Intelligent Systems Reference Library

Volume 61

Series editors

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Lakhmi C. Jain, University of Canberra, Canberra, Australia
e-mail: Lakhmi.Jain@unisa.edu.au

For further volumes:
http://www.springer.com/series/8578

About this Series

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included.

Animesh Adhikari · Jhimli Adhikari
Witold Pedrycz

Data Analysis and Pattern Recognition in Multiple Databases

Springer

Animesh Adhikari
Parvatibai Chowgule College
Margao
India

Jhimli Adhikari
Narayan Zantye College
Bicholim
India

Witold Pedrycz
Department of Electrical and Computer Engineering
University of Alberta
Edmonton
Canada

ISSN 1868-4394 ISSN 1868-4408 (electronic)
ISBN 978-3-319-03409-6 ISBN 978-3-319-03410-2 (eBook)
DOI 10.1007/978-3-319-03410-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013953634

© Springer International Publishing Switzerland 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To
Radha Ballav Adhikary, and
Manorama Adhikary

AA

Gouri Sankar Datta, and Rina Datta

JA

Ewa, Barbara, and Adam

WP

Foreword

I knew the authors of this book from reviewing their quality research papers and co-authored monograph many years ago. It is my honor to introduce you this new book concerning about multisourced data mining.

Analysis of big data from multiple sources is an emerging field of study. With the introduction of numerous channels of collecting data, analysis of multisourced data becomes a need of the hour. Several organizations such as banks, insurance companies, and shopping malls generate a huge amount of data coming from different sources. Quite commonly, data generated for a long period of time could be divided into a number of datasets to generate time/region specific knowledge.

A systematic analysis of multisourced data was started by S. Zhang, X. Wu, and C. Zhang in 2002. This field is gaining popularity as many people started working in this area by realizing its potential. A number of issues, such as big size of a database, presence of multiparty sensitive data, immovability of data from one place to another, and variety of data format originating from different sources have made data analysis and knowledge discovery more challenging.

Discovery of patterns and associations in multisourced data is a natural and important activity. Although classical patterns such as association rule, frequent itemset, and sequential pattern are of interest to study, patterns such as high frequency rule, heavy association rule, and exceptional pattern, specific to multidatabases, are also getting reported. They were selectively co-authored to the monograph "Developing Multi-Database Mining Applications" by A. Adhikari, P. Ramachandrarao, W. Pedrycz.

We understand that the classical data mining techniques have limitations in mining multiple large databases. Multisourced data mining techniques, such as local pattern analysis and pipelined feedback technique, provide a way to deal with multisourced large databases. We believe many specialized techniques will get reported to handle multisourced big databases.

Association rule mining (Chap. 2), global pattern discovery (Chap. 4), and mining patterns of select items (Chap. 6) provide different patterns discovery techniques in multisourced data. Some interesting item-based data analyzes are found in Chaps. 3 and 5. Interesting patterns, such as exceptional pattern, iceberg and periodic pattern are found in Chaps. 7–9, respectively. Chapter 10 discusses influence analysis between items. This book presents a recent research on mining multisourced data while highlighting contributions made in the past.

Over the time researchers have made efforts to discover techniques, approaches to deal with the new and challenging scenarios. Undoubtedly, in future, we will be witnessing new techniques to deal with massive, complex data originating from various sources.

China, April 2013 Shichao Zhang

Preface

Analysis of big data is a recent hot topic of research, and data analysis using multiple large databases is one such specific area. This area is gaining popularity, as the sources of multiple datasets are easily available due to significant technological advancements in capturing data and current requirements of our society.

We have noticed that multi-database mining is a different activity than mono-database mining. There are many challenges that one needs to tackle with. In many cases, data from multiple sources can not be moved to a single location. In this regard, the issues such as privacy-preservation, limited bandwidth offered by wireless channels, and retention of local features might play important roles. In these cases, local data are required to be mined locally.

