

Data Mining: Concepts and Techniques

Data Mining: Concepts and Techniques

Second Edition

Jiawei Han

and

Micheline Kamber

University of Illinois at Urbana-Champaign



ELSEVIER

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



MORGAN KAUFMANN PUBLISHERS

Publisher Diane Cerra
Publishing Services Manager Simon Crump
Editorial Assistant Asma Stephan
Cover Design
Cover Image
Cover Illustration
Text Design
Composition diacriTech
Technical Illustration Dartmouth Publishing, Inc.
Copyeditor Multiscience Press
Proofreader Multiscience Press
Indexer Multiscience Press
Interior printer Maple-Vail Book Manufacturing Group
Cover printer Phoenix Color

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.
© 2006 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

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, scanning, or otherwise—without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.co.uk. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>) by selecting "Customer Support" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data

Application submitted

ISBN 13: 978-1-55860-901-3
ISBN 10: 1-55860-901-6

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

Printed in the United States of America
06 07 08 09 10 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

Dedication

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.

Contents

About the Author xvii

Foreword xix

Preface xxi

Chapter 1	Introduction	1
1.1	What Motivated Data Mining? Why Is It Important?	1
1.2	So, What Is Data Mining?	5
1.3	Data Mining—On What Kind of Data?	9
1.3.1	Relational Databases	10
1.3.2	Data Warehouses	12
1.3.3	Transactional Databases	14
1.3.4	Advanced Data and Information Systems and Advanced Applications	15
1.4	Data Mining Functionalities—What Kinds of Patterns Can Be Mined?	21
1.4.1	Concept/Class Description: Characterization and Discrimination	21
1.4.2	Mining Frequent Patterns, Associations, and Correlations	23
1.4.3	Classification and Prediction	24
1.4.4	Cluster Analysis	25
1.4.5	Outlier Analysis	26
1.4.6	Evolution Analysis	27
1.5	Are All of the Patterns Interesting?	27
1.6	Classification of Data Mining Systems	29
1.7	Data Mining Task Primitives	31
1.8	Integration of a Data Mining System with a Database or Data Warehouse System	34
1.9	Major Issues in Data Mining	36

1.10	Summary	39
	Exercises	40
	Bibliographic Notes	42
Chapter 2	Data Preprocessing	47
2.1	Why Preprocess the Data?	48
2.2	Descriptive Data Summarization	51
2.2.1	Measuring the Central Tendency	51
2.2.2	Measuring the Dispersion of Data	53
2.2.3	Graphic Displays of Basic Descriptive Data Summaries	56
2.3	Data Cleaning	61
2.3.1	Missing Values	61
2.3.2	Noisy Data	62
2.3.3	Data Cleaning as a Process	65
2.4	Data Integration and Transformation	67
2.4.1	Data Integration	67
2.4.2	Data Transformation	70
2.5	Data Reduction	72
2.5.1	Data Cube Aggregation	73
2.5.2	Attribute Subset Selection	75
2.5.3	Dimensionality Reduction	77
2.5.4	Numerosity Reduction	80
2.6	Data Discretization and Concept Hierarchy Generation	86
2.6.1	Discretization and Concept Hierarchy Generation for Numerical Data	88
2.6.2	Concept Hierarchy Generation for Categorical Data	94
2.7	Summary	97
	Exercises	97
	Bibliographic Notes	101
Chapter 3	Data Warehouse and OLAP Technology: An Overview	105
3.1	What Is a Data Warehouse?	105
3.1.1	Differences between Operational Database Systems and Data Warehouses	108
3.1.2	But, Why Have a Separate Data Warehouse?	109
3.2	A Multidimensional Data Model	110
3.2.1	From Tables and Spreadsheets to Data Cubes	110
3.2.2	Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases	114
3.2.3	Examples for Defining Star, Snowflake, and Fact Constellation Schemas	117

