Contents

Preface to the Second Edition 5

1 Introduction 21

1.1 What Is Data Mining? 24
1.2 Motivating Challenges 25
1.3 The Origins of Data Mining 27
1.4 Data Mining Tasks 29
1.5 Scope and Organization of the Book 33
1.6 Bibliographic Notes 35
1.7 Exercises 41

2 Data 43

2.1 Types of Data 46
2.1.1 Attributes and Measurement 47
2.1.2 Types of Data Sets 54
2.2 Data Quality 62
2.2.1 Measurement and Data Collection Issues 62
2.2.2 Issues Related to Applications 69
2.3 Data Preprocessing 70
2.3.1 Aggregation 71
2.3.2 Sampling 72
2.3.3 Dimensionality Reduction 76
2.3.4 Feature Subset Selection 78
2.3.5 Feature Creation 81
2.3.6 Discretization and Binarization 83
2.3.7 Variable Transformation 89
2.4 Measures of Similarity and Dissimilarity 91
2.4.1 Basics 92
2.4.2 Similarity and Dissimilarity between Simple Attributes . 94
2.4.3 Dissimilarities between Data Objects 96
2.4.4 Similarities between Data Objects 98

2.4.5 Examples of Proximity Measures . . . 99
2.4.6 Mutual Information . . . 108
2.4.7 Kernel Functions* . . . 110
2.4.8 Bregman Divergence* . . . 114
2.4.9 Issues in Proximity Calculation . . . 116
2.4.10 Selecting the Right Proximity Measure . . . 118
2.5 Bibliographic Notes . . . 120
2.6 Exercises . . . 125

3 Classification: Basic Concepts and Techniques 133

3.1 Basic Concepts . . . 134
3.2 General Framework for Classification . . . 137
3.3 Decision Tree Classifier . . . 139
3.3.1 A Basic Algorithm to Build a Decision Tree . . . 141
3.3.2 Methods for Expressing Attribute Test Conditions . . . 144
3.3.3 Measures for Selecting an Attribute Test Condition . . . 147
3.3.4 Algorithm for Decision Tree Induction . . . 156
3.3.5 Example Application: Web Robot Detection . . . 158
3.3.6 Characteristics of Decision Tree Classifiers . . . 160
3.4 Model Overfitting . . . 167
3.4.1 Reasons for Model Overfitting . . . 169
3.5 Model Selection . . . 176
3.5.1 Using a Validation Set . . . 176
3.5.2 Incorporating Model Complexity . . . 177
3.5.3 Estimating Statistical Bounds . . . 182
3.5.4 Model Selection for Decision Trees . . . 182
3.6 Model Evaluation . . . 184
3.6.1 Holdout Method . . . 185
3.6.2 Cross-Validation . . . 185
3.7 Presence of Hyper-parameters . . . 188
3.7.1 Hyper-parameter Selection . . . 188
3.7.2 Nested Cross-Validation . . . 190
3.8 Pitfalls of Model Selection and Evaluation . . . 192
3.8.1 Overlap between Training and Test Sets . . . 192
3.8.2 Use of Validation Error as Generalization Error . . . 192
3.9 Model Comparison* . . . 193
3.9.1 Estimating the Confidence Interval for Accuracy . . . 194
3.9.2 Comparing the Performance of Two Models . . . 195
3.10 Bibliographic Notes . . . 196
3.11 Exercises . . . 205

4 Association Analysis: Basic Concepts and Algorithms 213

4.1 Preliminaries 214
4.2 Frequent Itemset Generation 218
4.2.1 The *Apriori* Principle 219
4.2.2 Frequent Itemset Generation in the *Apriori* Algorithm . 220
4.2.3 Candidate Generation and Pruning 224
4.2.4 Support Counting 229
4.2.5 Computational Complexity 233
4.3 Rule Generation 236
4.3.1 Confidence-Based Pruning 236
4.3.2 Rule Generation in *Apriori* Algorithm 237
4.3.3 An Example: Congressional Voting Records 238
4.4 Compact Representation of Frequent Itemsets 240
4.4.1 Maximal Frequent Itemsets 240
4.4.2 Closed Itemsets 242
4.5 Alternative Methods for Generating Frequent Itemsets* 245
4.6 FP-Growth Algorithm* 249
4.6.1 FP-Tree Representation 250
4.6.2 Frequent Itemset Generation in FP-Growth Algorithm . 253
4.7 Evaluation of Association Patterns 257
4.7.1 Objective Measures of Interestingness 258
4.7.2 Measures beyond Pairs of Binary Variables 270
4.7.3 Simpson's Paradox 272
4.8 Effect of Skewed Support Distribution 274
4.9 Bibliographic Notes 280
4.10 Exercises 294

5 Cluster Analysis: Basic Concepts and Algorithms 307

5.1 Overview 310
5.1.1 What Is Cluster Analysis? 310
5.1.2 Different Types of Clusterings 311
5.1.3 Different Types of Clusters 313
5.2 K-means 316
5.2.1 The Basic K-means Algorithm 317
5.2.2 K-means: Additional Issues 326
5.2.3 Bisecting K-means 329
5.2.4 K-means and Different Types of Clusters 330
5.2.5 Strengths and Weaknesses 331
5.2.6 K-means as an Optimization Problem 331

