005: PULSE: A Real Time System for Crowd Flow Prediction at Metropolitan Subway Stations

Ermal Toto, Elke Rundensteiner, Yanhua Li, Kajal Claypool, Richard Jordan, Mariya Ishutkina, Jun Luo, Fan Zhang

The fast pace of urbanization has given rise to complex transportation networks, such as subway systems, that deploy smart card readers generating detailed transactions of mobility. Predictions of human movement based on these transaction streams represents tremendous new opportunities from optimizing fleet allocation of on-demand transportation such as UBER and LYFT to dynamic pricing of services.

However, transportation research thus far has primarily focused on tackling other challenges from traffic congestion to network capacity. To take on this new opportunity, we propose a real-time framework, called PULSE (Prediction Framework For Usage Load on Subway SystEms), that offers accurate multi-granular arrival crowd flow prediction at subway stations. PULSE extracts and employs two types of features such as streaming features and station profile features. Streaming features are time-variant features including time, weather, and historical traffic at subway stations (as time-series of arrival/departure streams), where station profile features capture the time-invariant unique characteristics of stations, including each station's peak hour crowd flow, remoteness from the downtown area, and mean flow, etc. Then, given a future prediction interval, we design novel stream feature selection and model selection algorithms to select the most appropriate machine learning models for each target station and tune that model by choosing an optimal subset of stream traffic features from other stations. We evaluate our PULSE framework using real transaction data of 11 million passengers from a subway system in Shenzhen, China. The results demonstrate that PULSE greatly improves the accuracy of predictions at all subway stations by up to 49% over baseline algorithms.

INDUSTRIAL 3

ROOM 1000B

15:25 - 15:50 C006: Engine Misfire Detection With Pervasive Mobile Audio

Joshua Siegel, Sumeet Kumar, Isaac Ehrenberg, Sanjay Sarma

We address the problem of detecting whether an engine is misfiring by using machine learning techniques on transformed audio data collected from a smartphone. We recorded audio samples in an uncontrolled environment and extracted Fourier, Wavelet and Mel-frequency Cepstrum features from normal and abnormal engines. We then implemented Fisher score and Relief score based variable ranking to obtain an informative reduced feature set for training and testing classification algorithms. Using this feature set, we were able to obtain a model accuracy of over 99% using a linear SVM applied to outsample data. This application of machine learning to vehicle subsystem monitoring simplifies traditional engine diagnostics, aiding vehicle owners in the maintenance process and opening up new avenues for pervasive mobile sensing and automotive diagnostics.

15:50 - 16:15 C007: Do Street Fairs Boost Local Businesses? A Quasi-Experimental Analysis Using Social Network Data

Ke Zhang, Konstantinos Pelechrinis

Local businesses and retail stores are a crucial part of local economy. Local governments design policies for facilitating the growth of these businesses that can consequently have positive externalities on the local community. However, many times these policies have completely opposite from the expected results (e.g., free curb parking instead of helping businesses has been illustrated to actually hurt them due to small turnover per spot). Hence, it is important to evaluate the outcome of such policies in order to provide educated decisions for the future. In the era of social and ubiquitous computing, mobile social media, such as Foursquare, form a platform that can help towards this goal. Data from these platforms capture semantic information of human mobility from which we can distill the potential economic activities taking place. In this paper we focus on street fairs (e.g., arts festivals) and evaluate their ability to boost economic activities in their vicinity. In particular, we collected data from Foursquare for the three month period between June 2015 and August 2015 from the city of Pittsburgh. During these period several street fairs took place. Using these events as our case study we analyzed the data utilizing propensity score matching and a quasi-experimental technique inspired by the difference-in differences method. Our results indicate that street fairs provide positive externalities to nearby businesses. We further analyzed the spatial reach of this impact and we find that it can extend up to 0.6 miles from the epicenter of the event.

FRIDAY 23
SEPTEMBER 2016

TUTORIALS - MORNING

LARGE-SCALE LEARNING FROM DATA STREAMS IN EVOLVING ENVIRONMENTS

Organizers: Moamar Sayed-Mouchaweh, João Gama, Hamid Bouchachia
Time: 10:00 - 13:40
Room: 300A

This tutorial aims at discussing the problem of learning from data streams generated by evolving nonstationary processes. It will overview the advances of techniques, methods and tools that are dedicated to manage, exploit and interpret data streams in non-stationary environments. In particular, the event will examine the problems of modeling, prediction, and classification based on learning from data streams by: - Defining the problem of learning from data streams in evolving and complex non-stationary environments, its interests, and challenges, - Providing a general scheme of methods and techniques treating the problem of learning of data streams in evolving and complex non-stationary environments, - Presenting the majors methods and techniques treating the problem of learning of data streams in evolving and complex non-stationary environments, - Discussing the application of these methods and techniques in various real-world problems. This tutorial is combined with a workshop treating the same topic, which will be held in the afternoon.

