

# Contents

**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 | <b>Summary</b>             | <b>39</b> |
|      | <b>Exercises</b>           | <b>40</b> |
|      | <b>Bibliographic Notes</b> | <b>42</b> |

## Chapter 2 **Data Preprocessing** 47

|       |  |            |
|-------|--|------------|
| 2.1   | <b>Why Preprocess the Data?</b>                                    | <b>48</b>  |
| 2.2   | <b>Descriptive Data Summarization</b>                              | <b>51</b>  |
| 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   | <b>Data Cleaning</b>   | <b>61</b>  |
| 2.3.1 | Missing Values   | 61         |
| 2.3.2 | Noisy Data   | 62         |
| 2.3.3 | Data Cleaning as a Process   | 65         |
| 2.4   | <b>Data Integration and Transformation</b>                         | <b>67</b>  |
| 2.4.1 | Data Integration   | 67         |
| 2.4.2 | Data Transformation  | 70         |
| 2.5   | <b>Data Reduction</b>  | <b>72</b>  |
| 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   | <b>Data Discretization and Concept Hierarchy Generation</b>        | <b>86</b>  |
| 2.6.1 | Discretization and Concept Hierarchy Generation for Numerical Data | 88         |
| 2.6.2 | Concept Hierarchy Generation for Categorical Data                  | 94         |
| 2.7   | <b>Summary</b>   | <b>97</b>  |
|       | <b>Exercises</b>   | <b>97</b>  |
|       | <b>Bibliographic Notes</b>   | <b>101</b> |

## Chapter 3 **Data Warehouse and OLAP Technology: An Overview** 105

|       |  |            |
|-------|--|------------|
| 3.1   | <b>What Is a Data Warehouse?</b>   | <b>105</b> |
| 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   | <b>A Multidimensional Data Model</b>   | <b>110</b> |
| 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<br>Multidimensional Databases            | 126        |
| 3.3       | <b>Data Warehouse Architecture</b>  | <b>127</b> |
| 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<br>versus HOLAP                   | 135        |
| 3.4       | <b>Data Warehouse Implementation</b>  | <b>137</b> |
| 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       | <b>From Data Warehousing to Data Mining</b>                                 | <b>146</b> |
| 3.5.1     | Data Warehouse Usage  | 146        |
| 3.5.2     | From On-Line Analytical Processing<br>to On-Line Analytical Mining          | 148        |
| 3.6       | <b>Summary</b>  | <b>150</b> |
|           | <b>Exercises</b>  | <b>152</b> |
|           | <b>Bibliographic Notes</b>  | <b>154</b> |
| Chapter 4 | <b>Data Cube Computation and Data Generalization</b>                        | <b>157</b> |
| 4.1       | <b>Efficient Methods for Data Cube Computation</b>                          | <b>157</b> |
| 4.1.1     | A Road Map for the Materialization of Different Kinds<br>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<br>Downward               | 168        |
| 4.1.4     | Star-cubing: Computing Iceberg Cubes Using<br>a Dynamic Star-tree Structure | 173        |
| 4.1.5     | Precomputing Shell Fragments for Fast High-Dimensional<br>OLAP              | 178        |
| 4.1.6     | Computing Cubes with Complex Iceberg Conditions                             | 187        |
| 4.2       | <b>Further Development of Data Cube and OLAP<br/>Technology</b>             | <b>189</b> |
| 4.2.1     | Discovery-Driven Exploration of Data Cubes                                  | 189        |
| 4.2.2     | Complex Aggregation at Multiple Granularity:<br>Multifeature Cubes          | 192        |
| 4.2.3     | Constrained Gradient Analysis in Data Cubes                                 | 195        |

|       |   |            |
|-------|---|------------|
| 4.3   | <b>Attribute-Oriented Induction—An Alternative Method for Data Generalization and Concept Description</b> | <b>198</b> |
| 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   | <b>Summary</b>  | <b>218</b> |
|       | <b>Exercises</b>  | <b>219</b> |
|       | <b>Bibliographic Notes</b>  | <b>223</b> |

## Chapter 5 **Mining Frequent Patterns, Associations, and Correlations** 227

|       |   |            |
|-------|---|------------|
| 5.1   | <b>Basic Concepts and a Road Map</b>  | <b>227</b> |
| 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   | <b>Efficient and Scalable Frequent Itemset Mining Methods</b>                           | <b>234</b> |
| 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   | <b>Mining Various Kinds of Association Rules</b>  | <b>250</b> |
| 5.3.1 | Mining Multilevel Association Rules   | 250        |
| 5.3.2 | Mining Multidimensional Association Rules from Relational Databases and Data Warehouses | 254        |
| 5.4   | <b>From Association Mining to Correlation Analysis</b>                                  | <b>259</b> |
| 5.4.1 | Strong Rules Are Not Necessarily Interesting: An Example                                | 260        |
| 5.4.2 | From Association Analysis to Correlation Analysis                                       | 261        |
| 5.5   | <b>Constraint-Based Association Mining</b>  | <b>265</b> |
| 5.5.1 | Metarule-Guided Mining of Association Rules   | 266        |
| 5.5.2 | Constraint Pushing: Mining Guided by Rule Constraints                                   | 267        |
| 5.6   | <b>Summary</b>  | <b>272</b> |
|       | <b>Exercises</b>  | <b>274</b> |
|       | <b>Bibliographic Notes</b>  | <b>280</b> |