5.3 Agglomerative Hierarchical Clustering 336
5.3.1 Basic Agglomerative Hierarchical Clustering Algorithm 337
5.3.2 Specific Techniques 339
5.3.3 The Lance-Williams Formula for Cluster Proximity . . . 344
5.3.4 Key Issues in Hierarchical Clustering 345
5.3.5 Outliers 346
5.3.6 Strengths and Weaknesses 347
5.4 DBSCAN 347
5.4.1 Traditional Density: Center-Based Approach 347
5.4.2 The DBSCAN Algorithm 349
5.4.3 Strengths and Weaknesses 351
5.5 Cluster Evaluation 353
5.5.1 Overview 353
5.5.2 Unsupervised Cluster Evaluation Using Cohesion and Separation 356
5.5.3 Unsupervised Cluster Evaluation Using the Proximity Matrix 364
5.5.4 Unsupervised Evaluation of Hierarchical Clustering . . . 367
5.5.5 Determining the Correct Number of Clusters 369
5.5.6 Clustering Tendency 370
5.5.7 Supervised Measures of Cluster Validity 371
5.5.8 Assessing the Significance of Cluster Validity Measures . 376
5.5.9 Choosing a Cluster Validity Measure 378
5.6 Bibliographic Notes 379
5.7 Exercises 385

6 Classification: Alternative Techniques 395
6.1 Types of Classifiers 395
6.2 Rule-Based Classifier 397
6.2.1 How a Rule-Based Classifier Works 399
6.2.2 Properties of a Rule Set 400
6.2.3 Direct Methods for Rule Extraction 401
6.2.4 Indirect Methods for Rule Extraction 406
6.2.5 Characteristics of Rule-Based Classifiers 408
6.3 Nearest Neighbor Classifiers 410
6.3.1 Algorithm 411
6.3.2 Characteristics of Nearest Neighbor Classifiers 412
6.4 Naïve Bayes Classifier 414
6.4.1 Basics of Probability Theory 415
6.4.2 Naïve Bayes Assumption 420

6.5 Bayesian Networks 429
6.5.1 Graphical Representation 429
6.5.2 Inference and Learning 435
6.5.3 Characteristics of Bayesian Networks 444
6.6 Logistic Regression 445
6.6.1 Logistic Regression as a Generalized Linear Model 446
6.6.2 Learning Model Parameters 447
6.6.3 Characteristics of Logistic Regression 450
6.7 Artificial Neural Network (ANN) 451
6.7.1 Perceptron 452
6.7.2 Multi-layer Neural Network 456
6.7.3 Characteristics of ANN 463
6.8 Deep Learning 464
6.8.1 Using Synergistic Loss Functions 465
6.8.2 Using Responsive Activation Functions 468
6.8.3 Regularization 470
6.8.4 Initialization of Model Parameters 473
6.8.5 Characteristics of Deep Learning 477
6.9 Support Vector Machine (SVM) 478
6.9.1 Margin of a Separating Hyperplane 478
6.9.2 Linear SVM 480
6.9.3 Soft-margin SVM 486
6.9.4 Nonlinear SVM 492
6.9.5 Characteristics of SVM 496
6.10 Ensemble Methods 498
6.10.1 Rationale for Ensemble Method 499
6.10.2 Methods for Constructing an Ensemble Classifier 499
6.10.3 Bias-Variance Decomposition 502
6.10.4 Bagging 504
6.10.5 Boosting 507
6.10.6 Random Forests 512
6.10.7 Empirical Comparison among Ensemble Methods 514
6.11 Class Imbalance Problem 515
6.11.1 Building Classifiers with Class Imbalance 516
6.11.2 Evaluating Performance with Class Imbalance 520
6.11.3 Finding an Optimal Score Threshold 524
6.11.4 Aggregate Evaluation of Performance 525
6.12 Multiclass Problem 532
6.13 Bibliographic Notes 535
6.14 Exercises 547

7 Association Analysis: Advanced Concepts 559
7.1 Handling Categorical Attributes . . . 559
7.2 Handling Continuous Attributes . . . 562
7.2.1 Discretization-Based Methods . . . 562
7.2.2 Statistics-Based Methods . . . 566
7.2.3 Non-discretization Methods . . . 568
7.3 Handling a Concept Hierarchy . . . 570
7.4 Sequential Patterns . . . 572
7.4.1 Preliminaries . . . 573
7.4.2 Sequential Pattern Discovery . . . 576
7.4.3 Timing Constraints* . . . 581
7.4.4 Alternative Counting Schemes* . . . 585
7.5 Subgraph Patterns . . . 587
7.5.1 Preliminaries . . . 588
7.5.2 Frequent Subgraph Mining . . . 591
7.5.3 Candidate Generation . . . 595
7.5.4 Candidate Pruning . . . 601
7.5.5 Support Counting . . . 601
7.6 Infrequent Patterns* . . . 601
7.6.1 Negative Patterns . . . 602
7.6.2 Negatively Correlated Patterns . . . 603
7.6.3 Comparisons among Infrequent Patterns, Negative Patterns, and Negatively Correlated Patterns . . . 604
7.6.4 Techniques for Mining Interesting Infrequent Patterns . 606
7.6.5 Techniques Based on Mining Negative Patterns . . . 607
7.6.6 Techniques Based on Support Expectation . . . 609
7.7 Bibliographic Notes . . . 613
7.8 Exercises . . . 618

