

Third Edition



DATA MINING

Concepts and Techniques

MK
MORGAN KAUFMANN

Jiawei Han | Micheline Kamber | Jian Pei

Data Mining

Third Edition

The Morgan Kaufmann Series in Data Management Systems (Selected Titles)

Joe Celko's Data, Measurements, and Standards in SQL

Joe Celko

Information Modeling and Relational Databases, 2nd Edition

Terry Halpin, Tony Morgan

Joe Celko's Thinking in Sets

Joe Celko

Business Metadata

Bill Inmon, Bonnie O'Neil, Lowell Fryman

Unleashing Web 2.0

Gottfried Vossen, Stephan Hagemann

Enterprise Knowledge Management

David Loshin

The Practitioner's Guide to Data Quality Improvement

David Loshin

Business Process Change, 2nd Edition

Paul Harmon

IT Manager's Handbook, 2nd Edition

Bill Holtsnider, Brian Jaffe

Joe Celko's Puzzles and Answers, 2nd Edition

Joe Celko

Architecture and Patterns for IT Service Management, 2nd Edition, Resource Planning and Governance

Charles Betz

Joe Celko's Analytics and OLAP in SQL

Joe Celko

Data Preparation for Data Mining Using SAS

Mamdouh Refaat

Querying XML: XQuery, XPath, and SQL/ XML in Context

Jim Melton, Stephen Buxton

Data Mining: Concepts and Techniques, 3rd Edition

Jiawei Han, Micheline Kamber, Jian Pei

Database Modeling and Design: Logical Design, 5th Edition

Toby J. Teorey, Sam S. Lightstone, Thomas P. Nadeau, H. V. Jagadish

Foundations of Multidimensional and Metric Data Structures

Hanan Samet

Joe Celko's SQL for Smarties: Advanced SQL Programming, 4th Edition

Joe Celko

Moving Objects Databases

Ralf Hartmut Güting, Markus Schneider

Joe Celko's SQL Programming Style

Joe Celko

Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration

Earl Cox

Data Modeling Essentials, 3rd Edition

Graeme C. Simsion, Graham C. Witt

Developing High Quality Data Models

Matthew West

Location-Based Services

Jochen Schiller, Agnes Voisard

Managing Time in Relational Databases: How to Design, Update, and Query Temporal Data

Tom Johnston, Randall Weis

Database Modeling with Microsoft® Visio for Enterprise Architects

Terry Halpin, Ken Evans, Patrick Hallock, Bill Maclean

Designing Data-Intensive Web Applications

Stephano Ceri, Piero Fraternali, Aldo Bongio, Marco Brambilla, Sara Comai, Maristella Matera

Mining the Web: Discovering Knowledge from Hypertext Data

Soumen Chakrabarti

Advanced SQL: 1999—Understanding Object-Relational and Other Advanced Features

Jim Melton

Database Tuning: Principles, Experiments, and Troubleshooting Techniques

Dennis Shasha, Philippe Bonnet

SQL: 1999—Understanding Relational Language Components

Jim Melton, Alan R. Simon

Information Visualization in Data Mining and Knowledge Discovery

Edited by Usama Fayyad, Georges G. Grinstein, Andreas Wierse

Transactional Information Systems

Gerhard Weikum, Gottfried Vossen

Spatial Databases

Philippe Rigaux, Michel Scholl, and Agnes Voisard

Managing Reference Data in Enterprise Databases

Malcolm Chisholm

Understanding SQL and Java Together

Jim Melton, Andrew Eisenberg

Database: Principles, Programming, and Performance, 2nd Edition

Patrick and Elizabeth O'Neil

The Object Data Standard

Edited by R. G. G. Cattell, Douglas Barry

Data on the Web: From Relations to Semistructured Data and XML

Serge Abiteboul, Peter Buneman, Dan Suciu

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 3rd Edition

Ian Witten, Eibe Frank, Mark A. Hall

Joe Celko's Data and Databases: Concepts in Practice

Joe Celko

Developing Time-Oriented Database Applications in SQL

Richard T. Snodgrass

Web Farming for the Data Warehouse

Richard D. Hackathorn

Management of Heterogeneous and Autonomous Database Systems

Edited by Ahmed Elmagarmid, Marek Rusinkiewicz, Amit Sheth

Object-Relational DBMSs, 2nd Edition

Michael Stonebraker, Paul Brown, with Dorothy Moore

Universal Database Management: A Guide to Object/Relational Technology

Cynthia Maro Saracco

Readings in Database Systems, 3rd Edition

Edited by Michael Stonebraker, Joseph M. Hellerstein

Understanding SQL's Stored Procedures: A Complete Guide to SQL/PSM

Jim Melton

Principles of Multimedia Database Systems

V. S. Subrahmanian

Principles of Database Query Processing for Advanced Applications

Clement T. Yu, Weiyi Meng

Advanced Database Systems

Carlo Zaniolo, Stefano Ceri, Christos Faloutsos, Richard T. Snodgrass, V. S. Subrahmanian, Roberto Zicari

Principles of Transaction Processing, 2nd Edition

Philip A. Bernstein, Eric Newcomer

Using the New DB2: IBM's Object-Relational Database System

Don Chamberlin

Distributed Algorithms

Nancy A. Lynch

Active Database Systems: Triggers and Rules for Advanced Database Processing

Edited by Jennifer Widom, Stefano Ceri

Migrating Legacy Systems: Gateways, Interfaces, and the Incremental Approach

Michael L. Brodie, Michael Stonebraker

Atomic Transactions

Nancy Lynch, Michael Merritt, William Weihl, Alan Fekete

Query Processing for Advanced Database Systems

Edited by Johann Christoph Freytag, David Maier, Gottfried Vossen

Transaction Processing

Jim Gray, Andreas Reuter

Database Transaction Models for Advanced Applications

Edited by Ahmed K. Elmagarmid

A Guide to Developing Client/Server SQL Applications

Setrag Khoshafian, Arvola Chan, Anna Wong, Harry K. T. Wong

Data Mining Concepts and Techniques

Third Edition

Jiawei Han

University of Illinois at Urbana–Champaign

Micheline Kamber

Jian Pei

Simon Fraser University



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Morgan Kaufmann is an imprint of Elsevier



Morgan Kaufmann Publishers is an imprint of Elsevier.
225 Wyman Street, Waltham, MA 02451, USA

© 2012 by Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods or professional practices, may become necessary. Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information or methods described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

Han, Jiawei.

Data mining : concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed.

p. cm.

ISBN 978-0-12-381479-1

1. Data mining. I. Kamber, Micheline. II. Pei, Jian. III. Title.

QA76.9.D343H36 2011

006.3'12–dc22

2011010635

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

For information on all Morgan Kaufmann publications, visit our
Web site at www.mkp.com or www.elsevierdirect.com

Printed in the United States of America

11 12 13 14 15 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

To Y. Dora and Lawrence for your love and encouragement

J.H.

