2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016)

Montreal, Quebec, Canada
17-19 October 2016
Organizing Committee
Program Committee
Reviewers

Main Conference Papers

Classification
On the Evaluation of Outlier Detection and One-Class Classification Methods
Lorne Swersky, Henrique O. Marques, Jöerg Sander, Ricardo J.G.B. Campello,
and Arthur Zimek

Active Semi-Supervised Classification Based on Multiple Clustering Hierarchies
Antônio J.L. Batista, Ricardo J.G.B. Campello, and Jöerg Sander

Combining Static and Dynamic Features for Multivariate Sequence Classification
Anna Leontjeva and Ilya Kuzovkin

Correcting Relational Bias to Improve Classification in Sparsely-Labeled Networks
Joshua R. King and Luke K. McDowell

Hyperparameter Optimization Machines
Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme

Networks
Temporal Network Change Detection Using Network Centralities
Yoshitaro Yonamoto, Kai Morino, and Kenji Yamanishi

Harvester: Influence Optimization in Symmetric Interaction Networks
Sergei Ivanov and Panagiotis Karras
Pattern Matching Trajectories for Investigative Graph Searches ........................................71
  Benjamin W.K. Hung, Anura P. Jayasumana, and Vidarshana W. Bandara

A Framework for Description and Analysis of Sampling-Based Approximate
Triangle Counting Algorithms ........................................................................................................80
  Mostafa Haghir Chehreghani

Limiting the Diffusion of Information by a Selective PageRank-Preserving
Approach ..................................................................................................................................90
  Grigoris Loukides and Robert Gwadera

**Anonymity, Fraud, and Privacy**

An Exploratory Statistical Cusp Catastrophe Model .................................................................100
  Ding-Geng (Din) Chen, Xinguang (Jim) Chen, and Kai Zhang

Using Loglinear Model for Discrimination Discovery and Prevention .......................................110
  Yongkai Wu and Xintao Wu

Fraud Detection in Energy Consumption: A Supervised Approach ........................................120
  Bernat Coma-Puig, Josep Carmona, Ricard Gavaldà, Santiago Alcoverro,
  and Victor Martin

Anomaly Detection in Automobile Control Network Data with Long Short-Term
Memory Networks ....................................................................................................................130
  Adrian Taylor, Sylvain Leblanc, and Nathalie Japkowicz

Anonymizing NYC Taxi Data: Does It Matter? ........................................................................140
  Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T. Silva

**High-Dimensional data**

Infinite Langevin Mixture Modeling and Feature Selection ....................................................149
  Ola Amayri and Nizar Bouguila

Efficient Identification of Tanimoto Nearest Neighbors ........................................................156
  David C. Anastasiu and George Karypis

Parallel Least-Squares Policy Iteration ....................................................................................166
  Jun-Kun Wang and Shou-De Lin

Dilation of Chisini-Jensen-Shannon Divergence .....................................................................174
  Piyush Kumar Sharma and Gary Holness

Projecting "Better Than Randomly": How to Reduce the Dimensionality of Very
Large Datasets in a Way That Outperforms Random Projections ........................................184
  Michael Wojnowicz, Di Zhang, Glenn Chisholm, Xuan Zhao, and Matt Wolff
## Social Media and Crowd

Task Composition in Crowdsourcing .................................................................194  
  Sihem Amer-Yahia, Eric Gaussier, Vincent Leroy, Julien Pilourdault,  
  Ria Mae Borromeo, and Motomichi Toyama

On the Role of Mentions on Tweet Virality .....................................................204  
  Soumajit Pramanik, Qinna Wang, Maximilien Danisch, Sumanth Bandi,  
  Anand Kumar, Jean-Loup Guillaume, and Bivas Mitra

Mining Pre-Exposure Prophylaxis Trends in Social Media .............................214  
  Patrick Breen, Jane Kelly, Timothy Heckman, and Shannon Quinn

Overlapping Target Event and Story Line Detection of Online Newspaper  
  Articles ........................................................................................................222  
  Yifang Wei, Lisa Singh, Brian Gallagher, and David Buttler

Online Collaborative Prediction of Regional Vote Results ............................233  
  Vincent Etter, Mohammad Emtiyaz Khan, Matthias Grossglauser,  
  and Patrick Thiran

## Temporal Analytics

Continuous Monitoring of A/B Tests without Pain: Optional Stopping  
in Bayesian Testing .......................................................................................243  
  Alex Deng, Jiannan Lu, and Shouyuan Chen

Learning Temporal Dependence from Time-Series Data with Latent Variables  
.................................................................................................................253  
  Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran

Trend Detection Based Regret Minimization for Bandit Problems ................263  
  Paresh Nakhe and Rebecca Reiffenhäuser

A Symbolic Tree Model for Oil and Gas Production Prediction Using  
  Time-Series Production Data .....................................................................272  
  Bingjie Wei, Helen Pinto, and Xin Wang

Resampling Strategies for Imbalanced Time Series ......................................282  
  Nuno Moniz, Paula Branco, and Luís Torgo

## Scale

Performance Improvement of MapReduce Process by Promoting Deep Data  
  Locality ......................................................................................................292  
  Sungchul Lee, Ju-Yeon Jo, and Yoohwan Kim

