Exploiting the Power of Group Differences

Using Patterns to Solve Data Analysis Problems

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Exploratory Causal Analysis with Time Series Data
James M. McCracken
2016

Mining Human Mobility in Location-Based Social Networks
Huiji Gao and Huan Liu
2015

Mining Latent Entity Structures
Chi Wang and Jiawei Han
2015

Probabilistic Approaches to Recommendations
Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco
2014

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Exploiting the Power of Group Differences: Using Patterns to Solve Data Analysis Problems

Guozhu Dong

ISBN: 978-3-031-00785-9 paperback
ISBN: 978-3-031-01913-5 ebook
ISBN: 978-3-031-00108-6 hardcover

DOI 10.1007/978-3-031-01913-5

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #16
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of California, Santa Cruz*
Wei Wang, *University of California, Los Angeles*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Exploiting the Power of Group Differences

Using Patterns to Solve Data Analysis Problems

Guozhu Dong
Wright State University

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #16

ABSTRACT

This book presents pattern-based problem-solving methods for a variety of machine learning and data analysis problems. The methods are all based on techniques that exploit the power of group differences. They make use of group differences represented using emerging patterns (aka contrast patterns), which are patterns that match significantly different numbers of instances in different data groups. A large number of applications outside of the computing discipline are also included.

Emerging patterns (EPs) are useful in many ways. EPs can be used as features, as simple classifiers, as subpopulation signatures/characterizations, and as triggering conditions for alerts. EPs can be used in gene ranking for complex diseases since they capture multi-factor interactions. The length of EPs can be used to detect anomalies, outliers, and novelties. Emerging/contrast pattern-based methods for clustering analysis and outlier detection do not need distance metrics, avoiding pitfalls of the latter in exploratory analysis of high dimensional data. EP-based classifiers can achieve good accuracy even when the training datasets are tiny, making them useful for exploratory compound selection in drug design. EPs can serve as opportunities in opportunity-focused boosting and are useful for constructing powerful conditional ensembles. EP-based methods often produce interpretable models and results. In general, EPs are useful for classification, clustering, outlier detection, gene ranking for complex diseases, prediction model analysis and improvement, and so on.

EPs are useful for many tasks because they represent group differences, which have extraordinary power. Moreover, EPs represent multi-factor interactions, whose effective handling is of vital importance and is a major challenge in many disciplines.

Based on the results presented in this book, one can clearly say that patterns are useful, especially when they are linked to issues of interest.

We believe that many effective ways to exploit group differences' power still remain to be discovered. Hopefully this book will inspire readers to discover such new ways, besides showing them existing ways, to solve various challenging problems.

KEYWORDS

data mining, machine learning, data analytic, classification, regression, clustering, anomaly detection, outlier detection, intrusion detection, compound selection, complex disease analysis, extreme instance selection, factor ranking, prediction model analysis, group difference analysis, feature, multifactor interaction, diverse relationship, heterogeneity, boosting, ensemble, association rule, emerging pattern, contrast pattern, frequent pattern, distance metric, interpretability

To my wife for her love and support!

– Guozhu Dong

Contents

Acknowledgments

Many researchers directly contributed to results reported in this book, concerning (1) the development of computational methods that use emerging/contrast patterns to tackle various machine learning and data mining/analysis problems, and (2) the utilization of emerging/contrast patterns and methods to solve challenging problems outside of the computing discipline. Their work demonstrated the power of group differences and emerging/contrast patterns. Many other researchers indirectly contributed to results reported in this book through results on the mining of emerging/contrast patterns. I am grateful to all of those researchers.

I appreciate the helpful comments as well as useful references provided by James Bailey, Bruno Crémilleux, Ramamohanarao (Rao) Kotagiri, Jinyan Li, José Francisco Martínez-Trinidad, and Limsoon Wong. They helped improve the quality and completeness of this book.

I am grateful to Jiawei Han for helpful and encouraging suggestions, and to Diane D. Cerra and C.L. Tondo for assistance on various technical issues related to the preparation of this manuscript. I also appreciate the helpful comments from the reviewers, which helped improve the book.

Guozhu Dong
February 2019