**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   | <b>Accuracy and Error Measures</b>                          | <b>359</b> |
| 6.12.1 | Classifier Accuracy Measures                                | 360        |
| 6.12.2 | Predictor Error Measures                                    | 362        |
| 6.13   | <b>Evaluating the Accuracy of a Classifier or Predictor</b> | <b>363</b> |
| 6.13.1 | Holdout Method and Random Subsampling                       | 364        |
| 6.13.2 | Cross-validation  | 364        |
| 6.13.3 | Bootstrap   | 365        |
| 6.14   | <b>Ensemble Methods—Increasing the Accuracy</b>             | <b>366</b> |
| 6.14.1 | Bagging   | 366        |
| 6.14.2 | Boosting  | 367        |
| 6.15   | <b>Model Selection</b>                                      | <b>370</b> |
| 6.15.1 | Estimating Confidence Intervals                             | 370        |
| 6.15.2 | ROC Curves  | 372        |
| 6.16   | <b>Summary</b>  | <b>373</b> |
|        | <b>Exercises</b>  | <b>375</b> |
|        | <b>Bibliographic Notes</b>                                  | <b>378</b> |

## Chapter 7 **Cluster Analysis** 383

|       |   |            |
|-------|---|------------|
| 7.1   | <b>What Is Cluster Analysis?</b>  | <b>383</b> |
| 7.2   | <b>Types of Data in Cluster Analysis</b>  | <b>386</b> |
| 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   | <b>A Categorization of Major Clustering Methods</b>   | <b>398</b> |
| 7.4   | <b>Partitioning Methods</b>   | <b>401</b> |
| 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   | <b>Hierarchical Methods</b>   | <b>408</b> |
| 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   | <b>Density-Based Methods</b>  | <b>418</b> |
| 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    | <b>Grid-Based Methods</b>                                    | <b>424</b> |
| 7.7.1  | STING: STatistical INformation Grid                          | 425        |
| 7.7.2  | WaveCluster: Clustering Using Wavelet Transformation         | 427        |
| 7.8    | <b>Model-Based Clustering Methods</b>                        | <b>429</b> |
| 7.8.1  | Expectation-Maximization                                     | 429        |
| 7.8.2  | Conceptual Clustering  | 431        |
| 7.8.3  | Neural Network Approach                                      | 433        |
| 7.9    | <b>Clustering High-Dimensional Data</b>                      | <b>434</b> |
| 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   | <b>Constraint-Based Cluster Analysis</b>                     | <b>444</b> |
| 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   | <b>Outlier Analysis</b>                                      | <b>451</b> |
| 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   | <b>Summary</b>   | <b>460</b> |
|        | <b>Exercises</b>   | <b>461</b> |
|        | <b>Bibliographic Notes</b>                                   | <b>464</b> |

## Chapter 8 **Mining Stream, Time-Series, and Sequence Data** 467

|       |  |            |
|-------|--|------------|
| 8.1   | <b>Mining Data Streams</b>                                       | <b>468</b> |
| 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   | <b>Mining Time-Series Data</b>                                   | <b>489</b> |
| 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   | <b>Spatial Data Mining</b>   | <b>600</b> |
| 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   | <b>Multimedia Data Mining</b>  | <b>607</b> |
| 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   | <b>Text Mining</b>   | <b>614</b> |
| 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   | <b>Mining the World Wide Web</b>                                     | <b>628</b> |
| 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   | <b>Summary</b>   | <b>641</b> |
|        | <b>Exercises</b>   | <b>642</b> |
|        | <b>Bibliographic Notes</b>   | <b>645</b> |

## Chapter 11 Applications and Trends in Data Mining 649

|        |  |            |
|--------|--|------------|
| 11.1   | <b>Data Mining Applications</b>                | <b>649</b> |
| 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     | <b>Data Mining System Products and Research Prototypes</b>   | <b>660</b> |
| 11.2.1   | How to Choose a Data Mining System                           | 660        |
| 11.2.2   | Examples of Commercial Data Mining Systems                   | 663        |
| 11.3     | <b>Additional Themes on Data Mining</b>                      | <b>665</b> |
| 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     | <b>Social Impacts of Data Mining</b>                         | <b>675</b> |
| 11.4.1   | Ubiquitous and Invisible Data Mining                         | 675        |
| 11.4.2   | Data Mining, Privacy, and Data Security                      | 678        |
| 11.5     | <b>Trends in Data Mining</b>                                 | <b>681</b> |
| 11.6     | <b>Summary</b>   | <b>684</b> |
|          | <b>Exercises</b>   | <b>685</b> |
|          | <b>Bibliographic Notes</b>                                   | <b>687</b> |
| Appendix | <b>An Introduction to Microsoft's OLE DB for Data Mining</b> | <b>691</b> |
| A.1      | Model Creation   | 693        |
| A.2      | Model Training   | 695        |
| A.3      | Model Prediction and Browsing                                | 697        |
|          | <b>Bibliography</b>  | <b>703</b> |
|          | <b>Index</b>   | <b>745</b> |