Lecture Notes in Artificial Intelligence 10785

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel University of Alberta, Edmonton, Canada Yuzuru Tanaka Hokkaido University, Sapporo, Japan Wolfgang Wahlster DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann DFKI and Saarland University, Saarbrücken, Germany More information about this series at http://www.springer.com/series/1244

Annalisa Appice · Corrado Loglisci Giuseppe Manco · Elio Masciari Zbigniew W. Ras (Eds.)

New Frontiers in Mining Complex Patterns

6th International Workshop, NFMCP 2017 Held in Conjunction with ECML-PKDD 2017 Skopje, Macedonia, September 18–22, 2017 Revised Selected Papers



Editors Annalisa Appice University of Bari Aldo Moro Bari Italy

Corrado Loglisci University of Bari Aldo Moro Bari Italy

Giuseppe Manco CNR Rende Italy Elio Masciari CNR Rende Italy Zbigniew W. Ras University of North Carolina Charlotte, NC USA and Polish Japanese Academy of Information Technology Warsaw Poland

ISSN 0302-9743 ISSN 1611-3349 (electronic) Lecture Notes in Artificial Intelligence ISBN 978-3-319-78679-7 ISBN 978-3-319-78680-3 (eBook) https://doi.org/10.1007/978-3-319-78680-3

Library of Congress Control Number: 2018937389

LNCS Sublibrary: SL7 - Artificial Intelligence

© Springer International Publishing AG, part of Springer Nature 2018

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

Preface

Modern automatic systems are able to collect huge volumes of data, often with a complex structure (e.g., multi-table data, network data, Web data, time series and sequences, trees and hierarchies). Massive and complex data pose new challenges for current research in data mining. Specifically, they require new models and methods for their storage, management, and analysis, in order to deal with the following complexity factors:

- Data with a complex structure (e.g., multi-relational, time series and sequences, networks, and trees) as input or output of the data mining process
- Data collections with many examples and/or many dimensions, where data may be processed in (near) real time
- Partially labeled data
- Data which arrive continuously as a stream, at high rate, subject to concept drift

The 6th International Workshop on New Frontiers in Mining Complex Patterns (NFMCP 2017) was held in Skopje (Macedonia) in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2017) on September 22, 2017. The purpose of this workshop was to bring together researchers and practitioners of data mining who are interested in the latest developments in the analysis of complex and massive data sources, such as blogs, event or log data, medical data, spatiotemporal data, social networks, mobility data, sensor data, and streams. The workshop was aimed at discussing and introducing new algorithmic foundations and representation formalisms in complex pattern discovery. Finally, it encouraged the integration of recent results from existing fields, such as statistics, machine learning, and big data analytics. This book features a collection of revised and significantly extended versions of papers accepted for presentation at the workshop. These papers went through a rigorous review process to ensure compliance with Springer's high-quality publication standards. The individual contributions of this book illustrate advanced data mining techniques which take advantage of the informative richness of both complex data and massive data for efficient and effective identification of complex information units present in such data.

The book is composed of 13 chapters.

Chapter 1 introduces an efficient algorithm to analyze pharmacogenomic data and discover association rules between gene variants of a patient and drug-dependent adverse events.

Chapter 2 describes a classification-based approach for speech remediation. This is used to identify which portions of a speech can be deleted, in order to enhance the speech understandability in terms of both speech content and speech flow.

Chapter 3 presents a heterogeneous clustering algorithm that is able to predict possibly unknown lncRNA-disease relationships by analyzing complex heterogeneous biological networks.

Chapter 4 illustrates a probabilistic generative model, in order to address the problem of positive-unlabeled learning by considering a set of positive samples and a (usually larger) set of unlabeled ones.

Chapter 5 proposes a constraint programming approach that is combined with large neighborhood search, in order to efficiently identify homogeneous subsets of genes, which are similarly expressed across subsets of patients.

Chapter 6 investigates the use of the scaled correlation function, in order to derive the structure of a functional brain network from the activity series associated with the network nodes.

