

2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)

**Paris, France
19-21 October 2015**

Pages 1-568



**IEEE Catalog Number: CFP15DSB-POD
ISBN: 978-1-4673-8274-8**

**Copyright © 2015 by the Institute of Electrical and Electronic 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:	CFP15DSB-POD
ISBN 13:	978-1-4673-8274-8

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

Proceedings of the 2015 International Conference on Data Science and Advanced Analytics (DSAA'2015)

TABLE OF CONTENTS

Welcome from IEEE DSAA'2015 Chairs	III
DSAA'2015 Organizing Committee	V
DSAA'2015 Advisory/Steering Committees	VI
DSAA'2015 Program Committee	VII
Keynote Speeches	XI
Special Sessions	XVII
Invited Talks for DSAA Trends & Controversies	XXIX
Tutorials	XXXVII
DSAA Technical Papers in Main Conference Sessions	XL
DSAA Technical Papers in Special Sessions	XLVI

DSAA Main Conference Sessions

Timeseries and Sequences

Matrix Factorization Approach to Behavioral Mode Analysis from Acceleration Data.....	1
<i>Yehezkel S. Resheff</i>	
Random-shapelet : An Algorithm for Fast Shapelet Discovery.....	7
<i>Xavier Renard, Maria Rifqi, Walid Erray and Marcin Detyniecki</i>	
On-line Detection of Continuous Changes in Stochastic Processes.....	17
<i>Kohei Miyaguchi and Kenji Yamanishi</i>	
Duration Models for Activity Recognition and Prediction in Buildings using Hidden Markov Models.....	26
<i>Antonio Ridi, Nikos Zarkadis, Christophe Gisler and Jean Hennebert</i>	
Temporal Needleman Wunsch.....	36
<i>Haider Syed and Amar Das</i>	
Exploiting Big Data in Time Series Forecasting: A Cross-Sectional Approach	45

Claudio Hartmann, Martin Hahmann, Frank Rosenthal and Wolfgang Lehner

Compression Rate Distance Measure for Time Series.....	55
--	----

Vinh T. Vo and Duong Tuan Anh

Classification

Multi-class Learning using Data Driven ECOC with Deep Search and Re-Balancing.....	65
--	----

Nathalie Japkowicz, Vincent Barnabe-Lortie, Shawn Horvatic and Jie Zhou

Calibration of One Class SVM for MV-set Estimation.....	75
---	----

Albert Thomas, Vincent Feuillard and Alexandre Gramfort

Improved Risk Predictions via Sparse Imputation of Patient Conditions in Electronic Medical Records.....	84
--	----

Budhaditya Saha, Sunil Kumar Gupta and Svetha Venkatesh

Label Noise	94
-------------------	----

Bryce Nicholson, Victor Sheng and Jing Zhang

Hierarchical Label Partitioning for Large Scale Classification.....	103
---	-----

Raphael Puget and Nicolas Baskiotis

Detecting Human Emotion via Speech Recognition by Using Speech Spectrogram.....	113
---	-----

Sathit Prasomphan

An Effective and Economic Bi-level Approach to Ranking and Rating Spam Detection	123
--	-----

Sihong Xie, Qingbo Hu, Jingyuan Zhang and Philip S Yu

Clustering

Multi-Objective Clustering Ensemble for High-Dimensional Data based on Strength Pareto Evolutionary Algorithm.....	133
--	-----

Abdul Wahid, Xiaoying Gao and Peter Andreae

Spherical Wards Clustering and Generalized Voronoi Diagrams.....	142
--	-----

Marek Śmieja and Jacek Tabor

A New Consensus Function based on Dual-Similarity Measurements for Clustering Ensemble.....	152
---	-----

Tahani Alqurashi and Wenjia Wang

Relational Learning

Relational Active Learning for Link-Based Classification.....	161
---	-----

Luke McDowell

Scalable Relational Learning for Large Heterogeneous Networks.....	171
--	-----

Ryan Rossi and Rong Zhou

Integrating Spatial Information into Probabilistic Relational Models.....	181
---	-----

