

Chapman & Hall/CRC

Learning and Knowledge Discovery Series

CONTRAST DATA MINING

Concepts, Algorithms,
and Applications

Edited by
Guozhu Dong and James Bailey

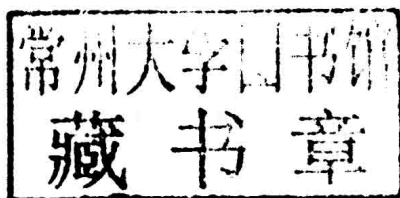


CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

CONTRAST DATA MINING

Concepts, Algorithms,
and Applications



Edited by
Guozhu Dong and James Bailey



CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an Informa business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2013 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
Version Date: 20120726

International Standard Book Number: 978-1-4398-5432-7 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

**CONTRAST
DATA MINING**
Concepts, Algorithms,
and Applications

Chapman & Hall/CRC

Data Mining and Knowledge Discovery Series

SERIES EDITOR

Vipin Kumar

University of Minnesota
Department of Computer Science and Engineering
Minneapolis, Minnesota, U.S.A.

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

UNDERSTANDING COMPLEX DATASETS:

DATA MINING WITH MATRIX DECOMPOSITIONS

David Skillicorn

COMPUTATIONAL METHODS OF FEATURE SELECTION

Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: ADVANCES IN ALGORITHMS, THEORY, AND APPLICATIONS

Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT

David Skillicorn

MULTIMEDIA DATA MINING: A SYSTEMATIC INTRODUCTION TO CONCEPTS AND THEORY

Zhongfei Zhang and Ruofei Zhang

NEXT GENERATION OF DATA MINING

Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

DATA MINING FOR DESIGN AND MARKETING

Yukio Ohsawa and Katsutoshi Yada

THE TOP TEN ALGORITHMS IN DATA MINING

Xindong Wu and Vipin Kumar

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, SECOND EDITION

Harvey J. Miller and Jiawei Han

TEXT MINING: CLASSIFICATION, CLUSTERING, AND APPLICATIONS

Ashok N. Srivastava and Mehran Sahami

BIOLOGICAL DATA MINING

Jake Y. Chen and Stefano Lonardi

试读结束：需要全本请在线购买：www.ertongbook.com

INFORMATION DISCOVERY ON ELECTRONIC HEALTH RECORDS

Vagelis Hristidis

TEMPORAL DATA MINING

Theophano Mitsa

RELATIONAL DATA CLUSTERING: MODELS, ALGORITHMS, AND APPLICATIONS

Bo Long, Zhongfei Zhang, and Philip S. Yu

KNOWLEDGE DISCOVERY FROM DATA STREAMS

João Gama

STATISTICAL DATA MINING USING SAS APPLICATIONS, SECOND EDITION

George Fernandez

INTRODUCTION TO PRIVACY-PRESERVING DATA PUBLISHING:
CONCEPTS AND TECHNIQUES

Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu

HANDBOOK OF EDUCATIONAL DATA MINING

Cristóbal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker

DATA MINING WITH R: LEARNING WITH CASE STUDIES

Luís Torgo

MINING SOFTWARE SPECIFICATIONS: METHODOLOGIES AND APPLICATIONS

David Lo, Siau-Cheng Khoo, Jiawei Han, and Chao Liu

DATA CLUSTERING IN C++: AN OBJECT-ORIENTED APPROACH

Guojun Gan

MUSIC DATA MINING

Tao Li, Mitsunori Ogihara, and George Tzanetakis

MACHINE LEARNING AND KNOWLEDGE DISCOVERY FOR
ENGINEERING SYSTEMS HEALTH MANAGEMENT

Ashok N. Srivastava and Jiawei Han

SPECTRAL FEATURE SELECTION FOR DATA MINING

Zheng Alan Zhao and Huan Liu

ADVANCES IN MACHINE LEARNING AND DATA MINING FOR ASTRONOMY

Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, and Ashok N. Srivastava

FOUNDATIONS OF PREDICTIVE ANALYTICS

James Wu and Stephen Coggeshall

INTELLIGENT TECHNOLOGIES FOR WEB APPLICATIONS

Priti Srinivas Sajja and Rajendra Akerkar

CONTRAST DATA MINING: CONCEPTS, ALGORITHMS, AND APPLICATIONS

Guozhu Dong and James Bailey

Dedication

To my wife Diana and my children. {G.D.}

To my wife Katherine. {J.B.}

To all contributing authors of the book
and to all researchers of the contrast mining field. {G.D. and J.B.}

Foreword

Contrast data mining is an important and focused subarea of data mining. Its aim is to find interesting contrast patterns that describe significant differences between datasets satisfying various contrasting conditions. The contrasting conditions can be defined on class, time, location, other “dimensions” of interest, or their combinations. The contrast patterns can represent nontrivial differences between classes, interesting changes over time, interesting trends in space, and so on.