To Erik, Kevan, Kian, and Mikael for your love and inspiration

M.K.

To my wife, Jennifer, and daughter, Jacqueline

J.P.

This page intentionally left blank

Contents

	Foreword	xix
	Foreword to Second Edition	xxi
	Preface	xxiii
	Acknowledgments	xxxi
	About the Authors	xxxv
Chapter 1	Introduction	1
1.1	Why Data Mining?	1
1.1.1	Moving toward the Information Age	1
1.1.2	Data Mining as the Evolution of Information Technology	2
1.2	What Is Data Mining?	5
1.3	What Kinds of Data Can Be Mined?	8
1.3.1	Database Data	9
1.3.2	Data Warehouses	10
1.3.3	Transactional Data	13
1.3.4	Other Kinds of Data	14
1.4	What Kinds of Patterns Can Be Mined?	15
1.4.1	Class/Concept Description: Characterization and Discrimination	15
1.4.2	Mining Frequent Patterns, Associations, and Correlations	17
1.4.3	Classification and Regression for Predictive Analysis	18
1.4.4	Cluster Analysis	19
1.4.5	Outlier Analysis	20
1.4.6	Are All Patterns Interesting?	21
1.5	Which Technologies Are Used?	23
1.5.1	Statistics	23
1.5.2	Machine Learning	24
1.5.3	Database Systems and Data Warehouses	26
1.5.4	Information Retrieval	26

1.6	Which Kinds of Applications Are Targeted?	27
1.6.1	Business Intelligence	27
1.6.2	Web Search Engines	28
1.7	Major Issues in Data Mining	29
1.7.1	Mining Methodology	29
1.7.2	User Interaction	30
1.7.3	Efficiency and Scalability	31
1.7.4	Diversity of Database Types	32
1.7.5	Data Mining and Society	32
1.8	Summary	33
1.9	Exercises	34
1.10	Bibliographic Notes	35
Chapter 2	Getting to Know Your Data	39
2.1	Data Objects and Attribute Types	40
2.1.1	What Is an Attribute?	40
2.1.2	Nominal Attributes	41
2.1.3	Binary Attributes	41
2.1.4	Ordinal Attributes	42
2.1.5	Numeric Attributes	43
2.1.6	Discrete versus Continuous Attributes	44
2.2	Basic Statistical Descriptions of Data	44
2.2.1	Measuring the Central Tendency: Mean, Median, and Mode	45
2.2.2	Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range	48
2.2.3	Graphic Displays of Basic Statistical Descriptions of Data	51
2.3	Data Visualization	56
2.3.1	Pixel-Oriented Visualization Techniques	57
2.3.2	Geometric Projection Visualization Techniques	58
2.3.3	Icon-Based Visualization Techniques	60
2.3.4	Hierarchical Visualization Techniques	63
2.3.5	Visualizing Complex Data and Relations	64
2.4	Measuring Data Similarity and Dissimilarity	65
2.4.1	Data Matrix versus Dissimilarity Matrix	67
2.4.2	Proximity Measures for Nominal Attributes	68
2.4.3	Proximity Measures for Binary Attributes	70
2.4.4	Dissimilarity of Numeric Data: Minkowski Distance	72
2.4.5	Proximity Measures for Ordinal Attributes	74
2.4.6	Dissimilarity for Attributes of Mixed Types	75
2.4.7	Cosine Similarity	77
2.5	Summary	79
2.6	Exercises	79
2.7	Bibliographic Notes	81

Chapter 3	Data Preprocessing	83
3.1	Data Preprocessing: An Overview	84
3.1.1	Data Quality: Why Preprocess the Data?	84
3.1.2	Major Tasks in Data Preprocessing	85
3.2	Data Cleaning	88
3.2.1	Missing Values	88
3.2.2	Noisy Data	89
3.2.3	Data Cleaning as a Process	91
3.3	Data Integration	93
3.3.1	Entity Identification Problem	94
3.3.2	Redundancy and Correlation Analysis	94
3.3.3	Tuple Duplication	98
3.3.4	Data Value Conflict Detection and Resolution	99
3.4	Data Reduction	99
3.4.1	Overview of Data Reduction Strategies	99
3.4.2	Wavelet Transforms	100
3.4.3	Principal Components Analysis	102
3.4.4	Attribute Subset Selection	103
3.4.5	Regression and Log-Linear Models: Parametric Data Reduction	105
3.4.6	Histograms	106
3.4.7	Clustering	108
3.4.8	Sampling	108
3.4.9	Data Cube Aggregation	110
3.5	Data Transformation and Data Discretization	111
3.5.1	Data Transformation Strategies Overview	112
3.5.2	Data Transformation by Normalization	113
3.5.3	Discretization by Binning	115
3.5.4	Discretization by Histogram Analysis	115
3.5.5	Discretization by Cluster, Decision Tree, and Correlation Analyses	116
3.5.6	Concept Hierarchy Generation for Nominal Data	117
3.6	Summary	120
3.7	Exercises	121
3.8	Bibliographic Notes	123
Chapter 4	Data Warehousing and Online Analytical Processing	125
4.1	Data Warehouse: Basic Concepts	125
4.1.1	What Is a Data Warehouse?	126
4.1.2	Differences between Operational Database Systems and Data Warehouses	128
4.1.3	But, Why Have a Separate Data Warehouse?	129

4.1.4	Data Warehousing: A Multitiered Architecture	130
4.1.5	Data Warehouse Models: Enterprise Warehouse, Data Mart, and Virtual Warehouse	132
4.1.6	Extraction, Transformation, and Loading	134
4.1.7	Metadata Repository	134
4.2	Data Warehouse Modeling: Data Cube and OLAP	135
4.2.1	Data Cube: A Multidimensional Data Model	136
4.2.2	Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models	139
4.2.3	Dimensions: The Role of Concept Hierarchies	142
4.2.4	Measures: Their Categorization and Computation	144
4.2.5	Typical OLAP Operations	146
4.2.6	A Starlet Query Model for Querying Multidimensional Databases	149
4.3	Data Warehouse Design and Usage	150
4.3.1	A Business Analysis Framework for Data Warehouse Design	150
4.3.2	Data Warehouse Design Process	151
4.3.3	Data Warehouse Usage for Information Processing	153
4.3.4	From Online Analytical Processing to Multidimensional Data Mining	155
4.4	Data Warehouse Implementation	156
4.4.1	Efficient Data Cube Computation: An Overview	156
4.4.2	Indexing OLAP Data: Bitmap Index and Join Index	160
4.4.3	Efficient Processing of OLAP Queries	163
4.4.4	OLAP Server Architectures: ROLAP versus MOLAP versus HOLAP	164
4.5	Data Generalization by Attribute-Oriented Induction	166
4.5.1	Attribute-Oriented Induction for Data Characterization	167
4.5.2	Efficient Implementation of Attribute-Oriented Induction	172
4.5.3	Attribute-Oriented Induction for Class Comparisons	175
4.6	Summary	178
4.7	Exercises	180
4.8	Bibliographic Notes	184
Chapter 5	Data Cube Technology	187
5.1	Data Cube Computation: Preliminary Concepts	188
5.1.1	Cube Materialization: Full Cube, Iceberg Cube, Closed Cube, and Cube Shell	188
5.1.2	General Strategies for Data Cube Computation	192
5.2	Data Cube Computation Methods	194
5.2.1	Multiway Array Aggregation for Full Cube Computation	195

