

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

**Montreal, Quebec, Canada
17-19 October 2016**



IEEE Catalog Number: CFP16DSB-POD
ISBN: 978-1-5090-5207-3

**Copyright © 2016 by the Institute of Electrical and Electronics Engineers, Inc
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

******This publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP16DSB-POD
ISBN (Print-On-Demand):	978-1-5090-5207-3
ISBN (Online):	978-1-5090-5206-6

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2016 IEEE International Conference on Data Science and Advanced Analytics

DSAA 2016

Table of Contents

Organizing Committee	xiii
Program Committee	xvii
Reviewers	xxi

Main Conference Papers

Classification

On the Evaluation of Outlier Detection and One-Class Classification Methods	1
<i>Lorne Swersky, Henrique O. Marques, Jörg Sander, Ricardo J.G.B. Campello, and Arthur Zimek</i>	
Active Semi-Supervised Classification Based on Multiple Clustering Hierarchies	11
<i>Antônio J.L. Batista, Ricardo J.G.B. Campello, and Jörg Sander</i>	
Combining Static and Dynamic Features for Multivariate Sequence Classification	21
<i>Anna Leontjeva and Ilya Kuzovkin</i>	
Correcting Relational Bias to Improve Classification in Sparsely-Labeled Networks	31
<i>Joshua R. King and Luke K. McDowell</i>	
Hyperparameter Optimization Machines	41
<i>Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme</i>	

Networks

Temporal Network Change Detection Using Network Centralities	51
<i>Yoshitaro Yonamoto, Kai Morino, and Kenji Yamanishi</i>	
Harvester: Influence Optimization in Symmetric Interaction Networks	61
<i>Sergei Ivanov and Panagiotis Karras</i>	

Pattern Matching Trajectories for Investigative Graph Searches	71
<i>Benjamin W.K. Hung, Anura P. Jayasumana, and Vidarshana W. Bandara</i>	
A Framework for Description and Analysis of Sampling-Based Approximate Triangle Counting Algorithms	80
<i>Mostafa Haghiri Chehreghani</i>	
Limiting the Diffusion of Information by a Selective PageRank-Preserving Approach	90
<i>Grigorios Loukides and Robert Gwadera</i>	

Anonymity, Fraud, and Privacy

An Exploratory Statistical Cusp Catastrophe Model	100
<i>Ding-Geng (Din) Chen, Xinguang (Jim) Chen, and Kai Zhang</i>	
Using Loglinear Model for Discrimination Discovery and Prevention	110
<i>Yongkai Wu and Xintao Wu</i>	
Fraud Detection in Energy Consumption: A Supervised Approach	120
<i>Bernat Coma-Puig, Josep Carmona, Ricard Gavaldà, Santiago Alcoverro, and Victor Martin</i>	
Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks	130
<i>Adrian Taylor, Sylvain Leblanc, and Nathalie Japkowicz</i>	
Anonymizing NYC Taxi Data: Does It Matter?	140
<i>Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T. Silva</i>	

High-Dimensional data

Infinite Langevin Mixture Modeling and Feature Selection	149
<i>Ola Amayri and Nizar Bouguila</i>	
Efficient Identification of Tanimoto Nearest Neighbors	156
<i>David C. Anastasiu and George Karypis</i>	
Parallel Least-Squares Policy Iteration	166
<i>Jun-Kun Wang and Shou-De Lin</i>	
Dilation of Chisini-Jensen-Shannon Divergence	174
<i>Piyush Kumar Sharma and Gary Holness</i>	
Projecting "Better Than Randomly": How to Reduce the Dimensionality of Very Large Datasets in a Way That Outperforms Random Projections	184
<i>Michael Wojnowicz, Di Zhang, Glenn Chisholm, Xuan Zhao, and Matt Wolff</i>	