Chapter 7 considers the problem of manufacturing defect identification. It compares two approaches that address the multiple instance problem when the traditional instance localization assumption is not met.

Chapter 8 focuses on the problem of designing ensemble strategies for defending the company's brands from an unauthorized use.

Chapter 9 investigates the task of electricity load forecasting through unsupervised ensemble learning of clustered time series data.

Chapter 10 tackles the problem of phenotype traits prediction using supervised and semi-supervised classification trees as well as supervised and semi-supervised random forests of classification trees.

Chapter 11 introduces an approach that constructs a label hierarchy as a decomposition of the output space of a classification problem, in order to improve the predictive performance.

Chapter 12 illustrates a non-parametric Bayesian approach, in order to fit a mixture model of Markov chains to user session data and devise behavioral patterns.

Chapter 13 studies the problem of community-based semantic subgroup discovery and leverages the structural properties of a complex network, in order to enhance the ontology-based subgroup identification.

We would like to thank all the authors who submitted papers for publication in this book and all the workshop participants and speakers. We are also grateful to the members of the Program Committee and external referee for their excellent work in reviewing submitted and revised contributions with expertise and patience. We would like to thank Hiroshi Motoda for his invited talk on "Which Is More Influential, 'Who' or 'When' for a User to Rate in Online Review Site?" A special thank you is due to both the ECML PKDD Workshop Chairs and to the ECML PKDD organizers who made the event possible. Last but not the least, we thank Alfred Hofmann of Springer for his continuous support.

February 2018

Annalisa Appice Corrado Loglisci Giuseppe Manco Elio Masciari Zbigniew Ras

Organization

Program Chairs

| Annalisa Appice | University of Bari Aldo Moro, Bari, Italy |
|------------------|--|
| Corrado Loglisci | University of Bari Aldo Moro, Bari, Italy |
| Giuseppe Manco | ICAR-CNR, Rende, Italy |
| Elio Masciari | ICAR-CNR, Rende, Italy |
| Zbigniew W. Ras | University of North Carolina, Charlotte, USA |
| | Polish-Japanese Academy of Information Technology, |
| | Warsaw, Poland |

Program Committee

| Martin Atzmueller | Tilburg University, Germany |
|----------------------|--|
| Elena Bellodi | University of Ferrara, Italy |
| Petr Berka | University of Economics of Prague, Czech Republic |
| Claudia Diamantini | Università Politecnica delle Marche, Italy |
| Hadi Fanaee Tork | University of Oslo, Norway |
| Bettina Fazzinga | CNR-ICAR, Italy |
| Filippo Furfaro | Università della Calabria, Italy |
| Stefano Ferilli | University of Bari, Italy |
| Dragi Kocev | Jozef Stefan Institute, Slovenia |
| Gjorgji Madjarov | Ss. Cyril and Methodius University, Macedonia |
| Mirjana Mazuran | Politecnico di Milano, Italy |
| Hiroshi Motoda | Osaka University and AFOSR/AOARD, Japan |
| Amedeo Napoli | LORIA, France |
| Ruggero G. Pensa | University of Turin, Italy |
| Gianvito Pio | University of Bari, Italy |
| Ettore Ritacco | CNR-ICAR, Italy |
| Samira Shaikh | University of North Carolina, Charlotte, USA |
| Jerzy Stefanowski | Poznan University of Technology, Poland |
| Irina Trubitsyna | University of Calabria, Italy |
| Herna Viktor | University of Ottawa, Canada |
| Alicja Wieczorkowska | Polish-Japanese Academy of Information Technology, |
| | Poland |
| Wlodek Zadrozny | University of North Carolina, Charlotte, USA |
| | |

Additional Reviewer

Massimo Guarascio

Which is More Influential, "Who" or "When" for a User to Rate in Online Review Site? (Invited Talk)