Contrast data mining has provided, and will continue to provide, a unique angle to examine certain challenging problems and to develop powerful methodologies for solving those challenging problems, both in data mining research and in various applications. For the former, contrast patterns have been used for classification, clustering, and discriminative pattern analysis. For the latter, contrast data mining has been used in a wide spectrum of applications, such as differentiating cancerous tissues from benign ones, distinguishing structures of toxic molecules from that of non-toxic ones, and characterizing the differences on the issues discussed in the blogs on U.S. presidential elections in 2008 and those discussed in 2012. Contrast data mining can be performed on many kinds of data, including relational, vector, transactional, numerical, textual, music, image, and multimedia data, as well as complex structured data, such as sequences, graphs, and networks.

There have been numerous research papers published in recent years, on contrast mining algorithms, on applying contrast patterns in classification, clustering, and discriminative pattern analysis, and on applying contrast patterns and contrast-pattern based classification and clustering to a wide range of problems in medicine, bioinformatics, chemoinformatics, crime analysis, blog analysis, and so on. This book, edited by two leading researchers on contrast mining, Professors Guozhu Dong and James Bailey, and contributed to by over 40 data mining researchers and application scientists, is a comprehensive and authoritative treatment of this research theme. It presents a systematic introduction and a thorough overview of the state-of-the-art for contrast data mining, including concepts, methodologies, algorithms, and applications.

I have high confidence that the book will appeal to a wide range of readers, including data mining researchers and developers who want to be informed about recent progress in this exciting and fruitful area of research, scientific researchers who seek to find new tools to solve challenging problems in their

own research domains, and graduate students who want to be inspired on problem solving techniques and who want to get help with identifying and solving novel data mining research problems in various domains.

I find the book enjoyable to read. I hope you will like it, too.

Jiawei Han

University of Illinois, Urbana-Champaign

March 19, 2012

Preface

Contrasting is one of the most basic types of analysis. Contrasting based analysis is routinely employed, often subconsciously, by all types of people. People use contrasting to better understand the world around them and the challenging problems they want to solve. People use contrasting to accurately assess the desirability of important situations, and to help them better avoid potentially harmful situations and embrace potentially beneficial ones.

Contrasting involves the comparison of one dataset against another. The datasets may represent data of different time periods, spatial locations, or classes, or they may represent data satisfying different conditions. Contrasting is often employed to compare cases with a desirable outcome against cases with an undesirable one, for example comparing the benign and diseased tissue classes of a cancer, or comparing students who graduate with university degrees against those who do not. Contrasting can identify patterns that capture changes and trends over time or space, or identify discriminative patterns that capture differences among contrasting classes or conditions.

Traditional methods for contrasting multiple datasets were often very simple so that they could be performed by hand. For example, one could compare the respective feature means, compare the respective attribute-value distributions, or compare the respective probabilities of simple patterns, in the datasets being contrasted. However, the simplicity of such approaches has limitations, as it is difficult to use them to identify specific patterns that offer novel and actionable insights, and identify desirable sets of discriminative patterns for building accurate and explainable classifiers.

Contrast data mining, a special and focused area of data mining, develops concepts and algorithmic tools to help us overcome the limitations of those simple approaches. Recently, especially in the last dozen or so years, a large number of research papers on the concepts and algorithms of contrast data mining, and a large number of papers on successful applications of contrast mining in a wide range of scientific and business domains, have been reported. However, those results were only available in widely scattered places. This book presents the results in one place, in a comprehensive and coordinated fashion, making them more accessible to a wider spectrum of readers.

The importance and usefulness, and the diversified nature of contrast mining, have been indicated not only by the large number of papers, but also by the many names that have been used for *contrast patterns*. For example, the following names have been used: change pattern, characterization rule,

class association rule, classification rule, concept drift, contrast set, difference pattern, discriminative association, discriminative interaction pattern, discriminative pattern, dissimilarity pattern, emerging pattern, gradient pattern, group difference, unusual subgroups, and generalized contrast patterns such as fuzzy/disjunctive emerging patterns and contrast inequalities/regressions.

This book is focused on the mining and utilization of contrast patterns. It is divided into seven parts.

Part I, Preliminaries and Measures on Contrasts, contains two chapters, on preliminaries and on statistical measures for contrast patterns, respectively.

