This year’s International Conference on Discovery Science (DS) was the 17th event in this series. Like in previous years, the conference was co-located with the International Conference on Algorithmic Learning Theory (ALT), which is already in its 25th year. Starting in 2001, ALT/DS is one of the longest-running series of co-located events in computer science. The unique combination of recent advances in the development and analysis of methods for discovering scientific knowledge, coming from machine learning, data mining, and intelligent data analysis, as well as their application in various scientific domains, on the one hand, with the algorithmic advances in machine learning theory, on the other hand, makes every instance of this joint event unique and attractive.

This volume contains the papers presented at the 17th International Conference on Discovery Science, while the papers of the 25th International Conference on Algorithmic Learning Theory are published by Springer in a companion volume (LNCS Vol. 8776). We had the pleasure of selecting contributions from 62 submissions by 178 authors from 22 countries. Each submission was reviewed by at least three Program Committee members. The program chairs eventually decided to accept 30 papers, yielding an acceptance rate of 48%. The program also included three invited talks and two tutorials. In the joint DS/ALT invited talk, Zoubin Ghahramani gave a presentation on “Building an Automated Statistician.” DS participants also had the opportunity to attend the ALT invited talk on “Cellular Tree Classifiers”, which was given by Luc Devroye. The two tutorial speakers were Anuška Ferligoj (“Social Network Analysis”) and Eyke Hüllermeier (“Online Preference Learning and Ranking”).

This year, both conferences were held in Bled, Slovenia, and were organized by the Jožef Stefan Institute (JSI) and the University of Ljubljana. We are very grateful to the Department of Knowledge Technologies (and the project MAESTRA) at JSI for sponsoring the conferences and providing administrative support. In particular, we thank the local arrangement chair, Mili Bauer, and her team, Tina Anžič, Nikola Simidjievski, and Jurica Levatič from JSI for their efforts in organizing the two conferences. We would like to thank the Office of Naval Research Global for the generous financial support provided under ONRG GRANT N62909-14-1-C195.

We would also like to thank all authors of submitted papers, the Program Committee members, and the additional reviewers for their efforts in evaluating the submitted papers, as well as the invited speakers and tutorial presenters. We are grateful to Sandra Zilles, Peter Auer, Alexander Clark and Thomas Zeugmann for ensuring a smooth coordination with ALT, Nikola Simidjievski for putting up and maintaining our website, and Andrei Voronkov for making
EasyChair freely available. Finally, special thanks go to the Discovery Science Steering Committee, in particular to its past and current chairs, Einoshin Suzuki and Akihiro Yamamoto, for entrusting us with the organization of the scientific program of this prestigious conference.

July 2014

Sašo Džeroski
Panče Panov
Dragi Kocev
Ljupčo Todorovski
Organization

ALT/DS General Chair

Ljupčo Todorovski  University of Ljubljana, Slovenia

Program Chair

Sašo Džeroski  Jožef Stefan Institute, Slovenia

Program Co-chairs

Panče Panov  Jožef Stefan Institute, Slovenia
Dragi Kocev  Jožef Stefan Institute, Slovenia

Local Organization Team

Mili Bauer  Jožef Stefan Institute, Slovenia
Tina Anžič  Jožef Stefan Institute, Slovenia
Nikola Simidjievski  Jožef Stefan Institute, Slovenia
Jurica Levatić  Jožef Stefan Institute, Slovenia

