

Communications in Computer and Information Science

1099

Commenced Publication in 2007

Founding and Former Series Editors:

Simone Diniz Junqueira Barbosa, Phoebe Chen, Alfredo Cuzzocrea,
Xiaoyong Du, Orhun Kara, Ting Liu, Krishna M. Sivalingam,
Dominik Ślęzak, Takashi Washio, Xiaokang Yang, and Junsong Yuan

Editorial Board Members

Joaquim Filipe 


Polytechnic Institute of Setúbal, Setúbal, Portugal

Ashish Ghosh

Indian Statistical Institute, Kolkata, India

Igor Kotenko 

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Raquel Oliveira Prates 

Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

Lizhu Zhou

Tsinghua University, Beijing, China


More information about this series at <http://www.springer.com/series/7899>

Henry Han · Tie Wei · Wenbin Liu ·
Fei Han (Eds.)

Recent Advances in Data Science

Third International Conference
on Data Science, Medicine, and Bioinformatics, IDMB 2019
Nanning, China, June 22–24, 2019
Revised Selected Papers

Editors

Henry Han 
Fordham University
New York, NY, USA

Wenbin Liu
Guangzhou University
Guangzhou, China

Tie Wei
Guangxi University
Nanning, China

Fei Han
Jiangsu University
Zhenjiang, China

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-981-15-8759-7 ISBN 978-981-15-8760-3 (eBook)
<https://doi.org/10.1007/978-981-15-8760-3>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

With the surge of big data and massive data, data science becomes a new exciting interdisciplinary field that is bringing challenges and opportunities to science, engineering, health, and business. The huge amount of data provide more available information and demand for more complicated and explainable knowledge discovery methods to decipher the information. For example, high-frequency trading data from algorithmic trading needs special algorithms and computing approaches to disclose the latent information encoded in the structured big data.

Data science is seeing the trend of problem solving from model-driven to data-driven. The model-driven was widely used in almost all fields because limited data is available. Researchers have to build different models from a theoretical standing point and apply it to ‘fit’ the limited data. It is likely that the models can have unrealistic assumptions that do not match real data well in modeling. Usually one or few types of data are considered in the models. Thus, the model-driven approaches may need to build different models to handle different types of data. As a result, they can have a hard time generalizing one result to another. Furthermore, the model-driven approaches cannot take advantage of a large amount of available data mainly because the modeling process is somewhat independent from data. On the other hand, the data-driven approaches derive and develop customized but more generalized models or algorithms from data. The data-driven approaches take advantage of the large amount of data and let data talk in decision making. In other words, data-driven approaches are driven and adjusted by data dynamically. The state-of-the-art artificial intelligence (AI) approaches (e.g. machine learning) can be viewed as a typical data-driven approach in data science to exploit the large amount of data.

The future of data science relies on how well AI methods are developed for different subfields. It is expected that different machine learning and deep learning methods are developed for specific type of data. For example, financial data science need their own deep learning methods to dig knowledge from financial data that generally has few independent variables but a huge amount of observations from time series. On the other hand, biological and health data science may want to derive the AI models that work well for high-dimensional data. Quite a lot of pioneering studies have been conducted in biological data science fields in recent years while other data science fields are starting to catch up.

This volume serves the proceeding of the International Conference on Data science, Medicine, and Bioinformatics (IDMB 2019), Nanning, China. It aims to report recent advances in data science, business data science, health and biological data science, and other data science fields. It will be a good guide for data scientists and practitioners in the fields. IDMB 2019 only accepted 35 high-quality papers among 93 submissions through a very rigorous reviewing process. There were 15 papers among the accepted papers recommended to the 3 IDMB 2019 associative SCI-indexed journals: BMC Bioinformatics, Computational Biology & Chemistry, and the Chinese Electronic

Journal, and published in two special journal issues. The volume consists of 20 cutting-edge papers devoted to state-of-the-art data science research in business data science, biological and health data science, and novel data science theory and applications.

