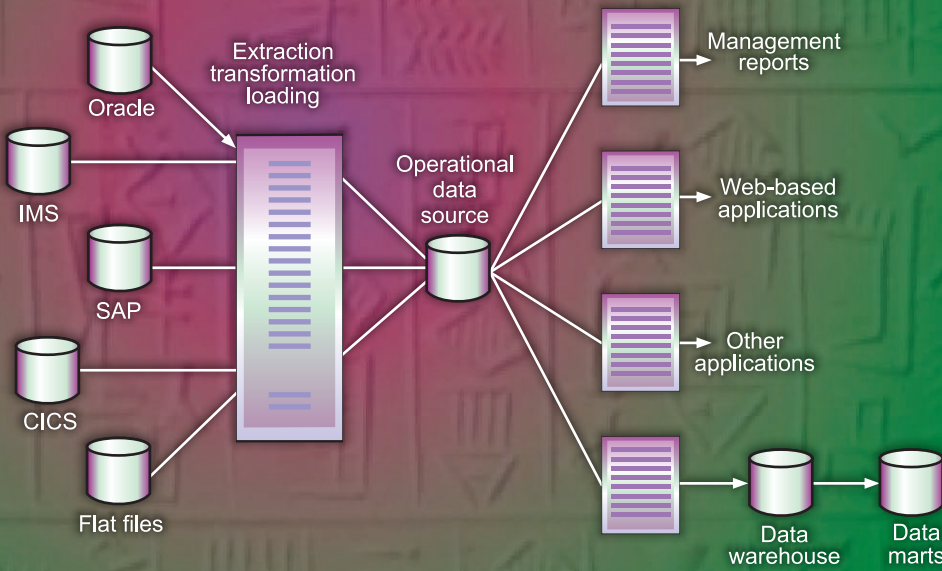


THIRD EDITION

**Eastern
Economy
Edition**



INTRODUCTION TO DATA MINING WITH CASE STUDIES



G.K. GUPTA

Introduction to Data Mining with Case Studies

Introduction to
Data Mining
with
Case Studies

THIRD EDITION

G.K. GUPTA

Adjunct Professor of Computer Science
Monash University
Clayton, Australia

PHI Learning Private Limited

Delhi-110092

2014

INTRODUCTION TO DATA MINING WITH CASE STUDIES, Third Edition

G.K. Gupta

© 2014 by PHI Learning Private Limited, Delhi. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

ISBN-978-81-203-5002-1

The export rights of this book are vested solely with the publisher.

Eighth Printing (Third Edition)

...

...

July, 2014

Published by Asoke K. Ghosh, PHI Learning Private Limited, Rimjhim House, 111, Patparganj Industrial Estate, Delhi-110092 and Printed by Baba Barkha Nath Printers, Bahadurgarh, Haryana-124507.

To

the memory of

Professor C.S. Wallace

Foundation Professor and Head
Department of Computer Science

Monash University

26 October 1933–7 August 2004



Contents

<i>Preface</i>	<i>xiii</i>
<i>Preface to the Second Edition</i>	<i>xv</i>
<i>Preface to the First Edition</i>	<i>xvii</i>
Chapter 1 INTRODUCTION	1-60
<i>Learning Objectives</i>	1
1.1 Introduction	1
<i>Chapter Overview</i>	2
1.2 What is Data Mining?	3
1.3 Why Data Mining Now?	4
1.4 The Data Mining Process—Software Development Approach	10
1.5 The Data Mining Process—The CRISP-DM Approach	11
1.6 Data Mining Applications	15
1.7 Data Mining Techniques	18
1.8 Practical Examples of Data Mining	21
1.9 The Future of Data Mining	28
1.10 Guidelines for Successful Data Mining	29
1.11 Limitations of Data Mining	30
1.12 Using WEKA Software in Class	31
1.13 Data Mining Software	31
<i>Summary</i>	34
<i>Review Questions</i>	35
<i>Exercises</i>	35
<i>Multiple Choice Questions</i>	36
<i>Bibliography</i>	38
Case Study 1A Data Mining Techniques for Optimizing Inventories for Electronic Commerce	41
Case Study 1B Crime Data Mining: A General Framework and Some Examples	52