Rajani Chulyadyo and Philippe Leray

Recommendation and Advertising

From One Star to Three Stars: Upgrading Legacy Open Data Using Crowdsourcing.....	189
<i>Satoshi Oyama, Yukino Baba, Ikki Ohmukai, Hiroaki Dokoshi and Hisashi Kashima</i>	
An Approach to Cover More Advertisers in Adwords.....	198
<i>Amar Budhiraja and P. Krishna Reddy</i>	
Learning From Missing Data Using Selection Bias in Movie Recommendation.....	208
<i>Claire Vernade and Olivier Cappé</i>	
An Accurate Rating Aggregation Method for Generating Item Reputation.....	217
<i>Ahmad Abdel-Hafez, Yue Xu and Audun Jøsang</i>	

Risk and Reliable Prediction

The NIST Data Science Program.....	225
<i>Bonnie J. Dorr, Craig Greenberg, Peter Fontana, Mark Przybocki, Marion Le Bras, Cathryn Ploehn, Oleg Aulov, Martial Michel, E. Jim Golden and Wo Chang</i>	
Predictive Reliability Mining for Early Warnings in Populations of Connected Machines.....	235
<i>Karamjit Singh, Gautam Shroff and Puneet Agarwal</i>	
Towards Predictable and Risk-Free Enterprise Systems.....	245
<i>Mayank Shrivastava, Maitreya Natu and Vaishali Sadaphal</i>	
Classification with a reject option under Concept Drift: the Droplets Algorithm.....	252
<i>Pierre-Xavier Loeffel, Christophe Marsala and Marcin Detyniecki</i>	

Network Analysis

The Layered Structure of Company Share Networks.....	261
<i>Andrea Romei, Salvatore Ruggieri and Franco Turini</i>	
A Time-Varying Graph Unifying Model	271
<i>Klaus Wehmuth, Artur Ziviani and Eric Fleury</i>	
Sensor Network Partitioning based on Homogeneity	281
<i>Yi Xu, Zhongfei Zhang, Yaqing Zhang and Philip S. Yu</i>	

Data Reduction and Factorization

Selecting Representative Instances from Datasets	291
<i>Hamid Mirisaei, Ahlame Douzal and Alexandre Termier</i>	
Improved Algorithms for Exact and Approximate Boolean Matrix Decomposition	301
<i>Yi Sun, Shiwei Ye, Yuan Sun and Tsunehiko Kameda</i>	
Modeling Recurrent Distributions in Streams Using Possible Worlds	311
<i>Michael Geilke, Andreas Karwath and Stefan Kramer</i>	

Manifolds and Metric Learning

Information Preserving and Locally Isometric&Conformal Embedding via Tangent Manifold Learning	320
<i>Alexander Bernstein, Alexander Kuleshov and Yury Yanovich</i>	
The HIM Glocal Metric and Kernel for Network Comparison and Classification	329
<i>Giuseppe Jurman, Roberto Visintainer, Michele Filosi, Samantha Riccadonna and Cesare Furlanello</i>	
Learning Hyperparameter Optimization Initializations	339
<i>Giuseppe Jurman, Roberto Visintainer, Michele Filosi, Martin Wistuba, Nicolas Schilling and Lars Schmidt-Thieme</i>	
Fast and User-friendly Non-linear Principal Manifold Learning by Method of Elastic Maps	349
<i>Andrei Zinovyev and Alexander Gorban</i>	
Efficient Metric Learning for the Analysis of Motion Data	358
<i>Babak Hosseini and Barbara Hammer</i>	
Modeling Temporal Dependencies in Data Using a DBN-LSTM	368
<i>Raunaq Vohra, Kratarth Goel and J. K. Sahoo</i>	