	3.2.4	Measures: Their Categorization and Computation	119
	3.2.5	Concept Hierarchies	121
	3.2.6	OLAP Operations in the Multidimensional Data Model	123
	3.2.7	A Starlet Query Model for Querying Multidimensional Databases	126
	3.3	Data Warehouse Architecture	127
	3.3.1	Steps for the Design and Construction of Data Warehouses	128
	3.3.2	A Three-Tier Data Warehouse Architecture	130
	3.3.3	Data Warehouse Back-End Tools and Utilities	134
	3.3.4	Metadata Repository	134
	3.3.5	Types of OLAP Servers: ROLAP versus MOLAP versus HOLAP	135
	3.4	Data Warehouse Implementation	137
	3.4.1	Efficient Computation of Data Cubes	137
	3.4.2	Indexing OLAP Data	141
	3.4.3	Efficient Processing of OLAP Queries	144
	3.5	From Data Warehousing to Data Mining	146
	3.5.1	Data Warehouse Usage	146
	3.5.2	From On-Line Analytical Processing to On-Line Analytical Mining	148
	3.6	Summary	150
		Exercises	152
		Bibliographic Notes	154
Chapter 4		Data Cube Computation and Data Generalization	157
	4.1	Efficient Methods for Data Cube Computation	157
	4.1.1	A Road Map for the Materialization of Different Kinds of Cubes	158
	4.1.2	Multiway Array Aggregation for Full Cube Computation	164
	4.1.3	BUC: Computing Iceberg Cubes from the Apex Cuboid Downward	168
	4.1.4	Star-cubing: Computing Iceberg Cubes Using a Dynamic Star-tree Structure	173
	4.1.5	Precomputing Shell Fragments for Fast High-Dimensional OLAP	178
	4.1.6	Computing Cubes with Complex Iceberg Conditions	187
	4.2	Further Development of Data Cube and OLAP Technology	189
	4.2.1	Discovery-Driven Exploration of Data Cubes	189
	4.2.2	Complex Aggregation at Multiple Granularity: Multifeature Cubes	192
	4.2.3	Constrained Gradient Analysis in Data Cubes	195

4.3	Attribute-Oriented Induction—An Alternative Method for Data Generalization and Concept Description	198
4.3.1	Attribute-Oriented Induction for Data Characterization	199
4.3.2	Efficient Implementation of Attribute-Oriented Induction	205
4.3.3	Presentation of the Derived Generalization	206
4.3.4	Mining Class Comparisons: Discriminating between Different Classes	210
4.3.5	Class Description: Presentation of Both Characterization and Comparison	215
4.4	Summary	218
	Exercises	219
	Bibliographic Notes	223
Chapter 5	Mining Frequent Patterns, Associations, and Correlations	227
5.1	Basic Concepts and a Road Map	227
5.1.1	Market Basket Analysis: A Motivating Example	228
5.1.2	Frequent Itemsets, Closed Itemsets, and Association Rules	230
5.1.3	Frequent Pattern Mining: A Road Map	232
5.2	Efficient and Scalable Frequent Itemset Mining Methods	234
5.2.1	The Apriori Algorithm: Finding Frequent Itemsets Using Candidate Generation	234
5.2.2	Generating Association Rules from Frequent Itemsets	239
5.2.3	Improving the Efficiency of Apriori	240
5.2.4	Mining Frequent Itemsets without Candidate Generation	242
5.2.5	Mining Frequent Itemsets Using Vertical Data Format	245
5.2.6	Mining Closed Frequent Itemsets	248
5.3	Mining Various Kinds of Association Rules	250
5.3.1	Mining Multilevel Association Rules	250
5.3.2	Mining Multidimensional Association Rules from Relational Databases and Data Warehouses	254
5.4	From Association Mining to Correlation Analysis	259
5.4.1	Strong Rules Are Not Necessarily Interesting: An Example	260
5.4.2	From Association Analysis to Correlation Analysis	261
5.5	Constraint-Based Association Mining	265
5.5.1	Metarule-Guided Mining of Association Rules	266
5.5.2	Constraint Pushing: Mining Guided by Rule Constraints	267
5.6	Summary	272
	Exercises	274
	Bibliographic Notes	280

