



Data Mining and Data Warehousing

S.K. Mourya • Shalu Gupta

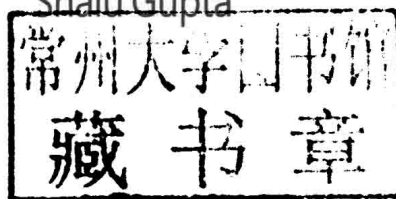


Alpha
Science

Data Mining and Data Warehousing

S. K. Mourya

Shalu Gupta



Alpha Science International Ltd.

Oxford, U.K.

Data Mining and Data Warehousing

214 pgs. | 85 figs. | 7 tbls.

S. K. Mourya

Shalu Gupta

Department of Computer Science and Engineering
MGM's College of Engineering & Technology
Noida

Copyright © 2013

ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

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, photocopying, recording or otherwise, without prior written permission of the publisher.

Printed from the camera-ready copy provided by the Authors.

ISBN 978-1-84265-757-7

Printed in India

Data Mining and Data Warehousing

Dedicated

*To our parents, grand parents
and
most beloved
Shilu
and our children*

**S. K. Mourya
Shalu Gupta**

Preface

The explosion of information technology, which continues to expand data driven markets and business, has made data mining an even more relevant topic of study. Books on data mining tend to be either broad or introductory or focus on some very specific technical aspect of the field.

Data Mining and Data Warehousing in nine chapters explores in depth the core of data mining (classification, clustering and association rules) by offering overviews that include both analysis and insight. Written for graduate students from various parts of the country studying Data Mining courses. The book is an ideal companion to either an introductory Data Mining textbook or a technical Data Mining book.

Unlike many other books that mainly focus on the modeling part, this volume discusses all the important—and often neglected—parts before and after modeling.

The book is organized as follows. It is divided into nine chapters.

In **Chapter 1**, a brief introduction to data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behaviour of customers, products and processes, various forms of data preprocessing, data cleaning, missing values, noisy data etc. In this chapter various classifications and various issues of data mining have also been explained with examples to illustrate them.

Chapter 2 deals with explaining data preprocessing and the various other needs of data processing. Forms of data preprocessing, e.g., data cleaning, missing values, noisy data. In this chapter we have also discussed how to handle inconsistent data, data integration and transformation.

Chapter 3 deals with explanation how statistics measures are used in large databases through measuring of central tendency, measuring dispersion of data in data mining, some graphical techniques used in data analysis of continuous data, etc. Further in the chapter we have also discussed data cube approach (OLAP).

Chapter 4 deals with discovery of frequent patterns, association, and correlation relationships among huge amounts of data, how it is useful in selective marketing, decision analysis, and business management. A popular area of application is market basket analysis which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence). Association rule mining that consists of first finding frequent item sets from which strong association rules in the form of $A \Rightarrow B$ are generated, has also been dealt with.

In **Chapter 5**, we have explained how classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. While classification predicts categorical labels (classes), prediction models deal with continuous-valued functions.

Chapter 6 explains that many clustering algorithms have been developed. These can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data (including frequent pattern-based methods), and constraint based methods. Some algorithms may belong to more than one category.

In **Chapter 7**, we study a well-accepted definition of the data warehouse and see why more and more organizations are building data warehouses for the analysis of their data. In particular, we study the *data cube*, a multidimensional data model for data warehouses, architecture and design. Through this chapter we can study that clearly, the presence of a data warehouse is a very useful precursor to data mining, and if it is not available, many of the steps involved in data warehousing will have to be undertaken to prepare the data for mining.

Chapter 8 presents an overview of OLAP technology aggregation, efficient query facility and its multidimensional aspects. Such an overview is essential for understanding the overall data mining and knowledge discovery process.

Chapter 9 deals with various issues of privacy and security of data emerged at a relatively early stage in the development of data mining. This development is not at all surprising given that all activities of data mining revolve around data and many sensitive issues of accessibility or possible reconstruction of data records exist, along with backup and testing concerns.

S. K. Mourya
Shalu Gupta

Acknowledgements

"Few successful endeavors have ever been made by one person alone, and this book is no exception."

We would like to thank the people and organizations that supported us in the production of this book and the authors are greatly indebted to all. There are several individuals and organizations whose support demands special mention and they are listed in the following.

Our special thanks to MGM's family Mr. Kamalkishor N. Kadam (Chairman), Dr. Mrs. Geeta S. Lathkar (Director), Prof. Sunil Wagh (VP) for being our inspiration and enabling us to publish this book. Above all, we want to thank our colleagues of Computer Science and Engineering Department, MGM-Noida, our family and friends who supported and encouraged us in spite of all the time it took us away from them.

We also wish to thank Mr. N. K. Mehra, Director, Narosa Publishing House Pvt. Ltd., for his enthusiasm, patience, and support during our writing of this book. Our Production Editor, and his staff for their conscientious efforts regarding production, editing, proof-reading and design etc. also deserve our special thanks.