Behavioral Data Analysis

Behavioral Entropy and Profitability in Retail	372
<i>Riccardo Guidotti, Michele Coscia, Dino Pedreschi and Diego Pennacchioli</i>	
Discovering Characterizing Regions For Consumer Products	382
<i>Shashwat Mishra, Vincent Leroy and Sihem Amer-Yahia</i>	
The Harsh Rule of the Goals”: Big Data Analytics and Football Team Success	392
<i>Paolo Cintia, Luca Pappalardo, Dino Pedreschi, Fosca Giannotti and Marco Malvaldi</i>	
When Cyberathletes Conceal Their Game: Clustering Confusion Matrices to Identify Avatar Aliases	402
<i>Olivier Cavadenti, Victor Codocedo, Jean-Francois Boulicaut and Mehdi Kaytoue</i>	
Data Science Foundry for MOOCs	412
<i>Sebastien Boyer, Ben U. Gelman, Benjamin Schreck, Kalyan Veeramachaneni</i>	

Databases

TSMH Graph Cube: A Novel Framework for Large Scale Multi-Dimensional Network Analysis	422
<i>Pengsen Wang, Bin Wu and Bai Wang</i>	
A Mining Based Approach for Efficient Enumeration of Algebraic Structures	432
<i>Majid Khan, Nazeeruddin Mohammad, Shahabuddin Muhammad and Asif Ali</i>	
FACTORBASE: Multi-Relational Structure Learning with SQL All The Way	438
<i>Zhensong Qian and Oliver Schulte</i>	

Text Mining

A Text Clock Context Informations based Multiple Web Contents Extraction	448
<i>Wonmoon Song and Myungwon Kim</i>	
LDA Based Semi-supervised Learning from Streaming Short Text	456
<i>Ji-De Chen and Hung-Yu Kao</i>	
A Context-Aware Approach to Detection of Short Irrelevant Texts	464
<i>Sihong Xie, Jing Wang, Mohammad S Amin, Baoshi Yan, Anmol Bhasin, Philip S Yu and Clement T Yu</i>	
Document Similarity Analysis via Involving Both Explicit and Implicit Semantic Couplings.....	474
<i>Qianqian Chen, Liang Hu, Jia Xu, Wei Liu and Longbing Cao</i>	

Machine Learning

Learning Better while Sending Less: Communication-Efficient Online Semi-Supervised Learning in Client-Server Settings.....	484
<i>Han Xiao, Shou-De Lin, Mi-Yen Yeh, Phillip Gibbons and Claudia Eckert</i>	
EXP3 with Drift Detection for the Switching Bandit Problem	494
<i>Robin Allesiardo and Raphaël Féraud</i>	
Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces	501
<i>Frederic Cazals and Alix Lhéritier</i>	
Multitask Learning with Selective Cross-Task Transfer for Predicting Bleeding and other Important Patient Outcomes.....	511
<i>Che Ngufor, Sudhindra Upadhyaya, Dennis Murphree, Daryl Kor and Jyotishman Pathak</i>	

Pattern Mining

P-N-RMiner: A Generic Framework for Mining Interesting Structured Relational Patterns	519
<i>Jefrey Lijffijt, Eirini Spyropoulou and Tijl De Bie</i>	
Exploring Distributed RDF Resources using Formal Concept Analysis	529
<i>Mehwish Alam and Amedeo Napoli</i>	
Cohesion Based Co-location Pattern Mining	539
<i>Cheng Zhou, Boris Cule and Bart Goethals</i>	
Mining Actionable Combined Patterns of High Utility and Frequency.....	549
<i>Jingyu Shao, Junfu Yin, Wei Liu and Longbing Cao</i>	

Social Data Analysis

Nonparametric Discovery of Online Mental Health-Related Communities	559
<i>Bo Dao, Thin Nguyen, Svetha Venkatesh and Dinh Phung</i>	
Human Judgments In Hiring Decisions Based On Online Social Network Profiles	569
<i>Yoram Bachrach</i>	
A Model-Based Approach for Identifying Spammers in Social Networks	579