Chapter 2 DATA UNDERSTANDING AND DATA PREPARATION	61–90
<i>Learning Objectives</i>	61
2.1 Introduction	61
<i>Chapter Overview</i>	62
2.2 Data Collection and Pre-processing	62
2.3 Outliers	70
2.4 Mining Outliers	72
2.5 Missing Data	74
2.6 Types of Data	75
2.7 Computing Distance	77
2.8 Data Summarising Using Basic Statistical Measurements	79
2.9 Displaying Data Graphically	82
2.10 Multidimensional Data Visualisation	85
<i>Summary</i>	85
<i>Review Questions</i>	85
<i>Exercises</i>	86
<i>Multiple Choice Questions</i>	87
<i>Bibliography</i>	90
Chapter 3 ASSOCIATION RULES MINING	91–151
<i>Learning Objectives</i>	91
3.1 Introduction	91
<i>Chapter Overview</i>	92
3.2 Basics	92
3.3 The Task and a Naïve Algorithm	94
3.4 The Apriori Algorithm	97
3.5 Improving the Efficiency of the Apriori Algorithm	110
3.6 Apriori-TID	111
3.7 Direct Hashing and Pruning (DHP)	114
3.8 Dynamic Itemset Counting (DIC)	117
3.9 Mining Frequent Patterns without Candidate Generation (FP–Growth)	118
3.10 Performance Evaluation of Algorithms	123
<i>Summary</i>	123
<i>Review Questions</i>	124
<i>Exercises</i>	125
<i>Multiple Choice Questions</i>	127
<i>Project 1—Using ARM for Table 1.1</i>	129
<i>Project 2—Designing a Shopping Mall</i>	129
<i>Project 3—Distributed ARM</i>	130
<i>Bibliography</i>	131
<i>Rakesh Agrawal—Inventor of Association Rules Mining</i>	133
Case Study 3 Mining Customer Value: From Association Rules to Direct Marketing	134

Chapter 4 CLASSIFICATION	152–215
<i>Learning Objectives</i>	152
4.1 Introduction	152
<i>Chapter Overview</i>	153
4.2 Decision Tree	154
4.3 Building a Decision Tree—The Tree Induction Algorithm	156
4.4 Split Algorithm Based on Information Theory	157
4.5 Split Algorithm Based on the Gini Index	163
4.6 Overfitting and Pruning	169
4.7 Decision Tree Rules	169
4.8 Decision Tree Summary	170
4.9 Naïve Bayes Method	171
4.10 Estimating Predictive Accuracy of Classification Methods	174
4.11 Improving Accuracy of Classification Methods	177
4.12 Other Evaluation Criteria for Classification Methods	178
4.13 Classification Software	179
<i>Summary</i>	181
<i>Review Questions</i>	181
<i>Exercises</i>	182
<i>Multiple Choice Questions</i>	184
<i>Projects</i>	185
<i>Bibliography</i>	187
<i>Ross Quinlan—Leading researcher in Decision Trees</i>	189
Case Study 4A KDD for Insurance Risk Assessment: A Case Study	190
Case Study 4B A Data Mining Approach for Retailing Bank Customer Attrition Analysis	198
Chapter 5 CLUSTER ANALYSIS	216–269
<i>Learning Objectives</i>	216
5.1 Introduction	216
<i>Chapter Overview</i>	219
5.2 Desired Features of Cluster Analysis	219
5.3 Types of Cluster Analysis Methods	220
5.4 Partitional Methods	221
5.5 Hierarchical Methods	228
5.6 Density-Based Methods	238
5.7 Dealing with Large Databases	239
5.8 Quality and Validity of Cluster Analysis Methods	241
5.9 Cluster Analysis Software	243
<i>Summary</i>	243
<i>Review Questions</i>	244
<i>Exercises</i>	245
<i>Multiple Choice Questions</i>	245
<i>Projects</i>	247
<i>Bibliography</i>	250
Case Study 5 Efficient Clustering of Very Large Document Collections	252

Chapter 6	WEB DATA MINING	270–331
<i>Learning Objectives</i>	270	
6.1 Introduction	270	
<i>Overview</i>	272	
6.2 Web Mining	272	
6.3 Web Terminology and Characteristics	273	
6.4 Locality and Hierarchy in the Web	278	
6.5 Web Content Mining	280	
6.6 Web Usage Mining	286	
6.7 Web Structure Mining	288	
6.8 Web Mining Software	295	
<i>Summary</i>	296	
<i>Review Questions</i>	296	
<i>Exercises</i>	297	
<i>Multiple Choice Questions</i>	297	
<i>Bibliography</i>	300	
<i>Tim Berners-Lee—Inventor of the World Wide Web</i>	303	
Case Study 6 Lessons and Challenges from Mining Retail E-Commerce Data		304
Chapter 7	SEARCH ENGINES AND QUERY MINING	332–381
<i>Learning Objectives</i>	332	
7.1 Introduction	332	
<i>Chapter Overview</i>	333	
7.2 Differences between Web Search and Information Retrieval		333
7.3 Characteristics of Search Engines	334	
7.4 Search Engine Functionality	338	
7.5 Search Engine Architecture	339	
7.6 Ranking of Web Pages	346	
7.7 Search Query Mining	352	
7.8 Individual Privacy and Query Data Mining	356	
<i>Summary</i>	357	
<i>Review Questions</i>	357	
<i>Exercises</i>	357	
<i>Multiple Choice Questions</i>	359	
<i>Project</i>	360	
<i>Bibliography</i>	362	
Case Study 7 The Anatomy of a Large-Scale Hypertextual Web Search Engine		364
Chapter 8	DATA WAREHOUSING	382–425
<i>Learning Objectives</i>	382	
8.1 Introduction	382	
8.2 Operational Data Stores	385	
8.3 Data Warehouses	387	

