Intelligent Systems Reference Library

Volume 149

Series editors

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Lakhmi C. Jain, Faculty of Engineering and Information Technology, Centre for Artificial Intelligence, University of Technology, Sydney, NSW, Australia;
Faculty of Science, Technology and Mathematics, University of Canberra, Canberra, ACT, Australia;
KES International, Shoreham-by-Sea, UK
e-mail: jainlakhmi@gmail.com; jainlc2002@yahoo.co.uk

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

More information about this series at http://www.springer.com/series/8578

George A. Tsihrintzis · Dionisios N. Sotiropoulos
Lakhmi C. Jain
Editors

Machine Learning Paradigms

Advances in Data Analytics

Springer

Editors
George A. Tsihrintzis
University of Piraeus
Piraeus
Greece

Dionisios N. Sotiropoulos
University of Piraeus
Piraeus
Greece

Lakhmi C. Jain
Faculty of Engineering and Information Technology, Centre for Artificial Intelligence
University of Technology
Sydney, NSW
Australia

and

Faculty of Science, Technology and Mathematics
University of Canberra
Canberra, ACT
Australia

and

KES International
Shoreham-by-Sea
UK

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-319-94029-8 ISBN 978-3-319-94030-4 (eBook)
https://doi.org/10.1007/978-3-319-94030-4

Library of Congress Control Number: 2018944351

© Springer International Publishing AG, part of Springer Nature 2019
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, express 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.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my wife and colleague, Prof.-Dr. Maria Virvou, and to our daughters, Evina, Konstantina and Andreani

George A. Tsihrintzis

To my beloved family and friends

Dionisios N. Sotiropoulos

To my beloved family

Lakhmi C. Jain

Foreword

In 1959, Arthur Samuel (1901–1990) published *Some Studies in Machine Learning Using the Game of Checkers* [1]. The paper was one of the earliest uses of the words "machine learning" [2]. He wrote, "As a result of these experiments one can say with some certainty that it is now possible to devise learning schemes which will greatly outperform an average person and that such learning schemes may eventually be economically feasible as applied to real-life problems" [1, p. 548]. His program together with IBM's first stored-program computer, the 701, demonstrated this statement by winning a game of checkers against a human expert in Connecticut. Since that time, games have provided fertile research ground for artificial intelligence, and in 1996, the Chinook Project for checkers was recognized by the Guinness Book of World Records as the first computer program to win a human world championship [3]. The day to apply machine learning to challenging real-world problems is here and now.

What is "machine learning"? Several suggested definitions are discussed on the IBM community site [4], including "The purpose of machine learning is to learn from training data in order to make as good as possible predictions on new, unseen, data." This definition suggests some of the challenges with machine learning. The program needs to build a model based on training data that includes the correct answer (i.e., supervised learning) and that minimizes the error in predicting new data. Alternatively, algorithms may look for structure in the data and group similar clusters (i.e., unsupervised learning). Too closely mirroring the training data results in overfitting and poor results with unknown data, and too little fitting results in unacceptable errors in predictions. In addition, as updated known data become available, the model may need to be re-adjusted to retain generalizability. Thus, the methods used in machine learning are constantly being researched and assessed against real-life data from various fields along with the computer technologies needed to implement them.

This book applies and assesses machine learning for classes of important real-life problems, an area often referred to as "data analytics." Authors have contributed leading research in the areas of medical, biological and signal sciences; social studies and social interactions; traffic, computer and power networks; and digital forensics. The book also looks to the future for research areas that may yield theoretical advances. The editors have provided a valuable and much-needed collection of leading research in machine learning and data analytics that will increasingly impact each of us in our everyday lives. New and experienced researchers, practitioners, and those interested in machine learning will be inspired by the innovative ideas contained in its pages.