5.2.2	BUC: Computing Iceberg Cubes from the Apex Cuboid Downward	200
5.2.3	Star-Cubing: Computing Iceberg Cubes Using a Dynamic Star-Tree Structure	204
5.2.4	Precomputing Shell Fragments for Fast High-Dimensional OLAP	210
5.3	Processing Advanced Kinds of Queries by Exploring Cube Technology	218
5.3.1	Sampling Cubes: OLAP-Based Mining on Sampling Data	218
5.3.2	Ranking Cubes: Efficient Computation of Top-k Queries	225
5.4	Multidimensional Data Analysis in Cube Space	227
5.4.1	Prediction Cubes: Prediction Mining in Cube Space	227
5.4.2	Multifeature Cubes: Complex Aggregation at Multiple Granularities	230
5.4.3	Exception-Based, Discovery-Driven Cube Space Exploration	231
5.5	Summary	234
5.6	Exercises	235
5.7	Bibliographic Notes	240
Chapter 6	Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods	243
6.1	Basic Concepts	243
6.1.1	Market Basket Analysis: A Motivating Example	244
6.1.2	Frequent Itemsets, Closed Itemsets, and Association Rules	246
6.2	Frequent Itemset Mining Methods	248
6.2.1	Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation	248
6.2.2	Generating Association Rules from Frequent Itemsets	254
6.2.3	Improving the Efficiency of Apriori	254
6.2.4	A Pattern-Growth Approach for Mining Frequent Itemsets	257
6.2.5	Mining Frequent Itemsets Using Vertical Data Format	259
6.2.6	Mining Closed and Max Patterns	262
6.3	Which Patterns Are Interesting?—Pattern Evaluation Methods	264
6.3.1	Strong Rules Are Not Necessarily Interesting	264
6.3.2	From Association Analysis to Correlation Analysis	265
6.3.3	A Comparison of Pattern Evaluation Measures	267
6.4	Summary	271
6.5	Exercises	273
6.6	Bibliographic Notes	276

Chapter 7	Advanced Pattern Mining	279
7.1	Pattern Mining: A Road Map	279
7.2	Pattern Mining in Multilevel, Multidimensional Space	283
7.2.1	Mining Multilevel Associations	283
7.2.2	Mining Multidimensional Associations	287
7.2.3	Mining Quantitative Association Rules	289
7.2.4	Mining Rare Patterns and Negative Patterns	291
7.3	Constraint-Based Frequent Pattern Mining	294
7.3.1	Metarule-Guided Mining of Association Rules	295
7.3.2	Constraint-Based Pattern Generation: Pruning Pattern Space and Pruning Data Space	296
7.4	Mining High-Dimensional Data and Colossal Patterns	301
7.4.1	Mining Colossal Patterns by Pattern-Fusion	302
7.5	Mining Compressed or Approximate Patterns	307
7.5.1	Mining Compressed Patterns by Pattern Clustering	308
7.5.2	Extracting Redundancy-Aware Top-k Patterns	310
7.6	Pattern Exploration and Application	313
7.6.1	Semantic Annotation of Frequent Patterns	313
7.6.2	Applications of Pattern Mining	317
7.7	Summary	319
7.8	Exercises	321
7.9	Bibliographic Notes	323
Chapter 8	Classification: Basic Concepts	327
8.1	Basic Concepts	327
8.1.1	What Is Classification?	327
8.1.2	General Approach to Classification	328
8.2	Decision Tree Induction	330
8.2.1	Decision Tree Induction	332
8.2.2	Attribute Selection Measures	336
8.2.3	Tree Pruning	344
8.2.4	Scalability and Decision Tree Induction	347
8.2.5	Visual Mining for Decision Tree Induction	348
8.3	Bayes Classification Methods	350
8.3.1	Bayes' Theorem	350
8.3.2	Naïve Bayesian Classification	351
8.4	Rule-Based Classification	355
8.4.1	Using IF-THEN Rules for Classification	355
8.4.2	Rule Extraction from a Decision Tree	357
8.4.3	Rule Induction Using a Sequential Covering Algorithm	359

8.5	Model Evaluation and Selection	364
8.5.1	Metrics for Evaluating Classifier Performance	364
8.5.2	Holdout Method and Random Subsampling	370
8.5.3	Cross-Validation	370
8.5.4	Bootstrap	371
8.5.5	Model Selection Using Statistical Tests of Significance	372
8.5.6	Comparing Classifiers Based on Cost–Benefit and ROC Curves	373
8.6	Techniques to Improve Classification Accuracy	377
8.6.1	Introducing Ensemble Methods	378
8.6.2	Bagging	379
8.6.3	Boosting and AdaBoost	380
8.6.4	Random Forests	382
8.6.5	Improving Classification Accuracy of Class-Imbalanced Data	383
8.7	Summary	385
8.8	Exercises	386
8.9	Bibliographic Notes	389
Chapter 9	Classification: Advanced Methods	393
9.1	Bayesian Belief Networks	393
9.1.1	Concepts and Mechanisms	394
9.1.2	Training Bayesian Belief Networks	396
9.2	Classification by Backpropagation	398
9.2.1	A Multilayer Feed-Forward Neural Network	398
9.2.2	Defining a Network Topology	400
9.2.3	Backpropagation	400
9.2.4	Inside the Black Box: Backpropagation and Interpretability	406
9.3	Support Vector Machines	408
9.3.1	The Case When the Data Are Linearly Separable	408
9.3.2	The Case When the Data Are Linearly Inseparable	413
9.4	Classification Using Frequent Patterns	415
9.4.1	Associative Classification	416
9.4.2	Discriminative Frequent Pattern–Based Classification	419
9.5	Lazy Learners (or Learning from Your Neighbors)	422
9.5.1	k-Nearest-Neighbor Classifiers	423
9.5.2	Case-Based Reasoning	425
9.6	Other Classification Methods	426
9.6.1	Genetic Algorithms	426
9.6.2	Rough Set Approach	427
9.6.3	Fuzzy Set Approaches	428
9.7	Additional Topics Regarding Classification	429
9.7.1	Multiclass Classification	430