8 Cluster Analysis: Additional Issues and Algorithms 633
8.1 Characteristics of Data, Clusters, and Clustering Algorithms . 634
8.1.1 Example: Comparing K-means and DBSCAN . . . 634
8.1.2 Data Characteristics . . . 635
8.1.3 Cluster Characteristics . . . 637
8.1.4 General Characteristics of Clustering Algorithms . . . 639
8.2 Prototype-Based Clustering . . . 641
8.2.1 Fuzzy Clustering . . . 641
8.2.2 Clustering Using Mixture Models . . . 647
8.2.3 Self-Organizing Maps (SOM) . . . 657
8.3 Density-Based Clustering . . . 664

8.3.1 Grid-Based Clustering 664
8.3.2 Subspace Clustering 668
8.3.3 DENCLUE: A Kernel-Based Scheme for Density-Based Clustering 672
8.4 Graph-Based Clustering 676
8.4.1 Sparsification 677
8.4.2 Minimum Spanning Tree (MST) Clustering 678
8.4.3 OPOSSUM: Optimal Partitioning of Sparse Similarities Using METIS 679
8.4.4 Chameleon: Hierarchical Clustering with Dynamic Modeling 680
8.4.5 Spectral Clustering 686
8.4.6 Shared Nearest Neighbor Similarity 693
8.4.7 The Jarvis-Patrick Clustering Algorithm 696
8.4.8 SNN Density 698
8.4.9 SNN Density-Based Clustering 699
8.5 Scalable Clustering Algorithms 701
8.5.1 Scalability: General Issues and Approaches 701
8.5.2 BIRCH 704
8.5.3 CURE 706
8.6 Which Clustering Algorithm? 710
8.7 Bibliographic Notes 713
8.8 Exercises 719

9 Anomaly Detection 723
9.1 Characteristics of Anomaly Detection Problems 725
9.1.1 A Definition of an Anomaly 725
9.1.2 Nature of Data 726
9.1.3 How Anomaly Detection is Used 727
9.2 Characteristics of Anomaly Detection Methods 728
9.3 Statistical Approaches 730
9.3.1 Using Parametric Models 730
9.3.2 Using Non-parametric Models 734
9.3.3 Modeling Normal and Anomalous Classes 735
9.3.4 Assessing Statistical Significance 737
9.3.5 Strengths and Weaknesses 738
9.4 Proximity-based Approaches 739
9.4.1 Distance-based Anomaly Score 739
9.4.2 Density-based Anomaly Score 740
9.4.3 Relative Density-based Anomaly Score 742
9.4.4 Strengths and Weaknesses 743

9.5 Clustering-based Approaches 744
9.5.1 Finding Anomalous Clusters 744
9.5.2 Finding Anomalous Instances 745
9.5.3 Strengths and Weaknesses 748
9.6 Reconstruction-based Approaches 748
9.6.1 Strengths and Weaknesses 751
9.7 One-class Classification 752
9.7.1 Use of Kernels 753
9.7.2 The Origin Trick 754
9.7.3 Strengths and Weaknesses 758
9.8 Information Theoretic Approaches 758
9.8.1 Strengths and Weaknesses 760
9.9 Evaluation of Anomaly Detection 760
9.10 Bibliographic Notes 762
9.11 Exercises 769

10 Avoiding False Discoveries 775
10.1 Preliminaries: Statistical Testing 776
10.1.1 Significance Testing 776
10.1.2 Hypothesis Testing 781
10.1.3 Multiple Hypothesis Testing 787
10.1.4 Pitfalls in Statistical Testing 796
10.2 Modeling Null and Alternative Distributions 798
10.2.1 Generating Synthetic Data Sets 801
10.2.2 Randomizing Class Labels 802
10.2.3 Resampling Instances 802
10.2.4 Modeling the Distribution of the Test Statistic 803
10.3 Statistical Testing for Classification 803
10.3.1 Evaluating Classification Performance 803
10.3.2 Binary Classification as Multiple Hypothesis Testing . . 805
10.3.3 Multiple Hypothesis Testing in Model Selection 806
10.4 Statistical Testing for Association Analysis 807
10.4.1 Using Statistical Models 808
10.4.2 Using Randomization Methods 814
10.5 Statistical Testing for Cluster Analysis 815
10.5.1 Generating a Null Distribution for Internal Indices . . . 816
10.5.2 Generating a Null Distribution for External Indices . . . 818
10.5.3 Enrichment 818
10.6 Statistical Testing for Anomaly Detection 820
10.7 Bibliographic Notes 823
10.8 Exercises 828

Author Index 836

Subject Index 849

Copyright Permissions 859