Social Media and Crowd

Task Composition in Crowdsourcing	194
<i>Sihem Amer-Yahia, Eric Gaussier, Vincent Leroy, Julien Pilourdault, Ria Mae Borromeo, and Motomichi Toyama</i>	
On the Role of Mentions on Tweet Virality	204
<i>Soumajit Pramanik, Qinna Wang, Maximilien Danisch, Sumanth Bandi, Anand Kumar, Jean-Loup Guillaume, and Bivas Mitra</i>	
Mining Pre-Exposure Prophylaxis Trends in Social Media	214
<i>Patrick Breen, Jane Kelly, Timothy Heckman, and Shannon Quinn</i>	
Overlapping Target Event and Story Line Detection of Online Newspaper Articles	222
<i>Yifang Wei, Lisa Singh, Brian Gallagher, and David Buttler</i>	
Online Collaborative Prediction of Regional Vote Results	233
<i>Vincent Etter, Mohammad Emtiyaz Khan, Matthias Grossglauser, and Patrick Thiran</i>	

Temporal Analytics

Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing	243
<i>Alex Deng, Jiannan Lu, and Shouyuan Chen</i>	
Learning Temporal Dependence from Time-Series Data with Latent Variables	253
<i>Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran</i>	
Trend Detection Based Regret Minimization for Bandit Problems	263
<i>Paresh Nakhe and Rebecca Reiffenhäuser</i>	
A Symbolic Tree Model for Oil and Gas Production Prediction Using Time-Series Production Data	272
<i>Bingjie Wei, Helen Pinto, and Xin Wang</i>	
Resampling Strategies for Imbalanced Time Series	282
<i>Nuno Moniz, Paula Branco, and Luís Torgo</i>	

Scale

Performance Improvement of MapReduce Process by Promoting Deep Data Locality	292
<i>Sungchul Lee, Ju-Yeon Jo, and Yoohwan Kim</i>	
Closest Interval Join Using MapReduce	302
<i>Qiang Zhang, Andy He, Chris Liu, and Eric Lo</i>	
EM*: An EM Algorithm for Big Data	312
<i>Hasan Kurban, Mark Jenne, and Mehmet M. Dalkilic</i>	

Efficient Sampling-Based ADMM for Distributed Data	321
<i>Jun-Kun Wang and Shou-De Lin</i>	
A Parallel Framework for Grid-Based Bottom-Up Subspace Clustering	331
<i>Poonam Goyal, Sonal Kumari, Shubham Singh, Vivek Kishore, Sundar S. Balasubramaniam, and Navneet Goyal</i>	

Search and Mining

Impact of Query Sample Selection Bias on Information Retrieval System Ranking	341
<i>Massimo Melucci</i>	
Mining Research Problems from Scientific Literature	351
<i>Chanakya Aalla and Vikram Pudi</i>	
Perceived, Projected, and True Investment Expertise: Not All Experts Provide Expert Recommendations	361
<i>Amit Shavit and Sameena Shah</i>	
A Multi-Granularity Pattern-Based Sequence Classification Framework for Educational Data	370
<i>Mohammad Jaber, Peter T. Wood, Panagiotis Papapetrou, and Ana González-Marcos</i>	

Relational and Structured Data

Inconsistent Node Flattening for Improving Top-Down Hierarchical Classification	379
<i>Azad Naik and Huzefa Rangwala</i>	
Learning Multifaceted Latent Activities from Heterogeneous Mobile Data	389
<i>Thanh-Binh Nguyen, Vu Nguyen, Thuong Nguyen, Svetha Venkatesh, Mohan Kumar, and Dinh Phung</i>	
The Synthetic Data Vault	399
<i>Neha Patki, Roy Wedge, and Kalyan Veeramachaneni</i>	
A Decision Tree-Based Approach for Categorizing Spatial Database Query Results	411
<i>Xiangfu Meng, Xiaoyan Zhang, Jinguang Sun, Lin Li, Changzheng Xing, and Chongchun Bi</i>	
The Semantic Knowledge Graph: A Compact, Auto-Generated Model for Real-Time Traversal and Ranking of any Relationship within a Domain	420
<i>Trey Grainger, Khalifeh Aljadda, Mohammed Korayem, and Andries Smith</i>	