We thank all authors, committee members, and session chairs involved in IDMB 2019, as well as folks that kindly provided their support to IDMB 2019. In particular, we thank the business school of Guangxi University and National Science Foundation of China (No. 71562001) for their valuable support alongside Dr. Tie Wei's great leadership.

Henry Han
Tie Wei
Wenbin Liu
Fei Han

Organization

IDMB 2019 Chairs

General Chair

Henry Han Fordham University, USA

Co-chair

Tie Wei Guangxi University, China

IDMB 2019 Steering Committee

Henry Han	Fordham University, USA
Fei Han	Jiangsu University, China
Wentian Li	Feinstein Institute for Medical Research, USA
Xiangrong Liu	Xiamen University, China
Wenbin Liu	Guangzhou University, China
Ke Men	Xi'an Medical University, China

IDMB 2019 Program Committee

Yongsheng Bai	University of Michigan, USA
Yaodong Chen	Northwest University, China
Jeff Forest	Slippery Rock University, USA
Qiannong Gu	Ball State University, USA
Deqing Li	University of Maryland, USA
Junyi Li	Harbin Institute of Technology, China
Zhen Li	Texas Woman's University, USA
Zhou Ji	Columbia University, USA
Ye Tian	Case Western Reserve University, USA
Guimin Qin	Xidian University, China
Chaoyong Qin	Guangxi University, China
Stellar Tao	UT Health Center, USA
Hongling Wang	ACT, USA
John Wu	Nanjing Tech University, China
Xiaoguang Tian	Purdue University, USA
Zhenning Xu	University of Southern Maine, USA
Yan Yu	St John's University, USA
Honggang Zhang	University of Massachusetts, USA
Jeff Zhang	California State University, USA
Yanbang Zhang	Northwestern Polytechnical University, China

IDMB 2019 Reviewers

Wei Chen	Fordham University, USA
Zhihua Chen	Guangzhou University, China
Gang Fang	Guangzhou University, China
Chunmei Feng	Harbin Institute of Technology, China
Heming Huang	Qinghai Normal University, China
Tao Huang	Chinese Academy of Sciences, China
Jing Jiang	Jiangsu University, China
Zheng Kou	Guangzhou University, China
Fangjun Kuang	Wenzhou business College, China
Chun Li	Hannai Normal University, China
Dongdong Li	South China University of Technology, China
Yang Li	Shandong University, China
Yihao Li	Fordham University, USA
Jinxing Liu	Qufu Normal University, China
Nan Mi	Texas Tech University, USA
Baoli Peng	The Chinese University of Hong Kong, China
Xianya Qin	Fordham University, USA
Junliang Shang	Qufu Normal University, China
Xuequn Shang	Northwestern Polytechnical University, China
Yuan Shang	University of Arizona, USA
Xiaohong Shi	Xi'an Technological University, China
Yansen Su	Anhui University, China
Juan Wang	Guangxi University, China
Juan Wang	Qufu Normal University, China
Xinzeng Wang	Shandong University of Science and Technology, China
Zhuo Wang	Shanghai Jiaotong University, China
Yi Wu	Georgia Institute of Technology, USA
Juanying Xie	Shaanxi Normal University, China
Jie Xu	St John's University, USA
Peng Xu	Guangzhou University, China
Tingyun Yan	Fordham University, USA
Xuemei Yang	Xianyang Normal University, China
Lianping Yang	Northeastern University, China
Shengli Zhang	Xidian University, China
Liang Zhao	University of Michigan, USA