Chapter 6	Classification and Prediction	285
6.1	What Is Classification? What Is Prediction?	285
6.2	Issues Regarding Classification and Prediction	289
6.2.1	Preparing the Data for Classification and Prediction	289
6.2.2	Comparing Classification and Prediction Methods	290
6.3	Classification by Decision Tree Induction	291
6.3.1	Decision Tree Induction	292
6.3.2	Attribute Selection Measures	296
6.3.3	Tree Pruning	304
6.3.4	Scalability and Decision Tree Induction	306
6.4	Bayesian Classification	310
6.4.1	Bayes' Theorem	310
6.4.2	Naïve Bayesian Classification	311
6.4.3	Bayesian Belief Networks	315
6.4.4	Training Bayesian Belief Networks	317
6.5	Rule-Based Classification	318
6.5.1	Using IF-THEN Rules for Classification	319
6.5.2	Rule Extraction from a Decision Tree	321
6.5.3	Rule Induction Using a Sequential Covering Algorithm	322
6.6	Classification by Backpropagation	327
6.6.1	A Multilayer Feed-Forward Neural Network	328
6.6.2	Defining a Network Topology	329
6.6.3	Backpropagation	329
6.6.4	Inside the Black Box: Backpropagation and Interpretability	334
6.7	Support Vector Machines	337
6.7.1	The Case When the Data Are Linearly Separable	337
6.7.2	The Case When the Data Are Linearly Inseparable	342
6.8	Associative Classification: Classification by Association Rule Analysis	344
6.9	Lazy Learners (or Learning from Your Neighbors)	347
6.9.1	k -Nearest-Neighbor Classifiers	348
6.9.2	Case-Based Reasoning	350
6.10	Other Classification Methods	351
6.10.1	Genetic Algorithms	351
6.10.2	Rough Set Approach	351
6.10.3	Fuzzy Set Approaches	352
6.11	Prediction	354
6.11.1	Linear Regression	355
6.11.2	Nonlinear Regression	357
6.11.3	Other Regression-Based Methods	358

6.12	Accuracy and Error Measures	359
6.12.1	Classifier Accuracy Measures	360
6.12.2	Predictor Error Measures	362
6.13	Evaluating the Accuracy of a Classifier or Predictor	363
6.13.1	Holdout Method and Random Subsampling	364
6.13.2	Cross-validation	364
6.13.3	Bootstrap	365
6.14	Ensemble Methods—Increasing the Accuracy	366
6.14.1	Bagging	366
6.14.2	Boosting	367
6.15	Model Selection	370
6.15.1	Estimating Confidence Intervals	370
6.15.2	ROC Curves	372
6.16	Summary	373
	Exercises	375
	Bibliographic Notes	378
Chapter 7	Cluster Analysis	383
7.1	What Is Cluster Analysis?	383
7.2	Types of Data in Cluster Analysis	386
7.2.1	Interval-Scaled Variables	387
7.2.2	Binary Variables	389
7.2.3	Categorical, Ordinal, and Ratio-Scaled Variables	392
7.2.4	Variables of Mixed Types	395
7.2.5	Vector Objects	397
7.3	A Categorization of Major Clustering Methods	398
7.4	Partitioning Methods	401
7.4.1	Classical Partitioning Methods: k -Means and k -Medoids	402
7.4.2	Partitioning Methods in Large Databases: From k -Medoids to CLARANS	407
7.5	Hierarchical Methods	408
7.5.1	Agglomerative and Divisive Hierarchical Clustering	408
7.5.2	BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies	412
7.5.3	ROCK: A Hierarchical Clustering Algorithm for Categorical Attributes	414
7.5.4	Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling	416
7.6	Density-Based Methods	418
7.6.1	DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density	418