Predictive Analytics

Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering	430
<i>James Max Kanter, Owen Gillespie, and Kalyan Veeramachaneni</i>	
What Would a Data Scientist Ask? Automatically Formulating and Solving Predictive Problems	440
<i>Benjamin Schreck and Kalyan Veeramachaneni</i>	
Detecting Inaccurate Predictions of Pediatric Surgical Durations	452
<i>Zhengyuan Zhou, Daniel Miller, Neal Master, David Scheinker, Nicholas Bambos, and Peter Glynn</i>	
Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data	458
<i>Luca Oneto, Emanuele Fumeo, Giorgio Clerico, Renzo Canepa, Federico Papa, Carlo Dambra, Nadia Mazzino, and Davide Anguita</i>	
Waiting to Be Sold: Prediction of Time-Dependent House Selling Probability	468
<i>Mansurul Bhuiyan and Mohammad Al Hasan</i>	

Business Intelligence

Customer Simulation for Direct Marketing Experiments	478
<i>Yegor Tkachenko, Mykel J. Kochenderfer, and Krzysztof Kluza</i>	
Behavior-Oriented Time Segmentation for Mining Individualized Rules of Mobile Phone Users	488
<i>Iqbal H. Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han</i>	
Online Experimentation Diagnosis and Troubleshooting Beyond AA Validation	498
<i>Zhenyu Zhao, Miao Chen, Don Matheson, and Maria Stone</i>	
Role Models: Mining Role Transitions Data in IT Project Management	508
<i>Girish Keshav Palshikar, Sachin Pawar, and Nitin Ramrakhiani</i>	
Deconstructing Domain Names to Reveal Latent Topics	518
<i>Cheryl J. Flynn, Kenneth E. Shirley, and Wei Wang</i>	

E-Commerce

Reserve Price Optimization at Scale	528
<i>Daniel Austin, Sam Seljan, Julius Monello, and Stephanie Tzeng</i>	
Uncovering the Bitcoin Blockchain: An Analysis of the Full Users Graph	537
<i>Damiano Di Francesco Maesa, Andrea Marino, and Laura Ricci</i>	
Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns	547
<i>Martin Kirchgessner, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra</i>	

Special Sessions Papers

Big Behavioral Data Analytics

Data-Driven Sales Leads Prediction for Everything-as-a-Service in the Cloud	557
<i>Chul Sung, Bo Zhang, Chunhui Y. Higgins, and Yoonsuck Choe</i>	
Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles	564
<i>África Perriáñez, Alain Saas, Anna Guitart, and Colin Magne</i>	
Web Behavior Analysis Using Sparse Non-Negative Matrix Factorization	574
<i>Akihiro Demachi, Shin Matsushima, and Kenji Yamanishi</i>	
Evidence-Based Behavioral Model for Calendar Schedules of Individual Mobile Phone Users	584
<i>Iqbal H. Sarker, Muhammad Ashad Kabir, Alan Colman, and Jun Han</i>	

Data Science for Agricultural Decision Support Systems

Disease Detection and Severity Estimation in Cotton Plant from Unconstrained Images	594
<i>Aditya Parikh, Mehul S. Raval, Chandrasinh Parmar, and Sanjay Chaudhary</i>	
Digital Knowledge Ecosystem for Achieving Sustainable Agriculture Production: A Case Study from Sri Lanka	602
<i>Athula Ginige, Anusha I. Walisadeera, Tamara Ginige, Lasanthi De Silva, Pasquale Di Giovanni, Maneesh Mathai, Jeevani Goonetillake, Gihan Wikramanayake, Giuliana Vitiello, Monica Sebillio, Genoveffa Tortora, Deborah Richards, and Ramesh Jain</i>	

Environmental and Geo-spatial Data Analytics

Efficient Large Scale Clustering Based on Data Partitioning	612
<i>Malika Bendechache, M-Tahar Kechadi, and Nhien-An Le-Khac</i>	
Traffic Risk Mining Using Partially Ordered Non-Negative Matrix Factorization	622
<i>Taito Lee, Shin Matsushima, and Kenji Yamanishi</i>	
On the Use of Ontology as A Priori Knowledge into Constrained Clustering	632
<i>Hatim Chahdi, Nistor Grozavu, Isabelle Mougnot, Laure Berti-Equille, and Younès Bennani</i>	
Maritime Pattern Extraction from AIS Data Using a Genetic Algorithm	642
<i>Andrej Dobrkovic, Maria-Eugenia Iacob, and Jos Van Hillegersberg</i>	

