ECML-PKDD 2012 is organised by the Intelligent Systems Laboratory of the University of Bristol, United Kingdom.

We gratefully acknowledge the financial support of our Platinum sponsors:

WINTON

University of BRISTOL

PASCAL2

our Gold sponsors:

hp

rapid-i REPORT THE FUTURE

our Silver sponsors:

CAMBRIDGE UNIVERSITY PRESS

Google

YAHOO! LABS

KNIME

our Bronze sponsors:

www.eternalS.eu

IBM

Microsoft Research

our Award sponsors:

Machine Learning

Data Mining and Knowledge Discovery

ECML-PKDD 2012

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

24 - 28 September | Bristol | UK

CONFERENCE PROGRAMME

Background image: Wills Memorial Building, University of Bristol by zedphoto.com

ECML PKDD 2012

INTELLIGENT SYSTEMS LAB

University of BRISTOL

ECML-PKDD 2012

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

24 - 28 September | Bristol | UK

CONFERENCE

Start Time		Monday 24/09/2012	Tuesday 25/09/2012				Wednesday 26/09/2012	
9.00	9.00	Tutorials and Workshops *(Wills)*	Invited Talk **Pieter Abbeel** *(Chem-LT1)*				Plenary **Ten-Year Award Talk** *(Chem-LT1)*	
	10.00		*Coffee break*				*Coffee break*	
10.30	10.30	*Coffee break*	Tue1A **Bayesian Learning and Graphical Models** *(Chem-LT1)*	Tue1B **Association Rules and Frequent Patterns** *(Chem-LT2)*	Tue1C **Reinforcement Learning and Planning I** *(Chem-LT3)*	Tue1D **Industry Track** *(Chem-LT4)*	Wed1A **Distance-Based Methods and Kernels** *(Chem-LT1)*	Wed1B **Time Series and Temporal Data Mining** *(Chem-LT2)*
11.00		Tutorials and Workshops *(Wills)*						
12.00	12.10	*Lunch break*	*Lunch break*				*Lunch break*	
13.30	14.00	Tutorials and Workshops *(Wills)*	Tue2A **Graphs, Trees, Sequences and Strings I** *(Chem-LT1)*	Tue2B **Rankings and Recommendations** *(Chem-LT2)*	Tue2C **Ensemble Methods** *(Chem-LT3)*	Tue2D **Industry Track** *(Chem-LT4)*	Wed2A **Social Network Mining I** *(Chem-LT1)*	Wed2B **Large-Scale, Distributed and Parallel Mining and Learning I** *(Chem-LT2)*
	15.40		*Coffee break*				*Coffee break*	
16.00	16.10	*Coffee break*	Tue3A **Dimensionality Reduction, Feature Selection and Extraction** *(Chem-LT1)*	Tue3B **Multi-Relational Mining and Learning** *(Chem-LT2)*	Tue3C **Semi-Supervised and Transductive Learning** *(Chem-LT3)*	Tue3D **Demo Spotlights** *(Chem-LT4)*	Wed3A **Data Mining Process** *(Chem-LT1)*	Wed3B **Sensor Data** *(Chem-LT2)*
16.30		Opening and Awards *(Wills-GreatHall)*						
17.00		Invited Talk **Padhraic Smyth** *(Wills-GreatHall)*						
	17.50							
18.00	19.00	Welcome Reception *(Bristol Museum)*	**Poster session and software demos** *(Victoria Rooms)*				**Conference dinner** *(Thistle Hotel)*	

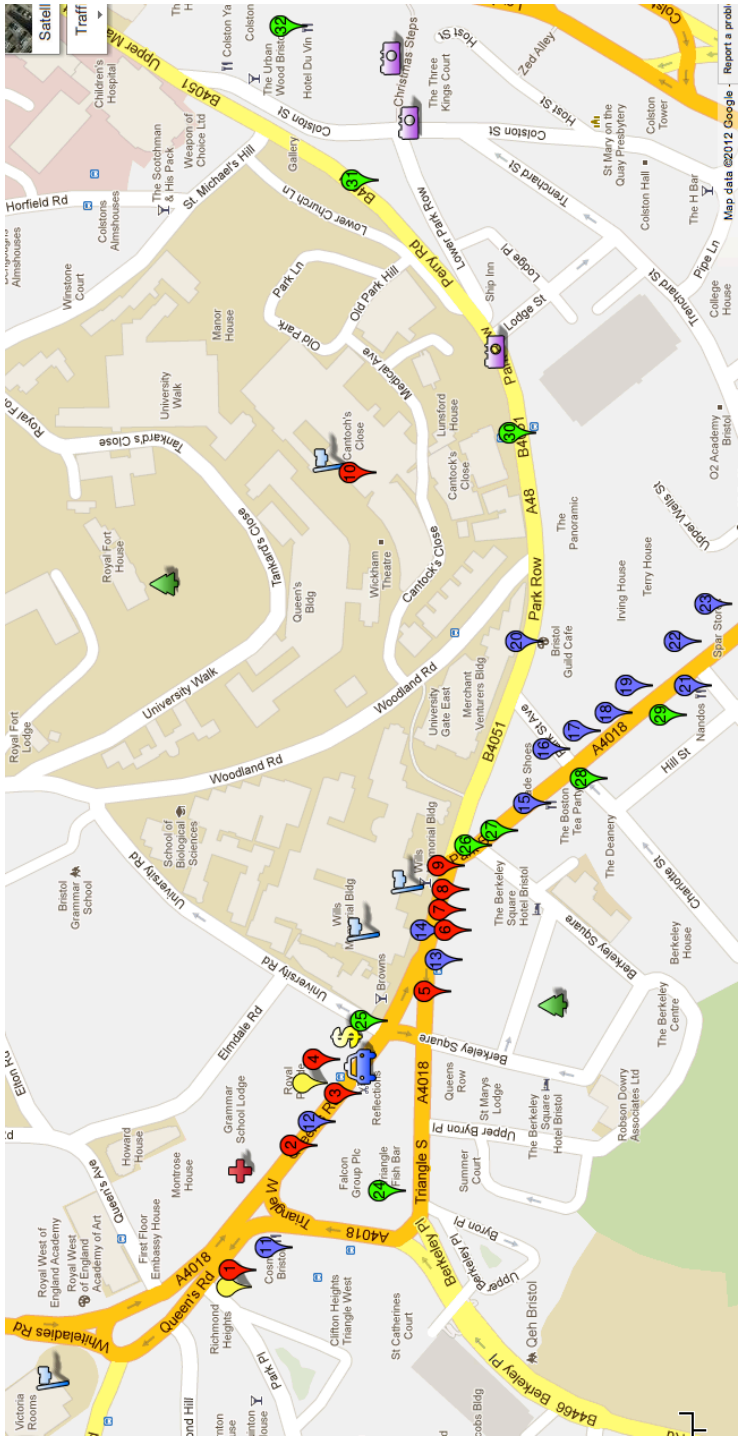
KEY

Tue2C
Ensemble Methods
(Chem-LT3)

- session ID
- session title
- colour code
- location name

- regular session
- plenary/invited talk
- poster session
- opening/invited talk
- reception
- workshop / tutorial
- conference dinner
- farewell
- other

ECMLPKDD2012.net



Places to eat

See http://www.ecmlpkdd2012.net/map for more

This is only a selection of places close to the conference venues. See the map on the next page for locations.

Quick Lunch / Take Away: Sandwich, drink and snacks. Sandwiches ≤ £5. No seating or very limited number of seats.

Affordable Restaurants: Reasonably priced hot meals. Seating. Price range for a main dish £5– £10.

Moderate / Expensive Restaurants: More expensive meals. Price range for a main dish ≥ £10 (normally between £10 and £15).

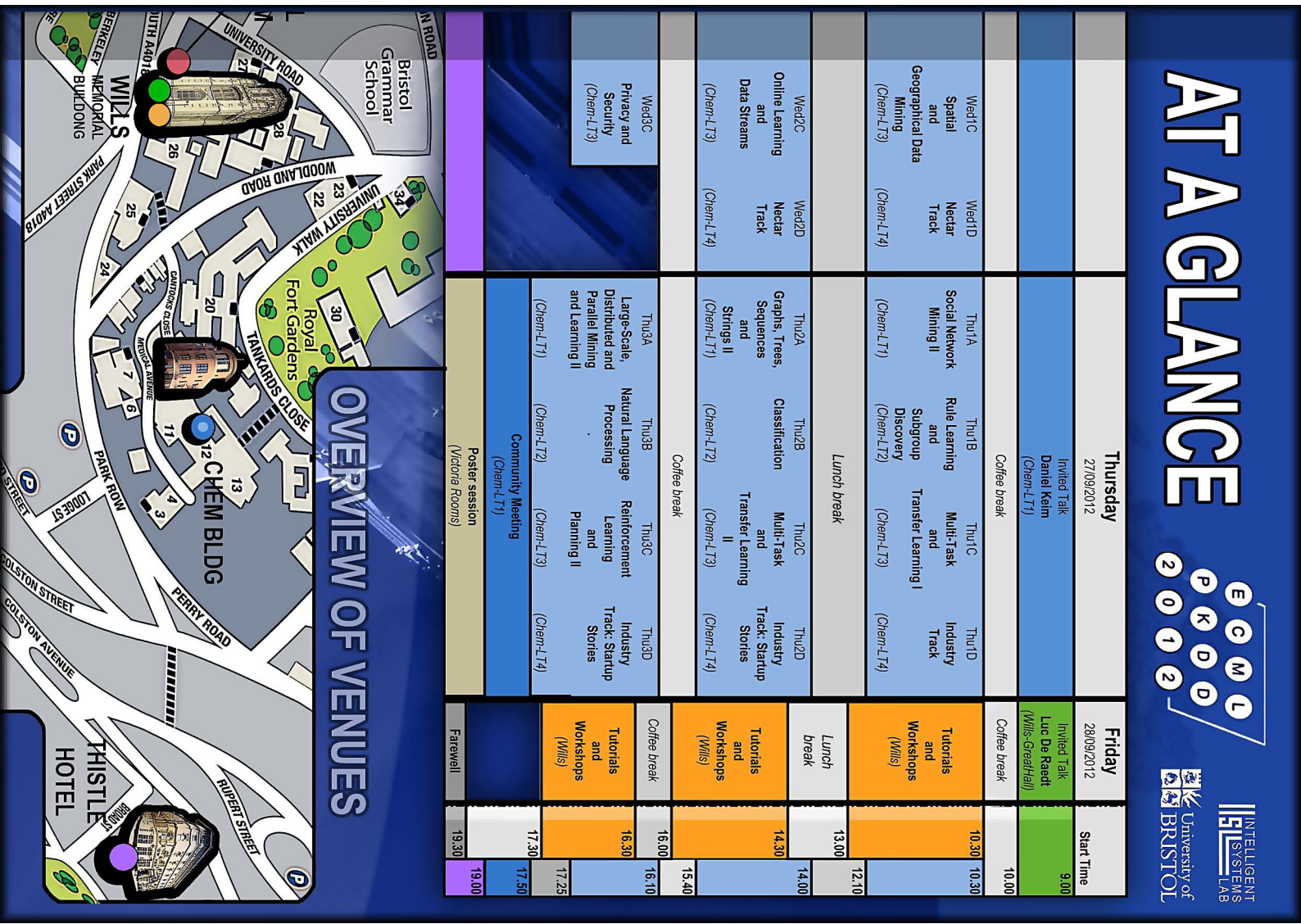
Quick Lunch / Take away		Affordable		Moderate / Expensive	
1	Waitrose Supermarket	11	Cosmo (Pan Asian buffet)	24	Oz (Turkish and Mediterranean)
2	Pappa Costa	12	Wagamama (Japanese)	25	Browns Bar & Brasserie (Fine dining)
3	Hermanos Cafe Bar	13	Pret-a-manger	26	Jamie's Italian (Fine dining)
4	Sainsbury's Supermarket	14	Wetherspoons (Pub)	27	Café Rouge (French)
5	Greggs	15	The Boston Tea Party	28	Goldbrick House (Fine dining)
6	Caffè Nero	16	Gourmet Burger Kitchen	29	Ask Restaurants (Italian)
7	Subway	17	Toro Pan Asian Cuisine	30	Mamma Mia (Italian)
8	Caffe Gusto	18	HK Dinner (Chinese)	31	Zero Degrees (Microbrewery and Restaurant)
9	Boulangerie	19	Mission Burrito (Mexican)	32	Hotel du Vin (Fine dining)
10	Chemistry Cafe	20	Bristol Guild Café		
		21	Nando's (Portuguese fast food)		
		22	All In One Bar & Restaurant		
		23	Antix Bar & Restaurant		

Other places of interest

- Boots pharmacy is marked with a +
- A cash machine is marked with a $. A bank is between 4 and 25.
- A taxi rank is marked with a car. Or phone 0117 926 4001 or 0117 955 5000.
- Conference venues are marked with flags.
- Parks are marked with trees.
- 2 gyms are marked with yellow bubbles:
  - Anytime Fitness offers a free 3-day trial. Separately, a guest pass is £8. To book a guest pass you must ring 0117 927 7225 in advance. Next to 2.
  - Cannons Health Club offers a free 2-day trial. Located next to 1.
- Bristol City Museum and Art Gallery is just West of the Wills Memorial Building. Open Monday to Friday: 10am–5pm, Saturday, Sunday: 10am–6pm. Free entry.
- The Red Lodge historic house is between 30 and 31 and is open Wednesday, Thursday, Saturday & Sunday 10:30am–4pm. Free entry.
- The beautiful gardens of Royal Fort House are a 2-minute walk North from Chemistry. (Go up the stairs, turn left and go up the steps.)

Miscellaneous Information

- Exchange rates:
  1 pound = 1.25 euros 1 euro = 0.79 pounds
  1 pound = 1.60 US dollars 1 dollar = 0.62 pounds
- Tipping in the UK:
  - Bars, cafes, buffets and coffee shops: nothing.
  - Taxis: between nothing and £1.
  - Restaurants with waiters: 10-15%.
- Google maps has 'Wills Memorial Bldg' in two places. The one on the left is actually the City Museum.



Sentiment Discovery from Affective Data

Mohamed Medhat Gaber, Mihaela Cocea, Stephan Weibelzahl, Ernestina Menasalvas and Cyril Labbé

09:00–17:00 Room: Wills, 1.5

The current expansion of social media leads to masses of affective data related to people's emotions, sentiments and opinions. Knowledge discovery from such data is an emerging area of research in the past few years, with a potential number of applications of paramount importance to business organisations, individual users and governments. Data mining and machine learning techniques are used to discover knowledge from various types of affective data such as ratings, text or browsing data. Although research in this area has grown considerably in the recent years, knowledge discovery from affective data is in its infancy state with more open issues and challenges which often requires interdisciplinary approaches.

10:30–10:35 **Welcome Address**
Mohamed Medhat Gaber

10:35–11:30 **Invited Talk: Multi-modal Affective Data Analytics**
Mykola Pechenizkiy

11:30–12:00 **New Features for Sentiment Analysis: Do Sentences Matter?**
Gizem Gezici, Berrin Yanikoglu, Dilek Tapucu and Yücel Saygin

12:00–12:30 **Product Reputation Model: An Opinion Mining Based Approach**
Ahmad Abdel-Hafez, Yue Xu and Dian Tjondronegoro

12:30–13:00 **Building Word-Emotion Mapping Dictionary for Online News**
Yanghui Rao, Xiaojun Quan, Liu Wenyin, Qing Li and Mingliang Chen

13:00–14:30 **Lunch (not provided)**

14:45–15:00 **Predicting Emotion Labels for Chinese Microblog Texts**
Zheng Yuan and Matthew Purver

15:00–15:15 **Feature Weighting Strategies in Sentiment Analysis**
Olena Kummer and Jacques Savoy

15:15–15:45 **Sentimentor: Sentiment Analysis of Twitter Streams**
Gulden Uchyigit and James Sepencer

15:45–16:00 **Mining for Opinions Across Domains: A Cross-Language Study**
Anna Kravchenko

16:00–16:30 **Coffee Break**

16:30–17:00 **Comparative Experiments for Multilingual Sentiment Analysis using Machine Translation**
Alexandra Balahur and Marco Turchi

17:00–17:15 **Towards an Abstractive Opinion Summarisation of Multiple Reviews in the Tourism Domain**
Cyril Labbé and François Portet

Collective Learning and Inference on Structured Data

Balaraman Ravindran, Kristian Kersting, Sriraam Natarajan and Subramanian Shivashankar

10:30–17:30 Room: Wills, 3.30

Classical ML techniques assume the data to be iid, but the real world data is inherently relational and can generally be represented using graphs or some variants of them. The importance of modelling structured data is evident from its increasing presence: WWW, social networks, organizational network, image, protein sequence, relational data etc. This field has been recently receiving a lot of attention in the community under different themes depending on the problem addressed and the nature of solution. Variants include iterative classification, structured prediction, relational learning, etc. While there are other issues such as learning the network structure, CoLISD focuses on the within-network learning and inference tasks with special emphasis on collective inference.

10:30–10:40 **Welcome Address**
Organizers

10:40–11:40 **Invited Talk by**
Charles Sutton

11:40–12:05 **Characterizing Retweeting Behaviors in Twitter: On the use of Text vs. Concepts**
Sofus Macskassy

12:05–12.30 **Ranking Mechanisms for Maximizing Spread, Trust in Signed Social Networks**
Ramasuri Narayanam, N.N.R. Ranga Suri, Vikas K. Garg and M. Narasimha Murty

12:30–12:55 **Link Prediction via Generalized Coupled Tensor Factorisation**
Beyza Ermis, Evrim Acar and A. Taylan Cemgil

12:55–13:00 **Poster Highlights - Node Classification in Partially Labeled Social Networks**
S Shamshu Dharwez, Subramanian Shivashankar and Balaraman Ravindran

13:00–14:10 **Lunch (not provided)**

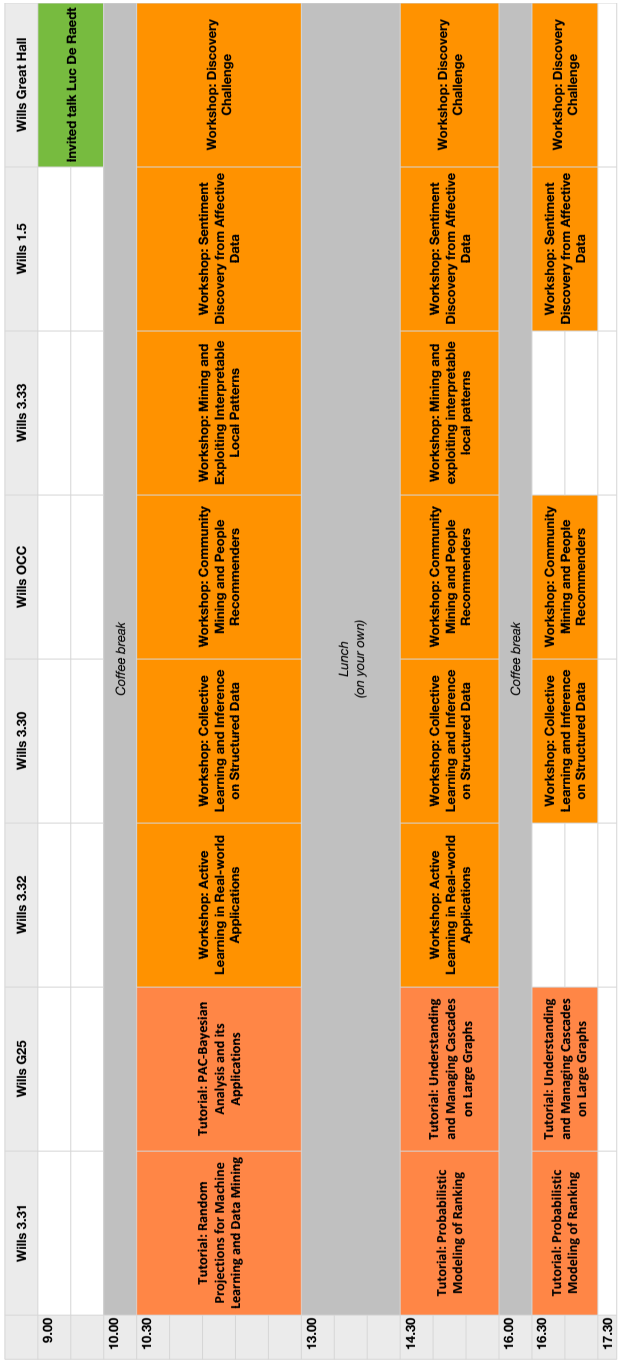
14:10–15:10 **Invited Talk by**
Sebastian Riedel

15:10–16:10 **Invited Talk by**
Christopher Re

16:10–16:35 **Coffee Break**

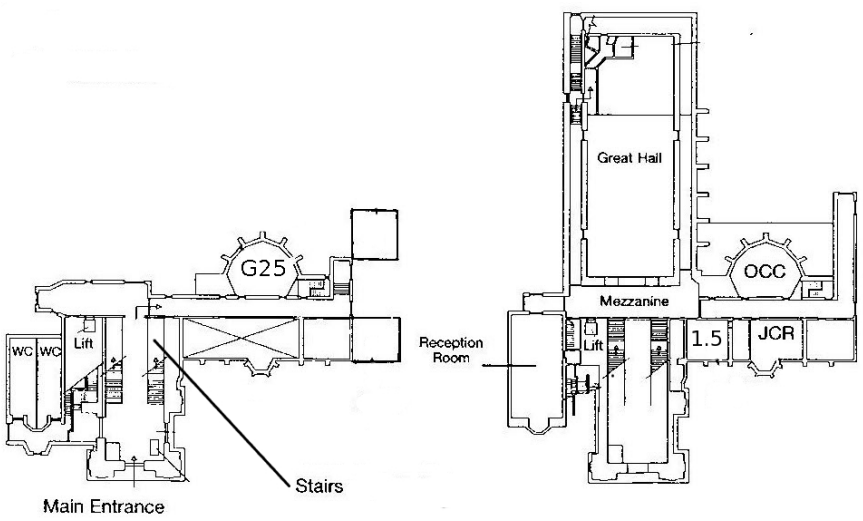
16:35–17:00 **Supervised Blockmodelling**
Leto Peel

17:00–17:30 **Discussions**
Participants



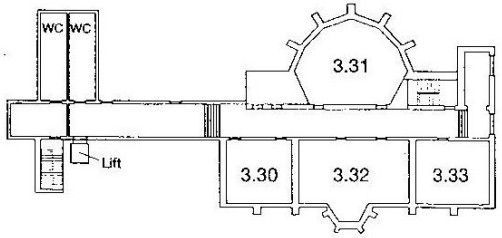
Wills Memorial Building

All Monday and Friday sessions except Monday welcome reception.



Ground Floor

First Floor



Third Floor

Community Mining and People Recommenders

Jaakko Hollmén, Panagiotis Papapetrou and Luiz Augusto Pizzato

10:00–17:00 Room: Wills, OCC

Data mining and knowledge discovery in social networks has advanced significantly over the past several years, due to the availability of a large variety of offline and online social network systems. The focus of COMMPER 2012 is on social networks with special focus on community mining and people recommenders. Community minding involves topics such as the analysis of scientific communities and collaboration networks, including bibliometrics, and the formation of teams. People recommenders focus on the all topics where recommender systems are used to enable connections among users, such systems can be found on all types of social networks such as photo sharing websites, expert search, mentoring systems and online dating.

10:30–10:45 **Welcome Address**
Papapetrou Panagiotis, Hollmén Jaakko and Luiz Augusto Pizzato

10:45–11:45 **Keynote Talk: Entity Selection and Ranking for Data-mining Applications**
Evimaria Terzi

11:45–12:15 **Discussion: Challenges in the Area of Community Mining and People Recommenders**
Panagiotis Papapetrou, Jaakko Hollmén and Luiz Augusto Pizzato

12:15–12:45 **Mining Dynamic Networks: The Importance of Pre-processing on Downstream Analytics**
Sofus Macskassy

12:45–13:15 **Mining the Dynamics of Scientific Publication Networks for Collaboration Recommendation**
Rushed Kanawati

13:15–14:45 **Lunch (not provided)**

14:45–15:15 **Supporting Community Mining and People Recommendations in a Social Internetworking Scenario**
Francesco Buccafurri, Gianluca Lax, Biagio Liberto, Antonino Nocera and Domenico Ursino

15:15–15:45 **Testing People To People Recommender in a Live Environment**
Michael Meisel, Stefan Dahms and Andreas Ittner

15:45–16:15 **Identifying Topical Twitter Communities via User List Aggregation**
Derek Greene, Derek O'Callaghan and Padraig Cunningham

16:15–16:45 **Social Citation: Finding Roles in Social Networks. An Analysis of TV-Series Web Forums**
Nikolay Anokhin, James Lanagan and Julien Velcin

16:45–17:30 **Discussion: Challenges in the Area of Community Mining and People Recommenders**
Panagiotis Papapetrou, Jaakko Hollmén and Luiz Augusto Pizzato

18:00 **Workshop Dinner and Drinks at a Local Pub**
To join us please register your interest with one of the organisers by the end of the lunch break.

Mining and Exploiting Interpretable Local Patterns

Henrik Grosskreutz, Nikos Karacapilidis and Stefan Rüping

10:30–16:00 Room: Wills, 3.33

Local patterns, like itemsets, correlations, contrast sets or subgroups, are valuable nuggets for a variety of applications. Among others, they can been used for classification, regression or outlier detection tasks. One particular characteristic which makes them stand out from other machine learning tools, however, is that (most) local patterns can directly be read and interpreted by end users lacking a profound machine-learning background.

