DATA MINING

Concepts, Models, Methods, and Algorithms

THIRD EDITION

Mehmed Kantardzic

IEEE PRESS

WILEY

CONTENTS

Preface xiii
Preface to the Second Edition xv
Preface to the First Edition xvii

1 Data-Mining Concepts 1

1.1 Introduction 2
1.2 Data-Mining Roots 4
1.3 Data-Mining Process 6
1.4 From Data Collection to Data Preprocessing 10
1.5 Data Warehouses for Data Mining 15
1.6 From Big Data to Data Science 18
1.7 Business Aspects of Data Mining: Why a Data-Mining Project Fails? 22
1.8 Organization of This Book 26
1.9 Review Questions and Problems 28
1.10 References for Further Study 30

2 Preparing the Data 33

2.1 Representation of Raw Data 34
2.2 Characteristics of Raw Data 38
2.3 Transformation of Raw Data 40
2.4 Missing Data 43
2.5 Time-Dependent Data 44
2.6 Outlier Analysis 49
2.7 Review Questions and Problems 56
2.8 References for Further Study 59

3 Data Reduction 61
3.1 Dimensions of Large Data Sets 62
3.2 Features Reduction 64
3.3 Relief Algorithm 75
3.4 Entropy Measure for Ranking Features 77
3.5 Principal Component Analysis 80
3.6 Value Reduction 83
3.7 Feature Discretization: ChiMerge Technique 86
3.8 Case Reduction 90
3.9 Review Questions and Problems 93
3.10 References for Further Study 95

4 Learning from Data 97
4.1 Learning Machine 99
4.2 Statistical Learning Theory 104
4.3 Types of Learning Methods 110
4.4 Common Learning Tasks 112
4.5 Support Vector Machines 117
4.6 Semi-Supervised Support Vector Machines (S3VM) 131
4.7 kNN: Nearest Neighbor Classifier 134
4.8 Model Selection vs. Generalization 138
4.9 Model Estimation 142
4.10 Imbalanced Data Classification 150
4.11 90% Accuracy … Now What? 154
4.12 Review Questions and Problems 158
4.13 References for Further Study 161

5 Statistical Methods 165
5.1 Statistical Inference 166
5.2 Assessing Differences in Data Sets 168
5.3 Bayesian Inference 172
5.4 Predictive Regression 175
5.5 Analysis of Variance 181
5.6 Logistic Regression 184
5.7 Log-Linear Models 185
5.8 Linear Discriminant Analysis 189
5.9 Review Questions and Problems 191
5.10 References for Further Study 194

6 Decision Trees and Decision Rules 197

6.1 Decision Trees 199
6.2 *C4.5 Algorithm*: Generating a Decision Tree 201
6.3 Unknown Attribute Values 209
6.4 Pruning Decision Trees 214
6.5 *C4.5 Algorithm*: Generating Decision Rules 215
6.6 Cart Algorithm and Gini Index 219
6.7 Limitations of Decision Trees and Decision Rules 222
6.8 Review Questions and Problems 225
6.9 References for Further Study 229

7 Artificial Neural Networks 231

7.1 Model of an Artificial Neuron 233
7.2 Architectures of Artificial Neural Networks 237
7.3 Learning Process 239
7.4 Learning Tasks Using Anns 243
7.5 Multilayer Perceptrons 245
7.6 Competitive Networks and Competitive Learning 255
7.7 Self-Organizing Maps 259
7.8 Deep Learning 264
7.9 Convolutional Neural Networks (CNNs) 270
7.10 Review Questions and Problems 273
7.11 References for Further Study 276

8 Ensemble Learning 279

8.1 Ensemble Learning Methodologies 280
8.2 Combination Schemes for Multiple Learners 285
8.3 Bagging and Boosting 286
8.4 AdaBoost 288
8.5 Review Questions and Problems 290
8.6 References for Further Study 293