	9.7.2	Semi-Supervised Classification	432
	9.7.3	Active Learning	433
	9.7.4	Transfer Learning	434
	9.8	Summary	436
	9.9	Exercises	438
	9.10	Bibliographic Notes	439
Chapter 10		Cluster Analysis: Basic Concepts and Methods	443
	10.1	Cluster Analysis	444
	10.1.1	What Is Cluster Analysis?	444
	10.1.2	Requirements for Cluster Analysis	445
	10.1.3	Overview of Basic Clustering Methods	448
	10.2	Partitioning Methods	451
	10.2.1	k-Means: A Centroid-Based Technique	451
	10.2.2	k-Medoids: A Representative Object-Based Technique	454
	10.3	Hierarchical Methods	457
	10.3.1	Agglomerative versus Divisive Hierarchical Clustering	459
	10.3.2	Distance Measures in Algorithmic Methods	461
	10.3.3	BIRCH: Multiphase Hierarchical Clustering Using Clustering Feature Trees	462
	10.3.4	Chameleon: Multiphase Hierarchical Clustering Using Dynamic Modeling	466
	10.3.5	Probabilistic Hierarchical Clustering	467
	10.4	Density-Based Methods	471
	10.4.1	DBSCAN: Density-Based Clustering Based on Connected Regions with High Density	471
	10.4.2	OPTICS: Ordering Points to Identify the Clustering Structure	473
	10.4.3	DENCLUE: Clustering Based on Density Distribution Functions	476
	10.5	Grid-Based Methods	479
	10.5.1	STING: STatistical INformation Grid	479
	10.5.2	CLIQUE: An Apriori-like Subspace Clustering Method	481
	10.6	Evaluation of Clustering	483
	10.6.1	Assessing Clustering Tendency	484
	10.6.2	Determining the Number of Clusters	486
	10.6.3	Measuring Clustering Quality	487
	10.7	Summary	490
	10.8	Exercises	491
	10.9	Bibliographic Notes	494
Chapter 11		Advanced Cluster Analysis	497
	11.1	Probabilistic Model-Based Clustering	497
	11.1.1	Fuzzy Clusters	499

11.1.2	Probabilistic Model-Based Clusters	501
11.1.3	Expectation-Maximization Algorithm	505
11.2	Clustering High-Dimensional Data	508
11.2.1	Clustering High-Dimensional Data: Problems, Challenges, and Major Methodologies	508
11.2.2	Subspace Clustering Methods	510
11.2.3	Biclustering	512
11.2.4	Dimensionality Reduction Methods and Spectral Clustering	519
11.3	Clustering Graph and Network Data	522
11.3.1	Applications and Challenges	523
11.3.2	Similarity Measures	525
11.3.3	Graph Clustering Methods	528
11.4	Clustering with Constraints	532
11.4.1	Categorization of Constraints	533
11.4.2	Methods for Clustering with Constraints	535
11.5	Summary	538
11.6	Exercises	539
11.7	Bibliographic Notes	540
Chapter 12	Outlier Detection	543
12.1	Outliers and Outlier Analysis	544
12.1.1	What Are Outliers?	544
12.1.2	Types of Outliers	545
12.1.3	Challenges of Outlier Detection	548
12.2	Outlier Detection Methods	549
12.2.1	Supervised, Semi-Supervised, and Unsupervised Methods	549
12.2.2	Statistical Methods, Proximity-Based Methods, and Clustering-Based Methods	551
12.3	Statistical Approaches	553
12.3.1	Parametric Methods	553
12.3.2	Nonparametric Methods	558
12.4	Proximity-Based Approaches	560
12.4.1	Distance-Based Outlier Detection and a Nested Loop Method	561
12.4.2	A Grid-Based Method	562
12.4.3	Density-Based Outlier Detection	564
12.5	Clustering-Based Approaches	567
12.6	Classification-Based Approaches	571
12.7	Mining Contextual and Collective Outliers	573
12.7.1	Transforming Contextual Outlier Detection to Conventional Outlier Detection	573

	12.7.2 Modeling Normal Behavior with Respect to Contexts	574
	12.7.3 Mining Collective Outliers	575
12.8	Outlier Detection in High-Dimensional Data	576
	12.8.1 Extending Conventional Outlier Detection	577
	12.8.2 Finding Outliers in Subspaces	578
	12.8.3 Modeling High-Dimensional Outliers	579
12.9	Summary	581
12.10	Exercises	582
12.11	Bibliographic Notes	583
Chapter 13	Data Mining Trends and Research Frontiers	585
13.1	Mining Complex Data Types	585
	13.1.1 Mining Sequence Data: Time-Series, Symbolic Sequences, and Biological Sequences	586
	13.1.2 Mining Graphs and Networks	591
	13.1.3 Mining Other Kinds of Data	595
13.2	Other Methodologies of Data Mining	598
	13.2.1 Statistical Data Mining	598
	13.2.2 Views on Data Mining Foundations	600
	13.2.3 Visual and Audio Data Mining	602
13.3	Data Mining Applications	607
	13.3.1 Data Mining for Financial Data Analysis	607
	13.3.2 Data Mining for Retail and Telecommunication Industries	609
	13.3.3 Data Mining in Science and Engineering	611
	13.3.4 Data Mining for Intrusion Detection and Prevention	614
	13.3.5 Data Mining and Recommender Systems	615
13.4	Data Mining and Society	618
	13.4.1 Ubiquitous and Invisible Data Mining	618
	13.4.2 Privacy, Security, and Social Impacts of Data Mining	620
13.5	Data Mining Trends	622
13.6	Summary	625
13.7	Exercises	626
13.8	Bibliographic Notes	628
	Bibliography	633
	Index	673

Foreword

Analyzing large amounts of data is a necessity. Even popular science books, like “super crunchers,” give compelling cases where large amounts of data yield discoveries and intuitions that surprise even experts. Every enterprise benefits from collecting and analyzing its data: Hospitals can spot trends and anomalies in their patient records, search engines can do better ranking and ad placement, and environmental and public health agencies can spot patterns and abnormalities in their data. The list continues, with cybersecurity and computer network intrusion detection; monitoring of the energy consumption of household appliances; pattern analysis in bioinformatics and pharmaceutical data; financial and business intelligence data; spotting trends in blogs, Twitter, and many more. Storage is inexpensive and getting even less so, as are data sensors. Thus, collecting and storing data is easier than ever before.

The problem then becomes *how to analyze* the data. This is exactly the focus of this Third Edition of the book. Jiawei, Micheline, and Jian give encyclopedic coverage of all the related methods, from the classic topics of clustering and classification, to database methods (e.g., association rules, data cubes) to more recent and advanced topics (e.g., SVD/PCA, wavelets, support vector machines).

The exposition is extremely accessible to beginners and advanced readers alike. The book gives the fundamental material first and the more advanced material in follow-up chapters. It also has numerous rhetorical questions, which I found extremely helpful for maintaining focus.

