

Exploiting the Power of Group Differences

Using Patterns to Solve Data Analysis Problems

Synthesis Lectures on Data Mining and Knowledge Discovery

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*

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.

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

Guozhu Dong

2019

Mining Structures of Factual Knowledge from Text

Xiang Ren and Jiawei Han

2018

Individual and Collective Graph Mining: Principles, Algorithms, and Applications

Danai Koutra and Christos Faloutsos

2017

Phrase Mining from Massive Text and Its Applications

Jialu Liu, Jingbo Shang, and Jiawei Han

2017

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

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2019

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

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	xv
1 Introduction and Overview	1
1.1 Importance of Group Differences	1
1.2 Summary of Chapters	2
1.2.1 Reading Order of the Chapters	5
1.3 Known Uses of Group Differences via Emerging Patterns	5
1.4 Unique Properties of Emerging Pattern Based Methods	6
1.5 Scenarios Where Emerging Patterns Are Especially Useful	7
1.6 Related Topics Not Covered in This Book	8
2 General Preliminaries	9
2.1 Attributes, Features, and Variables	9
2.2 Data Instances and Datasets	9
2.3 Attribute Binning and Discretization	10
2.4 Patterns, Matching Datasets, Supports, and Frequent Patterns	12
2.5 Equivalence Classes, Closed Patterns, Minimal Generators, and Borders . . .	13
2.6 Illustrating Examples	13
3 Emerging Patterns and a Flexible Mining Algorithm	15
3.1 Setting for Group Difference Analysis	15
3.2 Basics of Emerging Patterns	15
3.3 BorderDiff: A Simple, Flexible Emerging Pattern Mining Algorithm	18
3.4 What Emerging Patterns Can Represent	20
3.5 Comparison with Association Rules, Confidence, and Odds Ratio	21
3.6 Pointers to Sections Illustrating Uses of Emerging Patterns	23
3.7 Traditional Analysis of Group Differences	24
3.8 Discussion of Related Issues	25

4	CAEP: Classification by Aggregating Multiple Matching Emerging Patterns .	27
4.1	Background Materials on Classification	28
4.2	The CAEP Approach	28
4.2.1	CAEP's Class-Likelihood Computation	29
4.2.2	CAEP's Likelihood Normalization	29
4.2.3	Emerging Pattern Set Selection	30
4.2.4	The CAEP Training and Testing Algorithms	31
4.3	A Small Illustrating Example	31
4.4	Experiments and Applications by Other Researchers	33
4.5	Strengths and Uniqueness of CAEP	33
4.5.1	Strengths of CAEP	33
4.5.2	Uniqueness of CAEP	34
4.6	DeEPs: Instance-Based Classification Using Emerging Patterns	34
4.7	Relationship with Other Rule/Pattern-Based Classifiers	35
4.8	Discussion	36
5	CAEP for Classification on Tiny Training Datasets, Compound Selection, and Instance Selection	37
5.1	CAEP Performs Well on Tiny Training Data	37
5.1.1	Details on Data Used for Compound Selection	38
5.2	Using CAEP for Compound Selection	39
5.3	Iterative Algorithm for Extreme Instance Selection	40
5.4	Semi-Supervised Extreme Instance Selection vs. Semi-Supervised Learning	40
6	OCLEP: One-Class Intrusion Detection and Anomaly Detection	43
6.1	Background on Intrusion Detection, Anomaly Detection, and Outlier Detection	44
6.2	OCLEP: Emerging Pattern Length-Based Intrusion Detection	44
6.2.1	An Observation on Emerging Pattern's Length	45
6.2.2	What Emerging Patterns to Use and Their Mining	45
6.2.3	OCLEP's Training and Testing Algorithms	46
6.3	Experimental Evaluation of OCLEP	48
6.3.1	Details of the NSL-KDD Dataset	48
6.3.2	Intrusion Detection on the NSL-KDD Dataset	48
6.3.3	Masquerader Detection on Command Sequences	48
6.4	Discussion	49