Part II, Contrast Mining Algorithms, contains five chapters: Chapters 3 and 4 are on mining emerging patterns using tree-based structures or tree-based searches, and using Zero-Suppressed Binary Decision Diagrams, respectively. Chapter 5 is on efficient direct mining of selective discriminative patterns for classification. Chapter 6 is on mining emerging patterns from structured data, such as sequences and graphs. Chapter 7 is on incremental maintenance of emerging patterns.

Part III, Generalized Contrasts, Emerging Data Cubes, and Rough Sets, contains three chapters: Chapter 8 is on more expressive contrast patterns (such as disjunctive/fuzzy emerging patterns, and contrast inequalities). Chapter 9 is on emerging data cube representations for OLAP data mining. Chapter 10 relates jumping emerging patterns with rough set theory.

Part IV, Contrast Mining for Classification and Clustering, contains four chapters: Chapter 11 gives an overview and analysis of contrast pattern based classification. Chapter 12 is on using emerging patterns in outlier and rare-class prediction. Chapter 13 is on enhancing traditional classifiers using emerging patterns. Chapter 14 presents CPC — Contrast Pattern Based Clustering Algorithm — together with a brief discussion on the CPCQ clustering quality index, which is based on the quality, abundance, and diversity of contrast patterns.

Part V, Contrast Mining for Bioinformatics and Chemoinformatics, contains five chapters: Chapter 15 is on emerging pattern based rules characterizing subtypes of leukemia. Chapter 16 is on discriminating gene transfer and microarray concordance analysis. Chapter 17 is on mining optimal emerging patterns when there are thousands of genes or features. Chapter 18 is on the theory and applications of emerging chemical patterns. Chapter 19 is on emerging molecule patterns as structural alerts for computational toxicology.

Part VI, Contrast Mining for Special Application Domains, contains five chapters: Chapter 20 is on emerging patterns and classification for spatial and image data. Chapter 21 is on geospatial contrast mining with applications on vegetation, biodiversity, and election-voting analysis. Chapter 22 is on mining emerging patterns for activity recognition. Chapter 23 is on emerging pattern based prediction of heart diseases and powerline safety. Chapter 24 is on emerging pattern based crime spots analysis and rental price prediction.

Part VII, Survey of Other Papers, contains one chapter: Chapter 25 gives

an overview of results on contrast mining and applications, with a focus on papers not already cited in the other chapters of the book. The chapter includes citations of papers that present algorithms on mining changes and model shift, on mining conditional contrasts, on mining niche patterns, on discovering holes and bumps, on discovering changes and emerging trends in tourism and in music, on understanding retail customer behavior, on using patterns to analyze and improve genetic algorithms, on using patterns to preserve privacy and protect network security, and on summarizing knowledge level differences between datasets.

The 25 chapters of this book were written by more than 40 authors who conduct research in a diverse range of disciplines, including architecture engineering, bioinformatics, biology, chemoinformatics, computer science, life-science informatics, medicine, and systems engineering and engineering management. The cited papers of the book deal with topics in much wider range of disciplines. It is also interesting to note that the book's authors are from a dozen countries, namely Australia, Canada, China, Cuba, Denmark, France, Germany, Japan, Korea, Poland, Singapore, and the USA.

The 25 chapters demonstrate many useful and powerful capabilities of contrast mining. For example, contrast patterns can be used to characterize disease classes. They can capture discriminative gene group interactions, and can help define interaction based importance of genes, for cancers. They can be used to build accurate and explainable classifiers that perform well for balanced classification as well as for imbalanced classification, to perform outlier detection, to enhance traditional classifiers, to serve as feature sets of traditional classifiers, and to measure clustering quality and to construct clusters without distance functions. They can be used in compound selection for drug design and in molecule toxicity analysis, in crime spot analysis and in heart disease diagnosis, in rental price prediction and in powerline safety analysis, in activity recognition, and in image and spatial data analysis. In general, contrast mining is useful for diversified application domains involving many different data types.

A very interesting virtue of contrast mining is that contrast-pattern aggregation based classification can be effective when very few, as few as three, training examples per class are available. This virtue is especially useful for situations where training data may be hard to obtain, for instance for drug lead selection. Another interesting characteristic is that length statistics of minimal jumping emerging patterns can be used to detect outliers, allowing the use of one number as a measure to detect intruders. Using such a minimal model is advantageous, since it is hard for intruders to discover and emulate the model of the normal user in order to evade detection. A third interesting trait of contrast mining is the ability to use the collective quality and diversity of contrast patterns to measure clustering quality and to form clusters, without relying on a distance function, which is often hard to define appropriately in clustering-like exploratory data analysis. As you read the chapters of the

book, you will notice many other powerful aspects of contrast patterns, which make them very useful in solving many challenging problems.