This descriptive nature of local patterns makes them useful as a source of information for decision making. For example, in the analysis of clinical data, understandable models can help the clinician in understanding his data and thus making an informed decision about patient treatment. In addition, understandable knowledge can help domain experts to discuss the analysis results and collaboratively find a good, interesting solution in data-intensive settings to help guide the learner when complex background knowledge prevents the system from finding a good model without further input.

In this workshop, we wish to investigate typical use cases and key requirements for the successful usage of local pattern mining in applications where next to the statistical performance of models, the understandability and interestingness of the models is the key success factor. Here, we are particularly interested in settings where the data to be mined is large and complex, preventing investigations of the data without (semi-)automatic analysis tools.

Key questions to be investigated in the workshop are: which pattern language is adequate both for the representation of local phenomena and for the interpretation by the user; How can the raw set of local patterns be reduced to a representative and manageable subset? What conditions must be satisfied for a pattern to be actionable? How can feedback about the understandability and interestingness of partial models be given back to the system and how can the search be controlled? What is the best way to deal with the (typically exponential) size of the pattern space? How to design scalable algorithms?

10:30–10:45 **Welcome and Introduction**
Stefan Rüping and Nikos Karacapilidis

10:45–11:30 **Invited Talk**
Toon Calders

11:30–11:50 **Sequential Pattern Analysis in a Student Database**
R. Campagni, Donatella Merlini and Renzo Sprugnoli

11:50–12:10 **Density and Non-Grid based Subspace Clustering via Kernel Density Estimation**
Jing Zhang, Lantao Yu, Hanqi Zhu and Gang Sun

12:10–12:30 **Contextual Models for User Interaction on the Web**
Peter Haider, Luca Chliarandini, Ulf Brefeld and Alejandro Jaimes

12:30–13:15 **Invited Talk**
Igor Trajkovski

13:15–14:45 **Lunch (not provided)**

14:45–15:30 **Challenge: Results and Systems Description**
Stefan Rüping

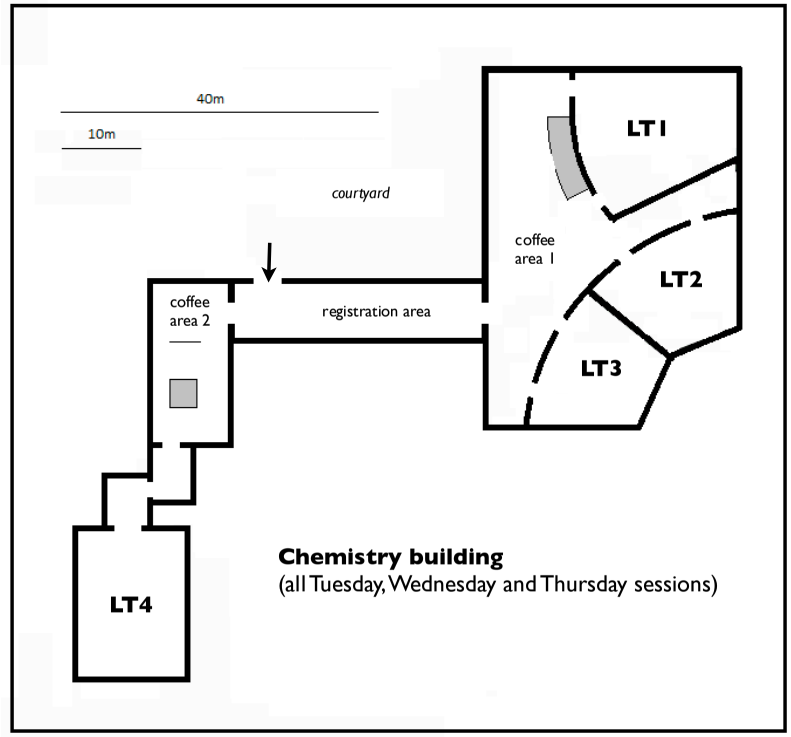
15:30–15:50 **The Dicode Project**
Nikos Karacapilidis

15:50–16:00 **Wrap-up**
All

This workshop is organized in the context of the EU Collaborative Project “DICODE - Mastering Data-Intensive Collaboration and Decision” which is co-funded by the European Commission under the contract FP7-ICT-257184.

Chemistry Building (Chem)

All Tuesday, Wednesday, Thursday sessions except poster sessions and conference dinner.



Welcome

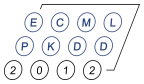
It is our great pleasure to welcome you to ECML-PKDD 2012 in Bristol, UK. We have worked hard – together with a local team and a large number of dedicated chairs who are listed in the conference proceedings – to bring you a high-quality and varied scientific programme. This booklet has been put together to help you find your way through the conference and make the most of your stay in Bristol.

In addition to this programme booklet we have developed two online tools that provide personalised recommendations based on your publication profile on DBLP. One is accessible via a web browser and the other is an iPhone app available from the App Store; surf to `www.ecmlpkdd2012.net/apps/` for the links. These conference planners are brought to you by the PreSift team: Ruthie Hambley, Tom Kelly, Louise Millard, Andrew Pickin and Simon Price. We hope you find them useful – feel free to email `PreSift@cs.bris.ac.uk` with your feedback.

No conference would be possible without a team of dedicated volunteers, and we gratefully acknowledge the help of the following MSc students, PhD students and postdocs:

Reem Al-Otaibi	David Kirchheimer	Andrew Pickin
Qipeng Chen	Akis Kontonasios	Seyed Ali Sadegh Zadeh
Ilias Flaounas	Tom Lansdall-Welfare	Nick Sampson
Noor Hafhizah	Daniel Lewis	Raul Santos-Rodriguez
Osian Haines	Kevin Lloyd	Pranshu Saxena
Joe Heath	Foteini Markatopoulou	Rakhi Singh
Elena Hensinger	Louise Millard	Eirini Spyropoulou
Patricia Jiménez Aguirre	Pablo Munoz Espada	Saatviga Sudhahar
Nanlin Jin	Ricardo Ñanculef	Claudio Taranto
Namita Kalantri	Yizhao Ni	Konstantin Tretjakov
Georgios Kalogiros	Athanasia Notta	Bowen Yan

During the week they can be recognised by their blue T-shirts with the ECML-PKDD 2012 logo (you can pick up your own T-shirt on Friday). Please talk to them if you have any questions, they are there to help.



Peter Flach, Tijl De Bie and Nello Cristianini
`ECMLPKDD2012pcchairs@cs.bris.ac.uk`

Active Learning in Real-world Applications

Laurent Candillier, Max Chevalier and Vincent Lemaire

10:30–16:00 Room: Wills, 3.32

This workshop aims to offer a meeting opportunity for academics and industry-related researchers, belonging to the various communities of Computational Intelligence, Machine Learning, Experimental Design and Data Mining to discuss new areas of active learning, and to bridge the gap between data acquisition or experimentation and model building. How active sampling, incremental learning and data acquisition, can contribute towards the design and modeling of highly intelligent machine learning systems?

Machine learning indicates methods and algorithms which allow a model to learn a behavior thanks to examples. Active learning gathers methods which select examples used to build a training dataset for the predictive model. All the strategies aim to use a set of examples as small as possible and to select the most informative examples.

When designing active learning algorithms for real-world data, some specific issues are raised. The main ones are scalability and practicability. Methods must be able to handle high volumes of data, in spaces of possibly high-dimension, and the process for labeling new examples by an expert must be optimized.

We encouraged papers that describe applications of active learning in real-world. The industrial context, the main difficulties met and the original solution developed, had to be described. An associated challenge (http://www.nomao.com/labs/challenge) has been conjointly organized on a practical application of active learning. The challenge and the results obtained will be presented.

10:30–11:30 **Invited Talk: Active Learning for Discovery in the Laboratory: Characterising Biomolecular Systems**
Chris Lovell

11:30–12:15 **Challenge Presentation: Design and Analysis of the Nomao challenge: Active Learning in the Real-World**
Laurent Candillier and Vincent Lemaire

12:15–12:45 **Challenge winner: Batch-Mode Active Learning by Using Misclassified Data**
Tengyu Sun and Jie Zhou

12:45–13:15 **Position paper: Programmer's Active Learning: A Broader Perspective of Choices for Real-World Classification Tasks that Matter**
George Forman

13:15–14:45 **Lunch (not provided)**

14:45–15:15 **Incorporating Density in Active Learning with Application to Ranking**
Wenbin Cai and Ya Zhang

15:15–15:45 **Active Learning in the Spatial-domain for Landslide Mapping in Remote Sensing Images**
Andre Stumpf, Nicolas Lachiche, Jean-Philippe Malet, Norman Kerle and Anne Puissant

15:45–16:00 **Participants Open Discussion**

Discovery Challenge: Pascal Large Scale Hierarchical Classification Workshop

Ion Androutsopoulos, Thierry Artières, Patrick Gallinari, Eric Gaussier
Aris Kosmopoulos, George Paliouras and Ioannis Partalas

10:30–17:30 Room: Wills, Great Hall

Hierarchies are becoming ever more popular for the organization of documents, particularly on the Web (e.g. Web directories). Along with their widespread use comes the need for automated classification of new documents to the categories in the hierarchy. Research on large-scale classification so far has focused on large numbers of documents and/or large numbers of features, with a limited number of categories. However, hierarchical category systems such as DMOZ or Wikipedia contain several hundreds of thousands categories raising new challenges. Approaching this problem, researchers have either extended existing large-scale classifiers, or have developed new models and methods.

LSHC3 is the third in the series of workshops on large-scale hierarchical classification. The workshop aims at bringing together researchers and practitioners of large-scale category systems and discussing recent trends and solutions.

10:30–11:30 **Invited Talk: Large-scale Learning with Gauge Regularization for Visual Classification**
Zaid Harchaoui

11:30–12:15 **Presentation of the Challenge**
Workshop Organizers

12:15–12:45 **A k-NN Method for Large Scale Hierarchical Text Classification at LSHTC3**
Xiaogang Han, Shaohua Li and Zhiqi Shen

12:45–13:00 **Multi-stage Rocchio Classification for Large-scale Multi-labeled Text data**
Dong-Hyun Lee

13:00–14:30 **Lunch (not provided)**

14:30–15:00 **TTI's System for the LSHTC3 Challenge**
Yutaka Sasaki and Davy Weissenbacher

15:00–15:30 **Ensembles of Sparse Multinomial Classifiers for Scalable Text Classification**
Antti Puurula and Albert Bifet

15:30–16:00 **A Fast Dendrogram Refinement Approach for Unsupervised Expansion of Hierarchies**
Ricardo Marcondes Marcacini, Everton A. Cherman, Jean Metz and Solange O. Rezende

16:00–16:30 **Coffee break**

16:30–17:00 **Regularization Framework for Large Scale Hierarchical Classification**
Siddharth Gopal, Yiming Yang and Alexandru Niculescu-Mizil

17:00–17:30 **Discussion**

Instructions for Participants

General Information

- For general enquiries or projector/computer/internet problems please see the registration desk.
- Internet access:
  - Your badge has a login name, password and instructions for the visitornet wireless system.
  - If you already have an eduroam account you can use that.
  - Please see the registration desk for an account to use the 12 terminals in chemistry's coffee area 2.
- University security: 0117 928-7848.
- Emergency services (Police, Fire and Rescue Service, Emergency Medical Service): 999.
- For help making a phone call you can reach the UK operator on 100 and the international one on 155.

Instructions for Speakers

- Every room will have a digital projector and computer.
- Speakers can bring their own laptop.
- A volunteer will be on hand to help in every room.
- Please arrive 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.
- If a speaker is missing from your session please do not alter the schedule of talks, so people do not miss them.
- The time allocated to each speaker, including time to set up and time for questions, is:
  - research track: 25 minutes.
  - demo track: 10 minutes.
  - industry and nectar tracks: 50 minutes.
- If the session chair is not present we ask the last speaker of the session to act as chair.

Instructions for Session Chairs

- Please arrive 10 minutes before the session starts to check the equipment works.
- Please stick to the schedule. If a speaker is absent please announce a short break.
- Please moderate questions.
- Please do not start sessions or talks early as attendees may be moving between locations.

Instructions for Poster Presenters

- There will be two poster sessions: one on Tuesday evening (for talks in sessions on Tuesday or on Wednesday morning); and another on Thursday evening (for talks in sessions on Thursday or on Wednesday afternoon).
- Both poster sessions will take place in the Auditorium of the Victoria Rooms from 7pm onwards, with buffet-style food and drink served from about 7:30.
- Presenters should arrive in advance to set up their poster from 6pm.
- Each presenter will be provided with a 2m × 1m poster board and drawing pins.
- Poster boards will be grouped according to sessions; maps and labels will be provided so you can easily find the poster boards allocated to your session; and a helper will be on hand to assist you.
- Presenters should remove their poster at the end of the session.

Instructions for Software Demos

- The Demo Session will be held together with the Tuesday poster session in the Auditorium of the Victoria Rooms from 7pm onwards, with buffet-style food and drink served from about 7:30.
- Presenters should arrive in advance to set up their demos from 6pm.
- Each demo will be provided with a desk, two chairs, and a (UK-style) power socket.
- Wireless internet will be available.
- Demonstration desks will be set up next to the stage area in the auditorium.
- Presenters should remove their demo at the end of the session.

Advanced Topics in Data Stream Mining

Albert Bifet, Joao Gama, Richard Gavalda, Georg Krempl, Mykola Pechenizkiy, Bernhard Pfahringer, Myra Spiliopoulou and Indrė Žliobaitė

09:00–12:00 Room: Wills, 3.31

Nowadays, the quantity of data that is created every two days is estimated to be 5 exabytes. This amount of data is similar to the amount of data created from the dawn of time up until 2003. Moreover, it was estimated that 2007 was the first year in which it was not possible to store all the data that we are producing. This massive amount of real time streaming data opens new challenging discovery tasks. Some of them are already addressed with mature algorithms, while new challenges emerge, including learning on not one but multiple streams. This tutorial has two parts. The first part gives an introduction to recent advances in algorithmic techniques and tools to cope with challenges on stream mining. The second part discusses state of the art research on mining multiple streams – distributed streams and interdependent relational streams.

Concept drift plays a central role in this tutorial. In the first part, we address it in the context of conventional one-stream mining to set the scene. In the second part, we recapitulate on it after introducing multiple-stream mining, and we also consider machine learning methods that are appropriate for incremental data and slow streams.

The first part, 'Mining One Stream', will be presented by Albert Bifet, Ricardo Gavalda, Mykola Pechenizkiy, Bernhard Pfahringer, and Indrė Žliobaitė.

The second part, 'Mining Multiple Streams' will be presented by Joao Gama, Myra Spiliopoulou, and Georg Krempl.

Advanced Topics in Ensemble Learning

Ioannis Partalas, Alberto Suarez, Daniel Hernandez Lobato, Grigorios Tsoumakas and Gonzalo Martinez

09:00–12:00 Room: Wills, G25

Ensemble methods are widely used by the machine learning community because they lead to improvements in accuracy and robustness in many prediction problems of practical interest. Furthermore, they offer appealing solutions to several interesting learning problems, such as dealing with small or large datasets, performing data fusion and modeling concept drift. Notwithstanding, using ensemble methods also poses some challenges. Specifically, in many problems very large ensembles need to be generated to achieve good generalization performance. Similarly, the ensemble prediction is typically generated by querying all the classifiers contained in the ensemble, a process that can be computationally expensive. In this tutorial we give a brief introduction to ensemble methods for machine learning and describe ensemble pruning techniques to deal with the shortcomings described. We also introduce advanced research topics of current interest, such as the problem of determining the optimal ensemble size or techniques to make ensembles scale to large learning problems.

Probabilistic Modeling of Ranking

Jose A. Lozano and Ekhiñe Irurozki

14:30–17:00 Room: Wills, 3.31

Rankings and permutations have become, nowadays, ubiquitous. They appear in numerous areas of computer systems: information retrieval, recommender systems, identity tracking or chemical compound classification, etc. Dealing with rankings, and particularly with rankings of many objects is a complex computational task as the number of permutations of n objects scales factorially in n. Recently a number of approaches have come to the machine learning arena to address this kind of data. Most of these approaches are based on the building of a probability distribution over the space of rankings. However, given the complexity of storing, learning and making inference on this kind of models, different simplifying assumptions have been considered: the use of parametric models, models based on low-order statistics, models based on kernels and the definition and use of notions of independence and conditional independence in the space of permutations. In this tutorial we will review the literature on the topic, explaining the different approaches in detail that have emerged in the literature, putting them in relation with other non-probabilistic ranking models and giving a collection of open problems in the area. In addition we will present the most relevant applications in the field as well as the most common benchmark datasets and software.

Understanding and Managing Cascades on Large Graphs

B. Aditya Prakash and Christos Faloutsos

14:30–17:00 Room: Wills, G25

How do contagions spread in population networks? Which group should we market to, for maximizing product penetration? Will a given YouTube video go viral? Who are the best people to vaccinate? What happens when two products compete? The objective of this tutorial is to provide an intuitive and concise overview of most important theoretical results and algorithms to help us understand and manipulate such propagation-style processes on large networks. The tutorial will contain three parts: (a) Theoretical results on the behavior of fundamental models; (b) Scalable Algorithms for changing the behavior of these processes e.g., for immunization, marketing etc.; and (c) Empirical Studies of diffusion on blogs and on-line websites like Twitter. We finally conclude with future research directions. The problems we focus on are central in surprisingly diverse areas: from computer science and engineering, epidemiology and public health, product marketing to information dissemination. Our emphasis is on intuition behind each topic, and guidelines for the practitioner.

PAC-Bayesian Analysis and Its Applications

Yevgeny Seldin, François Laviolette and John Shawe-Taylor

10:30–13:00 Room: Wills, G25

PAC-Bayesian analysis is a basic and very general tool for data-dependent analysis in machine learning. By now, it has been applied in such diverse areas as supervised learning, unsupervised learning, and reinforcement learning, leading to state-of-the-art algorithms and accompanying generalization bounds. PAC-Bayesian analysis, in a sense, takes the best out of Bayesian methods and PAC learning and puts it together: (1) it provides an easy way to exploit prior knowledge (like Bayesian methods); (2) it provides strict and explicit generalization guarantees (like VC theory); and (3) it is data-dependent and provides an easy and strict way of exploiting benign conditions (like Rademacher complexities). In addition, PAC-Bayesian bounds directly lead to efficient learning algorithms.

We will start with a general introduction to PAC-Bayesian analysis, which should be accessible to an average student, who is familiar with machine learning at the basic level. Then, we will survey multiple forms of PAC-Bayesian bounds and their numerous applications in different fields (including supervised and unsupervised learning, finite and continuous domains, and the very recent extension to martingales and reinforcement learning). Some of these applications will be explained in more details, while others will be surveyed at a high level. We will also describe the relations and distinctions between PAC-Bayesian analysis, Bayesian learning, VC theory, and Rademacher complexities. We will discuss the role, value, and shortcomings of frequentist bounds that are inspired by Bayesian analysis.

This tutorial builds upon a related tutorial given at ICML-2012. While there will naturally be certain overlap in the material, certain aspects will be presented differently and a stronger stress on applications will be provided.

Random Projections for Machine Learning and Data Mining: Theory and Applications

Ata Kaban and Bob Durrant

10:30–13:00 Room: Wills, 3.31

Random projections have been used as a dimensionality reduction technique for large data since the appearance of Arriaga and Vempala's seminal FOCS 1999 paper, and they continue to find applications both within the field of Machine Learning and elsewhere. Starting with some motivating examples from Machine Learning and Data Mining, this tutorial will review some key theoretical properties of random projections, and the practical applications this theory has inspired. In particular, we will cover the Johnson-Lindenstrauss lemma which gives conditions under which random projections approximately preserve geometric properties of data and give related applications, discuss the field of Compressed Sensing from which one can derive guarantees for techniques working with sparse data which we illustrate in the setting of SVM, discuss more recent theory giving guarantees for linear classifiers and regressors working with non-sparse data, and finally we take a look at random projections as regularizers.

Decomposing Binary Matrices: Where Linear Algebra Meets Combinatorial Data Mining

Pauli Miettinen

13:30–16:00 Room: Wills, 3.31

The tutorial studies the connection between matrix factorization methods and data mining on binary data (e.g. pattern set mining). On one hand, it shows how many data mining methods can be modeled as discrete matrix factorizations. This change of conceptual framework provides new tools and concepts that help the data miner to generalize the algorithms, develop new algorithms, and analyse the algorithms in new ways. On the other hand, it shows how discretization of matrix factorizations provides new tools and methods for traditional matrix factorization applications.

This tutorial considers the problems of representing a binary matrix as a product of two binary matrices using three different algebras: the normal algebra (where 1+1=2), the modulo-2 algebra (where 1+1=0), and the Boolean algebra (where 1+1=1). In all of these cases, the requirement of binary factor matrices renders most of our normal matrix factorization tools less useful, calling for more combinatorial approaches. The combinatorial nature of the problems further increase when one moves from normal algebra to modulo-2 algebra to Boolean algebra. Nevertheless, all these problems show traces of both linear algebra and combinatorics and solving them requires techniques from both fields. A core idea of this tutorial is to show how ideas and concepts from linear algebra can be translated to combinatorial settings (and vice versa) and how that can be used to solve the factorization problems (and therefore other problems modeled as factorization problems).

The tutorial covers both theory and algorithms for these factorizations. The theory-part concentrates on building the connections (and pointing non-obvious differences) between the discrete and continuous factorizations. Central to this discussion is the concept of matrix rank, and how ranks under different algebras relate to each other. This discussion is also indented to give more intuition on how the different binary factorizations behave. Also the computational complexity of the factorizations is studied at this part, as is the complexity of an important sub-problem: finding the least-error projections. Throughout this discussion the concepts discussed are also related to concepts in pattern mining (e.g. the Boolean rank of a matrix is the minimum number of tiles needed to cover all 1s in the data).

The second part of the tutorial concentrates on algorithms for the problems. More details will be given on few algorithms, but the main concentration is on explaining the ideas and intuitions behind – and similarities and differences between – the algorithms. In addition to the algorithms for the standard problems, the discussion will include methods for selecting the rank and special results with sparse matrices.

Applications of these factorizations are highlighted when appropriate, but the focus of this tutorial is not on applications. It is assumed that the participants generally know applications for discrete data mining methods or matrix factorizations.

Mining Deep Web Repositories

G. Das and N. Zhang

13:30–16:00 Room: Wills, G25

With the proliferation of online repositories (e.g., databases or document corpora) hidden behind proprietary web interfaces, e.g., keyword-/form-based search and hierarchical/graph-based browsing interfaces, efficient ways of enabling machine learning and data mining tasks over contents in such hidden repositories are of increasing importance. There are two key challenges: one on the proper understanding of interfaces, and the other on learning/mining over a properly understood interface. There are three ways to enable efficient machine learning and data mining over deep web data – (1) crawling the deep web repository before applying conventional mining techniques, (2) sampling the deep web repository before learning/mining the retrieved samples, at the expense of additional error introduces by sampling, and (3) estimating aggregates over slices of data in the deep web repository, and then using the estimated aggregates to support machine learning or data mining tasks. In this tutorial, we focus on the fundamental developments in the field, including web interface understanding, crawling, sampling, and aggregate estimation over web repositories with various types of interfaces and containing structured or unstructured data. We also discuss the potential changes required for machine learning and data mining algorithms should one choose to use the second and third methods described above. Our goal is to encourage the audience to initiate their own research in these exciting areas.

Monday Workshop

Mining Ubiquitous and Social Environments

Martin Atzmueller and Andreas Hotho

09:00–16:00 Room: Wills, OCC

The goal of this workshop is to promote an interdisciplinary forum for researchers working in the fields of ubiquitous computing, social web, Web 2.0, and social networks which are interested in utilizing data mining in a ubiquitous setting. The workshop features contributions adopting state-of-the-art mining algorithms on ubiquitous social data. We want to accelerate the process of identifying the power of advanced data mining operating on data collected in ubiquitous and social environments, as well as the process of advancing data mining through lessons learned in analyzing these new data.

09:00–09:15 **Welcome**

09:15–10:30 **Extracting Social and Community Intelligence by Mining Large Scale Digital Footprints**
Daqing Zhang

10:30–11:00 **Coffee Break**

11:00–11:35 **Guiding User Groupings - Learning and Combining Classification for Itemset Structuring**
Mathias Verbeke, Ilija Subasic and Bettina Berendt

11:35–11:55 **A Fast and Simple Method for Profiling a Population of Twitter Users**
Hideki Asoh, Kazushi Ikeda and Chihiro Ono

11:55–12:15 **An Analysis of Interactions Within and Between Extreme Right Communities in Social Media**
Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy and Padraig Cunningham

12:15–13:45 **Lunch (not provided)**

13:45–14:05 **Geographically-Aware Mining of AIS Messages to Track Ship Itineraries**
Annalisa Appice, Donato Malerba and Antonietta Lanza

14:05–14:40 **Robust Language Identification in Short, Noisy Texts: Improvements to LIGA**
John Vogel and David Tresner-Kirsch

14:40–15:15 **Identifying Time-Respecting Subgraphs in Temporal Networks**
Ursula Redmond, Martin Harrigan and Padraig Cunningham

15:15–16:00 **Discussion + Closing**

Friday Invited Talk

Declarative Modeling for Machine Learning and Data Mining

Luc De Raedt

University of Leuven, Belgium
luc.deraedt@cs.kuleuven.be
http://people.cs.kuleuven.be/~luc.deraedt/

Friday 09:00–10:00 Room: Wills, Great Hall

Abstract

Despite the popularity of machine learning and data mining today, it remains challenging to develop applications and software that incorporates machine learning or data mining techniques. This is because machine learning and data mining have focussed on developing high-performance algorithms for solving particular tasks rather than on developing general principles and techniques. I propose to alleviate these problems by applying the constraint programming methodology to machine learning and data mining and to specify machine learning and data mining problems as constraint satisfaction and optimization problems. What is essential is that the user be provided with a way to declaratively specify what the machine learning or data mining problem is rather than having to outline how that solution needs to be computed. This corresponds to a model + solver-based approach to machine learning and data mining, in which the user specifies the problem in a high level modeling language and the system automatically transforms such models into a format that can be used by a solver to efficiently generate a solution. This should be much easier for the user than having to implement or adapt an algorithm that computes a particular solution to a specific problem. Throughout the talk, I shall use illustrations from our work on constraint programming for itemset mining and probabilistic programming.

