1. Record Nr. UNINA9910830267803321

Autore	Kantardzic Mehmed
Titolo	Data mining : concepts, models, methods, and algorithms / / Mehmed Kantardzic
Pubbl/distr/stampa	Hoboken, New Jersey : , : John Wiley, , c2011 [Piscataqay, New Jersey] : , : IEEE Xplore, , [2011]
ISBN	1-283-23974-4 9786613239747 1-118-02913-5 1-118-02912-7 1-118-02914-3
Edizione	[2nd ed.]
Descrizione fisica	1 online resource (554 p.)
Disciplina	005.741 006.3/12
Soggetti	Data mining
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Description based upon print version of record.
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	Preface to the Second Edition xiii -- Preface to the First Edition xv -- 1 DATA-MINING CONCEPTS 1 -- 1.1 Introduction 1 -- 1.2 Data-Mining Roots 4 -- 1.3 Data-Mining Process 6 -- 1.4 Large Data Sets 9 -- 1.5 Data Warehouses for Data Mining 14 -- 1.6 Business Aspects of Data Mining: Why a Data-Mining Project Fails 17 -- 1.7 Organization of This Book 21 -- 1.8 Review Questions and Problems 23 -- 1.9 References for Further Study 24 -- 2 PREPARING THE DATA 26 -- 2.1 Representation of Raw Data 26 -- 2.2 Characteristics of Raw Data 31 -- 2.3 Transformation of Raw Data 33 -- 2.4 Missing Data 36 -- 2.5 Time-Dependent Data 37 -- 2.6 Outlier Analysis 41 -- 2.7 Review Questions and Problems 48 -- 2.8 References for Further Study 51 -- 3 DATA REDUCTION 53 -- 3.1 Dimensions of Large Data Sets 54 -- 3.2 Feature Reduction 56 -- 3.3 Relief Algorithm 66 -- 3.4 Entropy Measure for Ranking Features 68 -- 3.5 PCA 70 -- 3.6 Value Reduction 73 -- 3.7 Feature Discretization: ChiMerge Technique 77 -- 3.8 Case Reduction 80 -- 3.9 Review Questions and Problems 83 -- 3.10 References for Further Study 85 -- 4 LEARNING FROM DATA 87 -- 4.1

Learning Machine 89 -- 4.2 SLT 93 -- 4.3 Types of Learning Methods 99 -- 4.4 Common Learning Tasks 101 -- 4.5 SVMs 105 -- 4.6 kNN: Nearest Neighbor Classifi er 118 -- 4.7 Model Selection versus Generalization 122 -- 4.8 Model Estimation 126 -- 4.9 90% Accuracy: Now What? 132 -- 4.10 Review Questions and Problems 136 -- 4.11 References for Further Study 138 -- 5 STATISTICAL METHODS 140 -- 5.1 Statistical Inference 141 -- 5.2 Assessing Differences in Data Sets 143 -- 5.3 Bayesian Inference 146 -- 5.4 Predictive Regression 149 -- 5.5 ANOVA 155 -- 5.6 Logistic Regression 157 -- 5.7 Log-Linear Models 158 -- 5.8 LDA 162 -- 5.9 Review Questions and Problems 164 -- 5.10 References for Further Study 167 -- 6 DECISION TREES AND DECISION RULES 169 -- 6.1 Decision Trees 171 -- 6.2 C4.5 Algorithm: Generating a Decision Tree 173 -- 6.3 Unknown Attribute Values 180 -- 6.4 Pruning Decision Trees 184.
6.5 C4.5 Algorithm: Generating Decision Rules 185 -- 6.6 CART Algorithm & Gini Index 189 -- 6.7 Limitations of Decision Trees and Decision Rules 192 -- 6.8 Review Questions and Problems 194 -- 6.9 References for Further Study 198 -- 7 ARTIFICIAL NEURAL NETWORKS 199 -- 7.1 Model of an Artifi cial Neuron 201 -- 7.2 Architectures of ANNs 205 -- 7.3 Learning Process 207 -- 7.4 Learning Tasks Using ANNs 210 -- 7.5 Multilayer Perceptrons (MLPs) 213 -- 7.6 Competitive Networks and Competitive Learning 221 -- 7.7 SOMs 225 -- 7.8 Review Questions and Problems 231 -- 7.9 References for Further Study 233 -- 8 ENSEMBLE LEARNING 235 -- 8.1 Ensemble-Learning Methodologies 236 -- 8.2 Combination Schemes for Multiple Learners 240 -- 8.3 Bagging and Boosting 241 -- 8.4 AdaBoost 243 -- 8.5 Review Questions and Problems 245 -- 8.6 References for Further Study 247 -- 9 CLUSTER ANALYSIS 249 -- 9.1 Clustering Concepts 250 -- 9.2 Similarity Measures 253 -- 9.3 Agglomerative Hierarchical Clustering 259 -- 9.4 Partitional Clustering 263 -- 9.5 Incremental Clustering 266 -- 9.6 DBSCAN Algorithm 270 -- 9.7 BIRCH Algorithm 272 -- 9.8 Clustering Validation 275 -- 9.9 Review Questions and Problems 275 -- 9.10 References for Further Study 279 -- 10 ASSOCIATION RULES 280 -- 10.1 Market-Basket Analysis 281 -- 10.2 Algorithm Apriori 283 -- 10.3 From Frequent Itemsets to Association Rules 285 -- 10.4 Improving the Effi ciency of the Apriori Algorithm 286 -- 10.5 FP Growth Method 288 -- 10.6 Associative-Classifi cation Method 290 -- 10.7 Multidimensional Association-Rules Mining 293 -- 10.8 Review Questions and Problems 295 -- 10.9 References for Further Study 298 -- 11 WEB MINING AND TEXT MINING 300 -- 11.1 Web Mining 300 -- 11.2 Web Content, Structure, and Usage Mining 302 -- 11.3 HITS and LOGSOM Algorithms 305 -- 11.4 Mining Path-Traversal Patterns 310 -- 11.5 PageRank Algorithm 313 -- 11.6 Text Mining 316 -- 11.7 Latent Semantic Analysis (LSA) 320 -- 11.8 Review Questions and Problems 324 -- 11.9 References for Further Study 326.
12 ADVANCES IN DATA MINING 328 -- 12.1 Graph Mining 329 -- 12.2 Temporal Data Mining 343 -- 12.3 Spatial Data Mining (SDM) 357 -- 12.4 Distributed Data Mining (DDM) 360 -- 12.5 Correlation Does Not Imply Causality 369 -- 12.6 Privacy, Security, and Legal Aspects of Data Mining 376 -- 12.7 Review Questions and Problems 381 -- 12.8 References for Further Study 382 -- 13 GENETIC ALGORITHMS 385 -- 13.1 Fundamentals of GAs 386 -- 13.2 Optimization Using GAs 388 -- 13.3 A Simple Illustration of a GA 394 -- 13.4 Schemata 399 -- 13.5 TSP 402 -- 13.6 Machine Learning Using GAs 404 -- 13.7 GAs for Clustering 409 -- 13.8 Review Questions and Problems 411 -- 13.9 References for Further Study 413 -- 14 FUZZY SETS AND FUZZY LOGIC 414 -- 14.1 Fuzzy Sets 415 -- 14.2 Fuzzy-Set Operations 420 -- 14.3 Extension Principle and Fuzzy Relations 425 -- 14.4 Fuzzy Logic and