7	CPCQ: Contrast Pattern Based Clustering-Quality Evaluation	51
7.1	Background on Clustering-Quality Evaluation	51
7.2	CPCQ's Rationale	52
7.3	Measuring Quality of CPs	52
7.4	Measuring Diversity of High-Quality CPs	54
7.5	Defining CPCQ	54
7.6	Mining CPs and Computing the Best N Groups of CPs to Maximize CPCQ Values	55
7.7	Experimental Evaluation of CPCQ	56
7.8	Discussion	56
8	CPC: Pattern-Based Clustering Maximizing CPCQ	57
8.1	Notations	57
8.2	Background on Clustering and Clustering Evaluation	58
8.3	Problem Setting and Guiding Ideas for CPC	58
8.4	Main Technical Measures	59
8.4.1	MPQ Between Two Patterns	59
8.4.2	MPQ Between a Pattern and a Pattern Set	60
8.5	The CPC Algorithm	61
8.6	General Experimental Evaluation of CPC	62
8.7	Text Data Analysis on Blogs Using CPC	63
8.8	Discussion	64
9	IBIG: Ranking Genes and Attributes for Complex Diseases and Complex Problems	65
9.1	Basics of the Gene-Ranking Problem	66
9.2	Background on Complex Diseases	67
9.3	Capturing Interactions Using Jumping Emerging Patterns	68
9.4	The IBIG Approach	69
9.4.1	High-Level View of the IBIG Approach	69
9.4.2	IBIG Gene Ranking based on a Set of Emerging Patterns	69
9.4.3	Gene Clubs and Computing Gene Clubs	70
9.4.4	The Iterative IBIG Algorithm: IBIGi	71
9.5	Experimental Findings on IBIG on Colon Cancer Data	72
9.5.1	High-Quality JEPs Often Involve Lowly IG-Ranked Genes and IBIGi Can Find Many of Them	72

9.5.2	Significant Gene-Rank Differences Between IG and IBIG	73
9.6	Discussion	75
10	CPXR and CPXC: Pattern Aided Prediction Modeling and Prediction Model Analysis	77
10.1	Background Materials	78
10.2	Pattern Aided Prediction Models	78
10.2.1	Fitting Local Models for Logical Subpopulations	78
10.2.2	Pattern Aided Prediction Models	79
10.3	CPXP: Contrast Pattern Aided Prediction	80
10.4	Relationship with Boosting and Ensemble Member Selection	82
10.5	Diverse Predictor-Response Relationships	82
10.6	Uses of CPXR and CPXC in Experiments	83
10.6.1	Experiments on Commonly Used Datasets	83
10.6.2	Applications for Agriculture and Healthcare Predictions	84
10.7	Subpopulationwise Conditional Correlation Analysis	85
10.8	Discussion	86
11	Other Approaches and Applications Using Emerging Patterns	87
11.1	Compound Activity Analysis	87
11.2	Structure-Activity Relationship Exploration and Analysis	88
11.3	Metabolite Biomarker Discovery	88
11.4	Structural Alerts for Molecular Toxicity	89
11.5	Identifying Disease Subtypes, and Disease Treatment Planning	89
11.6	Safety and Street Crime Analysis	89
11.7	Characterizing Music Families	90
11.8	Identifying Interaction Terms: Adverse Drug Reaction Analysis	90
11.9	Coupled Hidden Markov Model for Critical Patient Care	91
11.10	Pose-Based Human Activity Recognition	91
11.11	Protein Complex Detection	92
11.12	Inhibitor Prediction Combining FCA and JEP	92
11.13	Instant Activity Recognition in Video Sequences	93
11.14	Birth Defect Detection	93
11.15	Surgery Stage Identification and Feedback Delivery	93
11.16	Sensor-Based Activity Recognition	94
11.17	Online Banking Fraud Detection	94

11.18 Other EP-based Classification Approaches and Studies 95

11.19 Emerging Patterns for Classification over Streaming Data 96

11.20 Other Studies and Applications 96

11.21 Summary of Uses: Application Domain Perspective 98

11.22 Discussion 99

Bibliography 101

Author’s Biography 125

Index 127

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