Farnoosh Fathaliani and Mohamed Bouguessa

Dynamics of Multi-Campaign Propagation in Online Social Networks588

Thejaswi M, Sriniketh Vijayaraghavan, Avinash Das and P Santhi Thilagam

Quantification in Social Networks596

Letizia Milli, Anna Monreale, Giulio Rossetti, Dino Pedreschi, Fosca Giannotti and Fabrizio Sebastiani

Discovering and Tracking Influencer-Influencee Relationships Between Online Communities606

Belkacem Chikhaoui, Mauricio Chiazzaro, and Shengrui Wang

Mobility Data Analysis

A Mixture Model Clustering Approach for Temporal Passenger Pattern Characterization in Public Transport615

Anne-Sarah Briand, Etienne Côme, Mohamed Khalil El Mahrsi and Latifa Oukhellou

Evaluating and Predicting Energy Consumption of Data Mining Algorithms on Mobile Devices625

Carmela Comito and Domenico Talia

Hermes_sem: a Semantic-aware Framework for the Management and Analysis of our Life Steps.....633

Nikos Pelekis, Stylianos Sideridis and Yannis Theodoridis

Scalable Extraction of Timeline Information from Road Traffic Data using MapReduce643

Ardi Imawan, Fadhilah Putri, Seonga An, Han-You Jeong and Joonho Kwon

IOHMM for Location Prediction with Missing Data651

Jiawei Hu, Yanfeng Wang and Ya Zhang

Random Walk Based Context-Aware Activity Recommendation for Location Based Social Networks661

Hakan Bagci and Pinar Karagoz

Finance, Market and Analytics

A Talent Management Tool using Propensity to Leave Analytics670

Karthikeyan Natesan Ramamurthy, Moninder Singh, Yichong Yu, Jessica Aspis, Matthew Iames, Michael Peran and Qin Held

Financial Crisis and Global Market Couplings680

Wei Cao, Yves Demazeau and Longbing Cao

Sentiment and Stock Market Volatility Predictive Modelling - A Hybrid Approach690

Rapheal Olaniyan, Daniel Stamate, Doina Logofătu and Lahcen Ouarbya

Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market700

Koosha Golmohammadi and Osmar Zaiane

Feature Analysis

An Unsupervised Attribute Clustering Algorithm for Unsupervised Feature Selection710

Pei-Yuan Zhou and Keith C.C. Chan

Deep Feature Synthesis: Towards Automating Data Science Endeavors	717
<i>James Max Kanter, Kalyan Veeramachaneni</i>	
Exploiting Feature Relationships Towards Stable Feature Selection	727
<i>Iman Kamkar, Sunil Gupta, Dinh Phung and Svetha Venkatesh</i>	
Context Weight Considered for Implicit Feature Extracting	737
<i>Jie Chen</i>	
Mining High-Utility Itemsets with Various Discount Strategies	742
<i>Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong and Vincent S. Tseng</i>	

DSAA Special Sessions

Bioinformatics, Health and Medical Analytics (BHMA)