We have used the first two editions as textbooks in data mining courses at Carnegie Mellon and plan to continue to do so with this Third Edition. The new version has significant additions: Notably, it has more than 100 citations to works from 2006 onward, focusing on more recent material such as graphs and social networks, sensor networks, and outlier detection. This book has a new section for visualization, has expanded outlier detection into a whole chapter, and has separate chapters for advanced

methods—for example, pattern mining with top- k patterns and more and clustering methods with biclustering and graph clustering.

Overall, it is an excellent book on classic and modern data mining methods, and it is ideal not only for teaching but also as a reference book.

Christos Faloutsos
Carnegie Mellon University

Foreword to Second Edition

We are deluged by data—scientific data, medical data, demographic data, financial data, and marketing data. People have no time to look at this data. Human attention has become the precious resource. So, we must find ways to automatically analyze the data, to automatically classify it, to automatically summarize it, to automatically discover and characterize trends in it, and to automatically flag anomalies. This is one of the most active and exciting areas of the database research community. Researchers in areas including statistics, visualization, artificial intelligence, and machine learning are contributing to this field. The breadth of the field makes it difficult to grasp the extraordinary progress over the last few decades.

Six years ago, Jiawei Han's and Micheline Kamber's seminal textbook organized and presented Data Mining. It heralded a golden age of innovation in the field. This revision of their book reflects that progress; more than half of the references and historical notes are to recent work. The field has matured with many new and improved algorithms, and has broadened to include many more datatypes: streams, sequences, graphs, time-series, geospatial, audio, images, and video. We are certainly not at the end of the golden age—indeed research and commercial interest in data mining continues to grow—but we are all fortunate to have this modern compendium.

The book gives quick introductions to database and data mining concepts with particular emphasis on data analysis. It then covers in a chapter-by-chapter tour the concepts and techniques that underlie classification, prediction, association, and clustering. These topics are presented with examples, a tour of the best algorithms for each problem class, and with pragmatic rules of thumb about when to apply each technique. The Socratic presentation style is both very readable and very informative. I certainly learned a lot from reading the first edition and got re-educated and updated in reading the second edition.

Jiawei Han and Micheline Kamber have been leading contributors to data mining research. This is the text they use with their students to bring them up to speed on

the field. The field is evolving very rapidly, but this book is a quick way to learn the basic ideas, and to understand where the field is today. I found it very informative and stimulating, and believe you will too.

Jim Gray
In his memory

Preface

The computerization of our society has substantially enhanced our capabilities for both generating and collecting data from diverse sources. A tremendous amount of data has flooded almost every aspect of our lives. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called *data mining*, and its various applications. Data mining, also popularly referred to as *knowledge discovery from data (KDD)*, is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams.

This book explores the concepts and techniques of *knowledge discovery* and *data mining*. As a multidisciplinary field, data mining draws on work from areas including statistics, machine learning, pattern recognition, database technology, information retrieval, network science, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We focus on issues relating to the feasibility, usefulness, effectiveness, and scalability of techniques for the discovery of patterns hidden in *large data sets*. As a result, this book is not intended as an introduction to statistics, machine learning, database systems, or other such areas, although we do provide some background knowledge to facilitate the reader's comprehension of their respective roles in data mining. Rather, the book is a comprehensive introduction to data mining. It is useful for computing science students, application developers, and business professionals, as well as researchers involved in any of the disciplines previously listed.

Data mining emerged during the late 1980s, made great strides during the 1990s, and continues to flourish into the new millennium. This book presents an overall picture of the field, introducing interesting data mining techniques and systems and discussing applications and research directions. An important motivation for writing this book was the need to build an organized framework for the study of data mining—a challenging task, owing to the extensive multidisciplinary nature of this fast-developing field. We hope that this book will encourage people with different backgrounds and experiences to exchange their views regarding data mining so as to contribute toward the further promotion and shaping of this exciting and dynamic field.

Organization of the Book

Since the publication of the first two editions of this book, great progress has been made in the field of data mining. Many new data mining methodologies, systems, and applications have been developed, especially for handling new kinds of data, including information networks, graphs, complex structures, and data streams, as well as text, Web, multimedia, time-series, and spatiotemporal data. Such fast development and rich, new technical contents make it difficult to cover the full spectrum of the field in a single book. Instead of continuously expanding the coverage of this book, we have decided to cover the core material in sufficient scope and depth, and leave the handling of complex data types to a separate forthcoming book.

The third edition substantially revises the first two editions of the book, with numerous enhancements and a reorganization of the technical contents. The core technical material, which handles mining on general data types, is expanded and substantially enhanced. Several individual chapters for topics from the second edition (e.g., data pre-processing, frequent pattern mining, classification, and clustering) are now augmented and each split into two chapters for this new edition. For these topics, one chapter encapsulates the basic concepts and techniques while the other presents advanced concepts and methods.

Chapters from the second edition on mining complex data types (e.g., stream data, sequence data, graph-structured data, social network data, and multirelational data, as well as text, Web, multimedia, and spatiotemporal data) are now reserved for a new book that will be dedicated to *advanced topics in data mining*. Still, to support readers in learning such advanced topics, we have placed an electronic version of the relevant chapters from the second edition onto the book's web site as companion material for the third edition.

The chapters of the third edition are described briefly as follows, with emphasis on the new material.

Chapter 1 provides an *introduction* to the multidisciplinary field of data mining. It discusses the evolutionary path of information technology, which has led to the need for data mining, and the importance of its applications. It examines the data types to be mined, including relational, transactional, and data warehouse data, as well as complex data types such as time-series, sequences, data streams, spatiotemporal data, multimedia data, text data, graphs, social networks, and Web data. The chapter presents a general classification of data mining tasks, based on the kinds of knowledge to be mined, the kinds of technologies used, and the kinds of applications that are targeted. Finally, major challenges in the field are discussed.

Chapter 2 introduces the *general data features*. It first discusses data objects and attribute types and then introduces typical measures for basic statistical data descriptions. It overviews data visualization techniques for various kinds of data. In addition to methods of numeric data visualization, methods for visualizing text, tags, graphs, and multidimensional data are introduced. Chapter 2 also introduces ways to measure similarity and dissimilarity for various kinds of data.

Chapter 3 introduces *techniques for data preprocessing*. It first introduces the concept of data quality and then discusses methods for data cleaning, data integration, data reduction, data transformation, and data discretization.

Chapters 4 and 5 provide a solid introduction to *data warehouses*, *OLAP* (online analytical processing), and *data cube technology*. **Chapter 4** introduces the basic concepts, modeling, design architectures, and general implementations of data warehouses and OLAP, as well as the relationship between data warehousing and other data generalization methods. **Chapter 5** takes an in-depth look at data cube technology, presenting a detailed study of methods of data cube computation, including Star-Cubing and high-dimensional OLAP methods. Further explorations of data cube and OLAP technologies are discussed, such as sampling cubes, ranking cubes, prediction cubes, multifeature cubes for complex analysis queries, and discovery-driven cube exploration.

