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 Starnet 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