Bio

Luc De Raedt is a full professor (of research) at the University of Leuven (KU Leuven) in the Department of Computer Science and a former chair of Machine Learning at the Albert-Ludwigs-University in Freiburg. Luc De Raedt has been working in the areas of artificial intelligence and computer science, especially on computational logic, machine learning and data mining, probabilistic reasoning and constraint programming and their applications in bio- and chemoinformatics, vision and robotics, natural language processing, and engineering. His work has typically crossed boundaries between different research areas, often working towards an integration of their principles. He is well-known for his early work on inductive logic programming (combining logic with learning). Since 2000, he has been working towards a further integration of logical and relational learning with probabilistic reasoning (statistical relational learning and probabilistic programming) and on inductive querying in databases. During the last three years he has been fascinated by the possibility of combining constraint programming principles with data mining and machine learning. He is currently coordinating a European IST FET project in this area (ICON – Inductive Constraint Programming) and is the program chair of the 20th European Conference on Artificial Intelligence (Montpellier, 2012). He was a program co-chair of ICML 2005 and ECML/PKDD 2001.

Room: Chem, LT3

Thu3C: Reinforcement Learning and Planning II — Chair: Myra Spiliopoulou

16:10–16:35 **Bootstrapping Monte Carlo Tree Search with an Imperfect Heuristic**
Truong-Huy Dinh Nguyen, Wee-Sun Lee and Tze-Yun Leong

We consider the problem of using a heuristic policy to improve the value approximation by the Upper Confidence Bound applied in Trees (UCT) algorithm in non-adversarial settings such as planning with large-state space Markov Decision Processes. Current improvements to UCT focus on either changing the action selection formula at the internal nodes or the rollout policy at the leaf nodes of the search tree. In this work, we propose to add an auxiliary arm to each of the internal nodes, and always use the heuristic policy to roll out simulations at the auxiliary arms. The method aims to get fast convergence to optimal values at states where the heuristic policy is optimal, while retaining similar approximation as the original UCT at other states. We show that bootstrapping with the proposed method in the new algorithm, UCT-Aux, performs better compared to the original UCT algorithm and its variants in two benchmark experiment settings. We also examine conditions under which UCT-Aux works well.

16:35–17:00 **Fast Reinforcement Learning with Large Action Sets Using Error-Correcting Output Codes for MDP Factorization**
Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux and Patrick Gallinari

The use of Reinforcement Learning in real-world scenarios is strongly limited by issues of scale. Most RL learning algorithms are unable to deal with problems composed of hundreds or sometimes even dozens of possible actions, and therefore cannot be applied to many real-world problems. We consider the RL problem in the supervised classification framework where the optimal policy is obtained through a multiclass classifier, the set of classes being the set of actions of the problem. We introduce error-correcting output codes (ECOCs) in this setting and propose two new methods for reducing complexity when using rollouts-based approaches. The first method consists in using an ECOC-based classifier as the multiclass classifier, reducing the learning complexity from $\mathcal{O}(A^2)$ to $\mathcal{O}(A\log(A))$. We then propose a novel method that profits from the ECOC's coding dictionary to split the initial MDP into $\mathcal{O}(\log(A))$ separate two-action MDPs. This second method reduces learning complexity even further, from $\mathcal{O}(A^2)$ to $\mathcal{O}(\log(A))$, thus rendering problems with large action sets tractable. We finish by experimentally demonstrating the advantages of our approach on a set of benchmark problems, both in speed and performance.

17:00–17:25 **Learning Policies For Battery Usage Optimization in Electric Vehicles**
Stefano Ermon, Yexiang Xue, Carla Gomes and Bart Selman

The high cost, limited capacity, and long recharge time of batteries pose a number of obstacles for the widespread adoption of electric vehicles. Multi-battery systems that combine a standard battery with supercapacitors are currently one of the most promising ways to increase battery lifespan and reduce operating costs. However, their performance crucially depends on how they are designed and operated.

In this paper, we formalize the problem of optimizing real-time energy management of multi-battery systems as a stochastic planning problem, and we propose a novel solution based on a combination of optimization, machine learning and data-mining techniques. We evaluate the performance of our intelligent energy management system on various large datasets of commuter trips crowdsourced in the United States. We show that our policy significantly outperforms the leading algorithms that were previously proposed as part of an open algorithmic challenge.

17:25–17:50 **Structured Apprenticeship Learning**
Abdeslam Boularias, Oliver Krömer and Jan Peters

We propose a graph-based algorithm for apprenticeship learning when the reward features are noisy. Previous apprenticeship learning techniques learn a reward function by using only local state features. This can be a limitation in practice, as often some features are misspecified or subject to measurement noise. Our graphical framework, inspired from the work on Markov Random Fields, allows to alleviate this problem by propagating information between states, and rewarding policies that choose similar actions in adjacent states. We demonstrate the advantage of the proposed approach on grid-world navigation problems, and on the problem of teaching a robot to grasp novel objects in simulation.

Instant Interactive Data Mining

Jilles Vreeken, Matthijs van Leeuwen, Siegfried Nijssen, Nikolaj Tatti, Anton Dries and Bart Goethals

09:00–16:00 Room: Wills, 3.30

The goal of the Instant and Interactive Data Mining workshop (IID) is to address the development of data mining techniques that allow users to interactively explore their data, receiving near-instant updates to every requested refinement. While both *Instant* and *Stream* mining start from different perspectives and operate under different constraints, there is a significant overlap in techniques and developments in both settings. Therefore, we aim to bring together researchers interested in instant and adaptive methods, whether for use in interactive systems or in the processing of large streams of evolving data.

09:00–09:05 **Welcome address**
Jilles Vreeken

09:05–10:05 **Keynote: Real-World Interactive Machine Learning of Customer Support Logs at Hewlett-Packard**
George Forman

10:05–10:30 **A Case of Visual and Interactive Data Analysis: Geospatial Redescription Mining**
Esther Galbrun and Pauli Miettinen

10:30–11:00 **Coffee Break**

11:00–11:20 **Online Estimation of Discrete Densities using Classifier Chains**
Michael Geilke, Eibe Frank and Stefan Kramer

11:20–11:40 **iST-MRF: Interactive Spatio-Temporal Probabilistic Models for Sensor Networks**
Nico Piatkowski

11:40–12:00 **From Block-based Ensembles to Online Learners In Changing Data Streams: If- and How-To**
Dariusz Brzezinski and Jerzy Stefanowski

12:00–13:30 **Lunch (not provided)**

13:30–14:30 **Keynote: Real Data Mining for Real Users: Instant, Interactive – A Dream?**
Michael Berthold

14:30–14:55 **Towards Exploratory Search of Scientific Information**
Dorota Glowacka, Ksenia Konyushkova, Tuukka Ruotsalo and Samuel Kaski

14:55–15:20 **Towards Real-Time Machine Learning**
Andreas Hapfelmeier, Christian Mertes, Jana Schmidt and Stefan Kramer

15:20–15:45 **Instant Selection of High Contrast Projections in Multi-dimensional Data Streams**
Andrei Vanea, Emmanuel Müller, Fabian Keller and Klemens Böhm

15:45–16:00 **Discussion & Closing**
Matthijs van Leeuwen

New Frontiers in Mining Complex Patterns

Annalisa Appice, Michelangelo Ceci, Corrado Loglisci, Giuseppe Manco, Elio Masciari and Zbigniew W. Ras

09:00–16:00 Room: Wills, 3.32

This workshop is concerned with emerging technologies and applications where complex patterns in expressive languages are extracted from new prominent data sources like blogs, event or log data, biological data, spatio-temporal data, social networks, mobility data, sensor data and streams, and so on. Advanced techniques, which preserve the informative richness of data and allow us to efficiently and efficaciously identify complex information units present in such data, will be presented.

09:00–09:10 **WELCOME**

09:10–10:00 **Invited Talk: Advances in Semantic Data Mining**
Nada Lavrač

THEORY

10:00–10:15 **Towards the Definition of Learning Systems with Configurable Operators and Heuristics**
Fernando Martínez-Plumed, Cèsar Ferri, José Hernández-Orallo and María José Ramírez-Quintana

10:15–10:30 **Reducing Examples in Relational Learning with Bounded-Treewidth Hypotheses**
Ondřej Kuželka, Andrea Szabóová and Filip Železný

10:30–11:00 **Coffee Break**

MISC COMPLEX DATA

11:00–11:15 **Mining Complex Event Patterns in Computer Networks**
Dietmar Seipel, Philipp Neubeck, Stefan Köhler and Martin Atzmueller

11:15–11:30 **Learning with Aggregation and Correlation in the Presence of Large Fluctuations**
Eric Paquet, Herna Lydia Viktor and Hongyu Guo

11:30–11:45 **Machine Learning as an Objective Approach to Understanding Musical Origin**
Claire Q and Ross King

11:45–12:00 **Object-driven Action Rules and their Application to Hypernasality Treatment**
Ayman Hajja, Alicja Wieczorkowska, Zbigniew Ras and Ryszard Gubrynowicz

12:00–12:10 **Pre-filtering Features in Random Forests for Microarray Data Classification**
Nicoletta Dessì, Gabriele Milia and Barbara Pes

12:10–13:25 **Lunch (not provided)**

TRAJECTORIES AND TIME SERIES

13:25–13:40 **MatArcs: An Exploratory Data Analysis of Recurring Patterns in Multivariate Time Series**
Julia Gaebler, Stephan Spiegel and Sahin Albayrak

13:40–13:55 **Effectively Grouping Trajectory Streams**
Gianni Costa, Giuseppe Manco and Elio Masciari

13:55–14:10 **Healthcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets**
Elias Egho, Dino Ienco, Nicolas Jay, Amedeo Napoli, Pascal Poncelet, Catherine Quantin, Chedy Raissi and Maguelonne Teisseire

14:10–14:25 **Graph-Based Approaches to Clustering Network-Constrained Trajectory Data**
Mohamed Khalil El Mahrsi and Fabrice Rossi

14:25–14:30 **Break**

Room: Chem, LT2

Thu3B: Natural Language Processing

Chair: Eric Gaussier

16:10–16:35 **Collective Information Extraction with Context-Specific Consistencies**
Peter Kluegl, Martin Toepfer, Florian Lemmerich, Andreas Hotho and Frank Puppe

Conditional Random Fields (CRFs) have been widely used for information extraction from free texts as well as from semi-structured documents. Interesting entities in semi-structured domains are often consistently structured within a certain context or document. However, their actual compositions vary and are possibly inconsistent among different contexts. We present two collective information extraction approaches based on CRFs for exploiting these context-specific consistencies. The first approach extends linear-chain CRFs by additional factors specified by a classifier, which learns such consistencies during inference. In a second extended approach, we propose a variant of skip-chain CRFs, which enables the model to transfer long-range evidence about the consistency of the entities. The practical relevance of the presented work for real-world information extraction systems is highlighted in an empirical study. Both approaches achieve a considerable error reduction.

16:35–17:00 **Supervised Learning of Semantic Relatedness**
Ran El-Yaniv and David Yanay

We propose and study a novel supervised approach to learning statistical semantic relatedness models from subjectively annotated training examples. The proposed semantic model consists of parameterized co-occurrence statistics associated with textual units of a large background knowledge corpus. We present an efficient algorithm for learning such semantic models from a training sample of relatedness preferences. Our method is corpus independent and can essentially rely on any sufficiently large (unstructured) collection of coherent texts. Moreover, the approach facilitates the fitting of semantic models for specific users or groups of users. We present the results of extensive range of experiments from small to large scale, indicating that the proposed method is effective and competitive with the state-of-the-art.

17:00–17:25 **Unsupervised Bayesian Part of Speech Inference with Particle Gibbs**
Gregory Dubbin and Phil Blunsom

As linguistic models incorporate more subtle nuances of language and its structure, standard inference techniques can fall behind. These models are often tightly coupled such that they defy clever dynamic programming tricks. Here we demonstrate that Sequential Monte Carlo approaches, i.e. particle filters, are well suited to approximating such models. We implement two particle filters, which jointly sample either sentences or word types, and incorporate them into a Particle Gibbs sampler for Bayesian inference of syntactic part-of-speech categories. We analyze the behavior of the samplers and compare them to an exact block sentence sampler, a local sampler, and an existing heuristic word type sampler. We also explore the benefits of mixing Particle Gibbs and standard samplers.

17:25–17:50 **WikiSent: Weakly Supervised Sentiment Analysis through Extractive Summarization with Wikipedia**
Subhabrata Mukherjee and Pushpak Bhattacharyya

This paper describes a weakly supervised system for sentiment analysis in the movie review domain. The objective is to classify a movie review into a polarity class, positive or negative, based on those sentences bearing opinion on the movie alone, leaving out other irrelevant text. Wikipedia incorporates the world knowledge of movie-specific features in the system which is used to obtain an extractive summary of the review, consisting of the reviewerâĂŹs opinions about the specific aspects of the movie. This filters out the concepts which are irrelevant or objective with respect to the given movie. The proposed system, WikiSent, does not require any labeled data for training. It achieves a better or comparable accuracy to the existing semi-supervised and unsupervised systems in the domain, on the same dataset. We also perform a general movie review trend analysis using WikiSent.

Room: Chem, LT1

Thu3A: Large-Scale, Distributed and Parallel Mining and Learning II

Chair: Sofus A. Macskassy

16:10–16:35 **CC-MR – Finding Connected Components in Huge Graphs with MapReduce**
Thomas Seidl, Brigitte Boden and Sergej Fries

The detection of connected components in graphs is a well-known problem arising in a large number of applications including data mining, analysis of social networks, image analysis and a lot of other related problems. In spite of the existing very efficient serial algorithms, this problem remains a subject of research due to increasing data amounts produced by modern information systems which cannot be handled by single workstations. Only highly parallelized approaches on multi-core-servers or computer clusters are able to deal with these large-scale data sets. In this work we present a solution for this problem for distributed memory architectures, and provide an implementation for the well-known MapReduce framework developed by Google. Our algorithm CC-MR significantly outperforms the existing approaches for the MapReduce framework in terms of the number of necessary iterations, communication costs and execution runtime, as we show in our experimental evaluation on synthetic and real-world data. Furthermore, we present a technique for accelerating our implementation for datasets with very heterogeneous component sizes as they often appear in real data sets.

16:35–17:00 **Fast Near Neighbor Search in High-Dimensional Binary Data**
Anshumali Shrivastava and Ping Li

Numerous applications in search, databases, machine learning, and computer vision, can benefit from efficient algorithms for near neighbor search. This paper proposes a simple framework for *fast near neighbor search in high-dimensional binary data*, which are common in practice (e.g., text). We develop a very simple and effective strategy for sub-linear time near neighbor search, by creating hash tables directly using the bits generated by b-bit minwise hashing. The advantages of our method are demonstrated through thorough comparisons with two strong baselines: *spectral hashing* and *sign (1-bit) random projections*.

17:00–17:25 **Fully Sparse Topic Models**
Khoat Than and Tu Bao Ho

In this paper, we propose Fully Sparse Topic Model (FSTM) for modeling large collections of documents. Three key properties of the model are: (1) the inference algorithm converges in linear time, (2) learning of topics is simply a multiplication of two sparse matrices, (3) it provides a principled way to directly trade off sparsity of solutions against inference quality and running time. These properties enable us to speedily learn sparse topics and to infer sparse latent representations of documents, and help significantly save memory for storage. We show that inference in FSTM is actually MAP inference with an implicit prior. Extensive experiments show that FSTM can perform substantially better than various existing topic models by different performance measures. Finally, our parallel implementation can handily learn thousands of topics from large corpora with millions of terms.

17:25–17:50 **Massively Parallel Feature Selection: An Approach Based on Variance Preservation**
Zheng Zhao, James Cox, David Duling and Warren Sarle

Advances in computer technologies have enabled corporations to accumulate data at an unprecedented speed. Large-scale business data might contain billions of observations and thousands of features, which easily brings their scale to the level of terabytes. Most traditional feature selection algorithms are designed for a centralized computing architecture. Their usability significantly deteriorates when data size exceeds hundreds of gigabytes. High-performance distributed computing frameworks and protocols, such as the Message Passing Interface (MPI) and MapReduce, have been proposed to facilitate software development on grid infrastructures, enabling analysts to process large-scale problems efficiently. This paper presents a novel large-scale feature selection algorithm that is based on variance analysis. The algorithm selects features by evaluating their abilities to explain data variance. It supports both supervised and unsupervised feature selection and can be readily implemented in most distributed computing environments. The algorithm was developed as a SAS High-Performance Analytics procedure, which can read data in distributed form and perform parallel feature selection in both symmetric multiprocessing mode and massively parallel processing mode. Experimental results demonstrated the superior performance of the proposed method for large scale feature selection.

GRAPHS AND NETWORKS

14:30–14:45 **Finding the Most Descriptive Substructures in Graphs with Numeric Labels**
Michael Davis, Weiru Liu and Paul Miller

14:45–15:00 **Learning in Probabilistic Graphs exploiting Language-Constrained Patterns**
Claudio Taranto, Nicola Di Mauro and Floriana Esposito

15:00–15:15 **Improving Robustness and Flexibility of Concept Taxonomy Learning from Text**
Fabio Leuzzi, Stefano Ferilli, Claudio Taranto and Fulvio Rotella

15:15–15:30 **Discovering Evolution Chains for Link Mining in Dynamic Networks**
Corrado Loglisci, Michelangelo Ceci and Donato Malerba

15:30–15:45 **SISO: a conceptual framework for the construction of "stereotypical maps" in a Social Internetworking**
Scenario Francesco Buccafurri, Gianluca Lax, Antonino Nocera and Domenico Ursino

15:45–15:55 **Context-Aware Prediction on Business Process Executions**
Francesco Folino, Massimo Guarascio and Luigi Pontieri

15:55–16:00 **FINAL REMARKS**

The Silver Lining: Learning from Unexpected Results

Joaquin Vanschoren and Wouter Duivesteijn

09:30–16:00 Room: Wills, 3.33

This workshop is dedicated to the proposition that insight often begins with unexpected results. Successful methods do not simply fall from the sky: they are discovered based on clues gathered by trying several ideas, learning from surprising results, and building an understanding of what works, what does not, and why. Unexpected results chart the boundaries of our knowledge: they identify errors, reveal false assumptions, and force us to dig deeper.

This process is rarely mentioned in the machine learning and data mining discourse, meaning that such insight is essentially lost. Ironically, while we have long understood that learning from only positive results is substantially harder than learning from both positives and negatives, there exists a publication bias that favours (incremental) successes over novel discoveries of why some ideas, while intuitive and plausible, do not work.

Good science consists of carefully designed experiments, systematic procedures, and honest evaluations, even if not all results are positive. As a scientific area where empirical methods dominate, it is a given that people try many ideas and obtain surprising results in the experimental stage. We can learn a lot if we analyze scientifically why an intuitive and plausible experiment did not work as expected.

Just as every cloud has a silver lining, these unexpected results define the actual boundaries of our field: they highlight what we do not yet understand, and often point to interesting and open problems that ought to be explored. With this workshop, we want to give a voice to those unexpected results that deserve wider dissemination.

09:30–10:30 **Opening Keynote: Unexpected Results in Monte-Carlo Tree Search**
Olivier Teytaud (Université Paris-Sud)

10:30–11:00 **Coffee Break**

11:00–11:30 **Generation of an Empirical Theory for Positive-Versus-Negative Ensemble Classification**
Patricia Lutu (University of Pretoria)

11:30–12:00 **How to Make the Most of Result Evaluation?**
Ana Costa e Silva (University of Porto)

12:00–13:30 **Lunch (not provided)**

13:30–14:30 **Keynote: On the Search for and Appreciation of Unexpected Results in Data Mining Research (or: Science – we Might be Doing it Wrong)**
Albrecht Zimmermann (University of Leuven)

14:30–15:00 **Adventures in Feature Selection on an Industrial Dataset... and Ensuing General Discoveries**
George Forman (Hewlett-Packard Labs)

15:00–16:00 **Panel Discussion**

Room: Chem, LT3

Thu2C: Multi-Task and Transfer Learning II

Chair: George Paliouras

14:00–14:25 **Discriminative Factor Alignment across Heterogeneous Feature Space**
Fangwei Hu, Tianqi Chen, Nathan N. Liu, Qiang Yang and Yong Yu

Transfer learning as a new machine learning paradigm has gained increasing attention lately. In situations where the training data in a target domain are not sufficient to learn predictive models effectively, transfer learning leverages auxiliary source data from related domains for learning. While most of the existing works in this area are only focused on using the source data with the same representational structure as the target data, in this paper, we push this boundary further by extending transfer between text and images.

We integrate documents , tags and images to build a heterogeneous transfer learning factor alignment model and apply it to improve the performance of tag recommendation. Many algorithms for tag recommendation have been proposed, but many of them have problem; the algorithm may not perform well under cold start conditions or for items from the long tail of the tag frequency distribution. However, with the help of documents, our algorithm handles these problems and generally outperforms other tag recommendation methods, especially the non-transfer factor alignment model.

14:25–14:50 **Learning to Perceive Two-Dimensional Displays Using Probabilistic Grammars**
Nan Li, William W. Cohen and Kenneth R. Koedinger

People learn to read and understand various displays (e.g., tables on webpages and software user interfaces) every day. How do humans learn to process such displays? Can computers be efficiently taught to understand and use such displays? In this paper, we use statistical learning to model how humans learn to perceive visual displays. We extend an existing probabilistic context-free grammar learner to support learning within a two-dimensional space by incorporating spatial and temporal information. Experimental results in both synthetic domains and real world domains show that the proposed learning algorithm is effective in acquiring user interface layout. Furthermore, we evaluate the effectiveness of the proposed algorithm within an intelligent tutoring agent, *SimStudent*, by integrating the learned display representation into the agent. Experimental results in learning complex problem solving skills in three domains show that the learned display representation is as good as one created by a human expert, in that skill learning using the learned representation is as effective as using a manually created representation.

14:50–15:15 **Sparse Gaussian Processes for Multi-task Learning**
Yuyang Wang and Roni Khardon

Multi-task learning models using Gaussian processes (GP) have been recently developed and successfully applied in various applications. The main difficulty with this approach is the computational cost of inference using the union of examples from all tasks. The paper investigates this problem for the grouped mixed-effect GP model where each individual response is given by a fixed-effect, taken from one of a set of unknown groups, plus a random individual effect function that captures variations among individuals. Such models have been widely used in previous work but no sparse solutions have been developed. The paper presents the first sparse solution for such problems, showing how the sparse approximation can be obtained by maximizing a variational lower bound on the marginal likelihood, generalizing ideas from single-task Gaussian processes to handle the mixed-effect model as well as grouping. Experiments using artificial and real data validate the approach showing that it can recover the performance of inference with the full sample, that it outperforms baseline methods, and that it outperforms state of the art sparse solutions for other multi-task GP formulations.

15:15–15:40 **Transfer Spectral Clustering**
Wenhao Jiang and Fu-lai Chung

Transferring knowledge from auxiliary datasets has been proved useful in machine learning tasks. Its adoption in clustering however is still limited. Despite of its superior performance, spectral clustering has not yet been incorporated with knowledge transfer or transfer learning. In this paper, we make such an attempt and propose a new algorithm called transfer spectral clustering (TSC). It involves not only the data manifold information of the clustering task but also the feature manifold information shared between related clustering tasks. Furthermore, it makes use of co-clustering to achieve and control the knowledge transfer among tasks. As demonstrated by the experimental results, TSC can greatly improve the clustering performance by effectively using auxiliary unlabeled data when compared with other state-of-the-art clustering algorithms.

Room: Chem, LT2

Thu2B: Classification

Chair: Alessandro Moschitti

14:00–14:25 **A Note on Extending Generalization Bounds for Binary Large-Margin Classifiers to Multiple Classes**
Ürün Dogan, Tobias Glasmachers and Christian Igel

A generic way to extend generalization bounds for binary large-margin classifiers to large-margin multi-category classifiers is presented. The simple proceeding leads to surprisingly tight bounds showing the same $\tilde{O}(d^2)$ scaling in the number d of classes as state-of-the-art results. The approach is exemplified by extending a textbook bound based on Rademacher complexity, which leads to a multi-class bound depending on the sum of the margin violations of the classifier.