Chapters 6 and 7 present methods for *mining frequent patterns, associations, and correlations* in large data sets. **Chapter 6** introduces fundamental concepts, such as market basket analysis, with many techniques for frequent itemset mining presented in an organized way. These range from the basic Apriori algorithm and its variations to more advanced methods that improve efficiency, including the frequent pattern growth approach, frequent pattern mining with vertical data format, and mining closed and max frequent itemsets. The chapter also discusses pattern evaluation methods and introduces measures for mining correlated patterns. **Chapter 7** is on advanced pattern mining methods. It discusses methods for pattern mining in multi-level and multidimensional space, mining rare and negative patterns, mining colossal patterns and high-dimensional data, constraint-based pattern mining, and mining compressed or approximate patterns. It also introduces methods for pattern exploration and application, including semantic annotation of frequent patterns.

Chapters 8 and 9 describe methods for *data classification*. Due to the importance and diversity of classification methods, the contents are partitioned into two chapters. **Chapter 8** introduces basic concepts and methods for classification, including decision tree induction, Bayes classification, and rule-based classification. It also discusses model evaluation and selection methods and methods for improving classification accuracy, including ensemble methods and how to handle imbalanced data. **Chapter 9** discusses advanced methods for classification, including Bayesian belief networks, the neural network technique of backpropagation, support vector machines, classification using frequent patterns, *k*-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough set theory, and fuzzy set approaches. Additional topics include multiclass classification, semi-supervised classification, active learning, and transfer learning.

Cluster analysis forms the topic of Chapters 10 and 11. **Chapter 10** introduces the basic concepts and methods for data clustering, including an overview of basic cluster analysis methods, partitioning methods, hierarchical methods, density-based methods, and grid-based methods. It also introduces methods for the evaluation of clustering. **Chapter 11** discusses advanced methods for clustering, including probabilistic model-based clustering, clustering high-dimensional data, clustering graph and network data, and clustering with constraints.

Chapter 12 is dedicated to *outlier detection*. It introduces the basic concepts of outliers and outlier analysis and discusses various outlier detection methods from the view of degree of supervision (i.e., supervised, semi-supervised, and unsupervised methods), as well as from the view of approaches (i.e., statistical methods, proximity-based methods, clustering-based methods, and classification-based methods). It also discusses methods for mining contextual and collective outliers, and for outlier detection in high-dimensional data.

Finally, in **Chapter 13**, we discuss *trends, applications, and research frontiers* in data mining. We briefly cover mining complex data types, including mining sequence data (e.g., time series, symbolic sequences, and biological sequences), mining graphs and networks, and mining spatial, multimedia, text, and Web data. In-depth treatment of data mining methods for such data is left to a book on advanced topics in data mining, the writing of which is in progress. The chapter then moves ahead to cover other data mining methodologies, including statistical data mining, foundations of data mining, visual and audio data mining, as well as data mining applications. It discusses data mining for financial data analysis, for industries like retail and telecommunication, for use in science and engineering, and for intrusion detection and prevention. It also discusses the relationship between data mining and recommender systems. Because data mining is present in many aspects of daily life, we discuss issues regarding data mining and society, including ubiquitous and invisible data mining, as well as privacy, security, and the social impacts of data mining. We conclude our study by looking at data mining trends.

Throughout the text, *italic* font is used to emphasize terms that are defined, while **bold** font is used to highlight or summarize main ideas. Sans serif font is used for reserved words. Bold italic font is used to represent multidimensional quantities.

This book has several strong features that set it apart from other texts on data mining. It presents a very broad yet in-depth coverage of the principles of data mining. The chapters are written to be as self-contained as possible, so they may be read in order of interest by the reader. Advanced chapters offer a larger-scale view and may be considered optional for interested readers. All of the major methods of data mining are presented. The book presents important topics in data mining regarding multidimensional OLAP analysis, which is often overlooked or minimally treated in other data mining books. The book also maintains web sites with a number of online resources to aid instructors, students, and professionals in the field. These are described further in the following.

To the Instructor

This book is designed to give a broad, yet detailed overview of the data mining field. It can be used to teach an introductory course on data mining at an advanced undergraduate level or at the first-year graduate level. Sample course syllabi are provided on the book's web sites (www.cs.uiuc.edu/~hanj/bk3 and www.booksite.mkp.com/datamining3e) in addition to extensive teaching resources such as lecture slides, instructors' manuals, and reading lists (see p. xxix).

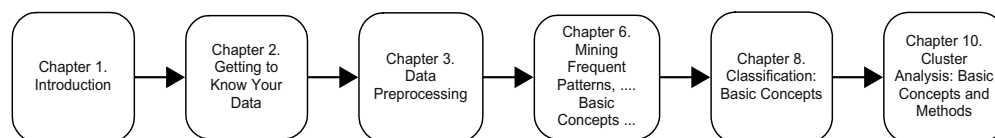


Figure P.1 A suggested sequence of chapters for a short introductory course.

Depending on the length of the instruction period, the background of students, and your interests, you may select subsets of chapters to teach in various sequential orderings. For example, if you would like to give only a short introduction to students on data mining, you may follow the suggested sequence in Figure P.1. Notice that depending on the need, you can also omit some sections or subsections in a chapter if desired.

Depending on the length of the course and its technical scope, you may choose to selectively add more chapters to this preliminary sequence. For example, instructors who are more interested in advanced classification methods may first add “Chapter 9. Classification: Advanced Methods”; those more interested in pattern mining may choose to include “Chapter 7. Advanced Pattern Mining”; whereas those interested in OLAP and data cube technology may like to add “Chapter 4. Data Warehousing and Online Analytical Processing” and “Chapter 5. Data Cube Technology.”

Alternatively, you may choose to teach the whole book in a two-course sequence that covers all of the chapters in the book, plus, when time permits, some advanced topics such as graph and network mining. Material for such advanced topics may be selected from the companion chapters available from the book’s web site, accompanied with a set of selected research papers.

Individual chapters in this book can also be used for tutorials or for special topics in related courses, such as machine learning, pattern recognition, data warehousing, and intelligent data analysis.

Each chapter ends with a set of exercises, suitable as assigned homework. The exercises are either short questions that test basic mastery of the material covered, longer questions that require analytical thinking, or implementation projects. Some exercises can also be used as research discussion topics. The bibliographic notes at the end of each chapter can be used to find the research literature that contains the origin of the concepts and methods presented, in-depth treatment of related topics, and possible extensions.

To the Student

We hope that this textbook will spark your interest in the young yet fast-evolving field of data mining. We have attempted to present the material in a clear manner, with careful explanation of the topics covered. Each chapter ends with a summary describing the main points. We have included many figures and illustrations throughout the text to make the book more enjoyable and reader-friendly. Although this book was designed as a textbook, we have tried to organize it so that it will also be useful to you as a reference

book or handbook, should you later decide to perform in-depth research in the related fields or pursue a career in data mining.

