



ECML PKDD 2011

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



5-9 SEPTEMBER 2011, ATHENS-GREECE
www.ecmlpkdd2011.org

	Mon 5		Tue 6		Wed 7		Thu 8		Fri 9
09:00-10:30	Workshops & Tutorials	09:00-09:10	Announcements	09:00-09:10	Announcements	09:00-09:10	Announcements		Invited talk by Heikki Mannila
		09:10-10:10	Invited talk by Rakesh Agrawal	09:10-10:10	Invited talk by László Barabási		Supervised Learning II	09:00-10:00	
10:30-11:00	Coffee Break	10:10-10:40	Coffee Break	10:10-10:40	Coffee break		Preference Learning and Ranking	10:00-10:30	Coffee break
			Classification & Prediction		Data Mining Theory & Foundations		Semi-Supervised and Transductive Learning		Workshops
11:00-12:30	Workshops & Tutorials	10:40-12:30	Frequent Sets and Patterns	10:40-12:30	Learning from Social and Information Networks I		Coffee Break	10:30-13:45	Tutorials
			Active & Online learning		Spectral Clustering & Graph Mining	11:00-11:30	Invited talk by Marco Gori		Industrial Track
12:30-14:00	Lunch (on your own)	12:30-14:00	Lunch	12:30-14:00	Lunch	12:30-14:00	Lunch	13:45-15:00	Lunch (on your own)
			Applications of Data Mining		Learning from Social and Information Networks II		Feature Selection, Extraction, and Construction		Workshops
14:00-15:30	Workshops & Tutorials	14:00-15:50	Classification & Bayesian Networks	14:00-15:50	Model Selection & Statistical Learning	14:00-15:50	Text Mining & Recommender Systems	15:00-16:30	Tutorials
			Ensemble Learning		Relational learning and Inductive Logic Programming		Reinforcement learning		Industrial Track
15:30-16:00	Coffee Break	15:50-16:20	Coffee Break	15:50-16:20	Coffee Break	15:50-16:20	Coffee Break	16:30-17:00	Coffee Break
16:00-17:30	Workshops & Tutorials		Learning from Time Series Data		Unsupervised Learning & Dimensionality Reduction	16:20-17:20	Demos		Workshops
		16:20-18:10	Matrix and Tensor Analysis	16:20-18:10	Graphical & Hidden Markov Models			17:00-18:30	Tutorials
19:00-19:30	Openings & awards ceremony		Clustering		Supervised Learning II	17:20-19:00	Community meeting		Industrial Track
19:30-20:30	Invited Talk by Christopher Bishop	17:30	Poster Reception	19:30-20:00	10-years award				
20:30	Welcome Reception			20:00-21:00	Invited talk by Andrei Broder	20:00	Guided Tour at the Acropolis Museum		
				21:00	Conference Banquet		Farewell party		

ECML PKDD 2011 Conference Brochure

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

5-9 September 2011, Athens, Greece
www.ecmlpkdd2011.org

Brochure created by:
Marianna Siganou

Editors:
Despina Kopanaki
Maria Halkidi
[University of Piraeus, Greece]

Contents

1.	Welcome message from General co-Chairs	3
2.	Welcome message from Program Committee Chairs	5
3.	Sponsors	7
4.	Useful Information	8
5.	Internet Access	9
6.	The Conference Venue	9
7.	Social Events	12
8.	Program at a glance	18
9.	Monday at a glance	19
10.	Tuesday at a glance	19
11.	Wednesday at a glance	20
12.	Thursday at a glance	20
13.	Friday at a glance	21
14.	Invited Talks	22
15.	Industrial Session	29
16.	Awards	30
17.	Technical Sessions	31
18.	Abstracts	39
19.	Demos	83
20.	Tutorials	86
21.	Discovery Challenge	89
22.	Workshops	91
23.	Athens	103
24.	Transportation	110
25.	Conference Organization	113
26.	Program Committee	115
27.	My notes	118

1. Welcome message from General co-Chairs

Welcome to the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2011) held in Athens, Greece from September 5th to 9th, 2011. ECML PKDD is an annual conference that provides an international forum for the discussion of the latest high quality research results in all areas related to machine learning and knowledge discovery in databases as well as other innovative application domains. Over the years it has evolved as one of the largest and most selective international conferences in machine learning and data mining, the only one that provides a common forum for these two closely related fields.

The ECML PKDD 2011 includes all the scientific events and activities of big conferences. The scientific program consists of technical presentations of accepted papers, plenary talks by distinguished keynote speakers, workshops and tutorials, a discovery challenge track, demo and industrial tracks. Moreover, two co-located workshops are organized on related research topics. We expect that all those scientific activities provide opportunities for knowledge dissemination, fruitful discussions and exchange of ideas among people both from academy and industry. Moreover, we hope that this conference will continue to offer a unique forum that stimulates and encourages close interaction among researchers working on machine learning and data mining.

We are very happy to have the conference back to Greece after 1995 when ECML was successfully organized in Heraklion, Crete. However, this is the first time that the joint ECML PKDD event is organized in Greece and, more specifically, in Athens, with the conference venue boasting a superb location under the Acropolis and in front of the Temple of Zeus. Besides the scientific activities, the conference offers the delegates an attractive bunch of social activities, such as a welcome reception at the roof garden of the conference venue directly facing the Acropolis hill, a poster session at “Technopolis” Gazi industrial park, the conference banquet at the conference venue, and a farewell party at the new Acropolis Museum, one of the most impressive archaeological museums worldwide, that includes a guided tour at the museum exhibits.

Several people worked hard together as a superb dedicated team to ensure the successful organization of this conference. First, we would like to express our thanks and deep gratitude to the PC Chairs Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba and Michalis Vazirgiannis. They efficiently carried out the enormous task of coordinating the rigorous hierarchical double-blind review process that resulted in a rich, while at the same time, very selective scientific program. Their contribution was crucial and essential in all phases and aspects of the conference organization and it was by no means restricted only to the paper review process. We would also like to thank the Area Chairs and Program Committee members for the valuable assistance they offered to the PC Chairs in timely completing the review process under strict deadlines. Special thanks should also be given to the Workshop Co-Chairs, Bart Goethals and Katharina Morik, the Tutorial Co-Chairs, Fosca Giannotti and Maguelonne Teisseire, the Discovery Challenge Co-Chairs, Alexandros Kalousis and Vassilis Plachouras, the Industrial Session Co-Chairs, Alexandros Ntoulas and Michail Vlachos, the Demo Track Co-Chairs, Michelangelo Ceci and Spiros Papadimitriou, and the Best Paper Award Co-Chairs, Sunita Sarawagi and Michèle Sebag. We further thank the keynote speakers, workshop organizers, the tutorial presenters and the organizers of the discovery challenge.

Furthermore, we are indebted to the Publicity Co-Chairs, Annalisa Appice and Grigo-

rios Tsoumakas, who developed and timely implemented an effective dissemination plan and supported the Program Chairs in the production of proceedings, and also to Margarita Karkali for the development, support and timely update of the conference webpage. We further thank the members of the ECML PKDD Steering Committee for their valuable help and guidance.

The conference was financially supported by the following generous sponsors who are worthy of special acknowledgment: Google, Pascal2 Network, Xerox, Yahoo Labs, COST-MOVE Action, Rapid-I, FP7-MODAP Project, Athena RIC / Institute for the Management of Information Systems, Hellenic Artificial Intelligence Society, Marathon Data Systems, Transinsight. Additional support was generously provided by Sony, Springer, and the UNESCO Privacy Chair Program. This support has given us the opportunity to specify low registration rates, provide video-recording services and support students through travel grants for attending the conference. The substantial effort of the Sponsorship Co-Chairs, Ina Lauth and Ioannis Kopanakis, was crucial in order to attract those sponsorships, therefore, they deserve our special thanks. Special thanks should also be given to the five Organizing Institutions, namely, University of Bari “Aldo Moro”, Athens University of Economics and Business, University of Athens, University of Ioannina and University of Piraeus for supporting in multiple ways our task.

We would like to especially acknowledge the members of the Local Organization team, Maria Halkidi, Despina Kopanaki and Nikos Pelekis, for making all necessary local arrangements and Triaena Tours & Congress S.A. for efficiently handling finance and registrations. The essential contribution of the student volunteers also deserves special acknowledgment.

Finally, we are indebted to all researchers who considered this conference as a forum for presenting and disseminating their research work, as well as to all conference participants hoping that this event will stimulate further expansion of research and industrial activities in machine learning and data mining.

Aristidis Likas
Yannis Theodoridis

2. Welcome message from Program Committee Chairs

The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2011) took place in Athens, Greece, during September 5-9, 2011. This year we have completed the first decade since the junction between the European Conference on Machine Learning and the Principles and Practice of Knowledge Discovery in Data Bases conferences, which as independent conferences date back to 1986 and 1997, respectively. During this decade there has been an increasing integration of the two fields, as reflected by the rising number of submissions of top-quality research results. In 2008 a single ECML PKDD Steering Committee was established, which gathered senior members of both communities.

The ECML PKDD conference is a highly selective conference in both areas, the leading forum where researchers in machine learning and data mining can meet, present their work, exchange ideas, gain new knowledge and perspectives, and motivate the development of new interesting research results. Although traditionally based in Europe, ECML PKDD is also a truly international conference with rich and diverse participation.

In 2011, as in previous years, ECML PKDD followed a full week schedule, from Monday to Friday. It featured six plenary invited talks by Rakesh Agrawal, Albert-László Barabási, Christopher Bishop, Andrei Broder, Marco Gori and Heikki Mannila. Monday and Friday were devoted to workshops selected by Katharina Morik and Bart Goethals, and tutorials, organized and selected by Fosca Giannotti and Maguelonne Teisseire. There was also an interesting industrial session, managed by Alexandros Ntoulas and Michalis Vlachos which welcomed distinguished speakers from the ML and DM industry: Vasilis Aggelis, Radu Jurca, Neel Sundaresan and Olivier Verscheure. The 2011 discovery challenge was organized by Alexandros Kalousis and Vassilis Plachouras.

The main conference program unfolded from Tuesday to Thursday, where 121 papers selected among 599 full-paper submissions were presented in the technical parallel sessions and in a poster session open to all accepted papers. The acceptance rate of 20% supports the traditionally high standards of the joint conference. The selection process was assisted by 35 Area Chairs, each supervising the reviews and discussions of about 17 papers, and by 270 members of the Program Committee, with the help of 197 additional reviewers. While the selection process was made particularly intense due to the very high number of submissions, we are grateful and heartily thank all Area Chairs, members of the Program Committee, and additional reviewers for their commitment and hard work during the tight reviewing period.

The composition of the paper topics covered a wide spectrum of issues. A significant portion of the accepted papers dealt with core issues such as supervised and unsupervised learning with some innovative contributions in fundamental issues such as cluster-ability of a dataset.

Other fundamental issues tackled by accepted papers include dimensionality reduction, distance and similarity learning, model learning and matrix/tensor analysis. In addition, there was a significant cluster of papers with valuable contributions on graph mining, graphical models, hidden Markov models, kernel methods, active and ensemble learning, semi-supervised and transductive learning, mining sparse representations, model learning, inductive logic programming, and statistical learning.

A significant part of the program covered novel and timely applications of data mining and machine learning in industrial domains, including: privacy-preserving and discrimi-

nation-aware mining, spatiotemporal data mining, text mining, topic modeling, learning from environmental and scientific data, Web mining and Web search, link analysis, bio/medical data, data Streams and sensor data, ontology-based data, relational data mining, learning from time series data, time series data.

In the past three editions of the joint conference, the two Springer journals Machine Learning and Data Mining and Knowledge Discovery published the top papers in two special issues printed in advance of the conference. These papers were not included in the conference proceedings, so there was no double publication of the same work. A novelty introduced this year was the post-conference publication of the special issues in order to guarantee the expected high-standard reviews for top-quality journals. Therefore, authors of selected machine learning and data mining papers were invited to submit a significantly extended version of their paper to the special issues. The selection was made by Program Chairs on the basis of their exceptional scientific quality and high impact on the field, as indicated by conference reviewers.

Following an earlier tradition, the Best Paper Chairs Suita Sarawak and Michele Sebag contributed to the selection of papers deserving the Best Paper Awards and Best Student Paper Awards in Machine Learning and in Data Mining, sponsored by Springer. As ECML PKDD completes 10 years of joint organization, the PC chairs, together with the steering committee, initiated a 10-year Awards series. This award is established for the author(s), whose paper appeared in the ECML PKDD conference 10 years ago, and had the most impact on the machine learning and data mining research since then. This year's, first award, committee consisted of three PC co-chairs (Dimitrios Gunopulos, Donato Malerba and Michalis Vazirgiannis) and three Steering Committee members (Wray Buntine, Bart Goethals and Michèle Sebag).

The conference also featured a demo track, managed by Michelangelo Ceci and Spiros Papadimitriou; 11 demos out of 21 submitted were selected by a Demo Track Program Committee, presenting prototype systems that illustrate the high impact of machine learning and data mining application in technology. The demo descriptions are included in the proceedings. We further thank the members of the Demo Track Program Committee for their efforts in timely reviewing submitted demos.

Finally, we would like to thank the General Chairs, Aristidis Likas and Yannis Theodoridis, for their critical role in the success of the conference, the Tutorial, Workshop, Demo, Industrial Session, Discovery Challenge, Best Paper, and Local Chairs, the Area Chairs and all reviewers, for their voluntary, highly dedicated and exceptional work, and the ECML PKDD Steering Committee for their help and support. Our last and warmest thanks go to all the invited speakers, the speakers, all the attendees, and especially to the authors who chose to submit their work to the ECML PKDD conference and thus enabled us to build up this memorable scientific event.

Dimitrios Gunopulos
Thomas Hofmann
Donato Malerba
Michalis Vazirgiannis

3. Sponsors

We wish to express our gratitude to the sponsors of ECML PKDD 2011 for their essential contribution to the conference: Google, Pascal2 Network, Xerox, Yahoo Lab, COST/MOVE (Knowledge Discovery from Moving Objects), FP7/MODAP (Mobility, Data Mining, and Privacy), Rapid-I (report the future), Athena/IMIS (Institute for the Management of Information Systems), EETN (Hellenic Artificial Intelligence Society), MDS Marathon Data Systems, Transinsight, SONY, UNESCO Chair in Data Privacy, Springer, Machine Learning Journal, Data Mining and Knowledge Discovery, Università degli Studi di Bari “Aldo Moro”, Athens University of Economics and Business, University of Ioannina, National and Kapodistrian University of Athens, and the University of Piraeus.

PLATINUM SPONSORS



GOLD SPONSORS



SILVER SPONSORS



BRONZE SPONSORS



ORGANIZING INSTITUTIONS



ADDITIONAL SUPPORTERS



4. Useful Information



Phone prefix for Greece:	+30
Congress Secretariat on-site:	210 9288400

Emergencies

Medical Emergency Number	112
Ambulance Service	166
Marine Police Immediate intervention	108
Police-Immediate Response	100
Immediate social help	197
Counterterrorism agency	1014
Police Departments, tel. Call center	1033, 10400
Emergencies Hospitals, Pharmacies	14944
Poisoning center	210 7793777
Fire Brigade	199
Tourist police	171
Athens traffic police	210 5284000
Forest fire service	191

Transportation

Athens International Airport “Eleftherios Venizelos”	210 3530000
Boats time-tables	14944
Port Office of Piraeus	210 4147800
Port Office of Rafina	22940 22300
O.A.S.A Urban transportation information desk	185
K.T.E.L. time-tables	14505
Proastiakos Suburban Railway	1110
Athens radio taxis	210 921-7942, 643-3400

Service of Lost or Stolen Credit Cards

American Express	210 339 7250
Diners	210 929 0200
Eurocard	210 950 3673
Mastercard	00800118870303
Visa	00800116380304

5. Internet Access

During the conference, free wireless internet access will be provided to the participants. Username and Password will be given at the conference secretariat (registration desk).

6. The Conference Venue

Athens Royal Olympic Hotel is a family run five star property in the centre of Athens. It lays just in front of the famous Temple of Zeus and the National Gardens. It is underneath the Acropolis and only 2 minutes walk to the new Athens Acropolis Museum.

After its complete renovation that finished in 2009, the Royal Olympic was transformed to an art hotel very elegantly decorated and more important very well looked after in every detail. One of the aspects given particular attention to, was to create a very personal hotel and as much environmentally friendly as possible.



Getting there

Athens Royal Olympic Hotel is located in 28-34 *Athanasiou Diakou Str.*

The closest metro station is “*Acropolis*”. From “*Acropolis*” Station the Hotel is 200m away, walking along *Athanasiou Diakou Str.*

From Athens International Airport “Eleftherios Venizelos” to the conference venue by metro:

You will board the Metro from the Airport’s Station and get off at Syntagma Station. At Syntagma Station you switch lines in the direction of Ag. Dimitrios and get off at the first Station, the “*Acropolis Station*”.

The conference will take place in the following rooms:

- » Olympia Hall (main conference room): basement floor.
- » Attica Hall: ground floor
- » Kallirhoe: ground floor
- » Conference Rooms I-V

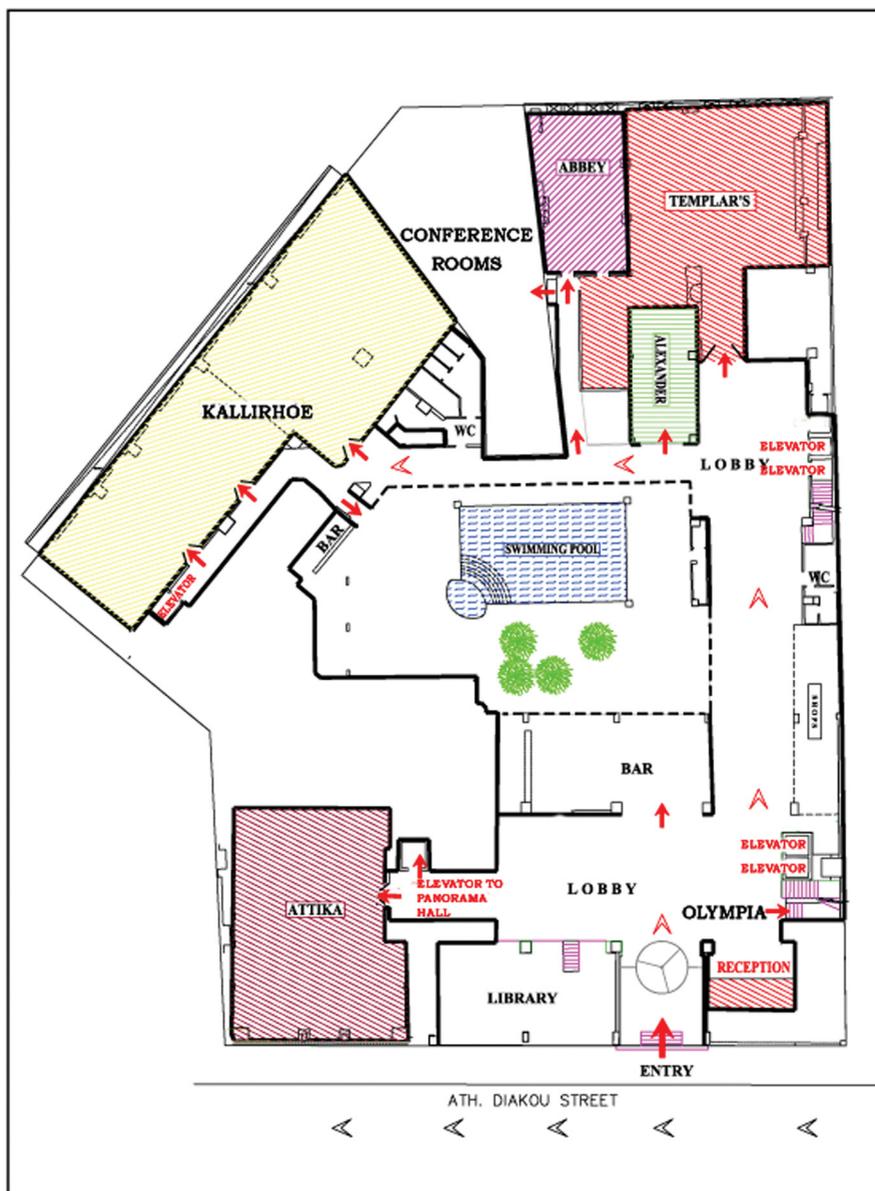
The following map indicates the location of each room.

6.1 Conference Secretariat

The conference secretariat (registration desk) will be located in Olympia Foyer in the basement of Royal Olympic Hotel.

Operation Hours:

- » Monday 5th, 08:00 – 19:30
- » Tuesday 6th, 08:00 – 18:00
- » Wednesday 7th, 08:30 – 20:00
- » Thursday 8th, 08:30 – 17:30
- » Friday 9th, 08:30 – 18:30



ROYAL OLYMPIC HOTEL

LOBBY - ΑΙΘΟΥΣΕΣ

7. Social Events

Monday: Welcome Reception



On Monday 5th at 20:30, a welcome reception will be held at “Ioannis Restaurant” in the roof garden of Royal Olympic, the main conference venue, facing the Acropolis hill.

i Please remember to bring your entrance ticket to this event.

Sponsored by the Machine Learning journal in honor of its 25th Anniversary



Tuesday: Poster Reception



TECHNOPOLIS
CITY OF ATHENS

On Tuesday 6th of September at 19:30, the poster reception will take place at “Technopolis”, inside Athens Gazi Industrial Park.

i Please remember to bring your entrance ticket to this event.



The “Technopolis” of the City of Athens, an industrial museum of incomparable architecture - among the most interesting in the world, has been transformed into a multipurpose cultural space. The centre has assisted in the upgrading of a historic Athens district and the creation of yet another positive element in Athens’ cultural identity.

It is housed in the city’s former gasworks, on a site known as Gazi spanning three hectares, next to the Kerameikos area and in close proximity to the Acropolis. The gasworks was gradually transformed into an education centre and host venue for various events. The visitor can take a stroll through the site filled with images, knowledge and emotions.

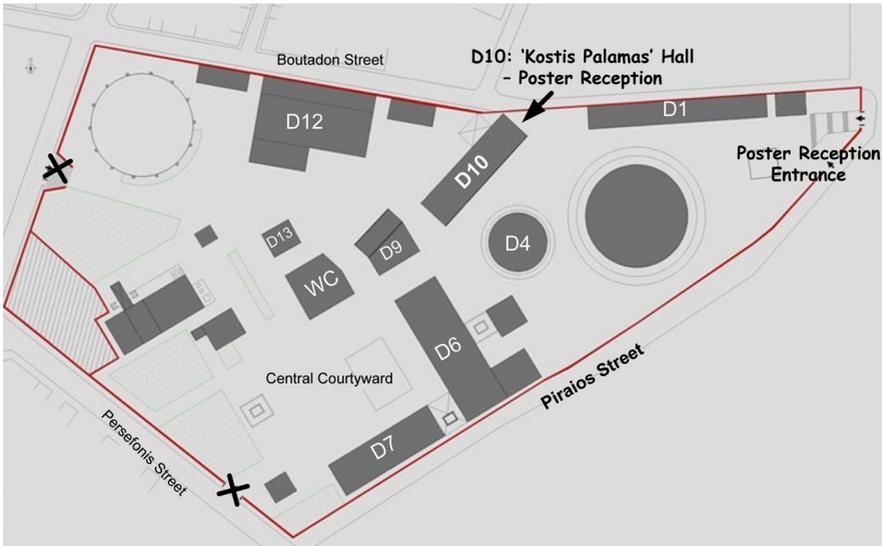
The charm of a bygone area, conveyed through stacks, enormous cauldrons (gasometers), chimneys and ovens, “conspires” with reverence to establish the site as a “factory” for the protection and production of art. Etymologically, the word “gas” (derived from the ancient German galist, later geist) means spirit.

In operation since **1999**, Technopolis is dedicated to the memory of the unforgettable composer **Manos Hadjidakis**.

The poster session will take place in **D10: ‘Kostis Palamas’ Hall**.



This site constituted the so-called “purgatory” where the gas, having undergone the freezing process, was channelled through pipes to the “presses” for the final stage of cleaning.

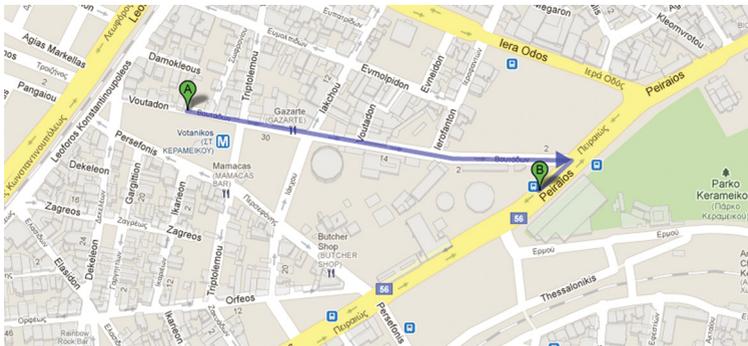


Getting there

Address: 100 Pireos St, Athens

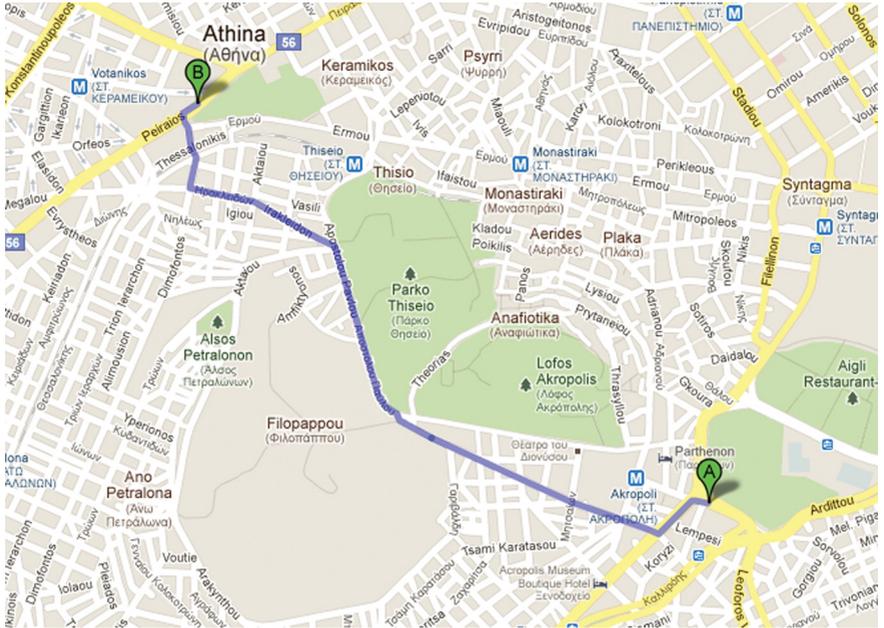
By metro: The main entrance of “Technopolis” is located approximately 500 meters away from metro station ‘Kerameikos’. Head east onto *Voutadon Str.* towards *Sofroniou Str.* Turn right onto *Pireos Str.*

i *Make sure that you enter “Technopolis” from its entrance at Pireos Str.*



By foot: If the weather is nice (and most probably, since it is September it will be...) we suggest walking from the conference venue to “Technopolis”, a 30 minutes relaxed walk down to *Apostolou Pavlou* pedestrian, viewing the Acropolis hill on your right and the Filopappou hill on your left.

i *Again, make sure that you enter “Technopolis” from its entrance at Pireos Str.*



Wednesday: Conference Banquet



On Wednesday 6th at 21:00, the Conference banquet will be held at “Ioannis Restaurant” in the roof garden of Royal Olympic, the main conference venue.

i Please remember to bring your entrance ticket to this event.

Thursday: Acropolis Museum and Farewell Party

On Thursday 7th at 20:00, a farewell party will take place at the Acropolis Museum, one of the most impressive archaeological museums worldwide. Our evening will start with a guided tour at the museum exhibits, followed by a cocktail at the museum's restaurant, with a breathtaking view of Acropolis.

 Please remember to bring your entrance ticket to this event.



The Acropolis Museum is considered to be one of the most important museums in the world, since it houses the unique collection of original sculptural masterpieces of Archaic and Classical Greek art from the sacred rock and citadel of ancient Athens.

These are mainly freestanding votive sculptures and important groups of architectural sculptures, which decorated the buildings erected on the Acropolis in the Archaic and Classical periods. The display also includes clay votive offerings. Other finds from the Acropolis, such as vases, bronze objects and relief sculptures, are displayed in the National Archaeological Museum, while the inscriptions are kept in the Epigraphical Museum. A noteworthy absence for the Acropolis Museum is the sculptures removed by Lord Elgin in the nineteenth century and currently displayed in the British Museum.

The museum is directly linked to the archaeological site of the Acropolis and to the extensive conservation work carried out on the sacred rock. It is under the supervision of the First Ephorate of Prehistoric and Classical Antiquities of the Ministry of Culture, which also oversees the construction of the New Acropolis Museum.

Today, the new Acropolis Museum has a total area of 25,000 square meters, with exhibition space of over 14,000 square meters, ten times more than that of the old museum on the Hill of the Acropolis. The new Museum offers all the amenities expected in an international museum of the 21st century.

 A guided tour at the Acropolis Museum is scheduled to take place, dedicated to ECML PKDD participants.

Getting there

Address: *15 Dionysiou Areopagitou Str., Athens*

The Acropolis Museum is located in the historical area of Makriyianni, southeast of the Rock of the Acropolis, on *Dionysiou Areopagitou Str.*, only a few meters from the Acropolis hill. The Museum entrance is located at the one end of the pedestrian walkway of *Dionysiou Areopagitou Str.*, which constitutes the central route for the unified network of the city's archaeological sites and by its own is considered one of the highlights of Athens. The 'Acropolis' Metro station is just on the east side of the Museum site.