THE DATA SCIENTIST'S GUIDE FOR WRITING PAPERS

Organizers: : Nikolaj Tatti
Time: 10:00 - 13:40
Room: 300B

The goal of this tutorial is to provide guidelines and good practices on how to write scientific papers, with emphasis on writing technical sections. The tutorial consists of two parts. In the first part we will go over general philosophy for designing and typesetting mathematical notation, describing experiments, typesetting pseudo-code and tables, as well as, writing proofs. In addition, we will highlight typical mistakes done in computer science papers. The second part will focus on designing and typesetting high-quality visualizations. Here, our goal is two-fold: (i) we provide tools and ideas for designing complex non-standard visualizations, (ii) we provide guidelines on how to typeset standard plots, and highlight common errors done in data mining papers.

TUTORIALS – AFTERNOON

AN INTRODUCTION TO REDESCRIPTION MINING

Organizers: Esther Galbrun, Pauli Miettinen
Time: 14:40 - 18:20
Room: Meeting

A biologist interested in bioclimatic habitats of species needs to find geographical areas that admit two characterizations, one in terms of their climatic profile and one in terms of the occupying species. For an ethnographer, matching the terms used by individuals of an ethnic group to call one another to their genealogical relationships might help elucidate the meaning of kinship terms they use.

These are just two examples of a general problem setting where we need to identify correspondences between data that have different nature (species vs. climate, kinship terminology vs. genealogical linkage). To identify the correspondences over binary data sets, Ramakrishnan et al. proposed redescription mining in 2004. Subsequent research has extended the problem formulation to more complex correspondences and data types, making it applicable to wide variety of data analysis tasks.

This tutorial is targeted at attendees with no prior knowledge of redescription mining as well as seasoned experts on the field. We will present the intuition behind redescription mining, formulation variants, and links to existing data analysis techniques such as classification and subgroup discovery. We will also discuss the current algorithms, not forgetting applications, for instance, to the identification of bioclimatic niches or in circuit design.

PROBABILISTIC LOGICS IN MACHINE LEARNING

Organizers: Fabrizio Riguzzi
Time: 14:40 - 18:20
Room: Stampa

The combination of logic and probability proved very useful for modeling domains with complex and uncertain relationships among entities. Machine learning approaches based on such combinations have recently achieved important results, originating the fields of Statistical Relational Learning, Probabilistic Inductive Logic Programming and, more generally, Statistical Relational Artificial Intelligence. This tutorial will briefly introduce probabilistic logic programming and probabilistic description logics and overview the main systems for learning these formalisms both in terms of parameters and of structure. The tutorial includes a significant hands-on experience with the systems ProbLog2, PITA, TRILL and SLIPCOVER using their online interfaces.

LEARNING BAYESIAN NETWORKS FOR COMPLEX RELATIONAL DATA

Organizers: Oliver Schulte, Ted Kirkpatrick
Time: 14:40 - 18:20
Room: 300B

Many organizations maintain critical data in a relational database. The tutorial describes the statistical and algorithmic challenges of constructing models from such data, compared to the single-table data that is the traditional emphasis of machine learning. We extend the popular model of Bayesian networks to relational data, integrating probabilistic associations across all tables in the database. Extensive research has shown that such models, accounting for relationships between instances of the same entity (such as actors featured in the same movie) and between instances of different entities (such as ratings of movies by viewers), have greater predictive accuracy than models learned from a single table. We illustrate these challenges and their solutions in a real-life running example, the Internet Movie Database. The tutorial shows how Bayesian networks support several relational learning tasks, such as querying multi-relational frequencies, linkbased classification, and relational anomaly detection. We describe how standard machine learning concepts can be extended and adapted for relational data, such as model likelihood, sufficient statistics, model selection, and statistical consistency. The tutorial addresses researchers with a background in machine learning who wish to apply graphical models to relational data.