Baltimore, USA

Gloria Phillips-Wren, Ph.D.
Professor, Loyola University Maryland

References

1. Samuel, A.: Some studies in machine learning using the game of checkers. IBM J. 3(3), 210–229 (1959)
2. McCarthy, J., Feigenbaum, E.: In memoriam—Arthur Samuel: Pioneer in machine learning. AI Mag. 11(3), 10–11 (1990)
3. Chinook,: Arthur Samuel's legacy. Accessed on 28 May 2018, from https://webdocs.cs.ualberta.ca/~chinook/project/legacy.html (2018)
4. Puget, J-C.: What is machine learning? IBM Community, May 18 from https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Is_Machine_Learning?lang=en (2016). Accessed on 28 May 2018

Preface

At the dawn of the fourth Industrial Revolution, *data analytics* is emerging as a force that drives towards dramatic changes in our daily lives, the workplace and human relations. Synergies between physical, digital, biological and energy sciences and technologies, sewn together by *non-traditional data collection and analysis*, drive the digital economy at all levels and offer new, previously unavailable opportunities.

The need for data analytics arises in most modern scientific disciplines, including engineering, natural, computer and information sciences, economics, business, commerce, environment, healthcare and life sciences. The book at hand explores some of the emerging scientific and technological areas in which data analytics arises as a need and, thus, may play a significant role in the years to come.

Coming as the third volume under the general title *Machine Learning Paradigms* and following two related monographs, the book includes an editorial note (Chap. 1) and an additional twelve (12) chapters and is divided into five parts, namely: (1) *Data Analytics in the Medical, Biological and Signal Sciences*, (2) *Data Analytics in Social Studies and Social Interactions*, (3) *Data Analytics in Traffic, Computer and Power Networks*, (4) *Data Analytics for Digital Forensics* and (5) *Theoretical Advances and Tools for Data Analytics*.

This research book is directed towards professors, researchers, scientists, engineers and students of all disciplines. We hope that they all will find it useful in their works and researches.

We are grateful to the authors and the reviewers for their excellent contributions and visionary ideas. We are also thankful to Springer for agreeing to publish this book. Last, but not least, we are grateful to the Springer staff for their excellent work in producing this book.

Piraeus, Greece	George A. Tsihrintzis
Piraeus, Greece	Dionisios N. Sotiropoulos
Sydney, Australia/Canberra, Australia	Lakhmi C. Jain

Contents

1 Machine Learning Paradigms: Advances in Data Analytics 1
George A. Tsihrintzis, Dionisios N. Sotiropoulos and Lakhmi C. Jain
Bibliography 4

Part I Data Analytics in the Medical, Biological and Signal Sciences

2 A Recommender System of Medical Reports Leveraging Cognitive Computing and Frame Semantics 7
Danilo Dessì, Diego Reforgiato Recupero, Gianni Fenu and Sergio Consoli
2.1 Introduction 8
2.2 State of the Art 10
2.2.1 Biomedical Information Retrieval 10
2.2.2 Biomedical Classification 11
2.2.3 Biomedical Clustering 12
2.2.4 Biomedical Recommendation 13
2.2.5 Cognitive Computing and IBM Watson 13
2.2.6 Frame Semantics and Framester 14
2.3 Architecture of Our System 14
2.3.1 Content Analyzer Module 15
2.3.2 Machine Learning Module 19
2.3.3 Recommendation Module 21
2.4 Experiments 21
2.4.1 The Test Dataset 21
2.4.2 Experiment Setup 22
2.4.3 Recommendation Module Setup 24
2.4.4 Results 24
2.5 Conclusion and Future Trends 28
References 29

3 Classification Methods in Image Analysis with a Special Focus on Medical Analytics 31
Lucio Amelio and Alessia Amelio
3.1 Introduction 31
3.2 Background 33
3.3 Feature Representation for Image Classification 36
3.3.1 Global Features 36
3.3.2 Local Features 39
3.3.3 Bag of Visual Words 40
3.3.4 Pixel-Level Features 40
3.4 Security and Biometrics 41
3.4.1 Supervised Classification 42
3.5 Aerospace and Satellite Monitoring 43
3.5.1 Supervised Classification 44
3.5.2 Unsupervised Classification 45
3.6 Document Analysis and Language Understanding 46
3.6.1 Supervised Classification 47
3.6.2 Unsupervised Classification 49
3.7 Information Retrieval 51
3.7.1 Supervised Classification 51
3.7.2 Unsupervised Classification 51
3.8 Classification in Image-Based Medical Analytics 52
3.8.1 Diagnostic Inspective Acquisition Imaging 53
3.8.2 Nuclear Medicine Imaging 55
3.8.3 Clinical Radiology Imaging 58
3.8.4 Horizon of the Research and Future Challenges 61
3.9 Conclusions 62
3.10 Further Readings 63
References 63