Hiroshi Motoda

Osaka University and AFOSR/AOARD, Japan

Abstract. At its heart the act of reviewing is very subjective, but in reality many factors influence user's decision. This can be called social influence bias. We pick two factors. "Who" and "When" and discuss which factor is more influential when a user posts his/her own rate after reading the past review scores in an online review system. We show that a simple model can learn the factor metric quite efficiently from a vast amount of data that is available in many online review systems and clarify that there is no universal solution and the influential factor depends on each dataset. We use a weighted multinomial generative model that takes account of each user's influence over other users. We consider two kinds of users: real and virtual, in accordance with the two factors, and assign an influence metric to each. In the former each user has its own metric, but in the latter the metric is assigned to the order of review posting actions (rating). Both metrics are learnable quite efficiently with a few tens of iterations by log-likelihood maximization. Goodness of metric is evaluated by the generalization capability. The proposed method was evaluated and confirmed effective by five review datasets. Different datasets give different results. Some dataset clearly indicates that user influence is more dominant than the order influence while the results are the other way around for some other dataset, and vet other dataset indicates that both factors are not relevant. The third one indicates that the decision is very subjective, i.e., independent of others' review. We tried to characterize the datasets, but were only partially successful. For datasets where user influence is dominant, we often observe that high metric users have strong positive correlations with three more basic metrics: 1) the number of reviews a user made, 2) the number of the user's followers who rate the same item, 3) the fraction of the user's followers who gave the similar rate, but this is not always true. We also observe that the majority of users is normal (average) and there are two small groups of users, each with high metric value and low metric value. Early adopters are not necessarily influential.

Contents

| Learning Association Rules for Pharmacogenomic Studies Giuseppe Agapito, Pietro H. Guzzi, and Mario Cannataro | 1 |
|---|-----|
| Segment-Removal Based Stuttered Speech Remediation Pierre Arbajian, Ayman Hajja, Zbigniew W. Raś, and Alicja A. Wieczorkowska | 16 |
| Identifying IncRNA-Disease Relationships via Heterogeneous Clustering Emanuele Pio Barracchia, Gianvito Pio, Donato Malerba, and Michelangelo Ceci | 35 |
| Density Estimators for Positive-Unlabeled Learning Teresa M. A. Basile, Nicola Di Mauro, Floriana Esposito, Stefano Ferilli, and Antonio Vergari | 49 |
| Combinatorial Optimization Algorithms to Mine a Sub-Matrix of Maximal Sum | 65 |
| A Scaled-Correlation Based Approach for Defining and Analyzing Functional Networks Samuel Dolean, Mihaela Dînşoreanu, Raul Cristian Mureşan, Attila Geiszt, Rodica Potolea, and Ioana Ţincaş | 80 |
| Complex Localization in the Multiple Instance Learning Context Dan-Ovidiu Graur, Răzvan-Alexandru Mariş, Rodica Potolea, Mihaela Dînşoreanu, and Camelia Lemnaru | 93 |
| Integrating a Framework for Discovering Alternative App Stores in a Mobile App Monitoring Platform | 107 |
| Usefulness of Unsupervised Ensemble Learning Methods for Time Series Forecasting of Aggregated or Clustered Load Peter Laurinec and Mária Lucká | 122 |
| Phenotype Prediction with Semi-supervised Classification Trees Jurica Levatić, Maria Brbić, Tomaž Stepišnik Perdih, Dragi Kocev, Vedrana Vidulin, Tomislav Šmuc, Fran Supek, and Sašo Džeroski | 138 |

| Structuring the Output Space in Multi-label Classification | |
|---|-----|
| by Using Feature Ranking | |
| Stevanche Nikoloski, Dragi Kocev, and Sašo Džeroski | |
| Infinite Mixtures of Markov Chains. | |
| Sun Reubou, Ancene Bouberry, Thorsten Straje, and Og Brejeu | 100 |
| Blaž Škrlj, Jan Kralj, Anže Vavpetič, and Nada Lavrač | |
| Author Index | 197 |