Design of an NGS MicroRNA Predictor Using Multi-layer Hierarchical MapReduce Framework.....	752
<i>Ren-Hao Pan, Lin-Yu Tseng, I-En Liao, Chien-Lung Chan, K. Robert Lai and Kai-Biao Lin</i>	
An Ensemble of Machine Learning and Anti-learning Methods for Predicting Tumour Patient Survival Rates.....	760
<i>Chris Roadknight and Uwe Aickelin</i>	
Patient Classification based on Expanded Query using 5-gram Collocation and Binary Tree.....	768
<i>Jaya Sil and Indrani Bhattacharya</i>	
Improved Approach for Protein Function Prediction by Exploiting Prominent Proteins.....	778
<i>Satheesh Kumar and P. Krishna Reddy</i>	
MIAT: A Novel Attribute Selection Approach to Better Predict Upper Gastrointestinal Cancer	785
<i>Avi Rosenfeld</i>	
Modeling Heterogeneous Clinical Sequence Data in Semantic Space for Adverse Drug Event Detection	792
<i>Aron Henriksson, Jing Zhao, Henrik Boström and Hercules Dalianis</i>	
A Data-driven Analytics Approach in the Study of Pneumonia's Fatalities	800
<i>Maribel Yasmina Santos, António Carneiro, Artur Teles de Araújo</i>	
Cascading Adverse Drug Event Detection in Electronic Health Records	810
<i>Jing Zhao, Aron Henriksson and Henrik Bostrom</i>	
Characterizing Chronic Disease and Polymedication Prescription Patterns from Electronic Health Records.....	818
<i>Martí Zamora, Manel Baradad, Ester Amado, Sílvia Cordoní, Esther Limón, Juliana Ribera, Marta Arias and Ricard Gavalda</i>	
Ensemble of Deep Long Short Term Memory Networks for Labelling Origin of Replication Sequences	827
<i>Urminder Singh, Sucheta Chouhan, A Krishnamachari and Lovekesh Vig</i>	
Anomaly Detection in ECG Time Signals through Long Short-Term Memory based Recurrent Neural Network Architecture	834

Big Behavioral Data Analytics 1 (BBDA)

Predicting Online Video Engagement Using Clickstreams	841
<i>Everaldo Aguiar, Saurabh Nagrecha and Nitesh Chawla</i>	
Profit Maximizing Logistic Regression Modeling for Customer Churn Prediction	851
<i>Eugen Stripling, Seppe Vanden Broucke, Katrien Antonio, Bart Baesens and Monique Snoeck</i>	
A Fusion-based Approach for User Activities Recognition on Smart Phones	861
<i>Gunarto Sindoro Njoo, Xiao Wen Ruan, Kuo-Wei Hsu and Wen-Chih Peng</i>	
Tracking the Evolution of Community Structures in Time-Evolving Social Networks	871
<i>Etienne Gael Tajeuna, Mohamed Bouguessa and Shengrui Wang</i>	
Mining Relationships in Learning-oriented Social Networks	881
<i>María Estrella Sousa Vieira, Jose Carlos López Ardao, Manuel Fernández Veiga, Miguel Rodríguez Pérez and Cándido López García</i>	

Environmental and Geo-spatial Data Analytics (EnGeoData) 1

Constrained Spectral Clustering for Regionalization: Exploring the Trade-off between Spatial Contiguity and Landscape Homogeneity	890
<i>Shuai Yuan, Pang-Ning Tan, Kendra Spence Cheruvilil, Sarah M. Collins and Patricia A. Soranno</i>	
Multifactorial Uncertainty Assessment for Monitoring Population Abundance using Computer Vision	900
<i>Emma Beauxis-Aussalet and Lynda Hardman</i>	
Detection, Tracking, and Visualization of Spatial Event Clusters for Real Time Monitoring	910
<i>Natalia Andrienko, Gennady Andrienko, Georg Fuchs, Salvatore Rinzivillo and Hans-Dieter Betz</i>	
Big Data from Cellular Networks: How to Estimate Energy Demand at real-time	920
<i>Davide Tosi, Stefano Marzorati, Mario La Rosa, Giovanna Dondossola and Roberta Terruggia</i>	

Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis (SentISData)

Sentiment Analysis as a Text Categorization Task: A Study on Feature and Algorithm Selection for Italian Language.....	930
<i>Berardina Nadja De Carolis, Stefano Ferilli, Domenico Redavid and Floriana Esposito</i>	
Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies.....	940
<i>Sudipto Shankar Dasgupta, Swaminathan Natarajan, Kiran Kumar Kaipa, Sujay Bhattacharjee, Arun Viswanathan and Sairam Yeturi.</i>	
Debate on Political Reforms in Twitter: A Hashtag-driven Analysis of Political Polarization	943
<i>Cristina Bosco, Mirko Lai, Patti Viviana and Daniela Virone</i>	
Hot Spot Detection - an Interactive Cluster Heat Map for Sentiment Analysis	952
<i>Patrick Hennig, Philipp Berger, Maximilian Brehm, Bastien Grasnack, Jonathan Herdt and Christoph Meinel</i>	

