

2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017)

**Tokyo, Japan
19 – 21 October 2017**



**IEEE Catalog Number: CFP17DSB-POD
ISBN: 978-1-5090-5005-5**

**Copyright © 2017 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 is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP17DSB-POD
ISBN (Print-On-Demand):	978-1-5090-5005-5
ISBN (Online):	978-1-5090-5004-8

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

2017 International Conference on Data Science and Advanced Analytics

DSAA 2017

Table of Contents

Message from General Chairs.....	xiii
Message from Program Chairs.....	xv
Conference Organization.....	xvii
Program Committee.....	xix
Reviewers.....	xxv

Classification and Regression

The k-Nearest Representatives Classifier: A Distance-Based Classifier with Strong Generalization Bounds	1
<i>Cyrus Cousins and Eli Upfal</i>	
Cyclic Classifier Chain for Cost-Sensitive Multilabel Classification	11
<i>Yi-An Lin and Hsuan-Tien Lin</i>	
Learning Low-Rank Document Embeddings with Weighted Nuclear Norm Regularization	21
<i>Lukas Pfahler, Katharina Morik, Frederik Elwert, Samira Tabti, and Volkhart Krech</i>	
Learning Through Utility Optimization in Regression Tasks	30
<i>Paula Branco, Luís Torgo, Rita P. Ribeiro, Eibe Frank, Bernhard Pfahringer, and Markus Michael Rau</i>	

Image and Behavior Modeling

Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring	40
<i>Hung Nguyen, Sarah J. Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G. Ritchie, and Dinh Phung</i>	
Nazr-CNN: Fine-Grained Classification of UAV Imagery for Damage Assessment	50
<i>Nazia Attari, Ferda Ofli, Mohammad Awad, Ji Lucas, and Sanjay Chawla</i>	

Website Navigation Behavior Analysis for Bot Detection	60
<i>Rabih Haidar and Shady Elbassuoni</i>	
Inform Product Change through Experimentation with Data-Driven Behavioral Segmentation	69
<i>Zhenyu Zhao, Yan He, and Miao Chen</i>	

Evolving Networks

Scalable RFM-enriched Representation Learning for Churn Prediction	79
<i>Sandra Mitrovic, Gaurav Singh, Bart Baesens, Wilfried Lemahieu, and Jochen de Weerd</i>	
A Comparative Study of Different Approaches for Tracking Communities in Evolving Social Networks	89
<i>Ziwei He, Etienne Gael Tajeuna, Shengrui Wang, and Mohamed Bouguessa</i>	
The Initialization and Parameter Setting Problem in Tensor Decomposition-Based Link Prediction	99
<i>Sofia da Silva Fernandes, Hadi Fanaee Tork, and João Manuel Portela da Gama</i>	

Feature Exploration and Classification

Combining Instance and Feature Neighbors for Efficient Multi-label Classification	109
<i>Len Feremans, Boris Cule, Celine Vens, and Bart Goethals</i>	
Expert Estimates for Feature Relevance are Imperfect	119
<i>Patrick M. de Boer, Marcel C. Böhler, and Abraham Bernstein</i>	
Multi-label Learning with Label-Specific Features via Clustering Ensemble	129
<i>Wang Zhan and Min-Ling Zhang</i>	
Customizing Travel Packages with Interactive Composite Items	137
<i>Manish Singh, Ria Mae Borromeo, Anas Hosami, Sihem Amer-Yahia, and Shady Elbassuoni</i>	

Network Analysis and Topic Modeling

Materials Science Literature-Patent Relevance Search: A Heterogeneous Network Analysis Approach	146
<i>Pingjie Tang, Jed Pitera, Dmitry Zubarev, and Nitesh V. Chawla</i>	
NDlib: Studying Network Diffusion Dynamics	155
<i>Giulio Rossetti, Letizia Milli, Salvatore Rinzivillo, Alina Sirbu, Dino Pedreschi, and Fosca Giannotti</i>	
Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation	165
<i>Shaheen Syed and Marco Spruit</i>	

Incremental Author Name Disambiguation for Scientific Citation Data	175
<i>Zhengqiao Zhao, Jason Rollins, Linge Bai, and Gail Rosen</i>	