9 Cluster Analysis 295

9.1 Clustering Concepts 296
9.2 Similarity Measures 299
9.3 Agglomerative Hierarchical Clustering 306
9.4 Partitional Clustering 310
9.5 Incremental Clustering 313
9.6 DBSCAN Algorithm 317

9.7 BIRCH Algorithm 320
9.8 Clustering Validation 323
9.9 Review Questions and Problems 328
9.10 References for Further Study 333

10 Association Rules 335
10.1 Market-Basket Analysis 337
10.2 Algorithm *Apriori* 338
10.3 From Frequent Itemsets to Association Rules 340
10.4 Improving the Efficiency of the *Apriori* Algorithm 342
10.5 Frequent Pattern Growth Method 344
10.6 Associative-Classification Method 346
10.7 Multidimensional Association Rule Mining 349
10.8 Review Questions and Problems 351
10.9 References for Further Study 355

11 Web Mining and Text Mining 357
11.1 Web Mining 358
11.2 Web Content, Structure, and Usage Mining 360
11.3 Hits and Logsom Algorithms 362
11.4 Mining Path-Traversal Patterns 368
11.5 PageRank Algorithm 371
11.6 Recommender Systems 374
11.7 Text Mining 375
11.8 Latent Semantic Analysis 379
11.9 Review Questions and Problems 385
11.10 References for Further Study 388

12 Advances in Data Mining 391
12.1 Graph Mining 392
12.2 Temporal Data Mining 406
12.3 Spatial Data Mining 422
12.4 Distributed Data Mining 426
12.5 Correlation Does not Imply Causality! 435
12.6 Privacy, Security, and Legal Aspects of Data Mining 442
12.7 Cloud Computing Based on Hadoop and Map/Reduce 449
12.8 Reinforcement Learning 454
12.9 Review Questions and Problems 459
12.10 References for Further Study 461

13 Genetic Algorithms 465

13.1 Fundamentals of Genetic Algorithms 466
13.2 Optimization Using Genetic Algorithms 468
13.3 A Simple Illustration of a Genetic Algorithm 474
13.4 Schemata 480
13.5 Traveling Salesman Problem 483
13.6 Machine Learning Using Genetic Algorithms 485
13.7 Genetic Algorithms for Clustering 490
13.8 Review Questions and Problems 493
13.9 References for Further Study 494

14 Fuzzy Sets and Fuzzy Logic 497

14.1 Fuzzy Sets 498
14.2 Fuzzy Set Operations 504
14.3 Extension Principle and Fuzzy Relations 509
14.4 Fuzzy Logic and Fuzzy Inference Systems 513
14.5 Multifactorial Evaluation 518
14.6 Extracting Fuzzy Models from Data 521
14.7 Data Mining and Fuzzy Sets 526
14.8 Review Questions and Problems 528
14.9 References for Further Study 530

15 Visualization Methods 533

15.1 Perception and Visualization 534
15.2 Scientific Visualization and Information Visualization 535
15.3 Parallel Coordinates 542
15.4 Radial Visualization 544
15.5 Visualization Using Self-Organizing Maps 547
15.6 Visualization Systems for Data Mining 549
15.7 Review Questions and Problems 554
15.8 References for Further Study 555

Appendix A: Information on Data Mining 559

A.1 Data-Mining Journals 559
A.2 Data-Mining Conferences 564
A.3 Data-Mining Forums/Blogs 568
A.4 Data Sets 570

A.5 Comercially and Publicly Available Tools 574
A.6 Web Site Links 583

Appendix B: Data-Mining Applications 589
B.1 Data Mining for Financial Data Analyses 589
B.2 Data Mining for the Telecomunication Industry 593
B.3 Data Mining for the Retail Industry 596
B.4 Data Mining in Healthcare and Biomedical Research 599
B.5 Data Mining in Science and Engineering 602
B.6 Pitfalls of Data Mining 605

Bibliography 607
Index 633