Perhaps the most important contribution of contrast mining will come when we no longer need to use the naive Bayes or similar simplifying approaches to handle the challenge of high dimensional data, when we have developed the methodology to systematically analyze, and accurately use, sets of multi-feature contrast patterns instead. We believe that contrast mining has made useful progress in this direction, and we hope that results reported in this book will help researchers make progress on this important problem. Success in this direction will have a large impact on the understanding and handling of intrinsically complex processes, such as complex diseases whose behaviors are influenced by the interaction of multiple genetic and environmental factors.

We envision that, in the not too distant future, the field of contrast data mining will become mature. Then, other disciplines such as biology, medicine, and physics will refer to contrast mining and use methods from the contrast mining toolbox, in the same way that they now use methods such as logistic regression and PCA. We also foresee that, as the world moves towards ubiquitous computing, people may some day have a *contrasting app* on their iPhone-like device, which, when pointed at two types of things, can answer the question “in what ways do these two types differ?”

This book demonstrates that contrast mining has been a fruitful field for research on data mining methodology and for research on utilizing contrast mining to solve real-life problems. There are still many interesting research questions that deserve our attention, both in developing contrast mining methodology within the realm of computer science and in utilizing contrast mining to solve challenging problems in domains outside of computer science. Let us join together in exploring the concepts, algorithms, techniques, and applications of contrast data mining, to quickly realize its full potential.

Guozhu Dong, Wright State University
James Bailey, The University of Melbourne
March 2012

Contents

Foreword	xix
Preface	xxi
I Preliminaries and Statistical Contrast Measures	1
1 Preliminaries	3
<i>Guozhu Dong</i>	
1.1 Datasets of Various Data Types	3
1.2 Data Preprocessing	4
1.3 Patterns and Models	6
1.4 Contrast Patterns and Models	8
2 Statistical Measures for Contrast Patterns	13
<i>James Bailey</i>	
2.1 Introduction	13
2.1.1 Terminology	14
2.2 Measures for Assessing Quality of Discrete Contrast Patterns	15
2.3 Measures for Assessing Quality of Continuous Valued Contrast Patterns	18
2.4 Feature Construction and Selection: PCA and Discriminative Methods	19
2.5 Summary	20
II Contrast Mining Algorithms	21
3 Mining Emerging Patterns Using Tree Structures or Tree Based Searches	23
<i>James Bailey and Kotagiri Ramamohanarao</i>	
3.1 Introduction	23
3.1.1 Terminology	24
3.2 Ratio Tree Structure for Mining Jumping Emerging Patterns	25
3.3 Contrast Pattern Tree Structure	27
3.4 Tree Based Contrast Pattern Mining with Equivalence Classes	28
3.5 Summary and Conclusion	29

4	Mining Emerging Patterns Using Zero-Suppressed Binary Decision Diagrams	31
	<i>James Bailey and Elsa Loekito</i>	
4.1	Introduction	31
4.2	Background on Binary Decision Diagrams and ZBDDs . . .	32
4.3	Mining Emerging Patterns Using ZBDDs	35
4.4	Discussion and Summary	38
5	Efficient Direct Mining of Selective Discriminative Patterns for Classification	39
	<i>Hong Cheng, Jiawei Han, Xifeng Yan, and Philip S. Yu</i>	
5.1	Introduction	40
5.2	DDPMine: Direct Discriminative Pattern Mining	42
5.2.1	Branch-and-Bound Search	42
5.2.2	Training Instance Elimination	44
5.2.2.1	Progressively Shrinking FP-Tree	46
5.2.2.2	Feature Coverage	46
5.2.3	Efficiency Analysis	48
5.2.4	Summary	49
5.3	Harmony: Efficiently Mining The Best Rules For Classification	49
5.3.1	Rule Enumeration	50
5.3.2	Ordering of the Local Items	51
5.3.3	Search Space Pruning	53
5.3.4	Summary	54
5.4	Performance Comparison Between DDPMine and Harmony .	55
5.5	Related Work	56
5.5.1	M ^b T: Direct Mining Discriminative Patterns via Model-based Search Tree	56
5.5.2	NDPMine: Direct Mining Discriminative Numerical Features	56
5.5.3	uHarmony: Mining Discriminative Patterns from Uncertain Data	57
5.5.4	Applications of Discriminative Pattern Based Classification	57
5.5.5	Discriminative Frequent Pattern Based Classification vs. Traditional Classification	58
5.6	Conclusions	58
6	Mining Emerging Patterns from Structured Data	59
	<i>James Bailey</i>	
6.1	Introduction	59
6.2	Contrasts in Sequence Data: Distinguishing Sequence Patterns	60
6.2.1	Definitions	61
6.2.2	Mining Approach	62