

Advanced Information and Knowledge Processing

Springer-Verlag London Ltd.

Also in this series

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker and Ron Young
Knowledge Asset Management
1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi and Dimitrios Gunopulos
Uncertainty Handling and Quality Assessment in Data Mining
1-85233-655-2

Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho
Ontological Engineering
1-85233-551-3

Arno Scharl
Environmental Online Communication
1-85233-783-4

Shichao Zhang, Chengqi Zhang and
Xindong Wu

Knowledge Discovery in Multiple Databases

With 21 Figures



Springer

Shichao Zhang, PhD, MSc
Chengqi Zhang, PhD, MSc, BSc, DSc
FIT, University of Technology Sydney, Australia

Xindong Wu, PhD, MSc
Department of Computer Science, University of Vermont, USA

Series Editors
Xindong Wu
Lakhmi Jain

British Library Cataloguing in Publication Data
Zhang, Shichao

Knowledge discovery in multiple databases. – (Advanced information and knowledge processing)

1. Data mining 2. Distributed databases

I. Title II. Zhang, Chengqi III. Wu, Xindong
005.7'58

ISBN 978-1-4471-1050-7

Library of Congress Cataloging-in-Publication Data
Zhang, Shichao.

Knowledge discovery in multiple databases / Shichao Zhang, Chengqi Zhang, Xindong Wu.
p. cm.

Includes bibliographic references and index.

ISBN 978-1-4471-1050-7 ISBN 978-0-85729-388-6 (eBook)

DOI 10.1007/978-0-85729-388-6

1. Database management. 2. Database searching. I. Zhang, Chengqi, 1957– II. Wu, Xindong
III. Title

QA76.9.D3Z54 2004

005.74–dc22

2004048100

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

AI&KP ISSN 1610-3947

ISBN 978-1-4471-1050-7

springeronline.com

© Springer-Verlag London 2004

Originally published by Springer-Verlag London Limited in 2004

Softcover reprint of the hardcover 1st edition 2004

The use of registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by authors

34/3830-543210 Printed on acid-free paper SPIN 10894388

Preface

Many organizations have an urgent need of mining their multiple databases inherently distributed in branches (distributed data). In particular, as the Web is rapidly becoming an information flood, individuals and organizations can take into account low-cost information and knowledge on the Internet when making decisions. How to efficiently identify quality knowledge from different data sources has become a significant challenge.

This challenge has attracted a great many researchers including the authors who have developed a local pattern analysis, a new strategy for discovering some kinds of potentially useful patterns that cannot be mined in traditional multi-database mining techniques. Local pattern analysis delivers high-performance pattern discovery from multiple databases. There has been considerable progress made on multi-database mining in such areas as hierarchical meta-learning, collective mining, database classification, and peculiarity discovery. While these techniques continue to be future topics of interest concerning multi-database mining, this book focuses on these interesting issues under the framework of local pattern analysis.

The book is intended for researchers and students in data mining, distributed data analysis, machine learning, and anyone else who is interested in multi-database mining. It is also appropriate for use as a text supplement for broader courses that might also involve knowledge discovery in databases and data mining.

The book consists of ten chapters. Chapter 1 states the multi-database mining problem and its importance. Chapter 2 lays a common foundation for subsequent material. This includes the preliminaries on data mining and multi-database mining, as well as necessary concepts, previous efforts, and applications. Chapter 3 introduces the framework of local pattern analysis. The later chapters are essentially self-contained and may be read selectively, and in any order. Chapters 4, 5, and 6 develop techniques for preprocessing the data in multi-databases. Chapters 7, 8, and 9 presents techniques for identifying interesting patterns from multi-databases based on local pattern analysis. And Chapter 10 presents a summary of the previous chapters and demonstrates some open problems.

Beginners should read Chapters 1 and 2 before selectively reading other chapters. Although the opening problems are very important, techniques in

other chapters may be helpful for experienced readers who want to attack such problems.

Shichao Zhang

March 2004

Chengqi Zhang

Xindong Wu

Acknowledgments

We are deeply indebted to Jenny Wolkowicki for the carefully proofreading, as well as many colleagues for the advice and support they gave during the writing of this book. We are especially grateful to Tony King for his editorial efforts when he worked with Springer.

For many suggested improvements and discussions on the material, we thank Professor Geoffrey Webb from Monash University, Mr. Zili Zhang from Deakin University, Dr. Huan Liu from Arizona State University, President Hong Liang and Ms. Yanchun Zhou from Guangxi Teachers University, Ms. Li Liu and Mr. Xiaowei Yan from the University of Technology, Sydney, Professor Xiaopei Luo from the Chinese Academy of Sciences, and Professor Guoxi Fan from the Education Bureau of Quanzhou.

