

Advanced Information and Knowledge Processing

Series Editors

Professor Lakhmi Jain
Lakhmi.jain@unisa.edu.au

Professor Xindong Wu
xwu@cems.uvm.edu

For other titles published in this series, go to
<http://www.springer.com/series/4738>

Animesh Adhikari · Pralhad Ramachandrarao ·
Witold Pedrycz

Developing Multi-database Mining Applications

 Springer

Animesh Adhikari
Department of Computer Science
Smt. Parvatibal Chowgule
College
Margoa-403602
India
animeshadhikari@yahoo.com

Pralhad Ramachandrarao
Department of Computer Science &
Technology
Goa University
Goa-403206
India
pralhaad@rediffmail.com

Witold Pedrycz
Department of Electrical & Computer
Engineering
University of Alberta
9107 116 Street
Edmonton AB T6G 2V4
Canada
pedrycz@ece.ualberta.ca

AI&KP ISSN 1610-3947
ISBN 978-1-84996-043-4 e-ISBN 978-1-84996-044-1
DOI 10.1007/978-1-84996-044-1
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010922804

© Springer-Verlag London Limited 2010

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 licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

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.

Printed on acid-free paper

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

To Jhimli and Sohom

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Distributed Data Mining	3
1.3	Existing Multi-database Mining Approaches	5
1.3.1	Local Pattern Analysis	5
1.3.2	Sampling	6
1.3.3	Re-mining	6
1.4	Applications of Multi-database Mining	7
1.5	Improving Multi-database Mining	8
1.5.1	Various Issues of Developing Effective Multi-database Mining Applications	8
1.6	Experimental Settings	10
1.7	Future Directions	10
	References	12
2	An Extended Model of Local Pattern Analysis	15
2.1	Introduction	15
2.2	Some Extreme Types of Association Rule in Multiple Databases	16
2.3	An Extended Model of Local Pattern Analysis for Synthesizing Global Patterns from Local Patterns in Different Databases	19
2.4	An Application: Synthesizing Heavy Association Rules in Multiple Real Databases	21
2.4.1	Related Work	21
2.4.2	Synthesizing an Association Rule	22
2.4.3	Error Calculation	28
2.4.4	Experiments	29
2.5	Conclusions	34
	References	34
3	Mining Multiple Large Databases	37
3.1	Introduction	37
3.2	Multi-database Mining Using Local Pattern Analysis	38
3.3	Generalized Multi-database Mining Techniques	39

3.3.1	Local Pattern Analysis	39
3.3.2	<i>Partition</i> Algorithm	39
3.3.3	<i>IdentifyExPattern</i> Algorithm	40
3.3.4	<i>RuleSynthesizing</i> Algorithm	40
3.4	Specialized Multi-database Mining Techniques	41
3.4.1	Mining Multiple Real Databases	41
3.4.2	Mining Multiple Databases for the Purpose of Studying a Set of Items	42
3.4.3	Study of Temporal Patterns in Multiple Databases	42
3.5	Mining Multiple Databases Using Pipelined Feedback Model (PFM)	43
3.5.1	Algorithm Design	44
3.6	Error Evaluation	45
3.7	Experiments	46
3.8	Conclusions	47
	References	49
4	Mining Patterns of Select Items in Multiple Databases	51
4.1	Introduction	51
4.2	Mining Global Patterns of Select Items	53
4.3	Overall Association Between Two Items in a Database	55
4.4	An Application: Study of Select Items in Multiple Databases Through Grouping	58
4.4.1	Properties of Different Measures	59
4.4.2	Grouping of Frequent Items	61
4.4.3	Experiments	65
4.5	Related Work	69
4.6	Conclusions	69
	References	69
5	Enhancing Quality of Knowledge Synthesized from Multi-database Mining	71
5.1	Introduction	71
5.2	Related Work	74
5.3	Simple Bit Vector (SBV) Coding	76
5.3.1	Dealing with Databases Containing Large Number of Items	77
5.4	Antecedent-Consequent Pair (ACP) Coding	79
5.4.1	Indexing Rule Codes	82
5.4.2	Storing Rulebases in Secondary Memory	86
5.4.3	Space Efficiency of Our Approach	88
5.5	Experiments	90
5.6	Conclusions	92
	References	93

- 6 Efficient Clustering of Databases Induced by Local Patterns** 95
 - 6.1 Introduction 95
 - 6.2 Problem Statement 97
 - 6.2.1 Related Work 98
 - 6.3 Clustering Databases 99
 - 6.3.1 Finding the Best Non-trivial Partition 110
 - 6.3.2 Efficiency of Clustering Technique 113
 - 6.4 Experiments 116
 - 6.5 Conclusions 118
 - References 119

- 7 A Framework for Developing Effective Multi-database Mining Applications** 121
 - 7.1 Introduction 121
 - 7.2 Shortcomings of the Existing Approaches to Multi-database Mining 122
 - 7.3 Improving Multi-database Mining Applications 122
 - 7.3.1 Preparation of Data Warehouses 123
 - 7.3.2 Choosing Appropriate Technique of Multi-database Mining 123
 - 7.3.3 Synthesis of Patterns 124
 - 7.3.4 Selection of Databases 124
 - 7.3.5 Representing Efficiently Patterns Space 125
 - 7.3.6 Designing an Appropriate Measure of Similarity 126
 - 7.3.7 Designing Better Algorithm for Problem Solving 126
 - 7.4 Conclusions 126
 - References 127

- Index** 129