14:25–14:50 **Extension of the Rocchio Classification Method to Multi-modal Categorization of Documents in Social Media**
Amin Mantrach and Jean-Michel Renders

Most of the approaches in multi-view categorization use early fusion, late fusion or co-training strategies. We propose here a novel classification method that is able to efficiently capture the interactions across the different modes. This method is a multi-modal extension of the Rocchio classification algorithm – very popular in the Information Retrieval community. The extension consists of simultaneously maintaining different "centroid" representations for each class, in particular "cross-media" centroids that correspond to pairs of modes. To classify new data points, different scores are derived from similarity measures between the new data point and these different centroids; a global classification score is finally obtained by suitably aggregating the individual scores. This method outperforms the multi-view logistic regression approach (using either the early fusion or the late fusion strategies) on a social media corpus - namely the ENRON email collection - on two very different categorization tasks (folder classification and recipient prediction).

14:50–15:15 **Label-Noise Robust Logistic Regression and Its Applications**
Jakramate Bootkrajang and Ata Kabán

The classical problem of learning a classifier relies on a set of labelled examples, without ever questioning the correctness of the provided label assignments. However, there is an increasing realisation that labelling errors are not uncommon in real situations. In this paper we consider a label-noise robust version of the logistic regression and multinomial logistic regression classifiers and develop the following contributions: (i) We derive efficient multiplicative updates to estimate the label flipping probabilities, and we give a proof of convergence for our algorithm. (ii) We develop a novel sparsity-promoting regularisation approach which allows us to tackle challenging high dimensional noisy settings. (iii) Finally, we throughly evaluate the performance of our approach in synthetic experiments and we demonstrate several real applications including gene expression analysis, class topology discovery and learning from crowdsourcing data.

15:15–15:40 **Sentiment Classification with Supervised Sequence Embedding**
Dmitriy Bespalov, Yanjun Qi, Bing Bai and Ali Shokoufandeh

In this paper, we introduce a novel approach for modeling n-grams in a latent space learned from supervised signals. The proposed procedure uses only unigram features to model short phrases (n-grams) in the latent space. The phrases are then combined to form document-level latent representation for a given text, where position of an n-gram in the document is used to compute corresponding combining weight. The resulting two-stage supervised embedding is then coupled with a classifier to form an end-to-end system that we apply to the large-scale sentiment classification task. The proposed model does not require feature selection to retain effective features during pre-processing, and its parameter space grows linearly with size of n-gram. We present comparative evaluations of this method using two large-scale datasets for sentiment classification in online reviews (Amazon and TripAdvisor). The proposed method outperforms standard baselines that rely on bag-of-words representation populated with n-gram features.

Learning and Discovery in Symbolic Systems Biology

Oliver Ray and Katsumi Inoue

09:00–16:00 Room: Wills, 1.5

Symbolic Systems Biology is a rapidly emerging field involving the application of formal logic-based methods to Systems Biology and Bioinformatics. The main aim of the workshop is to explore how machine learning and knowledge discovery techniques can be used within such formalisms in order to further the practical utility of Symbolic Systems Biology. A secondary aim is to identify ways in which purely symbolic methods have been or could be combined with numerical techniques and applied to emerging problems in experimental and synthetic biology. Finally, we wish to provide a forum to discuss recent developments within symbolic system biology that enhance the power or usability of its constituent approaches.

09:00–09:05 **Welcome Address**

09:05–09:40 **Invited Talk: Equation Discovery for Systems Biology**
Saso Dzeroski

09:40–10:05 **Biocircuit Design with Equation Discovery**
Jovan Tanevski, Nikola Simidjievski and Saso Dzeroski

10:05–10:30 **Symbolic Systems Biology: a Roadmap**
Oliver Ray and Marcus Tindall

10:30–11:00 **Coffee Break**

11:00–11:25 **Abducing Biological Regulatory Networks from Process Hitting models**
Maxime Folschette, Loic Pauleve, Katsumi Inoue, Morgan Magnin and Olivier Roux

11:25–11:50 **Logical Modeling of Cancer and Chemoprevention**
Antonis Kakas, Sotiris Lazarou, Christiana Neophytou and Andreas Constantinou

11:50–12:15 **Learning a Causal Network from Temporal Cancer Gene Expression Data: A Logical Abduction Approach**
Hiroaki Watanabe, Stephen Muggleton, Richard Currie, Pooja Jain, Dianhuan Lin, Jianzhong Chen, Michael Sternberg, Charles Baxter, Jose Domingo Salazar and Stuart Dunbar

12:15–13:30 **Lunch (not provided)**

13:30–13:55 **Learning Monadic and Dyadic Relations: Three Case Studies in Systems Biology**
Michiel Stock, Tapio Pahikkala, Antti Airola, Tapio Salakoski, Bernard De Baets and Willem Waegeman

13:55–14:20 **Ongoing Work on Applying Multi-clause ILP to Identify Metabolic Control Points**
Dianhuan Lin, Jianzhong Chen, Hiroaki Watanabe, Pooja Jain, Charles Baxter, Richard Currie, Jose Domingo Salazar, Stuart Dunbar, Mark Earll, Michael Sternberg and Stephen Muggleton

14:20–14:45 **Towards Automatic Construction and Corroboration of Food Webs**
Alireza Tamaddoni-Nezhad, Ghazal Afroozi Milani, David Bohan, Stuart Dunbar, Alan Raybould and Stephen Muggleton

14:45–15:10 **Panel Discussion**

15:10–15:35 **Towards a Logic-based Method to Infer Provenance-aware Molecular Networks**
Zahira Aslaoui-Errafi, Sarah Cohen-Boulakia, Christine Froidevaux, Pauline Gloaguen, Anne Poupon, Adrien Rougny and Meriem Yahiaoui

15:35–16:00 **Towards a Graphical Rule Editor for the Pathway Logic Assistant**
Amna Abbas, Carolyn Talcott, Merrill Knapp and Oliver Ray

Analyzing Text and Social Network Data with Probabilistic Models

Padhraic Smyth

University of California, Irvine, USA
smyth@ics.uci.edu
http://www.ics.uci.edu/~smyth/

Monday 17:00–18:00 Room: Wills, Great Hall

Abstract

Exploring and understanding large text and social network data sets is of increasing interest across multiple fields, in computer science, social science, history, medicine, and more. This talk will present an overview of recent work using probabilistic latent variable models to analyze such data. Latent variable models have a long tradition in data analysis and typically hypothesize the existence of simple unobserved phenomena to explain relatively complex observed data. In the past decade there has been substantial work on extending the scope of these approaches from relatively small simple data sets to much more complex text and network data. We will discuss the basic concepts behind these developments, reviewing key ideas, recent advances, and open issues. In addition we will highlight common ideas that lie beneath the surface of different approaches including links (for example) to work in matrix factorization. The concluding part of the talk will focus more specifically on recent work with temporal social networks, specifically data in the form of time-stamped events between nodes (such as emails exchanged among individuals over time).

Bio

Padhraic Smyth is a Professor at the University of California, Irvine, in the Department of Computer Science with a joint appointment in Statistics, and is also Director of the Center for Machine Learning and Intelligent Systems at UC Irvine. His research interests include machine learning, data mining, pattern recognition, and applied statistics and he has published over 150 papers on these topics. He was a recipient of best paper awards at the 2002 and 1997 ACM SIGKDD Conferences, received the ACM SIGKDD Innovation Award in 2009, and was named a AAAI Fellow in 2010. He is co-author of Modeling the Internet and the Web: Probabilistic Methods and Algorithms (with Pierre Baldi and Paolo Frasconi in 2003), and co-author of Principles of Data Mining, MIT Press (with David Hand and Heikki Mannila in 2001). Padhraic has served in editorial and advisory positions for journals such as the Journal of Machine Learning Research, the Journal of the American Statistical Association, and the IEEE Transactions on Knowledge and Data Engineering. While at UC Irvine he has received research funding from agencies such as NSF, NIH, IARPA, NASA, and DOE, and from companies such as Google, IBM, Yahoo!, Experian, and Microsoft. In addition to his academic research he is also active in industry consulting, working with companies such as eBay, Yahoo!, Microsoft, Oracle, Nokia, and AT&T, as well as serving as scientific advisor to local startups in Orange County. He also served as an academic advisor to Netflix for the Netflix prize competition from 2006 to 2009. Padhraic received a first class honors degree in Electronic Engineering from National University of Ireland (Galway) in 1984, and the MSEE and PhD degrees (in 1985 and 1988 respectively) in Electrical Engineering from the California Institute of Technology. From 1988 to 1996 he was a Technical Group Leader at the Jet Propulsion Laboratory, Pasadena, and has been on the faculty at UC Irvine since 1996.

Room: Chem, LT1
Thu2A: Graphs, Trees, Sequences and Strings II Chair: Albert Bifet

14:00–14:25 **An Efficiently Computable Support Measure for Frequent Subgraph Pattern Mining**
Yuyi Wang and Jan Ramon

Graph support measures are functions measuring how frequently a given subgraph pattern occurs in a given database graph. An important class of support measures relies on overlap graphs. A major advantage of the overlap graph based approaches is that they combine anti-monotonicity with counting occurrences of a pattern which are independent according to certain criteria. However, existing overlap graph based support measures are expensive to compute.

In this paper, we propose a new support measure which is based on a new notion of independence. We show that our measure is the solution to a linear program which is usually sparse, and using interior point methods can be computed efficiently. We show experimentally that for large networks, in contrast to earlier overlap graph based proposals, pattern mining based on our support measure is feasible.

14:25–14:50 **Efficient Graph Kernels by Randomization**
Marion Neumann, Novi Patricia, Roman Garnett and Kristian Kersting

Learning from complex data is becoming increasingly important, and graph kernels have recently evolved into a rapidly developing branch of learning on structured data. However, previously proposed kernels rely on having discrete node label information. In this paper, we explore the power of continuous node-level features for propagation-based graph kernels. Specifically, propagation kernels exploit node label distributions from propagation schemes like label propagation, which naturally enables the construction of graph kernels for partially labeled graphs. In order to efficiently extract graph features from continuous node label distributions, and in general from continuous vector-valued node attributes, we utilize randomized techniques, which easily allow for deriving similarity measures based on propagated information. We show that propagation kernels utilizing locality-sensitive hashing reduce the runtime of existing graph kernels by several orders of magnitude. We evaluate the performance of various propagation kernels on real-world bioinformatics and image benchmark datasets.

14:50–15:15 **Graph Mining for Object Tracking in Videos**
Fabien Diot, Elisa Fromont, Baptiste Jeudy, Emmanuel Marilly and Olivier Martinot

This paper shows a concrete example of the use of graph mining for tracking objects in videos with moving cameras and without any contextual information on the objects to track. To make the mining algorithm efficient, we benefit from a video representation based on dynamic (evolving through time) planar graphs. We then define a number of constraints to efficiently find our so-called spatio-temporal graph patterns. Those patterns are linked through an occurrences graph to allow us to tackle occlusion or graph features instability problems in the video. Experiments on synthetic and real videos show that our method is effective and allows us to find relevant patterns for our tracking application.

15:15–15:40 **Hypergraph Learning with Hyperedge Expansion**
Li Pu and Boi Faltings

We propose a new formulation called *hyperedge expansion* (HE) for hypergraph learning. The HE expansion transforms the hypergraph into a directed graph on the hyperedge level. Compared to the existing works (e.g. star expansion or normalized hypergraph cut), the learning results with HE expansion would be less sensitive to the vertex distribution among clusters, especially in the case that cluster sizes are unbalanced. Because of the special structure of the auxiliary directed graph, the linear eigenvalue problem of the Laplacian can be transformed into a quadratic eigenvalue problem, which has some special properties suitable for semi-supervised learning and clustering problems. We show in the experiments that the new algorithms based on the HE expansion achieves statistically significant gains in classification performance and good scalability for the co-occurrence data.

Room: Chem, LT3

Thu1C: Multi-Task and Transfer Learning I — Chair: George Forman

10:30–10:55 **Efficient Training of Graph-Regularized Multitask SVMs**
Christian Widmer, Marius Kloft, Nico Görnitz and Gunnar Rätsch

We present an optimization framework for graph-regularized multi-task SVMs based on the *primal* formulation of the problem. Previous approaches employ a so-called multi-task kernel (MTK) and thus are inapplicable when the numbers of training examples n is large (typically $n < 20{,}000$, even for just a few tasks). In this paper, we present a primal optimization criterion, allowing for general loss functions, and derive its dual representation. Building on the work of Hsieh et al., we derive an algorithm for optimizing the large-margin objective and prove its convergence. Our computational experiments show a speedup of up to *three orders of magnitude* over LibSVM and SVMLight for several standard benchmarks as well as challenging data sets from the application domain of computational biology. Combining our optimization methodology with the COFFIN large-scale learning framework, we are able to train a multi-task SVM using over 1,000,000 training points stemming from 4 different tasks. An efficient C++ implementation of our algorithm is being made publicly available as a part of the SHOGUN machine learning toolbox.

10:55–11:20 **Geometry Preserving Multi-task Metric Learning**
Peipei Yang, Kaizhu Huang and Cheng-Lin Liu

Multi-task learning has been widely studied in machine learning due to its capability to improve the performance of multiple related learning problems. However, few researchers have applied it on the important metric learning problem. In this paper, we propose to couple multiple related metric learning tasks with von Neumann divergence. On one hand, the novel regularized approach extends previous methods from the vector regularization to a general matrix regularization framework; on the other hand and more importantly, by exploiting von Neumann divergence as the regularizer, the new multi-task metric learning has the capability to well preserve the data geometry. This leads to more appropriate propagation of side-information among tasks and provides potential for further improving the performance. We propose the concept of *geometry preserving probability (PG)* and show that our framework leads to a larger PG in theory. In addition, our formulation proves to be jointly convex and the global optimal solution can be guaranteed. A series of experiments across very different disciplines verify that our proposed algorithm can consistently outperform the current methods.

11:20–11:45 **Learning and Inference in Probabilistic Classifier Chains with Beam Search**
Abhishek Kumar, Shankar Vembu, Aditya Krishna Menon and Charles Elkan

Multilabel learning is an extension of binary classification that is both challenging and practically important. Recently, a method for multilabel learning called *probabilistic classifier chains* (PCCs) was proposed with numerous appealing properties, such as conceptual simplicity, flexibility, and theoretical justification. However, PCCs suffer from the *computational* issue of having inference that is exponential in the number of tags, and the *practical* issue of being sensitive to the suitable ordering of the tags while training. In this paper, we show how the classical technique of *beam search* may be used to solve both these problems. Specifically, we show how to use beam search to perform tractable test time inference, and how to integrate beam search with training to determine a suitable tag ordering. Experimental results on a range of multilabel datasets show that these proposed changes dramatically extend the practical viability of PCCs.

11:45–12:10 **Learning Multiple Tasks with Boosted Decision Trees**
Jean Baptiste Faddoul, Boris Chidlovskii, Rémi Gilleron and Fabien Torre

We address the problem of multi-task learning with no label correspondence among tasks. Learning multiple related tasks simultaneously, by exploiting their shared knowledge can improve the predictive performance on every task. We develop the multi-task Adaboost environment with Multi-Task Decision Trees as weak classifiers. We first adapt the well known decision tree learning to the multi-task setting. We revise the information gain rule for learning decision trees in the multi-task setting. We use this feature to develop a novel criterion for learning Multi-Task Decision Trees. The criterion guides the tree construction by learning the decision rules from data of different tasks, and representing different degrees of task relatedness. We then modify MT-Adaboost to combine Multi-task Decision Trees as weak learners. We experimentally validate the advantage of the new technique; we report results of experiments conducted on several multi-task datasets, including the Enron email set and Spam Filtering collection.

Machine Learning for Robotics

Pieter Abbeel

University of California, Berkeley, USA
pabbeel@cs.berkeley.edu
http://www.cs.berkeley.edu/~pabbeel/

Tuesday 09:00–10:00 Room: Chem, LT1

Abstract

Robots are typically far less capable in autonomous mode than in tele-operated mode. The few exceptions tend to stem from long days (and more often weeks, or even years) of expert engineering for a specific robot and its operating environment. Current control methodology is quite slow and labor intensive. I believe advances in machine learning have the potential to revolutionize robotics. In this talk, I will present new machine learning techniques we have developed that are tailored to robotics. I will describe in depth "Apprenticeship learning", a new approach to high-performance robot control based on learning for control from ensembles of expert human demonstrations. Our initial work in apprenticeship learning has enabled the most advanced helicopter aerobatics to-date, including maneuvers such as chaos, tic-tocs, and auto-rotation landings which only exceptional expert human pilots can fly. Our most recent work in apprenticeship learning is providing traction on learning to perform challenging robotic manipulation tasks, such as knot-tying. I will also briefly highlight three other machine learning for robotics developments: Inverse reinforcement learning and its application to quadruped locomotion, Safe exploration in reinforcement learning which enables robots to learn on their own, and Learning for perception with application to robotic laundry.

Bio

Pieter Abbeel received a BS/MS in Electrical Engineering from KU Leuven (Belgium) and received his Ph.D. degree in Computer Science from Stanford University in 2008. He joined the faculty at UC Berkeley in Fall 2008, with an appointment in the Department of Electrical Engineering and Computer Sciences. He has won various awards, including best paper awards at ICML and ICRA, the Sloan Fellowship, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) award, the Okawa Foundation award, the 2011's TR35, and the IEEE Robotics and Automation Society (RAS) Early Career Award. He has developed apprenticeship learning algorithms which have enabled advanced helicopter aerobatics, including maneuvers such as tic-tocs, chaos and auto-rotation, which only exceptional human pilots can perform. His group has also enabled the first end-to-end completion of reliably picking up a crumpled laundry article and folding it. His work has been featured in many popular press outlets, including BBC, New York Times, MIT Technology Review, Discovery Channel, SmartPlanet and Wired. His current research focuses on robotics and machine learning with a particular focus on challenges in personal robotics, surgical robotics and connectomics.

Room: Chem, LT1

Tue1A: Bayesian Learning and Graphical Models

Chair: Jaakko Hollmén

10:30–10:55 **An Experimental Comparison of Hybrid Algorithms for Bayesian Network Structure Learning**
Maxime Gasse, Alex Aussem and Haytham Elghazel

10:55–11:20 **Bayesian Network Classifiers with Reduced Precision Parameters**
Sebastian Tschiatschek, Peter Reinprecht, Manfred Mücke and Franz Pernkopf

11:20–11:45 **Combining Subjective Probabilities and Data in Training Markov Logic Networks**
Tivadar Pápai, Shalini Ghosh and Henry Kautz

11:45–12:10 **Score-Based Bayesian Skill Learning**
Shengbo Guo, Scott Sanner, Thore Graepel and Wray Buntine

Room: Chem, LT2

Tue1B: Association Rules and Frequent Patterns

Chair: Matthijs van Leeuwen

10:30–10:55 **Discovering Descriptive Tile Trees by Mining Optimal Geometric Subtiles**
Nikolaj Tatti and Jilles Vreeken

10:55–11:20 **Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees**
Matteo Riondato and Eli Upfal

11:20–11:45 **General Algorithms for Mining Closed Flexible Patterns under Various Equivalence Relations**
Tomohiro I, Yuki Enokuma, Hideo Bannai and Masayuki Takeda

11:45–12:10 **Smoothing Categorical Data**
Arno Siebes and René Kersten

Room: Chem, LT3

Tue1C: Reinforcement Learning and Planning I

Chair: Balaraman Ravindran

10:30–10:55 **Adaptive Planning for Markov Decision Processes with Uncertain Transition Models via Incremental Feature Dependency Discovery**
N. Kemal Ure, Alborz Geramifard, Girish Chowdhary and Jonathan P. How

10:55–11:20 **APRIL: Active Preference Learning-Based Reinforcement Learning**
Riad Akrour, Marc Schoenauer and Michèle Sebag

11:20–11:45 **Autonomous Data-Driven Decision-Making in Smart Electricity Markets**
Markus Peters, Wolfgang Ketter, Maytal Saar-Tsechansky and John Collins

11:45–12:10 **Bayesian Nonparametric Inverse Reinforcement Learning**
Bernard Michini and Jonathan P. How

Room: Chem, LT4

Tue1D: Industry Track on Big Data

Chair: Cedric Archambeau

10:30–11:20 **Large-Scale Language Learning**
Slav Petrov (Google)

11:20–12:10 **Finding Low-Dimensional Structure in High-Dimensional Visual Data**
David Wipf (Microsoft Research Asia)

Room: Chem, LT2

Thu1B: Rule Learning and Subgroup Discovery

Chair: Eyke Hüllermeier

10:30–10:55 **A Bayesian Approach for Classification Rule Mining in Quantitative Databases**
Dominique Gay and Marc Boullé

We suggest a new framework for classification rule mining in quantitative data sets founded on Bayes theory – without univariate preprocessing of attributes. We introduce a space of rule models and a prior distribution defined on this model space. As a result, we obtain the definition of a parameter-free criterion for classification rules. We show that the new criterion identifies interesting classification rules while being highly resilient to spurious patterns. We develop a new parameter-free algorithm to mine locally optimal classification rules efficiently. The mined rules are directly used as new features in a classification process based on a selective naive Bayes classifier. The resulting classifier demonstrates higher inductive performance than state-of-the-art rule-based classifiers.

10:55–11:20 **A Bayesian Scoring Technique for Mining Predictive and Non-Spurious Rules**
Iyad Batal, Gregory Cooper and Milos Hauskrecht

Rule mining is an important class of data mining methods for discovering interesting patterns in data. The success of a rule mining method heavily depends on the evaluation function that is used to assess the quality of the rules. In this work, we propose a new rule evaluation score - the Predictive and Non-Spurious Rules (PNSR) score. This score relies on Bayesian inference to evaluate the quality of the rules and considers the structure of the rules to filter out spurious rules. We present an efficient algorithm for finding rules with high PNSR scores. The experiments demonstrate that our method is able to cover and explain the data with a much smaller rule set than existing methods.

11:20–11:45 **Generic Pattern Trees for Exhaustive Exceptional Model Mining**
Florian Lemmerich, Martin Becker and Martin Atzmueller

Exceptional model mining has been proposed as a variant of subgroup discovery especially focusing on complex target concepts. Currently, efficient mining algorithms are limited to heuristic (non exhaustive) methods. In this paper, we propose a novel approach for fast exhaustive exceptional model mining: We introduce the concept of valuation bases as an intermediate condensed data representation, and present the general GP-growth algorithm based on FP-growth. Furthermore, we discuss the scope of the proposed approach by drawing an analogy to data stream mining and provide examples for several different model classes. Runtime experiments show improvements of more than an order of magnitude in comparison to a naive exhaustive depth-first search.

11:45–12:10 **Handling Time Changing Data with Adaptive Very Fast Decision Rules**
Petr Kosina and João Gama

Data streams are usually characterized by changes in the underlying distribution generating data. Therefore algorithms designed to work with data streams should be able to detect changes and quickly adapt the decision model. Rules are one of the most interpretable and flexible models for data mining prediction tasks. In this paper we present the Adaptive Very Fast Decision Rules (AVFDR), an on-line, any-time and one-pass algorithm for learning decision rules in the context of time changing data. AVFDR can learn ordered and unordered rule sets. It is able to adapt the decision model via incremental induction and specialization of rules. Detecting local drifts takes advantage of the modularity of rule sets. In AVFDR, each individual rule monitors the evolution of performance metrics to detect concept drift. AVFDR prunes rules that detect drift. This explicit change detection mechanism provides useful information about the dynamics of the process generating data, faster adaption to changes and generates compact rule sets. The experimental evaluation shows this method is able to learn fast and compact rule sets from evolving streams in comparison to alternative methods.

forces of topic diffusion in SCN (Scientific Collaboration Network). Hence, we build the model upon the explanatory variables representing above two driving forces. Extensive experimental results show that our model can consistently achieves good predicting performance. Such results are independent of the tested topics and significantly better than that of state-of-the-art competitor.