Big Data and Disaster Management / Advanced Informatic Measurement Using Statistics, Machine Learning and Pattern Recognition

Supercharging Crowd Dynamics Estimation in Disasters via Spatio-Temporal Deep Neural Network	184
<i>Fang-Zhou Jiang, Lei Zhong, Kanchana Thilakarathna, Aruna Seneviratne, Kiyoshi Takano, Shigeki Yamada, and Yusheng Ji</i>	
Geo-Spatial Multimedia Sentiment Analysis in Disasters	193
<i>Abdullah Alfarrarjeh, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi</i>	
Situational Awareness from Social Media Photographs Using Automated Image Captioning	203
<i>João Monteiro, Asanobu Kitamoto, and Bruno Martins</i>	
Machine Learning Independent of Population Distributions for Measurement	212
<i>Takashi Washio, Gaku Imamura, and Genki Yoshikawa</i>	

Time Series Modeling and Forecast

A Dynamic Factor Machine Learning Method for Multi-variate and Multi-step-Ahead Forecasting	222
<i>Gianluca Bontempi, Yann-Aël Le Borgne, and Jacopo de Stefani</i>	
CSAR: The Cross-Sectional Autoregression Model	232
<i>Claudio Hartmann, Martin Hahmann, Dirk Habich, and Wolfgang Lehner</i>	
Dynamic and Heterogeneous Ensembles for Time Series Forecasting	242
<i>Vitor Cerqueira, Luis Torgo, Mariana Oliveira, and Bernhard Pfahringer</i>	
Forward-Backward Smoothing for Hidden Markov Models of Point Pattern Data	252
<i>Nhan Dam, Dinh Phung, Ba-Ngu Vo, and Viet Huynh</i>	

Search and Sequence Modeling

RadiusSketch: Massively Distributed Indexing of Time Series	262
<i>Djamel Edine Yagoubi, Reza Akbarinia, Florent Masseglia, and Dennis Shasha</i>	
BJR-Tree: Fast Skyline Computation Algorithm for Serendipitous Searching Problems	272
<i>Kenichi Koizumi, Peter Eades, Kei Hiraki, and Mary Inaba</i>	
A Directional Change Based Trading Strategy with Dynamic Thresholds	283
<i>Nora Alkhamees and Maria Fasli</i>	

Subsequence Search Considering Duration and Relations of Events in Time	
Interval-Based Events Sequences	293
<i>Cheng-Wei Yang, Bijay Prasad Jaysawal, and Jen-Wei Huang</i>	

Environmental and Geo-Spatial Data Analytics I

There's a Path for Everyone: A Data-Driven Personal Model Reproducing	
Mobility Agendas	303
<i>Riccardo Guidotti, Roberto Trasarti, Mirco Nanni, Fosca Giannotti,</i> <i>and Dino Pedreschi</i>	
Heterogeneous Information Integration for Mountain Augmented Reality	
Mobile Apps	313
<i>Darian Frajberg, Piero Fraternali, and Rocio Nahime Torres</i>	
Predictive Classification of Water Consumption Time Series Using	
Non-homogeneous Markov Models	323
<i>Milad Leyli Abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz,</i> <i>Pierre Mandel, Cédric Féliers, and Olivier Chesneau</i>	
DP-POIRS: A Diversified and Personalized Point-of-Interest Recommendation	
System	332
<i>Xiangfu Meng, Yanhuan Tang, and Xiaoyan Zhang</i>	

Statistical Approaches

On the Jeffreys-Lindley Paradox and the Looming Reproducibility Crisis	
in Machine Learning	334
<i>Daniel Berrar and Werner Dubitzky</i>	
M3A: Model, MetaModel and Anomaly Detection for Inter-arrivals of Web	
Searches and Postings	341
<i>Da-Cheng Juan, Neil Shah, Mingyu Tang, Zhiliang Qian, Diana Marculescu,</i> <i>and Christos Faloutsos</i>	
A Consistency-Based Multimodal Graph Embedding Method	
for Dimensionality Reduction	351
<i>Ilias Kalamaras, Anastasios Drosou, Eleftheria Polychronidou,</i> <i>and Dimitrios Tzovaras</i>	
Sample, Estimate, Tune: Scaling Bayesian Auto-Tuning of Data Science	
Pipelines	361
<i>Alec Anderson, Sebastien Dubois, Alfredo Cuesta-infante,</i> <i>and Kalyan Veeramachaneni</i>	