Monitoring the Twitter sentiment during the Bulgarian elections	961
<i>Jasmina Smailovic, Janez Kranjc, Miha Grcar, Martin Znidarsic and Igor Mozetic</i>	
Using Emotions to Predict User Interest Areas in Online Social Networks	971
<i>Yoad Lewenberg, Yoram Bachrach and Svitlana Volkova</i>	
Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers	981
<i>Elisabetta Fersini, Federico Alberto Pozzi and Enza Messina</i>	
Environmental and Geo-spatial Data Analytics (EnGeoData) 2	
Traffic Risk Mining from Heterogeneous Road Statistics	989
<i>Koichi Moriya, Shin Matsushima and Kenji Yamanishi</i>	
Learning Urban Users' Choices to Improve Trip Recommendations	999
<i>Boris Chidlovskii</i>	
RIMM: A Novel Map Matching Model With Rotational Invariance	1008
<i>Junpeng Bao, Yuepeng Zhang, Qian Cao, Jun Zeng and De Zhang</i>	
Assistive Classification for Improving the Efficiency of Avian Species Richness Surveys	1015
<i>Liang Zhang, Michael Towsey, Phil Eichinski, Jinglan Zhang and Paul Roe</i>	
Big Data, Distributed Technologies and Intelligent Agents (BDIA)	
Data-driven Semantic Concept Analysis for Automatic Actionable Ontology Design.....	1021
<i>Vladimir Gorodetsky and Olga Tushkanova</i>	
MapReduce-based K-Prototypes Clustering Method for Big Data	1030
<i>Mohamed Aymen Ben Hajkacem, Chiheb Eddine Ben N'cir and Nadia Essoussi</i>	
Cluster-Based Data Oriented Hashing	1037
<i>Sanaa Chafik, Imane Daoudi, Mounim A. El Yacoubi and Hamid El Ouardi</i>	
Data Oriented Constructive Mining and Multi-Agent Simulation (DOCMAS)	
Error Detection of Oceanic Observation Data Using Sequential Labeling.....	1044
<i>Satoshi Ono, Haruki Matsuyama, Ken-Ichi Fukui and Shigeki Hosoda</i>	
Statistical and Mathematical Tools for Data Science (SMTDS)	
Constrained Independence for Detecting Interesting Patterns	1052
<i>Thomas Delacroix, AHCÈNE BOUBEKKI, Philippe Lenca and Stéphane Lallich</i>	
A Ranking-based approach for Hierarchical Classification	1062
<i>Azad Naik and Huzefa Rangwala</i>	
A Combination of CUSUM-EWMA for Anomaly Detection in Time Series Data	1072
<i>Vyron Christodoulou and Yaxin Bi</i>	

Multi-Target Regression from High-Speed Data Streams with Adaptive Model Rules	1080
<i>João Duarte and João Gama</i>	
Scalable Image Annotation Using a Product Compressive Sampling Approach	1090
<i>Anastasios Maronidis, Elisavet Chatzilari, Spiros Nikolopoulos and Ioannis Kompatsiaris</i>	
Graph-based Semi-supervised Learning for Time Series Classification	1100
<i>Zhao Xu and Koichi Funaya</i>	
 Exploratory Computing (EC)	
IQ4EC: Intensional Answers as a Support to Exploratory Computing	1106
<i>Mirjana Mazuran, Elisa Quintarelli and Letizia Tanca</i>	
TESS: Temporal Event Sequence Summarization	1116
<i>Dominique Gay, Romain Guigourès, Marc Boullé and Fabrice Clérot</i>	
Succinctly Summarizing Machine Usage via Multi-subspace Clustering of Multi-sensor Data	1126
<i>Sarmimala Saikia, Gautam Shroff, Puneet Agarwal and Ashwin Srinivasan</i>	