WORKSHOPS – FULL DAY

DLPM: DEEP LEARNING FOR PRECISION MEDICINE

Organizers: Bertram MüllerMyhsok, Cesare Furlanello
Time: 10:00 – 18:20
Room: Belvedere
Website: https://dlpm2016.fbk.eu/

Deep Learning is expected to have a disruptive impact for functional genomics, with applications of high industrial and ethical relevance in pharmacogenomics and toxicogenomics. Examples in miRNA prediction already demonstrated the potential for deriving implicit features with high predictive accuracy. Following the emerging thread of interest emerged at the NIPS MLCB 2016 workshop, we wish to discuss about the best options for the adoption of deep learning models, both for improved accuracy as well as for better biomedical understanding from identification of patterns from internal features. Questions such as end to end modeling from structure to functionality and biological impact as well as architectures for integration of genotype, expression and epigenetics would be of immediate interest for the workshop. We aim to create a connection between machine learning experts and leaders in the Precision Medicine initiatives in Europe and the USA. In particular, the workshop aims to link experts from the FDA SEQC2 initiative on Precision Medicine, which will pave the way for defining optimal procedures for the development of actionable drugs that can target phenotypeselected patient groups. We wish to discuss also technical challenges such as working with very large cohorts (e.g. from 60K to 300K subjects In molecular psychiatry studies) that are now amenable for modeling with deep learning. Further, family cohorts will challenge machine learning and bioinformatics experts for new efficient solutions. In summary, both methodological aspects from deep learning, machine learning, information technology, statistics as well as actual applications, pitfalls and (medical) needs are to be featured.

PROGRAM

SESSION I

10:00 - 10:10 Cesare Furlanello - DLPM2016 Opening
10:10 - 10:40 Bertram Müller–Myhsok - Machine Learning, Deep Learning and Precision Medicine
10:40 - 11.20 Invited talk: Munir Pirmohamed - Precision Medicine - current state and urgent needs

11:20 - 11:40 Coffee Break

SESSION II

11:40 - 12:20 Invited talk: Joshua Xu - Sequencing data quality control project phase 2 (SEQC2): a new FDA-led consortium effort to advance precision medicine
12:20 - 13:00 Invited talk: Abraham Heifets - AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery
13:00 - 13:10 Gurpreet Ghattaoraya - Opportunities to Implement Deep Learning within the Field of Immune-Mediated Adverse Drug Reactions
13:10 - 13:30 Pavel Kisilev - Learning to describe medical image findings using multi-task-loss CNN

13:30 - 14:40 Lunch Break

SESSION III

14:40 - 15:20 Invited talk: Djork–Arné Clevert - Rectified Factor Networks for Learning Sparse Representations in Pharmacogenomics
15:20 – 16:00 Invited talk: Kristel van Steen - Steering between holistic and reductionist approaches to tackle modern problems in the omics era
16:00 – 16:20 Cristobal Estaban - Combining Static and Dynamic Information for Clinical Event Prediction

16:20 – 16:40 Coffee Break

SESSION IV

16:40 - 16:50 Calogero Zarbo - Integrating Deep Learning with the SEQC Data Analysis Plan for predictive biomarkers of Clinical Endpoints in Neuroblastoma
16:50 - 17:30 Gunnar Raetsch - Drugs, Patients and Mutations: Data Science Approaches for Oncology
17:30 – 18:20 Closing discussion - Coordinators: Leming Shi

Wine & Cheese

WORKSHOPS – FULL DAY

MML: 9TH INTERNATIONAL WORKSHOP ON MUSIC AND MACHINE LEARNING

Organizers: Rafael Ramirez, Darrell Conklin, José M. Iñesta
Time: 10:00 – 18:20
Room: 100A
Website: https://sites.google.com/site/musicmachinelearning16/

Machine learning has permeated nearly every area of music informatics, driven by a profusion of recordings available in digital audio formats, steady improvements to the accessibility and quality of symbolic corpora, availability of powerful algorithms in standard machine learning toolboxes, and theoretical advances in machine learning and data mining. As complexity of the problems investigated by researchers on machine learning and music increases, there is a need to develop new algorithms and methods to solve these problems. As a consequence, research on machine learning and music is an active and growing field reflected in international meetings such as the prior iterations of the International Workshop on Machine Learning and Music (MML). The expected outcome of the workshop is to promote fruitful multidisciplinary collaboration among researchers (computer scientists, musicians, and musicologists) who are using machine learning techniques in musical applications, by providing the opportunity to discuss ongoing work in the area. The expected attendees are active researchers in machine learning and music who have special interest in content-based music processing. Researchers from other disciplines (e.g. Cognitive Science, Musicology, and Music Psychology) are also welcome to contribute to the workshop.