What do you need to know to read this book?

- You should have some knowledge of the concepts and terminology associated with statistics, database systems, and machine learning. However, we do try to provide enough background of the basics, so that if you are not so familiar with these fields or your memory is a bit rusty, you will not have trouble following the discussions in the book.
- You should have some programming experience. In particular, you should be able to read pseudocode and understand simple data structures such as multidimensional arrays.

To the Professional

This book was designed to cover a wide range of topics in the data mining field. As a result, it is an excellent handbook on the subject. Because each chapter is designed to be as standalone as possible, you can focus on the topics that most interest you. The book can be used by application programmers and information service managers who wish to learn about the key ideas of data mining on their own. The book would also be useful for technical data analysis staff in banking, insurance, medicine, and retailing industries who are interested in applying data mining solutions to their businesses. Moreover, the book may serve as a comprehensive survey of the data mining field, which may also benefit researchers who would like to advance the state-of-the-art in data mining and extend the scope of data mining applications.

The techniques and algorithms presented are of practical utility. Rather than selecting algorithms that perform well on small “toy” data sets, the algorithms described in the book are geared for the discovery of patterns and knowledge hidden in large, real data sets. Algorithms presented in the book are illustrated in pseudocode. The pseudocode is similar to the C programming language, yet is designed so that it should be easy to follow by programmers unfamiliar with C or C++. If you wish to implement any of the algorithms, you should find the translation of our pseudocode into the programming language of your choice to be a fairly straightforward task.

Book Web Sites with Resources

The book has a web site at www.cs.uiuc.edu/~hanj/bk3 and another with Morgan Kaufmann Publishers at www.booksite.mkp.com/datamining3e. These web sites contain many supplemental materials for readers of this book or anyone else with an interest in data mining. The resources include the following:

- **Slide presentations for each chapter.** Lecture notes in Microsoft PowerPoint slides are available for each chapter.

- **Companion chapters on advanced data mining.** Chapters 8 to 10 of the second edition of the book, which cover mining complex data types, are available on the book's web sites for readers who are interested in learning more about such advanced topics, beyond the themes covered in this book.
- **Instructors' manual.** This complete set of answers to the exercises in the book is available only to instructors from the publisher's web site.
- **Course syllabi and lecture plans.** These are given for undergraduate and graduate versions of introductory and advanced courses on data mining, which use the text and slides.
- **Supplemental reading lists with hyperlinks.** Seminal papers for supplemental reading are organized per chapter.
- **Links to data mining data sets and software.** We provide a set of links to data mining data sets and sites that contain interesting data mining software packages, such as IlliMine from the University of Illinois at Urbana-Champaign (<http://illimine.cs.uiuc.edu>).
- **Sample assignments, exams, and course projects.** A set of sample assignments, exams, and course projects is available to instructors from the publisher's web site.
- **Figures from the book.** This may help you to make your own slides for your classroom teaching.
- **Contents** of the book in PDF format.
- **Errata on the different printings of the book.** We encourage you to point out any errors in this book. Once the error is confirmed, we will update the errata list and include acknowledgment of your contribution.

Comments or suggestions can be sent to hanj@cs.uiuc.edu. We would be happy to hear from you.

This page intentionally left blank

Acknowledgments

Third Edition of the Book

We would like to express our grateful thanks to all of the previous and current members of the Data Mining Group at UIUC, the faculty and students in the Data and Information Systems (DAIS) Laboratory in the Department of Computer Science at the University of Illinois at Urbana-Champaign, and many friends and colleagues, whose constant support and encouragement have made our work on this edition a rewarding experience. We would also like to thank students in CS412 and CS512 classes at UIUC of the 2010–2011 academic year, who carefully went through the early drafts of this book, identified many errors, and suggested various improvements.

We also wish to thank David Bevans and Rick Adams at Morgan Kaufmann Publishers, for their enthusiasm, patience, and support during our writing of this edition of the book. We thank Marilyn Rash, the Project Manager, and her team members, for keeping us on schedule.

We are also grateful for the invaluable feedback from all of the reviewers. Moreover, we would like to thank U.S. National Science Foundation, NASA, U.S. Air Force Office of Scientific Research, U.S. Army Research Laboratory, and Natural Science and Engineering Research Council of Canada (NSERC), as well as IBM Research, Microsoft Research, Google, Yahoo! Research, Boeing, HP Labs, and other industry research labs for their support of our research in the form of research grants, contracts, and gifts. Such research support deepens our understanding of the subjects discussed in this book. Finally, we thank our families for their wholehearted support throughout this project.

Second Edition of the Book

We would like to express our grateful thanks to all of the previous and current members of the Data Mining Group at UIUC, the faculty and students in the Data and

Information Systems (DAIS) Laboratory in the Department of Computer Science at the University of Illinois at Urbana-Champaign, and many friends and colleagues, whose constant support and encouragement have made our work on this edition a rewarding experience. These include Gul Agha, Rakesh Agrawal, Loretta Auvil, Peter Bajcsy, Geneva Belford, Deng Cai, Y. Dora Cai, Roy Cambell, Kevin C.-C. Chang, Surajit Chaudhuri, Chen Chen, Yixin Chen, Yuguo Chen, Hong Cheng, David Cheung, Shengnan Cong, Gerald DeJong, AnHai Doan, Guozhu Dong, Charios Ermopoulos, Martin Ester, Christos Faloutsos, Wei Fan, Jack C. Feng, Ada Fu, Michael Garland, Johannes Gehrke, Hector Gonzalez, Mehdi Harandi, Thomas Huang, Wen Jin, Chulyun Kim, Sangkyum Kim, Won Kim, Won-Young Kim, David Kuck, Young-Koo Lee, Harris Lewin, Xiaolei Li, Yifan Li, Chao Liu, Han Liu, Huan Liu, Hongyan Liu, Lei Liu, Ying Lu, Klara Nahrstedt, David Padua, Jian Pei, Lenny Pitt, Daniel Reed, Dan Roth, Bruce Schatz, Zheng Shao, Marc Snir, Zhaohui Tang, Bhavani M. Thuraisingham, Josep Torrellas, Peter Tzvetkov, Benjamin W. Wah, Haixun Wang, Jianyong Wang, Ke Wang, Muyuan Wang, Wei Wang, Michael Welge, Marianne Winslett, Ouri Wolfson, Andrew Wu, Tianyi Wu, Dong Xin, Xifeng Yan, Jiong Yang, Xiaoxin Yin, Hwanjo Yu, Jeffrey X. Yu, Philip S. Yu, Maria Zemankova, ChengXiang Zhai, Yuanyuan Zhou, and Wei Zou.