Program Committee

Albert Bifet  HUAWEI Noahs Ark Lab, Hong Kong and University of Waikato, New Zealand
Hendrik Blockeel  KU Leuven, Belgium
Ivan Bratko  University of Ljubljana, Slovenia
Michelangelo Ceci  Università degli Studi di Bari Aldo Moro, Italy
Simon Colton  Goldsmiths College, University of London, UK
Bruno Cremilleux  Université de Caen, France
Luc De Raedt  KU Leuven, Belgium
Ivica Dimitrovski  Ss. Cyril and Methodius University, Macedonia
Tapio Elomaa  Tampere University of Technology, Finland
Bogdan Filipič  Jožef Stefan Institute, Slovenia
Peter Flach  University of Bristol, UK
Johannes Fürnkranz  TU Darmstadt, Germany
Mohamed Gaber  University of Portsmouth, UK
João Gama  University of Porto and INESC Porto, Portugal
Dragan Gamberger  Rudjer Boskovic Institute, Croatia
Yolanda Gil University of Southern California, Information Sciences Institute, Marina del Rey, CA, USA
Makoto Haraguchi Hokkaido University, Japan
Kouichi Hirata Kyushu Institute of Technology, Japan
Jaakko Hollmén Aalto University, Finland
Geoffrey Holmes University of Waikato, New Zealand
Alipio Jorge University of Porto and INESC Porto, Portugal
Kristian Kersting Technical University of Dortmund and Fraunhofer IAIS, Bonn, Germany
Masahiro Kimura Ryukoku University, Japan
Ross King University of Manchester, UK
Stefan Kramer Johannes Gutenberg University Mainz, Germany
Nada Lavrač Jožef Stefan Institute, Slovenia
Philippe Lenca Telecom Bretagne, France
Gjorgji Madjarov Ss. Cyril and Methodius University, Macedonia
Donato Malerba Università degli Studi di Bari Aldo Moro, Italy
Dunja Mladenic Jožef Stefan Institute, Slovenia
Stephen Muggleton Imperial College, London, UK
Zoran Obradovic Temple University, Philadelphia, PA, USA
Bernhard Pfahringer University of Waikato, New Zealand
Marko Robnik-Sikonja University of Ljubljana, Slovenia
Juho Rousu Aalto University, Finland
Kazumi Saito University of Shizuoka, Japan
Ivica Slavkov Centre for Genomic Regulation, Barcelona, Spain
Tomislav Šmuc Rudjer Bosković Institute, Croatia
Larisa Soldatova Brunel University London, UK
Einoshin Suzuki Kyushu University, Japan
Maguelonne Teisseire Cemagref - UMR Tetis, Montpellier, France
Grigorios Tsoumakas Aristotle University, Thessaloniki, Greece
Takashi Washio ISIR, Osaka University, Japan
Min-Ling Zhang Southeast University, Nanjing, China
Zhi-Hua Zhou Nanjing University, China
Indrė Žliobaitė Aalto University, Finland
Blaž Zupan University of Ljubljana, Slovenia

Additional Reviewers

Mariam Adedoyin-Olowe João Gomes
Reem Al-Otaibi Chao Han
Darko Aleksovski Sheng-Jun Huang
Jérôme Azé Inaki Inza
Behrouz Bababki Tetsuji Kuboyama
Janez Brank Meelis Kull
Vladimir Kuzmanovski
Fabiana Pasqua Lanotte
Carlos Martin-Dancausa
Louise Millard
Alexandra Moraru
Benjamin Negrevergne
Inna Novalija
Hai Phan
Gianvito Pio
Marc Plantevit
Pascal Poncelet
John Puentes
Hani Ragab-Hassen
Dusan Ramljak

Shoumik Roychoudhury
Hiroshi Sakamoto
Rui Sarmento
Francesco Serafino
Nikola Simidjievski
Jasmina Smailović
Arnaud Soulet
Jovan Tanevski
Aneta Trajanov
Niall Twomey
Alexey Uversky
João Vinagre
Bernard Ženko
Marinka Žitnik
Invited Talks
(Abstracts)
Abstract. We live in an era of abundant data and there is an increasing need for methods to automate data analysis and statistics. I will describe the “Automated Statistician”, a project which aims to automate the exploratory analysis and modelling of data. Our approach starts by defining a large space of related probabilistic models via a grammar over models, and then uses Bayesian marginal likelihood computations to search over this space for one or a few good models of the data. The aim is to find models which have both good predictive performance, and are somewhat interpretable. Our initial work has focused on the learning of unknown nonparametric regression functions, and on learning models of time series data, both using Gaussian processes. Once a good model has been found, the Automated Statistician generates a natural language summary of the analysis, producing a 10-15 page report with plots and tables describing the analysis. I will discuss challenges such as: how to trade off predictive performance and interpretability, how to translate complex statistical concepts into natural language text that is understandable by a numerate non-statistician, and how to integrate model checking. This is joint work with James Lloyd and David Duvenaud (Cambridge) and Roger Grosse and Josh Tenenbaum (MIT).
Social Network Analysis

Anuška Ferligoj
Faculty of Social Sciences,
University of Ljubljana
anuska.ferligoj@fdv.uni-lj.si

Abstract. Social network analysis has attracted considerable interest from the social and behavioral science communities in recent decades. Much of this interest can be attributed to the focus of social network analysis on relationship among units, and on the patterns of these relationships. Social network analysis is a rapidly expanding and changing field with a broad range of approaches, methods, models and substantive applications. In the talk, special attention will be given to:
1. General introduction to social network analysis:
   – What are social networks?
   – Data collection issues.
   – Basic network concepts: network representation; types of networks; size and density.
   – Walks and paths in networks: length and value of path; the shortest path, k-neighbours; acyclic networks.
   – Connectivity: weakly, strongly and bi-connected components; contraction; extraction.
2. Overview of tasks and corresponding methods:
   – Network/node properties: centrality (degree, closeness, betweenness); hubs and authorities.
   – Cohesion: triads, cliques, cores, islands.
   – Partitioning: blockmodeling (direct and indirect approaches; structural, regular equivalence; generalised blockmodeling); clustering.
   – Statistical models.
3. Software for social network analysis (UCINET, PAJEK, …)