We would like to express our gratitude to the many people who saw us through this book and to all of the reviewers for their invaluable feedback.

S. K. Mourya
Shalu Gupta

Contents

<i>Preface</i>	<i>vii</i>
<i>Acknowledgements</i>	<i>ix</i>
1. Introduction	1.1
1.1 What is Data, Information and Knowledge?	1.2
1.2 What is Motivated Data Mining?	1.3
1.3 Data Mining: Overview	1.3
1.3.1 Data collections and data availability	1.4
1.3.2 Some alternative terms for data mining	1.5
1.3.3 Steps for data processing	1.5
1.4 Typical Architecture of a Data Mining System	1.7
1.5 Data Mining: What kind of Data can be Mined?	1.8
1.5.1 Flat files	1.8
1.5.2 Relational databases	1.9
1.5.3 Data warehouses	1.9
1.5.4 Transaction databases	1.10
1.5.5 Multimedia databases	1.10
1.5.6 Spatial databases	1.10
1.5.7 Time-series databases	1.11
1.5.8 World Wide Web	1.11
1.6 Data Mining: What can be Discovered?	1.11
1.7 Classification of Data Mining Systems	1.12
1.8 Major Issues in Data Mining	1.13
<i>Summary</i>	<i>1.14</i>
<i>Exercises</i>	<i>1.14</i>
2. Data Preprocessing	2.1
2.1 Need of Data Processing	2.1
2.2 Form of Data Preprocessing	2.2
2.3 Data Cleaning	2.3
2.3.1 Missing values	2.3
2.3.2 Noisy data	2.4
2.3.3 How to handle inconsistent data?	2.5
2.4 Data Integration and Transformation	2.5

2.4.1	Data integration	2.5
2.4.2	Data transformation	2.6
2.5	Data Reduction	2.7
2.5.1	Data cube aggregation	2.7
2.5.2	Attribute subset selection	2.7
2.5.3	Dimensionality reduction	2.7
2.5.4	Numerosity reduction	2.7
2.5.5	Discretization and concept hierarchy generation	2.8
	<i>Summary</i>	2.8
	<i>Exercises</i>	2.8
3.	Statistics and Concept Description in Data Mining	3.1
3.1	Overview	3.2
3.2	Statistics Measures in Large Databases	3.2
3.2.1	Measuring of Central Tendency	3.2
3.2.2	Measuring Dispersion of Data	3.4
3.3	Graphical Techniques used in Data Analysis of Continuous Data	3.6
3.4	Concept/Class Description: Characterization and Discrimination	3.9
3.4.1	Methods for concept description	3.10
3.4.2	Differences between concept description in large databases and on-line analytical processing	3.11
3.5	Data Generalization and Summarization based Characterization	3.12
3.5.1	Data cube approach (OLAP)	3.12
3.5.2	Attribute-oriented induction approach (AOI)	3.13
3.6	Mining Class Comparisons: Discrimination between Classes	3.14
	<i>Summary</i>	3.15
	<i>Exercises</i>	3.15
4.	Association Rule Mining	4.1
4.1	Introduction	4.2
4.1.1	Frequent pattern analysis	4.2
4.1.2	What is market basket analysis (MBA)?	4.3
4.2	Concepts of Association Rule Mining	4.4
4.3	Frequent Pattern Mining in Association Rules	4.6
4.4	Mining Single-dimensional Boolean Association Rules	4.7
4.4.1	Apriori algorithm	4.8
4.4.2	Improving the efficiency of the Apriori Rules	4.9
4.5	Mining Multi-level Association Rules from Transaction Database	4.11
4.5.1	Multi-level association rules	4.13
4.5.2	Approaches to mining multi-level association rules	4.14
4.6	Mining Multi-dimensional Association Rules from Relational Databases and Data Warehouses	4.14
4.6.1	Multi-dimensional association rules	4.15
	<i>Summary</i>	4.15
	<i>Exercises</i>	4.16