PROGRAM

10:00 - 10:05 Welcome and opening remarks

10:05 - 11:20 Oral Session I: Recommendation

Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Paiva - Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset

Arthur Flexer and Jeff Stevens - Mutual Proximity Graphs for Music Recommendation

Laura Pollacci, Riccardo Guidotti, and Giulio Rossetti – "Are we playing like Music-Stars?" Placing Emerging Artists on the Italian Music Scene

11:20 - 11:40 Coffee break

11:40 - 13:00 Oral Session II: Classification

Izaro Goienetxea, Kerstin Neubarth, and Darrell Conklin - Melody Classification with Pattern Covering

Jose J. Valero-Mas, Emmanouil Benetos, and José M. Iñesta - Classification-based Note Tracking for Automatic Music Transcription

Kerstin Neubarth and Darrell Conklin - Supervised Descriptive Folk Music Analysis Integrating Global and Event Features

13:00 - 13:40 Poster Craze

David Rizo, Jorge Calvo-Zaragoza, José M. Iñesta, and Plácido R. Illescas - Hidden Markov Models for Functional Analysis

Raymond Whorley and Darrell Conklin - A Transformational Method for Chorale Generation

Grigore Burloiu - Online Score-agnostic Tempo Models for Automatic Accompaniment

Matevž Pesek, Aleš Leonardis, and Matija Marolt - SymCHMMerge - Hypothesis Refinement for Pattern Discovery with a Compositional Hierarchical Model

Tetsuro Kitahara and Masaki Matsubara - Extracting Melodic Contour Using Wavelet-based Multi-resolution Analysis

Sergio Giraldo and Rafael Ramirez - A Genetic Approach for Evaluation of Computational Models for Expressive Music Performance in Jazz Guitar

Bass and Melody Part Identification in Multi-track MIDI Files

Octavio Vicente, Pedro J. Ponce de León, and José M. Iñesta

13:40 - 14:40 Lunch

14:40 - 16:20 Poster Session and networking

16:20 - 16:40 Coffee break

16:40 - 17:30 Oral Session III: Tempo and onset

Jose Luis Diez Antich, Mattia Paterna, Ricard Marxer, and Hendrik Purwins - Unsupervised Learning of Structural Representation of Percussive Audio Using a Hierarchical Dirichlet Process Hidden Markov Model

Sergio Giraldo, Rafael Ramirez, and William Rollin - Onset Detection using Machine Learning Ensemble Methods

17:30 Closing remarks

FRIDAY 23
SEPTEMBER 2016

WORKSHOPS – MORNING

DMNLP: 3RD EDITION OF THE WORKSHOP ON INTERACTIONS BETWEEN DATA MINING AND NATURAL LANGUAGE PROCESSING

Organizers: Peggy Cellier, Thierry Charnois, Adreas Hotho, Stan Matwin, Marie-Francine Moens, Yannick Toussaint
Time: 10:00 – 13:40
Room: Stampa
Website: http://dmnlp.loria.fr/

Recently, a new field has emerged taking benefit of both Data Mining and NLP domains. The objective of DMNLP is thus to provide a forum to discuss how Data Mining can be interesting for NLP tasks, providing symbolic knowledge, but also how NLP can enhance data mining approaches by providing richer and/or more complex information to mine and by integrating linguistic knowledge directly in the mining process. The workshop aims at bringing together researchers from both communities in order to stimulate discussions about the cross-fertilization of those two research fields. The idea of this workshop is to discuss future directions and new challenges emerging from the cross-fertilization of Data Mining and NLP and in the same time initiate collaborations between researchers of both communities.