References

Cellular Tree Classifiers

Gérard Biau and Luc Devroye

1 Sorbonne Universités, UPMC Univ Paris 06, France  
2 Institut universitaire de France  
3 McGill University, Canada

Abstract. Suppose that binary classification is done by a tree method in which the leaves of a tree correspond to a partition of d-space. Within a partition, a majority vote is used. Suppose furthermore that this tree must be constructed recursively by implementing just two functions, so that the construction can be carried out in parallel by using “cells”: first of all, given input data, a cell must decide whether it will become a leaf or an internal node in the tree. Secondly, if it decides on an internal node, it must decide how to partition the space linearly. Data are then split into two parts and sent downstream to two new independent cells. We discuss the design and properties of such classifiers.

Online Preference Learning and Ranking*

Eyke Hüllermeier

Department of Computer Science
University of Paderborn, Germany
eyke@upb.de

Abstract. A primary goal of this tutorial is to survey the field of preference learning [7], which has recently emerged as a new branch of machine learning, in its current stage of development. Starting with a systematic overview of different types of preference learning problems, methods to tackle these problems, and metrics for evaluating the performance of preference models induced from data, the presentation will focus on theoretical and algorithmic aspects of ranking problems [6, 8, 10]. In particular, recent approaches to preference-based online learning with bandit algorithms will be covered in some depth [12, 13, 11, 2, 9, 1, 3–5, 14].

References


* The full version of this paper can be found in Peter Auer, Alexander Clark, Sandra Zilles, and Thomas Zeugmann, Proceedings of the 25th International Conference on Algorithmic Learning Theory (ALT-14), Lecture Notes in Computer Science Vol. 8776, Springer, 2014.
**Table of Contents**

Explaining Mixture Models through Semantic Pattern Mining and Banded Matrix Visualization .................................................. 1  
Prem Raj Adhikari, Anže Vavpetič, Jan Kralj, Nada Lavrač, and Jaakko Hollmén

Big Data Analysis of StockTwits to Predict Sentiments in the Stock Market ........................................................................... 13  
Alya Al Nasseri, Allan Tucker, and Sergio de Cesare

Synthetic Sequence Generator for Recommender Systems – Memory Biased Random Walk on a Sequence Multilayer Network .......... 25  
Nino Antulov-Fantulin, Matko Bošnjak, Vinko Zlatić, Miha Grčar, and Tomislav Šmuc

Predicting Sepsis Severity from Limited Temporal Observations ....... 37  
Xi Hang Cao, Ivan Stojkovic, and Zoran Obradovic

Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining ................................................. 49  
Michelangelo Ceci, Pasqua Fabiana Lanotte, Fabio Fumarola, Dario Pietro Cavallo, and Donato Malerba

Antipattern Discovery in Ethiopian Bagana Songs......................... 62  
Darrell Conklin and Stéphanie Weisser

Categorize, Cluster, and Classify: A 3-C Strategy for Scientific Discovery in the Medical Informatics Platform of the Human Brain Project ....................................................................................... 73  
Tal Galili, Alexis Mitelpunkt, Netta Shachar, Mira Marcus-Kalish, and Yoav Benjamini

Multilayer Clustering: A Discovery Experiment on Country Level Trading Data ................................................................. 87  
Dragan Gamberger, Matej Mihelčič, and Nada Lavrač

Medical Document Mining Combining Image Exploration and Text Characterization ................................................................. 99  
Nicolau Gonçalves, Erkki Oja, and Ricardo Vigário