Deng Cai and ChengXiang Zhai have contributed to the text mining and Web mining sections, Xifeng Yan to the graph mining section, and Xiaoxin Yin to the multirelational data mining section. Hong Cheng, Charios Ermopoulos, Hector Gonzalez, David J. Hill, Chulyun Kim, Sangkyum Kim, Chao Liu, Hongyan Liu, Kasif Manzoor, Tianyi Wu, Xifeng Yan, and Xiaoxin Yin have contributed to the proofreading of the individual chapters of the manuscript.

We also wish to thank Diane Cerra, our Publisher at Morgan Kaufmann Publishers, for her constant enthusiasm, patience, and support during our writing of this book. We are indebted to Alan Rose, the book Production Project Manager, for his tireless and ever-prompt communications with us to sort out all details of the production process. We are grateful for the invaluable feedback from all of the reviewers. Finally, we thank our families for their wholehearted support throughout this project.

First Edition of the Book

We would like to express our sincere thanks to all those who have worked or are currently working with us on data mining–related research and/or the DBMiner project, or have provided us with various support in data mining. These include Rakesh Agrawal, Stella Atkins, Yvan Bedard, Binay Bhattacharya, (Yandong) Dora Cai, Nick Cercone, Surajit Chaudhuri, Sonny H. S. Chee, Jianping Chen, Ming-Syan Chen, Qing Chen, Qiming Chen, Shan Cheng, David Cheung, Shi Cong, Son Dao, Umeshwar Dayal, James Delgrande, Guozhu Dong, Carole Edwards, Max Egenhofer, Martin Ester, Usama Fayyad, Ling Feng, Ada Fu, Yongjian Fu, Daphne Gelbart, Randy Goebel, Jim Gray, Robert Grossman, Wan Gong, Yike Guo, Eli Hagen, Howard Hamilton, Jing He, Larry Henschen, Jean Hou, Mei-Chun Hsu, Kan Hu, Haiming Huang, Yue Huang, Julia Itskevitch, Wen Jin, Tiko Kameda, Hiroyuki Kawano, Rizwan Kheraj, Eddie Kim, Won Kim, Krzysztof Koperski, Hans-Peter Kriegel, Vipin Kumar, Laks V. S. Lakshmanan, Joyce

Man Lam, James Lau, Deyi Li, George (Wenmin) Li, Jin Li, Ze-Nian Li, Nancy Liao, Gang Liu, Junqiang Liu, Ling Liu, Alan (Yijun) Lu, Hongjun Lu, Tong Lu, Wei Lu, Xuebin Lu, Wo-Shun Luk, Heikki Mannila, Runying Mao, Abhay Mehta, Gabor Melli, Alberto Mendelzon, Tim Merrett, Harvey Miller, Drew Miners, Behzad Mortazavi-Asl, Richard Muntz, Raymond T. Ng, Vicent Ng, Shojiro Nishio, Beng-Chin Ooi, Tamer Ozsu, Jian Pei, Gregory Piatetsky-Shapiro, Helen Pinto, Fred Popowich, Aymnmohamed Rajan, Peter Scheuermann, Shashi Shekhar, Wei-Min Shen, Avi Silberschatz, Evangelos Simoudis, Nebojsa Stefanovic, Yin Jenny Tam, Simon Tang, Zhaoxue Tang, Dick Tsur, Anthony K. H. Tung, Ke Wang, Wei Wang, Zhaoxia Wang, Tony Wind, Lara Winstone, Ju Wu, Betty (Bin) Xia, Cindy M. Xin, Xiaowei Xu, Qiang Yang, Yiwen Yin, Clement Yu, Jeffrey Yu, Philip S. Yu, Osmar R. Zaiane, Carlo Zaniolo, Shuhua Zhang, Zhong Zhang, Yvonne Zheng, Xiaofang Zhou, and Hua Zhu.

We are also grateful to Jean Hou, Helen Pinto, Lara Winstone, and Hua Zhu for their help with some of the original figures in this book, and to Eugene Belchev for his careful proofreading of each chapter.

We also wish to thank Diane Cerra, our Executive Editor at Morgan Kaufmann Publishers, for her enthusiasm, patience, and support during our writing of this book, as well as Howard Severson, our Production Editor, and his staff for their conscientious efforts regarding production. We are indebted to all of the reviewers for their invaluable feedback. Finally, we thank our families for their wholehearted support throughout this project.

This page intentionally left blank

About the Authors

Jiawei Han is a Bliss Professor of Engineering in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He has received numerous awards for his contributions on research into knowledge discovery and data mining, including ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), and IEEE W. Wallace McDowell Award (2009). He is a Fellow of ACM and IEEE. He served as founding Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data* (2006–2011) and as an editorial board member of several journals, including *IEEE Transactions on Knowledge and Data Engineering* and *Data Mining and Knowledge Discovery*.

Micheline Kamber has a master's degree in computer science (specializing in artificial intelligence) from Concordia University in Montreal, Quebec. She was an NSERC Scholar and has worked as a researcher at McGill University, Simon Fraser University, and in Switzerland. Her background in data mining and passion for writing in easy-to-understand terms help make this text a favorite of professionals, instructors, and students.

Jian Pei is currently an associate professor at the School of Computing Science, Simon Fraser University in British Columbia. He received a Ph.D. degree in computing science from Simon Fraser University in 2002 under Dr. Jiawei Han's supervision. He has published prolifically in the premier academic forums on data mining, databases, Web searching, and information retrieval and actively served the academic community. His publications have received thousands of citations and several prestigious awards. He is an associate editor of several data mining and data analytics journals.

This page intentionally left blank

Introduction

This book is an introduction to the young and fast-growing field of *data mining* (also known as *knowledge discovery from data*, or *KDD* for short). The book focuses on fundamental data mining concepts and techniques for discovering interesting patterns from data in various applications. In particular, we emphasize prominent techniques for developing effective, efficient, and scalable data mining tools.

This chapter is organized as follows. In Section 1.1, you will learn why data mining is in high demand and how it is part of the natural evolution of information technology. Section 1.2 defines data mining with respect to the knowledge discovery process. Next, you will learn about data mining from many aspects, such as the kinds of data that can be mined (Section 1.3), the kinds of knowledge to be mined (Section 1.4), the kinds of technologies to be used (Section 1.5), and targeted applications (Section 1.6). In this way, you will gain a multidimensional view of data mining. Finally, Section 1.7 outlines major data mining research and development issues.

Why Data Mining?

Necessity, who is the mother of invention. – Plato

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. Section 1.1.1 looks at how data mining can meet this need by providing tools to discover knowledge from data. In Section 1.1.2, we observe how data mining can be viewed as a result of the natural evolution of information technology.

1.1.1 Moving toward the Information Age

“*We are living in the information age*” is a popular saying; however, *we are actually living in the data age*. Terabytes or petabytes¹ of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business,

¹A petabyte is a unit of information or computer storage equal to 1 quadrillion bytes, or a thousand terabytes, or 1 million gigabytes.