PROGRAM

10:00 - 10:05 Opening Session

10:05 - 11:35 Session I

Alexander Dlikman and Mark Last - Using Machine Learning Methods and Linguistic Features in Single-Document Extractive Summarization

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger and Fotis Jannidis - Prediction of Happy Endings in German Novels

Julien Ah-Pine and Edmundo Pavel Soriano Morales - A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis

11:35 - 11:50 Coffe Break

11:50 - 13:20 Session II

Julien Velcin, Mathieu Roche and Pascal Poncelet - Shallow Text Clustering Does Not Mean Weak Topics: How Topic Identification Can Leverage Bigram Features

Christian Pölitz - Topic Models with Sparse and Group-Sparsity Inducing Priors

Fu Xianghua, Wu Haiying and Cui Laizhong - Topic Sentiment Joint Model with Word Embeddings

13:40 - 14:40 lunch break

DARE: 4TH INTERNATIONAL WORKSHOP ON DATA ANALYTICS FOR RENEWABLE ENERGY INTEGRATION

Organizers: Wei Lee Woon, Zeyar Aung, Stuart Madnick
Time: 10:00 – 13:40
Room: Presidenza
Website: http://dare2016.dnagroup.org/

In recent years, climate change, the depletion of natural resources and the resultant rises in energy costs have led to an increased focus on renewable sources of energy like wind and solar. A lot of research has been devoted to the technologies used to extract energy from these sources; however, once generated, this energy would have to be stored and distributed to consumers in a way that is efficient and cost effective – which would mean making use of traditional power distribution networks. Unfortunately these networks are frequently designed with large, centralized power stations in mind, while in contrast renewable energy sources tend to be spatially distributed and subject to large temporal variations in generation capacity. As such, a concerted research effort is required to integrate these resources into the existing infrastructure. This research is inherently multidisciplinary as it utilizes techniques from a large range of disciplines. Apart from electrical engineering, addressing this challenge will involve a particularly large number of computational aspects which call for the judicious use of techniques drawn from the domains of data analytics, pattern recognition and machine learning. Examples of relevant issues which are of critical importance include:

— Forecasting of renewable energy supply and demand
— Automated fault prediction, detection and identification
— Mining of electricity users' data to design customized tariffs and marketing plans
— Cyber-Security of Smart Grid facilities
— Decision support and coordination for supporting demand response applications

PROGRAM

10:00 - 10:05	Welcome Address
10:10 - 10:30	Alexander Kogler and Patrick Traxler - Locating Faults in Photovoltaic Systems Data
10:35 - 10:55	Thierry Zufferey, Andreas Ulbig, Stephan Koch and Gabriela Hug - Forecasting of Smart Meter Time Series Based on Neural Networks
11:00 - 11:20	Armin Alibasic, Wei Lee Woon and Zeyar Aung - Cyber Security for Smart Cities: Trends, Opportunities, and Challenges
11:20 - 11:40	Coffee Break
11:40 - 12:00	Alejandro Catalina, Alberto Torres-Barrán and José R. Dorronsoro - Machine Learning Prediction of Photovoltaic Energy from Satellite Sources
12:05 - 12:25	Carlos D. Zuluaga and Mauricio Álvarez - Approximate Probabilistic Power Flow
12:30 - 12:50	Robert Ulbricht, Anna Thoß, Hilko Donker, Gunter Gräfe and Wolfgang Lehner, Robotron - Dealing with Uncertainty: An Empirical Study on the Relevance of Renewable Energy Forecasting Methods
12:55 - 13:15	Stuart Madnick, Mohammad S. Jalali, Michael Siegel, Yang Lee, Diane Strong, Richard Wang, Wee Horng Ang, Vicki Deng, Dinsha Mistree - Measuring Stakeholders' Perceptions of Cybersecurity for Renewable Energy Systems
13:20 - 13:40	Björn Wolff, Oliver Kramer and Detlev Heinemann - Selection of numerical weather forecast features for PV power predictions with Random Forests

FRIDAY 23 SEPTEMBER 2016

WORKSHOPS – MORNING

MLLS: 4TH WORKSHOP ON MACHINE LEARNING IN LIFE SCIENCES

Organizers: Bartosz Krawczyk, Michał Woźniak
Time: 10:00 – 13:40
Room: Meeting
Website: http://mlls.kssk.pwr.edu.pl/