5. Classification and Prediction	5.1
5.1 Classification	5.2
5.2 Prediction	5.3
5.3 Why are Classification and Prediction Important?	5.3
5.4 What is Test Data?	5.4
5.5 Issues Regarding Classification and Prediction	5.4
5.5.1 Preparing the data for classification and prediction	5.4
5.5.2 Comparing classification and prediction methods	5.4
5.6 Decision Tree	5.5
5.6.1 How a decision tree works	5.6
5.6.2 Decision tree induction	5.7
5.6.3 What is decision tree learning algorithm?	5.7
5.6.4 ID3	5.8
5.7 Bayesian Classification	5.11
5.7.1 Naïve Bayesian classifiers	5.12
5.7.2 Bayesian networks	5.14
5.8 Neural Networks	5.15
5.8.1 Feed forward	5.16
5.8.2 Backpropagation	5.16
5.9 K-nearest Neighbour Classifiers	5.18
5.10 Genetic Algorithm	5.20
<i>Summary</i>	5.21
<i>Exercises</i>	5.21
6. Cluster Analysis	6.1
6.1 Cluster Analysis: Overview	6.2
6.2 Stages of Clustering Process	6.4
6.3 Where do We Need Clustering/Application Areas?	6.4
6.4 Characteristics of Clustering Techniques in Data Mining	6.5
6.5 Data types in Cluster Analysis	6.6
6.5.1 Data matrix (or object-by-variable structure)	6.6
6.5.2 Dissimilarity matrix (or object-by-variable structure)	6.7
6.6 Categories of Clustering Methods	6.7
6.6.1 Partitioning methods	6.8
6.6.2 Hierarchical clustering	6.13
6.6.3 Density based methods	6.16
6.6.4 Grid based methods	6.18
6.6.5 Model based method: Statistical approach, neural network approach and outlier analysis	6.20
<i>Summary</i>	6.21
<i>Exercises</i>	6.21

7. Data Warehousing Concepts	7.1
7.1 What is a Data Warehouse?	7.2
7.1.1 Types of data warehouse	7.2
7.1.2 Data warehouse access tools	7.3
7.1.3 Data warehouse advantages	7.6
7.1.4 Differences between operational database systems and data warehouses	7.6
7.1.5 Difference between database (DB) and data warehouse	7.7
7.1.6 Transaction database vs. operational database	7.8
7.2 Multidimensional Data Model	7.8
7.2.1 Data cubes	7.8
7.2.2 Star schema	7.9
7.2.3 Snowflake schema	7.10
7.2.4 Fact constellation schema	7.11
7.2.5 Concept hierarchy	7.12
7.3 Data Warehouse Design	7.12
7.3.1 The process of data warehouse design	7.13
7.4 Architecture of Data Warehouse	7.13
7.4.1 Two-tier architecture of data warehouse	7.14
7.4.2 Three-tier architecture	7.14
7.5 What is Data Mart?	7.15
<i>Summary</i>	7.17
<i>Exercises</i>	7.17
8. OLAP Technology and Aggregation	8.1
8.1 Aggregation	8.2
8.1.1 Cube aggregation	8.4
8.1.2 Historical information	8.4
8.2 Query Facility	8.5
8.2.1 Efficient processing of OLAP queries	8.6
8.2.2 Transformation of complex SQL queries	8.6
8.2.3 System building blocks for an efficient query system	8.7
8.2.4 Query performance without impacting transaction processing	8.7
8.3 Online Analytical Processing (OLAP)	8.7
8.3.1 OLAP/OLAM architecture	8.8
8.3.2 Characteristics of OLAP	8.8
8.3.3 OLAP cube life-cycle	8.10
8.3.4 OLAP functions/analytical operations	8.11
8.3.4.1 Roll-up/Drill-up or consolidate	8.12
8.3.4.2 Drill-down	8.12
8.3.4.3 Slice and dice	8.12
8.3.4.4 Pivot (rotate)	8.14
8.3.5 OLAP tools	8.14
8.3.6 Types of OLAP servers	8.15

8.3.6.1	Multidimensional OLAP (MOLAP): Cube based	8.15
8.3.6.2	Relational OLAP (ROLAP): Star schema based	8.17
8.3.6.3	Comparison between MOLAP and ROLAP	8.18
8.3.6.4	Hybrid OLAP (HOLAP)	8.19
8.4	Data Mining vs. OLAP	8.19
	<i>Summary</i>	8.20
	<i>Exercises</i>	8.21
9.	Data Mining Security, Backup, Recovery	9.1
9.1	Data Mining Interfaces	9.2
9.1.1	Programmatic interfaces	9.2
9.1.2	Graphical user interface	9.2
9.2	Data Mining and Security	9.2
9.2.1	Identifying the data	9.3
9.2.2	Classifying the data	9.3
9.2.3	Quantifying the value of data	9.4
9.2.4	Identifying data vulnerabilities	9.5
9.2.5	Identifying protective measures and their costs	9.6
9.2.6	Why is security necessary for a data warehouse?	9.7
9.3	Backup and Recovery	9.8
9.4	Tuning Data Warehouse	9.9
9.5	Testing Data Warehouse	9.11
9.5.1	Data warehouse testing responsibilities	9.11
9.5.2	Business requirements and testing	9.11
9.5.3	Data warehousing test plan	9.11
9.5.4	Challenges of data warehouse testing	9.12
9.5.5	Categories of data warehouse testing	9.12
	<i>Summary</i>	9.14
	<i>Exercises</i>	9.14
	<i>Model Paper</i>	<i>MP.1</i>
	<i>Glossary</i>	<i>G.1</i>
	<i>Index</i>	<i>I.1</i>