Contents

Business Data Science: Fintech, Management, and Analytics

Locally Linear Embedding for High-Frequency Trading Marker Discovery. . .	3
<i>Henry Han, Jie Teng, Junruo Xia, Yunhan Wang, Zihao Guo, and Deqing Li</i>	
Implied Volatility Pricing with Selective Learning.	18
<i>Henry Han, Haofeng Huang, Jiayin Hu, and Fangjun Kuang</i>	
A High-Performance Basketball Game Forecast Using Magic Feature Extraction.	35
<i>Tiange Li and Henry Han</i>	
A Study on the Effect of Flexible Human Resource Practice on Group Citizenship Behavior.	51
<i>Ying Huang, Ziyao Li, and Jing Wang</i>	
Patent Text Mining on Technology Evolution Path in 5G Communication Field	65
<i>Tingting Liu, Tie Wei, Henry Han, and Tiange Li</i>	
Research on Advertising Click-Through Rate Prediction Model Based on Ensemble Learning.	82
<i>Xiaojuan He, Wenjie Pan, and Hong Cheng</i>	

Health and Biological Data Science

An Integrated Robust Graph Regularized Non-negative Matrix Factorization for Multi-dimensional Genomic Data Analysis.	97
<i>Yong-Jing Hao, Mi-Xiao Hou, Rong Zhu, and Jin-Xing Liu</i>	
Hyper-graph Robust Non-negative Matrix Factorization Method for Cancer Sample Clustering and Feature Selection	112
<i>Cui-Na Jiao, Tian-Ru Wu, Jin-Xing Liu, and Xiang-Zhen Kong</i>	
Association of PvuII and XbaI Polymorphisms in Estrogen Receptor Alpha (ESR1) Gene with the Chronic Hepatitis B Virus Infection.	126
<i>Ke Men, Wen Ren, Xia Wang, Tianjian Men, Ping Li, Kejun Ma, and Mengyan Gao</i>	

Inferring Communities and Key Genes of Triple Negative Breast Cancer Based on Robust Principal Component Analysis and Network Analysis	137
<i>Qian Ding, Yan Sun, Junliang Shang, Yuanyuan Zhang, Feng Li, and Jin-Xing Liu</i>	
A Novel Multiple Sequence Alignment Algorithm Based on Artificial Bee Colony and Particle Swarm Optimization.	152
<i>Fangjun Kuang and Siyang Zhang</i>	
A Flexible and Comprehensive Platform for Analyzing Gene Expression Data	170
<i>Bolin Chen, Chenfei Wang, Li Gao, and Xuequn Shang</i>	
Classification of Liver Cancer Images Based on Deep Learning	184
<i>Hui Ye, Qiaojun Chen, Haimei Wu, and Dong Cao</i>	
DSNPCMF: Predicting MiRNA-Disease Associations with Collaborative Matrix Factorization Based on Double Sparse and Nearest Profile	196
<i>Meng-Meng Yin, Zhen Cui, Jin-Xing Liu, Ying-Lian Gao, and Xiang-Zhen Kong</i>	
Novel Data Science Theory and Applications	
Exploring Blockchain in Speech Recognition	211
<i>Xuemei Yang and Heming Huang</i>	
Automatic Segmentation of Mitochondria from EM Images via Hierarchical Context Forest	221
<i>Jiajin Yi, Zhimin Yuan, and Jialin Peng</i>	
Generating the n -tuples of Natural Numbers by Enzymatic Numerical P System	234
<i>Zhiqiang Zhang, Yuan Kong, Zhihua Chen, Yunyun Niu, and Mingyuan Ma</i>	
DNA Self-assembly Computing Model for the Course Timetabling Problem	245
<i>Zheng Kou, Zhibao Xing, Wenfei Lan, and Xiaoli Qiang</i>	
Optimization of GenoCAD Design Based on AMMAS	254
<i>Yingjie Wang and Yafei Dong</i>	

Node Isolated Strategy Based on Network Performance Gain Function:
Security Defense Trade-Off Strategy Between Information Transmission
and Information Security

272

*Gang Wang, Shiwei Lu, Yun Feng, Wenbin Liu,
and Runnian Ma*

Author Index

287