Life sciences, ranging from medicine, biology and genetics to biochemistry and pharmacology have developed rapidly in previous years. Computerization of those domains allowed to gather and store enormous collections of data. Analysis of such vast amounts of information without any support is impossible for human being. Therefore recently machine learning and pattern recognition methods have attracted the attention of broad spectrum of experts from life sciences domain. The aim of this Workshop is to stress the importance of interdisciplinary collaboration between life and computer sciences and to provide an international forum for both practitioners seeking new cutting-edge tools for solving their domain problems and theoreticians seeking interesting and real-life applications for their novel algorithms. We are interested in novel machine learning technologies, designed to tackle complex medical, biological, chemical or environmental data that take into consideration the specific background knowledge and interactions between the considered problems. We look for novel applications of machine learning and pattern recognition tools to contemporary life sciences problems, that will shed light on their strengths and weaknesses. We are interested in new methods for data visualization and methods for accessible presentation of results of machine learning analysis to life scientists. We welcome new findings in the intelligent processing of non-stationary medical, biological and chemical data and in proposals for efficient fusion of information coming from multiple sources. Papers on efficient analysis and classification of bid data (understood as both massive volumes and high-dimensionality problems) will be of special interest to this Workshop.

PROGRAM

10:00 - 11.25	Session I
10:00 - 10:10	Welcome message from the Workshop Chairs
10:10 - 11:10	Invited talk: Jerzy Stefanowski - Evaluation of Interestingness and Interaction of Attribute-Value Conditions in Discovered Rules: Applications in Medical Data Analysis
11:10 - 11:25	Open panel discussion
11:40 - 13:40	Session II
11:40 - 12:00	S. Polsterl, N. Navab, A. Katouzian - An Efficient Training Algorithm for Kernel Survival Support Vector Machines
12:00 - 12:20	P. Rubbens, R. Props, N. Boon, W. Waegeman - Learning in silico communities to perform flow cytometric identification of synthetic bacterial communities
12:20 - 12:40	S. Nakoneczny, M. Smieja - Natural language processing methods in biological activity prediction
12:40 - 13:00	V. Vidulin, M. Brbic, F. Supek, T. Smuc - Evaluation of Fusion Approaches in Large-scale Bio-annotation Setting
13:00 - 13:20	B. Krawczyk, M. Wozniak - Online Weighted Naive Bayes for Automated Fetal State Assessment
13:20 - 13:40	P. Ksieniewicz, M. Wozniak - Imbalance medical data classification using Exposer Classifier Ensemble Medical Data Analysis

WORKSHOPS – AFTERNOON

DADA: 1ST INTERNATIONAL WORKSHOP ON DOMAIN ADAPTATION FOR DIALOG AGENTS

Organizers: Himanshu Sharad Bhatt, Sandipan Dandapat, Shourya Roy, Björn Gambäck
Time: 14:40 – 18:20
Room: 100B
Website: https://sites.google.com/site/ecmldaworkshop/home

Traditional statistical machine learning algorithms assume that training and test data are independently and identically distributed (i.i.d.) samples drawn from a distribution and perform well under this assumption. In real world applications, this assumption often does not hold true and the performance of a machine learning based system is severely affected when the data distribution in the test domain differs from that in the training domain. Under such scenario, the system has to be trained from beginning on the new data available from the test domain. However, the test data is unlabelled and it requires human effort to label the data, which is expensive and it is not a pragmatic solution. On the other hand, recent advances in transfer learning and domain adaptation (TL/DA) techniques allow domains, tasks, and distributions used in training and testing to be different, but related. It works in contrast to traditional supervised techniques on the principle of transferring learned knowledge across domains, thus, minimising the need for annotated training data from the test domain and learning new models from scratch. TL/DA has found many interesting applications; however, the main focus of this workshop is towards one of the recently emerging dialog system's technology. Research in dialog systems has focused on the semantics and pragmatics of dialog, hand-crafted designs, computational linguistics, evaluation of dialog systems, and standardization across research community (with continuing efforts in venues like SemDial, SIGDial, and YRRSDS). Dialog systems across different applications encounter huge variability in the nature of task/application (e.g., from help desks to technical support) and the user's behavior (e.g., cooperativeness, novice) and thus recent trends in dialog systems reflect a transition from hand-crafted designs to statistical machine learning methods to address such variability. Automatic learning of dialog strategies and related components of dialog systems have been a leading research area for the last few years with several papers having been published in leading conferences such as ACL, IJCAI, ECML, ICML, NAACL, EMNLP, and CIKM. Statistical machine learning based dialog systems that can have dialogs like humans, require large amounts of labelled data to train models. Consequently, expanding the scope of a dialog system from one domain to another requires significant human labelling and model building efforts. This has been one of the major bottlenecks in expansion of dialog systems to different domains and/or tasks. The focus of the proposed workshop is to provide an engaging venue to researchers for proposing novel solutions involving applications of transfer learning and domain adaptation for next generation dialog systems, discussing technical challenges, and new possibilities in the field.