From the conference venue: Getting to the Acropolis Museum from the Royal Olympic Hotel is only 500 meters walk. Turn left onto *Athanassiou Diakou Str.*, cross *Syngrou Ave.*, continue straight onto *Athanassiou Diakou Str.*, turn right onto *Stratigou Makriyianni Str.*, and turn left onto *Dionysiou Areopagitou Str.*

8. Program at a glance

	Mon 5		Tue 6		Wed 7		Thu 8		Fri 9
09:00-10:30	Workshops & Tutorials	09:00-09:10	Announcements	09:00-09:10	Announcements	09:00-09:10	Announcements	09:00-10:00	Invited talk by Heikki Mannila
		09:10-10:10	Invited talk by Rakesh Agrawal	09:10-10:10	Invited talk by László Barabási		Supervised Learning II		
10:30-11:00	Coffee Break	10:10-10:40	Coffee Break	10:10-10:40	Coffee break	09:10-11:00	Preference Learning and Ranking	10:00-10:30	Coffee break
11:00-12:30	Workshops & Tutorials	10:40-12:30	Classification & Prediction	10:40-12:30	Data Mining Theory & Foundations		Semi-Supervised and Transductive Learning	10:30-13:45	Workshops
			Frequent Sets and Patterns		Learning from Social and Information Networks I	11:00-11:30	Coffee Break		Tutorials
			Active & Online learning		Spectral Clustering & Graph Mining	11:30-12:30	Invited talk by Marco Gori		Industrial Track
12:30-14:00	Lunch (on your own)	12:30-14:00	Lunch	12:30-14:00	Lunch	12:30-14:00	Lunch	13:45-15:00	Lunch (on your own)
14:00-15:30	Workshops & Tutorials	14:00-15:50	Applications of Data Mining	14:00-15:50	Learning from Social and Information Networks II	14:00-15:50	Feature Selection, Extraction, and Construction	15:00-16:30	Workshops
			Classification & Bayesian Networks		Model Selection & Statistical Learning		Text Mining & Recommender Systems		Tutorials
			Ensemble Learning		Relational learning and Inductive Logic Programming		Reinforcement learning		Industrial Track
15:30-16:00	Coffee Break	15:50-16:20	Coffee Break	15:50-16:20	Coffee Break	15:50-16:20	Coffee Break	16:30-17:00	Coffee Break
16:00-17:30	Workshops & Tutorials	16:20-18:10	Learning from Time Series Data	16:20-18:10	Unsupervised Learning & Dimensionality Reduction	16:20-17:20	Demos	17:00-18:30	Workshops
19:00-19:30	Openings & awards ceremony		Matrix and Tensor Analysis		Graphical & Hidden Markov Models		Community meeting		Tutorials
			Clustering		Supervised Learning II	17:20-19:00			Industrial Track
19:30-20:30	Invited Talk by Christopher Bishop	17:30	Poster Reception	19:30-20:00	10-years award				
20:30	Welcome Reception			20:00-21:00	Invited talk by Andrei Broder	20:00	Guided Tour at the Acropolis Museum		
				21:00	Conference Banquet		Farewell party		



9. Monday at a glance

Monday 5 at a glance								
	Olympia Hall	Attica Hall	Kallirhoe Hall	Conference Room I	Conference Room II	Conference Room III	Conference Room IV	Conference Room V
09:00-10:30	Tutorial «Privacy Challenges and Solutions for Medical Data Sharing»	Tutorial «Introduction to causal discovery: A Bayesian Networks approach»	Workshop «LSHTC»	Workshop «ISEW»	Workshop «MUSE»	Discovery Challenge	Workshop MultiClust	Workshop «DMFGP»
10:30-11:00	Coffee Break							
11:00-12:30	Tutorial «Privacy Challenges and Solutions for Medical Data Sharing»	Tutorial «Introduction to causal discovery: A Bayesian Networks approach»	Workshop «LSHTC»	Workshop «ISEW»	Workshop «MUSE»	Discovery Challenge	Workshop MultiClust	Workshop «DMFGP»
12:30-14:00	Lunch (on your own)							
14:00-15:30	Tutorial «Mining complex dynamic data»	Tutorial «Factorizing Gigantic Matrices»	Workshop «LSHTC»	Workshop «ISEW»	Workshop «MUSE»			
15:30-16:00	Coffee Break							
16:00-17:30	Tutorial «Mining complex dynamic data»	Tutorial «Factorizing Gigantic Matrices»	Workshop «LSHTC»	Workshop «ISEW»	Workshop «MUSE»			
19:00-19:30	Openings & Awards Ceremony (Olympia Hall)							
19:30-20:30	Invited Talk by Christopher Bishop (Olympia Hall)							
20:30	Welcome Reception (Ioannis Restaurant - Roof Garden)							

10. Tuesday at a glance

Tuesday 6 at a glance			
	Olympia Hall	Attica Hall	Kallirhoe Hall
09:00-09:10	Announcements		
09:10-10:10	Invited Talk by Rakesh Agrawal (Olympia Hall)		
10:10-10:40	Coffee Break		
10:40-12:30	S1. Classification & Prediction	S2. Frequent Sets and Patterns	S3. Active & Online learning
12:30-14:00	Lunch		
14:00-15:50	S4. Classification & Bayesian Networks	S5. Applications of Data Mining	S6. Ensemble Learning
15:50-16:20	Coffee Break		
16:20-18:10	S7. Clustering	S8. Matrix and Tensor Analysis	S9. Learning from Time Series Data
19:30	Poster Reception (Technopolis)		

11. Wednesday at a glance

Wednesday 7 at a glance			
	Olympia Hall	Attica Hall	Kallirhoe Hall
09:00-09:10	Announcements		
09:10-10:10	Invited Talk by Albert-László Barabási (Olympia Hall)		
10:10-10:40	Coffee Break		
10:40-12:30	S10. Learning from Social and Information Networks I	S11. Spectral Clustering & Graph Mining	S12. Data Mining Theory & Foundations
12:30-14:00	Lunch		
14:00-15:50	S13. Learning from Social and Information Networks II	S14. Relational learning and Inductive Logic Programming	S15. Model Selection & Statistical Learning
15:50-16:20	Coffee Break		
16:20-18:10	S16. Graphical & Hidden Markov Models	S17. Supervised Learning I	S18. Unsupervised Learning & Dimensionality Reduction
19:30-20:00	10-years Award Ceremony (Olympia Hall)		
20:00-21:00	Invited Talk by Andrei Broder (Olympia Hall)		
21:00	Conference Banquet (Ioannis Restaurant - Roof Garden)		

12. Thursday at a glance

Thursday 8 at a glance			
	Olympia Hall	Attica Hall	Kallirhoe Hall
09:00-09:10	Announcements		
09:10-11:00	S19. Supervised Learning II	S20. Semi-Supervised and Transductive Learning	S21. Preference Learning and Ranking
11:00-11:30	Coffee Break		
11:30-12:30	Invited Talk by Marco Gori (Olympia Hall)		
12:30-14:00	Lunch		
14:00-15:50	S22. Feature Selection, Extraction, and Construction	S23. Text Mining & Recommender Systems	S24. Reinforcement learning
15:50-16:20	Coffee Break		
16:20-17:20	Demos (Attica Hall)		
17:20-19:00	Community Meeting (Olympia Hall)		
20:00	Guided Tour at the Acropolis Museum Farewell party		

13. Friday at a glance

Friday 9 at a glance							
	Olympia Hall	Attica Hall	Conference Room I	Conference Room II	Conference Room III	Conference Room IV	Conference Room V
09:00-10:00	Invited Talk by Heikki Mannila (Olympia Hall)						
10:00-10:30	Coffee Break						
10:30-13:45	Industrial Session	Tutorial «Mining Complex Entities from Heterogeneous Information Networks»	Workshop «CoLISD»	Workshop «PlanSoKD»	Workshop «KD-HCM»	Workshop «MLDMG»	Workshop «NEMO»
13:45-15:00	Lunch (on your own)						
15:00-16:30	Industrial Session	Tutorial «Semantic Data Mining»	Workshop «CoLISD»	Workshop «MIND»	Workshop «KD-HCM»	Workshop «MLDMG»	Workshop «NEMO»
16:30-17:00	Coffee Break						
17:00-18:30	Industrial Session	Tutorial «Semantic Data Mining»	Workshop «CoLISD»	Workshop «MIND»	Workshop «KD-HCM»	Workshop «MLDMG»	Workshop «NEMO»

14. Invited Talks

Embracing Uncertainty: Applied Machine Learning Comes of Age **Christopher Bishop, Microsoft Research Labs, Cambridge, UK**

Location: Monday 5th, 19:30, Olympia Hall



Abstract: Over the last decade the number of deployed applications of machine learning has grown rapidly, with examples in domains ranging from recommendation systems and web search, to spam filters and voice recognition. Most recently, the Kinect 3D full-body motion sensor, which relies crucially on machine learning, has become the fastest-selling consumer electronics device in history. Developments such as the advent of widespread internet connectivity, with its centralisation of data storage, as well as new algorithms for computationally efficient probabilistic inference, will create many new opportunities for machine learning over the coming years. The talk will be illustrated with tutorial examples, live demonstrations, and real-world case studies.

Bio: Chris Bishop is a Distinguished Scientist at Microsoft Research Cambridge, where he leads the Machine Learning and Perception group. He is also Professor of Computer Science at the University of Edinburgh, and Vice President of the Royal Institution of Great Britain. He is a Fellow of the Royal Academy of Engineering, a Fellow of the Royal Society of Edinburgh, and a Fellow of Darwin College Cambridge. His research interests include probabilistic approaches to machine learning, as well as their practical application. Chris is the author of the leading textbook “Neural Networks for Pattern Recognition” (Oxford University Press, 1995) which has over 15,000 citations, and which helped to bring statistical concepts into the mainstream of the machine learning field. His latest textbook “Pattern Recognition and Machine Learning” (Springer, 2006) has over 4,000 citations, and has been widely adopted. In 2008 he presented the 180th series of annual Royal Institution Christmas Lectures, with the title “Hi-tech Trek: the Quest for the Ultimate Computer”, to a television audience of close to 5 million.

Enriching Education Through Data Mining

Rakesh Agrawal, Microsoft Search Labs, California, USA

Location: Tuesday 6th, 9:10, Olympia Hall



Abstract: Education is acknowledged to be the primary vehicle for improving the economic well-being of people [1,6]. Textbooks have a direct bearing on the quality of education imparted to the students as they are the primary conduits for delivering content knowledge [9]. They are also indispensable for fostering teacher learning and constitute a key component of the ongoing professional development of the teachers [5,8].

Many textbooks, particularly from emerging countries, lack clear and adequate coverage of important concepts [7]. In this talk, we present our early explorations into developing a data mining based approach for enhancing the quality of textbooks. We discuss techniques for algorithmically augmenting different sections of a book with links to selective content mined from the Web. For finding authoritative articles, we first identify the set of key concept phrases contained in a section. Using these phrases, we find web (Wikipedia) articles that represent the central concepts presented in the section and augment the section with links to them [4]. We also describe a framework for finding images that are most relevant to a section of the textbook, while respecting global relevancy to the entire chapter to which the section belongs. We pose this problem of matching images to sections in a textbook chapter as an optimization problem and present an efficient algorithm for solving it [2].

We also present a diagnostic tool for identifying those sections of a book that are not well-written and hence should be candidates for enrichment. We propose a probabilistic decision model for this purpose, which is based on syntactic complexity of the writing and the newly introduced notion of the dispersion of key concepts mentioned in the section. The model is learned using a tune set which is automatically generated in a novel way. This procedure maps sampled text book sections to the closest versions of Wikipedia articles having similar content and uses the maturity of those versions to assign need-for-enrichment labels. The maturity of a version is computed by considering the revision history of the corresponding Wikipedia article and convolving the changes in size with a smoothing filter [3].

We also provide the results of applying the proposed techniques to a corpus of widely-used, high school textbooks published by the National Council of Educational Research and Training (NCERT), India. We consider books from grades IX-XII, covering four broad subject areas, namely, Sciences, Social Sciences, Commerce, and Mathematics. The preliminary results are encouraging and indicate that developing technological approaches to enhancing the quality of textbooks could be a promising direction for research for our field.

[1] World Bank Knowledge for Development. World Development Report 1998/99, 1998.

[2] R.Agrawal, S.Gollapudi, A.Kannan, and K.Kenthapadi. Enriching textbooks with web images. Working paper, 2011.

[3] R.Agrawal, S.Gollapudi, A.Kannan, and K.Kenthapadi. Identifying enrichment candidates in textbooks. In WWW, 2011.

[4] R.Agrawal, S.Gollapudi, K.Kenthapadi, N.Srivastava, and R.Velu. Enriching textbooks through data mining. In First Annual ACM Symposium on Computing for Development (ACM DEV), 2010.

- [5] J.Gillies and J.Quijada. Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries. USAID Educational Quality Improvement Program (EQUIP2), 2008.
- [6] E.A. Hanushek and L.Woessmann. The role of education quality for economic growth. Policy Research Department Working Paper 4122, World Bank, 2007.
- [7] R.Mohammad and R.Kumari. Effective use of textbooks: A neglected aspect of education in Pakistan. Journal of Education for International Development, 3(1), 2007.
- [8] J.Oakes and M.Saunders. Education's most basic tools: Access to textbooks and instructional materials in California's public schools. Teachers College Record, 106(10), 2004.
- [9] M.Stein, C.Stuen, D.Carnine, and R.M. Long. Textbook evaluation and adoption. Reading & Writing Quarterly, 17(1), 2001.

Bio: Dr. Rakesh Agrawal is a Microsoft Technical Fellow working at the Search Labs in Microsoft Research in Silicon Valley.

Rakesh is a Member of the National Academy of Engineering, a Fellow of ACM, and a Fellow of IEEE. He is the recipient of the 2010 IIT-Roorkee Distinguished Alumni Award, ACM-SIGKDD First Innovation Award, ACM-SIGMOD Edgar F. Codd Innovations Award, ACM-SIGMOD Test of Time Award, VLDB 10-Yr Most Influential Paper Award, and the Computerworld First Horizon Award. Scientific American named him to the list of 50 top scientists and technologists in 2003.

Rakesh has been granted more than 60 patents and has published more than 150 research papers. He has written the first and second highest cited papers in the fields of databases and data mining. His work has been featured in New York Times Year in Review, New York Times Science section, and several other publications.

Before joining Microsoft in March 2006, Rakesh worked as an IBM Fellow at the IBM Almaden Research Center. Earlier, he was with the Bell Laboratories, Murray Hill from 1983 to 1989. He also worked for three years at the Bharat Heavy Electricals Ltd. in India. He received the M.S. and Ph.D. degrees in Computer Science from the University of Wisconsin-Madison in 1983. He also holds a B.E. degree in Electronics and Communication Engineering from IIT-Roorkee, and a two-year Post Graduate Diploma in Industrial Engineering from the National Institute of Industrial Engineering (NITIE), Bombay.

Human Dynamics: From Human Mobility to Predictability

Albert-László Barabási, Center of Complex Networks Research,

Northeastern University and Department of Medicine, Harvard University.

Location: Wednesday 7th, 9:10, Olympia Hall



Abstract: A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to understand human activity patterns. I will discuss recent effort to explore human activity patterns, using the mobility of individuals as a proxy. As an application, I will show that by measuring the entropy of each individual's trajectory, we find can explore the underlying predictability of human mobility, raising fundamental questions on how predictable we really are. I will also discuss the interplay between human mobility, social links, and the predictive power of data mining.

Bio: Albert-László Barabási is a Distinguished University Professor at Northeastern University, where he directs the Center for Complex Network Research, and holds appointments in the Departments of Physics, Computer Science and Biology, as well as in the Department of Medicine, Harvard Medical School and Brigham and Women Hospital, and is a member of the Center for Cancer Systems Biology at Dana Farber Cancer Institute. A Hungarian born native of Transylvania, Romania, he received his Masters in Theoretical Physics at the Eötvös University in Budapest, Hungary and was awarded a Ph.D. three years later at Boston University. After a year at the IBM T.J. Watson Research Center, he joined Notre Dame as an Assistant Professor, and in 2001 was promoted to the Professor and the Emil T. Hofman Chair. Barabási recently released on April 29th his newest book “Bursts: The Hidden Pattern Behind Everything We Do” (Dutton, 2010) available in five languages. He has also authored “Linked: The New Science of Networks” (Perseus, 2002), currently available in eleven languages, is co-author of “Fractal Concepts in Surface Growth” (Cambridge, 1995), and the co-editor of “The Structure and Dynamics of Networks” (Princeton, 2005). His work lead to the discovery of scale-free networks in 1999, and proposed the Barabási-Albert model to explain their widespread emergence in natural, technological and social systems, from the cellular telephone to the WWW or online communities. His work on complex networks have been widely featured in the media, including the cover of Nature, Science News and many other journals, and written about in Science, Science News, New York Times, USA Today, Washington Post, American Scientist, Discover, Business Week, Die Zeit, El Pais, Le Monde, London's Daily Telegraph, National Geographic, The Chronicle of Higher Education, New Scientist, and La Repubblica, among others. He has been interviewed by BBC Radio, National Public Radio, CBS and ABC News, CNN, NBC, and many other media outlets.

Highly dimensional problems in Computational Advertising

Andrei Broder, Yahoo! Research

Location: Wednesday 7th, 19:30, Olympia Hall



Abstract: The central problem of Computational Advertising is to find the “best match” between a given user in a given context and a suitable advertisement. The context could be a user entering a query in a search engine (“sponsored search”), a user reading a web page (“content match” and “display ads”), a user interacting with a portable device, and so on. The information about the user can vary from scarily detailed to practically nil. The number of potential advertisements might be in the

billions. The number of contexts is unbound. Thus, depending on the definition of “best match” this problem leads to a variety of massive optimization and search problems, with complicated constraints. The solution to these problems provides the scientific and technical underpinnings of the online advertising industry, an industry estimated to surpass 28 billion dollars in US alone in 2011.

An essential aspect of this problem is predicting the impact of an ad on users’ behavior, whether immediate and easily quantifiable (e.g. clicking on ad or buying a product on line) or delayed and harder to measure (e.g. off-line buying or changes in brand perception). To this end, the three components of the problem -- users, contexts, and ads -- are represented as high dimensional objects and terabytes of data documenting the interactions among them are collected every day. Nevertheless, considering the representation difficulty, the dimensionality of the problem and the rarity of the events of interest, the prediction problem remains a huge challenge.

The goal of this talk is twofold: to present a short introduction to Computational Advertising and survey several high dimensional problems at the core of this emerging scientific discipline.

Bio: Andrei Broder is a Yahoo! Fellow and Vice President for Computational Advertising. Previously he was an IBM Distinguished Engineer and the CTO of the Institute for Search and Text Analysis in IBM Research. From 1999 until 2002 he was Vice President for Research and Chief Scientist at the AltaVista Company. He was graduated Summa cum Laude from Technion, the Israeli Institute of Technology, and obtained his M.Sc. and Ph.D. in Computer Science at Stanford University under Don Knuth. His current research interests are centered on computational advertising, web search, context-driven information supply, and randomized algorithms. He has authored more than a hundred papers and was awarded thirty patents. He is a member of the US National Academy of Engineering, a fellow of ACM and of IEEE, and past chair of the IEEE Technical Committee on Mathematical Foundations of Computing.

Learning from constraints

Marco Gori, University of Siena, Italy

Location: Thursday 8th, 11:40, Olympia Hall



Abstract: In this talk, I propose a functional framework to understand the emergence of intelligence in agents exposed to examples and knowledge granules. The theory is based on the abstract notion of constraint, which provides a representation of knowledge granules gained from the interaction with the environment. I give a picture of the “agent body” in terms of representation theorems by extending the classic framework of kernel machines in such a way to incorporate logic formalisms, like first-order logic. This is made possible by the unification of continuous and discrete computational mechanisms in the same functional framework, so as any stimulus, like supervised examples and logic predicates, is translated into a constraint. The learning, which is based on constrained variational calculus, is either guided by a parsimonious match of the constraints or by unsupervised mechanisms expressed in terms of the minimization of the entropy.

I show some experiments with different kinds of symbolic and sub-symbolic constraints, and then I give insights on the adoption of the proposed framework in computer vision. It is shown that in most interesting tasks the learning from constraints naturally leads to “deep architectures”, that emerge when following the developmental principle of focusing attention on “easy constraints”, at each stage. Interestingly, this suggests that stage-based learning, as discussed in developmental psychology, might not be primarily the outcome of biology, but it could be instead the consequence of optimization principles and complexity issues that hold regardless of the “body.”

Bio: Marco Gori received the Ph.D. degree in 1990 from Università di Bologna, Italy, working partly at the School of Computer Science (McGill University, Montreal). In 1992, he became an Associate Professor of Computer Science at Università di Firenze and, in November 1995, he joined the Università di Siena, where he is currently full professor of computer science.

His main interests are in machine learning with applications to pattern recognition, Web mining, and game playing. He is especially interested in bridging logic and learning and in the connections between symbolic and sub-symbolic representation of information. He is the leader of the WebCrow project for automatic solving of crosswords that outperformed human competitors in an official competition which took place within the ECAI-06 conference. As a follow up of this grand challenge, he founded QuestIt, a spin-off company of the University of Siena, working in the field of question-answering. He is co-author of the book “Web Dragons: Inside the myths of search engines technologies,” Morgan Kauffman (Elsevier), 2006.

Dr. Gori serves (has served) as an Associate Editor of a number of technical journals related to his areas of expertise, he has been the recipient of best paper awards, and keynote speakers in a number of international conferences. He was the Chairman of the Italian Chapter of the IEEE Computational Intelligence Society, and the President of the Italian Association for Artificial Intelligence. He is in the list of top Italian scientists kept by VIA-Academy (http://www.topitalianscientists.org/top_italian_scientists.aspx) based on the h-index and he is a fellow of the ECCAI and of the IEEE.

Permutation structure in 0-1 data

Heikki Mannila, Department of Computer Science, University of Helsinki, Finland

Location: Friday 9th, 9:00, Olympia Hall



Abstract: Multidimensional 0-1 data occurs in many domains. Typically one assumes that the order of rows and columns has no importance. However, in some applications, e.g., in ecology, there is structure in the data that becomes visible only when the rows and columns are permuted in a certain way. Examples of such structure are different forms of nestedness and bandedness. I review some of the applications, intuitions, results, and open problems in this area.

Bio: Prof. Heikki Mannila received his Ph.D. in computer science in 1985 from the University of Helsinki. He has been a professor of computer science at the University of Helsinki and Helsinki University of Technology, and a researcher at Microsoft Research and Nokia Research Centre. Currently, he is vice president for academic affairs at Aalto University. His research area of is algorithms for data analysis, and applications in science and in industry. Heikki Mannila is the author of two books and over 190 refereed articles in computer science and related areas. He received the ACM SIGKDD Innovation award in 2003 and the IEEE ICDM research contributions award in 2009.

15. Industrial Session

ECML PKDD 2011 Industrial Session will consist of invited presentations on selected topics in machine learning and data mining from industry perspective. It will be held in Olympia Hall, on Friday 9th, starting at 10:30:

- 10:30-11:30: Olivier Verscheure, IBM Dublin Research Lab, Smarter Cities Technology Centre
- 11:30-12:30: Vasilis Aggelis, Pireus Bank
- 12:30-12:45: *Short Break*
- 12:45-13:45: Radu Jurca, Google Maps, Zurich
- 13:45-15:00: *Lunch Break (on your own)*
- 15:00-16:00: Neel Sundaresan, Head of eBay Research Labs

Speakers and talks

Olivier Verscheure, IBM Dublin Research Lab, Smarter Cities Technology Centre

Smart Cities: How Data Mining and Optimization Can Shape Future Cities

By 2050, an estimated 70% of the world's population will live in cities – up from 13% in 1900. Already, cities consume an estimated 75% of the world's energy, emit more than 80% of greenhouse gases, and lose as much as 20% of their water supply due to infrastructure leaks. As their urban populations continue to grow and these metrics increase, civic leaders face an unprecedented series of challenges to scale and optimize their infrastructures.

Vasilis Aggelis, Pireus Bank, Greece

Reading Customers' Needs and Expectations with Analytics

Customers are the greatest asset for every bank. Do we know them in whole? Are we ready to fulfill their needs and expectations? Use of analytics is one of the keys in order to make better our relation with customers. In advance, analytics can bring gains both for customers and banks. Customer segmentation, targeted cross- and up-sell campaigns, data mining utilization are tools that drive in great results and contribute to customer centric turn.

Radu Jurca, Google Maps, Zurich

Algorithms and Challenges on the GeoWeb

A substantial number of queries addressed nowadays to online search engines have a geographical dimension. People look up addresses on a map, but are also interested in events happening nearby, or inquire information about products, shops or attractions in a particular area. It is no longer enough to index and display geographical information; one should instead geographically organize the world's information. This is the mission of Google's GeoWeb, and several teams inside Google focus on solving this problem. This talk gives an overview of the main challenges characterizing this endeavor, and offers a glimpse into some of the solutions we built.

Neel Sundaresan, Head of eBay Research Labs

Data Science and Machine Learning at Scale

Large Social Commerce Network sites like eBay have to constantly grapple with building scalable machine learning algorithms for search ranking, recommender systems, classification, and others. Large data availability is both a boon and curse. While it offers a lot more diverse observation, the same diversity with sparsity and lack of reliable labeled data at scale introduces new challenges. Also, availability of large data helps take advantage of correlational factors while requiring creativity in discarding irrelevant data. In this talk we will discuss all of this and more from the context of eBay's large data problems.

16. Awards

ECML PKDD 10 years award

As ECML PKDD celebrates already its 10th joint organization, it was decided to establish the "ECML PKDD 10 year award" offering a prize to the paper that was published 10 years ago in the proceedings of ECML PKDD and proved to be important in terms of scientific or other impact.

The inaugural award will be given to Peter Turney for his paper: "*Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL*", published in the ECML/PKDD 2001 proceedings.



Abstract: This paper presents a simple unsupervised learning algorithm for recognizing synonyms, based on statistical data acquired by querying a Web search engine. The algorithm, called PMI-IR, uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words. PMI-IR is empirically evaluated using 80 synonym test questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym test questions from a collection of tests for students of English as a Second Language (ESL). On both tests, the algorithm obtains a score of 74%. PMI-

IR is contrasted with Latent Semantic Analysis (LSA), which achieves a score of 64% on the same 80 TOEFL questions. The paper discusses potential applications of the new unsupervised learning algorithm and some implications of the results for LSA and LSI (Latent Semantic Indexing).

17. Technical Sessions

Session 1: Classification & Prediction

Tuesday 6, 10:40 - 12:30, Olympia Hall

Differentiating Code from Data in x86 Binaries

Richard Wartell, Yan Zhou, Kevin W. Hamlen, Murat Kantarcioglu, Bhavani Thuraisingham

Focused Multi-task Learning Using Gaussian Processes

Gayle Leen, Jaakko Peltonen, Samuel Kaski

On the Stratification of Multi-Label Data

Konstantinos Sechidis, Grigorios Tsoumakas, Ioannis Vlahavas

Learning Monotone Nonlinear Models using the Choquet Integral

Ali Fallah Tehrani, Weiwei Cheng, Krzysztof Dembczyński, Eyke Hüllermeier

Compact Coding for Hyperplane Classifiers in Heterogeneous Environment

Hao Shao, Bin Tong, Einoshin Suzuki

Session 2: Frequent Sets and Patterns

Tuesday 6, 10:40 - 12:30, Attica Hall

Fast and Memory-Efficient Discovery of the Top-k Relevant Subgroups in a Reduced Candidate Space

Henrik Grosskreutz, Daniel Paurat

Constrained Logistic Regression for Discriminative Pattern Mining

Rajul Anand, Chandan K. Reddy

Mining Actionable Partial Orders in Collections of Sequences

Robert Gwadera, Gianluca Antonini, Abderrahim Labbi

Efficient Mining of Top Correlated Patterns Based on Null-Invariant Measures

Sangkyum Kim, Marina Barsky, Jiawei Han

Non-Redundant Subgroup Discovery in Large and Complex Data

Matthijs van Leeuwen, Arno Knobbe

Session 3: Active & Online learning

Tuesday 6, 10:40 - 12:30, Kallirhoe Hall

Frequency-aware Truncated methods for Sparse Online Learning

Hidekazu Oiwa, Shin Matsushima, Hiroshi Nakagawa

Discriminative Experimental Design

Yu Zhang, Dit-Yan Yeung

Manifold Coarse Graining for Online Semi-Supervised Learning

Mehrdad Farajtabar, Amirreza Shaban, Hamid Rabiee, Mohammad Hossein Rohban

Active learning with evolving streaming data

Indrè Žliobaitė, Albert Bifet, Bernhard Pfahringer, Geoff Holmes

Online Structure Learning for Markov Logic Networks

Tuyen N. Huynh, Raymond J. Mooney

Session 4: Classification & Bayesian Networks

Tuesday 6, 14:00 - 15:50, Olympia Hall

Ancestor Relations in the Presence of Unobserved Variables

Pekka Parviainen, Mikko Koivisto

A Robust Ranking Methodology based on Diverse Calibration of AdaBoost

Róbert Busa-Fekete, Balázs Kégl, Tamás Élétető, György Szarvas

Efficiently approximating Markov tree bagging for high-dimensional density estimation

François Schnitzler, Sourour Ammar, Philippe Leray, Pierre Geurts, Louis Wehenkel

A boosting approach to multiview classification with cooperation

Sokol Koço, Cécile Capponi

ShareBoost: Boosting for Multi-View Learning with Performance Guarantees

Jing Peng, Costin Barbu, Guna Seetharaman, Wei Fan, Xian Wu, Kannappan Palaniappan

Session 5: Applications of Data Mining

Tuesday 6, 14:00 - 15:50, Attica Hall

Image Classification for Age-related Macular Degeneration Screening using Hierarchical Image Decompositions and Graph Mining

Mohd Hanafi Ahmad Hijazi, Frans Coenen, Chuntao Jiang, Yalin Zheng

Label Noise-Tolerant Hidden Markov Models for Segmentation: Application to ECGs

Benoît Frénay, Gaël De Lannoy, Michel Verleysen

Resource-Aware On-Line RFID Localization Using Proximity Data

Christoph Scholz, Martin Atzmueller, Gerd Stumme, Stephan Doerfel, Andreas Hotho

PTMSearch: a Greedy Tree Traversal Algorithm for finding Protein Post-Translational Modifications in Tandem Mass Spectra

Attila Kertész-Farkas, Beáta Reiz, Michael P. Myers, Sándor Pongor

A Novel Framework for Locating Software Faults Using Latent Divergences

Shounak Roychowdhury, Sarfraz Khurshid

Session 6: Ensemble Learning

Tuesday 6, 14:00 - 15:50, Kallirhoe Hall

Tracking Concept Change with Incremental Boosting by Minimization of the Evolving Exponential Loss

Mihajlo Grbovic, Slobodan Vucetic

Aggregating Independent and Dependent Models to Learn Multi-label Classifiers

Elena Montañés, José Ramón Quevedo, Juan José del Coz

Multi-Label Ensemble Learning

Chuan Shi, Xiangnan Kong, Philip S. Yu, Bai Wang

On oblique random forests

Bjoern H. Menze, B. Michael Kelm, Daniel N. Splitthoff, Ullrich Koethe, Fred A. Hamprecht

Novel Fusion Methods for Pattern Recognition

Muhammad Awais, Fei Yan, Krystian Mikolajczyk, Josef Kittler

Session 7: Clustering

Tuesday 6, 16:20 - 18:10, Olympia Hall

The Minimum Code Length for Clustering Using the Gray Code

Mahito Sugiyama, Akihiro Yamamoto

Fast approximate text document clustering using Compressive Sampling

Laurence A. F. Park

Clustering Rankings in the Fourier Domain

Stéphan Cléménçon, Romaric Gaudel, Jérémie Jakubowicz

Is there a best quality metric for graph clusters?