Room: Chem, LT1

Tue2A: Graphs, Trees, Sequences and Strings I

Chair: José L. Balcázar

14:00–14:25 **A Family of Feed-forward Models for Protein Sequence Classification**
Sam Blasiak, Huzefa Rangwala and Kathryn Laskey

14:25–14:50 **Nearly Exact Mining of Frequent Trees in Large Networks**
Ashraf M. Kibriya and Jan Ramon

14:50–15:15 **Reachability Analysis and Modeling of Dynamic Event Networks**
Kathy Macropol and Ambuj Singh

15:15–15:40 **An Efficiently Computable Support Measure for Frequent Subgraph Pattern Mining**
Yuyi Wang and Jan Ramon

Room: Chem, LT2

Tue2B: Rankings and Recommendations

Chair: Carlos Soares

14:00–14:25 **A Live Comparison of Methods for Personalized Article Recommendation at Forbes.com**
Evan Kirshenbaum, George Forman and Michael Dugan

14:25–14:50 **Active Evaluation of Ranking Functions Based on Graded Relevance**
Christoph Sawade, Steffen Bickel, Timo von Oertzen, Tobias Scheffer and Niels Landwehr

14:50–15:15 **Fast ALS-Based Tensor Factorization for Context-Aware Recommendation from Implicit Feedback**
Balázs Hidasi and Domonkos Tikk

15:15–15:40 **Probability Estimation for Multi-class Classification Based on Label Ranking**
Weiwei Cheng and Eyke Hüllermeier

Room: Chem, LT3

Tue2C: Ensemble Methods

Chair: Herna Viktor

14:00–14:25 **Boosting Nearest Neighbors for the Efficient Estimation of Posteriors**
Roberto D'Ambrosio, Richard Nock, Wafa Bel Haj Ali, Frank Nielsen and Michel Barlaud

14:25–14:50 **Diversity Regularized Ensemble Pruning**
Nan Li, Yang Yu and Zhi-Hua Zhou

14:50–15:15 **Ensembles on Random Patches**
Gilles Louppe and Pierre Geurts

15:15–15:40 **Multi-Task Boosting by Exploiting Task Relationships**
Yu Zhang and Dit-Yan Yeung

Room: Chem, LT4

Tue2D: Industry Track on Big Data

Chair: Cedric Archambeau

14:00–14:50 **Demand Management for On-street Parking: Data, Analysis, and Actions**
Onno Zoeter (Xerox Research Centre Europe)

14:50–15:40 **Trusting Cloudy Data**
Simon Shiu (HP Labs)

Room: Chem, LT1

Tue3A: Dimensionality Reduction, Feature Selection and Extraction

Chair: Katharina Morik

16:10–16:35 **Embedding Monte Carlo Search of Features in Tree-Based Ensemble Methods**
Francis Maes, Pierre Geurts and Louis Wehenkel

16:35–17:00 **Hypergraph Spectra for Semi-supervised Feature Selection**
Zhihong Zhang, Edwin R. Hancock and Xiao Bai

17:00–17:25 **Learning Neighborhoods for Metric Learning**
Jun Wang, Adam Woznica and Alexandros Kalousis

17:25–17:50 **PCA, Eigenvector Localization and Clustering for Side-Channel Attacks on Cryptographic Hardware Devices**
Dimitrios Mavroeidis, Lejla Batina, Twan van Laarhoven and Elena Marchiori

Room: Chem, LT2

Tue3B: Multi-Relational Mining and Learning

Chair: Pauli Miettinen

16:10–16:35 **Author Name Disambiguation Using a New Categorical Distribution Similarity**
Shaohua Li, Gao Cong and Chunyan Miao

16:35–17:00 **Lifted Online Training of Relational Models with Stochastic Gradient Methods**
Babak Ahmadi, Kristian Kersting and Sriraam Natarajan

17:00–17:25 **Scalable Relation Prediction Exploiting Both Intrarelational Correlation and Contextual Information**
Xueyan Jiang, Volker Tresp, Yi Huang, Maximilian Nickel and Hans-Peter Kriegel

17:25–17:50 **Relational Differential Prediction**
Houssam Nassif, Vítor Santos Costa, Elizabeth S. Burnside and David Page

Room: Chem, LT3

Tue3C: Semi-Supervised and Transductive Learning

Chair: Florence d'Alché

16:10–16:35 **Bidirectional Semi-supervised Learning with Graphs**
Tomoharu Iwata and Kevin Duh

16:35–17:00 **Graph-Based Transduction with Confidence**
Matan Orbach and Koby Crammer

17:00–17:25 **Maximum Consistency Preferential Random Walks**
Deguang Kong and Chris Ding

17:25–17:50 **Semi-supervised Multi-label Classification: A Simultaneous Large-Margin, Subspace Learning Approach**
Yuhong Guo and Dale Schuurmans

Room: Chem, LT4

Tue3D: Demo Spotlights

Chair: Bettina Berendt and Myra Spiliopoulou

16:10–16:20 **VIKAMINE – Open-Source Subgroup Discovery, Pattern Mining, and Analytics**
Martin Atzmueller and Florian Lemmerich

16:20–16:30 **Association Rule Mining Following the Web Search Paradigm**
Radek Škrabal, Milan Šimůnek, Stanislav Vojíř, Andrej Hazucha, Tomáš Marek, David Chudán and Tomáš Kliegr

session continues on next page

Room: Chem, LT1

Thu1A: Social Network Mining II

Chair: Francesco Bonchi

10:30–10:55 **On Approximation of Real-World Influence Spread**
Yu Yang, Enhong Chen, Qi Liu, Biao Xiang, Tong Xu and Shafqat Ali Shad

To find the most influential nodes for viral marketing, several models have been proposed to describe the influence propagation process. Among them, the *Independent Cascade (IC) Model* is most widely-studied. However, under IC model, computing influence spread (i.e., the expected number of nodes that will be influenced) for each given seed set has been proved to be #P-hard. To that end, in this paper, we propose *GS* algorithm for quick approximation of influence spread by solving a linear system, based on the fact that propagation probabilities in real-world social networks are usually quite small. Furthermore, for better approximation, we study the structural defect problem existing in networks, and correspondingly, propose enhanced algorithms, *GSbyStep* and *SSSbyStep*, by incorporating the *Maximum Influence Path* heuristic. Our algorithms are evaluated by extensive experiments on four social networks. Experimental results show that our algorithms can get better approximations to the IC model than the state-of-the-arts.

10:55–11:20 **Opinion Formation by Voter Model with Temporal Decay Dynamics**
Masahiro Kimura, Kazumi Saito, Kouzou Ohara and Hiroshi Motoda

Social networks play an important role for spreading information and forming opinions. A variety of voter models have been defined that help analyze how people make decisions based on their neighbors' decisions. In these studies, common practice has been to use the latest decisions in opinion formation process. However, people may decide their opinions by taking account not only of their neighbors' latest opinions, but also of their neighbors' past opinions. To incorporate this effect, we enhance the original voter model and define the temporal decay voter (TDV) model incorporating a temporary decay function with parameters, and propose an efficient method of learning these parameters from the observed opinion diffusion data. We further propose an efficient method of selecting the most appropriate decay function from among the candidate functions each with the optimized parameter values. We adopt three functions as the typical candidates: the exponential decay, the power-law decay, and no decay, and evaluate the proposed method (parameter learning and model selection) through extensive experiments. We, first, experimentally demonstrate, by using synthetic data, the effectiveness of the proposed method, and then we analyze the real opinion diffusion data from a Japanese word-of-mouth communication site for cosmetics using three decay functions above, and show that most opinions conform to the TDV model of the power-law decay function.

11:20–11:45 **Viral Marketing for Product Cross-Sell through Social Networks**
Ramasuri Narayanam and Amit A. Nanavati

The well known *influence maximization problem* (or viral marketing through social networks) deals with selecting a few influential initial seeds to maximize the awareness of product(s) over the social network. In this paper, we introduce a novel and generalized version of the influence maximization problem that considers simultaneously the following three practical aspects: (i) Often cross-sell among products is possible, (ii) Product specific costs (and benefits) for promoting the products have to be considered, and (iii) Since a company often has budget constraints, the initial seeds have to be chosen within a given budget. We refer to this generalized problem setting as *Budgeted Influence Maximization with Cross-sell of Products (B-IMCP)*. To the best of our knowledge, we are not aware of any work in the literature that addresses the B-IMCP problem which is the subject matter of this paper. Given a fixed budget, one of the key issues associated with the B-IMCP problem is to choose the initial seeds within this budget not only for the individual products, but also for promoting cross-sell phenomenon among these products. In particular, the following are the specific contributions of this paper: (i) We propose an influence propagation model to capture both the cross-sell phenomenon and product specific costs and benefits; (ii) As the B-IMCP problem is NP-hard computationally, we present a simple greedy approximation algorithm and then derive the approximation guarantee of this greedy algorithm by drawing upon the results from the theory of matroids; (iii) We then outline two efficient heuristics based on well known concepts in the literature. Finally, we experimentally evaluate the proposed approach for the B-IMCP problem using a few well known social network data sets such as WikiVote data set, Epinions, and Telecom call detail records data.

11:45–12:10 **Which Topic will You Follow?**
Deqing Yang, Yanghua Xiao, Bo Xu, Hanghang Tong, Wei Wang and Sheng Huang

Who are the most appropriate candidates to receive a call-for-paper or call-for-participation? What session topics should we propose for a conference of next year? To answer these questions, we need to precisely predict research topics of authors. In this paper, we build a MLR (Multiple Logistic Regression) model to predict the topic-following behavior of an author. By empirical studies, we find that social influence and homophily are two fundamental driving

Room: Chem, LT1

Thu3A: Large-Scale, Distributed and Parallel Mining and Learning II

Chair: Sofus A. Macskassy

16:10–16:35 **CC-MR – Finding Connected Components in Huge Graphs with MapReduce**
Thomas Seidl, Brigitte Boden and Sergej Fries

16:35–17:00 **Fast Near Neighbor Search in High-Dimensional Binary Data**
Anshumali Shrivastava and Ping Li

17:00–17:25 **Fully Sparse Topic Models**
Khoat Than and Tu Bao Ho

17:25–17:50 **Massively Parallel Feature Selection: An Approach Based on Variance Preservation**
Zheng Zhao, James Cox, David Duling and Warren Sarle

Room: Chem, LT2

Thu3B: Natural Language Processing

Chair: Eric Gaussier

16:10–16:35 **Collective Information Extraction with Context-Specific Consistencies**
Peter Kluegl, Martin Toepfer, Florian Lemmerich, Andreas Hotho and Frank Puppe

16:35–17:00 **Supervised Learning of Semantic Relatedness**
Ran El-Yaniv and David Yanay

17:00–17:25 **Unsupervised Bayesian Part of Speech Inference with Particle Gibbs**
Gregory Dubbin and Phil Blunsom

17:25–17:50 **WikiSent: Weakly Supervised Sentiment Analysis through Extractive Summarization with Wikipedia**
Subhabrata Mukherjee and Pushpak Bhattacharyya

Room: Chem, LT3

Thu3C: Reinforcement Learning and Planning II

Chair: Myra Spiliopoulou

16:10–16:35 **Bootstrapping Monte Carlo Tree Search with an Imperfect Heuristic**
Truong-Huy Dinh Nguyen, Wee-Sun Lee and Tze-Yun Leong

16:35–17:00 **Fast Reinforcement Learning with Large Action Sets Using Error-Correcting Output Codes for MDP Factorization**
Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux and Patrick Gallinari

17:00–17:25 **Learning Policies For Battery Usage Optimization in Electric Vehicles**
Stefano Ermon, Yexiang Xue, Carla Gomes and Bart Selman

17:25–17:50 **Structured Apprenticeship Learning**
Abdeslam Boularias, Oliver Krömer and Jan Peters

Room: Chem, LT4

Thu3D: Industry Track: Start-up Stories

Chair: Cedric Archambeau

16:10–17:00 **"Choice is Good, Choosing is a Chore" – Choosing the Right Database Platform for Effective Knowledge Discovery**
Alastair Page (JustOneDB)

Room: LT4

Tue3D: Demo Spotlights (continued)

Chair: Bettina Berendt and Myra Spiliopoulou

16:30–16:40 ***OutRules*: A Framework for Outlier Descriptions in Multiple Context Spaces**
Emmanuel Müller, Fabian Keller, Sebastian Blanc and Klemens Böhm

16:40–16:50 **Knowledge Discovery through Symbolic Regression with HeuristicLab**
Gabriel Kronberger, Stefan Wagner, Michael Kommenda, Andreas Beham, Andreas Scheibenpflug and Michael Affenzeller

16:50–17:00 **An Aspect-Lexicon Creation and Evaluation Tool for Sentiment Analysis Researchers**
Mus'ab Husaini, Ahmet Koçyiğit, Dilek Tapucu, Berrin Yanikoglu and Yücel Saygın

17:00–17:10 **ASV Monitor: Creating Comparability of Machine Learning Methods for Content Analysis**
Andreas Niekler, Patrick Jähnichen and Gerhard Heyer

17:10–17:20 **TopicExplorer: Exploring Document Collections with Topic Models**
Alexander Hinneburg, Rico Preiss and René Schröder

17:20–17:30 **Extracting Trajectories through an Efficient and Unifying Spatio-temporal Pattern Mining System**
Phan Nhat Hai, Dino Ienco, Pascal Poncelet and Maguelonne Teisseire

17:30–17:40 **Scientific Workflow Management with ADAMS**
Peter Reutemann and Joaquin Vanschoren

17:40–17:50 **ClowdFlows: A Cloud Based Scientific Workflow Platform**
Janez Kranjc, Vid Podpečan and Nada Lavrač

Room: Chem, LT1

Tue1A: Bayesian Learning and Graphical Models

Chair: Jaakko Hollmén

10:30–10:55 **An Experimental Comparison of Hybrid Algorithms for Bayesian Network Structure Learning**
Maxime Gasse, Alex Aussem and Haytham Elghazel

We present a novel hybrid algorithm for Bayesian network structure learning, called Hybrid HPC (H2PC). It first reconstructs the skeleton of a Bayesian network and then performs a Bayesian-scoring greedy hill-climbing search to orient the edges. It is based on a subroutine called HPC, that combines ideas from incremental and divide-and-conquer constraint-based methods to learn the parents and children of a target variable. We conduct an experimental comparison of H2PC against Max-Min Hill-Climbing (MMHC), which is currently the most powerful state-of-the-art algorithm for Bayesian network structure learning, on several benchmarks with various data sizes. Our extensive experiments show that H2PC outperforms MMHC both in terms of goodness of fit to new data and in terms of the quality of the network structure itself, which is closer to the true dependence structure of the data, without increasing the computational burden involved. The source code (in *R*) of H2PC as well as all data sets used for the empirical tests are publicly available.

10:55–11:20 **Bayesian Network Classifiers with Reduced Precision Parameters**
Sebastian Tschiatschek, Peter Reinprecht, Manfred Mücke and Franz Pernkopf

Bayesian network classifiers (BNCs) are probabilistic classifiers showing good performance in many applications. They consist of a directed acyclic graph and a set of conditional probabilities associated with the nodes of the graph. These conditional probabilities are also referred to as parameters of the BNCs. According to common believe, these classifiers are insensitive to deviations of the conditional probabilities under certain conditions. The first condition is that these probabilities are not too extreme, i.e. not too close to 0 or 1. The second is that the posterior over the classes is significantly different. In this paper, we investigate the effect of precision reduction of the parameters on the classification performance of BNCs. The probabilities are either determined generatively or discriminatively. Discriminative probabilities are typically more extreme. However, our results indicate that BNCs with discriminatively optimized parameters are almost as robust to precision reduction as BNCs with generatively optimized parameters. Furthermore, even large precision reduction does not decrease classification performance significantly. Our results allow the implementation of BNCs with less computational complexity. This supports application in embedded systems using floating-point numbers with small bit-width. Reduced bit-widths further enable to represent BNCs in the integer domain while maintaining the classification performance.

11:20–11:45 **Combining Subjective Probabilities and Data in Training Markov Logic Networks**
Tivadar Pápai, Shalini Ghosh and Henry Kautz

Markov logic is a rich language that allows one to specify a knowledge base as a set of weighted first-order logic formulas, and to define a probability distribution over truth assignments to ground atoms using this knowledge base. Usually, the weight of a formula cannot be related to the probability of the formula without taking into account the weights of the other formulas. In general, this is not an issue, since the weights are learned from training data. However, in many domains (*e.g.* healthcare, dependable systems, *etc.*), only little or no training data may be available, but one has access to a domain expert whose knowledge is available in the form of subjective probabilities. Within the framework of Bayesian statistics, we present a formalism for using a domain expert's knowledge for weight learning. Our approach defines priors that are different from and more general than previously used Gaussian priors over weights. We show how one can learn weights in an MLN by combining subjective probabilities and training data, without requiring that the domain expert provides consistent knowledge. Additionally, we also provide a formalism for capturing conditional subjective probabilities, which are often easier to obtain and more reliable than non-conditional probabilities. We demonstrate the effectiveness of our approach by extensive experiments in a domain that models failure dependencies in a cyber-physical system. Moreover, we demonstrate the advantages of using our proposed prior over that of using non-zero mean Gaussian priors in a commonly cited social network MLN testbed.

11:45–12:10 **Score-Based Bayesian Skill Learning**
Shengbo Guo, Scott Sanner, Thore Graepel and Wray Buntine

We extend the Bayesian skill rating system of TrueSkill to accommodate score-based match outcomes. TrueSkill has proven to be a very effective algorithm for matchmaking — the process of pairing competitors based on similar skill-level — in competitive online gaming. However, for the case of two teams/players, TrueSkill only learns from win, lose, or draw outcomes and cannot use additional match outcome information such as scores. To address this deficiency, we propose novel Bayesian graphical models as extensions of TrueSkill that (1) model player's offence and defence

Room: Chem, LT1

Thu2A: Graphs, Trees, Sequences and Strings II

Chair: Albert Bifet

14:00–14:25 **An Efficiently Computable Support Measure for Frequent Subgraph Pattern Mining**
Yuyi Wang and Jan Ramon

14:25–14:50 **Efficient Graph Kernels by Randomization**
Marion Neumann, Novi Patricia, Roman Garnett and Kristian Kersting

14:50–15:15 **Graph Mining for Object Tracking in Videos**
Fabien Diot, Elisa Fromont, Baptiste Jeudy, Emmanuel Marilly and Olivier Martinot

15:15–15:40 **Hypergraph Learning with Hyperedge Expansion**
Li Pu and Boi Faltings

Room: Chem, LT2

Thu2B: Classification

Chair: Alessandro Moschitti

14:00–14:25 **A Note on Extending Generalization Bounds for Binary Large-Margin Classifiers to Multiple Classes**
Ürün Dogan, Tobias Glasmachers and Christian Igel

14:25–14:50 **Extension of the Rocchio Classification Method to Multi-modal Categorization of Documents in Social Media**
Amin Mantrach and Jean-Michel Renders

14:50–15:15 **Label-Noise Robust Logistic Regression and Its Applications**
Jakramate Bootkrajang and Ata Kabán

15:15–15:40 **Sentiment Classification with Supervised Sequence Embedding**
Dmitriy Bespalov, Yanjun Qi, Bing Bai and Ali Shokoufandeh

Room: Chem, LT3

Thu2C: Multi-Task and Transfer Learning II

Chair: George Paliouras

14:00–14:25 **Discriminative Factor Alignment across Heterogeneous Feature Space**
Fangwei Hu, Tianqi Chen, Nathan N. Liu, Qiang Yang and Yong Yu

14:25–14:50 **Learning to Perceive Two-Dimensional Displays Using Probabilistic Grammars**
Nan Li, William W. Cohen and Kenneth R. Koedinger

14:50–15:15 **Sparse Gaussian Processes for Multi-task Learning**
Yuyang Wang and Roni Khardon

15:15–15:40 **Transfer Spectral Clustering**
Wenhao Jiang and Fu-lai Chung

Room: Chem, LT4

Thu2D: Industry Track: Startup Stories

Chair: David Barber

14:00–14:50 **Somebody Needs your Algorithm – Cloud'N'Sci.fi**
Petri Myllymaki (University of Helsinki, Ekahau, Cloud'N'Sci) and Pauli Misikangas (Ekahau, Cloud'N'Sci)

14:50–15:40 **Machine Learning at PeerIndex: Telling stories about users and their influence**
Ferenc Huszar (Peerindex)

Room: Chem, LT1

Thu1A: Social Network Mining II

Chair: Francesco Bonchi

10:30–10:55 **On Approximation of Real-World Influence Spread**
Yu Yang, Enhong Chen, Qi Liu, Biao Xiang, Tong Xu and Shafqat Ali Shad

10:55–11:20 **Opinion Formation by Voter Model with Temporal Decay Dynamics**
Masahiro Kimura, Kazumi Saito, Kouzou Ohara and Hiroshi Motoda

11:20–11:45 **Viral Marketing for Product Cross-Sell through Social Networks**
Ramasuri Narayanam and Amit A. Nanavati

11:45–12:10 **Which Topic will You Follow?**
Deqing Yang, Yanghua Xiao, Bo Xu, Hanghang Tong, Wei Wang and Sheng Huang

Room: Chem, LT2

Thu1B: Rule Learning and Subgroup Discovery

Chair: Eyke Hüllermeier

10:30–10:55 **A Bayesian Approach for Classification Rule Mining in Quantitative Databases**
Dominique Gay and Marc Boullé

10:55–11:20 **A Bayesian Scoring Technique for Mining Predictive and Non-Spurious Rules**
Iyad Batal, Gregory Cooper and Milos Hauskrecht

11:20–11:45 **Generic Pattern Trees for Exhaustive Exceptional Model Mining**
Florian Lemmerich, Martin Becker and Martin Atzmueller

11:45–12:10 **Handling Time Changing Data with Adaptive Very Fast Decision Rules**
Petr Kosina and João Gama

Room: Chem, LT3

Thu1C: Multi-Task and Transfer Learning I

Chair: George Forman

10:30–10:55 **Efficient Training of Graph-Regularized Multitask SVMs**
Christian Widmer, Marius Kloft, Nico Görnitz and Gunnar Rätsch

10:55–11:20 **Geometry Preserving Multi-task Metric Learning**
Peipei Yang, Kaizhu Huang and Cheng-Lin Liu

11:20–11:45 **Learning and Inference in Probabilistic Classifier Chains with Beam Search**
Abhishek Kumar, Shankar Vembu, Aditya Krishna Menon and Charles Elkan

11:45–12:10 **Learning Multiple Tasks with Boosted Decision Trees**
Jean Baptiste Faddoul, Boris Chidlovskii, Rémi Gilleron and Fabien Torre

Room: Chem, LT4

Thu1D: Industry Track: Start-up Stories

Chair: David Barber

10:30–10:55 **Who is Buying What in the UK? Mining e-commerce Data**
Jurgen Van Gael (Rangespan)

10:55–12:10 **Automated real-time intelligent marketing**
Jason McFall (Causata)

skills separately and (2) model how these offence and defence skills interact to generate score-based match outcomes. We derive efficient (approximate) Bayesian inference methods for inferring latent skills in these new models and evaluate them on three real data sets including Halo 2 XBox Live matches. Empirical evaluations demonstrate that the new score-based models (a) provide more accurate win/loss probability estimates than TrueSkill when training data is limited, (b) provide competitive and often better win/loss classification performance than TrueSkill, and (c) provide reasonable score outcome predictions with an appropriate choice of likelihood — prediction for which TrueSkill was not designed, but which can be useful in many applications.

Room: Chem, LT2

Tue1B: Association Rules and Frequent Patterns

Chair: Matthijs van Leeuwen

10:30–10:55 **Discovering Descriptive Tile Trees by Mining Optimal Geometric Subtiles**
Nikolaj Tatti and Jilles Vreeken

When analysing binary data, the ease at which one can interpret results is very important. Many existing methods, however, discover either models that are difficult to read, or return so many results interpretation becomes impossible. Here, we study a fully automated approach for mining easily interpretable models for binary data. We model data hierarchically with noisy tiles—rectangles with significantly different density than their parent tile. To identify good trees, we employ the Minimum Description Length principle.

We propose STIJL, a greedy any-time algorithm for mining good tile trees from binary data. Iteratively, it finds the locally *optimal* addition to the current tree, allowing overlap with tiles of the same parent. A major result of this paper is that we find the optimal tile in only $\Theta(NM \min(N, M))$ time. STIJL can either be employed as a top-k miner, or by MDL we can identify the tree that describes the data best.

Experiments show we find succinct models that accurately summarise the data, and, by their hierarchical property are easily interpretable.

10:55–11:20 **Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees**
Matteo Riondato and Eli Upfal

The tasks of extracting (top-K) Frequent Itemsets (FI's) and Association Rules (AR's) are fundamental primitives in data mining and database applications. Exact algorithms for these problems exist and are widely used, but their running time is hindered by the need of scanning the entire dataset, possibly multiple times. High quality approximations of FI's and AR's are sufficient for most practical uses, and a number of recent works explored the application of sampling for fast discovery of approximate solutions to the problems. However, these works do not provide satisfactory performance guarantees on the quality of the approximation, due to the difficulty of bounding the probability of under- or over-sampling any one of an unknown number of frequent itemsets. In this work we circumvent this issue by applying the statistical concept of *Vapnik-Chervonenkis (VC) dimension* to develop a novel technique for providing tight bounds on the sample size that guarantees approximation within user-specified parameters. Our technique applies both to absolute and to relative approximations of (top-K) FI's and AR's. The resulting sample size is linearly dependent on the VC-dimension of a range space associated with the dataset to be mined. The main theoretical contribution of this work is a characterization of the VC-dimension of this range space and a proof that it is upper bounded by an easy-to-compute characteristic quantity of the dataset which we call *d-index*, namely the maximum integer d such that the dataset contains at least d transactions of length at least d. We show that this bound is strict for a large class of datasets. The resulting sample size for an absolute (resp. relative) (ε,δ)-approximation of the collection of FI's is $O(\frac{1}{\varepsilon^2}(d+\log\frac{1}{\delta}))$ (resp. $O(\frac{2+\varepsilon}{\varepsilon^2(2-\varepsilon)\theta}(d\log\frac{2+\varepsilon}{(2-\varepsilon)\theta}+\log\frac{1}{\delta}))$) transactions, which is a significant improvement over previous known results. We present an extensive experimental evaluation of our technique on real and artificial datasets, demonstrating the practicality of our methods, and showing that they achieve even higher quality approximations than what is guaranteed by the analysis.

11:20–11:45 **General Algorithms for Mining Closed Flexible Patterns under Various Equivalence Relations**
Tomohiro I, Yuki Enokuma, Hideo Bannai and Masayuki Takeda

We address the closed pattern discovery problem in sequential databases for the class of *flexible* patterns. We propose two techniques of coarsening existing equivalence relations on the set of patterns to obtain new equivalence relations. Our new algorithm GenCloFlex is a generalization of MaxFlex proposed by Arimura and Uno (2007) that was designed for a particular equivalence relation. GenCloFlex can cope with existing, as well as new equivalence relations, and we investigate the computational complexities of the algorithm for respective equivalence relations. Then, we present an improved algorithm GenCloFlex+ based on new pruning techniques, which improve the delay time per output for some of the equivalence relations. By computational experiments on synthetic data, we show that most of the redundancies in the mined patterns are removed using the proposed equivalence relations.

11:45–12:10 **Smoothing Categorical Data**
Arno Siebes and René Kersten

Global models of a dataset reflect not only the large scale structure of the data distribution, they also reflect small(er) scale structure. Hence, if one wants to see the large scale structure, one should somehow subtract this smaller scale structure from the model.