PROGRAM

14:40 - 14:50	Introduction and Overview
	Session I
14:50 - 15:50	Keynote Talk : Hary C Bunt - Dialogue Annotation and Domain Adaptation
15:50 - 16:20	Invited talk 1 : Marie-Francine Moens - Domain Adaptation in Natural Language Understanding
16:20 - 16:40	Coffee Break
	Session II
16:40 - 17:10	Invited talk 2 : Giuseppe Riccardi - Towards Personal Helthcare Agents (PHA)
17:10 - 17:30	Björn Gambäck and Lars Bungum - Linguistic Domains and Adaptable Companionable Agents
17:30 - 17:50	Simon Keizer and Verena Rieser - The MaDrIgAL project: Multi-Dimensional Interaction Management and Adaptive Learning
17:50 - 18:10	Alexandros Papangelis and Yannis Stylianou - Multi-domain Spoken Dialogue Systems using Domain-Independent Parameterization
18:10 - 18:35	Invited talk 3: Talk by Yannick Toussaint
18:35 - 19:00	Invited Talk 4: Talk by Prof Oliver Lemon

FRIDAY 23 SEPTEMBER 2016

WORKSHOPS – AFTERNOON

STREAMEVOLV: LEARNING FROM DATA STREAMS IN LARGE-SCALE EVOLVING ENVIRONMENTS: CHALLENGES, METHODS AND APPLICATIONS

Organizers: Moamar Sayed-Mouchaweh, João Gama, Hamid Bouchachia, Rita Ribeiro
Time: 14:40 – 18:20
Room: 300A
Website: https://sites.google.com/site/streamevolv2016/

The volume of data is rapidly increasing due to the development of the technology of information and communication. This data comes mostly in the form of streams. Learning from this ever-growing amount of data requires flexible learning models that self-adapt over time. In addition, these models must take into account many constraints: (pseudo) real-time processing, high-velocity, and dynamic multi-form change such as concept drift and novelty. This workshop welcomes novel research about learning from data streams in evolving environments. It will provide the researchers and participants with a forum for exchanging ideas, presenting recent advances and discussing challenges related to data streams processing. It solicits original work, already completed or in progress. Position papers are also considered. This workshop is combined with a tutorial treating the same topic and taking place the morning of the same day.

PROGRAM

14:40 - 16:20	Session I
14:40 - 14:55	Fabiola Pereira, Sandra de Amo and João Gama - Detecting Events in Evolving Social Networks through Node Centrality Analysis
15:00 - 15:15	João Vinagre, Alipio M. Jorge and Joao Gama - Online bagging for recommendation with incremental matrix factorization
15:20 - 15:35	Ricardo Sousa and João Gama - First Principle Models Based Dataset Generation for Multi-Target Regression and Multi-Label Classification Evaluation
15:40 - 15:55	Hugo Cardoso and João Mendes-Moreira - Improving Human Activity Classification through Online Semi-Supervised Learning
16:20 - 16:40	Coffee Break
16:40 - 18:20	Session II
16:40 - 16:55	Olena Rudenko, Markus Endres, Patrick Roocks and Werner Kießling - A Preference-based Stream Analyzer
17:00 - 17:15	Manuel Mourato, João Moreira and Tânia Correia - Online Failure Prevention from Connected Heating Systems
17:20 - 17:35	Saad Mohamad, Moamar Sayed-Mouchaweh and Abdelhamid Bouchachia - Active Learning for Data Streams under Concept Drift and concept evolution
17:40 - 17:55	Waqas Jamil, Yuri Kaliniskan and Abdelhamid Bouchachia - Aggregation Algorithm vs. Average For Time Series Prediction
17:00 – 17:15	Julie Soulas - Frequent episode mining over the latest window using approximate support counting

DISCOVERY CHALLENGES – MORNING

NETCLA: THE ECML-PKDD NETWORK CLASSIFICATION CHALLENGE

Discovery Challenge Chairs: Elio Masciari, Alessandro Moschitti
Challenge Organizers: Daniele Bonadiman, Susanne Greiner, Luca di Stefano, Olga Uryupina
Time: 10:00 – 11:20
Room: 100B
Website: http://www.neteye-blog.com/netcla-the-ecml-pkdd-network-classification-challenge/

In recent years, there have been many proposals pushing for the use of Machine Learning (ML) in automatic network management. This challenge is one of the first explorations of ML for automatic network analysis. Our goal is to promote the use of ML for network-related tasks in general and, at the same time, to assess the participants' ability to quickly build a learning-based system showing a reliable performance. Additionally, one difficulty of using ML for network-related applications is the lack of datasets for training and evaluating different algorithms. The challenge provides one of the few datasets for this field, which may become a reference point for future and more advanced research.
As this is one of the first initiative in network classification, we started with a relatively simple multi-class single label classification task, where the labels are standard applications and signals are static network parameters. A more detailed description can be found on the challenge website.