Hélio Almeida, Dorgival Guedes, Wagner Meira Jr, Mohammed Zaki

α -Clusterable Sets

Gerasimos S. Antzoulatos, Michael N. Vrahatis

Session 8: Matrix and Tensor Analysis

Tuesday 6, 16:20 - 18:10, Attica Hall

Tensor Factorization Using Auxiliary Information

Atsuhiko Narita, Kohei Hayashi, Ryota Tomioka, Hisashi Kashima

Bayesian Matrix Co-Factorization: Variational Algorithm and Cramér-Rao Bound

Jiho Yoo, Seungjin Choi

Generalized Dictionary Learning for Symmetric Positive Definite Matrices with Application to Nearest Neighbor Retrieval

Suvrit Sra, Anoop Cherian

Link prediction via matrix factorization

Aditya Krishna Menon, Charles Elkan

Multi-Subspace Representation and Discovery

Dijun Luo, Feiping Nie, Chris Ding, Heng Huang

Session 9: Learning from Time Series Data

Tuesday 6, 16:20 - 18:10, Kallirhoe Hall

Motion segmentation by a model-based clustering approach of incomplete trajectories

Vasileios Karavasiliis, Konstantinos Blekas, Christophoros Nikou

Unsupervised Modeling of Partially Observable Environments

Vincent Graziano, Jan Koutník, Jürgen Schmidhuber

Artemis: Assessing the Similarity of Event-interval Sequences

Orestis Kostakis, Panagiotis Papapetrou, Jaakko Hollmén

Discovering Temporal Bisociations for Linking Concepts over Time

Corrado Loglisci, Michelangelo Ceci

ShiftTree: an Interpretable Model-Based Approach for Time Series Classification

Balázs Hidasi, Csaba Gáspár-Papanek

Session 10: Learning from Social and Information Networks I

Wednesday 7, 10:40 - 12:30, Olympia Hall

Peer and Authority Pressure in Information-Propagation Models

Aris Anagnostopoulos, George Brova, Evimaria Terzi

Active Learning of Model Parameters for Influence Maximization

Tianyu Cao, Xindong Wu, Tony Xiaohua Hu, Song Wang

A Shapley value Approach for Influence Attribution

Panagiotis Papapetrou, Aris Gionis, Heikki Mannila

Influence and Passivity in Social Media

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, Bernardo A. Huberman

Learning Recommendations in Social Media Systems By Weighting Multiple Relations

Boris Chidlovskii

Session 11: Spectral Clustering & Graph Mining

Wednesday 7, 10:40 - 12:30, Attica Hall

Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms

Danai Koutra, Tai-You Ke, U Kang, Duen Horng Polo Chau, Hsing-Kuo Kenneth Pao, Christos Faloutsos

Privacy Preserving Semi-Supervised Learning for Labeled Graphs

Hiroimi Arai, Jun Sakuma

Eigenvector Sensitive Feature Selection For Spectral Clustering

Yi Jiang, Jiangtao Ren

DB-CSC: A density-based approach for subspace clustering in graphs with feature vectors

Stephan Günnemann, Brigitte Boden, Thomas Seidl

Parallel Structural Graph Clustering

Madeleine Seeland, Simon A. Berger, Alexandros Stamatakis, Stefan Kramer

Session 12: Data Mining Theory & Foundations

Wednesday 7, 10:40 - 12:30, Kallirhoe Hall

The VC-Dimension of SQL Queries and Selectivity Estimation Through Sampling

Matteo Riondato, Mert Akdere, Ugur Çetintemel, Stanley B. Zdonik, Eli Upfal

Smooth Receiver Operating Characteristics (smROC) Curves

William Klement, Peter Flach, Nathalie Japkowicz, Stan Matwin

Active Supervised Domain Adaptation

Avishek Saha, Piyush Rai, Hal Daumé III, Suresh Venkatasubramanian, Scott DuVall
Comparing Apples and Oranges - Measuring Differences between Data Mining Results

Nikolaj Tatti, Jilles Vreeken

Learning Good Edit Similarities with Generalization Guarantees

Aurélien Bellet, Amaury Habrard, Marc Sebban

Session 13: Learning from Social and Information Networks II

Wednesday 7, 14:00 - 15:50, Olympia Hall

Toward a Fair Review-Management System

Theodoros Lappas, Evimaria Terzi

Learning to Infer Social Ties in Large Networks

Wenbin Tang, Honglei Zhuang, Jie Tang

A Community-Based Pseudolikelihood Approach for Relationship Labeling in Social Networks

Huaiyu Wan, Youfang Lin, Zhihao Wu, Houkuan Huang

Graph Evolution via Social Diffusion Processes

Dijun Luo, Chris Ding, Heng Huang

Mining Research Topic-related Influence between Academia and Industry

Dan He

Session 14: Relational learning and Inductive Logic Programming

Wednesday 7, 14:00 - 15:50, Attica Hall

Correcting Bias in Statistical Tests for Network Classifier Evaluation

Tao Wang, Jennifer Neville, Brian Gallagher, Tina Eliassi-Rad

Abductive Plan Recognition by Extending Bayesian Logic Programs

Sindhu Raghavan, Raymond J. Mooney

Learning the Parameters of Probabilistic Logic Programs from Interpretations

Bernd Gutmann, Ingo Thon, Luc De Raedt

Gaussian Logic for Predictive Classification

Ondřej Kuželka, Andrea Szabóová, Matěj Holec, Filip Železný

Learning First-Order Definite Theories via Object-Based Queries

Joseph Selman, Alan Fern

Session 15: Model Selection & Statistical Learning

Wednesday 7, 14:00 - 15:50, Kallirhoe Hall

A selecting-the-best method for budgeted model selection

Gianluca Bontempi, Olivier Caelen

Aspects of Semi-Supervised and Active Learning in Conditional Random Fields

Nataliya Sokolovska

Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process

Changyou Chen, Lan Du, Wray Buntine

Comparing Probabilistic Models for Melodic Sequences

Athina Spiliopoulou, Amos Storkey

Multimodal nonlinear filtering using Gauss-Hermite Quadrature

Hannes P. Saal, Nicolas Heess, Sethu Vijayakumar

Session 16: Graphical & Hidden Markov Models

Wednesday 7, 16:20 - 18:10, Olympia Hall

Fourier-Information Duality in the Identity Management Problem

Xiaoye Jiang, Jonathan Huang, Leonidas Guibas

An Alternating Direction Method for Dual MAP LP Relaxation

Ofer Meshi, Amir Globerson

Restricted Deep Belief Networks for Multi-View Learning

Yoonseop Kang, Seungjin Choi

A Spectral Learning Algorithm for Finite State Transducers

Borja Balle, Ariadna Quattoni, Xavier Carreras

Common Substructure Learning of Multiple Graphical Gaussian Models

Satoshi Hara, Takashi Washio

Session 17: Supervised Learning I

Wednesday 7, 16:20 - 18:10, Attica Hall

Generalized Agreement Statistics over Fixed Group of Experts

Mohak Shah

Larger Residuals Less Work: Active Document Scheduling for Latent Dirichlet Allocation

Mirwaes Wahabzada, Kristian Kersting

Datum-Wise Classification: A Sequential Approach to Sparsity

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, Patrick Gallinari

Transfer Learning With Adaptive Regularizers

Ulrich Rückert, Marius Kloft

Network Regression with Predictive Clustering Trees

Daniela Stojanova, Michelangelo Ceci, Annalisa Appice, Sašo Džeroski

Learning from Inconsistent and Unreliable Annotators by a Gaussian Mixture Model and Bayesian Information Criterion

Ping Zhang, Zoran Obradovic

Session 18: Unsupervised Learning & Dimensionality Reduction

Wednesday 7, 16:20 - 18:10, Kallirhoe Hall

Minimum Neighbor Distance Estimators of Intrinsic Dimension

Gabriele Lombardi, Alessandro Rozza, Claudio Ceruti, Elena Casiraghi, Paola Campadelli

Online Clustering of High-Dimensional Trajectories under Concept Drift

Georg Kreml, Zaigham Faraz Siddiqui, Myra Spiliopoulou

Linear Discriminant Dimensionality Reduction

Quanquan Gu, Zhenhui Li, Jiawei Han

The Minimum Transfer Cost Principle for Model-Order Selection

Mario Frank, Morteza Haghir Chehreghani, Joachim Buhmann

Higher Order Contractive auto-encoder

Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, Xavier Glorot

Session 19: Supervised Learning II

Thursday 8, 9:10 - 11:00, Olympia Hall

Regularized Sparse Kernel Slow Feature Analysis

Wendelin Böhmer, Steffen Grünewälder, Hannes Nickisch, Klaus Obermayer

Kernels for Link Prediction with Latent Feature Models

Canh Hao Nguyen, Hiroshi Mamitsuka

PerTurbo: a new classification algorithm based on the spectrum perturbations of the Laplace-Beltrami operator

Nicolas Courty, Thomas Burger, Johann Laurent

Fast Support Vector Machines for Structural Kernels

Aliaksei Severyn, Alessandro Moschitti

Building Sparse Support Vector Machines for Multi-Instance Classification

Zhouyu Fu, Guojun Lu, Kai Ming Ting, Dengsheng Zhang

Session 20: Semi-Supervised and Transductive Learning

Thursday 8, 9:10 - 11:00, Attica Hall

Learning from Label Proportions by Optimizing Cluster Model Selection

Marco Stolpe, Katharina Morik

Adaptive Boosting for Transfer Learning using Dynamic Updates

Samir Al-Stouhi, Chandan K. Reddy

Learning from Partially Annotated Sequences

Eraldo R. Fernandes, Ulf Brefeld

Constraint selection for semi-supervised topological clustering

Kais Allab, Khalid Benabdeslem

COSNet: a Cost Sensitive Neural Network for Semi-supervised Learning in Graphs

Alberto Bertoni, Marco Frasca, Giorgio Valentini

Session 21: Preference Learning and Ranking

Thursday 8, 9:10 - 11:00, Kallirhoe Hall

Direct Policy Ranking with Robot Data Streams

Riad Akrou, Marc Schoenauer, Michèle Sebag

Multiview Semi-Supervised Learning for Ranking Multilingual Documents

Nicolas Usunier, Massih-Reza Amini, Cyril Goutte

Preference-based policy iteration: Leveraging preference learning for reinforcement learning

Weiwei Cheng, Johannes Fürnkranz, Eyke Hüllermeier, Sang-Hyeun Park

Rule-Based Active Sampling for Learning to Rank

Rodrigo Silva, Marcos Gonçalves, Adriano Veloso

A Geometric Approach to Find Nondominated Policies to Imprecise Reward MDPs

Valdinei Freire da Silva, Anna Helena Reali Costa

Session 22: Feature Selection, Extraction, and Construction

Thursday 8, 14:00 - 15:50, Olympia Hall

Feature Selection Stability Assessment based on the Jensen-Shannon Divergence

Roberto Guzmán-Martínez, Rocío Alaiz-Rodríguez

Fast projections onto $L_{1,q}$ -norm balls for grouped feature selection

Suvrit Sra

A Novel Stability based Feature Selection Framework for k-means Clustering

Dimitrios Mavroeidis, Elena Marchiori

Constrained Laplacian Score for semi-supervised feature selection

Khalid Benabdeslem, Mohammed Hindawi

Feature Selection for Transfer Learning

Selen Uguroglu, Jaime Carbonell

Session 23: Text Mining & Recommender Systems

Thursday 8, 14:00 - 15:50, Attica Hall

Expertise finding using topic models -- the expert--tag--topic model

Gregor Heinrich

Analyzing Word Frequencies in Large Text Corpora using Inter-arrival Times and Bootstrapping

Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, Heikki Mannila

An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering

Nicola Barbieri, Giuseppe Manco

A Game Theoretic Framework for Data Privacy Preservation in Recommender Systems

Maria Halkidi, Iordanis Koutsopoulos

iDVS: An Interactive Multi-Document Visual Summarization System

Yi Zhang, Dingding Wang, Tao Li

Session 24: Reinforcement learning

Thursday 8, 14:00 - 15:50, Kallirhoe Hall



Preference elicitation and inverse reinforcement learning

Constantin Rothkopf, Christos Dimitrakakis

Sparse Kernel-SARSA(λ) with an Eligibility Trace

Matthew Robards, Peter Sunehag, Scott Sanner, Bhaskara Marthi

Analyzing and Escaping Local Optima in Planning as Inference for Partially Observable Domains

Pascal Poupart, Tobias Lang, Marc Toussaint

Lagrange Dual Decomposition for Finite Horizon Markov Decision Processes

Thomas Furnstun, David Barber

Reinforcement Learning Through Global Stochastic Search in N-MDPs

Matteo Leonetti, Luca Iocchi, Subramanian Ramamoorthy

18. Abstracts

Session 1: Classification & Prediction

Tuesday 6, 10:40 - 12:30, Olympia Hall

Differentiating Code from Data in x86 Binaries

Richard Wartell, Yan Zhou, Kevin W. Hamlen, Murat Kantarcioglu, Bhavani Thuraisingham

Robust, static disassembly is an important part of achieving high coverage for many binary code analyses, such as reverse engineering, malware analysis, reference monitor in-lining, and software fault isolation. However, one of the major difficulties current disassemblers face is differentiating code from data when they are interleaved. This paper presents a machine learning-based disassembly algorithm that segments an x86 binary into subsequences of bytes and then classifies each subsequence as code or data. The algorithm builds a language model from a set of pre-tagged binaries using a statistical data compression technique. It sequentially scans a new binary executable and sets a breaking point at each potential code-to-code and code-to-data/data-to-code transition. The classification of each segment as code or data is based on the minimum cross-entropy. Experimental results are presented to demonstrate the effectiveness of the algorithm.

Focused Multi-task Learning Using Gaussian Processes

Gayle Leen, Jaakko Peltonen, Samuel Kaski

Given a learning task for a data set, learning it together with related tasks (data sets) can improve performance. Gaussian process models have been applied to such multi-task learning scenarios, based on joint priors for functions underlying the tasks. In previous Gaussian process approaches, all tasks have been assumed to be of equal importance, whereas in transfer learning the goal is *asymmetric*: to enhance performance on a target task given all other tasks. In both settings, transfer learning and joint modelling, *negative transfer* is a key problem: performance may actually decrease if the tasks are not related closely enough. In this paper, we propose a Gaussian process model for the asymmetric setting, which learns to “explain away” non-related variation in the additional tasks, in order to focus on improving performance on the target task. In experiments, our model

improves performance compared to single-task learning, symmetric multi-task learning using hierarchical Dirichlet processes, and transfer learning based on predictive structure learning.

On the Stratification of Multi-Label Data

Konstantinos Sechidis, Grigorios Tsoumakas, Ioannis Vlahavas

Stratified sampling is a sampling method that takes into account the existence of disjoint groups within a population and produces samples where the proportion of these groups is maintained. In single-label classification tasks, groups are differentiated based on the value of the target variable. In multi-label learning tasks, however, where there are multiple target variables, it is not clear how stratified sampling could/should be performed. This paper investigates stratification in the multi-label data context. It considers two stratification methods for multi-label data and empirically compares them along with random sampling on a number of datasets and based on a number of evaluation criteria. The results reveal some interesting conclusions with respect to the utility of each method for particular types of multi-label datasets.

Learning Monotone Nonlinear Models using the Choquet Integral

Ali Fallah Tehrani, Weiwei Cheng, Krzysztof Dembczyński, Eyke Hüllermeier

The learning of predictive models that guarantee monotonicity in the input variables has received increasing attention in machine learning in recent years. While the incorporation of monotonicity constraints is rather simple for certain types of models, it may become a more intricate problem for others. By trend, the difficulty of ensuring monotonicity increases with the flexibility or, say, nonlinearity of a model. In this paper, we advocate the so-called Choquet integral as a tool for learning monotone nonlinear models. While being widely used as a flexible aggregation operator in different fields, such as multiple criteria decision making, the Choquet integral is much less known in machine learning so far. Apart from combining monotonicity and flexibility in a mathematically sound and elegant manner, the Choquet integral has additional features making it attractive from a machine learning point of view. Notably, it offers measures for quantifying the importance of individual predictor variables and the interaction between groups of variables. As a concrete application of the Choquet integral, we propose a generalization of logistic regression. The basic idea of our approach, referred to as choquistic regression, is to replace the linear function of predictor variables, which is commonly used in logistic regression to model the log odds of the positive class, by the Choquet integral.

Compact Coding for Hyperplane Classifiers in Heterogeneous Environment

Hao Shao, Bin Tong, Einoshin Suzuki

Transfer learning techniques have witnessed a significant development in real applications where the knowledge from previous tasks are required to reduce the high cost of inquiring the labeled information for the target task. However, how to avoid *negative transfer* which happens due to different distributions of tasks in heterogeneous environment is still an open problem. In order to handle this kind of issue, we propose a Compact Coding method for Hyperplane Classifiers (CCHC) under a *two-level* framework in inductive transfer learning setting. Unlike traditional methods, we measure the similarities among tasks from the *macro level* perspective through minimum encoding. Particularly speaking, the degree of the similarity is represented by the relevant code length of the class boundary of each

source task with respect to the target task. In addition, informative parts of the source tasks are adaptively selected in the *micro level* viewpoint to make the choice of the specific source task more accurate. Extensive experiments show the effectiveness of our algorithm in terms of the classification accuracy in both UCI and text data sets.

Session 2: Frequent Sets and Patterns

Tuesday 6, 10:40 - 12:30, Attica Hall

Fast and Memory-Efficient Discovery of the Top-k Relevant Subgroups in a Reduced Candidate Space

Henrik Grosskreutz, Daniel Paurat

We consider a modified version of the top-k subgroup discovery task, where subgroups dominated by other subgroups are discarded. The advantage of this modified task, known as relevant subgroup discovery, is that it avoids redundancy in the outcome. Although it has been applied in many applications, so far no efficient exact algorithm for this task has been proposed. Most existing solutions do not guarantee the exact solution (as a result of the use of non-admissible heuristics), while the only exact solution relies on the explicit storage of the whole search space, which results in prohibitively large memory requirements.

In this paper, we present a new top-k relevant subgroup discovery algorithm which overcomes these shortcomings. Our solution is based on the fact that if an iterative deepening approach is applied, the relevance check - which is the root of the problems of all other approaches - can be realized based solely on the best k subgroups visited so far. The approach also allows for the integration of admissible pruning techniques like optimistic estimate pruning. The result is a fast, memory-efficient algorithm which clearly outperforms existing top-k relevant subgroup discovery approaches. Moreover, we analytically and empirically show that it is competitive with simpler approaches which do not consider the relevance criterion.

Constrained Logistic Regression for Discriminative Pattern Mining

Rajul Anand, Chandan K. Reddy

Analyzing differences in multivariate datasets is a challenging problem. This topic was earlier studied by finding changes in the distribution differences either in the form of patterns representing conjunction of attribute value pairs or univariate statistical analysis for each attribute in order to highlight the differences. All such methods focus only on change in attributes in some form and do not implicitly consider the class labels associated with the data. In this paper, we pose the difference in distribution in a supervised scenario where the change in the data distribution is measured in terms of the change in the corresponding classification boundary. We propose a new constrained logistic regression model to measure such a difference between multivariate data distributions based on the predictive models induced on them. Using our constrained models, we measure the difference in the data distributions using the changes in the classification boundary of these models. We demonstrate the advantages of the proposed work over other methods available in the literature using both synthetic and real-world datasets.

Mining Actionable Partial Orders in Collections of Sequences

Robert Gwadera, Gianluca Antonini, Abderrahim Labbi

Mining frequent partial orders from a collection of sequences was introduced as an alternative to mining frequent sequential patterns in order to provide a more compact/understandable representation. The motivation was that a single partial order can represent the same ordering information between items in the collection as a set of sequential patterns (set of totally ordered sets of items). However, in practice, a discovered set of frequent partial orders is still too large for an effective usage. We address this problem by proposing a method for ranking partial orders with respect to significance that extends our previous work on ranking sequential patterns. In experiments, conducted on a collection of visits to a website of a multinational technology and consulting firm we show the applicability of our framework to discover partial orders of frequently visited webpages that can be actionable in optimizing effectiveness of web-based marketing.

Efficient Mining of Top Correlated Patterns Based on Null-Invariant Measures

Sangkyum Kim, Marina Barsky, Jiawei Han

Mining strong correlations from transactional databases often leads to more meaningful results than mining association rules. In such mining, null (transaction)-invariance is an important property of the correlation measures. Unfortunately, some useful null-invariant measures such as *Kulczynski* and *Cosine*, which can discover correlations even for the very unbalanced cases, lack the (anti)-monotonicity property. Thus, they could only be applied to frequent itemsets as the post-evaluation step. For large datasets and for low supports, this approach is computationally prohibitive. This paper presents new properties for all known null-invariant measures. Based on these properties, we develop efficient pruning techniques and design the Apriori-like algorithm NICOMINER for mining strongly correlated patterns *directly*. We develop both the threshold-bounded and the top-k variations of the algorithm, where top-k is used when the optimal correlation threshold is not known in advance and to give user control over the output size. We test NICOMINER on real-life datasets from different application domains, using *Cosine* as an example of the null-invariant correlation measure. We show that NICOMINER outperforms support-based approach more than an order of magnitude, and that it is very useful for discovering top correlations in itemsets with low support.

Non-Redundant Subgroup Discovery in Large and Complex Data

Matthijs van Leeuwen, Arno Knobbe

Large and complex data is challenging for most existing discovery algorithms, for several reasons. First of all, such data leads to enormous hypothesis spaces, making exhaustive search infeasible. Second, many variants of essentially the same pattern exist, due to (numeric) attributes of high cardinality, correlated attributes, and so on. This causes top-k mining algorithms to return highly redundant result sets, while ignoring many potentially interesting results.

These problems are particularly apparent with Subgroup Discovery and its generalisation, Exceptional Model Mining. To address this, we introduce *subgroup set mining*: one should not consider individual subgroups, but sets of subgroups. We consider three degrees of redundancy, and propose corresponding heuristic selection strategies in order to eliminate redundancy. By incorporating these strategies in a beam search, the balance between exploration and exploitation is improved.

Session 3: Active & Online learning

Tuesday 6, 10:40 - 12:30, Kallirhoe Hall

Frequency-aware Truncated methods for Sparse Online Learning

Hidekazu Oiwa, Shin Matsushima, Hiroshi Nakagawa

Online supervised learning with L_1 -regularization has gained attention recently because it generally requires less computational time and a smaller space of complexity than batch-type learning methods. However, a simple L_1 -regularization method used in an online setting has the side effect that rare features tend to be truncated more than necessary. In fact, feature frequency is highly skewed in many applications. We developed a new family of L_1 -regularization methods based on the previous updates for loss minimization in linear online learning settings. Our methods can identify and retain low-frequency occurrence but informative features at the same computational cost and convergence rate as previous works. Moreover, we combined our methods with a cumulative penalty model to derive more robust models over noisy data. We applied our methods to several datasets and empirically evaluated the performance of our algorithms. Experimental results showed that our frequency-aware truncated models improved the prediction accuracy.

Discriminative Experimental Design

Yu Zhang, Dit-Yan Yeung

Since labeling data is often both laborious and costly, the labeled data available in many applications is rather limited. Active learning is a learning approach which actively selects unlabeled data points to label as a way to alleviate the labeled data deficiency problem. In this paper, we extend a previous active learning method called transductive experimental design (TED) by proposing a new unlabeled data selection criterion. Our method, called discriminative experimental design (DED), incorporates both margin-based discriminative information and data distribution information and hence it can be seen as a discriminative extension of TED. We report experiments conducted on some benchmark data sets to demonstrate the effectiveness of DED.

Manifold Coarse Graining for Online Semi-Supervised Learning

Mehrdad Farajtabar, Amirreza Shaban, Hamid Rabiee, Mohammad Hossein Rohban

When the number of labeled data is not sufficient, Semi-Supervised Learning (SSL) methods utilize unlabeled data to enhance classification. Recently, many SSL methods have been developed based on the manifold assumption in a batch mode. However, when data arrive sequentially and in large quantities, both computation and storage limitations become a bottleneck. In this paper, we present a new semi-supervised coarse graining (CG) algorithm to reduce the required number of data points for preserving the manifold structure. First, an equivalent formulation of Label Propagation (LP) is derived. Then a novel spectral view of the Harmonic Solution (HS) is proposed. Finally an algorithm to reduce the number of data points while preserving the manifold structure is provided and a theoretical analysis on preservation of the LP properties is presented. Experimental results on real world datasets show that the proposed method outperforms the state of the art coarse graining algorithm in different settings.

Active learning with evolving streaming data

Indrè Žliobaitė, Albert Bifet, Bernhard Pfahringer, Geoff Holmes

In learning to classify streaming data, obtaining the true labels may require major effort and may incur excessive cost. Active learning focuses on learning an accurate model with as few labels as possible. Streaming data poses additional challenges for active learning, since the data distribution may change over time (concept drift) and classifiers need to adapt. Conventional active learning strategies concentrate on querying the most uncertain instances, which are typically concentrated around the decision boundary. If changes do not occur close to the boundary, they will be missed and classifiers will fail to adapt. In this paper we develop two active learning strategies for streaming data that explicitly handle concept drift. They are based on uncertainty, dynamic allocation of labeling efforts over time and randomization of the search space. We empirically demonstrate that these strategies react well to changes that can occur anywhere in the instance space and unexpectedly.

Online Structure Learning for Markov Logic Networks

Tuyen N. Huynh, Raymond J. Mooney

Most existing learning methods for Markov Logic Networks (MLNs) use batch training, which becomes computationally expensive and eventually infeasible for large datasets with thousands of training examples which may not even all fit in main memory. To address this issue, previous work has used online learning to train MLNs. However, they all assume that the model's structure (set of logical clauses) is given, and only learn the model's parameters. However, the input structure is usually incomplete, so it should also be updated. In this work, we present OSL-the first algorithm that performs both online structure and parameter learning for MLNs. Experimental results on two real-world datasets for natural-language field segmentation show that OSL outperforms systems that cannot revise structure.

Session 4: Classification & Bayesian Networks

Tuesday 6, 14:00 - 15:50, Olympia Hall

Ancestor Relations in the Presence of Unobserved Variables

Pekka Parviainen, Mikko Koivisto

Bayesian networks (BNs) are an appealing model for causal and non-causal dependencies among a set of variables. Learning BNs from observational data is challenging due to the nonidentifiability of the network structure and model misspecification in the presence of unobserved (latent) variables. Here, we investigate the prospects of Bayesian learning of ancestor relations, including arcs, in the presence and absence of unobserved variables. An exact dynamic programming algorithm to compute the respective posterior probabilities is developed, under the complete data assumption. Our experimental results show that ancestor relations between observed variables, arcs in particular, can be learned with good power even when a majority of the involved variables are unobserved. For comparison, deduction of ancestor relations from single maximum a posteriori network structures or their Markov equivalence class appears somewhat inferior to Bayesian averaging. We also discuss some shortcomings of applying existing conditional independence test based methods for learning ancestor relations.

A Robust Ranking Methodology based on Diverse Calibration of AdaBoost

Róbert Busa-Fekete, Balázs Kégl, Tamás Éllető, György Szarvas

In *subset ranking*, the goal is to learn a ranking function that approximates a gold standard partial ordering of a set of objects (in our case, relevance labels of a set of documents retrieved for the same query). In this paper we introduce a learning to rank approach to subset ranking based on multi-class classification. Our technique can be summarized in three major steps. First, a multi-class classification model (AdaBoost.MH) is trained to predict the relevance label of each object. Second, the trained model is calibrated using various calibration techniques to obtain diverse class probability estimates. Finally, the Bayes-scoring function (which optimizes the popular Information Retrieval performance measure NDCG), is approximated through mixing these estimates into an ultimate scoring function. An important novelty of our approach is that many different methods are applied to estimate the same probability distribution, and all these hypotheses are combined into an improved model. It is well known that mixing different conditional distributions according to a prior is usually more efficient than selecting one “optimal” distribution. Accordingly, using all the calibration techniques, our approach does not require the estimation of the best suited calibration method and is therefore less prone to overfitting. In an experimental study, our method outperformed many standard ranking algorithms on the LETOR benchmark datasets, most of which are based on significantly more complex learning to rank algorithms than ours.

Efficiently approximating Markov tree bagging for high-dimensional density estimation

François Schnitzler, Sourour Ammar, Philippe Leray, Pierre Geurts, Louis Wehenkel

We consider algorithms for generating *Mixtures of Bagged Markov Trees*, for density estimation. In problems defined over many variables and when few observations are available, those mixtures generally outperform a single Markov tree maximizing the data likelihood, but are far more expensive to compute. In this paper, we describe new algorithms for approximating such models, with the aim of *speeding up learning without sacrificing accuracy*. More specifically, we propose to use a filtering step obtained as a by-product from computing a first Markov tree, so as to avoid considering poor candidate edges in the subsequently generated trees. We compare these algorithms (on synthetic data sets) to Mixtures of Bagged Markov Trees, as well as to a single Markov tree derived by the classical Chow-Liu algorithm and to a recently proposed randomized scheme used for building tree mixtures.

A boosting approach to multiview classification with cooperation

Sokol Koço, Cécile Capponi

In many fields, such as bioinformatics or multimedia, data may be described using different sets of features (or views) which carry either global or local information. Some learning tasks make use of these several views in order to improve overall predictive power of classifiers through fusion-based methods. Usually, these approaches rely on a weighted combination of classifiers (or selected descriptions), where classifiers are learned independently. One drawback of these methods is that the classifier learned on one view does not communicate its failures within the other views. This paper deals with a novel approach to integrate multiview information. The proposed algorithm, named Mumbo, is based on boosting. Within the boosting scheme, Mumbo maintains one distribution

of examples on each view, and at each round, it learns one weak classifier on each view. Within a view, the distribution of examples evolves both with the ability of the dedicated classifier to deal with examples of the corresponding features space, and with the ability of classifiers in other views to process the same examples within their own description spaces. Hence, the principle is to slightly remove the hard examples from the learning space of one view, while their weights get higher in the other views. This way, we expect that examples are urged to be processed by the most appropriate views, when possible. At the end of the iterative learning process, a final classifier is computed by a weighted combination of selected weak classifiers.

This paper provides the Mumbo algorithm in a multiclass and multiview setting, based on recent theoretical advances in boosting. The boosting properties of Mumbo are proved, as well as some results on its generalization capabilities. Several experimental results are reported which point out that complementary views may actually cooperate under some assumptions.

ShareBoost: Boosting for Multi-View Learning with Performance Guarantees

Jing Peng, Costin Barbu, Guna Seetharaman, Wei Fan, Xian Wu, Kannappan Palaniappan

Algorithms combining multi-view information are known to exponentially quicken classification, and have been applied to many fields. However, they lack the ability to mine most discriminant information sources (or data types) for making predictions. In this paper, we propose an algorithm based on boosting to address these problems. The proposed algorithm builds base classifiers independently from each data type (view) that provides a partial view about an object of interest. Different from AdaBoost, where each view has its own re-sampling weight, our algorithm uses a single re-sampling distribution for all views at each boosting round. This distribution is determined by the view whose training error is minimal. This shared sampling mechanism restricts noise to individual views, thereby reducing sensitivity to noise. Furthermore, in order to establish performance guarantees, we introduce a randomized version of the algorithm, where a winning view is chosen probabilistically. As a result, it can be cast within a multi-armed bandit framework, which allows us to show that with high probability the algorithm seeks out most discriminant views of data for making predictions. We provide experimental results that show its performance against noise and competing techniques.

Session 5: Applications of Data Mining

Tuesday 6, 14:00 - 15:50, Attica Hall

Image Classification for Age-related Macular Degeneration Screening using Hierarchical Image Decompositions and Graph Mining

Mohd Hanafi Ahmad Hijazi, Frans Coenen, Chuntao Jiang, Yalin Zheng

Age-related Macular Degeneration (AMD) is the most common cause of adult blindness in the developed world. This paper describes a new image mining technique to perform automated detection of AMD from colour fundus photographs. The technique comprises a novel hierarchical image decomposition mechanism founded on a circular and angular partitioning. The resulting decomposition is then stored in a tree structure to which a weighted frequent sub-tree mining algorithm is applied. The identified sub-graphs are

then incorporated into a feature vector representation (one vector per image) to which classification techniques can be applied. The results show that the proposed approach performs both efficiently and accurately.

Label Noise-Tolerant Hidden Markov Models for Segmentation: Application to ECGs

Benôit Fréney, Gaël De Lannoy, Michel Verleysen

The performance of traditional classification models can adversely be impacted by the presence of label noise in training observations. The pioneer work of Lawrence and Schölkopf tackled this issue in datasets with independent observations by incorporating a statistical noise model within the inference algorithm. In this paper, the specific case of label noise in non-independent observations is rather considered. For this purpose, a label noise-tolerant expectation-maximisation algorithm is proposed in the frame of hidden Markov models. Experiments are carried on both healthy and pathological electrocardiogram signals with distinct types of additional artificial label noise. Results show that the proposed label noise-tolerant inference algorithm can improve the segmentation performances in the presence of label noise.

Resource-Aware On-Line RFID Localization Using Proximity Data

Christoph Scholz, Martin Atzmueller, Gerd Stumme, Stephan Doerfel, Andreas Hotho

This paper focuses on resource-aware and cost-effective indoor-localization at room-level using RFID technology. In addition to the tracking information of people wearing active RFID tags, we also include information about their proximity contacts. We present an evaluation using real-world data collected during a conference: We complement state-of-the-art machine learning approaches with strategies utilizing the proximity data in order to improve a core localization technique further.

PTMSearch: a Greedy Tree Traversal Algorithm for finding Protein Post-Translational Modifications in Tandem Mass Spectra

Attila Kertész-Farkas, Beáta Reiz, Michael P. Myers, Sándor Pongor

Peptide identification by tandem mass spectrometry (MS/MS) and database searching is becoming the standard high-throughput technology in many areas of the life sciences. The analysis of post-translational modifications (PTMs) is a major source of complications in this area, which calls for efficient computational approaches. In this paper we describe PTMSearch, a novel algorithm in which the PTM search space is represented by a tree structure, and a greedy traversal algorithm is used to identify a path within the tree that corresponds to the PTMs that best fit the input data. Tests on simulated and real (experimental) PTMs show that the algorithm performs well in terms of speed and accuracy. Estimates are given for the error caused by the greedy heuristics, for the size of the search space and a scheme is presented for the calculation of statistical significance.

A Novel Framework for Locating Software Faults Using Latent Divergences

Shounak Roychowdhury, Sarfraz Khurshid

Fault localization, i.e., identifying erroneous lines of code in a buggy program, is a tedious process, which often requires considerable manual effort and is costly. Recent years have seen much progress in techniques for automated fault localization, specifically using program spectra – executions of failed and passed test runs provide a basis for isolating the faults. Despite the progress, fault localization in large programs remains a challenging

problem, because even inspecting a small fraction of the lines of code in a large problem can require substantial manual effort. This paper presents a novel framework for fault localization based on latent divergences – an effective method for feature selection in machine learning. Our insight is that the problem of fault localization can be reduced to the problem of feature selection, where lines of code correspond to features. We also present an experimental evaluation of our framework using the Siemens suite of subject programs, which are a standard benchmark for studying fault localization techniques in software engineering. The results show that our framework enables more accurate fault localization than existing techniques.

Session 6: Ensemble Learning

Tuesday 6, 14:00 - 15:50, Kallirhoe Hall

Tracking Concept Change with Incremental Boosting by Minimization of the Evolving Exponential Loss

Mihajlo Grbovic, Slobodan Vucetic

Methods involving ensembles of classifiers, such as bagging and boosting, are popular due to the strong theoretical guarantees for their performance and their superior results. Ensemble methods are typically designed by assuming the training data set is static and completely available at training time. As such, they are not suitable for online and incremental learning. In this paper we propose *IBoost*, an extension of *AdaBoost* for incremental learning via optimization of an exponential cost function which changes over time as the training data changes. The resulting algorithm is flexible and allows a user to customize it based on the computational constraints of the particular application. The new algorithm was evaluated on stream learning in presence of concept change. Experimental results showed that *IBoost* achieves better performance than the original *AdaBoost* trained from scratch each time the data set changes, and that it also outperforms previously proposed *Online Coordinate Boost*, *Online Boost* and its non-stationary modifications, *Fast and Light Boosting*, *ADWIN Online Bagging* and *DWM algorithms*.

