

Preface

Nowadays, data mining and knowledge discovery are advanced research fields with numerous algorithms and studies to extract patterns and models from data in different forms. Although most historical data mining approaches look for patterns in tabular data, there are also numerous recent studies where the focus is on data with a complex structure (e.g., multi-relational data, XML data, web data, time series and sequences, graphs, and trees). Complex data pose new challenges for current research in data mining and knowledge discovery with respect to storing, managing, and mining these sets of complex data.

The Third International Workshop on New Frontiers in Mining Complex Patterns (NFMCP 2014) was held in Nancy in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2014) on September 19, 2014. It was aimed at bringing together researchers and practitioners of data mining and knowledge discovery who are interested in the advances and latest developments in the area of extracting nuggets of knowledge from complex data sources.

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-Verlag's high-quality publication standards. The individual contributions of this book illustrate advanced data mining techniques which 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 four parts and a total of 13 chapters.

Part I focuses on **Classification and Regression** by illustrating some complex predictive problems. It consists of two chapters. Chapter 1 presents ensembles of predictive clustering trees, which are learned in a self-training fashion for output spaces consisting of multiple numerical values. Chapter 2 compares different clustering algorithms for constructing the label hierarchies (in a data-driven manner) in multi-label classification.

Part II analyzes issues posed by **Clustering** in the presence of complex data. It consists of three chapters. Chapter 3 studies the problem of predicting patients' negative side effects. It describes a system that measures the similarity of a new patient to existing clusters, and makes a personalized decision on the patient's most likely negative side effects. Chapter 4 proposes a dual decomposition approach for correlation clustering and multicut segmentation, in order to address the problem of distributing the computation in the parallel implementation of a learning algorithm. Chapter 5 investigates the adoption of cluster analysis to build accurate classifiers from imbalanced datasets.

Part III presents algorithms and applications where complex patterns are discovered from **Data Streams and Sequences**. It contains four chapters. Chapter 6 focuses on ROC analysis with imbalanced data streams and proposes an efficient incremental

algorithm to compute AUC using constant time and memory. Chapter 7 studies the problem of mining frequent patterns from positional data streams in a continuous setting. Chapter 8 presents a grouping technique to visualize the influential actors of a network data stream. Chapter 9 illustrates a new approach to mine dependencies between sequences of interval-based events.

Finally, Part IV gives a general overview of **Applications** in mobile, organizational, and music scenarios. It contains four chapters. Chapter 10 describes a case study that uses a process mining methodology to extract meaningful collaboration behavioral patterns in research activities. Chapter 11 proposes an approach based on First-Order Logic to learn complex process models extended with conditions, which are exploited to detect and manage anomalies in a real case study. Chapter 12 presents an approach based on sequence mining for location prediction of mobile phone users. Chapter 13 addresses the problem of using binary random forests as a classification tool to identify pitch-and-instrument combination in short audio frames of polyphonic recordings of classical music.

We would like to thank all the authors who submitted papers for publishing in this book and all the workshop participants and speakers. We are also grateful to the members of the Program Committee and to the external referees for their excellent work in reviewing submitted and revised contributions with expertise and patience. We would like to thank Thomas Gärtner for his invited talk on “Sampling and Presenting Patterns from Structured Data.” Special thanks are due to both the ECML PKDD Workshop Chairs and to the members of ECML PKDD Organizers who made the event possible. We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944). Last but not least, we thank Alfred Hofmann of Springer for his continuous support.

February 2015

Annalisa Appice
Michelangelo Ceci
Corrado Loglisci
Giuseppe Manco
Elio Masciari
Zbigniew W. Ras

New Frontiers in Mining Complex Patterns

Third International Workshop, NFMCP 2014, Held in

Conjunction with ECML-PKDD 2014, Nancy, France,

September 19, 2014, Revised Selected Papers

Appice, A.; Ceci, M.; Loglisci, C.; Manco, G.; Masciari, E.;

Rás, Z.W. (Eds.)

2015, XII, 211 p. 61 illus., Softcover

ISBN: 978-3-319-17875-2