Solving Problems with Visual Analytics: Challenges and Applications

Daniel Keim

University of Konstanz
Daniel.Keim@uni-konstanz.de
www.informatik.uni-konstanz.de/arbeitsgruppen/infovis/mitglieder/prof-dr-daniel-keim/

Thursday 09:00–10:00 Room: Chem, LT1

Abstract

Never before in history data is generated and collected at such high volumes as it is today. As the volumes of data available to business people, scientists, and the public increase, their effective use becomes more challenging. Keeping up to date with the flood of data, using standard tools for data analysis and exploration, is fraught with difficulty. The field of visual analytics seeks to provide people with better and more effective ways to explore and understand large datasets, while also enabling them to act upon their findings immediately. Visual analytics integrates the analytic capabilities of the computer and the perceptual and intellectual abilities of the human analyst, allowing novel discoveries and empowering individuals to take control of the analytical process. Visual analytics enables unexpected insights, which may lead to beneficial and profitable innovation. The talk presents the challenges of visual analytics and exemplifies them with several application examples, which illustrate the exiting potential of current visual analysis techniques but also their limitations.

Bio

Daniel A. Keim is full professor and head of the Information Visualization and Data Analysis Research Group at the University of Konstanz, Germany. He has been actively involved in information visualization and data analysis research for about 20 years and developed a number of novel visual analysis techniques for very large data sets with applications to a wide range of application areas including financial analysis, network analysis, geo-spatial analysis, as well as text and multimedia analysis. His research resulted in two recent books "Solving problems with Visual Analytics" and "Interactive Data Visualization" which he both co-authored. Dr. Keim has been program co-chair of the IEEE InfoVis and IEEE VAST symposia as well as the SIGKDD conference, and he is or was member of the IEEE InfoVis, IEEE VAST, and EuroVis steering committees. He is an associate editor of Palgrave's Information Visualization Journal (since 2001) and has been an associate editor of the IEEE Transactions on Visualization and Computer Graphics (1999–2004), the IEEE Transactions on Knowledge and Data Engineering (2002–2007), and the Knowledge and Information System Journal (2006–2011). He is coordinator of the German Strategic Research Initiative (SPP) on Scalable Visual Analytics and he was the scientific coordinator of the EU Coordination Action on Visual Analytics called VisMaster. Dr. Keim got his Ph.D. and habilitation degrees in computer science from the University of Munich. Before joining the University of Konstanz, Dr. Keim was associate professor at the University of Halle, Germany and Technology Consultant at AT&T Shannon Research Labs, NJ, USA.

Room: Chem, LT3

Wed3C: Privacy and Security

Chair: Patrick Gallinari

16:10–16:35 ***AUDIO*: An Integrity Auditing Framework of Outlier-Mining-as-a-Service Systems**
Ruilin Liu, Hui (Wendy) Wang, Anna Monreale, Dino Pedreschi, Fosca Giannotti and Wenge Guo

Spurred by developments such as cloud computing, there has been considerable recent interest in the data-mining-as-a-service paradigm. Users lacking in expertise or computational resources can outsource their data and mining needs to a third-party service provider (server). Outsourcing, however, raises issues about *result integrity*: how can the data owner verify that the mining results returned by the server are correct? In this paper, we present AUDIO, an integrity auditing framework for the specific task of distance-based outlier mining outsourcing. It provides efficient and practical verification approaches to check both completeness and correctness of the mining results. The key idea of our approach is to insert a small amount of artificial tuples into the outsourced data; the artificial tuples will produce artificial outliers and non-outliers that do not exist in the original dataset. The server's answer is verified by analyzing the presence of artificial outliers/non-outliers, obtaining a probabilistic guarantee of correctness and completeness of the mining result. Our empirical results show the effectiveness and efficiency of our method.

16:35–17:00 **Differentially Private Projected Histograms: Construction and Use for Prediction**
Staal A. Vinterbo

Privacy concerns are among the major barriers to efficient secondary use of information and data on humans. Differential privacy is a relatively recent measure that has received much attention in machine learning as it quantifies individual risk using a strong cryptographically motivated notion of privacy. At the core of differential privacy lies the concept of information dissemination through a randomized process. One way of adding the needed randomness to any process is to pre-randomize the input. This can yield lower quality results than other more specialized approaches, but can be an attractive alternative when i. there does not exist a specialized differentially private alternative, or when ii. multiple processes applied in parallel can use the same pre-randomized input.

A simple way to do input randomization is to compute perturbed histograms, which essentially are noisy multiset membership functions. Unfortunately, computation of perturbed histograms is only efficient when the data stems from a low-dimensional discrete space. The restriction to discrete spaces can be mitigated by discretization; Lei presented in 2011 an analysis of discretization in the context of M-estimators. Here we address the restriction regarding the dimensionality of the data. In particular we present a differentially private approximation algorithm for selecting features that preserve conditional frequency densities, and use this to project data prior to computing differentially private histograms. The resulting projected histograms can be used as machine learning input and include the necessary randomness for differential privacy. We empirically validate the use of differentially private projected histograms for learning binary and multinomial logistic regression models using four real world data sets.

17:00–17:25 **Fairness-Aware Classifier with Prejudice Remover Regularizer**
Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh and Jun Sakuma

With the spread of data mining technologies and the accumulation of social data, such technologies and data are being used for determinations that seriously affect individuals' lives. For example, credit scoring is frequently determined based on the records of past credit data together with statistical prediction techniques. Needless to say, such determinations must be nondiscriminatory and fair in sensitive features, such as race, gender, religion, and so on. Several researchers have recently begun to attempt the development of analysis techniques that are aware of social fairness or discrimination. They have shown that simply avoiding the use of sensitive features is insufficient for eliminating biases in determinations, due to the indirect influence of sensitive information. In this paper, we first discuss three causes of unfairness in machine learning. We then propose a regularization approach that is applicable to any prediction algorithm with probabilistic discriminative models. We further apply this approach to logistic regression and empirically show its effectiveness and efficiency.

While for some kinds of model – such as boosted classifiers – it is easy to see the "important" components, for many kind of models this is far harder, if at all possible. In such cases one might try an implicit approach: simplify the data distribution without changing the large scale structure. That is, one might first smooth the local structure out of the dataset. Then induce a new model from this smoothed dataset. This new model should now reflect the large scale structure of the original dataset. In this paper we propose such a smoothing for categorical data and for one particular type of models, viz., code tables.

By experiments we show that our approach preserves the large scale structure of a dataset well. That is, the smoothed dataset is simpler while the original and smoothed datasets share the same large scale structure.

Room: Chem, LT3

Tue1C: Reinforcement Learning and Planning I

Chair: Balaraman Ravindran

10:30–10:55 **Adaptive Planning for Markov Decision Processes with Uncertain Transition Models via Incremental Feature Dependency Discovery**

N. Kemal Ure, Alborz Geramifard, Girish Chowdhary and Jonathan P. How

Solving large scale sequential decision making problems without prior knowledge of the state transition model is a key problem in the planning literature. One approach to tackle this problem is to learn the state transition model online using limited observed measurements. We present an adaptive function approximator (incremental Feature Dependency Discovery (iFDD)) that grows the set of features online to approximately represent the transition model. The approach leverages existing feature-dependencies to build a sparse representation of the state transition model. Theoretical analysis and numerical simulations in domains with state space sizes varying from thousands to millions are used to illustrate the benefit of using iFDD for incrementally building transition models in a planning framework.

10:55–11:20 **APRIL: Active Preference Learning-Based Reinforcement Learning**

Riad Akrour, Marc Schoenauer and Michèle Sebag

This paper focuses on reinforcement learning (RL) with limited prior knowledge. In the domain of swarm robotics for instance, the expert can hardly design a reward function or demonstrate the target behavior, forbidding the use of both standard RL and inverse reinforcement learning. Although with a limited expertise, the human expert is still often able to emit preferences and rank the agent demonstrations. Earlier work has presented an iterative preference-based RL framework: expert preferences are exploited to learn an approximate policy return, thus enabling the agent to achieve direct policy search. Iteratively, the agent selects a new candidate policy and demonstrates it; the expert ranks the new demonstration comparatively to the previous best one; the expert's ranking feedback enables the agent to refine the approximate policy return, and the process is iterated.

In this paper, preference-based reinforcement learning is combined with active ranking in order to decrease the number of ranking queries to the expert needed to yield a satisfactory policy. Experiments on the mountain car and the cancer treatment testbeds witness that a couple of dozen rankings enable to learn a competent policy.

11:20–11:45 **Autonomous Data-Driven Decision-Making in Smart Electricity Markets**

Markus Peters, Wolfgang Ketter, Maytal Saar-Tsechansky and John Collins

For the vision of a Smart Grid to materialize, substantial advances in intelligent decentralized control mechanisms are required. We propose a novel class of autonomous broker agents for retail electricity trading that can operate in a wide range of Smart Electricity Markets, and that are capable of deriving long-term, profit-maximizing policies. Our brokers use Reinforcement Learning with function approximation, they can accommodate arbitrary economic signals from their environments, and they learn efficiently over the large state spaces resulting from these signals. Our design is the first that can accommodate an offline training phase so as to automatically optimize the broker for particular market conditions. We demonstrate the performance of our design in a series of experiments using real-world energy market data, and find that it outperforms previous approaches by a significant margin.

11:45–12:10 **Bayesian Nonparametric Inverse Reinforcement Learning**

Bernard Michini and Jonathan P. How

Inverse reinforcement learning (IRL) is the task of learning the reward function of a Markov Decision Process (MDP) given the transition function and a set of observed demonstrations in the form of state-action pairs. Current IRL algorithms attempt to find a single reward function which explains the entire observation set. In practice, this leads to a computationally-costly search over a large (typically infinite) space of complex reward functions. This paper proposes the notion that if the observations can be partitioned into smaller groups, a class of much simpler reward functions can be used to explain each group. The proposed method uses a Bayesian nonparametric mixture model to automatically partition the data and find a set of simple reward functions corresponding to each partition. The simple rewards are interpreted intuitively as subgoals, which can be used to predict actions or analyze which states are important to the demonstrator. Experimental results are given for simple examples showing comparable performance to other IRL algorithms in nominal situations. Moreover, the proposed method handles cyclic tasks (where the agent begins and ends in the same state) that would break existing algorithms without modification. Finally, the new algorithm has a fundamentally different structure than previous methods, making it more computationally efficient in a real-world learning scenario where the state space is large but the demonstration set is small.

Room: Chem, LT2

Wed3B: Sensor Data

Chair: Martin Atzmueller

16:10–16:35 **MDL-Based Analysis of Time Series at Multiple Time-Scales**

Ugo Vespier, Arno Knobbe, Siegfried Nijssen and Joaquin Vanschoren

The behavior of many complex physical systems is affected by a variety of phenomena occurring at different temporal scales. Time series data produced by measuring properties of such systems often mirrors this fact by appearing as a composition of signals across different time scales. When the final goal of the analysis is to model the individual phenomena affecting a system, it is crucial to be able to recognize the right temporal scales and to separate the individual components of the data. In this paper, we approach this challenge through a combination of the Minimum Description Length (MDL) principle, feature selection strategies, and convolution techniques from the signal processing field. As a result, our algorithm produces a good decomposition of a given time series and, as a side effect, builds a compact representation of its identified components. Experiments demonstrate that our method manages to identify correctly both the number and the temporal scale of the components for real-world as well as artificial data and show the usefulness of our method as an exploratory tool for analyzing time series data.

16:35–17:00 **Separable Approximate Optimization of Support Vector Machines for Distributed Sensing**

Sangkyun Lee, Marco Stolpe and Katharina Morik

Sensor measurements from diverse locations connected with possibly low bandwidth communication channels pose a challenge of resource-restricted distributed data analyses. In such settings it would be desirable to perform learning in each location as much as possible, without transferring all data to a central node. Applying the support vector machines (SVMs) with nonlinear kernels becomes nontrivial, however.

In this paper, we present an efficient optimization scheme for training SVMs over such sensor networks. Our framework performs optimization independently in each node, using only the local features stored in the respective node. We make use of multiple local kernels and explicit approximations to the feature mappings induced by them. Together they allow us constructing a separable surrogate objective that provides an upper bound of the primal SVM objective. A central coordination is also designed to adjust the weights among local kernels for improved prediction, while minimizing communication cost.

17:00–17:25 **Unsupervised Inference of Auditory Attention from Biosensors**

Melih Kandemir, Arto Klami, Akos Vetek and Samuel Kaski

We study ways of automatically inferring the level of attention a user is paying to auditory content, with applications for example in automatic podcast highlighting and auto-pause, as well as in a selection mechanism in auditory interfaces. In particular, we demonstrate how the level of attention can be inferred in an unsupervised fashion, without requiring any labeled training data. The approach is based on measuring the (generalized) correlation or synchrony between the auditory content and physiological signals reflecting the state of the user. We hypothesize that the synchrony is higher when the user is paying attention to the content, and show empirically that the level of attention can indeed be inferred based on the correlation. In particular, we demonstrate that the novel method of time-varying Bayesian canonical correlation analysis gives unsupervised prediction accuracy comparable to having trained a supervised Gaussian process regression with labeled training data recorded from other users.

Room: Chem, LT1

Wed3A: Data Mining Process

Chair: Johannes Fürnkranz

16:10–16:35 **A Framework for Evaluating the Smoothness of Data-Mining Results**
Gaurav Misra, Behzad Golshan and Evimaria Terzi

The data-mining literature is rich in problems that are formalized as combinatorial-optimization problems. An indicative example is the *entity-selection* formulation that has been used to model the problem of selecting a subset of representative reviews from a review corpus or important nodes in a social network. Existing combinatorial algorithms for solving such entity-selection problems identify a set of entities (e.g., reviews or nodes) as important. Here, we consider the following question: how do small or large changes in the input dataset change the value or the structure of the such reported solutions?

We answer this question by developing a general framework for evaluating the *smoothness* (i.e., consistency) of the data-mining results obtained for the input dataset X. We do so by comparing these results with the results obtained for datasets that are within a small or a large distance from X. The algorithms we design allow us to perform such comparisons effectively and thus, approximate the results' smoothness efficiently. Our experimental evaluation on real datasets demonstrates the efficacy and the practical utility of our framework in a wide range of applications.

16:35–17:00 **Coupled Bayesian Sets Algorithm for Semi-supervised Learning and Information Extraction**
Saurabh Verma and Estevam R. Hruschka Jr.

Our inspiration comes from Nell (Never Ending Language Learning), a computer program running at Carnegie Mellon University to extract structured information from unstructured web pages. We consider the problem of semi-supervised learning approach to extract category instances (e.g. country(USA), city(New York)) from web pages, starting with a handful of labeled training examples of each category or relation, plus hundreds of millions of unlabeled web documents. Semi-supervised approaches using a small number of labeled examples together with many unlabeled examples are often unreliable as they frequently produce an internally consistent, but nevertheless, incorrect set of extractions. We believe that this problem can be overcome by simultaneously learning independent classifiers in a new approach named Coupled Bayesian Sets algorithm, based on Bayesian Sets, for many different categories and relations (in the presence of an ontology defining constraints that couple the training of these classifiers). Experimental results show that simultaneously learning a coupled collection of classifiers for random 11 categories resulted in much more accurate extractions than training classifiers through original Bayesian Sets algorithm, Naive Bayes, BaS-all and Coupled Pattern Learner (the category extractor used in NELL).

17:00–17:25 **Policy Iteration Based on a Learned Transition Model**
Vivek Ramavajjala and Charles Elkan

This paper investigates a reinforcement learning method that combines learning a model of the environment with least-squares policy iteration (LSPI). The LSPI algorithm learns a linear approximation of the optimal state-action value function; the idea studied here is to let this value function depend on a learned estimate of the expected next state instead of directly on the current state and action. This approach makes it easier to define useful basis functions, and hence to learn a useful linear approximation of the value function. Experiments show that the new algorithm, called NSPI for next-state policy iteration, performs well on two standard benchmarks, the well-known mountain car and inverted pendulum swing-up tasks. More importantly, the NSPI algorithm performs well, and better than a specialized recent method, on a resource management task known as the day-ahead wind commitment problem. This latter task has action and state spaces that are high-dimensional and continuous.

Room: Chem, LT1

Tue2A: Graphs, Trees, Sequences and Strings I

Chair: José L. Balcázar

14:00–14:25 **A Family of Feed-forward Models for Protein Sequence Classification**
Sam Blasiak, Huzefa Rangwala and Kathryn Laskey

Advances in sequencing have greatly outpaced experimental methods for determining a protein's structure and function. As a result, biologists increasingly rely on computational techniques to infer these properties of proteins from sequence information alone. We present a sequence classification framework that differs from the common SVM/kernel-based approach. We introduce a type of artificial neural network which we term the Subsequence Network (SN) that incorporates structural models over sequences in its lowest layer. These structural models, which we call Sequence Scoring Models (SSM), are similar to Hidden Markov Models and act as a mechanism to extract relevant features from sequences. In contrast to SVM/kernel methods, which only allow learning of linear discrimination weights, our feed-forward structure allows linear weights to be learned in conjunction with sequence-level features using standard optimization techniques.

14:25–14:50 **Nearly Exact Mining of Frequent Trees in Large Networks**
Ashraf M. Kibriya and Jan Ramon

Mining frequent patterns in a single network (graph) poses a number of challenges. Already only to match one path pattern to a network (upto subgraph isomorphism) is NP-complete. Matching algorithms that exist, become intractable even for reasonably small patterns, on networks which are large or have a high average degree. Based on recent advances in parameterized complexity theory, we propose a novel miner for rooted trees in networks. The miner, for a fixed parameter k (maximal pattern size), can mine all rooted trees with *delay* linear in the size of the network and only mildly exponential in the fixed parameter k (2^k). This allows us to mine tractably, rooted trees, in large networks such as the WWW or social networks. We establish the practical applicability of our miner, by presenting an experimental evaluation on both synthetic and real-world data.

14:50–15:15 **Reachability Analysis and Modeling of Dynamic Event Networks**
Kathy Macropol and Ambuj Singh

A wealth of graph data, from email and telephone graphs to Twitter networks, falls into the category of dynamic "event" networks. Edges in these networks represent brief events, and their analysis leads to multiple interesting and important topics, such as the prediction of road traffic or modeling of communication flow. In this paper, we analyze a novel new dynamic event graph property, the "Dynamic Reachability Set" (DRS), which characterizes reachability within graphs across time. We discover that DRS histograms of multiple real world dynamic event networks follow novel distribution patterns. From these patterns, we introduce a new generative dynamic graph model, DRS-Gen. DRS-Gen captures the dynamic graph properties of connectivity and reachability, as well as generates time values for its edges. To the best of our knowledge, DRS-Gen is the first such model which produces exact time values on edges, allowing us to understand simultaneity across multiple information flows.

15:15–15:40 **An Efficiently Computable Support Measure for Frequent Subgraph Pattern Mining**
Yuyi Wang and Jan Ramon

Graph support measures are functions measuring how frequently a given subgraph pattern occurs in a given database graph. An important class of support measures relies on overlap graphs. A major advantage of the overlap graph based approaches is that they combine anti-monotonicity with counting occurrences of a pattern which are independent according to certain criteria. However, existing overlap graph based support measures are expensive to compute.

In this paper, we propose a new support measure which is based on a new notion of independence. We show that our measure is the solution to a linear program which is usually sparse, and using interior point methods can be computed efficiently. We show experimentally that for large networks, in contrast to earlier overlap graph based proposals, pattern mining based on our support measure is feasible.

Room: Chem, LT2

Tue2B: Rankings and Recommendations

Chair: Carlos Soares

14:00–14:25 **A Live Comparison of Methods for Personalized Article Recommendation at Forbes.com**
Evan Kirshenbaum, George Forman and Michael Dugan

We present the results of a multi-phase study to optimize strategies for generating personalized article recommendations at the Forbes.com web site. In the first phase we compared the performance of a variety of recommendation methods on historical data. In the second phase we deployed a live system at Forbes.com for five months on a sample of 82,000 users, each randomly assigned to one of 20 methods. We analyze the live results both in terms of click-through rate (CTR) and user session lengths. The method with the best CTR was a hybrid of collaborative-filtering and a content-based method that leverages Wikipedia-based concept features, post-processed by a novel Bayesian remapping technique that we introduce. It both statistically significantly beat decayed popularity and increased CTR by 37%.

14:25–14:50 **Active Evaluation of Ranking Functions Based on Graded Relevance**
Christoph Sawade, Steffen Bickel, Timo von Oertzen, Tobias Scheffer and Niels Landwehr

Evaluating the quality of ranking functions is a core task in web search and other information retrieval domains. Because query distributions and item relevance change over time, ranking models often cannot be evaluated accurately on held-out training data. Instead, considerable effort is spent on manually labeling the relevance of query results for test queries in order to track ranking performance. We address the problem of estimating ranking performance as accurately as possible on a fixed labeling budget. Estimates are based on a set of most informative test queries selected by an active sampling distribution. Query labeling costs depend on the number of result items as well as item-specific attributes such as document length. We derive cost-optimal sampling distributions for the commonly used performance measures Discounted Cumulative Gain (DCG) and Expected Reciprocal Rank (ERR). Experiments on web search engine data illustrate significant reductions in labeling costs.

14:50–15:15 **Fast ALS-Based Tensor Factorization for Context-Aware Recommendation from Implicit Feedback**
Balázs Hidasi and Domonkos Tikk

Albeit the implicit feedback based recommendation problem—when only the user history is available but there are no ratings—is the most typical setting in real-world applications, it is much less researched than the explicit feedback case. State-of-the-art algorithms that are efficient on the explicit case cannot be straightforwardly transformed to the implicit case if scalability should be maintained. There are few implicit feedback benchmark datasets, therefore new ideas are usually experimented on explicit benchmarks. In this paper, we propose a generic context-aware implicit feedback recommender algorithm, coined iTALS. iTALS applies a fast, ALS-based tensor factorization learning method that scales linearly with the number of non-zero elements in the tensor. The method also allows us to incorporate various contextual information into the model while maintaining its computational efficiency. We present two context-aware implementation variants of iTALS. The first incorporates seasonality and enables to distinguish user behavior in different time intervals. The other views the user history as sequential information and has the ability to recognize usage pattern typical to certain group of items, e.g. to automatically tell apart product types that are typically purchased repetitively or once. Experiments performed on five implicit datasets (LastFM 1K, Grocery, VoD, and "implicitized" Netflix and MovieLens 10M) show that by integrating context-aware information with our factorization framework into the state-of-the-art implicit recommender algorithm the recommendation quality improves significantly.

15:15–15:40 **Probability Estimation for Multi-class Classification Based on Label Ranking**
Weiwei Cheng and Eyke Hüllermeier

We consider the problem of probability estimation in the setting of multi-class classification. While this problem has already been addressed in the literature, we tackle it from a novel perspective. Exploiting the close connection between probability estimation and ranking, our idea is to solve the former on the basis of the latter, taking advantage of recently developed methods for label ranking. More specifically, we argue that the Plackett-Luce ranking model is a very natural choice in this context, especially as it can be seen as a multinomial extension of the Bradley-Terry model. The latter provides the basis of pairwise coupling techniques, which arguably constitute the state-of-the-art in multi-class probability estimation. We explore the relationship between the pairwise and the ranking-based approach to probability estimation, both formally and empirically. Using synthetic and real-world data, we show that our method does not only enjoy nice theoretical properties, but is also competitive in terms of accuracy and efficiency.

Room: Chem, LT3

Wed2C: Online Learning and Data Streams

Chair: João Gama

14:00–14:25 **Adaptive Two-View Online Learning for Math Topic Classification**
Tam T. Nguyen, Kuiyu Chang and Siu Cheung Hui

Text categorization has been a popular research topic for years and has become more or less a practical technology. However, there exists little research on math topic classification. Math documents contain both textual data and math expressions. The text and math can be considered as two related but different views of a math document. The goal of online math topic classification is to automatically categorize a math document containing both mathematical expressions and textual content into an appropriate topic without the need for periodically retraining the classifier. To achieve this, it is essential to have a two-view online classification algorithm, which deals with the textual data view and the math expression view at the same time. In this paper, we propose a novel adaptive two-view online math document classifier based on the Passive Aggressive (PA) algorithm. The proposed approach is evaluated on real world math questions and answers from the Math Overflow question answering system. Compared to the baseline PA algorithm, our method's overall F-measure is improved by up to 3%. The improvement of our algorithm over the plain math expression view is almost 6%.

14:25–14:50 **BDUOL: Double Updating Online Learning on a Fixed Budget**
Peilin Zhao and Steven C.H. Hoi

Kernel-based online learning often exhibits promising empirical performance for various applications according to previous studies. However, it often suffers a main shortcoming, that is, the unbounded number of support vectors, making it unsuitable for handling large-scale datasets. In this paper, we investigate the problem of budget kernel-based online learning that aims to constrain the number of support vectors by a predefined budget when learning the kernel-based prediction function in the online learning process. Unlike the existing studies, we present a new framework of budget kernel-based online learning based on a recently proposed online learning method called "Double Updating Online Learning" (DUOL), which has shown state-of-the-art performance as compared with the other traditional kernel-based online learning algorithms. We analyze the theoretical underpinning of the proposed Budget Double Updating Online Learning (BDUOL) framework, and then propose several BDUOL algorithms by designing different budget maintenance strategies. We evaluate the empirical performance of the proposed BDUOL algorithms by comparing them with several well-known budget kernel-based online learning algorithms, in which encouraging results validate the efficacy of the proposed technique.