Mining Cohesive Itemsets in Graphs ............................................. 111  
Tayena Hendrickx, Boris Cule, and Bart Goethals
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining Rank Data</td>
<td>123</td>
</tr>
<tr>
<td>Sascha Henzgen and Eyke Hüllermeier</td>
<td></td>
</tr>
<tr>
<td>Link Prediction on the Semantic MEDLINE Network: An Approach to</td>
<td>135</td>
</tr>
<tr>
<td>Literature-Based Discovery</td>
<td></td>
</tr>
<tr>
<td>Andrej Kastrin, Thomas C. Rindflesch, and Dimitar Hristovski</td>
<td></td>
</tr>
<tr>
<td>Medical Image Retrieval Using Multimodal Data</td>
<td>144</td>
</tr>
<tr>
<td>Ivan Kitanovski, Ivica Dimitrovski, Gjorgji Madjarov, and Suzana Loskovska</td>
<td></td>
</tr>
<tr>
<td>Fast Computation of the Tree Edit Distance between Unordered Trees</td>
<td>156</td>
</tr>
<tr>
<td>Using IP Solvers</td>
<td></td>
</tr>
<tr>
<td>Seiichi Kondo, Keisuke Otaki, Madori Ikeda, and Akihiro Yamamoto</td>
<td></td>
</tr>
<tr>
<td>Probabilistic Active Learning: Towards Combining Versatility, Optimal</td>
<td>168</td>
</tr>
<tr>
<td>ity and Efficiency</td>
<td></td>
</tr>
<tr>
<td>Georg Krempl, Daniel Kottke, and Myra Spiliopoulou</td>
<td></td>
</tr>
<tr>
<td>Incremental Learning with Social Media Data to Predict Near Reality</td>
<td>180</td>
</tr>
<tr>
<td>Duc Kinh Le Tran, Cécile Bothorel, Pascal Cheung Mon Chan, and Yvon Kermarrec</td>
<td></td>
</tr>
<tr>
<td>Stacking Label Features for Learning Multilabel Rules</td>
<td>192</td>
</tr>
<tr>
<td>Eneldo Loza Mencía and Frederik Janssen</td>
<td></td>
</tr>
<tr>
<td>Selective Forgetting for Incremental Matrix Factorization in</td>
<td>204</td>
</tr>
<tr>
<td>Recommender Systems</td>
<td></td>
</tr>
<tr>
<td>Pawel Matuszyk and Myra Spiliopoulou</td>
<td></td>
</tr>
<tr>
<td>Providing Concise Database Covers Instantly by Recursive Tile</td>
<td>216</td>
</tr>
<tr>
<td>Sampling</td>
<td></td>
</tr>
<tr>
<td>Sandy Moens, Mario Boley, and Bart Goethals</td>
<td></td>
</tr>
<tr>
<td>Resampling-Based Framework for Estimating Node Centrality of Large</td>
<td>228</td>
</tr>
<tr>
<td>Social Network</td>
<td></td>
</tr>
<tr>
<td>Kouzou Ohara, Kazumi Saito, Masahiro Kimura, and Hiroshi Motoda</td>
<td></td>
</tr>
<tr>
<td>Detecting Maximum $k$-Plex with Iterative Proper $l$-Plex Search</td>
<td>240</td>
</tr>
<tr>
<td>Yoshiaki Okubo, Masanobu Matsudaira, and Makoto Haraguchi</td>
<td></td>
</tr>
<tr>
<td>Exploiting Bhattacharyya Similarity Measure to Diminish User Cold-</td>
<td>252</td>
</tr>
<tr>
<td>Start Problem in Sparse Data</td>
<td></td>
</tr>
<tr>
<td>Bidyut Kr. Patra, Raimo Launonen, Ville Ollikainen, and Sukumar Nandi</td>
<td></td>
</tr>
<tr>
<td>Failure Prediction – An Application in the Railway Industry</td>
<td>264</td>
</tr>
<tr>
<td>Pedro Pereira, Rita P. Ribeiro, and João Gama</td>
<td></td>
</tr>
</tbody>
</table>
Wind Power Forecasting Using Time Series Cluster Analysis ............... 276
Sonja Pravilovic, Annalisa Appice, Antonietta Lanza, and Donato Malerba

Feature Selection in Hierarchical Feature Spaces ............................. 288
Petar Ristoski and Heiko Paulheim

Incorporating Regime Metrics into Latent Variable Dynamic Models to Detect Early-Warning Signals of Functional Changes in Fisheries Ecology ................................................................. 301
Neda Trifonova, Daniel Duplisea, Andrew Kenny, David Maxwell, and Allan Tucker

An Efficient Algorithm for Enumerating Chordless Cycles and Chordless Paths ................................................................. 313
Takeaki Uno and Hiroko Satoh

Algorithm Selection on Data Streams ............................................. 325
Jan N. van Rijn, Geoffrey Holmes, Bernhard Pfahringer, and Joaquin Vanschoren

Sparse Coding for Key Node Selection over Networks ...................... 337
Ye Xu and Dan Rockmore

Variational Dependent Multi-output Gaussian Process Dynamical Systems ................................................................. 350
Jing Zhao and Shiliang Sun

Erratum

Categorize, Cluster, and Classify: A 3-C Strategy for Scientific Discovery in the Medical Informatics Platform of the Human Brain Project ................................................................. E1
Tal Galili, Alexis Mitelpunkt, Netta Shachar, Mira Marcus-Kalish, and Yoav Benjamini

Author Index .............................................................................. 363