7.6.2	OPTICS: Ordering Points to Identify the Clustering Structure	420
7.6.3	DENCLUE: Clustering Based on Density Distribution Functions	422
7.7	Grid-Based Methods	424
7.7.1	STING: STatistical INformation Grid	425
7.7.2	WaveCluster: Clustering Using Wavelet Transformation	427
7.8	Model-Based Clustering Methods	429
7.8.1	Expectation-Maximization	429
7.8.2	Conceptual Clustering	431
7.8.3	Neural Network Approach	433
7.9	Clustering High-Dimensional Data	434
7.9.1	CLIQUE: A Dimension-Growth Subspace Clustering Method	436
7.9.2	PROCLUS: A Dimension-Reduction Subspace Clustering Method	439
7.9.3	Frequent Pattern-Based Clustering Methods	440
7.10	Constraint-Based Cluster Analysis	444
7.10.1	Clustering with Obstacle Objects	446
7.10.2	User-Constrained Cluster Analysis	448
7.10.3	Semi-Supervised Cluster Analysis	449
7.11	Outlier Analysis	451
7.11.1	Statistical Distribution-Based Outlier Detection	452
7.11.2	Distance-Based Outlier Detection	454
7.11.3	Density-Based Local Outlier Detection	455
7.11.4	Deviation-Based Outlier Detection	458
7.12	Summary	460
	Exercises	461
	Bibliographic Notes	464
Chapter 8	Mining Stream, Time-Series, and Sequence Data	467
8.1	Mining Data Streams	468
8.1.1	Methodologies for Stream Data Processing and Stream Data Systems	469
8.1.2	Stream OLAP and Stream Data Cubes	474
8.1.3	Frequent-Pattern Mining in Data Streams	479
8.1.4	Classification of Dynamic Data Streams	481
8.1.5	Clustering Evolving Data Streams	486
8.2	Mining Time-Series Data	489
8.2.1	Trend Analysis	490
8.2.2	Similarity Search in Time-Series Analysis	493

8.3	Mining Sequence Patterns in Transactional Databases	498
8.3.1	Sequential Pattern Mining: Concepts and Primitives	498
8.3.2	Scalable Methods for Mining Sequential Patterns	500
8.3.3	Constraint-Based Mining of Sequential Patterns	509
8.3.4	Periodicity Analysis for Time-Related Sequence Data	512
8.4	Mining Sequence Patterns in Biological Data	513
8.4.1	Alignment of Biological Sequences	514
8.4.2	Hidden Markov Model for Biological Sequence Analysis	518
8.5	Summary	527
	Exercises	528
	Bibliographic Notes	531
Chapter 9	Graph Mining, Social Network Analysis, and Multirelational Data Mining	535
9.1	Graph Mining	535
9.1.1	Methods for Mining Frequent Subgraphs	536
9.1.2	Mining Variant and Constrained Substructure Patterns	545
9.1.3	Applications: Graph Indexing, Similarity Search, Classification, and Clustering	551
9.2	Social Network Analysis	556
9.2.1	What Is a Social Network?	556
9.2.2	Characteristics of Social Networks	557
9.2.3	Link Mining: Tasks and Challenges	561
9.2.4	Mining on Social Networks	565
9.3	Multirelational Data Mining	571
9.3.1	What Is Multirelational Data Mining?	571
9.3.2	ILP Approach to Multirelational Classification	573
9.3.3	Tuple ID Propagation	575
9.3.4	Multirelational Classification Using Tuple ID Propagation	577
9.3.5	Multirelational Clustering with User Guidance	580
9.4	Summary	584
	Exercises	586
	Bibliographic Notes	587
Chapter 10	Mining Object, Spatial, Multimedia, Text, and Web Data	591
10.1	Multidimensional Analysis and Descriptive Mining of Complex Data Objects	591
10.1.1	Generalization of Structured Data	592
10.1.2	Aggregation and Approximation in Spatial and Multimedia Data Generalization	593