Aggregating Independent and Dependent Models to Learn Multi-label Classifiers

Elena Montañés, José Ramón Quevedo, Juan José del Coz

The aim of multi-label classification is to automatically obtain models able to tag objects with the labels that better describe them. Despite it could seem like any other classification task, it is widely known that exploiting the presence of certain correlations between labels helps to improve the classification performance. In other words, object descriptions are usually not enough to induce good models, also label information must be taken into account. This paper presents an aggregated approach that combines two groups of classifiers, one assuming independence between labels, and the other considering fully conditional dependence among them. The framework proposed here can be applied not only for multi-label classification, but also in multi-label ranking tasks. Experiments carried out over several datasets endorse the superiority of our approach with regard to other methods in terms of some evaluation measures, keeping competitiveness in terms of others.

Multi-Label Ensemble Learning

Chuan Shi, Xiangnan Kong, Philip S. Yu, Bai Wang

Multi-label learning aims at predicting potentially multiple labels for a given instance. Conventional multi-label learning approaches focus on exploiting the label correlations to improve the accuracy of the learner by building an individual multi-label learner or a combined learner based upon a group of single-label learners. However, the generalization ability of such individual learner can be weak. It is well known that ensemble learning can effectively improve the generalization ability of learning systems by constructing multiple base learners and the performance of an ensemble is related to the both accuracy and diversity of base learners. In this paper, we study the problem of multi-label ensemble learning. Specifically, we aim at improving the generalization ability of multi-label learning systems by constructing a *group of multi-label base learners* which are both *accurate* and *diverse*. We propose a novel solution, called EnML, to effectively augment the accuracy as well as the diversity of multi-label base learners. In detail, we design two objective functions to evaluate the accuracy and diversity of multi-label base learners, respectively, and EnML simultaneously optimizes these two objectives with an evolutionary multi-objective optimization method. Experiments on real-world multi-label learning tasks validate the effectiveness of our approach against other well-established methods.

On oblique random forests

Bjoern H. Menze, B. Michael Kelm, Daniel N. Splitthoff, Ullrich Koethe, Fred A. Hamprecht

In his original paper on random forests, Breiman proposed two different decision tree ensembles: one generated from “orthogonal” trees with thresholds on individual features in every split, and one from “oblique” trees separating the feature space by randomly oriented hyperplanes. In spite of a rising interest in the random forest framework, however, ensembles built from orthogonal trees (RF) have gained most, if not all, attention so far. In the present work we propose to employ “oblique” random forests (oRF) built from multivariate trees which explicitly learn optimal split directions at internal nodes using linear discriminative models, rather than using random coefficients as the original oRF. This oRF outperforms RF, as well as other classifiers, on nearly all data sets but those with discrete factorial features. Learned node models perform distinctively better than random splits. An oRF feature importance score shows to be preferable over standard RF feature importance scores such as Gini or permutation importance. The topology of the oRF decision space appears to be smoother and better adapted to the data, resulting in improved generalization performance. Overall, the oRF propose here may be preferred over standard RF on most learning tasks involving numerical and spectral data.

Novel Fusion Methods for Pattern Recognition

Muhammad Awais, Fei Yan, Krystian Mikołajczyk, Josef Kittler

Over the last few years, several approaches have been proposed for information fusion including different variants of classifier level fusion (ensemble methods), stacking and multiple kernel learning (MKL). MKL has become a preferred choice for information fusion in object recognition. However, in the case of highly discriminative and complementary feature channels, it does not significantly improve upon its trivial baseline which averages the kernels. Alternative ways are stacking and classifier level fusion

(CLF) which rely on a two phase approach. There is a significant amount of work on linear programming formulations of ensemble methods particularly in the case of binary classification.

In this paper we propose a multiclass extension of binary ν -LPBoost, which learns the contribution of each class in each feature channel. The existing approaches of classifier fusion promote sparse features combinations, due to regularization based on l_1 -norm, and lead to a selection of a subset of feature channels, which is not good in the case of informative channels. Therefore, we generalize existing classifier fusion formulations to arbitrary l_p -norm for binary and multiclass problems which results in more effective use of complementary information. We also extended stacking for both binary and multiclass datasets. We present an extensive evaluation of the fusion methods on four datasets involving kernels that are all informative and achieve state-of-the-art results on all of them.

Session 7: Clustering

Tuesday 6, 16:20 - 18:10, Olympia Hall

The Minimum Code Length for Clustering Using the Gray Code

Mahito Sugiyama, Akihiro Yamamoto

We propose new approaches to exploit compression algorithms for clustering numerical data. Our first contribution is to design a measure that can score the quality of a given clustering result under the light of a *fixed* encoding scheme. We call this measure the *Minimum Code Length* (MCL). Our second contribution is to propose a general strategy to translate any encoding method into a cluster algorithm, which we call COOL (Coding-Oriented clustering). COOL has a low computational cost since it scales linearly with the data set size. The clustering results of COOL is also shown to minimize MCL. To illustrate further this approach, we consider the *Gray Code* as the encoding scheme to present G-COOL. G-COOL can find clusters of arbitrary shapes and remove noise. Moreover, it is robust to change in the input parameters; it requires only two lower bounds for the number of clusters and the size of each cluster, whereas most algorithms for finding arbitrarily shaped clusters work well only if all parameters are tuned appropriately. G-COOL is theoretically shown to achieve internal cohesion and external isolation and is experimentally shown to work well for both synthetic and real data sets.

Fast approximate text document clustering using Compressive Sampling

Laurence A. F. Park

Document clustering involves repetitive scanning of a document set, therefore as the size of the set increases, the time required for the clustering task increases and may even become impossible due to computational constraints. Compressive sampling is a feature sampling technique that allows us to perfectly reconstruct a vector from a small number of samples, provided that the vector is sparse in some known domain. In this article, we apply the theory behind compressive sampling to the document clustering problem using k-means clustering. We provide a method of computing high accuracy clusters in a fraction of the time it would have taken by directly clustering the documents. This is performed by using the Discrete Fourier Transform and the Discrete Cosine Transform. We provide empirical results showing that compressive sampling provides a 14 times increase in speed with little reduction in accuracy on 7,095 documents, and we also provide a very accurate

clustering of a 231,219 document set, providing 20 times increase in speed when compared to performing k-means clustering on the document set. This shows that compressive clustering is a very useful tool that can be used to quickly compute approximate clusters.

Clustering Rankings in the Fourier Domain

Stéphan Cléménçon, Romaric Gaudel, Jérémie Jakubowicz

It is the purpose of this paper to introduce a novel approach to clustering rank data on a set of possibly large cardinality $n \in \mathbf{N}^*$, relying upon Fourier representation of functions defined on the symmetric group \mathcal{G}_n . In the present setup, covering a wide variety of practical situations, rank data are viewed as distributions on \mathcal{G}_n . Cluster analysis aims at segmenting data into homogeneous subgroups, hopefully very dissimilar in a certain sense. Whereas considering dissimilarity measures/distances between distributions on the non commutative group \mathcal{G}_n , in a coordinate manner by viewing it as embedded in the set $[0,1]^{n!}$ for instance, hardly yields interpretable results and leads to face obvious computational issues, evaluating the closeness of groups of permutations in the Fourier domain may be much easier in contrast. Indeed, in a wide variety of situations, a few well-chosen Fourier (matrix) coefficients may permit to approximate efficiently two distributions on \mathcal{G}_n as well as their degree of dissimilarity, while describing global properties in an interpretable fashion. Following in the footsteps of recent advances in automatic feature selection in the context of unsupervised learning, we propose to cast the task of clustering rankings in terms of optimization of a criterion that can be expressed in the Fourier domain in a simple manner. The effectiveness of the method proposed is illustrated by numerical experiments based on artificial and real data.

Is there a best quality metric for graph clusters?

Hélio Almeida, Dorgival Guedes, Wagner Meira Jr, Mohammed Zaki

Graph clustering, the process of discovering groups of similar vertices in a graph, is a very interesting area of study, with applications in many different scenarios. One of the most important aspects of graph clustering is the evaluation of cluster quality, which is important not only to measure the effectiveness of clustering algorithms, but also to give insights on the dynamics of relationships in a given network. Many quality evaluation metrics for graph clustering have been proposed in the literature, but there is no consensus on how do they compare to each other and how well they perform on different kinds of graphs. In this work we study five major graph clustering quality metrics in terms of their formal biases and their behavior when applied to clusters found by four implementations of classic graph clustering algorithms on five large, real world graphs. Our results show that those popular quality metrics have strong biases toward incorrectly awarding good scores to some kinds of clusters, especially seen in larger networks. They also indicate that currently used clustering algorithms and quality metrics do not behave as expected when cluster structures are different from the more traditional, clique-like ones.

α -Clusterable Sets

Gerasimos S. Antzoulatos, Michael N. Vrahatis

In spite of the increasing interest into clustering research within the last decades, a unified clustering theory that is independent of a particular algorithm, or underlying the data structure and even the objective function has not be formulated so far. In the paper at hand, we take the first steps towards a theoretical foundation of clustering, by proposing a new

notion of “*clusterability*” of data sets based on the density of the data within a specific region. Specifically, we give a formal definition of what we call “ *α -clusterable*” set and we utilize this notion to prove that the principles proposed in Kleinberg’s impossibility theorem for clustering [25], are consistent. We further propose an unsupervised clustering algorithm which is based on the notion of α -clusterable set. The proposed algorithm exploits the ability of the well known and widely used particle swarm optimization [31] to maximize the recently proposed window density function [38]. The obtained clustering quality is compared favorably to the corresponding clustering quality of various other well-known clustering algorithms.

Session 8: Matrix and Tensor Analysis

Tuesday 6, 16:20 - 18:10, Attica Hall

Tensor Factorization Using Auxiliary Information

Atsuhiko Narita, Kohei Hayashi, Ryota Tomioka, Hisashi Kashima

Most of the existing analysis methods for tensors (or multiway arrays) only assume that tensors to be completed are of low rank. However, for example, when they are applied to tensor completion problems, their prediction accuracy tends to be significantly worse when only limited entries are observed. In this paper, we propose to use relationships among data as auxiliary information in addition to the low-rank assumption to improve the quality of tensor decomposition. We introduce two regularization approaches using graph Laplacians induced from the relationships, and design iterative algorithms for approximate solutions. Numerical experiments on tensor completion using synthetic and benchmark datasets show that the use of auxiliary information improves completion accuracy over the existing methods based only on the low-rank assumption, especially when observations are sparse.

Bayesian Matrix Co-Factorization: Variational Algorithm and Cramér-Rao Bound

Jiho Yoo, Seungjin Choi

Matrix factorization is a popular method for collaborative prediction, where unknown ratings are predicted by user and item factor matrices which are determined to approximate a user-item matrix as their product. Bayesian matrix factorization is preferred over other methods for collaborative filtering, since Bayesian approach alleviates overfitting, integrating out all model parameters using variational inference or sampling methods. However, Bayesian matrix factorization still suffers from the cold-start problem where predictions of ratings for new items or of new users’ preferences are required. In this paper we present *Bayesian matrix co-factorization* as an approach to exploiting side information such as content information and demographic user data, where multiple data matrices are jointly decomposed, i.e., each Bayesian decomposition is coupled by sharing some factor matrices. We derive variational inference algorithm for Bayesian matrix co-factorization. In addition, we compute Bayesian Cramér-Rao bound in the case of Gaussian likelihood, showing that Bayesian matrix co-factorization indeed improves the reconstruction over Bayesian factorization of single data matrix. Numerical experiments demonstrate the useful behavior of Bayesian matrix co-factorization in the case of cold-start problems.

Generalized Dictionary Learning for Symmetric Positive Definite Matrices with Application to Nearest Neighbor Retrieval

Suvrit Sra, Anoop Cherian

We introduce *Generalized Dictionary Learning* (GDL), a simple but practical framework for learning dictionaries over the manifold of positive definite matrices. We illustrate GDL by applying it to Nearest Neighbor (NN) retrieval, a task of fundamental importance in disciplines such as machine learning and computer vision. GDL distinguishes itself from traditional dictionary learning approaches by explicitly taking into account the manifold structure of the data. In particular, GDL allows performing “sparse coding” of positive definite matrices, which enables better NN retrieval. Experiments on several covariance matrix datasets show that GDL achieves performance rivaling state-of-the-art techniques.

Link prediction via matrix factorization

Aditya Krishna Menon, Charles Elkan

We propose to solve the link prediction problem in graphs using a supervised matrix factorization approach. The model learns latent features from the topological structure of a (possibly directed) graph, and is shown to make better predictions than popular unsupervised scores. We show how these latent features may be combined with optional explicit features for nodes or edges, which yields better performance than using either type of feature exclusively. Finally, we propose a novel approach to address the class imbalance problem which is common in link prediction by directly optimizing for a ranking loss. Our model is optimized with stochastic gradient descent and scales to large graphs. Results on several datasets show the efficacy of our approach.

Multi-Subspace Representation and Discovery

Dijun Luo, Feiping Nie, Chris Ding, Heng Huang

This paper presents the multi-subspace discovery problem and provides a theoretical solution which is guaranteed to recover the number of subspaces, the dimensions of each subspace, and the members of data points of each subspace simultaneously. We further propose a data representation model to handle noisy real world data. We develop a novel optimization approach to learn the presented model which is guaranteed to converge to global optimizers. As applications of our models, we first apply our solutions as preprocessing in a series of machine learning problems, including clustering, classification, and semi-supervised learning. We found that our method automatically obtains robust data presentation which preserves the affine subspace structures of high dimensional data and generate more accurate results in the learning tasks. We also establish a robust standalone classifier which directly utilizes our sparse and low rank representation model. Experimental results indicate our methods improve the quality of data by preprocessing and the standalone classifier outperforms some state-of-the-art learning approaches.

Session 9: Learning from Time Series Data

Tuesday 6, 16:20 - 18:10, Kallirhoe Hall

Motion segmentation by a model-based clustering approach of incomplete trajectories

Vasileios Karavasilis, Konstantinos Blekas, Christophoros Nikou

In this paper, we present a framework for visual object tracking based on clustering

trajectories of image key points extracted from a video. The main contribution of our method is that the trajectories are automatically extracted from the video sequence and they are provided directly to a model-based clustering approach. In most other methodologies, the latter constitutes a difficult part since the resulting feature trajectories have a short duration, as the key points disappear and reappear due to occlusion, illumination, viewpoint changes and noise. We present here a sparse, translation invariant regression mixture model for clustering trajectories of variable length. The overall scheme is converted into a Maximum A Posteriori approach, where the Expectation-Maximization (EM) algorithm is used for estimating the model parameters. The proposed method detects the different objects in the input image sequence by assigning each trajectory to a cluster, and simultaneously provides the motion of all objects. Numerical results demonstrate the ability of the proposed method to offer more accurate and robust solution in comparison with the mean shift tracker, especially in cases of occlusions.

Unsupervised Modeling of Partially Observable Environments

Vincent Graziano, Jan Koutnik, Jürgen Schmidhuber

We present an architecture based on self-organizing maps for learning a sensory layer in a learning system. The architecture, temporal network for transitions (TNT), enjoys the freedoms of unsupervised learning, works on-line, in non-episodic environments, is computationally light, and scales well. TNT generates a predictive model of its internal representation of the world, making planning methods available for both the exploitation and exploration of the environment. Experiments demonstrate that TNT learns nice representations of classical reinforcement learning mazes of varying size (up to 20 x 20) under conditions of high-noise and stochastic actions.

Artemis: Assessing the Similarity of Event-interval Sequences

Orestis Kostakis, Panagiotis Papapetrou, Jaakko Hollmén

In several application domains, such as sign language, medicine, and sensor networks, events are not necessarily instantaneous but they can have a time duration. Sequences of interval-based events may contain useful domain knowledge; thus, searching, indexing, and mining such sequences is crucial. We introduce two distance measures for comparing sequences of interval-based events which can be used for several data mining tasks such as classification and clustering. The first measure maps each sequence of interval-based events to a set of vectors that hold information about all concurrent events. These sets are then compared using an existing dynamic programming method. The second method, called Artemis, finds correspondence between intervals by mapping the two sequences into a bipartite graph. Similarity is inferred by employing the Hungarian algorithm. In addition, we present a linear-time lower-bound for Artemis. The performance of both measures is tested on data from three domains: sign language, medicine, and sensor networks. Experiments show the superiority of Artemis in terms of robustness to high levels of artificially introduced noise.

Discovering Temporal Bisociations for Linking Concepts over Time

Corrado Loglisci, Michelangelo Ceci

Bisociations represent interesting relationships between seemingly unconnected concepts from two or more contexts. Most of the existing approaches that permit the discovery of bisociations from data rely on the assumption that contexts are static or considered as

unchangeable domains. Actually, several real-world domains are intrinsically dynamic and can change over time. The same domain can change and can become completely different from what/how it was before: a dynamic domain observed at different time-points can present different representations and can be reasonably assimilated to a series of distinct static domains. In this work, we investigate the task of linking concepts from a dynamic domain through the discovery of bisociations which link concepts over time. This provides us with a means to unearth linkages which have not been discovered when observing the domain as static, but which may have developed over time, when considering the dynamic nature. We propose a computational solution which, assuming a time interval-based discretization of the domain, explores the spaces of association rules mined in the intervals and chains the rules on the basis of the concept generalization and information theory criteria. The application to the literature-based discovery shows how the method can re-discover known connections in biomedical terminology. Experiments and comparisons using alternative techniques highlight the additional peculiarities of this work.

ShiftTree: an Interpretable Model-Based Approach for Time Series Classification

Balázs Hidasi, Csaba Gáspár-Papanek

Efficient algorithms of time series data mining have the common denominator of utilizing the special time structure of the attributes of time series. To accommodate the information of time dimension into the process, we propose a novel instance-level cursor based indexing technique, which is combined with a decision tree algorithm. This is beneficial for several reasons: (a) it is insensitive to the time level noise (for example rendering, time shifting), (b) its working method can be interpreted, making the explanation of the classification process more understandable, and (c) it can manage time series of different length. The implemented algorithm named ShiftTree is compared to the well-known instance-based time series classifier 1-NN using different distance metrics, used over all 20 datasets of a public benchmark time series database and two more public time series datasets. On these benchmark datasets, our experiments show that the new model-based algorithm has an average accuracy slightly better than the most efficient instance-based methods, and there are multiple datasets where our model-based classifier exceeds the accuracy of instance-based methods. We also evaluated our algorithm via blind testing on the 20 datasets of the SIGKDD 2007 Time Series Classification Challenge. To improve the model accuracy and to avoid model overfitting, we provide forest methods as well.

Session 10: Learning from Social and Information Networks I

Wednesday 7, 10:40 - 12:30, Olympia Hall

Peer and Authority Pressure in Information-Propagation Models

Aris Anagnostopoulos, George Brova, Evimaria Terzi

Existing models of information diffusion assume that *peer influence* is the main reason for the observed propagation patterns. In this paper, we examine the role of *authority pressure* on the observed information cascades. We model this intuition by characterizing some nodes in the network as “authority” nodes. These are nodes that can influence large number of peers, while themselves cannot be influenced by peers. We propose a model that associates with every item two parameters that quantify the impact of the peer and the authority pressure on the item’s propagation. Given a network and the observed diffusion

patterns of the item, we learn these parameters from the data and characterize the item as peer- or authority-propagated. We also develop a randomization test that evaluates the statistical significance of our findings and makes our item characterization robust to noise. Our experiments with real data from online media and scientific-collaboration networks indicate that there is a strong signal of authority pressure in these networks.

Active Learning of Model Parameters for Influence Maximization

Tianyu Cao, Xindong Wu, Tony Xiaohua Hu, Song Wang

Previous research efforts on the influence maximization problem assume that the network model parameters are known beforehand. However, this is rarely true in real world networks. This paper deals with the situation when the network information diffusion parameters are unknown. To this end, we firstly examine the parameter sensitivity of a popular diffusion model in influence maximization, i.e., the *linear threshold model*, to motivate the necessity of learning the unknown model parameters. Experiments show that the influence maximization problem is sensitive to the model parameters under the linear threshold model. In the sequel, we formally define the problem of finding the model parameters for influence maximization as an active learning problem under the linear threshold model. We then propose a weighted sampling algorithm to solve this active learning problem. Extensive experimental evaluations on five popular network datasets demonstrate that the proposed weighted sampling algorithm outperforms pure random sampling in terms of both model accuracy and the proposed objective function.

A Shapley value Approach for Influence Attribution

Panagiotis Papapetrou, Aris Gionis, Heikki Mannila

Finding who and what is “important” is an ever-occurring question. Many methods that aim at characterizing important items or influential individuals have been developed in areas such as, bibliometrics, social-network analysis, link analysis, and web search. In this paper we study the problem of attributing influence scores to individuals who accomplish tasks in a collaborative manner. We assume that individuals build small teams, in different and diverse ways, in order to accomplish atomic tasks. For each task we are given an assessment of success or importance score, and the goal is to attribute those team-wise scores to the individuals. The challenge we face is that individuals in strong coalitions are favored against individuals in weaker coalitions, so the objective is to find fair attributions that account for such biasing. We propose an iterative algorithm for solving this problem that is based on the concept of Shapley value. The proposed method is applicable to a variety of scenarios, for example, attributing influence scores to scientists who collaborate in published articles, or employees of a company who participate in projects. Our method is evaluated on two real datasets: ISI Web of Science publication data and the Internet Movie Database.

Influence and Passivity in Social Media

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, Bernardo A. Huberman

The ever-increasing amount of information owing through Social Media forces the members of these networks to compete for attention and influence by relying on other people to spread their message. A large study of information propagation within Twitter reveals that the majority of users act as passive information consumers and do not forward the content to the network. Therefore, in order for individuals to become influential they

must not only obtain attention and thus be popular, but also overcome user passivity. We propose an algorithm that determines the influence and passivity of users based on their information forwarding activity. An evaluation performed with a 2.5 million user dataset shows that our influence measure is a good predictor of URL clicks, outperforming several other measures that do not explicitly take user passivity into account. We demonstrate that high popularity does not necessarily imply high influence and vice-versa.

Learning Recommendations in Social Media Systems By Weighting Multiple Relations

Boris Chidlovskii

We address the problem of item recommendation in social media sharing systems. We adopt a multi-relational framework capable to integrate different entity types available in the social media system and relations between the entities. We then model different recommendation tasks as weighted random walks in the relational graph. The main contribution of the paper is a novel method for learning the optimal contribution of each relation to a given recommendation task, by minimizing a loss function on the training dataset. We report results of the relation weight learning for two common tasks on the Flickr dataset, tag recommendation for images and contact recommendation for users.

Session 11: Spectral Clustering & Graph Mining

Wednesday 7, 10:40 - 12:30, Attica Hall

Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms

Danai Koutra, Tai-You Ke, U Kang, Duen Horng Polo Chau, Hsing-Kuo Kenneth Pao, Christos Faloutsos

If several friends of Smith have committed petty thefts, what would you say about Smith? Most people would not be surprised if Smith is a hardened criminal. **Guilt-by-association** methods combine weak signals to derive stronger ones, and have been extensively used for anomaly detection and classification in numerous settings (e.g., accounting fraud, cybersecurity, calling-card fraud).

The focus of this paper is to compare and contrast several very successful, guilt-by-association methods: *Random Walk with Restarts*, Semi-Supervised Learning, and *Belief Propagation* (BP).

Our main contributions are two-fold: (a) theoretically, we prove that all the methods result in a similar matrix inversion problem; (b) for practical applications, we developed FaBP, a fast algorithm that yields 2x speedup, equal or higher accuracy than BP, and is guaranteed to converge. We demonstrate these benefits using synthetic and real datasets, including YahooWeb, one of the largest graphs ever studied with BP.

Privacy Preserving Semi-Supervised Learning for Labeled Graphs

Hiroki Arai, Jun Sakuma

We propose a novel privacy preserving learning algorithm that achieves semi-supervised learning in graphs. In real world networks, such as disease infection over individuals, links (contact) and labels (infection) are often highly sensitive information. Although traditional semi-supervised learning methods play an important role in network data analysis, they fail to protect such sensitive information. Our solutions enable to predict labels of partially

labeled graphs without disclosure of labels and links, by incorporating cryptographic techniques into the label propagation algorithm. Even when labels included in the graph are kept private, the accuracy of our PPLP is equivalent to that of label propagation which is allowed to observe all labels in the graph. Empirical analysis showed that our solution is scalable compared with existing privacy preserving methods. The results with human contact networks showed that our protocol takes only about 10 seconds for computation and no sensitive information is disclosed through the protocol execution.

Eigenvector Sensitive Feature Selection For Spectral Clustering

Yi Jiang, Jiangtao Ren

Spectral clustering is one of the most popular methods for data clustering, and its performance is determined by the quality of the eigenvectors of the related graph Laplacian. Generally, graph Laplacian is constructed using the full features, which will degrade the quality of the related eigenvectors when there are a large number of noisy or irrelevant features in datasets. To solve this problem, we propose a novel unsupervised feature selection method inspired by perturbation analysis theory, which discusses the relationship between the perturbation of the eigenvectors of a matrix and its elements' perturbation. We evaluate the importance of each feature based on the average L1 norm of the perturbation of the first k eigenvectors of graph Laplacian corresponding to the k smallest positive eigenvalues, with respect to the feature's perturbation. Extensive experiments on several high-dimensional multi-class datasets demonstrate the good performance of our method compared with some state-of-the-art unsupervised feature selection methods.

DB-CSC: A density-based approach for subspace clustering in graphs with feature vectors

Stephan Günnemann, Brigitte Boden, Thomas Seidl

Data sources representing attribute information in combination with network information are widely available in today's applications. To realize the full potential for knowledge extraction, mining techniques like clustering should consider both information types simultaneously. Recent clustering approaches combine *subspace clustering* with *dense subgraph mining* to identify groups of objects that are similar in subsets of their attributes as well as densely connected within the network. While those approaches successfully circumvent the problem of full-space clustering, their limited cluster definitions are restricted to clusters of certain shapes.

In this work, we introduce a density-based cluster definition taking the attribute similarity in subspaces and the graph density into account. This novel cluster model enables us to detect clusters of arbitrary shape and size. We avoid redundancy in the result by selecting only the most interesting non-redundant clusters. Based on this model, we introduce the clustering algorithm DB-CSC. In thorough experiments we demonstrate the strength of DB-CSC in comparison to related approaches.

Parallel Structural Graph Clustering

Madeleine Seeland, Simon A. Berger, Alexandros Stamatakis, Stefan Kramer

We address the problem of clustering large graph databases according to scaffolds (i.e., large structural overlaps) that are shared between cluster members. In previous work, an online algorithm was proposed for this task that produces overlapping (non-disjoint) and non-exhaustive clusterings. In this paper, we parallelize this algorithm to take advantage

of high-performance parallel hardware and further improve the algorithm in three ways: a refined cluster membership test based on a set abstraction of graphs, sorting graphs according to size, to avoid cluster membership tests in the first place, and the definition of a cluster representative once the cluster scaffold is unique, to avoid cluster comparisons with all cluster members. In experiments on a large database of chemical structures, we show that running times can be reduced by a large factor for one parameter setting used in previous work. For harder parameter settings, it was possible to obtain results within reasonable time for 300,000 structures, compared to 10,000 structures in previous work. This shows that structural, scaffold-based clustering of smaller libraries for virtual screening is already feasible.

Session 12: Data Mining Theory & Foundations

Wednesday 7, 10:40 - 12:30, Kallirhoe Hall

The VC-Dimension of SQL Queries and Selectivity Estimation Through Sampling

Matteo Riondato, Mert Akdere, Ugur Çetintemel, Stanley B. Zdonik, Eli Upfal

We develop a novel method, based on the statistical concept of VC-dimension, for evaluating the selectivity (output cardinality) of SQL queries – a crucial step in optimizing the execution of large scale database and data-mining operations. The major theoretical contribution of this work, which is of independent interest, is an explicit bound on the VC-dimension of a range space defined by all possible outcomes of a collection (class) of queries. We prove that the VC-dimension is a function of the maximum number of Boolean operations in the selection predicate, and of the maximum number of select and join operations in any individual query in the collection, but it is neither a function of the number of queries in the collection nor of the size of the database. We develop a method based on this result: given a class of queries, it constructs a concise random sample of a database, such that with high probability the execution of any query in the class on the sample provides an accurate estimate for the selectivity of the query on the original large database. The error probability holds *simultaneously* for the selectivity estimates of all queries in the collection, thus the same sample can be used to evaluate the selectivity of multiple queries, and the sample needs to be refreshed only following major changes in the database. The sample representation computed by our method is typically sufficiently small to be stored in main memory. We present extensive experimental results, validating our theoretical analysis and demonstrating the advantage of our technique when compared to complex selectivity estimation techniques used in PostgreSQL and the Microsoft SQL Server.

Smooth Receiver Operating Characteristics (smROC) Curves

William Klement, Peter Flach, Nathalie Japkowicz, Stan Matwin

Supervised learning algorithms perform common tasks including classification, ranking, scoring, and probability estimation. We investigate how scoring information, often produced by these models, is utilized by an evaluation measure. The ROC curve represents a visualization of the ranking performance of classifiers. However, they ignore the scores which can be quite informative. While this ignored information is less precise than that given by probabilities, it is much more detailed than that conveyed by ranking. This paper presents a novel method to weight the ROC curve by these scores. We call it the Smooth

ROC (*smROC*) curve, and we demonstrate how it can be used to visualize the performance of learning models. We report experimental results to show that the *smROC* is appropriate for measuring performance similarities and differences between learning models, and is more sensitive to performance characteristics than the standard ROC curve.

Active Supervised Domain Adaptation

Avishek Saha, Piyush Rai, Hal Daumé III, Suresh Venkatasubramanian, Scott DuVall

In this paper, we harness the synergy between two important learning paradigms, namely, active learning and domain adaptation. We show how active learning in a target domain can leverage information from a different but related source domain. Our proposed framework, Active Learning Domain Adapted (ALDA), uses source domain knowledge to transfer information that facilitates active learning in the target domain. We propose two variants of Alda: a batch B-Alda and an online O-Alda. Empirical comparisons with numerous baselines on real-world datasets establish the efficacy of the proposed methods.

Comparing Apples and Oranges - Measuring Differences between Data Mining Results

Nikolaj Tatti, Jilles Vreeken

Deciding whether the results of two different mining algorithms provide significantly different information is an important open problem in exploratory data mining. Whether the goal is to select the most informative result for analysis, or decide which mining approach will likely provide the most novel insight, it is essential that we can tell how different the information is that two results provide.

In this paper we take a first step towards comparing exploratory results on binary data. We propose to meaningfully convert results into sets of noisy tiles, and compare between these sets by Maximum Entropy modeling and Kullback-Leibler divergence. The measure we construct this way is exible, and allows us to naturally include background knowledge, such that differences in results can be measured from the perspective of what a user already knows. Furthermore, adding to its interpretability, it coincides with Jaccard dissimilarity when we only consider exact tiles. Our approach provides a means to study and tell differences between results of different data mining methods. As an application, we show that it can also be used to identify which parts of results best redescribe other results. Experimental evaluation shows our measure gives meaningful results, correctly identifies methods that are similar in nature, and automatically provides sound redescrptions of results.

Learning Good Edit Similarities with Generalization Guarantees

Aurélien Bellet, Amaury Habrard, Marc Sebban

Similarity and distance functions are essential to many learning algorithms, thus training them has attracted a lot of interest. When it comes to dealing with structured data (e.g., strings or trees), *edit similarities* are widely used, and there exists a few methods for learning them. However, these methods offer no theoretical guarantee as to the generalization performance and discriminative power of the resulting similarities. Recently, a theory of learning with (ϵ, γ, τ) -good similarity functions was proposed. This new theory bridges the gap between the properties of a similarity function and its performance in classification. In this paper, we propose a novel edit similarity learning approach (*GESL*) driven by the idea of (ϵ, γ, τ) -goodness, which allows us to derive generalization guarantees using the notion of

uniform stability. We experimentally show that edit similarities learned with our method induce classification models that are both more accurate and sparser than those induced by the edit distance or edit similarities learned with a state-of-the-art method.

Session 13: Learning from Social and Information Networks II

Wednesday 7, 14:00 - 15:50, Olympia Hall

Toward a Fair Review-Management System

Theodoros Lappas, Evimaria Terzi

Item reviews are a valuable source of information for potential buyers, who are looking for information on a product's attributes before making a purchase decision. This search of information is often hindered by overwhelming numbers of available reviews, as well as low-quality and noisy content. While a significant amount of research has been devoted to filtering and organizing review corpora toward the benefit of the buyers, a crucial part of the reviewing process has been overlooked: *reviewer satisfaction*. As in every content-based system, the content-generators, in this case the reviewers, serve as the driving force. Therefore, keeping the reviewers satisfied and motivated to continue submitting high-quality content is essential. In this paper, we propose a system that helps potential buyers by focusing on high-quality and informative reviews, while keeping reviewers content and motivated.