Closest Interval Join Using MapReduce .........................................................302  
  Qiang Zhang, Andy He, Chris Liu, and Eric Lo

EM*: An EM Algorithm for Big Data ...........................................................312  
  Hasan Kurban, Mark Jenne, and Mehmet M. Dalkilic
Efficient Sampling-Based ADMM for Distributed Data
Jun-Kun Wang and Shou-De Lin

A Parallel Framework for Grid-Based Bottom-Up Subspace Clustering
Poonam Goyal, Sonal Kumari, Shubham Singh, Vivek Kishore,
Sundar S. Balasubramaniam, and Navneet Goyal

Search and Mining
Impact of Query Sample Selection Bias on Information Retrieval System
Ranking
Massimo Melucci

Mining Research Problems from Scientific Literature
Chanakya Aalla and Vikram Pudi

Perceived, Projected, and True Investment Expertise: Not All Experts Provide Expert Recommendations
Amit Shavit and Sameena Shah

A Multi-Granularity Pattern-Based Sequence Classification Framework for Educational Data
Mohammad Jaber, Peter T. Wood, Panagiotis Papapetrou,
and Ana González-Marcos

Relational and Structured Data
Inconsistent Node Flattening for Improving Top-Down Hierarchical Classification
Azad Naik and Huzefa Rangwala

Learning Multifaceted Latent Activities from Heterogeneous Mobile Data
Thanh-Binh Nguyen, Vu Nguyen, Thuong Nguyen, Svetla Venkatesh,
Mohan Kumar, and Dinh Phung

The Synthetic Data Vault
Neha Patki, Roy Wedge, and Kalyan Veeramachaneni

A Decision Tree-Based Approach for Categorizing Spatial Database Query Results
Xiangfu Meng, Xiaoyan Zhang, Jinguang Sun, Lin Li, Changzheng Xing,
and Chongchun Bi

The Semantic Knowledge Graph: A Compact, Auto-Generated Model for Real-Time Traversal and Ranking of any Relationship within a Domain
Trey Grainger, Khalifeh Aljadda, Mohammed Korayem, and Andries Smith
Predictive Analytics

Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering .................................................................430

James Max Kanter, Owen Gillespie, and Kalyan Veeramachaneni

What Would a Data Scientist Ask? Automatically Formulating and Solving Predictive Problems .................................................................440

Benjamin Schreck and Kalyan Veeramachaneni

Detecting Inaccurate Predictions of Pediatric Surgical Durations .................................................................452

Zhengyuan Zhou, Daniel Miller, Neal Master, David Scheinker,
Nicholas Bambos, and Peter Glynn

Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data .................................................................458

Luca Oneto, Emanuele Fumeo, Giorgio Clerico, Renzo Canepa, Federico Papa,
Carlo Dambra, Nadia Mazzino, and Davide Anguita

Waiting to Be Sold: Prediction of Time-Dependent House Selling Probability .................................................................468

Mansurul Bhuiyan and Mohammad Al Hasan

Business Intelligence

Customer Simulation for Direct Marketing Experiments .................................................................478

Yegor Tkachenko, Mykel J. Kochenderfer, and Krzysztof Kluza

Behavior-Oriented Time Segmentation for Mining Individualized Rules of Mobile Phone Users .................................................................488

Iqbal H. Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han

Online Experimentation Diagnosis and Troubleshooting Beyond AA Validation .................................................................498

Zhenyu Zhao, Miao Chen, Don Matheson, and Maria Stone

Role Models: Mining Role Transitions Data in IT Project Management .................................................................508

Girish Keshev Palshikar, Sachin Pawar, and Nitin Ramrakhiani

Deconstructing Domain Names to Reveal Latent Topics .................................................................518

Cheryl J. Flynn, Kenneth E. Shirley, and Wei Wang

E-Commerce

Reserve Price Optimization at Scale .................................................................528

Daniel Austin, Sam Seljan, Julius Monello, and Stephanie Tzeng

Uncovering the Bitcoin Blockchain: An Analysis of the Full Users Graph .................................................................537

Damiano Di Francesco Maesa, Andrea Marino, and Laura Ricci

Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns .................................................................547

Martin Kirchgessner, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra
Special Sessions Papers

Big Behavioral Data Analytics
Data-Driven Sales Leads Prediction for Everything-as-a-Service in the Cloud .................................................................557
Chul Sung, Bo Zhang, Chunhui Y. Higgins, and Yoonsuck Choe

Churn Prediction in Mobile Social Games: Towards a Complete Assessment
Using Survival Ensembles ....................................................................................................................................................................564
África Periáñez, Alain Saas, Anna Guitart, and Colin Magne

Web Behavior Analysis Using Sparse Non-Negative Matrix Factorization .............................................................574
Akihiro Demachi, Shin Matsushima, and Kenji Yamanishi

Evidence-Based Behavioral Model for Calendar Schedules of Individual Mobile Phone Users .......................................................584
Iqbal H. Sarker, Muhammad Ashad Kabir, Alan Colman, and Jun Han