	10.1.3	Generalization of Object Identifiers and Class/Subclass Hierarchies	594
	10.1.4	Generalization of Class Composition Hierarchies	595
	10.1.5	Construction and Mining of Object Cubes	596
	10.1.6	Generalization-Based Mining of Plan Databases by Divide-and-Conquer	596
10.2		Spatial Data Mining	600
	10.2.1	Spatial Data Cube Construction and Spatial OLAP	601
	10.2.2	Mining Spatial Association and Co-location Patterns	605
	10.2.3	Spatial Clustering Methods	606
	10.2.4	Spatial Classification and Spatial Trend Analysis	606
	10.2.5	Mining Raster Databases	607
10.3		Multimedia Data Mining	607
	10.3.1	Similarity Search in Multimedia Data	608
	10.3.2	Multidimensional Analysis of Multimedia Data	609
	10.3.3	Classification and Prediction Analysis of Multimedia Data	611
	10.3.4	Mining Associations in Multimedia Data	612
	10.3.5	Audio and Video Data Mining	613
10.4		Text Mining	614
	10.4.1	Text Data Analysis and Information Retrieval	615
	10.4.2	Dimensionality Reduction for Text	621
	10.4.3	Text Mining Approaches	624
10.5		Mining the World Wide Web	628
	10.5.1	Mining the Web Page Layout Structure	630
	10.5.2	Mining the Web's Link Structures to Identify Authoritative Web Pages	631
	10.5.3	Mining Multimedia Data on the Web	637
	10.5.4	Automatic Classification of Web Documents	638
	10.5.5	Web Usage Mining	640
10.6		Summary	641
		Exercises	642
		Bibliographic Notes	645
Chapter 11		Applications and Trends in Data Mining	649
	11.1	Data Mining Applications	649
		11.1.1 Data Mining for Financial Data Analysis	649
		11.1.2 Data Mining for the Retail Industry	651
		11.1.3 Data Mining for the Telecommunication Industry	652
		11.1.4 Data Mining for Biological Data Analysis	654
		11.1.5 Data Mining in Other Scientific Applications	657
		11.1.6 Data Mining for Intrusion Detection	658

11.2	Data Mining System Products and Research Prototypes	660
11.2.1	How to Choose a Data Mining System	660
11.2.2	Examples of Commercial Data Mining Systems	663
11.3	Additional Themes on Data Mining	665
11.3.1	Theoretical Foundations of Data Mining	665
11.3.2	Statistical Data Mining	666
11.3.3	Visual and Audio Data Mining	667
11.3.4	Data Mining and Collaborative Filtering	670
11.4	Social Impacts of Data Mining	675
11.4.1	Ubiquitous and Invisible Data Mining	675
11.4.2	Data Mining, Privacy, and Data Security	678
11.5	Trends in Data Mining	681
11.6	Summary	684
	Exercises	685
	Bibliographic Notes	687
Appendix	An Introduction to Microsoft's OLE DB for Data Mining	691
	A.1 Model Creation	693
	A.2 Model Training	695
	A.3 Model Prediction and Browsing	697
	Bibliography	703

About the Authors

Jiawei Han is Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Well known for his research in the areas of data mining and database systems, he has received many recognitions and awards for his contributions in the field, including the ACM Fellow and the 2004 ACM SIGKDD Innovations Award. He serves as Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data, and on the editorial boards for several scientific journals in the field. Micheline Kamber is a researcher who enjoys writing in easy-to-understand terms. She has a master's degree in computer science (specializing in artificial intelligence) from Concordia University, Canada.

Foreword

Jim Gray

Microsoft Research

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 a 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 such as statistics, visualization, artificial intelligence, and machine learning are contributing to this field. The breadth of the field makes it difficult to grasp its extraordinary progress over the last few years.

Jiawei Han and Micheline Kamber have done a wonderful job of organizing and presenting data mining in this very readable textbook. They begin by giving quick introductions to database and data mining concepts with particular emphasis on data analysis. They review the current product offerings by presenting a general framework that covers them all. They then cover, 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 pragmatic rules of thumb about when to apply each technique. I found this presentation style to be very readable, and I certainly learned a lot from reading the book. 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 I expect you will too.

Preface

Our capabilities of both generating and collecting data have been increasing rapidly. Contributing factors include the computerization of business, scientific, and government transactions; the widespread use of digital cameras, publication tools, and bar codes for most commercial products; and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information. 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 book explores the concepts and techniques of *data mining*, a promising and flourishing frontier in data and information systems and their 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 catchable in large databases, data warehouses, the Web, other massive information repositories, or data streams.

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

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 edition of this book, great progress has been made in the field of data mining. Many new data mining methods, systems, and applications have been developed. This new edition substantially revises the first edition of the book, with numerous enhancements and a reorganization of the technical contents of the entire book. In addition, several new chapters are included to address recent developments on mining complex types of data, including stream data, sequence data, graph structured data, social network data, and multirelational data.