Learning to Infer Social Ties in Large Networks

Wenbin Tang, Honglei Zhuang, Jie Tang

In online social networks, most relationships are lack of meaning labels (e.g., “colleague” and “intimate friends”), simply because users do not take the time to label them. An interesting question is: can we automatically infer the type of social relationships in a large network? what are the fundamental factors that imply the type of social relationships? In this work, we formalize the problem of social relationship learning into a semi-supervised framework, and propose a Partially-labeled Pairwise Factor Graph Model (PLP-FGM) for learning to infer the type of social ties. We tested the model on three different genres of data sets: Publication, Email and Mobile. Experimental results demonstrate that the proposed PLP-FGM model can accurately infer 92.7% of advisor-advisee relationships from the coauthor network (Publication), 88.0% of manager-subordinate relationships from the email network (Email), and 83.1% of the friendships from the mobile network (Mobile). Finally, we develop a distributed learning algorithm to scale up the model to real large networks.

A Community-Based Pseudolikelihood Approach for Relationship Labeling in Social Networks

Huaiyu Wan, Youfang Lin, Zhihao Wu, Houkuan Huang

A social network consists of people (or other social entities) connected by a set of social relationships. Awareness of the relationship types is very helpful for us to understand the structure and the characteristics of the social network. Traditional classifiers are not accurate enough for relationship labeling since they assume that all the labels are independent and identically distributed. A relational probabilistic model, relational Markov networks (RMNs), is introduced to labeling relationships, but the inefficient

parameter estimation makes it difficult to deploy in large-scale social networks. In this paper, we propose a community-based pseudolikelihood (CBPL) approach for relationship labeling. The community structure of a social network is used to assist in constructing the conditional random field, and this makes our approach reasonable and accurate. In addition, the computational simplicity of pseudolikelihood effectively resolves the time complexity problem which RMNs are suffering. We apply our approach on two real-world social networks, one is a terrorist relation network and the other is a phone call network we collected from encrypted call detail records. In our experiments, for avoiding losing links while splitting a closely connected social network into separate training and test subsets, we split the datasets according to the links rather than the individuals. The experimental results show that our approach performs well in terms of accuracy and efficiency.

Graph Evolution via Social Diffusion Processes

Dijun Luo, Chris Ding, Heng Huang

We present a new stochastic process, called as Social Diffusion Process (SDP), to address the graph modeling. Based on this model, we derive a graph evolution algorithm and a series of graph-based approaches to solve machine learning problems, including clustering and semi-supervised learning. SDP can be viewed as a special case of *Matthew effect*, which is a general phenomenon in nature and societies. We use social event as a metaphor of the intrinsic stochastic process for broad range of data. We evaluate our approaches in a large number of frequently used datasets and compare our approaches to other state-of-the-art techniques. Results show that our algorithm outperforms the existing methods in most cases. We also applying our algorithm into the functionality analysis of microRNA and discover biologically interesting cliques. Due to the broad availability of graph-based data, our new model and algorithm potentially have applications in wide range.

Mining Research Topic-related Influence between Academia and Industry

Dan He

Recently the problem of mining social influence has attracted lots of attention. Given a social network, researchers are interested in problems such as how influence, ideas, information propagate in the network. Similar problems have been proposed on co-authorship networks where the goal is to differentiate the social influences on research topic level and quantify the strength of the influence. In this work, we are interested in the problem of mining topic-specific influence between academia and industry. More specifically, given a co-authorship network, we want to identify which academia researcher is most influential to a given company on specific research topics. Given pairwise influences between researchers, we propose three models (simple additive model, weighted additive model and clustering-based additive model) to evaluate how influential a researcher is to a company. Finally, we illustrate the effectiveness of these three models on real large data set as well as on simulated data set.

Session 14: Relational learning and Inductive Logic Programming

Wednesday 7, 14:00 - 15:50, Attica Hall

Correcting Bias in Statistical Tests for Network Classifier Evaluation

Tao Wang, Jennifer Neville, Brian Gallagher, Tina Eliassi-Rad

It is difficult to directly apply conventional significance tests to compare the performance of network classification models because network data instances are not independent and identically distributed. Recent work [6] has shown that paired t-tests applied to overlapping network samples will result in unacceptably high levels (e.g., up to 50%) of Type I error (i.e., the tests lead to incorrect conclusions that models are different, when they are not). Thus, we need new strategies to accurately evaluate network classifiers. In this paper, we analyze the sources of bias (e.g. dependencies among network data instances) theoretically and propose analytical corrections to standard significance tests to reduce the Type I error rate to more acceptable levels, while maintaining reasonable levels of statistical power to detect true performance differences. We validate the effectiveness of the proposed corrections empirically on both synthetic and real networks.

Abductive Plan Recognition by Extending Bayesian Logic Programs

Sindhu Raghavan, Raymond J. Mooney

Plan recognition is the task of predicting an agent's top-level plans based on its observed actions. It is an abductive reasoning task that involves inferring cause from effect. Most existing approaches to plan recognition use either first-order logic or probabilistic graphical models. While the former cannot handle uncertainty, the latter cannot handle structured representations. In order to overcome these limitations, we develop an approach to plan recognition using Bayesian Logic Programs (BLPs), which combine first-order logic and Bayesian networks. Since BLPs employ logical deduction to construct the networks, they cannot be used effectively for plan recognition. Therefore, we extend BLPs to use logical abduction to construct Bayesian networks and call the resulting model Bayesian Abductive Logic Programs (BALPs). We learn the parameters in BALPs using the Expectation Maximization algorithm adapted for BLPs. Finally, we present an experimental evaluation of BALPs on three benchmark data sets and compare its performance with the state-of-the-art for plan recognition.

Learning the Parameters of Probabilistic Logic Programs from Interpretations

Bernd Gutmann, Ingo Thon, Luc De Raedt

ProbLog is a recently introduced probabilistic extension of the logic programming language Prolog, in which facts can be annotated with the probability that they hold. The advantage of this probabilistic language is that it naturally expresses a generative process over interpretations using a declarative model. Interpretations are relational descriptions or possible worlds. This paper introduces a novel parameter estimation algorithm LFI-ProbLog for learning ProbLog programs from partial interpretations. The algorithm is essentially a Soft-EM algorithm. It constructs a propositional logic formula for each interpretation that is used to estimate the marginals of the probabilistic parameters. The LFI-ProbLog algorithm has been experimentally evaluated on a number of data sets that justifies the approach and shows its effectiveness.

Gaussian Logic for Predictive Classification

Ondřej Kuželka, Andrea Szabóová, Matěj Holec, Filip Železný

We describe a statistical relational learning framework called Gaussian Logic capable to work efficiently with combinations of relational and numerical data. The framework assumes that, for a fixed relational structure, the numerical data can be modelled by a multivariate normal distribution. We demonstrate how the Gaussian Logic framework

can be applied to predictive classification problems. In experiments, we first show an application of the framework for the prediction of DNA-binding propensity of proteins. Next, we show how the Gaussian Logic framework can be used to find motifs describing highly correlated gene groups in gene-expression data which are then used in a set-level-based classification method.

Learning First-Order Definite Theories via Object-Based Queries

Joseph Selman, Alan Fern

We study the problem of exact learning of first-order definite theories via queries, toward the goal of allowing humans to more efficiently teach first-order concepts to computers. Prior work has shown that first order Horn theories can be learned using a polynomial number of membership and equivalence queries [6]. However, these query types are sometimes unnatural for humans to answer and only capture a small fraction of the information that a human teacher might be able to easily communicate. In this work, we enrich the types of information that can be provided by a human teacher and study the associated learning problem from a theoretical perspective. First, we consider allowing queries that ask the teacher for the relevant objects in a training example. Second, we examine a new query type, called a pairing query, where the teacher provides mappings between objects in two different examples. We present algorithms that leverage these new query types as well as restrictions applied to equivalence queries to significantly reduce or eliminate the required number of membership queries, while preserving polynomial learnability. In addition, we give learnability results for certain cases of imperfect teachers. These results show, in theory, the potential for incorporating object-based queries into first-order learning algorithms in order to reduce human teaching effort.

Session 15: Model Selection & Statistical Learning

Wednesday 7, 14:00 - 15:50, Kallirhoe Hall

A selecting-the-best method for budgeted model selection

Gianluca Bontempi, Olivier Caelen

The paper focuses on budgeted model selection, that is the selection between a set of alternative models when the ratio between the number of model assessments and the number of alternatives, though bigger than one, is low. We propose an approach based on the notion of probability of correct selection, a notion borrowed from the domain of Monte Carlo stochastic approximation. The idea is to estimate from data the probability that a greedy selection returns the best alternative and to define a sampling rule which maximises such quantity. Analytical results in the case of two alternatives are extended to a larger number of alternatives by using the Clark's approximation of the maximum of a set of random variables. Preliminary results on synthetic and real model selection tasks show that the technique is competitive with state-of-the-art algorithms, like the bandit UCB.

Aspects of Semi-Supervised and Active Learning in Conditional Random Fields

Nataliya Sokolovska

Conditional random fields are among the state-of-the-art approaches to structured output prediction, and the model has been adopted for various real-world problems. The supervised classification is expensive, since it is usually expensive to produce labelled

data. Unlabeled data are relatively cheap, but how to use it? Unlabeled data can be used to estimate marginal probability of observations, and we exploit this idea in our work. Introduction of unlabeled data and of probability of observations into a purely discriminative model is a challenging task.

We consider an extrapolation of a recently proposed semi-supervised criterion to the model of conditional random fields, and show its drawbacks. We discuss alternative usage of the marginal probability and propose a pool-based active learning approach based on quota sampling. We carry out experiments on synthetic as well as on standard natural language data sets, and we show that the proposed quota sampling active learning method is efficient.

Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process

Changyou Chen, Lan Du, Wray Buntine

Hierarchical modeling and reasoning are fundamental in machine intelligence, and for this the two-parameter Poisson-Dirichlet Process (PDP) plays an important role. The most popular MCMC sampling algorithm for the hierarchical PDP and hierarchical Dirichlet Process is to conduct an incremental sampling based on the Chinese restaurant metaphor, which originates from the Chinese restaurant process (CRP). In this paper, with the same metaphor, we propose a new table representation for the hierarchical PDPs by introducing an auxiliary latent variable, called table indicator, to record which customer takes responsibility for starting a new table. In this way, the new representation allows full exchangeability that is an essential condition for a correct Gibbs sampling algorithm. Based on this representation, we develop a block Gibbs sampling algorithm, which can jointly sample the data item and its table contribution. We test this out on the hierarchical Dirichlet process variant of latent Dirichlet allocation (HDP-LDA) developed by Teh, Jordan, Beal and Blei. Experiment results show that the proposed algorithm outperforms their “posterior sampling by direct assignment” algorithm in both out-of-sample perplexity and convergence speed. The representation can be used with many other hierarchical PDP models.

Comparing Probabilistic Models for Melodic Sequences

Athina Spiliopoulou, Amos Storkey

Modelling the real world complexity of music is a challenge for machine learning. We address the task of modeling melodic sequences from the same music genre. We perform a comparative analysis of two probabilistic models; a Dirichlet Variable Length Markov Model (Dirichlet-VMM) and a Time Convolutional Restricted Boltzmann Machine (TC-RBM). We show that the TC-RBM learns descriptive music features, such as underlying chords and typical melody transitions and dynamics. We assess the models for future prediction and compare their performance to a VMM, which is the current state of the art in melody generation. We show that both models perform significantly better than the VMM, with the Dirichlet-VMM marginally outperforming the TC-RBM. Finally, we evaluate the short order statistics of the models, using the Kullback-Leibler divergence between test sequences and model samples, and show that our proposed methods match the statistics of the music genre significantly better than the VMM.

Multimodal nonlinear filtering using Gauss-Hermite Quadrature

Hannes P. Saal, Nicolas Heess, Sethu Vijayakumar

In many filtering problems the exact posterior state distribution is not tractable and is therefore approximated using simpler parametric forms, such as single Gaussian distributions. In nonlinear filtering problems the posterior state distribution can, however, take complex shapes and even become multimodal so that single Gaussians are no longer sufficient. A standard solution to this problem is to use a bank of independent filters that individually represent the posterior with a single Gaussian and jointly form a mixture of Gaussians representation. Unfortunately, since the filters are optimized separately and interactions between the components consequently not taken into account, the resulting representation is typically poor. As an alternative we therefore propose to directly optimize the full approximating mixture distribution by minimizing the KL divergence to the true state posterior. For this purpose we describe a deterministic sampling approach that allows us to perform the intractable minimization approximately and at reasonable computational cost. We find that the proposed method models multimodal posterior distributions noticeably better than banks of independent filters even when the latter are allowed many more mixture components. We demonstrate the importance of accurately representing the posterior with a tractable number of components in an active learning scenario where we report faster convergence, both in terms of number of observations processed and in terms of computation time, and more reliable convergence on up to ten-dimensional problems.

Session 16: Graphical & Hidden Markov Models

Wednesday 7, 16:20 - 18:10, Olympia Hall

Fourier-Information Duality in the Identity Management Problem

Xiaoye Jiang, Jonathan Huang, Leonidas Guibas

We compare two recently proposed approaches for representing probability distributions over the space of permutations in the context of multi-target tracking. We show that these two representations, the Fourier approximation and the information form approximation can both be viewed as low dimensional projections of a true distribution, but with respect to different metrics. We identify the strengths and weaknesses of each approximation, and propose an algorithm for converting between the two forms, allowing for a *hybrid* approach that draws on the strengths of both representations. We show experimental evidence that there are situations where hybrid algorithms are favorable.

An Alternating Direction Method for Dual MAP LP Relaxation

Ofer Meshi, Amir Globerson

Maximum a-posteriori (MAP) estimation is an important task in many applications of probabilistic graphical models. Although finding an exact solution is generally intractable, approximations based on linear programming (LP) relaxation often provide good approximate solutions. In this paper we present an algorithm for solving the LP relaxation optimization problem. In order to overcome the lack of strict convexity, we apply an augmented Lagrangian method to the dual LP. The algorithm, based on the alternating direction method of multipliers (ADMM), is guaranteed to converge to the global optimum of the LP relaxation objective. Our experimental results show that this algorithm is competitive with other state-of-the-art algorithms for approximate MAP estimation.

Restricted Deep Belief Networks for Multi-View Learning

Yoonseop Kang, Seungjin Choi

Deep belief network (DBN) is a probabilistic generative model with multiple layers of hidden nodes and a layer of visible nodes, where parameterizations between layers obey harmonium or restricted Boltzmann machines (RBMs). In this paper we present restricted deep belief network (RDBN) for multi-view learning, where each layer of hidden nodes is composed of view-specific and shared hidden nodes, in order to learn individual and shared hidden spaces from multiple views of data. View-specific hidden nodes are connected to corresponding view-specific hidden nodes in the lower-layer or visible nodes involving a specific view, whereas shared hidden nodes follow inter-layer connections without restrictions as in standard DBNs. RDBN is trained using layer-wise contrastive divergence learning. Numerical experiments on synthetic and real-world datasets demonstrate the useful behavior of the RDBN, compared to the multi-wing harmonium (MWH) which is a two-layer undirected model.

A Spectral Learning Algorithm for Finite State Transducers

Borja Balle, Ariadna Quattoni, Xavier Carreras

Finite-State Transducers (FSTs) are a popular tool for modeling paired input-output sequences, and have numerous applications in real-world problems. Most training algorithms for learning FSTs rely on gradient-based or EM optimizations which can be computationally expensive and suffer from local optima issues. Recently, Hsu et al. [13] proposed a spectral method for learning Hidden Markov Models (HMMs) which is based on an Observable Operator Model (OOM) view of HMMs. Following this line of work we present a spectral algorithm to learn FSTs with strong PAC-style guarantees. To the best of our knowledge, ours is the first result of this type for FST learning. At its core, the algorithm is simple, and scalable to large data sets. We present experiments that validate the effectiveness of the algorithm on synthetic and real data.

Common Substructure Learning of Multiple Graphical Gaussian Models

Satoshi Hara, Takashi Washio

Learning underlying mechanisms of data generation is of great interest in the scientific and engineering fields amongst others. Finding dependency structures among variables in the data is one possible approach for the purpose, and is an important task in data mining. In this paper, we focus on learning dependency substructures shared by multiple datasets. In many scenarios, the nature of data varies due to a change in the surrounding conditions or non-stationary mechanisms over the multiple datasets. However, we can also assume that the change occurs only partially and some relations between variables remain unchanged. Moreover, we can expect that such commonness over the multiple datasets is closely related to the invariance of the underlying mechanism. For example, errors in engineering systems are usually caused by faults in the sub-systems with the other parts remaining healthy. In such situations, though anomalies are observed in sensor values, the underlying invariance of the healthy sub-systems is still captured by some steady dependency structures before and after the onset of the error. We propose a structure learning algorithm to find such invariances in the case of Graphical Gaussian Models (GGM). The proposed method is based on a block coordinate descent optimization, where subproblems can be solved efficiently by existing algorithms for *Lasso* and the *continuous quadratic knapsack problem*. We confirm the validity of our approach through numerical

simulations and also in applications with real world datasets extracted from the analysis of city-cycle fuel consumption and anomaly detection in car sensors.

Session 17: Supervised Learning I

Wednesday 7, 16:20 - 18:10, Attica Hall

Generalized Agreement Statistics over Fixed Group of Experts

Mohak Shah

Generalizations of chance corrected statistics to measure inter-expert agreement on class label assignments to the data instances have traditionally relied on the marginalization argument over a variable group of experts. Further, this argument has also resulted in agreement measures to evaluate the class predictions by an isolated classifier against the (multiple) labels assigned by the group of experts. We show that these measures are not necessarily suitable for application in the more typical fixed experts' group scenario. We also propose novel, more meaningful, less variable generalizations for quantifying both the inter-expert agreement over the fixed group and assessing a classifier's output against it in a multi-expert multi-class scenario by taking into account expert-specific biases and correlations.

Larger Residuals Less Work: Active Document Scheduling for Latent Dirichlet Allocation

Mirwaes Wahabzada, Kristian Kersting

Recently, there have been considerable advances in fast inference for latent Dirichlet allocation (LDA). In particular, stochastic optimization of the variational Bayes (VB) objective function with a natural gradient step was proved to converge and able to process massive document collections. To reduce noise in the gradient estimation, it considers multiple documents chosen uniformly at random. While it is widely recognized that the scheduling of documents in stochastic optimization may have significant consequences, this issue remains largely unexplored. In this work, we address this issue. Specifically, we propose residual LDA, a novel, easy-to-implement, LDA approach that schedules documents in an informed way. Intuitively, in each iteration, residual LDA actively selects documents that exert a disproportionately large influence on the current residual to compute the next update. On several real-world datasets, including 3M articles from Wikipedia, we demonstrate that residual LDA can handily analyze massive document collections and find topic models as good or better than those found with batch VB and randomly scheduled VB, and significantly faster.

Datum-Wise Classification: A Sequential Approach to Sparsity

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, Patrick Gallinari

We propose a novel classification technique whose aim is to select an appropriate representation for each datapoint, in contrast to the usual approach of selecting a representation encompassing the whole dataset. This *datum-wise* representation is found by using a sparsity inducing empirical risk, which is a relaxation of the standard L_0 regularized risk. The classification problem is modeled as a sequential decision process that sequentially chooses, for each datapoint, which features to use before classifying. Datum-Wise Classification extends naturally to multi-class tasks, and we describe

a specific case where our inference has equivalent complexity to a traditional linear classifier, while still using a variable number of features. We compare our classifier to classical L_1 regularized linear models (L_1 -SVM and LARS) on a set of common binary and multi-class datasets and show that for an equal average number of features used we can get improved performance using our method.

Transfer Learning With Adaptive Regularizers

Ulrich Rückert, Marius Kloft

The success of regularized risk minimization approaches to classification with linear models depends crucially on the selection of a regularization term that matches with the learning task at hand. If the necessary domain expertise is rare or hard to formalize, it may be difficult to find a good regularizer. On the other hand, if plenty of related or similar data is available, it is a natural approach to adjust the regularizer for the new learning problem based on the characteristics of the related data. In this paper, we study the problem of obtaining good parameter values for a l_2 -style regularizer with feature weights. We analytically investigate a moment-based method to obtain good values and give uniform convergence bounds for the prediction error on the target learning task. An empirical study shows that the approach can improve predictive accuracy considerably in the application domain of text classification.

Network Regression with Predictive Clustering Trees

Daniela Stojanova, Michelangelo Ceci, Annalisa Appice, Sašo Džeroski

Regression inference in network data is a challenging task in machine learning and data mining. Network data describe entities represented by nodes, which may be connected with (related to) each other by edges. Many network datasets are characterized by a form of autocorrelation where the values of the response variable at a given node depend on the values of the variables (predictor and response) at the nodes connected to the given node. This phenomenon is a direct violation of the assumption of independent (i.i.d.) observations: At the same time, it offers a unique opportunity to improve the performance of predictive models on network data, as inferences about one entity can be used to improve inferences about related entities. In this paper, we propose a data mining method that explicitly considers autocorrelation when building regression models from network data. The method is based on the concept of predictive clustering trees (PCTs), which can be used both for clustering and predictive tasks: PCTs are decision trees viewed as hierarchies of clusters and provide symbolic descriptions of the clusters. In addition, PCTs can be used for multi-objective prediction problems, including multi-target regression and multi-target classification. Empirical results on real world problems of network regression show that the proposed extension of PCTs performs better than traditional decision tree induction when autocorrelation is present in the data.

Learning from Inconsistent and Unreliable Annotators by a Gaussian Mixture Model and Bayesian Information Criterion

Ping Zhang, Zoran Obradovic

Supervised learning from multiple annotators is an increasingly important problem in machine learning and data mining. This paper develops a probabilistic approach to this problem when annotators are not only unreliable, but also have varying performance depending on the data. The proposed approach uses a Gaussian mixture model (GMM)

and Bayesian information criterion (BIC) to find the fittest model to approximate the distribution of the instances. Then the maximum a posterior (MAP) estimation of the hidden true labels and the maximum-likelihood (ML) estimation of quality of multiple annotators are provided alternately. Experiments on emotional speech classification and CASP9 protein disorder prediction tasks show performance improvement of the proposed approach as compared to the majority voting baseline and a previous data-independent approach. Moreover, the approach also provides more accurate estimates of individual annotators performance for each Gaussian component, thus paving the way for understanding the behaviors of each annotator.

Session 18: Unsupervised Learning & Dimensionality Reduction

Wednesday 7, 16:20 - 18:10, Kallirhoe Hall

Minimum Neighbor Distance Estimators of Intrinsic Dimension

Gabriele Lombardi, Alessandro Rozza, Claudio Ceruti, Elena Casiraghi, Paola Campadelli

Most of the machine learning techniques suffer the “curse of dimensionality” effect when applied to high dimensional data. To face this limitation, a common preprocessing step consists in employing a dimensionality reduction technique. In literature, a great deal of research work has been devoted to the development of algorithms performing this task. Often, these techniques require as parameter the number of dimensions to be retained; to this aim, they need to estimate the “intrinsic dimensionality” of the given dataset, which refers to the minimum number of degrees of freedom needed to capture all the information carried by the data. Although many estimation techniques have been proposed, most of them fail in case of noisy data or when the intrinsic dimensionality is too high. In this paper we present a family of estimators based on the probability density function of the normalized nearest neighbor distance. We evaluate the proposed techniques on both synthetic and real datasets comparing their performances with those obtained by state of the art algorithms; the achieved results prove that the proposed methods are promising.

Online Clustering of High-Dimensional Trajectories under Concept Drift

Georg Krempl, Zaigham Faraz Siddiqui, Myra Spiliopoulou

Historical transaction data are collected in many applications, e.g., patient histories recorded by physicians and customer transactions collected by companies. An important question is the learning of models upon *the primary objects* (patients, customers) rather than the transactions, especially when these models are subjected to drift.

We address this problem by combining advances of online clustering on multivariate data with the trajectory mining paradigm. We model the measurements of each individual primary object (e.g. its transactions), taken at irregular time intervals, as a trajectory in a high-dimensional feature space. Then, we cluster individuals with similar trajectories to identify sub-populations that evolve similarly, e.g. groups of customers that evolve similarly or groups of employees that have similar careers.

We assume that the multivariate trajectories are generated by drifting Gaussian Mixture Models. We study (i) an EM-based approach that clusters these trajectories incrementally as a reference method that has access to all the data for learning, and propose (ii) an online algorithm based on a Kalman filter that efficiently tracks the trajectories of Gaussian

clusters. We show that while both methods approximate the reference well, the algorithm based on a Kalman filter is faster by one order of magnitude compared to the EM-based approach.

Linear Discriminant Dimensionality Reduction

Quanquan Gu, Zhenhui Li, Jiawei Han

Fisher criterion has achieved great success in dimensionality reduction. Two representative methods based on Fisher criterion are *Fisher Score* and *Linear Discriminant Analysis* (LDA). The former is developed for feature selection while the latter is designed for subspace learning. In the past decade, these two approaches are often studied independently. In this paper, based on the observation that Fisher score and LDA are complementary, we propose to integrate Fisher score and LDA in a unified framework, namely *Linear Discriminant Dimensionality Reduction* (LDDR). We aim at finding a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. LDDR inherits the advantages of Fisher score and LDA and is able to do feature selection and subspace learning simultaneously. Both Fisher score and LDA can be seen as the special cases of the proposed method. The resultant optimization problem is a mixed integer programming, which is difficult to solve. It is relaxed into a $L_{2,1}$ -norm constrained least square problem and solved by accelerated proximal gradient descent algorithm. Experiments on benchmark face recognition data sets illustrate that the proposed method outperforms the state of the art methods arguably.

The Minimum Transfer Cost Principle for Model-Order Selection

Mario Frank, Morteza Haghir Chehreghani, Joachim Buhmann

The goal of model-order selection is to select a model variant that generalizes best from training data to unseen test data. In unsupervised learning without any labels, the computation of the generalization error of a solution poses a conceptual problem which we address in this paper. We formulate the principle of “*minimum transfer costs*” for model-order selection. This principle renders the concept of cross-validation applicable to unsupervised learning problems. As a substitute for labels, we introduce a mapping between objects of the training set to objects of the test set enabling the transfer of training solutions. Our method is explained and investigated by applying it to well-known problems such as singular-value decomposition, correlation clustering, Gaussian mixture-models, and k-means clustering. Our principle finds the optimal model complexity in controlled experiments and in real-world problems such as image denoising, role mining and detection of misconfigurations in access-control data.

Higher Order Contractive auto-encoder

Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, Xavier Glorot

We propose a novel regularizer when training an auto-encoder for unsupervised feature extraction. We explicitly encourage the latent representation to contract the input space by regularizing the norm of the Jacobian (analytically) and the Hessian (stochastically) of the encoder’s output with respect to its input, at the training points. While the penalty on the Jacobian’s norm ensures robustness to tiny corruption of samples in the input space, constraining the norm of the Hessian extends this robustness when moving further away from the sample. From a manifold learning perspective, balancing this regularization

with the auto-encoder's reconstruction objective yields a representation that varies most when moving along the data manifold in input space, and is most insensitive in directions orthogonal to the manifold. The second order regularization, using the Hessian, penalizes curvature, and thus favors smooth manifold. We show that our proposed technique, while remaining computationally efficient, yields representations that are significantly better suited for initializing deep architectures than previously proposed approaches, beating state-of-the-art performance on a number of datasets.

Session 19: Supervised Learning II

Thursday 8, 9:10 - 11:00, Olympia Hall

Regularized Sparse Kernel Slow Feature Analysis

Wendelin Böhmer, Steffen Grünewälder, Hannes Nickisch, Klaus Obermayer

This paper develops a kernelized *slow feature analysis* (SFA) algorithm. SFA is an unsupervised learning method to extract features which encode latent variables from time series. Generative relationships are usually complex, and current algorithms are either not powerful enough or tend to over-fit. We make use of the *kernel trick* in combination with *sparsification* to provide a powerful function class for large data sets. Sparsity is achieved by a novel *matching pursuit* approach that can be applied to other tasks as well. For small but complex data sets, however, the kernel SFA approach leads to over-fitting and numerical instabilities. To enforce a stable solution, we introduce *regularization* to the SFA objective. Versatility and performance of our method are demonstrated on audio and video data sets.

Kernels for Link Prediction with Latent Feature Models

Canh Hao Nguyen, Hiroshi Mamitsuka

Predicting new links in a network is a problem of interest in many application domains. Most of the prediction methods utilize information on the network's entities such as nodes to build a model of links. Network structures are usually not used except for the networks with similarity or relatedness semantics. In this work, we use network structures for link prediction with a more general network type with latent feature models. The problem is the difficulty to train these models directly for large data. We propose a method to solve this problem using kernels and cast the link prediction problem into a binary classification problem. The key idea is not to infer latent features explicitly, but to represent these features implicitly in the kernels, making the method scalable to large networks. In contrast to the other methods for latent feature models, our method inherits all the advantages of kernel framework: optimality, efficiency and nonlinearity. We apply our method to real data of protein-protein interactions to show the merits of our method.

PerTurbo: a new classification algorithm based on the spectrum perturbations of the Laplace-Beltrami operator

Nicolas Courty, Thomas Burger, Johann Laurent

PerTurbo, an original, non-parametric and efficient classification method is presented here. In our framework, the manifold of each class is characterized by its Laplace-Beltrami operator, which is evaluated with classical methods involving the graph Laplacian. The classification criterion is established thanks to a measure of the magnitude of the spectrum

perturbation of this operator. The first experiments show good performances against classical algorithms of the state-of-the-art. Moreover, from this measure is derived an efficient policy to design sampling queries in a context of active learning. Performances collected over toy examples and real world datasets assess the qualities of this strategy.

Fast Support Vector Machines for Structural Kernels

Aliaksei Severyn, Alessandro Moschitti

In this paper, we propose three important enhancements of the approximate cutting plane algorithm (CPA) to train Support Vector Machines with structural kernels: (i) we exploit a compact yet exact representation of cutting plane models using directed acyclic graphs to speed up both training and classification, (ii) we provide a parallel implementation, which makes the training scale almost linearly with the number of CPUs, and (iii) we propose an alternative sampling strategy to handle class-imbalanced problem and show that theoretical convergence bounds are preserved. The experimental evaluations on three diverse datasets demonstrate the soundness of our approach and the possibility to carry out fast learning and classification with structural kernels.

Building Sparse Support Vector Machines for Multi-Instance Classification

Zhouyu Fu, Guojun Lu, Kai Ming Ting, Dengsheng Zhang

We propose a direct approach to learning sparse Support Vector Machine (SVM) prediction models for Multi-Instance (MI) classification. The proposed sparse SVM is based on a “label-mean” formulation of MI classification which takes the average of predictions of individual instances for bag-level prediction. This leads to a convex optimization problem, which is essential for the tractability of the optimization problem arising from the sparse SVM formulation we derived subsequently, as well as the validity of the optimization strategy we employed to solve it. Based on the “label-mean” formulation, we can build sparse SVM models for MI classification and explicitly control their sparsities by enforcing the maximum number of expansions allowed in the prediction function. An effective optimization strategy is adopted to solve the formulated sparse learning problem which involves the learning of both the classifier and the expansion vectors. Experimental results on benchmark data sets have demonstrated that the proposed approach is effective in building very sparse SVM models while achieving comparable performance to the state-of-the-art MI classifiers.

Session 20: Semi-Supervised and Transductive Learning

Thursday 8, 9:10 - 11:00, Attica Hall

Learning from Label Proportions by Optimizing Cluster Model Selection

Marco Stolpe, Katharina Morik

In a supervised learning scenario, we learn a mapping from input to output values, based on labeled examples. Can we learn such a mapping also from groups of unlabeled observations, only knowing, for each group, the proportion of observations with a particular label? Solutions have real world applications. Here, we consider groups of steel sticks as samples in quality control. Since the steel sticks cannot be marked individually, for each group of sticks it is only known how many sticks of high (low) quality it contains. We want to predict the achieved quality for each stick before it reaches the final production

station and quality control, in order to save resources. We define the problem of learning from label proportions and present a solution based on clustering. Our method empirically shows a better prediction performance than recent approaches based on probabilistic SVMs, Kernel k-Means or conditional exponential models.