Contents

- 1. Importance of Multi-database Mining** 1
 - 1.1 Introduction 1
 - 1.2 Role of Multi-database Mining in Real-world Applications ... 2
 - 1.3 Multi-database Mining Problems 4
 - 1.4 Differences Between Mono- and Multi-database Mining 6
 - 1.4.1 Features of Data in Multi-databases 6
 - 1.4.2 Features of Patterns in Multi-databases 8
 - 1.5 Evolution of Multi-database Mining 9
 - 1.6 Limitations of Previous Techniques 12
 - 1.7 Process of Multi-database Mining 14
 - 1.7.1 Description of Multi-database Mining 14
 - 1.7.2 Practical Issues in the Process 16
 - 1.8 Features of the Defined Process 20
 - 1.9 Major Contributions of This Book 23
 - 1.10 Organization of the Book 24
- 2. Data Mining and Multi-database Mining** 27
 - 2.1 Introduction 27
 - 2.2 Knowledge Discovery in Databases 28
 - 2.2.1 Processing Steps of KDD 28
 - 2.2.2 Data Pre-processing 30
 - 2.2.3 Data Mining 31
 - 2.2.4 Post Data Mining 33
 - 2.2.5 Applications of KDD 34
 - 2.3 Association Rule Mining 36
 - 2.4 Research into Mining Mono-databases 41
 - 2.5 Research into Mining Multi-databases 51
 - 2.5.1 Parallel Data Mining 51
 - 2.5.2 Distributed Data Mining 52
 - 2.5.3 Application-dependent Database Selection 58
 - 2.5.4 Peculiarity-oriented Multi-database Mining 59
 - 2.6 Summary 61

3. Local Pattern Analysis	63
3.1 Introduction	63
3.2 Previous Multi-database Mining Techniques	64
3.3 Local Patterns	65
3.4 Local Instance Analysis Inspired by Competition in Sports...	67
3.5 The Structure of Patterns in Multi-database Environments	70
3.6 Effectiveness of Local Pattern Analysis	73
3.7 Summary	74
4. Identifying Quality Knowledge	75
4.1 Introduction	75
4.2 Problem Statement	76
4.2.1 Problems Faced by Traditional Multi-database Mining	76
4.2.2 Effectiveness of Identifying Quality Data	78
4.2.3 Needed Concepts	80
4.3 Nonstandard Interpretation	82
4.4 Proof Theory	88
4.5 Adding External Knowledge	91
4.6 The Use of the Framework	95
4.6.1 Applying to Real-world Applications	95
4.6.2 Evaluating Veridicality	96
4.7 Summary	100
5. Database Clustering	103
5.1 Introduction	103
5.2 Effectiveness of Classifying	104
5.3 Classifying Databases	107
5.3.1 Features in Databases	107
5.3.2 Similarity Measurement	108
5.3.3 Relevance of Databases and Classification	113
5.3.4 Ideal Classification and Goodness Measurement	115
5.4 Searching for a Good Classification	120
5.4.1 The First Step: Generating a Classification	121
5.4.2 The Second Step: Searching for a Good Classification	123
5.5 Algorithm Analysis	127
5.5.1 Procedure <i>GreedyClass</i>	127
5.5.2 Algorithm <i>GoodClass</i>	129
5.6 Evaluation of Application-independent Database Classification	130
5.6.1 Dataset Selection	130
5.6.2 Experimental Results	131
5.6.3 Analysis	134
5.7 Summary	135

6. Dealing with Inconsistency	137
6.1 Introduction	137
6.2 Problem Statement	138
6.3 Definitions of Formal Semantics	139
6.4 Weighted Majority	143
6.5 Mastering Local Pattern Sets	146
6.6 Examples of Synthesizing Local Pattern Sets	148
6.7 A Syntactic Characterization	150
6.8 Summary	155
7. Identifying High-vote Patterns	157
7.1 Introduction	157
7.2 Illustration of High-vote Patterns	158
7.3 Identifying High-vote Patterns	161
7.4 Algorithm Design	163
7.4.1 Searching for High-vote Patterns	164
7.4.2 Identifying High-vote Patterns: An Example	165
7.4.3 Algorithm Analysis	167
7.5 Identifying High-vote Patterns Using a Fuzzy Logic Controller	168
7.5.1 Needed Concepts in Fuzzy Logic	168
7.5.2 System Analysis	170
7.5.3 Setting Membership Functions for Input and Output Variables	171
7.5.4 Setting Fuzzy Rules	172
7.5.5 Fuzzification	174
7.5.6 Inference and Rule Composition	174
7.5.7 Defuzzification	176
7.5.8 Algorithm Design	177
7.6 High-vote Pattern Analysis	178
7.6.1 Normal Distribution	178
7.6.2 The Procedure of Clustering	179
7.7 Suggested Patterns	183
7.8 Summary	183
8. Identifying Exceptional Patterns	185
8.1 Introduction	185
8.2 Interesting Exceptional Patterns	186
8.2.1 Measuring the Interestingness	186
8.2.2 Behavior of Interest Measurements	189
8.3 Algorithm Design	189
8.3.1 Algorithm Design	189
8.3.2 Identifying Exceptions: An Example	192
8.3.3 Algorithm Analysis	193
8.4 Identifying Exceptions with a Fuzzy Logic Controller	195
8.5 Summary	195

9. Synthesizing Local Patterns by Weighting 197

9.1 Introduction 197

9.2 Problem Statement 198

9.3 Synthesizing Rules by Weighting 200

9.3.1 Weight of Evidence 200

9.3.2 Solving Weights of Databases 201

9.3.3 Algorithm Design 205

9.4 Improvement of Synthesizing Model 206

9.4.1 Effectiveness of Rule Selection 206

9.4.2 Process of Rule Selection 208

9.4.3 Optimized Algorithm 210

9.5 Algorithm Analysis 211

9.5.1 Procedure *RuleSelection* 211

9.5.2 Algorithm *RuleSynthesizing* 212

9.6 Summary 213

10. Conclusions and Future Work 215

10.1 Conclusions 215

10.2 Future Work 218

References 221

Subject Index 231