4 Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field 71
Carine Bou Rjeily, Georges Badr, Amir Hajjarm El Hassani and Emmanuel Andres
4.1 Introduction 72
4.2 Classical Data Mining Technics and Heart Diseases 73
4.2.1 Popular Data Mining Algorithms 74
4.2.2 Data Mining and Heart Diseases 76
4.3 Sequential Mining in Medical Domain 79
4.4 Sequential Mining 80
4.4.1 Important Terms and Notations 81
4.4.2 Sequential Patterns Mining 83
4.4.3 General and Specific Techniques Used by SPM Algorithms 84

4.4.4 Extensions of Sequential Pattern Mining Algorithms 88
4.5 Discussion 94
4.6 Conclusion 94
References 95

5 Machine Learning Methods for the Protein Fold Recognition Problem 101
Katarzyna Stapor, Irena Roterman-Konieczna and Piotr Fabian
5.1 Introduction 101
5.2 Supervised Learning 104
5.3 Deep Learning Methods in Pattern Recognition 109
5.4 Features of the Amino Acid Sequence 110
5.5 Protein Fold Machine Learning-Based Classification Methods 117
5.5.1 Datasets Used in the Described Experiments 117
5.5.2 Methods 117
5.6 Discussion, Conclusions and Future Work 121
References 124

6 Speech Analytics Based on Machine Learning 129
Grazina Korvel, Adam Kurowski, Bozena Kostek and Andrzej Czyzewski
6.1 Introduction 130
6.2 Speech Phoneme Signal Analysis 130
6.3 Speech Signal Pre-processing 132
6.4 Speech Information Retrieval Scheme 133
6.5 Feature Extraction 135
6.5.1 Time Domain Features 135
6.5.2 Frequency Domain Features 136
6.5.3 Mel-Frequency Cepstral Coefficients 142
6.6 Data Preparation for Deep Learning 146
6.7 Experiment Results 149
6.7.1 Feature Vector Applied to Vowel Classification 149
6.7.2 Feature Vector Applied to Allophone Classification 150
6.7.3 Convolutional Neural Networks Applied to Allophone Classification 151
6.7.4 Convolutional Neural Networks Applied to Vowel Classification 152
6.8 Conclusions 153
References 155

Part II Data Analytics in Social Studies and Social Interactions

7 Trends on Sentiment Analysis over Social Networks: Pre-processing Ramifications, Stand-Alone Classifiers and Ensemble Averaging 161
Christos Troussas, Akrivi Krouska and Maria Virvou
7.1 Introduction 162
7.2 Research Methodology 164
7.3 Twitter Datasets 165
7.4 Evaluation of Data Preprocessing Techniques 167
7.5 Evaluation of Stand-Alone Classifiers 172
7.6 Evaluation of Ensemble Classifiers 175
7.7 The Case of Sentiment Analysis in Social e-Learning 180
7.8 Conclusions and Future Work 185
References 185

8 Finding a Healthy Equilibrium of Geo-demographic Segments for a Telecom Business: Who Are Malicious Hot-Spotters? 187
J. Sidorova, O. Rosander, L. Skold, H. Grahn and L. Lundberg
8.1 Introduction 187
8.2 Geospatial and Geo-demographic Data 189
8.3 The Combinatorial Optimization Module 190
8.4 Infrastructure-Friendly and Stressing Clients 192
8.5 Experiments 194
8.6 Conclusions 195
References 195