Game Data Science

Using Players' Gameplay Action-Decision Profiles to Prescribe Training: Reducing Training Costs with Serious Games Analytics	652
<i>Christian Sebastian Loh and I-Hung Li</i>	
What Did I Do Wrong in My MOBA Game? Mining Patterns Discriminating Deviant Behaviours	662
<i>Olivier Cavadenti, Victor Codocedo, Jean-François Boulicaut, and Mehdi Kaytoue</i>	
On the Tiny Yet Real Happiness Phenomenon in the Mobile Games Market	672
<i>Po-Heng Chen, Yi-Pei Tu, and Kuan-Ta Chen</i>	

Health Data Science

Actitracker: A Smartphone-Based Activity Recognition System for Improving Health and Well-Being	682
<i>Gary M. Weiss, Jeffrey W. Lockhart, Tony T. Pulickal, Paul T. McHugh, Isaac H. Ronan, and Jessica L. Timko</i>	
The Highly Adaptive Lasso Estimator	689
<i>David Benkeser and Mark Van Der Laan</i>	
Meeting Health Care Research Needs in a Kimball Integrated Data Warehouse	697
<i>Robert Hart and Alex Mu-Hsing Kuo</i>	
MedCare: Leveraging Medication Similarity for Disease Prediction	706
<i>Dipanwita Dasgupta and Nitesh V. Chawla</i>	

Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis

Connecting Opinions to Opinion-Leaders: A Case Study on Brazilian Political Protests	716
<i>Leonardo Rocha, Fernando Mourão, Ramon Vieira, Alan Neves, Dárlinton Carvalho, Bortik Bandyopadhyay, Srinivasan Parthasarathy, and Renato Ferreira</i>	
Exploiting a Bootstrapping Approach for Automatic Annotation of Emotions in Texts	726
<i>Lea Canales, Carlo Strapparava, Ester Boldrini, and Patricio Martnez-Barco</i>	
Senpy: A Pragmatic Linked Sentiment Analysis Framework	735
<i>J. Fernando Sánchez-Rada, Carlos A. Iglesias, Ignacio Corcuera, and Óscar Araque</i>	
Word Segmentation Algorithms with Lexical Resources for Hashtag Classification	743
<i>Credell Simeon, Howard J. Hamilton, and Robert J. Hilderman</i>	

Statistical Learning for Data Science

A Distributed Decision Tree Algorithm and Its Implementation on Big Data Platforms	752
<i>Jingxiang Chen, Tao Wang, Ralph Abbey, and Joseph Pingetot</i>	
Analysing the History of Autism Spectrum Disorder Using Topic Models	762
<i>Adham Beykikhoshk, Dinh Phung, Ognjen Arandjelović, and Svetha Venkatesh</i>	
Sparse Linear Discriminant Analysis in Structured Covariates Space	772
<i>Sandra E. Safo and Qi Long</i>	
Informative Priors and Bayesian Computation	782
<i>Shirin Golchi</i>	
Causal Structure Learning with Reduced Partial Correlation Thresholding	790
<i>Arjun Sondhi and Ali Shojaie</i>	
Nonparametric Adjoint-Based Inference for Stochastic Differential Equations	798
<i>Harish S. Bhat and R.W.M.A. Madushani</i>	

Statistical and Mathematical tools for Data Mining

The Uniqueness and Greedy Method for Quadratic Compressive Sensing	808
<i>Jun Fan, Lingchen Kong, Liqun Wang, and Naihua Xiu</i>	
Robust Online Time Series Prediction with Recurrent Neural Networks	816
<i>Tian Guo, Zhao Xu, Xin Yao, Haifeng Chen, Karl Aberer, and Koichi Funaya</i>	
Author Index	826