14:50–15:15 **Improved Counter Based Algorithms for Frequent Pairs Mining in Transactional Data Streams**
Konstantin Kutzkov

A straightforward approach to frequent pairs mining in transactional streams is to generate all pairs occurring in transactions and apply a frequent items mining algorithm to the resulting stream. The well-known counter based algorithms FREQUENT and SPACE-SAVING are known to achieve a very good approximation when the frequencies of the items in the stream adhere to a skewed distribution.

Motivated by observations on real datasets, we present a general technique for applying FREQUENT and SPACE-SAVING to transactional data streams for the case when the transactions considerably vary in their lengths. Despite of its simplicity, we show through extensive experiments that our approach is considerably more efficient and precise than the naïve application of FREQUENT and SPACE-SAVING.

15:15–15:40 **Mirror Descent for Metric Learning: A Unified Approach**
Gautam Kunapuli and Jude Shavlik

Most metric learning methods are characterized by diverse loss functions and projection methods, which naturally begs the question: is there a wider framework that can generalize many of these methods? In addition, ever persistent issues are those of scalability to large data sets and the question of kernelizability. We propose a unified approach to Mahalanobis metric learning: an online regularized metric learning algorithm based on the ideas of composite objective mirror descent (COMID). The metric learning problem is formulated as a regularized positive semi-definite matrix learning problem, whose update rules can be derived using the COMID framework. This approach aims to be scalable, kernelizable, and admissible to many different types of Bregman and loss functions, which allows for the tailoring of several different classes of algorithms. The most novel contribution is the use of the trace norm, which yields a sparse metric in its eigenspectrum, thus simultaneously performing feature selection along with metric learning.

Room: Chem, LT2

Wed2B: Large-Scale, Distributed and Parallel Mining and Learning I

Chair: Yizhao Ni

14:00–14:25 **Learning Compact Class Codes for Fast Inference in Large Multi Class Classification**
M. Cissé, T. Artières and Patrick Gallinari

We describe a new approach for classification with a very large number of classes where we assume some class similarity information is available, e.g. through a hierarchical organization. The proposed method learns a compact binary code using such an existing similarity information defined on classes. Binary classifiers are then trained using this code and decoding is performed using a simple nearest neighbor rule. This strategy, related to Error Correcting Output Codes methods, is shown to perform similarly or better than the standard and efficient one-vs-all approach, with much lower inference complexity.

14:25–14:50 **ParCube: Sparse Parallelizable Tensor Decompositions**
Evangelos E. Papalexakis, Christos Faloutsos and Nicholas D. Sidiropoulos

How can we efficiently decompose a tensor into sparse factors, when the data does not fit in memory? Tensor decompositions have gained a steadily increasing popularity in data mining applications, however the current state-of-art decomposition algorithms operate on main memory and do not scale to truly large datasets. In this work, we propose ParCube, a new and highly parallelizable method for speeding up tensor decompositions that is well-suited to producing sparse approximations. Experiments with even moderately large data indicate over 90% sparser outputs and 14 times faster execution, with approximation error close to the current state of the art irrespective of computation and memory requirements. We provide theoretical guarantees for the algorithm's correctness and we experimentally validate our claims through extensive experiments, including four different real world datasets (ENRON, LBNL, FACEBOOK and NELL), demonstrating its effectiveness for data mining practitioners. In particular, we are the first to analyze the very large NELL dataset using a sparse tensor decomposition, demonstrating that ParCube enables us to handle effectively and efficiently very large datasets.

14:50–15:15 **Stochastic Coordinate Descent Methods for Regularized Smooth and Nonsmooth Losses**
Qing Tao, Kang Kong, Dejun Chu and Gaowei Wu

Stochastic Coordinate Descent (SCD) methods are among the first optimization schemes suggested for efficiently solving large scale problems. However, until now, there exists a gap between the convergence rate analysis and practical SCD algorithms for general smooth losses and there is no primal SCD algorithm for nonsmooth losses. In this paper, we discuss these issues using the recently developed structural optimization techniques. In particular, we first present a principled and practical SCD algorithm for regularized smooth losses, in which the one-variable subproblem is solved using the proximal gradient method and the adaptive componentwise Lipschitz constant is obtained employing the line search strategy. When the loss is nonsmooth, we present a novel SCD algorithm, in which the one-variable subproblem is solved using the dual averaging method. We show that our algorithms exploit the regularization structure and achieve several optimal convergence rates that are standard in the literature. The experiments demonstrate the expected efficiency of our SCD algorithms in both smooth and nonsmooth cases.

15:15–15:40 **Sublinear Algorithms for Penalized Logistic Regression in Massive Datasets**
Haoruo Peng, Zhengyu Wang, Edward Y. Chang, Shuchang Zhou and Zhihua Zhang

Penalized logistic regression (PLR) is a widely used supervised learning model. In this paper, we consider its applications in large-scale data problems and resort to a stochastic primal-dual approach for solving PLR. In particular, we employ a random sampling technique in the primal step and a multiplicative weights method in the dual step. This technique leads to an optimization method with sublinear dependency on both the volume and dimensionality of training data. We develop concrete algorithms for PLR with ℓ_2-norm and ℓ_1-norm penalties, respectively. Experimental results over several large-scale and high-dimensional datasets demonstrate both efficiency and accuracy of our algorithms.

Room: Chem, LT3

Tue2C: Ensemble Methods

Chair: Herna Viktor

14:00–14:25 **Boosting Nearest Neighbors for the Efficient Estimation of Posteriors**
Roberto D'Ambrosio, Richard Nock, Wafa Bel Haj Ali, Frank Nielsen and Michel Barlaud

It is an admitted fact that mainstream boosting algorithms like AdaBoost do not perform well to estimate class conditional probabilities. In this paper, we analyze, in the light of this problem, a recent algorithm, UNN, which leverages nearest neighbors while minimizing a convex loss. Our contribution is threefold. First, we show that there exists a subclass of surrogate losses, elsewhere called balanced, whose minimization brings simple and statistically efficient estimators for Bayes posteriors. Second, we show *explicit* convergence rates towards these estimators for UNN, for any such surrogate loss, under a Weak Learning Assumption which parallels that of classical boosting results. Third and last, we provide experiments and comparisons on synthetic and real datasets, including the challenging SUN computer vision database. Results clearly display that boosting nearest neighbors may provide highly accurate estimators, sometimes more than a hundred times more accurate than those of other contenders like support vector machines.

14:25–14:50 **Diversity Regularized Ensemble Pruning**
Nan Li, Yang Yu and Zhi-Hua Zhou

Diversity among individual classifiers is recognized to play a key role in ensemble, however, few theoretical properties are known for classification. In this paper, by focusing on the popular ensemble pruning setting (i.e., combining classifier by *voting* and measuring diversity in *pairwise* manner), we present a theoretical study on the effect of diversity on the generalization performance of voting in the PAC-learning framework. It is disclosed that the diversity is closely-related to the hypothesis space complexity, and encouraging diversity can be regarded to apply regularization on ensemble methods. Guided by this analysis, we apply explicit diversity regularization to ensemble pruning, and propose the *Diversity Regularized Ensemble Pruning* (DREP) method. Experimental results show the effectiveness of DREP.

14:50–15:15 **Ensembles on Random Patches**
Gilles Louppe and Pierre Geurts

In this paper, we consider supervised learning under the assumption that the available memory is small compared to the dataset size. This general framework is relevant in the context of big data, distributed databases and embedded systems. We investigate a very simple, yet effective, ensemble framework that builds each individual model of the ensemble from a random patch of data obtained by drawing random subsets of *both* instances and features from the whole dataset. We carry out an extensive and systematic evaluation of this method on 29 datasets, using decision tree-based estimators. With respect to popular ensemble methods, these experiments show that the proposed method provides on par performance in terms of accuracy while simultaneously lowering the memory needs, and attains significantly better performance when memory is severely constrained.

15:15–15:40 **Multi-Task Boosting by Exploiting Task Relationships**
Yu Zhang and Dit-Yan Yeung

Multi-task learning aims at improving the performance of one learning task with the help of other related tasks. It is particularly useful when each task has very limited labeled data. A central issue in multi-task learning is to learn and exploit the relationships between tasks. In this paper, we generalize boosting to the multi-task learning setting and propose a method called multi-task boosting (MTBoost). Different tasks in MTBoost share the same base learners but with different weights which are related to the estimated task relationships in each iteration. In MTBoost, unlike ordinary boosting methods, the base learners, weights and task covariances are learned together in an integrated fashion using an alternating optimization procedure. We conduct theoretical analysis on the convergence of MTBoost and also empirical analysis comparing it with several related methods.

Room: Chem, LT1

Tue3A: Dimensionality Reduction, Feature Selection and Extraction

Chair: Katharina Morik

16:10–16:35 **Embedding Monte Carlo Search of Features in Tree-Based Ensemble Methods**
Francis Maes, Pierre Geurts and Louis Wehenkel

Feature generation is the problem of automatically constructing good features for a given target learning problem. While most feature generation algorithms belong either to the *filter* or to the *wrapper* approach, this paper focuses on *embedded* feature generation. We propose a general scheme to embed feature generation in a wide range of tree-based learning algorithms, including single decision trees, random forests and tree boosting. It is based on the formalization of feature construction as a sequential decision making problem addressed by a tractable Monte Carlo search algorithm coupled with node splitting. This leads to fast algorithms that are applicable to large-scale problems. We empirically analyze the performances of these tree-based learners combined or not with the feature generation capability on several standard datasets.

16:35–17:00 **Hypergraph Spectra for Semi-supervised Feature Selection**
Zhihong Zhang, Edwin R. Hancock and Xiao Bai

In many data analysis tasks, one is often confronted with the problem of selecting features from very high dimensional data. Most existing feature selection methods focus on ranking individual features based on a utility criterion, and select the optimal feature set in a greedy manner. However, the feature combinations found in this way do not give optimal classification performance, since they neglect the correlations among features. While the labeled data required by supervised feature selection can be scarce, there is usually no shortage of unlabeled data. In this paper, we propose a novel hypergraph based semi-supervised feature selection algorithm to select relevant features using both labeled and unlabeled data. There are two main contributions in this paper. The first is that by incorporating multidimensional interaction information (MII) for higher order similarities measure, we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Thus, the structural information latent in the data can be more effectively modeled. Secondly, we derive a hypergraph subspace learning view of feature selection which casting the feature discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard face data-sets.

17:00–17:25 **Learning Neighborhoods for Metric Learning**
Jun Wang, Adam Woznica and Alexandros Kalousis

Metric learning methods have been shown to perform well on different learning tasks. Many of them rely on target neighborhood relationships that are computed in the original feature space and remain fixed throughout learning. As a result, the learned metric reflects the original neighborhood relations. We propose a novel formulation of the metric learning problem in which, in addition to the metric, the target neighborhood relations are also learned in a two-step iterative approach. The new formulation can be seen as a generalization of many existing metric learning methods. The formulation includes a target neighbor assignment rule that assigns different numbers of neighbors to instances according to their quality; 'high quality' instances get more neighbors. We experiment with two of its instantiations that correspond to the metric learning algorithms LMNN and MCML and compare it to other metric learning methods on a number of datasets. The experimental results show state-of-the-art performance and provide evidence that learning the neighborhood relations does improve predictive performance.

17:25–17:50 **PCA, Eigenvector Localization and Clustering for Side-Channel Attacks on Cryptographic Hardware Devices**
Dimitrios Mavroeidis, Lejla Batina, Twan van Laarhoven and Elena Marchiori

Spectral methods, ranging from traditional Principal Components Analysis to modern Laplacian matrix factorization, have proven to be a valuable tool for a wide range of diverse data mining applications. Commonly these methods are stated as optimization problems and employ the extremal (maximal or minimal) eigenvectors of a certain input matrix for deriving the appropriate statistical inferences. Interestingly, recent studies have questioned this "modus operandi" and revealed that useful information may also be present within low-order eigenvectors whose mass is concentrated (localized) in a small part of their indexes. An application context where localized low-order eigenvectors have been successfully employed is "Differential Power Analysis" (DPA). DPA is a well studied side-channel attack on cryptographic hardware devices (such as smart cards) that employs statistical analysis of the device's power consumption in order to retrieve the secret key of the cryptographic algorithm. In this work we propose a data mining

Room: Chem, LT1

Wed2A: Social Network Mining I

Chair: Jan Ramon

14:00–14:25 **Discovering Links among Social Networks**
Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino

Distinct social networks are interconnected via bridge users, who play thus a key role when crossing information is investigated in the context of Social Internetworking analysis. Unfortunately, not always users make their role of *bridge* explicit by specifying the so-called `me` edge (i.e., the edge connecting the accounts of the same user in two distinct social networks), missing thus a potentially very useful information. As a consequence, discovering missing `me` edges is an important problem to face in this context yet not so far investigated. In this paper, we propose a common-neighbors approach to detecting missing `me` edges, which returns good results in real life settings. Indeed, an experimental campaign shows both that the state-of-the-art common-neighbors approaches cannot be effectively applied to our problem and, conversely, that our approach returns precise and complete results.

14:25–14:50 **Efficient Bi-objective Team Formation in Social Networks**
Mehdi Kargar Aijun An, and Morteza Zihayat

We tackle the problem of finding a team of experts from a social network to complete a project that requires a set of skills. The social network is modeled as a graph. A node in the graph represents an expert and has a weight representing the monetary cost for using the expert service. Two nodes in the graph can be connected and the weight on the edge represents the communication cost between the two corresponding experts. Given a project, our objective is to find a team of experts that covers all the required skills and also minimizes the communication cost as well as the personnel cost of the project. To minimize both of the objectives, we define a new combined cost function which is based on the linear combination of the objectives (i.e. communication and personnel costs). We show that the problem of minimizing the combined cost function is an NP-hard problem. Thus, one approximation algorithm is proposed to solve the problem. The proposed approximation algorithm is bounded and the approximation ratio of the algorithm is proved in the paper. Three heuristic algorithms based on different intuitions are also proposed for solving the problem. Extensive experiments on real datasets demonstrate the effectiveness and scalability of the proposed algorithms.

14:50–15:15 **Feature-Enhanced Probabilistic Models for Diffusion Network Inference**
Liaoruo Wang, Stefano Ermon and John E. Hopcroft

Cascading processes, such as disease contagion, viral marketing, and information diffusion, are a pervasive phenomenon in many types of networks. The problem of devising intervention strategies to facilitate or inhibit such processes has recently received considerable attention. However, a major challenge is that the underlying network is often unknown. In this paper, we revisit the problem of inferring latent network structure given observations from a diffusion process, such as the spread of trending topics in social media. We define a family of novel probabilistic models that can explain recurrent cascading behavior, and take into account not only the time differences between events but also a richer set of additional features. We show that MAP inference is tractable and can therefore scale to very large real-world networks. Further, we demonstrate the effectiveness of our approach by inferring the underlying network structure of a subset of the popular Twitter following network by analyzing the topics of a large number of messages posted by users over a 10-month period. Experimental results show that our models accurately recover the links of the Twitter network, and significantly improve the performance over previous models based entirely on time.

15:15–15:40 **Influence Spread in Large-Scale Social Networks – A Belief Propagation Approach**
Huy Nguyen and Rong Zheng

Influence maximization is the problem of finding a small set of seed nodes in a social network that maximizes the spread of influence under a certain diffusion model. The Greedy algorithm for influence maximization first proposed by Kempe, later improved by Leskovec suffers from two sources of computational deficiency: 1) the need to evaluate many candidate nodes before selecting a new seed in each round, and 2) the calculation of the influence spread of any seed set relies on Monte-Carlo simulations. In this work, we tackle both problems by devising efficient algorithms to compute influence spread and determine the best candidate for seed selection. The fundamental insight behind the proposed algorithms is the linkage between influence spread determination and belief propagation on a directed acyclic graph (DAG). Experiments using real-world social network graphs with scales ranging from thousands to millions of edges demonstrate the superior performance of the proposed algorithms with moderate computation costs.

Room: Chem, LT3

Wed1C: Spatial and Geographical Data Mining

Chair: Willem Waegeman

10:30–10:55 **Inferring Geographic Coincidence in Ephemeral Social Networks**
Honglei Zhuang, Alvin Chin, Sen Wu, Wei Wang, Xia Wang and Jie Tang

We study users' behavioral patterns in ephemeral social networks, which are temporarily built based on events such as conferences. From the data distribution and social theory perspectives, we found several interesting patterns. For example, the duration of two random persons staying at the same place and at the same time obeys a two-stage power-law distribution. We develop a framework to infer the likelihood of two users to meet together, and we apply the framework to two mobile social networks: UbiComp and Reality. The former is formed by researchers attending UbiComp 2011 and the latter is a network of students published by MIT. On both networks, we validate the proposed predictive framework, which significantly improve the accuracy for predicting geographic coincidence by comparing with two baseline methods.

10:55–11:20 **Socioscope: Spatio-temporal Signal Recovery from Social Media**
Jun-Ming Xu, Aniruddha Bhargava, Robert Nowak and Xiaojin Zhu

Many real-world phenomena can be represented by a spatio-temporal signal: where, when, and how much. Social media is a tantalizing data source for those who wish to monitor such signals. Unlike most prior work, we assume that the target phenomenon is known and we are given a method to count its occurrences in social media. However, counting is plagued by sample bias, incomplete data, and, paradoxically, data scarcity – issues inadequately addressed by prior work. We formulate signal recovery as a Poisson point process estimation problem. We explicitly incorporate human population bias, time delays and spatial distortions, and spatio-temporal regularization into the model to address the noisy count issues. We present an efficient optimization algorithm and discuss its theoretical properties. We show that our model is more accurate than commonly-used baselines. Finally, we present a case study on wildlife roadkill monitoring, where our model produces qualitatively convincing results.

11:20–11:45 **Location Affiliation Networks: Bonding Social and Spatial Information**
Konstantinos Pelechrinis and Prashant Krishnamurthy

Location-based social networks (LBSNs) have recently attracted a lot of attention due to the number of novel services they can offer. Prior work on analysis of LBSNs has mainly focused on the social part of these systems. Even though it is important to know how different the structure of the social graph of an LBSN is as compared to the friendship-based social networks (SNs), it raises the interesting question of what kinds of linkages exist between locations and friendships. The main problem we are investigating is to identify such connections between the social and the spatial planes of an LBSN. In particular, in this paper we focus on answering the following general question "What are the bonds between the social and spatial information in an LBSN and what are the metrics that can reveal them?" In order to tackle this problem, we employ the idea of *affiliation networks*. Analyzing a dataset from a specific LBSN (Gowalla), we make two main interesting observations; (i) the social network exhibits *signs of homophily* with regards to the "places/venues" visited by the users, and (ii) the "nature" of the visited venues that are common to users is powerful and informative in revealing the social/spatial linkages. We further show that the "entropy" (or diversity) of a venue can be used to better connect spatial information with the existing social relations. The entropy records the diversity of a venue and requires only location history of users (it does not need temporal history). Finally, we provide a simple application of our findings for predicting existing friendship relations based on users' historic spatial information. We show that even with simple unsupervised learning models we can achieve significant improvement in prediction when we consider features that capture the "nature" of the venue as compared to the case where only apparent properties of the location history are used (e.g., number of common visits).

11:45–12:10 **Pedestrian Quantity Estimation with Trajectory Patterns**
Thomas Liebig, Zhao Xu, Michael May and Stefan Wrobel

In street-based mobility mining, traffic volume estimation receives increasing attention as it provides important applications such as emergency support systems, quality-of-service evaluation and billboard placement. In many real world scenarios, empirical measurements are usually sparse due to some constraints. On the other hand, pedestrians generally show some movement preferences, especially in closed environments, e.g., train stations. We propose a Gaussian process regression based method for traffic volume estimation, which incorporates topological information and prior knowledge on preferred trajectories with a trajectory pattern kernel. Our approach also enables effectively finding most informative sensor placements. We evaluate our method with synthetic German train station pedestrian data and real-world episodic movement data from the zoo of Duisburg. The empirical analysis demonstrates that incorporating trajectory patterns can largely improve the traffic prediction accuracy, especially when traffic networks are sparsely monitored.

(clustering) formulation of the DPA process and also provide a theoretical model that justifies and explains the utility of low-order eigenvectors. In our data mining formulation, we consider that the key-relevant information is modelled as a "low-signal" pattern that is embedded in a "high-noise" dataset. In this respect our results generalize beyond DPA and are applicable to analogous low-signal, hidden pattern problems. The experimental results using power trace measurements from a programmable smart card, verify our approach empirically.

Room: Chem, LT2

Tue3B: Multi-Relational Mining and Learning

Chair: Pauli Miettinen

16:10–16:35 **Author Name Disambiguation Using a New Categorical Distribution Similarity**
Shaohua Li, Gao Cong and Chunyan Miao

Author name ambiguity has been a long-standing problem which impairs the accuracy of publication retrieval and bibliometric methods. Most of the existing disambiguation methods are built on similarity measures, e.g., "Jaccard Coefficient", between two sets of papers to be disambiguated, each set represented by a set of categorical features, e.g., coauthors and published venues[1]. Such measures perform bad when the two sets are small, which is typical in Author Name Disambiguation. In this paper, we propose a novel categorical set similarity measure. We model an author's preference, e.g., to venues, using a categorical distribution, and derive a likelihood ratio to estimate the likelihood that the two sets are drawn from the same distribution. This likelihood ratio is used as the similarity measure to decide whether two sets belong to the same author. This measure is mathematically principled and verified to perform well even when the cardinalities of the two compared sets are small. Additionally, we propose a new method to estimate the number of distinct authors for a given name based on the name statistics extracted from a digital library. Experiment shows that our method significantly outperforms a baseline method, a widely used benchmark method, and a real system.

16:35–17:00 **Lifted Online Training of Relational Models with Stochastic Gradient Methods**
Babak Ahmadi, Kristian Kersting and Sriraam Natarajan

Lifted inference approaches have rendered large, previously intractable probabilistic inference problems quickly solvable by employing symmetries to handle whole sets of indistinguishable random variables. Still, in many if not most situations training relational models will not benefit from lifting: symmetries within models easily break since variables become correlated by virtue of depending asymmetrically on evidence. An appealing idea for such situations is to train and recombine local models. This breaks long-range dependencies and allows to exploit lifting within and across the local training tasks. Moreover, it naturally paves the way for online training for relational models. Specifically, we develop the first lifted stochastic gradient optimization method with gain vector adaptation, which processes each lifted piece one after the other. On several datasets, the resulting optimizer converges to the same quality solution over an order of magnitude faster, simply because unlike batch training it starts optimizing long before having seen the entire mega-example even once.

17:00–17:25 **Scalable Relation Prediction Exploiting Both Intrarelational Correlation and Contextual Information**
Xueyan Jiang, Volker Tresp, Yi Huang, Maximilian Nickel and Hans-Peter Kriegel

We consider the problem of predicting instantiated binary relations in a multi-relational setting and exploit both intrarelational correlations and contextual information. For the modular combination we discuss simple heuristics, additive models and an approach that can be motivated from a hierarchical Bayesian perspective. In the concrete examples we consider models that exploit contextual information both from the database and from contextual unstructured information, e.g., information extracted from textual documents describing the involved entities. By using low-rank approximations in the context models, the models perform latent semantic analyses and can generalize across specific terms, i.e., the model might use similar latent representations for semantically related terms. All the approaches we are considering have unique solutions. They can exploit sparse matrix algebra and are thus highly scalable and can easily be generalized to new entities. We evaluate the effectiveness of nonlinear interaction terms and reduce the number of terms by applying feature selection. For the optimization of the context model we use an alternating least squares approach. We experimentally analyze scalability. We validate our approach using two synthetic data sets and using two data sets derived from the Linked Open Data (LOD) cloud.

17:25–17:50 **Relational Differential Prediction**
Houssam Nassif, Vítor Santos Costa, Elizabeth S. Burnside and David Page

A typical classification problem involves building a model to correctly segregate instances of two or more classes. Such a model exhibits differential prediction with respect to given data subsets when its performance is significantly different over these subsets. Driven by a mammography application, we aim at learning rules that predict breast cancer stage while maximizing differential prediction over age-stratified data. In this work, we present the first multi-relational differential prediction (aka uplift modeling) system, and propose three different approaches to learn differential predictive rules within the Inductive Logic Programming framework. We first test and validate our methods on synthetic data, then apply them on a mammography dataset for breast cancer stage differential prediction rule discovery. We mine a novel rule linking calcification to *in situ* breast cancer in older women.

[1] Venues here refer to the journal or conference, such as *J. ACM* or *SIGIR*.

Room: Chem, LT2

Wed1B: Time Series and Temporal Data Mining

Chair: Indrė Žliobaitė

10:30–10:55 **Community Trend Outlier Detection Using Soft Temporal Pattern Mining**
Manish Gupta, Jing Gao, Yizhou Sun, and Jiawei Han

Numerous applications, such as bank transactions, road traffic, and news feeds, generate temporal datasets, in which data evolves continuously. To understand the temporal behavior and characteristics of the dataset and its elements, we need effective tools that can capture evolution of the objects. In this paper, we propose a novel and important problem in evolution behavior discovery. Given a series of snapshots of a temporal dataset, each of which consists of evolving communities, our goal is to find objects which evolve in a dramatically different way compared with the other community members. We define such objects as *community trend outliers*. It is a challenging problem as evolutionary patterns are hidden deeply in noisy evolving datasets and thus it is difficult to distinguish anomalous objects from normal ones. We propose an effective two-step procedure to detect community trend outliers. We first model the normal evolutionary behavior of communities across time using soft patterns discovered from the dataset. In the second step, we propose effective measures to evaluate chances of an object deviating from the normal evolutionary patterns. Experimental results on both synthetic and real datasets show that the proposed approach is highly effective in discovering interesting community trend outliers.

