

# Preface

Modern automatic systems are able to collect huge volumes of data, often with a complex structure (e.g., multi-table data, XML data, Web data, time series and sequences, graphs, and trees). This fact poses new challenges for current information systems with respect to storing, managing, and mining these big sets of complex data.

The 5th International Workshop on New Frontiers in Mining Complex Patterns (NFMCP 2016) was held in Riva del Garda in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2016) on September 19, 2016. The purpose of this workshop was to bring together researchers and practitioners in data mining who are interested in the advances and latest developments in the area of extracting patterns from big and complex data sources. The workshop was aimed at integrating recent results from existing fields such as data mining, statistics, machine learning, and relational databases to discuss and introduce new algorithmic foundations and representation formalisms in complex pattern discovery.

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 that preserve the informative richness of complex data and allow for efficient and effective identification of complex information units present in such data.

The book is composed of five parts and a total of 16 chapters.

Part I analyzes Feature Selection and Induction in the presence of complex data. It consists of two chapters. Chapter 1 introduces an unsupervised algorithm for feature construction based on tree ensembles. It defines an informative data representation that is able to handle complex data structures, combining information from multiple sources. Chapter 2 presents a graph-based algorithm for feature selection. It ranks features by identifying the most important ones into an arbitrary set of cues.

Part II focuses on Classification and Prediction by illustrating some complex predictive problems. It consists of five chapters. Chapter 3 tackles the problem of pruning rule classifiers, while retaining their descriptive properties. It uses confirmation measures as representatives of interestingness measures designed to select rules with desirable descriptive properties. Chapter 4 studies the problem of automatically recognizing speed changes from audio data recorded in controlled conditions. The classification of the audio data is performed using random forests, deep learning architectures and support vector machines. Chapter 5 describes a classification task that aims at determining whether two voices are spoken by the same person or not. It illustrates an algorithm that performs the classification by evaluating the dissimilarity between a speech sample and a set of known models. Chapter 6 investigates the problem of interpreting rules induced from imbalanced data. It proposes three different strategies that combine Bayesian confirmation measures, in order to select rules having

good descriptive characteristics. Chapter 7 addresses the problem of modeling trust network evolution through social communications among users in a social media site. It introduces a link prediction algorithm based on mediating-objects and analyzes the effect of time-decay in creating trust-links.

Part III analyzes issues posed by Clustering in the presence of complex data. It consists of four chapters. Chapter 8 investigates the adoption of cluster analysis to predict the primary medical procedure for a patient. The processed patients are clustered according to their set of diagnoses. This cluster knowledge is then used to identify other existing patients that are considered similar to the new patient. Chapter 9 describes a clustering algorithm allowing us to group features that are likely to take extreme values simultaneously. It exploits the graphical structure stemming from the definition of the clusters. Chapter 10 presents a latent-factor-based approach whose goal is to profile users according to their behavior. It considers the actions as set of features instead of single atomic elements. Chapter 11 proposes a multiview clustering methodology that determines clusters of patients with similar symptoms and detects patterns of medication changes that lead to the improvement or decline of patients' quality of life.

Part IV presents algorithms Pattern Discovery. It consists of three chapters. Chapter 12 introduces an approach to extract recurrent deviations from historical logging data and generate anomalous patterns representing high-level deviations. It applies a frequent subgraph mining technique together with an ad hoc conformance-checking technique. Chapter 13 investigates the task of detecting weather changes, which are periodically repeated over time and space. It introduces a spatiotemporal pattern to represent a periodic change and describes a computational solution to discover this kind of pattern. Chapter 14 investigates the problem of user authentication based on key-stroke timing pattern. It proposes a simple, robust, and nonparameterized nearest-neighbor regression-based feature-ranking algorithm for anomaly detection.

Finally, Part V gives a general overview of Applications in sensor network and game scenarios. It contains two chapters. Chapter 15 provides a formalization of a graph-based approach that extends a directed weighted graph using a sequential state transformation function. It interprets the graph to model state transition matrices and describes an algorithm for deriving these interpretations in large-scale real-world sensor networks. Chapter 16 checks whether, and to what extent, advanced process mining techniques can support efficient and effective knowledge discovery in chess playing. It also provides interesting insight into the game rules and strategies, and/or may support effective game playing in future matches.

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 additional reviewers for their excellent work in reviewing submitted and revised contributions with expertise and patience. We would like to thank Jaakko Hollmen for his invited talk on "On Model, Patterns, and Prediction." A special thanks is due to both the ECML PKDD Workshop Chairs and to the ECML PKDD organizers who made the event possible. We would like to acknowledge the support of the European Commission through the projects

MAESTRA—Learning from Massive, Incompletely Annotated, and Structured Data (Grant number ICT-2013-612944) and TOREADOR—Trustworthy Model-Aware Analytics Data Platform (Grant number H2020-688797). Last but not the least, we thank Alfred Hofmann of Springer for his continuous support.

March 2017

Annalisa Appice  
Michelangelo Ceci  
Corrado Loglisci  
Elio Masciari  
Zbigniew Ras

New Frontiers in Mining Complex Patterns  
5th International Workshop, NFMCP 2016, Held in  
Conjunction with ECML-PKDD 2016, Riva del Garda, Italy,  
September 19, 2016, Revised Selected Papers  
Appice, A.; Ceci, M.; Loglisci, C.; Masciari, E.; Ras, Z.W.  
(Eds.)  
2017, XIV, 263 p. 66 illus., Softcover  
ISBN: 978-3-319-61460-1