Part III Data Analytics in Traffic, Computer and Power Networks

9 Advanced Parametric Methods for Short-Term Traffic Forecasting in the Era of Big Data 199
George A. Gravvanis, Athanasios I. Salamanis and Christos K. Filelis-Papadopoulos
9.1 Introduction 200
9.2 Traffic Data 202
9.2.1 Traffic Network 202
9.2.2 Traffic Descriptors 203
9.2.3 Traffic Data Sources 205
9.3 Traffic Data Preprocessing 207
9.3.1 Time Series Formulation 207
9.3.2 Outlier Detection 209
9.3.3 Missing Data Imputation 211
9.3.4 Map-Matching 213
9.4 Parametric Short-Term Traffic Forecasting 214

9.4.1 Autoregressive Moving Average (ARMA) 214
9.4.2 Autoregressive Integrated Moving Average (ARIMA) 215
9.4.3 Space-Time ARIMA (STARIMA) 221
9.4.4 Lag-STARIMA 223
9.4.5 Graph-Based Lag-STARIMA (GBLS) 225
References 229

10 Network Traffic Analytics for Internet Service Providers—Application in Early Prediction of DDoS Attacks 233
Apostolos P. Leros and Antonios S. Andreatos
10.1 Introduction 234
10.2 The Procedure Adopted 235
10.2.1 Related Work 236
10.3 The Proposed Approach 237
10.3.1 Mathematical Formulation 238
10.3.2 State Space Model—Autoregressive Model—Discrete-Time Kalman Filter 239
10.4 Structure and Parameters of the MRSP Algorithm 241
10.5 Results and Performance of the MRSP Algorithm 244
10.6 Detecting Anomalies 256
10.6.1 DETECTING a DDoS ATTACK 256
10.6.2 Detecting an Anomaly 262
10.6.3 Final Remarks 265
10.7 Conclusions 266
References 266

11 Intelligent Data Analysis in Electric Power Engineering Applications 269
V. P. Androvitsaneas, K. Boulas and G. D. Dounias
11.1 Introduction 270
11.2 Intelligent Techniques in Ground Resistance Estimation 271
11.2.1 Grounding Systems 272
11.2.2 Application of ANN Methodologies for the Estimation of Ground Resistance 275
11.2.3 Wavelet Networks Modeling for the Estimation of Ground Resistance 288
11.2.4 Inductive Machine Learning 291
11.2.5 Genetic and Gene Expression Programming Versus Linear Regression Models 295
11.3 Estimation of Critical Flashover Voltage of Insulators 297
11.3.1 Problem Description 297
11.3.2 Genetic Algorithms 300
11.3.3 Application of ANNs 302

11.3.4 Multilayer Perceptron ANNs 304
11.3.5 Genetic Programming 305
11.3.6 Gravitational Search Algorithm Technique 306
11.4 Other Applications of Electric Power Systems 307
11.4.1 Load Forecasting 308
11.4.2 Lightning Performance Evaluation in Transmission Lines 308
11.5 Conclusions and Further Research 309
References 310

Part IV Data Analytics for Digital Forensics

12 Combining Genetic Algorithms and Neural Networks for File Forgery Detection 317
Konstantinos Karampidis, Ioannis Deligiannis and Giorgos Papadourakis
12.1 Introduction 317
12.1.1 McKemmish Predominant Model 318
12.1.2 Kent Predominant Model 318
12.1.3 Digital Evidences 319
12.1.4 File Type Identification. 321
12.2 Methodology of the Proposed Method 324
12.3 Experimental Setup and Results. 326
12.4 Conclusions 333
References 333

Part V Theoretical Advances and Tools for Data Analytics

13 Deep Learning Analytics 339
Nikolaos Passalis and Anastasios Tefas
13.1 Introduction 340
13.2 Preliminaries and Notation 341
13.3 Unsupervised Learning 342
13.3.1 Deep Autoencoders 343
13.3.2 Autoencoder Variants 350
13.4 Supervised Learning 352
13.4.1 Multilayer Perceptrons 354
13.4.2 Convolutional Neural Networks 359
13.4.3 Recurrent Neural Networks 362
13.5 Deep Learning Frameworks 364
13.6 Concluding Remarks 366
References 366