Although mining multiple large databases poses many challenges, it offers many opportunities. Many decisions are based on multiple large databases located in different geographical regions. Such decisions would be more valid if they are based on the data distributed over the regions. First systematic studies[1] were carried out by Shichao Zhang, Xindong Wu and Chengqi Zhang. We continued investigations[2] on mining multiple large databases. The present book is based on the current research on knowledge discovery in multiple related databases.

The authors of this book extend their gratitude to Professor Lakhmi Jain for recognizing our work, and we thank Springer for accepting our work. We thank Ms. Varddhene for her excellent typesetting work. We also thank Ms. Ramya for overseeing the entire production process.

India — Animesh Adhikari
India — Jhimli Adhikari
Canada — Witold Pedrycz

[1] Zhang S, Zhang C, Wu X (2004) Knowledge discovery in multiple databases. Springer.

[2] Adhikari A, Ramachandrarao P, Pedrycz W (2010) Developing multi-databases mining applications. Springer.

Contents

1 Introduction 1
1.1 Motivation 1
1.2 Distributed Data Mining 3
1.3 Multi-Database Mining Approaches 4
1.3.1 Local Pattern Analysis 5
1.3.2 Sampling 6
1.3.3 Re-Mining 6
1.4 Pre-Processing of Databases 6
1.4.1 Preparation of Data Warehouses 7
1.4.2 Temporal Aggregation 8
1.4.3 Partitioning Database 8
1.4.4 Database Thinning 9
1.4.5 Ordering of Databases 9
1.4.6 Selection of Databases 9
1.5 Patterns and Associations 10
1.6 Related Studies 13
1.7 Experimental Settings 14
1.8 Conclusions 14
References 16

2 Synthesizing Different Extreme Association Rules from Multiple Databases 21
2.1 Introduction 21
2.2 Some Extreme Types of Association Rule in Multiple Databases 23
2.3 Problem Statement 25
2.4 An Extended Model of Local Pattern Analysis for Synthesizing Global Patterns 25
2.5 Related Work 27
2.6 Synthesizing an Association Rule 29
2.6.1 Design of Algorithm 30
2.6.2 Error Calculation 34

2.7 Experiments . . . 35
2.7.1 Results of Experimental Studies . . . 37
2.7.2 Comparison with Existing Algorithm . . . 38
2.8 Conclusions . . . 40
References . . . 41

3 Clustering Items in Time-Stamped Databases Induced by Stability . . . 43
3.1 Introduction . . . 43
3.2 Related Work . . . 44
3.3 A Model of Mining Multiple Transactional Time-Stamped Databases . . . 45
3.4 Problem Statement . . . 47
3.5 Clustering Items . . . 49
3.5.1 Finding the Best Non-Trivial Partition . . . 50
3.5.2 Finding a Best Class . . . 54
3.6 Experiments . . . 55
3.7 Conclusions . . . 58
References . . . 59

4 Synthesizing Global Patterns in Multiple Large Data Sources . . . 61
4.1 Introduction . . . 61
4.2 Multi-Database Mining Using Local Pattern Analysis . . . 62
4.3 Generalized Multi-Database Mining Techniques . . . 63
4.3.1 Local Pattern Analysis . . . 63
4.3.2 Partition Algorithm . . . 64
4.3.3 IdentifyExPattern Algorithm . . . 64
4.3.4 RuleSynthesizing Algorithm . . . 64
4.4 Specialized Multi-Database Mining Techniques . . . 65
4.4.1 Mining Multiple Real Databases . . . 65
4.4.2 Mining Multiple Databases for the Purpose of Studying a Set of Items . . . 66
4.4.3 Study of Temporal Patterns in Multiple Databases . . . 66
4.5 Mining Multiple Databases Using Pipelined Feedback Technique . . . 66
4.5.1 Algorithm Design . . . 68
4.6 Error Evaluation . . . 68
4.7 Experiments . . . 69
4.8 Conclusions . . . 73
References . . . 73