PROGRAM

10:00 - 10:05 Opening
10:05 - 10:20 Olga Uryupina - Machine Learning for 5G applications
10:20 - 10:30 S. Greiner - NetEye: Monitoring User Network Experience
10:30 - 10:45 D. Bonadiman - Data, Task and System Statistics
10:45 - 11:20 Descriptions of Outstanding Participant Approaches

DISCOVERY CHALLENGES – MORNING

CQA CHALLENGE: LEARNING TO RE-RANK QUESTIONS FOR COMMUNITY QUESTION ANSWERING

Discovery Challenge Chairs: Elio Masciari, Alessandro Moschitti
cQA Challenge Chairs: Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Preslav Nakov
Time: 11:40 – 13:40
Room: 100B
Website: http://alt.qcri.org/ecml2016/

Due to the extended use of Web forums, such as Yahoo! Answers or Stackoverflow, there has been a renewed interest in Community Question Answering (cQA). cQA combines traditional question answering with a modern Web scenario, where users pose questions hoping to get the right answers from other users. The most critical problem arises when a new question is asked in the forum. If the user's question is similar (even semantically equivalent) to a previously posted question, she/he should not wait for answers or for another user to address her/him to the relevant thread already archived in the forum. An automatic system can search for previously-posted relevant questions and instantaneously provide the found information.
In this challenge, given a new question and a set of questions previously posted to a forum, together with their corresponding answer threads, a machine learning model must rank the forum questions according to their relevance against the new user question.
Even if this task involves both Natural Language Processing (NLP) and Information Retrieval, the challenge focuses on the machine learning aspects of reranking the relevant questions. Therefore, we provide both the initial rank and the feature representation of training and test examples to the participants. We extract features from the text of the user and forum questions using advanced NLP techniques, e.g., syntactic parsing. Most interestingly, we also provide the Gram matrices of tree kernels applied to advanced structural tree representation. A few other features express the relevance of the thread comments, associated with the forum questions, against the user question.
Participants are expected to exploit these data for building novel and effective machine learning models for reranking the initial question list in a better rank according to Mean Average Precision (MAP).

PROGRAM

11:40 - 11:45	Opening
11:45 - 12:00	G. Da San Martino - Task and Data Description
12:00 - 12:25	S. Filice - Structural Kernel Methods for Recognizing Similar Questions in cQA
12:25 - 12:35	A. Barron Cedeno - Statistics on the Challenge
12:35 - 12:55	J. Kubota, T. Unoki, K. Umezawa - SVM-based Modeling with Pairwise Transformation for Learning to Re-Rank
12:55 - 13:15	M. Nguyen, V. Phan, T. Nguyen, M. Nguyen - Learning to Rank Questions for Community Question Answering with Ranking SVM
13:15 - 13:25	Concluding remarks

FRIDAY 23	THURSDAY 22	WEDNESDAY 21	TUESDAY 20	MONDAY 19

Riva del Garda

1. Albergo Alle Porte
2. Ambassador Suite Hotel
3. Appartamenti Sembenini
4. Garda Sporting Club Hotel
5. Grand Hotel Liberty
6. Grand Hotel Riva
7. Hotel Bellavista
8. Hotel Bristol
9. Hotel Canarino
10. Hotel Gabry
11. Hotel Gardesana
12. Hotel Giardino Verdi
13. Hotel Luise
14. Hotel Portici
15. Hotel Royal
16. Hotel Rudy
17. Hotel Sole
18. Hotel Venezia
19. Hotel Villa Miravalle
20. Ostello Benacus

Torbole sul Garda

21. Hotel Lido Blu
22. Hotel Piccolo Mondo

NOTES

NOTES

Impaginazione
Marco Veneri

Finito di stampare nel mese di settembre 2016 da
Tipolitografia "La Reclame" - Trento