Acoustic and Video Recognition

What Makes a Video Memorable?	373
<i>Akankshya Kar, Prashasthi Mavin, Yogesh Ghaturlle, and Vani M.</i>	
Convolutional Neural Networks Based Multi-task Deep Learning for Movie Review Classification	382
<i>Xuanyi Li, Weimin Wu, and Hongye Su</i>	
Masked Conditional Neural Networks for Automatic Sound Events Recognition	389
<i>Fady Medhat, David Chesmore, and John Robinson</i>	
A Spatial-Cue-Based Probabilistic Model for Bird Song Scene Analysis	395
<i>Ryosuke Kojima, Osamu Sugiyama, Kotaro Hoshiba, Reiji Suzuki, and Kazuhiro Nakadai</i>	

New Applications I

The Data and Science behind GrabShare Carpooling	405
<i>Muchen Tang, Serene Ow, Wenqing Chen, Yang Cao, Kong-wei Lye, and Yaozhang Pan</i>	
Regression Based Model for Autosteering of a Car with Delayed Steering Response	412
<i>Vsevolod Nikulin, Albert Podusenko, Ivan Tanev, and Katsunori Shimohara</i>	
Ensemble-Based Location Tracking Using Passive RFID	420
<i>Hao-Ying Liang, Yun-Tung Shieh, Addicam Sanjay, Shao-Wen Yang, and Shou-De Lin</i>	
Leveraging on Predictive Analytics to Manage Clinic No Show and Improve Accessibility of Care	429
<i>Guanhua Lee, Sijia Wang, Fransiscus Dipuro, Jue Hou, Priyanka Grover, Lian Leng Low, Nan Liu, and Chui Yee Loke</i>	

Environmental and Geo-Spatial Data Analytics II

A Shape-Based Approach to Spatio-Temporal Data Analysis Using Satellite Imagery	439
<i>Darpan Baheti and K. S Rajan</i>	
Mobility Genome™- A Framework for Mobility Intelligence from Large-Scale Spatio-Temporal Data	449
<i>The Anh Dang, Jayakumaran Deepak, Jingxuan Wang, Shixin Luo, Yunye Jin, Yibin Ng, Aloysius Lim, and Ying Li</i>	
A Peak Detection Method to Uncover Events from Social Media	459
<i>Carmela Comito, Deborah Falcone, and Domenico Talia</i>	

Semantic Trajectory Modeling for Dynamic Built Environments	468
<i>Christophe Cruz</i>	

New Applications II

HiSPEED: A System for Mining Performance Appraisal Data and Text	477
<i>Girish Keshav Palshikar, Manoj Apte, Sachin Pawar, and Nitin Ramrakhiyani</i>	
Identification of Signal and Noise Components in Spacecraft Neutral Particle Data Using a Bi-Level Mixture Model	487
<i>Shin'Ya Nakano and Yoshifumi Futaana</i>	
A Collaborative Filtering-Based Two Stage Model with Item Dependency for Course Recommendation	496
<i>Eric L. Lee, Tsung-Ting Kuo, and Shou-De Lin</i>	
Enriching Course-Specific Regression Models with Content Features for Grade Prediction	504
<i>Qian Hu, Agoritsa Polyzou, George Karypis, and Huzefa Rangwala</i>	

Beyond IID: Non-IID Learning

Steganalysis Feature Subspace Selection Based on Fisher Criterion	514
<i>Chunfang Yang, Yi Zhang, Ping Wang, Xiangyang Luo, Fenlin Liu, and Jicang Lu</i>	
Coupled Bayesian Matrix Factorization in Recommender Systems	522
<i>Xueci Zhao, Chengzhang Zhu, and Lizhi Cheng</i>	
A Comparative Study of Performance Estimation Methods for Time Series Forecasting	529
<i>Vitor Cerqueira, Luis Torgo, Jasmina Smailović, and Igor Mozetič</i>	