5 Clustering Local Frequency Items in Multiple Data Sources 75
5.1 Introduction 75
5.2 Related Work 77
5.2.1 Measures of Association 77
5.2.2 Multi-Database Mining Techniques 78
5.2.3 Clustering Techniques 81
5.3 Problem Statement 82
5.4 Synthesizing Support of an Itemset 83
5.5 Clustering Local Frequency Items 86
5.5.1 Finding the Best Non-Trivial Partition 89
5.5.2 Error Analysis 93
5.6 Experimental Results 94
5.6.1 Overall Output 96
5.6.2 Synthesis of High Frequency Itemsets 97
5.6.3 Error Quantification 99
5.6.4 Average Error Versus γ 100
5.6.5 Average Error Versus α 101
5.6.6 Clustering Time Versus Number of Databases 103
5.6.7 Comparison with Existing Technique 104
5.7 Conclusions 106
References 107

6 Mining Patterns of Select Items in Different Data Sources 109
6.1 Introduction 109
6.2 Mining Global Patterns of Select Items 111
6.3 Overall Association Between Two Items in a Database 113
6.4 An Application: Study of Select Items in Multiple Databases Through Grouping 116
6.4.1 Properties of Different Measures 118
6.4.2 Grouping Frequent Items 120
6.4.3 Experiments 123
6.5 Related Work 127
6.6 Conclusions 128
References 128

7 Synthesizing Global Exceptional Patterns in Different Data Sources 131
7.1 Introduction 131
7.2 Exceptional Patterns in Multiple Data Sources 133
7.3 Problem Statement 139
7.4 Related Work 139
7.5 Synthesizing Support of an Itemset 140
7.6 Synthesizing Type II Global Exceptional Itemsets 141
7.7 Error Calculation 145

7.8 Experiments 146
7.8.1 Comparison with the Existing Algorithm 151
7.9 Conclusions 153
References 153

8 Mining Icebergs in Different Time-Stamped Data Sources 157
8.1 Introduction 157
8.2 Related Work 159
8.3 Notches in Sales Series 160
8.3.1 Identifying Notches 162
8.4 Generalized Notch 162
8.5 Iceberg Notch 163
8.6 Sales Series 164
8.6.1 Another View at Sales Series 164
8.6.2 Other Applications of Icebergs 165
8.7 Mining Icebergs in Time-Stamped Databases 166
8.7.1 Non-Incremental Approach 166
8.7.2 Incremental Approach 169
8.8 Significant Year 171
8.9 Experimental Studies 172
8.10 Conclusions 180
References 180

9 Mining Calendar-Based Periodic Patterns in Time-Stamped Data 183
9.1 Introduction 183
9.2 Related Work 186
9.3 Calendar-Based Periodic Patterns 187
9.3.1 Extending Certainty Factor 188
9.3.2 Extending Certainty Factor with Respect to Other Intervals 191
9.4 Mining Calendar-Based Periodic Patterns 192
9.4.1 Improving Mining Calendar-Based Periodic Patterns 193
9.4.2 Data Structure 193
9.4.3 A Modified Algorithm 195
9.5 Experimental Studies 198
9.5.1 Selection of Mininterval and Maxgap 202
9.5.2 Selection of Minsupp 204
9.5.3 Performance Analysis 205
9.6 Conclusions 207
References 207

10 Measuring Influence of an Item in Time-Stamped Databases 209
10.1 Introduction 209
10.2 Association Between Two Itemsets 211
10.3 Concept of Influence 212
10.3.1 Influence of an Itemset on Another Itemset 213
10.3.2 Properties of Influence Measures 214
10.3.3 Influence of an Item on a Set of Specific Items 215
10.3.4 Motivation 216
10.4 Problem Statement 218
10.5 Related Work 219
10.6 Design of Algorithms 219
10.6.1 Designing Algorithm for Measuring Overall Influence of an Item on Another Item 220
10.6.2 Designing Algorithm for Measuring Overall Influence of an Item on Each of the Specific Items 221
10.6.3 Designing Algorithm for Identifying Top Influential Items on a Set of Specific Items 221
10.7 Experiments 222
10.8 Conclusions 228
References 228

11 Summary and Conclusions 231
11.1 Changing Scenarios 231
11.2 Summary of Chapters 232
11.3 Selected Open Problems and Challenges 234
11.4 Conclusions 234
References 235

Index 237