The chapters 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 database technology, which has led to the need for data mining, and the importance of its applications. It examines the types of data to be mined, including relational, transactional, and data warehouse data, as well as complex types of data such as data streams, time-series, sequences, graphs, social networks, multirelational data, spatiotemporal data, multimedia data, text data, and Web data. The chapter presents a general classification of data mining tasks, based on the different kinds of knowledge to be mined. In comparison with the first edition, two new sections are introduced: Section 1.7 is on data mining primitives, which allow users to interactively communicate with data mining systems in order to direct the mining process, and Section 1.8 discusses the issues regarding how to integrate a data mining system with a database or data warehouse system. These two sections represent the condensed materials of Chapter 4, “*Data Mining Primitives, Languages and Architectures*,” in the first edition. Finally, major challenges in the field are discussed.

Chapter 2 introduces techniques for preprocessing the data before mining. This corresponds to Chapter 3 of the first edition. Because data preprocessing precedes the construction of data warehouses, we address this topic here, and then follow with an introduction to data warehouses in the subsequent chapter. This chapter describes various statistical methods for descriptive data summarization, including measuring both central tendency and dispersion of data. The description of data cleaning methods has been enhanced. Methods for data integration and transformation and data reduction are discussed, including the use of concept hierarchies for dynamic and static discretization. The automatic generation of concept hierarchies is also described.

Chapters 3 and 4 provide a solid introduction to data warehouse, OLAP (On-Line Analytical Processing), and data generalization. These two chapters correspond to Chapters 2 and 5 of the first edition, but with substantial enhancement regarding data

warehouse implementation methods. **Chapter 3** introduces the basic concepts, architectures and general implementations of data warehouse and on-line analytical processing, as well as the relationship between data warehousing and data mining. **Chapter 4** takes a more in-depth look at data warehouse and OLAP technology, presenting a detailed study of methods of data cube computation, including the recently developed star-cubing and high-dimensional OLAP methods. Further explorations of data warehouse and OLAP are discussed, such as discovery-driven cube exploration, multifeature cubes for complex data mining queries, and cube gradient analysis. Attribute-oriented induction, an alternative method for data generalization and concept description, is also discussed.

Chapter 5 presents methods for mining frequent patterns, associations, and correlations in transactional and relational databases and data warehouses. In addition to introducing the basic concepts, such as market basket analysis, many techniques for frequent itemset mining are presented in an organized way. These range from the basic Apriori algorithm and its variations to more advanced methods that improve on efficiency, including the frequent-pattern growth approach, frequent-pattern mining with vertical data format, and mining closed frequent itemsets. The chapter also presents techniques for mining multilevel association rules, multidimensional association rules, and quantitative association rules. In comparison with the previous edition, this chapter has placed greater emphasis on the generation of meaningful association and correlation rules. Strategies for constraint-based mining and the use of interestingness measures to focus the rule search are also described.

Chapter 6 describes methods for data classification and prediction, including decision tree induction, Bayesian classification, rule-based classification, the neural network technique of backpropagation, support vector machines, associative classification, k -nearest neighbor classifiers, case-based reasoning, genetic algorithms, rough set theory, and fuzzy set approaches. Methods of regression are introduced. Issues regarding accuracy and how to choose the best classifier or predictor are discussed. In comparison with the corresponding chapter in the first edition, the sections on rule-based classification and support vector machines are new, and the discussion of measuring and enhancing classification and prediction accuracy has been greatly expanded.

Cluster analysis forms the topic of **Chapter 7**. Several major data clustering approaches are presented, including partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. New sections in this edition introduce techniques for clustering high-dimensional data, as well as for constraint-based cluster analysis. Outlier analysis is also discussed.

Chapters 8 to 10 treat advanced topics in data mining and cover a large body of materials on recent progress in this frontier. These three chapters now replace our previous single chapter on advanced topics. **Chapter 8** focuses on the mining of stream data, time-series data, and sequence data (covering both transactional sequences and biological sequences). The basic data mining techniques (such as frequent-pattern mining, classification, clustering, and constraint-based mining) are extended for these types of data. **Chapter 9** discusses methods for graph and structural pattern mining, social network analysis and multirelational data mining. **Chapter 10** presents methods for

mining object, spatial, multimedia, text, and Web data, which cover a great deal of new progress in these areas.