10:55–11:20 **Data Structures for Detecting Rare Variations in Time Series**
Caio Valentim, Eduardo S. Laber and David Sotelo

In this paper we study, from both a theoretical and an experimental perspective, algorithms and data structures to process queries that help in the detection of rare variations over time intervals that occur in time series. Our research is strongly motivated by applications in financial domain.

11:20–11:45 **Invariant Time-Series Classification**
Josif Grabocka, Alexandros Nanopoulos and Lars Schmidt-Thieme

Time-series classification is a field of machine learning that has attracted considerable focus during the recent decades. The large number of time-series application areas ranges from medical diagnosis up to financial econometrics. Support Vector Machines (SVMs) are reported to perform non-optimally in the domain of time series, because they suffer detecting similarities in the lack of abundant training instances. In this study we present a novel time-series transformation method which significantly improves the performance of SVMs. Our novel transformation method is used to enlarge the training set through creating new transformed instances from the support vector instances. The new transformed instances encapsulate the necessary intra-class variations required to redefine the maximum margin decision boundary. The proposed transformation method utilizes the variance distributions from the intra-class warping maps to build transformation fields, which are applied to series instances using the Moving Least Squares algorithm. Extensive experimentations on 35 time series datasets demonstrate the superiority of the proposed method compared to both the Dynamic Time Warping version of the Nearest Neighbor and the SVMs classifiers, outperforming them in the majority of the experiments.

11:45–12:10 **Learning Bi-clustered Vector Autoregressive Models**
Tzu-Kuo Huang and Jeff Schneider

Vector Auto-regressive (VAR) models are useful for analyzing temporal dependencies among multivariate time series, known as *Granger causality*. There exist methods for learning sparse VAR models, leading directly to causal networks among the variables of interest. Another useful type of analysis comes from clustering methods, which summarize multiple time series by putting them into groups. We develop a methodology that integrates both types of analyses, motivated by the intuition that Granger causal relations in real-world time series may exhibit some clustering structure, in which case the estimation of both should be carried out together. Our methodology combines sparse learning and a nonparametric *bi-clustered* prior over the VAR model, conducting full Bayesian inference via blocked Gibbs sampling. Experiments on simulated and real data demonstrate improvements in both model estimation and clustering quality over standard alternatives, and in particular biologically more meaningful clusters in a T-cell activation gene expression time series dataset than those by other methods.

Room: Chem, LT1
Wed1A: Distance-Based Methods and Kernels — Chair: Stefan Rüping

10:30–10:55 **Classifying Stem Cell Differentiation Images by Information Distance**
Xianglilan Zhang, Hongnan Wang, Tony J. Collins, Zhigang Luo and Ming Li

The ability of stem cells holds great potential for drug discovery and cell replacement therapy. To realize this potential, effective high content screening for drug candidates is required. Analysis of images from high content screening typically requires DNA staining to identify cell nuclei to do cell segmentation before feature extraction and classification. However, DNA staining has negative effects on cell growth, and segmentation algorithms err when compound treatments cause nuclear or cell swelling/shrinkage. In this paper, we introduced a novel Information Distance Classification (IDC) method, requiring no segmentation or feature extraction; hence no DNA staining is needed. In classifying 480 candidate compounds that may be used to stimulate stem cell differentiation, the proposed IDC method was demonstrated to achieve a 3% higher F_1 score than conventional analysis. As far as we know, this is the first work to apply information distance in high content screening.

10:55–11:20 **Distance Metric Learning Revisited**
Qiong Cao, Yiming Ying and Peng Li

The success of many machine learning algorithms (e.g. the nearest neighborhood classification and k-means clustering) depends on the representation of the data as elements in a metric space. Learning an appropriate distance metric from data is usually superior to the default Euclidean distance. In this paper, we revisit the original model proposed by Xing et al. and propose a general formulation of learning a Mahalanobis distance from data. We prove that this novel formulation is equivalent to a convex optimization problem over the spectrahedron. Then, a gradient-based optimization algorithm is proposed to obtain the optimal solution which only needs the computation of the largest eigenvalue of a matrix per iteration. Finally, experiments on various UCI datasets and a benchmark face verification dataset called Labeled Faces in the Wild (LFW) demonstrate that the proposed method compares competitively to those state-of-the-art methods.

11:20–11:45 **Geodesic Analysis on the Gaussian RKHS Hypersphere**
Nicolas Courty, Thomas Burger and Pierre-François Marteau

Using kernels to embed non linear data into high dimensional spaces where linear analysis is possible has become utterly classical. In the case of the Gaussian kernel however, data are distributed on a hypersphere in the corresponding Reproducing Kernel Hilbert Space (RKHS). Inspired by previous works in non-linear statistics, this article investigates the use of dedicated tools to take into account this particular geometry. Within this geometrical interpretation of the kernel theory, Riemannian distances are preferred over Euclidean distances. It is shown that this amounts to consider a new kernel and its corresponding RKHS. Experiments on real publicly available datasets show the possible benefits of the method on clustering tasks, notably through the definition of a new variant of kernel k-means on the hypersphere. Classification problems are also considered in a classwise setting. In both cases, the results show improvements over standard techniques.

11:45–12:10 **The Bitvector Machine: A Fast and Robust Machine Learning Algorithm for Non-linear Problems**
Stefan Edelkamp and Martin Stommel

In this paper we present and evaluate a simple but effective machine learning algorithm that we call *Bitvector Machine*: Feature vectors are partitioned along component-wise quantiles and converted into bitvectors that are learned. It is shown that the method is efficient in both training and classification. The effectiveness of the method is analysed theoretically for best and worst-case scenarios. Experiments on high-dimensional synthetic and real world data show a huge speed boost compared to Support Vector Machines with RBF kernel. By tabulating kernel functions, computing medians in linear-time, and exploiting modern processor technology for advanced bitvector operations, we achieve a speed-up of 32 for classification and 48 for kernel evaluation compared to the popular LIBSVM. Although the method does not generally outperform a SVM with RBF kernel it achieves a high classification accuracy and has qualitative advantages over the linear classifier.

Room: Chem, LT3
Tue3C: Semi-Supervised and Transductive Learning — Chair: Florence d'Alché

16:10–16:35 **Bidirectional Semi-supervised Learning with Graphs**
Tomoharu Iwata and Kevin Duh

We present a machine learning task, which we call bidirectional semi-supervised learning, where label-only samples are given as well as labeled and unlabeled samples. A label-only sample contains the label information of the sample but not the feature information. Then, we propose a simple and effective graph-based method for bidirectional semi-supervised learning in multi-label classification. The proposed method assumes that correlated classes are likely to have the same labels among the similar samples. First, we construct a graph that represents similarities between samples using labeled and unlabeled samples in the same way with graph-based semi-supervised methods. Second, we construct another graph using labeled and label-only samples by connecting classes that are likely to co-occur, which represents correlations between classes. Then, we estimate labels of unlabeled samples by propagating labels over these two graphs. We can find a closed-form global solution for the label propagation by using matrix algebra. We demonstrate the effectiveness of the proposed method over supervised and semi-supervised learning methods with experiments using synthetic and multi-label text data sets.

16:35–17:00 **Graph-Based Transduction with Confidence**
Matan Orbach and Koby Crammer

We present a new multi-class graph-based transduction algorithm. Examples are associated with vertices in an undirected weighted graph and edge weights correspond to a similarity measure between examples. Typical algorithms in such a setting perform label propagation between neighbours, ignoring the quality, or estimated quality, in the labeling of various nodes. We introduce an additional quantity of confidence in label assignments, and learn them jointly with the weights, while using them to dynamically tune the influence of each vertex on its neighbours. We cast learning as a convex optimization problem, and derive an efficient iterative algorithm for solving it. Empirical evaluations on seven NLP data sets demonstrate our algorithm improves over other state-of-the-art graph-based transduction algorithms.

17:00–17:25 **Maximum Consistency Preferential Random Walks**
Deguang Kong and Chris Ding

Random walk plays a significant role in computer science. The popular PageRank algorithm uses random walk. Personalized random walks force random walk to "personalized views" of the graph according to users' preferences. In this paper, we show the close relations between different preferential random walks and label propagation methods used in semi-supervised learning. We further present a maximum consistency algorithm on these preferential random walk/label propagation methods to ensure maximum consistency from labeled data to unlabeled data. Extensive experimental results on 9 datasets provide performance comparisons of different preferential random walks/label propagation methods. They also indicate that the proposed maximum consistency algorithm clearly improves the classification accuracy over existing methods.

17:25–17:50 **Semi-supervised Multi-label Classification: A Simultaneous Large-Margin, Subspace Learning Approach**
Yuhong Guo and Dale Schuurmans

Labeled data is often sparse in common learning scenarios, either because it is too time consuming or too expensive to obtain, while unlabeled data is almost always plentiful. This asymmetry is exacerbated in *multi-label* learning, where the labeling process is more complex than in the single label case. Although it is important to consider *semi-supervised* methods for multi-label learning, as it is in other learning scenarios, surprisingly, few proposals have been investigated for this particular problem. In this paper, we present a new semi-supervised multi-label learning method that combines large-margin multi-label classification with unsupervised subspace learning. We propose an algorithm that learns a subspace representation of the labeled and unlabeled inputs, while simultaneously training a supervised large-margin multi-label classifier on the labeled portion. Although joint training of these two interacting components might appear intractable, we exploit recent developments in induced matrix norm optimization to show that these two problems can be solved jointly, globally and efficiently. In particular, we develop an efficient training procedure based on subgradient search and a simple coordinate descent strategy. An experimental evaluation demonstrates that semi-supervised subspace learning can improve the performance of corresponding supervised multi-label learning methods.

Room: Chem, LT4

Tue3D: Demo Spotlights

Chair: Bettina Berendt and Myra Spiliopoulou

16:10–16:20 **VIKAMINE – Open-Source Subgroup Discovery, Pattern Mining, and Analytics**
Martin Atzmueller and Florian Lemmerich

This paper presents an overview on the VIKAMINE[2] system for subgroup discovery, pattern mining and analytics. As of VIKAMINE version 2, it is implemented as rich-client platform (RCP) application, based on the Eclipse[3] framework. This provides for a highly-configurable environment, and allows modular extensions using plugins. We present the system, briefly discuss exemplary plugins, and provide a sketch of successful applications.

16:20–16:30 **Association Rule Mining Following the Web Search Paradigm**
Radek Škrabal, Milan Šimůnek, Stanislav Vojíř, Andrej Hazucha, Tomáš Marek, David Chudán and Tomáš Kliegr

I:ZI Miner (`sewebar.vse.cz/izi-miner`) is an association rule mining system with a user interface resembling a search engine. It brings to the web the notion of interactive pattern mining introduced by the MIME framework at ECML'11 and KDD'11. In comparison with MIME, I:ZI Miner discovers multi-valued attributes, supports the full range of logical connectives and 19 interest measures. A relevance feedback module is used to filter the rules based on previous user interactions.

16:30–16:40 ***OutRules*: A Framework for Outlier Descriptions in Multiple Context Spaces**
Emmanuel Müller, Fabian Keller, Sebastian Blanc and Klemens Böhm

Analyzing exceptional objects is an important mining task. It includes the identification of outliers but also the description of outlier properties in contrast to regular objects. However, existing *detection* approaches miss to provide important *descriptions* that allow human understanding of outlier reasons. In this work we present *OutRules*, a framework for outlier descriptions that enable an easy understanding of multiple outlier reasons in different contexts. We introduce outlier rules as a novel outlier description model. A rule illustrates the deviation of an outlier in contrast to its context that is considered to be normal. Our framework highlights the practical use of outlier rules and provides the basis for future development of outlier description models.

16:40–16:50 **Knowledge Discovery through Symbolic Regression with HeuristicLab**
Gabriel Kronberger, Stefan Wagner, Michael Kommenda, Andreas Beham, Andreas Scheibenpflug and Michael Affenzeller

This contribution describes how symbolic regression can be used for knowledge discovery with the open-source software HeuristicLab. HeuristicLab includes a large set of algorithms and problems for combinatorial optimization and for regression and classification, including symbolic regression with genetic programming. It provides a rich GUI to analyze and compare algorithms and identified models. This contribution mainly focuses on specific aspects of symbolic regression that are unique to HeuristicLab, in particular, the identification of relevant variables and model simplification.

16:50–17:00 **An Aspect-Lexicon Creation and Evaluation Tool for Sentiment Analysis Researchers**
Mus'ab Husaini, Ahmet Koçyiğit, Dilek Tapucu, Berrin Yanikoglu and Yücel Saygın

In this demo paper, we present SARE, a modular and extendable semi-automatic system that *1*) assists researchers in building gold-standard lexicons and evaluating their lexicon extraction algorithms; and *2*) provides a general and extendable sentiment analysis environment to help researchers analyze the behavior and errors of a core sentiment analysis engine using a particular lexicon.

17:00–17:10 **ASV Monitor: Creating Comparability of Machine Learning Methods for Content Analysis**
Andreas Niekler, Patrick Jähnichen and Gerhard Heyer

In this demonstration paper we present an application to compare and evaluate machine learning methods used for natural language processing within a content analysis framework. Our aim is to provide an example set of possible machine learning results for different inputs to increase the acceptance of using machine learning in settings that originally rely on manual treatment. We will demonstrate the possibility to compare machine learning algorithms regarding the outcome of the implemented approaches. The application allows the user to evaluate the benefit of using machine learning algorithms for content analysis by a visual comparison of their results.

[2] `http://www.vikamine.org`
[3] `http://www.eclipse.org`

Room: Chem, LT1

Wed3A: Data Mining Process

Chair: Johannes Fürnkranz

16:10–16:35 **A Framework for Evaluating the Smoothness of Data-Mining Results**
Gaurav Misra, Behzad Golshan and Evimaria Terzi

16:35–17:00 **Coupled Bayesian Sets Algorithm for Semi-supervised Learning and Information Extraction**
Saurabh Verma and Estevam R. Hruschka Jr.

17:00–17:25 **Policy Iteration Based on a Learned Transition Model**
Vivek Ramavajjala and Charles Elkan

Room: Chem, LT2

Wed3B: Sensor Data

Chair: Martin Atzmueller

16:10–16:35 **MDL-Based Analysis of Time Series at Multiple Time-Scales**
Ugo Vespier, Arno Knobbe, Siegfried Nijssen and Joaquin Vanschoren

16:35–17:00 **Separable Approximate Optimization of Support Vector Machines for Distributed Sensing**
Sangkyun Lee, Marco Stolpe and Katharina Morik

17:00–17:25 **Unsupervised Inference of Auditory Attention from Biosensors**
Melih Kandemir, Arto Klami, Akos Vetek and Samuel Kaski

Room: Chem, LT3

Wed3C: Privacy and Security

Chair: Patrick Gallinari

16:10–16:35 ***AUDIO*: An Integrity Auditing Framework of Outlier-Mining-as-a-Service Systems**
Ruilin Liu, Hui (Wendy) Wang, Anna Monreale, Dino Pedreschi, Fosca Giannotti and Wenge Guo

16:35–17:00 **Differentially Private Projected Histograms: Construction and Use for Prediction**
Staal A. Vinterbo

17:00–17:25 **Fairness-Aware Classifier with Prejudice Remover Regularizer**
Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh and Jun Sakuma

Room: Chem, LT1
Chair: Jan Ramon

Wed2A: Social Network Mining I

14:00–14:25 **Discovering Links among Social Networks**
Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino

14:25–14:50 **Efficient Bi-objective Team Formation in Social Networks**
Mehdi Kargar Aijun An, and Morteza Zihayat

14:50–15:15 **Feature-Enhanced Probabilistic Models for Diffusion Network Inference**
Liaoruo Wang, Stefano Ermon and John E. Hopcroft

15:15–15:40 **Influence Spread in Large-Scale Social Networks – A Belief Propagation Approach**
Huy Nguyen and Rong Zheng

Room: Chem, LT2
Chair: Yizhao Ni

Wed2B: Large-Scale, Distributed and Parallel Mining and Learning I

14:00–14:25 **Learning Compact Class Codes for Fast Inference in Large Multi Class Classification**
M. Cissé, T. Artières and Patrick Gallinari

14:25–14:50 **ParCube: Sparse Parallelizable Tensor Decompositions**
Evangelos E. Papalexakis, Christos Faloutsos and Nicholas D. Sidiropoulos

14:50–15:15 **Stochastic Coordinate Descent Methods for Regularized Smooth and Nonsmooth Losses**
Qing Tao, Kang Kong, Dejun Chu and Gaowei Wu

15:15–15:40 **Sublinear Algorithms for Penalized Logistic Regression in Massive Datasets**
Haoruo Peng, Zhengyu Wang, Edward Y. Chang, Shuchang Zhou and Zhihua Zhang

Room: Chem, LT3
Chair: João Gama

Wed2C: Online Learning and Data Streams

14:00–14:25 **Adaptive Two-View Online Learning for Math Topic Classification**
Tam T. Nguyen, Kuiyu Chang and Siu Cheung Hui

14:25–14:50 **BDUOL: Double Updating Online Learning on a Fixed Budget**
Peilin Zhao and Steven C.H. Hoi

14:50–15:15 **Improved Counter Based Algorithms for Frequent Pairs Mining in Transactional Data Streams**
Konstantin Kutzkov

15:15–15:40 **Mirror Descent for Metric Learning: A Unified Approach**
Gautam Kunapuli and Jude Shavlik

Room: Chem, LT4
Chair: Sami Kaski

Wed2D: Nectar Track

14:00–14:50 **Matrix Factorization as Search**
Kristian Kersting, Christian Bauckhage, Christian Thurau and Mirwaes Wahabzada

14:50–15:40 **Metal Binding in Proteins: Machine Learning Complements X-Ray Absorption Spectroscopy**
Marco Lippi, Andrea Passerini, Marco Punta and Paolo Frasconi

17:10–17:20 **TopicExplorer: Exploring Document Collections with Topic Models**
Alexander Hinneburg, Rico Preiss and René Schröder

The demo presents a prototype – called TopicExplorer– that combines topic modeling, key word search and visualization techniques to explore a large collection of Wikipedia documents. Topics derived by Latent Dirichlet Allocation are presented by top words. In addition, topics are accompanied by image thumbnails extracted from related Wikipedia documents to aid sense making of derived topics during browsing. Topics are shown in a linear order such that similar topics are close. Topics are mapped to color using that order. The auto-completion of search terms suggests words together with their color coded topics, which allows to explore the relation between search terms and topics. Retrieved documents are shown with color coded topics as well. Relevant documents and topics found during browsing can be put onto a shortlist. The tool can recommend further documents with respect to the average topic mixture of the shortlist.

17:20–17:30 **Extracting Trajectories through an Efficient and Unifying Spatio-temporal Pattern Mining System**
Phan Nhat Hai, Dino Ienco, Pascal Poncelet and Maguelonne Teisseire

Recent improvements in positioning technology has led to a much wider availability of massive moving object data. A crucial task is to find the moving objects that travel together. Usually, these object sets are called spatio-temporal patterns. Analyzing such data has been applied in many real world applications, e.g., in ecological study, vehicle control, mobile communication management, etc. However, few tools are available for flexible and scalable analysis of massive scale moving objects. Additionally, there is no framework devoted to efficiently manage multiple kinds of patterns at the same time. Motivated by this issue, we propose a framework, named GeT_Move, which is designed to extract and manage different kinds of spatio-temporal patterns concurrently. A user-friendly interface is provided to facilitate interactive exploration of mining results. Since GeT_Move is tested on many kinds of real data sets, it will benefit users to carry out versatile analysis on these kinds of data by exhibiting different kinds of patterns efficiently.

17:30–17:40 **Scientific Workflow Management with ADAMS**
Peter Reutemann and Joaquin Vanschoren

We demonstrate the Advanced Data mining And Machine learning System (ADAMS), a novel workflow engine designed for rapid prototyping and maintenance of complex knowledge workflows. ADAMS does not require the user to manually connect inputs to outputs on a large canvas. It uses a compact workflow representation, *control operators*, and a simple interface between operators, allowing them to be auto-connected. It contains an extensive library of operators for various types of analysis, and a convenient plug-in architecture to easily add new ones.

17:40–17:50 **ClowdFlows: A Cloud Based Scientific Workflow Platform**
Janez Kranjc, Vid Podpečan and Nada Lavrač

This paper presents an open cloud based platform for composition, execution, and sharing of interactive data mining workflows. It is based on the principles of service-oriented knowledge discovery, and features interactive scientific workflows. In contrast to comparable data mining platforms, our platform runs in all major Web browsers and platforms, including mobile devices. In terms of crowdsourcing, ClowdFlows provides researchers with an easy way to expose and share their work and results, as only an Internet connection and a Web browser are required to access the workflows from anywhere. Practitioners can use ClowdFlows to seamlessly integrate and join different implementations of algorithms, tools and Web services into a coherent workflow that can be executed in a cloud based application. ClowdFlows is also easily extensible during run-time by importing Web services and using them as new workflow components.

Ten-Year Award Talk: Non-Derivable Frequent Itemsets

Toon Calders

Eindhoven University of Technology, The Netherlands
http://wwwis.win.tue.nl/~tcalders/

Bart Goethals

University of Antwerp, Belgium
http://adrem.ua.ac.be/~goethals/

Wednesday 09:00–10:00 Room: Chem, LT1

Abstract

In 1993, in their ACM SIGMOD'93 paper, Agrawal et al. proposed the frequent itemset mining problem. For many years following the introduction of this problem, numerous studies on algorithms for this problem had resulted in significant performance improvements for mining all frequent itemsets. At the same time, however, it also became clear that problem definition in itself, was problematic: regardless of the algorithm being used, if the minimal support threshold is set too low, or the data is highly correlated, the number of frequent itemsets itself becomes prohibitively large. Furthermore, typically the collection of frequent itemsets contains a lot of redundancy.

To overcome this problem, several proposals were made to construct concise representations of the frequent itemsets, instead of mining all frequent itemsets; the most well-known being the *closed sets* proposed by Pasquier et al. at ICDT 1999. It is in this context that our ECML PKDD 2002 paper *Mining All Non-Derivable Frequent Itemsets* was written. The main goal of our paper was to identify redundancies in the set of all frequent itemsets and to exploit these redundancies in order to reduce the result of a mining operation. Deduction rules to derive tight bounds on the support of candidate itemsets were proposed, and it was shown how the deduction rules allow for constructing a minimal representation for all frequent itemsets.

In our presentation we present the most important properties and algorithms for mining non-derivable itemsets from our 2002 paper and follow-up works. We will also position the non-derivable itemsets as part of a much larger program of understanding the interactions of itemset frequencies as a special case of probabilistic reasoning, and end with the recently developed statistically based itemset interestingness measures as a natural next step in this endeavor to make pattern mining more meaningful.

Room: Chem, LT1
Chair: Stefan Rüping

Wed1A: Distance-Based Methods and Kernels

10:30–10:55 **Classifying Stem Cell Differentiation Images by Information Distance**
Xianglilan Zhang, Hongnan Wang, Tony J. Collins, Zhigang Luo and Ming Li

10:55–11:20 **Distance Metric Learning Revisited**
Qiong Cao, Yiming Ying and Peng Li

11:20–11:45 **Geodesic Analysis on the Gaussian RKHS Hypersphere**
Nicolas Courty, Thomas Burger and Pierre-François Marteau

11:45–12:10 **The Bitvector Machine: A Fast and Robust Machine Learning Algorithm for Non-linear Problems**
Stefan Edelkamp and Martin Stommel

Room: Chem, LT2
Chair: Indrė Žliobaitė

Wed1B: Time Series and Temporal Data Mining

10:30–10:55 **Community Trend Outlier Detection Using Soft Temporal Pattern Mining**
Manish Gupta, Jing Gao, Yizhou Sun, and Jiawei Han

10:55–11:20 **Data Structures for Detecting Rare Variations in Time Series**
Caio Valentim, Eduardo S. Laber and David Sotelo

11:20–11:45 **Invariant Time-Series Classification**
Josif Grabocka, Alexandros Nanopoulos and Lars Schmidt-Thieme

11:45–12:10 **Learning Bi-clustered Vector Autoregressive Models**
Tzu-Kuo Huang and Jeff Schneider

Room: Chem, LT3
Chair: Willem Waegeman

Wed1C: Spatial and Geographical Data Mining

10:30–10:55 **Inferring Geographic Coincidence in Ephemeral Social Networks**
Honglei Zhuang, Alvin Chin, Sen Wu, Wei Wang, Xia Wang and Jie Tang

10:55–11:20 **Socioscope: Spatio-temporal Signal Recovery from Social Media**
Jun-Ming Xu, Aniruddha Bhargava, Robert Nowak and Xiaojin Zhu

11:20–11:45 **Location Affiliation Networks: Bonding Social and Spatial Information**
Konstantinos Pelechrinis and Prashant Krishnamurthy

11:45–12:10 **Pedestrian Quantity Estimation with Trajectory Patterns**
Thomas Liebig, Zhao Xu, Michael May and Stefan Wrobel

Room: Chem, LT4
Chair: Thomas Gärtner

Wed1D: Nectar Track

10:30–11:20 **Learning Submodular Functions**
Maria-Florina Balcan and Nicholas J.A. Harvey

11:20–12:10 **Modelling Input Varying Correlations between Multiple Responses**
Andrew Gordon Wilson and Zoubin Ghahramani