Adaptive Boosting for Transfer Learning using Dynamic Updates

Samir Al-Stouhi, Chandan K. Reddy

Instance-based transfer learning methods utilize labeled examples from one domain to improve learning performance in another domain via knowledge transfer. Boosting-based transfer learning algorithms are a subset of such methods and have been applied successfully within the transfer learning community. In this paper, we address some of the weaknesses of such algorithms and extend the most popular transfer boosting algorithm, TrAdaBoost. We incorporate a dynamic factor into TrAdaBoost to make it meet its intended design of incorporating the advantages of both AdaBoost and the “Weighted Majority Algorithm”. We theoretically and empirically analyze the effect of this important factor on the boosting performance of TrAdaBoost and we apply it as a “correction factor” that significantly improves the classification performance. Our experimental results on several real-world datasets demonstrate the effectiveness of our framework in obtaining better classification results.

Learning from Partially Annotated Sequences

Eraldo R. Fernandes, Ulf Brefeld

We study sequential prediction models in cases where only fragments of the sequences are annotated with the ground-truth. The task does not match the standard semi-supervised setting and is highly relevant in areas such as natural language processing, where completely labeled instances are expensive and require editorial data. We propose to generalize the semi-supervised setting and devise a simple transductive loss-augmented perceptron to learn from inexpensive partially annotated sequences that could for instance be provided by laymen, the wisdom of the crowd, or even automatically. Experiments on mono- and cross-lingual named entity recognition tasks with automatically generated partially annotated sentences from Wikipedia demonstrate the effectiveness of the proposed approach. Our results show that learning from partially labeled data is never worse than standard supervised and semi-supervised approaches trained on data with the same ratio of labeled and unlabeled tokens.

Constraint selection for semi-supervised topological clustering

Kais Allab, Khalid Benabdeslem

In this paper, we propose to adapt the batch version of self-organizing map (SOM) to background information in clustering task. It deals with constrained clustering with SOM in a deterministic paradigm. In this context we adapt the appropriate topological clustering to pair-wise instance level constraints with the study of their informativeness and coherence properties for measuring their utility for the semi-supervised learning process. These measures will provide guidance in selecting the most useful constraint sets for the proposed algorithm. Experiments will be given over several databases for validating our approach in comparison with another constrained clustering ones.

COSNet: a Cost Sensitive Neural Network for Semi-supervised Learning in Graphs

Alberto Bertoni, Marco Frasca, Giorgio Valentini

The semi-supervised problem of learning node labels in graphs consists, given a partial graph labeling, in inferring the unknown labels of the unlabeled vertices. Several machine learning algorithms have been proposed for solving this problem, including Hopfield networks and label propagation methods; however, some issues have been only partially considered, e.g. the preservation of the prior knowledge and the unbalance between positive and negative labels. To address these items, we propose a Hopfield-based cost sensitive neural network algorithm (*COSNet*). The method factorizes the solution of the problem in two parts: 1) the subnetwork composed by the labelled vertices is considered, and the network parameters are estimated through a supervised algorithm; 2) the estimated parameters are extended to the subnetwork composed of the unlabeled vertices, and the attractor reached by the dynamics of this subnetwork allows to predict the labeling of the unlabeled vertices. The proposed method embeds in the neural algorithm the “a priori” knowledge coded in the labelled part of the graph, and separates node labels and neuron states, allowing to differentially weight positive and negative node labels. Moreover, *COSNet* introduces an efficient cost-sensitive strategy which allows to learn the near-optimal parameters of the network in order to take into account the unbalance between positive and negative node labels. Finally, the dynamics of the network is restricted to its unlabeled part, preserving the minimization of the overall objective function and significantly reducing the time complexity of the learning algorithm. *COSNet* has been applied to the genome-wide prediction of gene function in a model organism. The results, compared with those obtained by other semi-supervised label propagation algorithms and supervised machine learning methods, show the effectiveness of the proposed approach.

Session 21: Preference Learning and Ranking

Thursday 8, 9:10 - 11:00, Kallirhoe Hall

Direct Policy Ranking with Robot Data Streams

Riad Akrou, Marc Schoenauer, Michèle Sebag

Many machine learning approaches in robotics, based on reinforcement learning, inverse optimal control or direct policy learning, critically rely on robot simulators. This paper investigates a simulator-free direct policy learning, called *Preference-based Policy Learning* (PPL). PPL iterates a four-step process: the robot demonstrates a candidate policy; the expert ranks this policy comparatively to other ones according to her preferences; these preferences are used to learn a policy return estimate; the robot uses the policy return estimate to build new candidate policies, and the process is iterated until the desired behavior is obtained. PPL requires a good representation of the policy search space be available, enabling one to learn accurate policy return estimates and limiting the human ranking effort needed to yield a good policy. Furthermore, this representation cannot use informed features (e.g., how far the robot is from any target) due to the simulator-free setting. As a second contribution, this paper proposes a representation based on the agnostic exploitation of the robotic log.

The convergence of PPL is analytically studied and its experimental validation on two problems, involving a single robot in a maze and two interacting robots, is presented.

Multiview Semi-Supervised Learning for Ranking Multilingual Documents

Nicolas Usunier, Massih-Reza Amini, Cyril Goutte

We address the problem of learning to rank documents in a multilingual context, when reference ranking information is only partially available. We propose a multiview learning approach to this semi-supervised ranking task, where the translation of a document in a given language is considered as a view of the document. Although both multiview and semi-supervised learning of classifiers have been studied extensively in recent years, their application to the problem of ranking has received much less attention. We describe a semi-supervised multiview ranking algorithm that exploits a global agreement between view-specific ranking functions on a set of unlabeled observations. We show that our proposed algorithm achieves significant improvements over both semi-supervised multiview classification and semi-supervised single-view rankers on a large multilingual collection of Reuters news covering 5 languages. Our experiments also suggest that our approach is most effective when few labeled documents are available and the classes are imbalanced.

Preference-based policy iteration: Leveraging preference learning for reinforcement learning

Weimei Cheng, Johannes Fürnkranz, Eyke Hüllermeier, Sang-Hyeun Park

This paper makes a first step toward the integration of two subfields of machine learning, namely preference learning and reinforcement learning (RL). An important motivation for a “preference-based” approach to reinforcement learning is a possible extension of the type of feedback an agent may learn from. In particular, while conventional RL methods are essentially confined to deal with numerical rewards, there are many applications in which this type of information is not naturally available, and in which only qualitative reward signals are provided instead. Therefore, building on novel methods for preference learning, our general goal is to equip the RL agent with qualitative policy models, such as ranking functions that allow for sorting its available actions from most to least promising, as well as algorithms for learning such models from qualitative feedback. Concretely, in this paper, we build on an existing method for approximate policy iteration based on roll-outs. While this approach is based on the use of classification methods for generalization and policy learning, we make use of a specific type of preference learning method called label ranking. Advantages of our preference-based policy iteration method are illustrated by means of two case studies.

Rule-Based Active Sampling for Learning to Rank

Rodrigo Silva, Marcos Gonçalves, Adriano Veloso

Learning to rank (L2R) algorithms rely on a labeled training set to generate a ranking model that can be later used to rank new query results. Producing these labeled training sets is usually very costly as it requires human annotators to assess the relevance or order the elements in the training set. Recently, active learning alternatives have been proposed to reduce the labeling effort by selectively sampling an unlabeled set. In this paper we propose a novel rule-based active sampling method for Learning to Rank. Our method actively samples an unlabeled set, selecting new documents to be labeled based on how many relevance inference rules they generate given the previously selected and labeled examples. The smaller the number of generated rules, the more dissimilar and more “informative” is a document with regard to the current state of the labeled set. Differently

from previous solutions, our algorithm does not rely on an initial training seed and can be directly applied to an unlabeled dataset. Also in contrast to previous work, we have a clear stop criterion and do not need to empirically discover the best configuration by running a number of iterations on the validation or test sets. These characteristics make our algorithm highly practical. We demonstrate the effectiveness of our active sampling method on several benchmarking datasets, showing that a significant reduction in training size is possible. Our method selects as little as 1.1% and at most 2.2% of the original training sets, while providing competitive results when compared to state-of-the-art supervised L2R algorithms that use the complete training sets.

A Geometric Approach to Find Nondominated Policies to Imprecise Reward MDPs

Valdinei Freire da Silva, Anna Helena Realí Costa

Markov Decision Processes (MDPs) provide a mathematical framework for modelling decision-making of agents acting in stochastic environments, in which transitions probabilities model the environment dynamics and a reward function evaluates the agent's behaviour. Lately, however, special attention has been brought to the difficulty of modelling precisely the reward function, which has motivated research on MDP with imprecisely specified reward. Some of these works exploit the use of nondominated policies, which are optimal policies for some instantiation of the imprecise reward function. An algorithm that calculates nondominated policies is π Witness, and nondominated policies are used to take decision under the minimax regret evaluation. An interesting matter would be defining a small subset of nondominated policies so that the minimax regret can be calculated faster, but accurately. We modified π Witness to do so. We also present the π Hull algorithm to calculate nondominated policies adopting a geometric approach. Under the assumption that reward functions are linearly defined on a set of features, we show empirically that π Hull can be faster than our modified version of π Witness.

Session 22: Feature Selection, Extraction, and Construction

Thursday 8, 14:00 - 15:50, Olympia Hall

Feature Selection Stability Assessment based on the Jensen-Shannon Divergence

Roberto Guzmán-Martínez, Rocío Alaiiz-Rodríguez

Feature selection and ranking techniques play an important role in the analysis of high-dimensional data. In particular, their stability becomes crucial when the feature importance is later studied in order to better understand the underlying process. The fact that a small change in the dataset may affect the outcome of the feature selection/ranking algorithm has been long overlooked in the literature. We propose an information-theoretic approach, using the Jensen-Shannon divergence to assess this stability (or robustness). Unlike other measures, this new metric is suitable for different algorithm outcomes: full ranked lists, partial sublists (top-k lists) as well as the least studied partial ranked lists. This generalized metric attempts to measure the disagreement among a whole set of lists with the same size, following a probabilistic approach and being able to give more importance to the differences that appear at the top of the list. We illustrate and compare it with popular metrics like the Spearman rank correlation and the Kuncheva's index on feature selection/ranking outcomes artificially generated and on an spectral fat dataset with different filter-based feature selectors.

Fast projections onto $l_{1,q}$ -norm balls for grouped feature selection

Suvrit Sra

Joint sparsity is widely acknowledged as a powerful structural cue for performing feature selection in setups where variables are expected to demonstrate “grouped” behavior. Such grouped behavior is commonly modeled by Group-Lasso or Multitask Lasso-type problems, where feature selection is effected via $l_{1,q}$ -mixed-norms. Several particular formulations for modeling groupwise sparsity have received substantial attention in the literature; and in some cases, efficient algorithms are also available. Surprisingly, for *constrained* formulations of fundamental importance (e.g., regression with an $l_{1,\infty}$ -norm constraint), highly scalable methods seem to be missing. We address this deficiency by presenting a method based on spectral projected-gradient (SPG) that can tackle $l_{1,q}$ -constrained convex regression problems. The most crucial component of our method is an algorithm for projecting onto $l_{1,q}$ -norm balls. We present several numerical results which show that our methods attain up to 30X speedups on large $l_{1,\infty}$ -multitask lasso problems. Even more dramatic are the gains for just the $l_{1,\infty}$ -projection subproblem: we observe almost three orders of magnitude speedups compared against the currently standard method.

A Novel Stability based Feature Selection Framework for k-means Clustering

Dimitrios Mavroeidis, Elena Marchiori

Stability of a learning algorithm with respect to small input perturbations is an important property, as it implies the derived models to be robust with respect to the presence of noisy features and/or data sample fluctuations. In this paper we explore the effect of stability optimization in the standard feature selection process for the continuous (PCA-based) k-means clustering problem. Interestingly, we derive that stability maximization naturally introduces a trade-off between cluster separation and variance, leading to the selection of features that have a high cluster separation index that is not artificially inflated by the feature’s variance. The proposed algorithmic setup is based on a Sparse PCA approach, that selects the features that maximize stability in a greedy fashion. In our study, we also analyze several properties of Sparse PCA relevant to stability that promote Sparse PCA as a viable feature selection mechanism for clustering. The practical relevance of the proposed method is demonstrated in the context of cancer research, where we consider the problem of detecting potential tumor biomarkers using microarray gene expression data. The application of our method to a leukemia dataset shows that the tradeoff between cluster separation and variance leads to the selection of features corresponding to important biomarker genes. Some of them have relative low variance and are not detected without the direct optimization of stability in Sparse PCA based k-means.

Constrained Laplacian Score for semi-supervised feature selection

Khalid Benabdeslem, Mohammed Hindawi

In this paper, we address the problem of semi-supervised feature selection from high-dimensional data. It aims to select the most discriminative and informative features for data analysis. This is a recent addressed challenge in feature selection research when dealing with small labeled data sampled with large unlabeled data in the same set. We present a filter based approach by constraining the known Laplacian score. We evaluate the relevance of a feature according to its locality preserving and constraints preserving ability. The problem is then presented in the spectral graph theory framework with a study of the complexity of the proposed algorithm. Finally, experimental results will be provided

for validating our proposal in comparison with other known feature selection methods.

Feature Selection for Transfer Learning

Selen Uguroglu, Jaime Carbonell

Common assumption in most machine learning algorithms is that, labeled (source) data and unlabeled (target) data are sampled from the same distribution. However, many real world tasks violate this assumption: in temporal domains, feature distributions may vary over time, clinical studies may have sampling bias, or sometimes sufficient labeled data for the domain of interest does not exist, and labeled data from a related domain must be utilized. In such settings, knowing in which dimensions source and target data vary is extremely important to reduce the distance between domains and accurately transfer knowledge. In this paper, we present a novel method to identify variant and invariant features between two datasets. Our contribution is two fold: First, we present a novel transfer learning approach for domain adaptation, and second, we formalize the problem of finding differently distributed features as a convex optimization problem. Experimental studies on synthetic and benchmark real world datasets show that our approach outperform other transfer learning approaches, and it aids the prediction accuracy significantly.

Session 23: Text Mining & Recommender Systems

Thursday 8, 14:00 - 15:50, Attica Hall

Expertise finding using topic models -- the expert--tag--topic model

Gregor Heinrich

Presents an analysis of the structure of mixed-membership models into elementary blocks and their numerical properties. By associating such model structures with structures known or assumed in the data, we propose how models can be constructed in a controlled way, using the numerical properties of data likelihood and Gibbs full conditionals as predictors of model behavior. To illustrate this “bottom-up” design method, example models are constructed that may be used for expertise finding from labeled data.

Analyzing Word Frequencies in Large Text Corpora using Inter-arrival Times and Bootstrapping

Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolamäki, Heikki Mannila

Comparing frequency counts over texts or corpora is an important task in many applications and scientific disciplines. Given a text corpus, we want to test a hypothesis, such as “word X is frequent”, “word X has become more frequent over time”, or “word X is more frequent in male than in female speech”. For this purpose we need a null model of word frequencies. The commonly used bag-of-words model, which corresponds to a Bernoulli process with fixed parameter, does not account for any structure present in natural languages. Using this model for word frequencies results in large numbers of words being reported as unexpectedly frequent. We address how to take into account the inherent occurrence patterns of words in significance testing of word frequencies. Based on studies of words in two large corpora, we propose two methods for modeling word frequencies that both take into account the occurrence patterns of words and go beyond the bag-of-words assumption. The first method models word frequencies based on the spatial distribution of individual words in the language. The second method is based on bootstrapping and takes

into account only word frequency at the text level. The proposed methods are compared to the current gold standard in a series of experiments on both corpora. We find that words obey different spatial patterns in the language, ranging from bursty to non-bursty / uniform, independent of their frequency, showing that the traditional approach leads to many false positives.

An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering

Nicola Barbieri, Giuseppe Manco

In this work we perform an analysis of probabilistic approaches to recommendation upon a different validation perspective, which focuses on accuracy metrics such as recall and precision of the recommendation list. Traditionally, state-of-art approaches to recommendations consider the recommendation process from a “missing value prediction” perspective. This approach simplifies the model validation phase that is based on the minimization of standard error metrics such as RMSE. However, recent studies have pointed several limitations of this approach, showing that a lower RMSE does not necessarily imply improvements in terms of specific recommendations. We demonstrate that the underlying probabilistic framework offers several advantages over traditional methods, in terms of exibility in the generation of the recommendation list and consequently in the accuracy of recommendation.

A Game Theoretic Framework for Data Privacy Preservation in Recommender Systems

Maria Halkidi, Iordanis Koutsopoulos

We address the fundamental tradeoff between privacy preservation and high-quality recommendation stemming from a third party. Multiple users submit their ratings to a third party about items they have viewed. The third party aggregates the ratings and generates personalized recommendations for each user. The quality of recommendations for each user depends on submitted rating profiles from all users, including the user to which the recommendation is destined. Each user would like to declare a rating profile so as to preserve data privacy as much as possible, while not causing deterioration in the quality of the recommendation he would get, compared to the one he would get if he revealed his true private profile.

We employ game theory to model and study the interaction of users and we derive conditions and expressions for the Nash Equilibrium Point (NEP). This consists of the rating strategy of each user, such that no user can benefit in terms of improving its privacy by unilaterally deviating from that point. User strategies converge to the NEP after an iterative best-response strategy update. For a hybrid recommendation system, we find that the NEP strategy for each user in terms of privacy preservation is to declare false rating only for one item, the one that is highly ranked in his private profile and less correlated with items for which he anticipates recommendation. We also present various modes of cooperation by which users can mutually benefit.

iDVS: An Interactive Multi-Document Visual Summarization System

Yi Zhang, Dingding Wang, Tao Li

Multi-document summarization is a fundamental tool for understanding documents. Given a collection of documents, most of existing multi-document summarization

methods automatically generate a static summary for all the users using unsupervised learning techniques such as sentence ranking and clustering. However, these methods almost exclude human from the summarization process. They do not allow for user interaction and do not consider users' feedback which delivers valuable information and can be used as the guidance for summarization. Another limitation is that the generated summaries are displayed in textual format without visual representation. To address the above limitations, in this paper, we develop iDVS, a visualization-enabled multi-document summarization system with users' interaction, to improve the summarization performance using users' feedback and to assist users in document understanding using visualization techniques. In particular, iDVS uses a new semi-supervised document summarization method to dynamically select sentences based on users' interaction. To this regard, iDVS tightly integrates semi-supervised learning with interactive visualization for document summarization. Comprehensive experiments on multi-document summarization using benchmark datasets demonstrate the effectiveness of iDVS, and a user study is conducted to evaluate the users' satisfaction.

Session 24: Reinforcement learning

Thursday 8, 14:00 - 15:50, Kallirhoe Hall

Preference elicitation and inverse reinforcement learning

Constantin Rothkopf, Christos Dimitrakakis

We state the problem of inverse reinforcement learning in terms of preference elicitation, resulting in a principled (Bayesian) statistical formulation. This generalises previous work on Bayesian inverse reinforcement learning and allows us to obtain a posterior distribution on the agent's preferences, policy and optionally, the obtained reward sequence, from observations. We examine the relation of the resulting approach to other statistical methods for inverse reinforcement learning via analysis and experimental results. We show that preferences can be determined accurately, even if the observed agent's policy is sub-optimal with respect to its own preferences. In that case, significantly improved policies with respect to the agent's preferences are obtained, compared to both other methods and to the performance of the demonstrated policy.

Sparse Kernel-SARSA(λ) with an Eligibility Trace

Matthew Robards, Peter Sunehag, Scott Sanner, Bhaskara Marthi

We introduce the first online kernelized version of SARSA(λ) to permit sparsification for arbitrary λ for $0 \leq \lambda \leq 1$; this is possible via a novel kernelization of the eligibility trace that is maintained separately from the kernelized value function. This separation is crucial for preserving the functional structure of the eligibility trace when using sparse kernel projection techniques that are essential for memory efficiency and capacity control. The result is a simple and practical Kernel-SARSA(λ) algorithm for general $0 \leq \lambda \leq 1$ that is memory-efficient in comparison to standard SARSA(λ) (using various basis functions) on a range of domains including a real robotics task running on a Willow Garage PR2 robot.

Analyzing and Escaping Local Optima in Planning as Inference for Partially Observable Domains

Pascal Poupart, Tobias Lang, Marc Toussaint

Planning as inference recently emerged as a versatile approach to decision-theoretic planning and reinforcement learning for single and multi-agent systems in fully and partially observable domains with discrete and continuous variables. Since planning as inference essentially tackles a non-convex optimization problem when the states are partially observable, there is a need to develop techniques that can robustly escape local optima. We investigate the local optima of finite state controllers in single agent partially observable Markov decision processes (POMDPs) that are optimized by expectation maximization (EM). We show that EM converges to controllers that are optimal with respect to a one-step lookahead. To escape local optima, we propose two algorithms: the `_rst` one adds nodes to the controller to ensure optimality with respect to a multi-step lookahead, while the second one splits nodes in a greedy fashion to improve reward likelihood. The approaches are demonstrated empirically on benchmark problems.

Lagrange Dual Decomposition for Finite Horizon Markov Decision Processes

Thomas Furnston, David Barber

Solving finite-horizon Markov Decision Processes with stationary policies is a computationally difficult problem. Our dynamic dual decomposition approach uses Lagrange duality to decouple this hard problem into a sequence of tractable sub-problems. The resulting procedure is a straightforward modification of standard non-stationary Markov Decision Process solvers and gives an upper-bound on the total expected reward. The empirical performance of the method suggests that not only is it a rapidly convergent algorithm, but that it also performs favourably compared to standard planning algorithms such as policy gradients and lower-bound procedures such as Expectation Maximisation.

Reinforcement Learning Through Global Stochastic Search in N-MDPs

Matteo Leonetti, Luca Iocchi, Subramanian Ramamoorthy

Reinforcement Learning (RL) in either fully or partially observable domains usually poses a requirement on the knowledge representation in order to be sound: the underlying stochastic process must be Markovian. In many applications, including those involving interactions between multiple agents (e.g., humans and robots), sources of uncertainty affect rewards and transition dynamics in such a way that a Markovian representation would be computationally very expensive. An alternative formulation of the decision problem involves partially specified behaviors with choice points. While this reduces the complexity of the policy space that must be explored - something that is crucial for realistic autonomous agents that must bound search time - it does render the domain Non-Markovian. In this paper, we present a novel algorithm for reinforcement learning in Non-Markovian domains. Our algorithm, Stochastic Search Monte Carlo, performs a global stochastic search in policy space, shaping the distribution from which the next policy is selected by estimating an upper bound on the value of each action. We experimentally show how, in challenging domains for RL, high-level decisions in Non-Markovian processes can lead to a behavior that is at least as good as the one learned by traditional algorithms, and can be achieved with significantly fewer samples.

19. Demos

There will be three sessions of demo presentations, on Tuesday, on Wednesday, and on Thursday, starting around 10:00 right after the keynote talk in the morning, and lasting until lunch. They will take place in the Olympia Foyer. On Thursday 8th at 16:20 – 17:20 demos will be presented in a separate summary session, in Attica Hall.

Thursday 6th, 10:10 - 12:30

Celebrity Watch: Browsing News Content by Exploiting Social Intelligence

Omar Ali, Ilias Flaounas, Tjil De Bie, Nello Cristianini

Celebrity Watch is an automatically-generated website that presents up-to-date entertainment news from around the world. It demonstrates the application of many pattern analysis methods that allow us to autonomously monitor millions of news articles and hundreds of millions of references to people mentioned in them. We apply statistical methods to merge references into people, track their association to various topics of news, and generate social networks of their co-occurrences in articles. From this sea of data we select the forty most-relevant people and display them on the website, offering users a highly condensed view of the latest in entertainment news. The site updates itself throughout the day and is the final step in a large, fully-autonomous system that monitors online news media.

L-SME: a system for mining loosely structured motifs

Fabio Fassetti, Gianluigi Greco, Giorgio Terracina

We present L-SME, a system to efficiently identify loosely structured motifs in genome-wide applications. L-SME is innovative in three aspects. Firstly, it handles wider classes of motifs than earlier motif discovery systems, by supporting boxes swaps and skips in the motifs structure as well as various kinds of similarity functions. Secondly, in addition to the standard exact search, it supports search via randomization in which guarantees on the quality of the results can be given a-priori based on user-definable resource (time and space) constraints. Finally, L-SME comes equipped with an intuitive graphical interface through which the structure for the motifs of interest can be defined, the discovery method can be selected, and results can be visualized. The tool is flexible and scalable, by allowing genome-wide searches for very complex motifs and is freely accessible at <http://siloe.deis.unical.it/l-sme>. A detailed description of the algorithms underlying L-SME is available in [1].

Exploring City Structure from Georeferenced Photos Using Graph Centrality Measures

Katerina Vrotsou, Natalia Andrienko, Gennady Andrienko, Piotr Jankowski

We explore the potential of applying graph theory measures of centrality to the network of movements extracted from sequences of georeferenced photo captures in order to identify interesting places and explore city structure. We adopt a systematic procedure composed of a series of stages involving the combination of computational methods and interactive visual analytics techniques. The approach is demonstrated using a collection of Flickr photos from the Seattle metropolitan area.

MOA: a Real-time Analytics Open Source Framework

Albert Bifet, Geo Holmes, Bernhard Pfahringer, Jesse Read Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl

Massive Online Analysis (MOA) is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. MOA is designed to deal with the challenging problems of scaling up the implementation of state of the art algorithms to real world dataset sizes and of making algorithms comparable in benchmark streaming settings. It contains a collection of offline and online algorithms for classification, clustering and graph mining as well as tools for evaluation. For researchers the framework yields insights into advantages and disadvantages of different approaches and allows for the creation of benchmark streaming data sets through stored, shared and repeatable settings for the data feeds. Practitioners can use the framework to easily compare algorithms and apply them to real world data sets and settings. MOA supports bi-directional interaction with WEKA, the Waikato Environment for Knowledge Analysis. Besides providing algorithms and measures for evaluation and comparison, MOA is easily extensible with new contributions and allows for the creation of benchmark scenarios.

Wednesday 7th, 10:10 - 12:30

MetaData Retrieval: A Software Prototype for the Annotation of Maps with Social Metadata

Rosa Meo, Elena Roglia, Enrico Ponassi

MetaData Retrieval (MDR) is a software module for the enrichment of geo-referenced maps with metadata. Metadata are annotations on spatial locations that are taken from the Volunteered Graphical Information projects like OpenStreetMap and GeoNames.

The MDR user acts with a user-friendly GUI, a Query By Example in which the user specifies in a multi-dimensional data model the spatial objects for which new information are searched for. The request is translated into SQL queries for the database and in web service requests for OpenStreetMap and GeoNames. Downloaded annotations are checked and compared with the history for duplicate elimination. Annotations are presented to the user in the context of an interactive, georeferenced map and in a hierarchical, ontological structure, that is a facility for indexing and browsing. On demand, an annotation is stored in the system history. Finally, the user can filter the annotations that characterize a specified area by a statistical filter that compares the annotation frequency with the neighborhood.

InFeRno - an Intelligent Framework for Recognizing Pornographic Web Pages

Sotiris Karavarsamis, Nikos Ntarmos, Konstantinos Blekas

In this work we present InFeRno, an intelligent web pornography elimination system, classifying web pages based solely on their visual content. The main characteristics of our system include: (i) a powerful vector space with a small but sufficient number of features that manage to improve the discriminative ability of the SVM classifier; (ii) an extra class (*bikini*) that strengthens the performance of the classifier; (iii) an overall classification scheme that achieves high accuracy at considerably lower runtime costs compared to current state-of-the-art systems; and (iv) a full-fledged implementation of the proposed system capable of being integrated with ICAP-aware web proxy cache servers.

The MASH Project

Francois Fleuret, Philip Abbet, Charles Dubout, Leonidas Lefakis

It has been demonstrated repeatedly that combining multiple types of image features improves the performance of learning-based classification and regression. However, no tools exist to facilitate the creation of large pools of feature extractors by extended teams of contributors.

The MASH project aims at creating such tools. It is organized around the development of a collaborative web platform where participants can contribute feature extractors, browse a repository of existing ones, run image classification and goal-planning experiments, and participate in public large-scale experiments and contests.

The tools provided on the platform facilitate the analysis of experimental results. In particular, they rank the feature extractors according to their efficiency, and help to identify the failure mode of the prediction system.

Activity Recognition With Mobile Phones

Jordan Frank, Shie Mannor, Doina Precup

Our demonstration consists of a working activity and gait recognition system, implemented on a commercial smartphone. The activity recognition feature allows participants to train various activities, such as running, walking, or jumping, on the phone; the system can then identify when those activities are performed. The gait recognition feature learns particular characteristics of how participants walk, allowing the phone to identify the person carrying it.

Thursday 8th, 09:10 - 11:30

Traffic Jams detection using flock mining

Rebecca Ong, Fabio Pinelli, Trasarti Roberto, Mirco Nanni, Chiara Renso, Salvatore Rinzivillo, Fosca Giannotti

TRUMIT: A Tool to Support Large-Scale Mining of Text Association Rules

Robert Neumayer, George Tsatsaronis, Kjetil Nørvåg

Due to the nature of textual data the application of association rule mining in text corpora has attracted the focus of the research scientific community for years. In this paper we demonstrate a system that can efficiently mine association rules from text. The system annotates terms using several annotators, and extracts text association rules between terms or categories of terms. An additional contribution of this work is the inclusion of novel unsupervised evaluation measures for weighting and ranking the importance of the text rules. We demonstrate the functionalities of our system with two text collections, a set of *Wikileaks* documents, and one from TREC-7.

MIME: A Framework for Interactive Visual Pattern Mining

Bart Goethals, Sandy Moens, Jilles Vreeken

We present a framework for interactive visual pattern mining. Our system enables the user to browse through the data and patterns easily and intuitively, using a toolbox consisting of interestingness measures, mining algorithms and post-processing algorithms to assist in identifying interesting patterns. By mining interactively, we enable the user to combine

their subjective interestingness measure and background knowledge with a wide variety of objective measures to easily and quickly mine the most important and interesting patterns. Basically, we enable the user to become an essential part of the mining algorithm. Our demo currently applies to mining interesting itemsets and association rules, and its extension to episodes and decision trees is ongoing research.

20. Tutorials

Monday 5th, morning

Privacy Challenges and Solutions for Medical Data Sharing (Olympia Hall)

Aris Gkoulalas-Divanis, (IBM Research-Zurich), Grigorios Loukides, (Vanderbilt University, U.S.).

Various types of data, including demographics, clinical, and genomic information, are increasingly collected and stored in Electronic Medical Record (EMR) systems and biomedical research repositories. Such data have been traditionally used in automating the workflow of healthcare, but were recently recognized as an invaluable source for performing large-scale, low-cost biological, medical and healthcare analysis. These tasks are essential for the discovery of new drugs and therapies, and are a key step towards realizing the vision of personalized medicine. As a result, over \$50 Billion were pledged by the Obama administration in 2009 to promote technologies for managing and sharing medical data. Meanwhile, detailed medical data are increasingly disseminated beyond the institution they were collected by, in accordance with data sharing regulations, such as the policy of the National Institutes of Health (NIH) for genomic information. This, however, may pose serious threats to patients' privacy, which must be eliminated to comply with data sharing policies and legislation, such as the HIPAA privacy rule and the EU Directive 95/46/CE. In this tutorial, we will elaborate on the need of sharing medical data in a privacy-preserving way, review the existing policies and practices for sharing medical data, and present state-of-the-art approaches for ensuring that the disseminated data are privacy-protected and useful. Following that, we will highlight important open problems and future directions. More specifically, the tutorial will consist of three parts. The first part will provide an overview of successful practices and paradigms to share and use medical data in applications. We will focus on the analysis and mining tasks supported by different types of medical data, as well as on privacy threats that data sharing entails. The second part of the tutorial will survey approaches for privacy-preserving medical data sharing. We will address a number of important issues, such as capturing and balancing data utility and privacy in applications, and designing privacy techniques for different types of data and data sharing scenarios. We will also present interesting case studies using data from the US Census and the EMR system of the Vanderbilt University Medical Center, a state-of-the-art system that stores information about 2 Million patients over 15 years. In the third part of the tutorial, we will discuss important open problems and provide a roadmap for the future.

By the end of this tutorial, the attendees will have a basic understanding of the concepts and underlying principles used to disseminate medical data in a protected and useful form.

The tutorial will be accessible to computer science researchers and educators who are interested in data privacy, data mining, and information systems, as well as to industry developers. By focusing on open problems, we also hope to engage graduate students to conduct research in this emerging and interesting field.

Introduction to causal discovery: A Bayesian Networks approach (Attica Hall)

Ioannis Tsamardinos (University of Crete, Greece, Vanderbilt University, U.S.), Sofia Triantafyllou (University of Crete, Greece).