8.4	Data Warehouse Design	392
8.5	Guidelines for Data Warehouse Implementation	396
8.6	Data Warehouse Metadata	398
8.7	Software for ODS and Data Warehousing	399
	<i>Summary</i>	400
	<i>Review Questions</i>	401
	<i>Exercises</i>	402
	<i>Multiple Choice Questions</i>	402
	<i>Projects</i>	404
	<i>Bibliography</i>	405
	<i>Bill Inmon—Inventor of Data Warehouse</i>	407
Case Study 8	Data Warehouse Governance: Best Practices at Blue Cross and Blue Shield of North Carolina	408

Chapter 9 ONLINE ANALYTICAL PROCESSING (OLAP) 426–470

	<i>Learning Objectives</i>	426
9.1	Introduction	426
9.2	OLAP	427
9.3	Characteristics of OLAP Systems	429
9.4	Motivations for Using OLAP	432
9.5	Multidimensional View and Data Cube	433
9.6	Data Cube Implementations	439
9.7	Data Cube Operations	443
9.8	Guidelines for OLAP Implementation	447
9.9	OLAP Software	448
	<i>Summary</i>	449
	<i>Review Questions</i>	450
	<i>Exercises</i>	450
	<i>Multiple Choice Questions</i>	450
	<i>Bibliography</i>	453
	<i>Jim Gray (1944–2007)—Pioneer in Databases and OLAP</i>	454
Case Study 9	Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs	455

Chapter 10 INFORMATION PRIVACY AND DATA MINING 471–504

	<i>Learning Objectives</i>	471
10.1	Introduction	471
	<i>Chapter Overview</i>	472
10.2	What is Information Privacy?	472
10.3	Basic Principles to Protect Information Privacy	472
10.4	Privacy Legislation in India	475
10.5	Uses and Misuses of Data Mining	476
10.6	Primary Aims of Data Mining	478
10.7	Pitfalls of Data Mining	479

10.8	Why Current Privacy Principles are Ineffective for Data Mining?	480	
10.9	A Revised Set of Privacy Principles for Data Mining of Personal Information		482
10.10	Technological Solutions	483	
10.11	Examples of Use of Data Mining by the US Government		485
	<i>Summary</i>	488	
	<i>Review Questions</i>	488	
	<i>Exercises</i>	489	
	<i>Multiple Choice Questions</i>	489	
	<i>Bibliography</i>	491	
Case Study 10	Privacy Conflicts in CRM Services for Online Shops: A Case Study		493
	<i>Answers to Multiple Choice Questions</i>		505-506
	<i>Index</i>		507-514



Preface

The third edition of this book is a substantially revised version of the earlier editions. The first chapter has been rewritten and expanded. A new example has been added. Whereas the second chapter is completely new to this edition. It discusses the importance of data preprocessing in data mining. A number of issues are discussed in this chapter. An interesting example is included as an exercise at the end of the chapter. This example may also be used in Chapter 5 on Clustering. Chapter 3 has been revised and a new project has been included. Chapter 4 has been revised and so has been Chapter 5. Minor modifications have been made to Chapter 6. Whereas Chapter 7 has been revised substantially. A new section on Query Data Mining has been added to this chapter. Minor modifications have been made to Chapters 8 and 9. Finally, Chapter 10 has been revised substantially to focus on privacy developments in India.

Please continue to send me feedback about the book at my email address gkgupta@acm.org.

G.K. Gupta
gkgupta@acm.org

Introduction To Data Mining With Case Studies



Publisher : [PHI Learning](#)

ISBN : [9788120350021](#)

Author : [Gupta](#)

Type the URL : <http://www.kopykitab.com/product/10277>



Get this eBook