Finally, in **Chapter 11**, we summarize the concepts presented in this book and discuss applications and trends in data mining. New material has been added on data mining for biological and biomedical data analysis, other scientific applications, intrusion detection, and collaborative filtering. Social impacts of data mining, such as privacy and data security issues, are discussed, in addition to challenging research issues. Further discussion of ubiquitous data mining has also been added.

The **Appendix** provides an introduction to Microsoft's OLE DB for Data Mining (OLEDB for DM).

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 and system names.

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 from the spectrum of data mining, especially regarding several recent research topics on data stream mining, graph mining, social network analysis, and multirelational data mining. The chapters preceding the advanced topics are written to be as self-contained as possible, so they may be read in order of interest by the reader. All of the major methods of data mining are presented. Because we take a database point of view to data mining, the book also presents many important topics in data mining, such as scalable algorithms and multidimensional OLAP analysis, that are often overlooked or minimally treated in other books.

To the Instructor

This book is designed to give a broad, yet detailed overview of the field of data mining. It can be used to teach an *introductory* course on data mining at an advanced undergraduate level or at the first-year graduate level. In addition, it can also be used to teach an *advanced* course on data mining.

If you plan to use the book to teach an introductory course, you may find that the materials in Chapters 1 to 7 are essential, among which Chapter 4 may be omitted if you do not plan to cover the implementation methods for data cubing and on-line analytical processing in depth. Alternatively, you may omit some sections in Chapters 1 to 7 and use Chapter 11 as the final coverage of applications and trends on data mining.

If you plan to use the book to teach an advanced course on data mining, you may use Chapters 8 through 11. Moreover, additional materials and some recent research papers may supplement selected themes from among the advanced topics of these chapters.

Individual chapters in this book can also be used for tutorials or for special topics in related courses, such as database systems, machine learning, pattern recognition, 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. Extensive teaching aids are available from the book's websites, such as lecture slides, reading lists, and course syllabi.

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 in order 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 in order to read this book?

- You should have some knowledge of the concepts and terminology associated with database systems, statistics, and machine learning. However, we do try to provide enough background of the basics in these fields, 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 pseudo-code 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 field of data mining. As a result, it is an excellent handbook on the subject. Because each chapter is designed to be as stand-alone as possible, you can focus on the topics that most interest you. The book can be used by applications 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 also 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. In Chapter 11, we briefly discuss data mining systems in commercial use, as well as promising research prototypes. Algorithms presented in the book are illustrated in pseudo-code. The pseudo-code 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 pseudo-code into the programming language of your choice to be a fairly straightforward task.

Book Websites with Resources

The book has a website at www.cs.uiuc.edu/~hanj/bk2 and another with Morgan Kaufmann Publishers at www.mkp.com/datamining2e. These websites contain many supplemental materials for readers of this book or anyone else with an interest in data mining. The resources include:

- **Slide presentations per chapter.** Lecture notes in Microsoft PowerPoint slides are available for each chapter.
- **Artwork of the book.** This may help you to make your own slides for your classroom teaching.
- **Instructors' manual.** This complete set of answers to the exercises in the book is available only to instructors from the publisher's website.
- **Course syllabi and lecture plan.** 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 will provide a set of links to the data mining data sets and some sites containing interesting data mining software packages.
- **Sample assignments, exams, course projects.** A set of sample assignments, exams, and course projects will be made available to instructors from the publisher's website.
- **Table of contents of the book in PDF.**
- **Errata on the different printings of the book.** We welcome you to point out any errors in the book. Once the error is confirmed, we will update this errata list, associated with the acknowledgment of your contribution.

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

Acknowledgments for the 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, Aymn-mohamed Rajan, Peter Scheuermann, Shashi Shekhar, Wei-Min Shen, Avi Silberschatz, Evangelos Simoudis, Nebojsa Stefanovic, Yin Jenny Tam, Simon Tang, Zhaohui 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.

Acknowledgments for the 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, 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 Executive Editor at Morgan Kaufmann Publishers, for her 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.