Graph and Network

On Spectral Analysis of Directed Signed Graphs	539
<i>Yuemeng Li, Xintao Wu, and Aidong Lu</i>	
AnonML: Locally Private Machine Learning over a Network of Peers	549
<i>Bennett Cyphers and Kalyan Veeramachaneni</i>	
Maximizing Network Performance Based on Group Centrality by Creating Most Effective k-Links	561
<i>Kouzou Ohara, Kazumi Saito, Masahiro Kimura, and Hiroshi Motoda</i>	
Multi-task Network Embedding	571
<i>Linchuan Xu, Xiaokai Wei, Jiannong Cao, and Philip S. Yu</i>	

Network Service I

Multiple Social Role Embedding	581
<i>Linchuan Xu, Xiaokai Wei, Jiannong Cao, and Philip S. Yu</i>	
FeatureHub: Towards Collaborative Data Science	590
<i>Micah J. Smith, Roy Wedge, and Kalyan Veeramachaneni</i>	
Identifying Anomalous Nodes in Multidimensional Networks	601
<i>Amani Chouchane and Mohamed Bouguessa</i>	
Discovering Community Structure in Multilayer Networks	611
<i>Soumajit Pramanik, Raphael Tackx, Anchit Navelkar, Jean-Loup Guillaume, and Bivas Mitra</i>	

Outliers and Compression

Learning to Compress Unstructured Mesh Data from Simulations	621
<i>Chandrika Kamath</i>	
An Assessment of Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection	631
<i>Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, and Gianluca Bontempi</i>	
A Probabilistic, Mechanism-Independent Outlier Detection Method for Online Experimentation	640
<i>Yan He and Miao Chen</i>	

Data Science in Societal Debates / Game Data Science

News Consumption during the Italian Referendum: A Cross-Platform Analysis on Facebook and Twitter	648
<i>Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo</i>	
Feature Analysis for Fake Review Detection through Supervised Classification	658
<i>Julien Fontanarava, Gabriella Pasi, and Marco Viviani</i>	
Online k-Maxoids Clustering	667
<i>Rafet Sifa and Christian Bauckhage</i>	

Network Service II

Disentangled Link Prediction for Signed Social Networks via Disentangled Representation Learning	676
<i>Linchuan Xu, Xiaokai Wei, Jiannong Cao, and Philip S. Yu</i>	
Exploiting Digital DNA for the Analysis of Similarities in Twitter Behaviours	686
<i>Stefano Cresci, Roberto di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi</i>	

Where are You Going? Next Place Prediction from Twitter	696
<i>Carmela Comito</i>	

A Study of Stochastic Mixed Membership Models for Link Prediction in Social Networks	706
<i>Adrien Dulac, Eric Gaussier, and Christine Largeron</i>	

Estimating Dependency and Dimensions

Latent Dimensionality Estimation for Probabilistic Canonical Correlation Analysis Using Normalized Maximum Likelihood Code-Length	716
<i>Tomohiko Nakmaura, Tomoharu Iwata, and Kenji Yamanishi</i>	

A Novel Approach for Estimating Multiple Sparse Precision Matrices Using $\ell_{0,0}$ Regularization	726
<i>Duy Nhat Phan and Hoai An Le Thi</i>	

Copula-Based High Dimensional Cross-Market Dependence Modeling	734
<i>Jia Xu, Wei Wei, and Longbing Cao</i>	

Causal Patterns: Extraction of Multiple Causal Relationships by Mixture of Probabilistic Partial Canonical Correlation Analysis	744
<i>Hiroki Mori, Keisuke Kawano, and Hiroki Yokoyama</i>	

Data and Information Quality

SECODA: Segmentation- and Combination-Based Detection of Anomalies	755
<i>Ralph Foorthuis</i>	

Extended Methods to Handle Classification Biases	765
<i>Emma Beauxis-Aussalet and Lynda Hardman</i>	

Toward Optimal Streaming Feature Selection	775
<i>Noura Al Nuaimi and Mohammad M. Masud</i>	

Author Index	783
---------------------------	-----