The tutorial presents an introduction to basic assumptions and techniques for causal discovery from observational data with the use of graphs that represent conditional independence models. It first presents the basic theory of causal discovery such as the Causal Markov Condition, the Faithfulness Condition, and the d-separation criterion, as well as graphical models for representing causality such as Causal Bayesian Networks, Maximal Ancestral Graphs and Partial Ancestral Graphs. It then presents prototypical and state-of-the-art algorithms such as the PC, FCI and HITON for learning such models (global learning) or parts of such models (local learning) from data. The tutorial also discusses the connections of causality to feature selection and presents causality-based feature selection techniques. Apart from the theory and techniques, the tutorial illustrates the use of causal inference through case-studies of applications of causal discovery algorithms with a focus on applications to biomedical data. Finally, it discusses recently introduced directions in the field, such as the integrative analysis of studies using causal models.

Monday 5th, afternoon

Mining complex dynamic data (Olympia Hall)

Hans-Peter Kriege (University Munich, Germany), Irene Ntoutsi (Ludwig Maximilians University Munich, Germany), Myra Spiliopoulou (University Magdeburg, Germany), Grigoris Tsoumakas (Aristotle University of Thessaloniki, Greece), Arthur Zimek (Ludwig Maximilians University Munich, Germany).

In recent years, many applications require mining from richer data types than conventional data(base) records: the analysis of social networks requires the combination of activity recordings with content (e.g. resource descriptions and user records); recommendation engines require considering user ratings, customer transactions, item descriptions and user profiles; medical applications require the combination of different kinds of recordings on patients, including historical data on ailments and medication. At the same time, the mining tasks become more elaborate: the data are multi-faceted and adhere to many, orthogonal or overlapping concepts; the data accumulate or form streams; they are dynamic and call for adaptation of the mining models. In this tutorial, we discuss mining on complex data, putting the emphasis on learning and adaptation over streaming, dynamic data.

We consider three categories of complex data: data that adhere to multiple overlapping labels, high-dimensional data that contain interesting subspaces, and data that span across multiple tables. For each category, we first provide a comprehensive overview of static mining methods, and then focus on methods and example applications for dynamic data. For multi-label stream data, we focus on the example application of document (news) categorization; the core methods are stream classification with decision trees, prediction and ranking. For high-dimensional stream data, we focus on the example application of bio-

informatics and network intrusion; the core methods are stream subspace clustering and outlier detection. For multi-relational stream data, we consider two example applications: analysis of dynamic social networks, and analysis of evolving customer data; the core methods are tensor-based clustering, and multi-relational clustering and classification.

The target groups are: postgraduate students with solid background in data mining; research scholars who work on conventional stream mining and are confronted with applications on complex data; practitioners that own applications on complex and dynamic data.

Factorizing Gigantic Matrices (Attica Hall)

Christian Bauckhage, Kristian Kersting, Christian Thurau (Fraunhofer IAIS, Germany).

Low-rank approximations of data matrices have become an important tool in machine learning and data mining. They allow for embedding high dimensional data in lower dimensional spaces and can therefore mitigate effects due to noise, uncover latent relations, or facilitate further processing. These properties have been proven successful in many applications areas such as bio-informatics, computer vision, text processing, recommender systems, social network analysis, among others. Present day technologies are characterized by exponentially growing amounts of data. Recent advances in sensor technology, Internet applications, and communication networks call for methods that scale to very large and/or growing data matrices. In this tutorial, we discuss basic characteristics of matrix factorization and introduce several recent approaches that scale to modern massive data analysis problems.

Friday 9th, morning

Mining Complex Entities from Heterogeneous Information Networks (Attica Hall)

Fabio Ciravegna, Ziqi Zhang (University of Sheffield, UK)

Most research on information mining has focused on classic Information Extraction (IE) tasks, from structured and unstructured documents, like newspaper articles and web pages. In the last years however the staggering growth of social media as platform for sharing content has moved the focus towards a different type of extraction target. Social media pose a number of challenge to information extraction: contributions to social media sites like blogs, forums, Twitter, etc. are conversational in nature and thus tend to be brief and informal, containing imprecise, subjective and ambiguous information. The expanded context (who the author is, the social and geographical context, their social links, etc.) becomes relevant to disambiguate and interlink information.

Aim of this tutorial is to introduce and discuss issues, methodologies and technologies for extracting information from documents, with a particular focus on mining heterogeneous information networks (e.g. social websites) in order to mine complex entities.

Friday 9th, afternoon

Semantic Data Mining (Attica Hall)

Nada Lavrac (Jožef Stefan Institute, University of Ljubljana and University of Nova Gorica, Slovenia), Anze Vavpetic (Jožef Stefan Institute, University of Ljubljana,

Slovenia), Melanie Hilario (University of Geneva, Switzerland), Alexandros Kalousis (University of Geneva, Switzerland), Agnieszka Lawrynowicz (Poznan University of Technology, Poland), Jędrzej Potoniec (Poznan University of Technology, Poland)

The term semantic data mining denotes a data mining approach where domain ontologies are used as background knowledge. Such approach is motivated by large amounts of data that are increasingly becoming openly available and described using real-life ontologies represented in Semantic Web languages. This recently opened up the possibility for interesting large-scale and real-world semantic applications.

The tutorial will address the problems of how machine learning techniques can work directly on the richly structured Semantic Web data, exploit ontologies, and other Semantic Web technologies, what is the value added of machine learning methods exploiting ontologies, and what are the challenges for developers of semantic data mining methods. It will also contain demonstrations of tools supporting semantic data mining.

The tutorial will present the topic of semantic data mining from three complementary perspectives. Firstly, it will present a general framework for semantic data mining, and illustrate it with an example of a new method for semantic subgroup discovery. The second part of tutorial will cover the topic of learning from description logics (DL-learning). Finally, the third part of the tutorial will cover the topic of semantic meta-mining.

21. Discovery Challenge

The focus of the ECML PKDD 2011 Discovery Challenge is the recommendation of video lectures in the context of VideoLectures.net, a free and open access multimedia repository of video lectures, mainly of research and educational character. The challenge consists of two main tasks and a computational workflow contest. The prizes are sponsored by the EU FP7 project e-LICO.

General description: VideoLectures.net is a free and open access multimedia repository of video lectures, mainly of research and educational character. The lectures are given by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events from many fields of Science. The portal is aimed at promoting science, exchanging ideas and fostering knowledge sharing by providing high quality didactic contents not only to the scientific community but also to the general public. All lectures, accompanying documents, information and links are systematically selected and classified through the editorial process taking into account also users' comments.

The ECML PKDD 2011 Discovery Challenge is organized in order to improve the website's current recommender system. The challenge consists of two main tasks and a "side-by" contest. The provided data is for both of the tasks, and it is up to the contestants how it will be used for learning (building up) a recommender.

Due to the nature of the problem, each of the tasks has its own merit: task 1 simulates new-user and new-item recommendation (cold-start mode), task 2 simulates clickstream based recommendation (normal mode).

Data from VL.Net site does not include any explicit nor implicit user profiles. Due to the privacy-preserving constraints implicit profiles embodied in viewing sequences (click-

streams) have been transformed, so that no individual viewing sequence information can be revealed or reconstructed. There are however other viewing related data included: i) co-viewing frequencies ii) pooled viewing sequences, and iii) content related information available: lecture category taxonomy, names, descriptions and slide titles (where available), authors, institutions, lecture events and timestamps.

Tasks: The ECML-PKDD 2011 Discovery Challenge involves two main tasks and one “side-by” contest:

- Cold start problem
- Recommendation based on pooled lecture viewing sequences
- Computational workflow contest.

Dataset: The dataset used for the ECML-PKDD 2011 Discovery Challenge is based on a data snapshot from VideoLectures.net, which was taken in August 2010.

Prizes

The awards for each of the tracks are:

- 1500€ for the first place
- 700€ for the second place
- 300€ for the third place

The awards for the the best solution-workflow:

- 500€
- Free admission to RapidMiner Community Meeting and Conference 2012 for the best RapidMiner workflow

Awards Chairs: Tomislav Šmuc, Ingo Mierswa

The winner of Discovery Challenge is Alexander D'Yakonov (Overall winner: Track1 and Track2).

Sponsors

The Discovery Challenge is supported by the European Project e-LICO, Rapid-I, Viidea Ltd.

Organizers:

- Matko Bošnjak, Nino Antulov-Fantulin, Tomislav Šmuc, Ruđer Bošković Institute, Croatia
- Nada Lavrač, Mitja Jermol, Martin Žnidaršič, Jožef Stefan Institute, Slovenia
- Peter Keše, Viidea Ltd, Slovenia

22. Workshops

Monday, September 5th

Joint ECML PKDD - PASCAL Workshop on Large-Scale Hierarchical Classification - LSHTC - (Monday, full day, Kallirhoe Hall)

Organizers

- George Paliouras, NCSR “Demokritos”, Athens, Greece
- Eric Gaussier, LIG, Grenoble, France
- Aris Kosmopoulos, NCSR “Demokritos” & AUEB, Athens, Greece
- Ion Androutsopoulos, AUEB, Athens, Greece
- Thierry Artières, LIP6, Paris, France
- Patrick Gallinari, LIP6, Paris, France

Program

9:00 - 10:00 *Introduction (workshop organizers)*

10:00 – 10:30 *Session I*

- An Optimized K-Nearest Neighbor Algorithm for Large Scale Hierarchical Text Classification. *Xiaogang Han, Junfa Liu, Zhiqi Shen, Chunyan Miao*

10:30 - 11:00 *Coffee Break*

11:00 – 12:30 *Session II*

- 11:00 - 11:30 Towards Using Reranking in Hierarchical Classification. *Ju Qi, Richard Johansson, Alessandro Moschitti*
- 11:30 - 12:00 Voting using Minimally-Sized Decomposition Schemes. *Evgueni Smirnov, Matthijs Moed, Georgi Nalbantov, Ida Sprinkhuizen-Kuyper*
- 12:00 - 12:30 Learning efficient error correcting output codes for large hierarchical multi-class problems. *Cisse Mouhamadou Moustapha, Artières Thierry, Gallinari Patrick*

12:30 - 14:00 *Lunch (on your own)*

14:00 – 15:30 *Session III*

- 14:00 - 15:00 Invited Talk. *Axel Ngonga* : Linking the Web of Data.
- 15:00 - 15:30 ECHO at the LSHTC Pascal Challenge 2. *Christophe Brouard*

15:30 - 16:00: *Coffee break*

16:00 – 17:00 *Round Table & Open Discussion*

4th Workshop on Intelligent Techniques in Software Engineering - ISEW - (Monday, full day, Conference Room I)

Organizers

- Ioannis Stamelos, Aristotle University, Thessaloniki, Greece
- Stamatia Bibi, Aristotle University, Thessaloniki, Greece

Program

9:00 - 9:15 *Workshop Registration*

9:15 - 9:30 *Welcome, I. Stamelos*

9:30 – 10:30 *Session I*

- 9:30 - 10:00 Experienced Based Process Support via Dynamic Clustering and Progress Estimation. *Patrick Dohrmann*
- 10:00 - 10:30 Predicting User Actions in Software Processes. *Michael Deynet*

10:30 - 11:00 *Coffee Break*

11:00 – 12:30 *Session II*

- 11:00 - 11:30 Discovering patterns of correlation and similarities in software project data with the Circos visualization tool. *Makrina Viola Kosti, Sofia Lazaridou, Nikoleta Bourazani, Lefteris Angelis*
- 11:30 - 12:00 Open Source Software: How can design metrics facilitate architecture recovery. *Eleni Constantinou, George Kakarontzas, Ioannis Stamelos*
- 12:00 – 12:30 Open Discussion on ISEW emerging issues

12:30 - 14:00 *Lunch (on your own)*

14:00 – 15:30 *Open Discussion on ISEW emerging issues, cont'd, all participants*

15:30 - 16:00 *Coffee Break*

16:00 – 17:30 *Discussion on future joint research projects, all participants*

Workshop Mining Ubiquitous and Social Environments - *MUSE 2011* - (Monday, full day, Conference Room II)

Organizers

- Martin Atzmueller, University of Kassel, Germany
- Andreas Hotho, University of Wuerzburg, Germany

Program

9:00 - 9:20 *Opening*

9:20 – 10:30 *Session I*

- 9:20 - 9:50 Dealing with Collinearity in Learning Regression Models from Geographically Distributed Data. *Annalisa Appice, Michelangelo Ceci, Donato Malerba, Antonietta Lanza*
- 9:50 - 10:10 Spatio-Temporal Reconstruction of Un-Sampled Data in a Sensor Network. *Pietro Guccione, Anna Ciampi, Annalisa Appice, Donato Malerba, Angelo Muolo*
- 10:10 - 10:30 A Data Generator for Multi-Stream Data. *Zaigham Faraz Siddiqui, Myra Spiliopoulou, Panagiotis Symeonidis, Eleftherios Tiakas*

10:30 - 11:00 *Coffee Break*

11:00 – 12:20 *Session II*

- 11:00 - 11:30 The Generation of User Interest Profiles from Semantic Quiz Games. *Magnus Knuth, Nadine Ludwig, Lina Wolf, Harald Sack*
- 11:30 - 12:00: Face-to-Face Contacts during a Conference - Communities, Roles, and Key Players. *Martin Atzmueller, Stephan Doerfel, Andreas Hotho, Folke Mitzlaff, Gerd Stumme*
- 12:00 - 12:20 Graph Indexing Challenges in Detecting Misuse and Fraud in Network Data. *Ursula Redmond, Martin Harrigan, Padraig Cunningham*

12:20 - 14:00 *Lunch (on your own)*

14:00 – 15:30 *Session III*

- 14:00 - 15:00 Invited Talk
- 15:00 - 15:30 Discussion and Closing

ECML-PKDD Discovery Challenge Workshop - VideoLectures.Net Recommender System Challenge (Monday, morning, Conference Room III)

Organizers

- Matko Bošnjak, Ruder Bošković Institute, Croatia
- Nino Antulov-Fantulin, Ruder Bošković Institute, Croatia
- Tomislav Šmuc, Ruder Bošković Institute, Croatia
- Nada Lavrač, Jožef Stefan Institute, Slovenia
- Mitja Jermol, Jožef Stefan Institute, Slovenia
- Martin Žnidaršič, Jožef Stefan Institute, Slovenia
- Peter Keše, Viidea Ltd, Slovenia

Program

9:00 - 9:15 *Opening, winners and awards. DCW Organizers*

9:15-10:30 *Session I*

- 9:20 - 9:50 Two Recommendation Algorithms Based on Deformed Linear Combinations. *Alexander D'Yakonov (Overall winner: Track1 and Track2)*
- 9:50 - 10:10 A Content-based Approach for Cold-start Recommendations of Videolectures. *Eleftherios Spyromitros Xioufis, Emmanouela Stachtiari, Grigorios Tsoumakas, Ioannis Vlahavas*
- 10:10 - 10:30 Recommending VideoLectures with Linear Regression. *Martin Mozi-na, Aleksander Sadikov, Ivan Bratko*

10:30 - 11:00 *Coffee Break*

11:00 - 12:20 *Session II*

- 11:00 - 11:20 Lightweight Approach to the Cold Start Problem in the Video Lecture Recommendation. *Leo Iaquinta, Giovanni Semerar*
- 11:20 - 11:40 IRIT at VLNetChallenge. *Max Chevalier, Taoufiq Dkaki, Damien Dudognon, Josiane Mothe*
- 11:40 - 12:00 Using Co-view Information to Learn Lecture Recommendations. *Haibin Liu, Sujatha Das Gollapalli, Dongwon Lee, Prasenjit Mitra, C.Lee Giles*
- 12:00 - 12:20 Overview of the datasets and the challenge. *DCW Organizers*

2nd *MultiClust* Workshop on Discovering, Summarizing and Using Multiple Clusterings (Monday, morning, Conference Room IV)

Organizers

- Emmanuel Müller, Karlsruhe Institute of Technology (KIT), Germany
- Stephan Günemann, RWTH Aachen University, Germany
- Ira Assent, Aarhus University, Denmark
- Thomas Seidl, RWTH Aachen University, Germany

Program:

9:00-10:30 Session I

- 09:00 – 09:30 Invited Talk. *Michael Houle* : Combinatorial Approaches to Clustering and Feature Selection
- 09:30 – 09:45 Subjectively interesting alternative clusters. *Tijl De Bie*
- 09:45 – 10:00 When Pattern Met Subspace Cluster. *Jilles Vreeken, Arthur Zimek*
- 10:00 – 10:15 Fast Multidimensional Clustering of Categorical Data. *Tengfei Liu, Nevin L. Zhang, Kin Man Poon, Yi Wang, Hua Liu*
- 10:15 – 10:30 Factorial Clustering with an Application to Plant Distribution Data. *Manfred Jaeger, Simon Lyager, Michael Vandborg and Thomas Wohlgemuth*

10:30 - 11:00 Coffee Break

11:00 - 12:30 Session II

- 11:00 - 11:30 Invited Talk. *Bart Goethals* : Cartification: from Similarities to Itemset Frequencies
- 11:30 – 11:45 Browsing Robust Clustering-Alternatives. *Martin Hahmann, Dirk Habich, Wolfgang Lehner*
- 11:45 – 12:00 Evaluation of Multiple Clustering Solutions. *Hans-Peter Kriegel, Erich Schubert, Arthur Zimek*
- 12:00 – 12:15 Generating A Diverse Set Of High-Quality Clusterings. *Jeff M. Phillips, Parasaran Raman, Suresh Venkatasubramanian*
- 12:15 - 12:30 Discussion Panel

Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions - *DMFGP* - (Monday, morning, Conference Room V)

Organizers:

- José M. Peña, UPM, Spain
- A. Fazel Famili, NRC, Canada
- Ana Teresa Freitas, INESC-ID/IST, Portugal
- Jaakko Hollmén, Helsinki University of Technology, Finland
- Alexander Schliep, Rutgers University, USA
- Henrik Bostrom, Stockholm University, Sweden
- Min-Ling Zhang, South East University, Nanjing, China
- Evgenii Vityaev, Russian Academy of Science, Russia

Program:

9:00 - 10:30 Session I

- Identifying informative genes in highly imbalanced gene expression data. *Fazel Famili, Ziyang Liu, Sieu Phan*
- Comparing Functional Visualizations of Genes Hamid Ghou. *Nicholas Ho, Daniel R. Catchpoole, Paul J. Kennedy*
- Exploiting Gene Expression data for Pharmacogenomics analysis. *S. Gonzalez, Y. Cheng, V. Robles*
- Finding HCV NS5A Discriminative Motifs for Assesment of IFN/Ribavarin Therapy Effect. *Tu Bao Ho, Saori Kawasaki, Ngoc Tu Le, Tatsuo Kanda, Nhan Ley, Katsuhiko Takabayashi, Osamu Yokosuka*

10:30 - 11:00 Coffee Break

11:00 - 12:30 Session II

- A Genome-Wide Study of the Effect of Aging on Level-2 Gene-Ontology Categories in Mice Using Mixed Models Vincenzo Lagani. *Ioannis Tsamardinos, Magda Grammatikou, George Garinis*
- Characterizing and extracting the synaptic apposition surface for the analysis of synaptic geometry. *Juan Morales, Angel Rodriguez, Jose-Rodrigo Rodriguez, Javier DeFelipe, Angel Merchan-Perez*
- Non-Linear Computational Evolutionary Environment (NLCEE): Building High-level Knowledge in Complex Biological Databases. *Laurence Rodrigues do Amaral, Estevam Rafael Hruschka Junior*
- 12:10 - 12:30 Conclusion and Discussion

Friday, September 9th

Workshop on Collective Learning and Inference on Structured Data - *CoLISD* - (Friday, full day, Conference Room I)

Organizers

- Balaraman Ravindran, IIT Madras, Chennai, India
- Kristian Kersting, Fraunhofer IAIS, Germany
- Sriraam Natarajan, Wake Forest University Baptist Medical Center, North Carolina
- Indrajit Bhattacharya, Indian Institute of Science, Bangalore, India
- Subramanian. Shivashankar, IIT Madras, Ericsson R&D, Chennai, India

Program

10:30 - 10:35 *Welcome*

10:35 - 13:30 *Session I*

- 10:35 - 11:35 Invited Talk. *Jennifer Neville, Purdue University*
- 11:35 - 11:55 Joint Mode Estimation in Multi-Label Classification by Chaining. *Krzysztof Dembszynski, Willem Waegeman, Eyke Hüllermeier*
- 11:55 - 12:10 *Short Break*
- 12:10 - 12:30 A Graph-based Bagging. *Nils Murrugarra-Llerena, Alneu Lopes*
- 12:30 - 13:30 Invited Talk. *Paolo Frasconi, University of Florence* : kLog: a language for logical and relational learning with kernels

13:30 - 14:50 *Lunch (on your own)*

14:50 - 16:30 *Session II*

- 14:50 - 15:50 Invited Talk. *Marco Gori, Università di Siena*
- 15:50 - 16:10 Kernels for measuring similarity of EL++ description logic concepts. *Lukasz Jozefowski, Agnieszka Lawrynowicz, Joanna Jozefowska, Jędrzej Potoniec, Tomasz Lukaszewski*
- 16:10 - 16:30 Second Order Pseudolikelihood Learning in Relational Domain. *Krishna Kumar Tiwar, V. Vijaya Saradhi*

16:30 - 17:20 *Poster Session with Coffee Break*

- Learning Order in Boosting-based Classification of Structured Output Elements. *Tomasz Kajdanowicz, Przemyslaw Kazienko*
- An Adaptive Graph-Based K-Nearest Neighbor. *Nils Murrugarra-Llerena, Alneu Lopes*
- Parallel Inference on Structured Data with CRFs on GPUs. *Nico Piatkowski, Katharina Morik*
- Is Frequent Pattern Mining useful in building predictive models?. *Thashmee Karunaratne*
- Collective Learning of the Community Effect on Churn. *M. Saravanan, S. Bharanidharan, Balaraman Ravindran.*
- Machine Reading Using Markov Logic Networks for Collective Probabilistic Inference. *Shalini Ghosh, Natarajan Shankar, Sam Owre*

17:20 - 18:30 *Session III*

- 17:20 - 18:20 Invited Talk. *Amir Globerson, The Hebrew University of Jerusalem*
- 18:20 - 18:30 Concluding Remarks

Workshop on Planning to Learn and Service-Oriented Knowledge Discovery - *Plan-SoKD-11* - (Friday, morning, Conference Room II)

Organizers

- Jörg-Uwe Kietz, University of Zurich, Switzerland
- Simon Fischer, Rapid-I, Germany
- Nada Lavrač, Jozef Stefan Institute, Slovenia
- Vid Podpecan, Jozef Stefan Institute, Slovenia

Program

10:30 - 10:35 Welcome

10:35 - 12:40 Session I

- 10:35 - 11:00 A meta-mining infrastructure to support KD workflow optimization. *Phong Nguyen, Alexandros Kalousis, Melanie Hilario*
- 11:00 - 11:25 RMonto - towards KDD workflows for ontology-based data mining. *Jedrzej Potoniec, Agnieszka Lawrynowicz*
- 11:25 - 11:50 Using Ontologies in Semantic Data Mining with SEGS and g-SEGS. *Nada Lavrač, Anže Vavpetič, Larisa Soldatova, Igor Trajkovski, Petra Kralj Novak*
- 11:50 - 12:15 A Browser-based Platform for Service-Oriented Knowledge Discovery. *Janez Kranjc, Vid Podpečan, Nada Lavrač*
- 12:15 - 12:40 Web Services for Stream Mining: A Stream-Based Active Learning Use Case. *Martin Saveski, Miha Grčar*

12:40 - 13:45 Session II Plenary System Demos (10-15min each)

- A Browser-based Platform for Service-Oriented Knowledge Discovery
- Taverna: A workflow engine calling Data Mining Services (invited demo)
- Integration of an Intelligent Discovery Assistant into RapidMiner (invited demo)
- A meta-mining infrastructure to support KD workflow optimization
- Semantic Data Mining System g-SEGS
- 13:40 - 13:45 Closing (Future of PlanSoKD)

Workshop on Mining Complex Entities from Network and Biomedical Data - MIND 2011 - (Friday, afternoon, Conference Room II)

Organizers

- Stefano Ferilli, University of Bari, Italy
- Corrado Loglisci, University of Bari, Italy
- Michael Schroeder, Technical University of Dresden, Germany
- George Tsatsaronis, Technical University of Dresden, Germany
- Iraklis Varlamis, Harokopio University of Athens, Greece

Program

15:00 – 16:30 Session I: Mining Biomedical Data

- 15:00 – 15:30 Inter-Event Dependencies support Event Extraction from Biomedical Literature (full paper). *Roman Klinger, Sebastian Riedel, Andrew Mccallum*
- 15:30 – 16:00 An Experimental Evaluation of Lifted Gene Sets (full paper). *Ondrej Kuželka, Filip Zelezny*
- 16:00 – 16:30 Creating a focused corpus of factual outcomes from biomedical experiments (full paper). *James Eales, George Demetriou, Robert Stevens*

16:30 – 17:00 Poster Session with Coffee Break

17:00 – 18:30 Session II: Mining Networks over Time, Biomedical Applications, Conclusions

- 17:00 – 17:30 Learning to Rank by Transferring Knowledge Across Different Time Windows (full paper). *Lucrezia Macchia, Michelangelo Ceci, Donato Malerba*
- 17:30 – 17:50 Automated generation of meta-wikis: strategies and open challenges (short paper). *Salvatore Loguercio, Benjamin Good, Andrew Su*
- 17:50 - 18:30 Discussion and Closing Remarks

Workshop on Knowledge Discovery in Health Care and Medicine - *KD-HCM* - (Friday, full day, Conference Room III)

Organizers

- Huzefa Rangwala, George Mason University, US
- Andrea Tagarelli, University of Calabria, IT
- Nikil Wale, Pfizer, US
- George Karypis, University of Minnesota, US

Program

10:30 – 10:40 Introduction to Workshop: Goals and Description. Co-Chairs (Rangwala, Tagarelli, Wale, Karypis)

10:40 – 13:30 Session I

- 10:40 – 11:30 Invited Talk. *Prof. Dr. Stefan Kramer (Technische Universitaet Muenchen, Germany)*. Data Mining Methods for Predictive Toxicology
- 11:30 – 12:00 A Genetic Fuzzy Approach for Rule Extraction for Rule-Based Classification with Application to Medical Diagnosis. *Rahil Hosseini*
- 12:00 – 12:30 Identification of Adverse Drug Event Assertive Sentences in Medical Case Reports. *Harsha Gurulingappa, Juliane Fluck, Martin Hofmann-Apitius, Luca Toldo*
- 12:30 – 13:00 Analysis of Decision Tree Pruning Using Windowing in Medical Datasets with Different Class Distributions. *Pedro Perez, José Baranauskas*
- 13:00 – 13:30 CSDMSD: A Secure Framework to Anonymously Share and Distribute Medical-Sequencing Data. *Ahmed Al-Faresi, Duminda*

13:30 – 15:00 Lunch (on your own)

15:00 – 16:30 Session II

- 15:00 – 15:30 Preserving Data Privacy Using Coalitional Game Theory. *Srinivasa Chakravarthy L, Valli Kumari Vatsavayi*
- 15:30 – 16:00 Prediction of Surgery Duration using Empirical Anesthesia Protocols. *Rene Schult, Pawel Matuszyk, Myra Spiliopoulou*
- 16:00 – 16:30 Real-time prediction of an anesthetic monitor index using machine learning. *Gianluca Bontempi, Olivier Caelen, Olivier Cailloux, Abhilash Miranda, Luc Barvais, Djamal Ghoundiwal*

16:30 – 17:00 Coffee Break

17:00 – 18:00 Session III

- 17:00- 17:30 Nonparametric scoring methods as a support decision tool for medical diagnosis. The TreeRank algorithm and its variants. *Marine Depecker, Nicolas Vayatis, Stéphane Cléménçon*
- 17:30 – 18:00 Hybrid Intelligent Model for the Diagnosis of Impaired Glucose Tolerance. *Nahla Barakat, Mohamed Nabil Barakat*

Workshop on Machine Learning and Data Mining in and around Games - *MLDMG* - (Friday, full day, Conference Room IV)

Organizers

- Tom Croonenborghs, Katholieke Hogeschool Kempen, Belgium
- Kurt Driessens, Maastricht University, Netherlands
- Olana Missura, University of Bonn, Fraunhofer IAIS, Germany

Program

10:30 – 10:45 Introduction

10:45 – 13:30 Session I

- 10:45 – 11:10 Action Sequence Mining. *Sarah Breining, Hans-Peter Kriegel, Matthias Schubert, Andreas Züfle*
- 11:10 – 11:35 Learning Move Sequence Knowledge Having Configurational and Contextual Elements. *Arthur Cater*
- 11:35 – 12:00 Automatic Discretization of Actions and States in Monte-Carlo Tree Search. *Guy Van den Broeck, Kurt Driessens*
- *12:00 – 12:30 Short Break*
- 12:30 – 13:30 Invited Talk. *Jaideep Srivastava* : Behavioral Analytics. Data Mining as a Key Enabler of Computational Behavioral Sciences

13:30 – 15:00 Lunch (on your own)

15:00 – 16:30 Session II

- 15:00 – 15:25 Dynamic Difficulty Adjustment on Partially Ordered Sets. *Olana Missura, Thomas Gärtner*
- 15:25 – 15:50 Hedging Algorithms and Repeated Matrix Games. *Bruno Bouzy, Marc Métivier, Damien Pellier*
- 15:50 – 16:15 Semantic Representation of Action Games. *Costas Boletsis, Dimitra Chasanidou, Panagiotis Pandis, Katia Lida Kermanidis*
- 16:15 – 16:30 Closing

Workshop on Finding patterns of human behaviors in NETWORKS and MOBILITY data - NEMO - (Friday, full day, Conference Room V)

Organizers

- Albert-László Barabási, Northeastern University, USA
- Michele Berlingerio, ISTI-CNR Pisa, Italy
- Dino Pedreschi, University of Pisa, Italy
- Dashun Wang, Northeastern University, USA

Program

10:30 - 13:45: Session I

- Opening *Workshop Chairs*
- Where Traffic Meets DNA. *Ahmed Jawad, Kristian Kersting*
- Checking out checking in: observations on Foursquare usage patterns. *Martin Chorley, Gualtiero Colombo, Matthew Williams, Stuart Allen, Roger Whitaker*
- From mobility data to social attitudes: a complex network approach. *Sabrina Gaito, Gian Paolo Rossi, Matteo Zignani*
- Pedestrian RoutePrediction using Augmented Cover Trees. *Gavin Smith, James Goulding*
- Augmented Betweenness Centrality for Mobility Prediction in Transportation Networks. *Yaniv Altshuler, Rami Puzis, Yuval Elovici, Shlomo Bekhor, Alex (Sandy) Pentland*

13:45 - 15:00 Lunch (on your own)

15:00 - 16:30 Session II

- Clustering Multiple and Flexible Time Intervals in Sequential Patterns Towards Predictive Modeling of Human Gait Behavior. *Roberto Legaspi, Danaipat Sodkomkham, Kazuya Maruo, Kenichi Fukui, Koichi Moriyama, Satoshi Kurihara, Masayuki Numao*
- Generalized network community detection. *Lovro Šubelj, Marko Bajec*
- A Hybrid Multi-objective Evolutionary Algorithm for Community Detection in Complex Networks. *Babak Amiri*

16:30 - 17:00 Coffee Break

17:00 - 18:30 Interactive Panel. *Authors and Workshop Chairs*

23. Athens



The city with the most glorious history in the world, a city worshipped by gods and people, a magical city. The enchanting capital of Greece has always been a birthplace for civilization. It is the city where democracy was born and most of the wise men of ancient times. The most important civilization of ancient world flourished in Athens and relives through some of the world's most formidable edifices.

Who hasn't heard of the Acropolis of Athens? Photos and history of the most famous archaeological monument in Europe have made the world tour causing feelings of admiration by thousands of people. Acropolis is nominated to be one of the 7 wonders of modern world. In fact the trademark of Athens is one of the favorites. The Holy Rock of Acropolis dates back to the 5th BC, the famous Golden Age of Periklis. Athens met times of bloom and decline, but still shines under the Attic sky gazing the future. Still sparkling like the marbles of Parthenon and the limpid white of Pentelic marble.

Athens is situated in the prefecture of Attica and extends to the peninsula that reaches up to Central Greece. It is surrounded by mountains Ymmytos, Pendeli and Parnitha, northwards and eastwards, and the Saronic gulf southwards and westwards. The sun is shining over Athens all year round. The climate is one of the best in Europe, with mild winters and very hot summers, ideal for tourism. It is located just a few kilometers from the port of Piraeus, the central commercial port of the capital, and the shores of southern Attica.

Athens is constantly inhabited since Neolithic Age. The 5th century was the time of its ultimate bloom, when moral values and civilization surpassed city limits and became the mother land of western civilization. In the centuries that followed, many conquerors tried to take over Athens. In 1834 Athens was chosen to be the capital of the newly established Greek State. The city that now hosts more than 4,5 million people, was constructed around the Acropolis walls. Today it is the political, social, cultural, financial and commercial center of Greece.