Data Science for Agricultural Decision Support Systems
Disease Detection and Severity Estimation in Cotton Plant from Unconstrained Images ..........................................................594
Aditya Parikh, Mehul S. Raval, Chandrasinh Parmar, and Sanjay Chaudhary

Digital Knowledge Ecosystem for Achieving Sustainable Agriculture Production: A Case Study from Sri Lanka ..........................602
Athula Ginige, Anusha I. Walisadeera, Tamara Ginige, Lasanthi De Silva, Pasquale Di Giovanni, Maneesh Mathai, Jeevani Goonetillake, Gihan Wikramanayake, Giuliana Vitiello, Monica Sebillo, Genoveffa Tortora, Deborah Richards, and Ramesh Jain

Environmental and Geo-spatial Data Analytics
Efficient Large Scale Clustering Based on Data Partitioning .................................................................612
Malika Bendechache, M-Tahar Kechadi, and Nhien-An Le-Khac

Traffic Risk Mining Using Partially Ordered Non-Negative Matrix Factorization ................................................622
Taito Lee, Shin Matsushima, and Kenji Yamanishi

On the Use of Ontology as A Priori Knowledge into Constrained Clustering ...............................................................632
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille, and Younès Bennani

Maritime Pattern Extraction from AIS Data Using a Genetic Algorithm ..........................................................642
Andrej Dobrkovic, Maria-Eugenia Iacob, and Jos Van Hillegersberg
**Game Data Science**

Using Players' Gameplay Action-Decision Profiles to Prescribe Training: Reducing Training Costs with Serious Games Analytics .................................................................652
  Christian Sebastian Loh and I-Hung Li

What Did I Do Wrong in My MOBA Game? Mining Patterns Discriminating Deviant Behaviours ........................................................................................................662
  Olivier Cavadenti, Victor Codocedo, Jean-François Boulicaut, and Mehdi Kaytoue

On the Tiny Yet Real Happiness Phenomenon in the Mobile Games Market ........................................672
  Po-Heng Chen, Yi-Pei Tu, and Kuan-Ta Chen

**Health Data Science**

Actitracker: A Smartphone-Based Activity Recognition System for Improving Health and Well-Being ........................................................................................................................................682
  Gary M. Weiss, Jeffrey W. Lockhart, Tony T. Pulickal, Paul T. McHugh, Isaac H. Ronan, and Jessica L. Timko

The Highly Adaptive Lasso Estimator ..................................................................................................689
  David Benkeser and Mark Van Der Laan

Meeting Health Care Research Needs in a Kimball Integrated Data Warehouse ................................697
  Robert Hart and Alex Mu-Hsing Kuo

MedCare: Leveraging Medication Similarity for Disease Prediction ..................................................706
  Dipanwita Dasgupta and Nitesh V. Chawla

**Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis**

Connecting Opinions to Opinion-Leaders: A Case Study on Brazilian Political Protests .................................................................716
  Leonardo Rocha, Fernando Mourão, Ramon Vieira, Alan Neves, Dârlinton Carvalho, Bortik Bandyopadhyay, Srinivasan Parthasarathy, and Renato Ferreira

Exploiting a Bootstrapping Approach for Automatic Annotation of Emotions in Texts .................................................................726
  Lea Canales, Carlo Strapparava, Ester Boldrini, and Patricio Martinez-Barco

Senpy: A Pragmatic Linked Sentiment Analysis Framework ................................................................735
  J. Fernando Sánchez-Rada, Carlos A. Iglesias, Ignacio Corcuera, and Óscar Araque

Word Segmentation Algorithms with Lexical Resources for Hashtag Classification .................................................................743
  Credell Simeon, Howard J. Hamilton, and Robert J. Hilderman
**Statistical Learning for Data Science**

A Distributed Decision Tree Algorithm and Its Implementation on Big Data
Platforms ................................................................................................................................................... 752

  Jingxiang Chen, Tao Wang, Ralph Abbey, and Joseph Pingenot

Analysing the History of Autism Spectrum Disorder Using Topic Models ........................................... 762

  Adham Beykikhoshk, Dinh Phung, Ognjen Arandjelović, and Svetha Venkatesh

Sparse Linear Discriminant Analysis in Structured Covariates Space ..................................................... 772

  Sandra E. Safo and Qi Long

Informative Priors and Bayesian Computation .......................................................................................... 782

  Shirin Golchi

Causal Structure Learning with Reduced Partial Correlation Thresholding ............................................. 790

  Arjun Sondhi and Ali Shojaie

Nonparametric Adjoint-Based Inference for Stochastic Differential Equations ........................................ 798

  Harish S. Bhat and R.W.M.A. Madushani

**Statistical and Mathematical tools for Data Mining**

The Uniqueness and Greedy Method for Quadratic Compressive Sensing ............................................. 808

  Jun Fan, Lingchen Kong, Liqun Wang, and Naihua Xiu

Robust Online Time Series Prediction with Recurrent Neural Networks .................................................. 816

  Tian Guo, Zhao Xu, Xin Yao, Haifeng Chen, Karl Aberer, and Koichi Funaya

**Author Index** ............................................................................................................................................ 826