Athens is a city of different aspects. A walk around the famous historic triangle (Plaka, Thission, Psyri) the old neighborhoods, reveal the coexistence of different eras. Old mansions, well-preserved ones and other worn down by time. Luxurious department stores

and small intimate shops, fancy restaurants and traditional taverns. All have their place in this city.

The heart of Athens beats in Syntagma Square. Where Parliament and most of the Ministries are. Monastiraki, Kolonaki and Lycabettus Hill attract thousands of visitors all year round. A few kilometers from the historic center in Faliro, Glyfada, Voula and Vouliagmeni, you can enjoy the sea breeze. Or you can head up north and enjoy the fresh air at the more classy neighborhoods of Marousi, Melissa, Vrilissia and of course Kifisia.

Athens and Attica in general have the most important archaeological monuments (Acropolis, Odeion of Herodes Atticus, Olymbion, Roman Market, Panathinaiko Stadium or Kallimarmaro, The Temple of Poseidon in Sounio, etc). In the capital you will admire many imposing neoclassic buildings, true ornaments of the city (The Greek Parliament, Athens Academy and University, etc). Don't miss visiting the museums hosting unique treasures of our cultural inheritance (Archaeological Museum, Military Museum, Byzantine Museum, etc).

Athens has always attracted peoples' attention. During the 2004 Olympic Games proved that, despite all the slings and arrows of outrageous fortune, she never - not once - lost the talent. The return of Olympic Games to their birthplace was a great success.

The capital is famous, more than any other European capital, for its nightlife. Athens by night totally changes. The options for entertainment satisfy all tastes. The famous "bouzoukia" are the leaders in the Athenian entertainment. While the numerous theaters all around Athens offer a different type of entertainment. Athens is a divine city. Let it enchant you...

Local Attractions¹

The Acropolis



The sacred rock of the Acropolis and its most recognized monument, the Parthenon, have withstood the test of time. Natives have often commented on its commanding presence that is instilled in their daily life. But nothing compares to witnessing its grandeur up close and in person. The archeological park, known as the Unification of Archaeological Sites, (4.7 km or 3 miles wide) that surrounds the Acropolis, encompasses some of the world's most ancient treasures.

The New Acropolis Museum



Years in the making, this museum and its creative use of natural Greek light is the new gem of Athens and has been heralded as a masterpiece in itself. The permanent collections present finds and artifacts from the sacred hill of the Acropolis, while smaller «vignette» temporary exhibits offer insight on the whole. The cafe and museum shops are quite popular and are a must to visit as well.

National Archaeological Museum



One of the richest museums of ancient Greek art in the world, its collections span cultures that flourished in Greece from the pre-historic age and beyond. A bronze statue of Poseidon is here as are frescoes from ancient Thira. A comprehensive museum that is often overlooked.

Greek Parliament



The Greek Parliament and the Tomb of the Unknown Soldier. Every vacation portfolio should not be without a photo alongside the tall, commanding Presidential Guards, known as evzones or soliadés. Worth the wait is to witness the changing of the guards, a ten-minute ceremonial procedure that takes place every hour on the hour. The foustanela or skirt that is part of their uniform is made up of 400 pleats, each one symbolizing a year that Greece was under Turkish rule.

Panathenaic Stadium



This must-see monument opposite Zappeion Gardens on Vassilisis Konstantinou Avenue is the stadium that hosted the first modern Olympic Games in Athens in 1896. The stadium, first built in 330 B.C., is made of white marble from the mountain Penteli and has a seating capacity of 45,000 for the concerts and events.

Lycabettus Hill



At a height of 277 metres (approximately 1,000) feet Lycabettus Hill is perhaps the best spot in which to get an aerial view of the city. Visible from here is the Acropolis, the port of Pireaus, and the island of Aigina. If a mini-trek up is not appealing, take the cable car to the top (and back down). The entrance is on the corner of Aristipou and Ploutarchou streets. If you decide to walk down the forest path you will encounter Dexameni Square in Kolonaki, where you can grab a bite to eat.

Ancient Agora, Monastiraki



Its befitting that this monument the center of commercial and business life in ancient times would later give rise to the buzzing shopping district that surrounds it today. Of course, Monastiraki does not compare to the milieu of the Ancient Agora, but it still continues to inspire those who live, work and visit the area.

Plaka



With its undisputable charm, this area is one of the most frequented by visitors and natives alike. Plaka's winding pathways carry thousands of years of history. Walk amongst the buildings whose facades are dressed in 19th century neoclassical design and architecture. Dine at one or several of its restaurants. And explore the ancient monuments, contemporary museums and traditional souvenir shops throughout.

The Attica Coastline



Athens is surrounded by pristine beaches, where you can swim for many months during the year. Visit a beach in Athens and you are likely to feel like you're on a Greek island, as you are greeted with stretches of crystal sands, fine pebbles and blue, clean waters. The tram and bus take you to nearby, organized beaches (some offer water sports) in Faliro, Alimo, Kalamaki, Glyfada, Schinia and Varkiza in less than an hour.

Ideal for the whole family is a walk on the Flisvos Marina promenades a great destination for all ages, at any time of year.

Temple of Poseidon-Cape Sounion



Take a road trip to the southernmost tip of Attica for a breathtaking drive along the coastal highway and you are rewarded with a visit to one of the most fascinating temples in ancient history. It is no wonder that the ancient Greeks built the temple to their sea god Poseidon here in Sounion. Situated on a plateau on the top of a cliff it welcomes ships and sailors even today.

Gastronomic Neighborhoods²

Here is a breakdown of gastronomic districts that offer tastes which are sure to satisfy:

Sintagma Square: We start our gastronomic tour in Sintagma Square in the center of Athens and in the area around the tree-lined *Vassilisis Sofias Avenue*. There are several fast food and chic restaurants and cafés to stop by and enjoy a cup of coffee, a fresh salad or, in the summer time, a grilled fish or cocktail whilst overlooking the Athens cityscape. Several of the area's hotels offer fine dining at their restaurants with well-known chefs at the helm. Reservations are recommended. Choose to eat a light meal at the Benaki Museum or *Goulandris Museum of Cycladic Art*, two contemporary designed cafés, located along *Vassilisis Sofias Avenue*. Another option is to “do lunch” as corporate Athenians do at one of the nearby bistros around Sintagma Square.



Kolonaki: This popular elegant Athenian neighborhood is the safest choice for finding a great place to eat, drink or shop. Along the narrow, hilly sidewalks are fine and stylish eateries such as all-day tavernas, grills, international restaurants and brasseries, bistros, bars and an esplanade of outdoor cafés perfect for people-watching or relaxing after taking in the delightful shops.

Plaka, Theseion, Psirri, Monastiraki Plaka, located at the base of the *Kolonaki*, is also one of the more lively and traditional places to eat and enjoy traditional Greek culture all year round. In the spring and summer months, many of the restaurants in Plaka are set up in outdoor gardens where the aroma of jasmine blends with classical Greek melodies, for an experience that satisfies all senses. Just a bit further down is *Monastiraki* which is a maze of market-filled streets, perfect for buying souvenirs. It features many ouzeri or spots to enjoy a Greek *méze* (appetizer) with a bit of ouzo and traditional souvlaki and gyro, as well as many delightful Greek restaurants with international cuisine.



Don't forget to check out the neoclassical design of the *Monastiraki Metro Station*. The neighboring *Thision* and *Psiri* districts (take the metro to the *Thision* station) have expe-

rienced an explosion of fashionable restaurants and chic bars that have made the food and nightlife scene here one of the most popular in Athens today. The modern renditions of traditional Greek music tavernas with live bouzouki are here too. In the winter, cozy Athens restaurants offer finger-licking Greek mezedes (appetizers) and live music to entertain local Athenians young and old, day and night.

Kerameikos-Gazi: *Kerameikos* is another place where Athenians “in-the-know” discover multicultural hangouts ranging from rakadika or places where you enjoy the potent Cretan spirit, raki, with hot mezedes; *mageria* or places with homemade, oven-baked casseroles; as well as chic, modern, ethnic restaurants. In recent years, the downtown Gazi area has also developed into one of the trendiest Athenian neighborhoods. The area and its new inhabitants of artists have attracted fine restaurants.

Outside of Athens: The following municipalities are close enough to Athens to encourage a culinary visit.

- » *Piraeus:* A visit to the main port’s many waterfront tavernas for freshly caught fish of the day.
- » *Glyfada:* A coastal district easily accessible by tram, Glyfada is known to offer a splendid and concentrated shopping district, many outdoor clubs in the summer and Biftekoupoli, a “town” of tavernas with delectable meat-based menus.
- » *Kifissia:* Considered an Athenian suburb for the well-to-do, it offers numerous dining options as does the neighboring area, Kefalari.

Athens Museums

Acropolis Museum



Information can be found at *Social Events* page 16

Location: 2-4 Makriyianni Str

Tuesday to Sunday: 08:00 to 20:00

Monday: Closed.

The Museum is open every Friday until 22:00

<http://www.theacropolismuseum.gr/>

National Archeological Museum



The National Archaeological Museum is a must to visit. One of the richest museums concerning ancient Greek art in the world, its collections are representative of all the cultures that flourished in Greece: from the prehistoric age until the later age of Turkish dominance, including frescoes from prehistoric Thera and statues from the classical period, such as a bronze statue of Poseidon.

Location: 44 Patision Ave., Athens

Opening hours: Monday: 13:30-20:00

Tuesday-Sunday: 09:30-16:00

<http://www.namuseum.gr>

Museum of Cycladic Art



The Museum of Cycladic Art is dedicated to the study and promotion of ancient cultures of the Aegean and Cyprus, with special emphasis on Cycladic Art of the 3rd millennium BC.

It was founded in 1986, to house the collection of Nicholas and Dolly Goulandris. Since then it has grown in size to accommodate new acquisitions, obtained either through direct purchases or through donations by important

collectors and institutions.

Today, in the galleries of the MCA the visitor can approach three major subjects:

- » the Cycladic Culture of the Early Bronze Age (3200 - 2000 BC)
- » the Ancient Greek Art, from the Bronze Age to Late Roman times (2nd millenium BC - 4th century AD)
- » the Cypriot Culture from the Chalcolithic Age to the Early Christian period (4th millenium BC - 6th century AD)

Location: *4, Neophytou Douka Str.*

Monday - Wednesday - Friday - Saturday: 10:00 - 17:00

Thursday: 10:00 - 20:00

Sunday: 11:00 - 17:00

Tuesday: closed

<http://www.cycladic.gr>

Benaki Museum



This group of collections comprises many distinct categories totaling more than 30.000 items illustrating the character of the Greek world through a spectacular historical panorama: from antiquity and the age of Roman domination to the medieval Byzantine period; from the fall of Constantinople in 1453 and the centuries of Frankish and Ottoman occupation; to the outbreak of the struggle for independence in 1821; from the formation of the modern state of Greece in 1830, up to 1922, the year in which the Asia Minor disaster took place

Location: *Koumpari Str. & Vasilissis Sofias Ave.*

Monday, Wednesday: 9:00 - 17:00

Thursday: 9:00 - 24:00

Friday: 12:00 - 17:00

Saturday: 9:00 - 15:00

<http://www.benaki.gr>

National Museum of Contemporary Art



The National Museum of Contemporary Art, Athens, (EMST) was founded in 1997. Among its basic aims are: the creation of collections of works of contemporary Hellenic and international art, the promotion and presentation of advanced and experimental artistic tendencies, the enhancement of the aesthetic and artistic cultivation of the audience and the development of scientific research on subjects of contemporary art history and theory.

Location: 17-19 Vas. Georgiou B' and Rigillis Str.

Tuesday to Sunday: 11:00 - 19:00

Thursday: 11:00 - 22:00

Monday: Closed

<http://www.emst.gr>

Byzantine Museum



The Byzantine and Christian Museum of Athens is one of the most important public institutions in Greece, established in the early 20th century (1914) in order to collect, study, preserve and exhibit the Byzantine and Post-Byzantine cultural heritage in the Hellenic territory.

The museum collection contains an important number (approximately 30,000) of works of art such as icons, sculptures, ceramics, ecclesiastical textiles, paintings, jewelries and architectural elements (wall paintings and mosaics).

The permanent exhibition is divided in two main parts:

The first is devoted to Byzantium (4th -15th c. AD) and contains 1200 artifacts and the second entitled "From Byzantium to the modern era" presents 1500 artworks dating from the 15th to 20th century.

Location: 22 Vasilissis Sofias Ave.

Tuesday-Saturday: 8:00 - 20:00, Sunday: 08:00 - 20:00

<http://www.byzantinemuseum.gr>

War Museum



In 1964, the Hellenic State decided to found the War Museum, wishing to honor all those who fought for our country and its freedom. The design of the museum was undertaken by a team of distinguished scientists, headed by Professor Thoukidides Valentis of the National Technical University of Athens (N.T.U.A). On July 18, 1975, the President of the Hellenic Republic H.E. Constantine Tsatsos and the Minister of National Defense Evangelos Averof-Tositsas inaugurated the Museum. Its various activities include the publication of books, the establishment and maintenance of monuments and memorials and the aid to services and agencies all over Greece. The Museum's exhibition areas are distributed over four levels (floors) and present images of Greek history from antiquity to the present.

Location: 2 Rizari Str., Athens

Tuesday to Friday: 09:00 - 14:00

Sundays: 09:30 - 14:00

Closed: on Mondays.

<http://www.warmuseum.gr>

National Gallery

The National Art Gallery is one of Greece's main art institutions and features paintings and works of art from some of Greece's and Europe's best from the 19th and 20th centuries. Emphasis is given to popular Greek contemporary artists including Giannis Tsarouchis,

Domenikos Theotokopoulos (a.k.a. El Greco), Theodors Vrizakis, Nikolaos Kounelakis, Nikiforos Litras, Konstantinos Parthenis, Maleas, Giannis Moralis and others.

Location: *1 Michalakopoulou Str.*

Monday, Thursday, Friday, Saturday: 09:00 - 15:00

Tuesday: Closed

Wednesday: 15:00 - 21:00

Sunday: 10:00 - 14:00

<http://www.nationalgallery.gr>



Archeological Museum – Ancient Market



The museum of Ancient Market is accommodated in the restored Gallery [Attalou], monument of the 150 B.C. His exhibits have direct relation with the operation of democratic regime of Ancient Athens, since the Market was the heart of public life. The gallery (Attalou) was revealed in the excavations of the Archaeological Company between the years 1859-1902. At the period 1953-56 it was restored and it has been reconstructed in order to accommodate

the discoveries of excavations of Market. In 1957 the Greek State undertook the administration and the safety of the Museum and the archaeological space.

Location: *24 Andrianou Str., Athens*

Summer season from 08:00 to 19:30

<http://www.breathtakingathens.com>

24. Transportation

The Athens Mass Transit System consists of a large bus fleet, a trolleybus fleet that mainly serves the downtown area (city center), the city's Metro, a tram line connecting the southern suburbs to the city centre, and the Athens commuter rail service.

Bus: The urban buses that are under the authority of the OASA (organisation of urban transport of Athens) connect all Municipalities of Athens and Piraeus. The Athens metropolitan area is also served by trolleybuses — or electric buses.

Athens Tram: The tram network covers ten suburbs of Athens. This network runs from Syntagma Square to the southwestern suburb of Palaio Faliro, where the line splits in two branches; the first runs along the Athens coastline toward the southern suburb of Voula, while the other heads toward the Piraeus district of Neo Faliro.

Attiko Metro: The Athens Metro is more commonly known in Greece as the Attiko Metro joins together with the Athens-Piraeus Electric rail and the Proastiakos suburban and it has 2 lines. **The Red Line** (line 2) runs from Aghios Antonios to Aghios Dimitrios and covers a distance of 10.9 km. **The Blue Line** (line 3) runs from the western suburbs, namely the Egaleo station, through the central Monastiraki and Syntagma stations to Doukissis Plakentias avenue in the northeastern suburb of Halandri, covering a distance of 16 km, then

ascending to ground level and reaching Eleftherios Venizelos International Airport, using the Suburban Railway infrastructure.

Electric railway (ISAP): The *Electric Railway Company* (ISAP) line, which for many years served as Athens' primary urban rail transport. This is today the **Green Line** (line 1) of the Athens Metro network and unlike the red and blue routes running entirely underground, ISAP runs either above-ground or below-ground at different sections of its journey. It forms the oldest line of the Athens metro network connecting the port of Piraeus with the northern suburb of Kifissia.

Commuter/Suburban rail (Proastiakos): The Athens commuter rail service, referred to as the "Proastiakós", connects Eleftherios Venizelos International Airport to the city of Korinthos, west of Athens, via Larissa station, the city's central rail station and the port of Piraeus.

24.1 Ticket prices

Integrated Ticket: 1,40€ Reduced: 0.70€

This is valid for multiple trips on all public transport options, in urban zone (buses, trolleys, tram, metro, suburban railway), in any direction for up to 90 minutes. This ticket is not valid on express bus lines to the airport and on Varkiza-Saronida section of route E22.

One Way Ticket: 1.20€ Reduced: 0.60€

This is valid only on buses and trolleys for only one trip.

Points of sale:

- » at metro, tram, suburban railway stations
- » at blue/yellow ticket offices
- » at many newsstands.

Airport Tickets

Airport Express Bus Lines Ticket: 5€ Reduced: 2.50 €

Points of sale:

- » from the drivers of the airport buses
- » at all metro and blue ticket offices

The above tickets can be used for only one trip from or to the airport.

Metro and Suburban Railway

Metro and Suburban Railway Ticket: 8€ Reduced: 4€

Return Ticket (within 48 hours): 14€

Group Ticket for 2: 14€

Group Ticket for 3: 20€

Points of sale: at all metro and suburban railway stations.

The above tickets are valid for travel with all public transport means within 90 minutes.

Transportation to conference events

- » Monday's welcome reception will be held at Ioannis Restaurant in the roof garden of Royal Olympic, the main conference venue. The closest metro station is "Acropolis"

(red line).

- » Tuesday's poster session will be held in "Technopolis" which is located approximately 500 meters away from metro station "*Keramikos*" (blue line).
- » Wednesday's banquet will be held at Ioannis Restaurant in the roof garden of Royal Olympic, the main conference venue. The closest metro station is "*Acropolis*" (red line).
- » Thursday's farewell part will take place at the "*Acropolis Museum*". The closest metro station is "*Acropolis*" (red line).

25. Conference Organization

Organizing Institutions	Athens University of Economics and Business National & Kapodistrian University of Athens Google Inc., Zurich, Switzerland Dipartimento di Informatica, Bari Università degli Studi di Bari “Aldo Moro” University of Ioannina University of Piraeus
General co-Chairs	Aristidis Likas, <i>Department of Computer Science University of Ioannina, Greece</i> Yannis Theodoridis, <i>Department of Informatics, University of Piraeus, Greece</i>
Programme Committee Chairs	Thomas Hofmann, <i>Google Inc., Zurich, Switzerland</i> Donato Malerba, <i>Department of Computer Science, University of Bari, Italy</i> Dimitrios Gunopulos, <i>Department of Informatics and Telecommunications, University of Athens, Greece</i> Michalis Vazirgiannis, <i>Department of Informatics, Athens University of Economics & Business, Greece</i>
Workshop Chairs	Katharina Morik, <i>University of Dortmund, Germany</i> Bart Goethals, <i>Department of Mathematics and Computer Science, University of Antwerp, Belgium</i>
Tutorial Chairs	Fosca Giannotti, <i>Knowledge Discovery and Delivery (KDD) Lab, ISTI-CNR, Italy</i> Maguelonne Teisseire, <i>TETIS Lab. Departement of Information System & LIRMM Lab. Department of Computer Science, France</i>
Best Paper Award Chairs	Sunita Sarawagi, <i>Computer Science and Engineering, IIT Bombay, India</i> Michèle Sebag, <i>Laboratoire de Recherche en Informatique, CNRS, University of Paris-Sud, France</i>
Industrial Session Chairs	Alexandros Ntoulas, <i>Microsoft Research, USA</i> Michail Vlachos, <i>IBM Zurich Research Laboratory, Switzerland</i>
Demo Track Chairs	Michelangelo Ceci, <i>University of Bari “Aldo Moro”, Italy</i> Spiros Papadimitriou, <i>Google Research</i>

Discovery Challenge Chairs Alexandros Kalousis, *Artificial Intelligence Laboratory, Department of Computer Science, University of Geneva, Switzerland*
Vassilis Plachouras, *Athens University of Economics and Business, Greece*

Publicity Chairs Annalisa Appice, *Dipartimento di Informatica, University of Bari "Aldo Moro", Italy*
Grigorios Tsoumakas, *Department of Informatics, Aristotle University of Thessaloniki, Greece*

Sponsorship Chairs Ina Lauth, *IAIS Fraunhofer, Germany*
Ioannis Kopanakis, *Technological Educational Institute of Crete, Greece*

Organization Committee Maria Halkidi, *Department of Digital Systems, University of Piraeus, Greece*
Despina Kopanaki, *Department of Informatics, University of Piraeus, Greece*
Nikos Pelekis, *Department of Statistics and Insurance Science, University of Piraeus, Greece*

Steering Committee

José Balcázar
Francesco Bonchi
Wray Buntine
Walter Daelemans
Aristides Gionis
Bart Goethals
Katharina Morik
Dunja Mladenic
John Shawe-Taylor
Michèle Sebag

George Karypis
Ravi Kumar
James Kwok
Stan Matwin
Michael May
Taneli Mielikainen
Yücel Saygin
Arno Siebes
Jian Pei
Myra Spiliopoulou
Jie Tang
Evimaria Terzi
Bhavani M. Thuraisingham
Hannu Toivonen
Luis Torgo
Ioannis Tsamardinos
Panayiotis Tsaparas
Ioannis P. Vlahavas
Haixun Wang
Stefan Wrobel
Xindong Wu

Area Chairs

Elena Baralis
Hendrik Blockeel
Francesco Bonchi
Gautam Das
Janez Demsar
Amol Deshpande
Carlotta Domeniconi
Tapio Elomaa
Floriana Esposito
Fazel Famili
Wei Fan
Peter Flach
Johannes Furnkranz
Aristides Gionis

Finance & Registration

Triaena Tours & Congress S.A.
(Greece)

26. Program Committee

Foto Afrati
Aijun An
Aris Anagnostopoulos
Gennady Andrienko
Ion Androustopoulos
Annalisa Appice
Marta Arias
Ira Assent
Vassilis Athitsos
Martin Atzmueller
Jose Luis Balcazar
Daniel Barbara
Sugato Basu
Roberto Bayardo
Klaus Berberich
Bettina Berendt
Michele Berlingerio
Michael Berthold
Indrajit Bhattacharya
Marenglen Biba
Albert Bifet
Enrico Blanzieri
Konstantinos Blekas
Mario Boley
Zoran Bosnic
Marco Botta
Jean-Francois Boulicaut
Pavel Bradzil
Ulf Brefeld
Paula Brito
Wray Buntine
Toon Calders
Rui Camacho
Longbing Cao
Michelangelo Ceci
Tania Cerquitelli
Sharma Chakravarthy
Keith Chan
Vineet Chaoji
Keke Chen
Ling Chen
Xue-wen Chen
Weiwei Cheng
Yun Chi
Silvia Chiusano
Vassilis Christophides
Frans Coenen
James Cussens
Alfredo Cuzzocrea
Maria Damiani
Atish Das Sarma
Tijl De Bie
Jeroen De Knijf
Colin de la Higuera
Antonios Deligiannakis
Krzysztof Dembczynski
Anne Denton
Christian Desrosiers
Wei Ding
Ying Ding
Debora Donato
Kurt Driessens
Chris Drummond
Pierre Dupont
Saso Dzeroski
Tina Eliassi-Rad
Roberto Eposito
Nicola Fanizzi
Fabio Fassetti
Ad Feelders
Hakan Ferhatosmanoglou
Stefano Ferilli
Cesar Ferri
Daan Fierens
Eibe Frank
Enrique Frias-Martinez
Elisa Fromont
Efstratios Gallopoulos
Byron Gao
Jing Gao
Paolo Garza
Ricard Gavaldà
Floris Geerts
Pierre Geurts
Aris Gkoulalas-Divanis
Bart Goethals
Vivekanand Gopalkrishnan
Marco Gori
Henrik Grosskreutz
Maxim Gurevich

Maria Halkidi
Mohammad Hasan
Jose Hernandez-Orallo
Eyke Hüllermeier
Vasant Honavar
Andreas Hotho
Xiaohua Hu
Ming Hua
Minlie Huang
Marcus Hutter
Nathalie Japkowicz
Szymon Jaroszewicz
Daxin Jiang
Alipio Jorge
Theodore Kalamboukis
Alexandros Kalousis
Panagiotis Karras
Samuel Kaski
Ioannis Katakis
John Keane
Kristian Kersting
Latifur Khan
Joost Kok
Christian König
Irena Koprinska
Walter Kusters
Georgia Koutrika
Stefan Kramer
Raghuram Krishnapuram
Marzena Kryszkiewicz
Nicolas Lachiche
Nada L'Avrac
Wang-Chien Lee
Feifei Li
Jiuyong Li
Juanzi Li
Tao Li
Chih-Jen Lin
Hsuan-Tien Lin
Jessica Lin
Shou-de Lin
Song Lin
Helger Lipmaa
Bing Liu
Huan Liu
Yan Liu
Corrado Loglisci

Chang-Tien Lu
Ping Luo
Panagis Magdalinos
Giuseppe Manco
Yannis Manolopoulos
Simone Marinai
Dimitrios Mavroeidis
Ernestina Menasalvas
Rosa Meo
Pauli Miettinen
Dunja Mladenic
Marie-Francine Moens
Katharina Morik
Mirco Nanni
Alexandros Nanopoulos
Benjamin Nguyen
Frank Nielsen
Siegfried Nijssen
Richard Nock
Kjetil Norvag
Irene Ntoutsis
Salvatore Orlando
Gerhard Paass
George Paliouras
Apostolos Papadopoulos
Panagiotis Papapetrou
Stelios Pappas
Dimitris Pappas
Ioannis Partalas
Srinivasan Parthasarathy
Andrea Passerini
Vladimir Pavlovic
Dino Pedreschi
Nikos Pelekis
Jing Peng
Ruggero Pensa
Bernhard Pfahringer
Fabio Pinelli
Enric Plaza
George Potamias
Michalis Potamias
Doina Precup
Kunal Punera
Chedy Raissi
Jan Ramon
Huzefa Rangwala
Zbigniew Ras

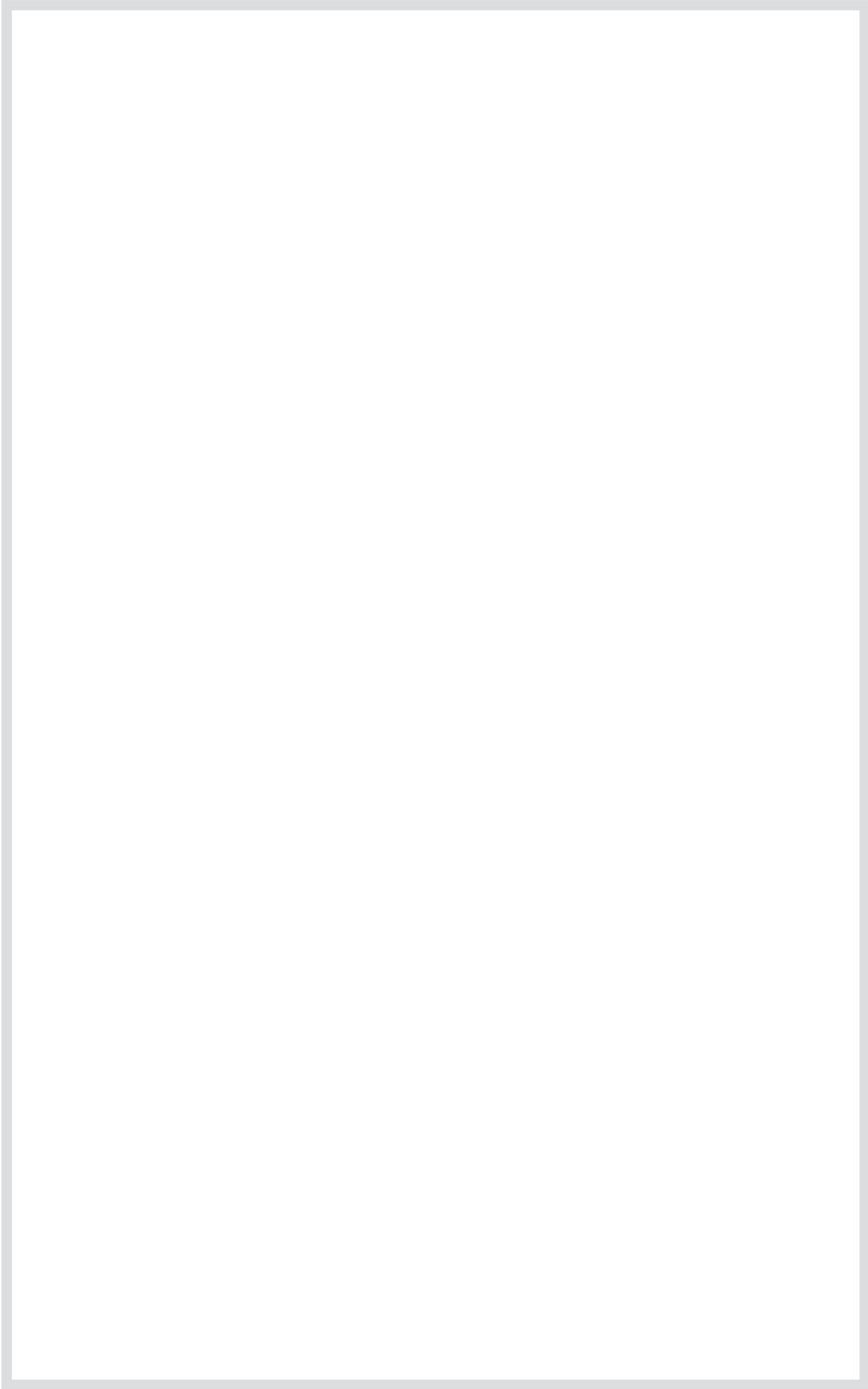
Ann Ratanamahatana
Jan Rauch
Matthias Renz
Christophe Rigotti
Fabrizio Riguzzi
Celine Robardet
Marko Robnik-Sikonja
Pedro Rodrigues
Fabrice Rossi
Juho Rousu
Celine Rouveirol
Ulrich Ruckert
Salvatore Ruggieri
Stefan Ruping
Lorenza Saitta
Ansaf Salleb-Aouissi
Claudio Sartori
Lars Schmidt-Thieme
Matthias Schubert
Michèle Sebad
Thomas Seidl
Prithviraj Sen
Andrzej Skowron
Carlos Soares
Yangqiu Song
Alessandro Sperduti
Jerzy Stefanowski
Jean-Marc Steyaert
Alberto Suarez
Johan Suykens
Einoshin Suzuki
Panagiotis Symeonidis
Marcin Szczuka
Andrea Tagarelli
Domenico Talia
Pang-Ning Tan
Letizia Tanca
Lei Tang
Dacheng Tao
Nikolaj Tatti
Martin Theobald
Dimitrios Thilikos
Jilei Tian
Ivor Tsang
Grigorios Tsoumakas
Theodoros Tzouramanis
Antti Ukkonen

Takeaki Uno
Athina Vakali
Giorgio Valentini
Maarten van Someren
Iraklis Varlamis
Julien Velcin
Celine Vens
Jean-Philippe Vert
Vassilios Verykios
Herna Viktor
Jilles Vreeken
Willem Waegeman
Jianyong Wang
Wei Wang
Xuanhui Wang
Hui Xiong
Jieping Ye
Jeffrey Yu
Philip Yu
Bianca Zadrozny
Gerson Zaverucha
Demetris Zeinalipour
Filip Zelezny
Changshui Zhang
Kai Zhang
Kun Zhang
Min-Ling Zhang
Nan Zhang
Shichao Zhang
Zhongfei Zhang
Junping Zhang
Ying Zhao
Bin Zhou
Zhi-Hua Zhou
Kenny Zhu
Xingquan Zhu
Djamel Zighed
Indre Zliobaite
Blaz Zupan

27. My notes

A large, empty rectangular box with a thin gray border, occupying most of the page. It is intended for the user to write their notes.











ATTIKO METRO
ATHENS METRO



PLATINUM SPONSOR



GOLD SPONSOR



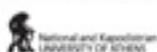
SILVER SPONSORS



BRONZE SPONSORS



ORGANIZING INSTITUTIONS



ADDITIONAL SUPPORTERS