Fuzzy Inference Systems 429 -- 14.5 Multifactorial Evaluation 433 -- 14.6 Extracting Fuzzy Models from Data 436 -- 14.7 Data Mining and Fuzzy Sets 441 -- 14.8 Review Questions and Problems 443 -- 14.9 References for Further Study 445 -- 15 VISUALIZATION METHODS 447 -- 15.1 Perception and Visualization 448 -- 15.2 Scientifi c Visualization and -- Information Visualization 449 -- 15.3 Parallel Coordinates 455 -- 15.4 Radial Visualization 458 -- 15.5 Visualization Using Self-Organizing Maps (SOMs) 460 -- 15.6 Visualization Systems for Data Mining 462 -- 15.7 Review Questions and Problems 467 -- 15.8 References for Further Study 468 -- Appendix A 470 -- A.1 Data-Mining Journals 470 -- A.2 Data-Mining Conferences 473 -- A.3 Data-Mining Forums/Blogs 477 -- A.4 Data Sets 478 -- A.5 Comercially and Publicly Available Tools 480 -- A.6 Web Site Links 489 -- Appendix B: Data-Mining Applications 496 -- B.1 Data Mining for Financial Data Analysis 496 -- B.2 Data Mining for the Telecomunications Industry 499 -- B.3 Data Mining for the Retail Industry 501 -- B.4 Data Mining in Health Care and Biomedical Research 503 -- B.5 Data Mining in Science and Engineering 506.
B.6 Pitfalls of Data Mining 509 -- Bibliography 510 -- Index 529.

Sommario/riassunto

Now updated--the systematic introductory guide to modern analysis of large data setsAs data sets continue to grow in size and complexity, there has been an inevitable move towards indirect, automatic, and intelligent data analysis in which the analyst works via more complex and sophisticated software tools. This book reviews state-of-the-art methodologies and techniques for analyzing enormous quantities of raw data in high-dimensional data spaces to extract new information for decision-making.This Second Edition of Data Mining: Concepts, Models, Methods, and Algorithms discusses data mining principles and then describes representative state-of-the-art methods and algorithms originating from different disciplines such as statistics, machine learning, neural networks, fuzzy logic, and evolutionary computation. Detailed algorithms are provided with necessary explanations and illustrative examples, and questions and exercises for practice at the end of each chapter. This new edition features the following new techniques/methodologies:. Support Vector Machines (SVM)--developed based on statistical learning theory, they have a large potential for applications in predictive data mining. Kohonen Maps (Self-Organizing Maps - SOM)--one of very applicative neural-networks-based methodologies for descriptive data mining and multi-dimensional data visualizations. DBSCAN, BIRCH, and distributed DBSCAN clustering algorithms--representatives of an important class of density-based clustering methodologies. Bayesian Networks (BN) methodology often used for causality modeling. Algorithms for measuring Betweeness and Centrality parameters in graphs, important for applications in mining large social networks. CART algorithm and Gini index in building decision trees. Bagging & Boosting approaches to ensemble-learning methodologies, with details of AdaBoost algorithm. Relief algorithm, one of the core feature selection algorithms inspired by instance-based learning. PageRank algorithm for mining and authority ranking of web pages. Latent Semantic Analysis (LSA) for text mining and measuring semantic similarities between text-based documents. New sections on temporal, spatial, web, text, parallel, and distributed data mining. More emphasis on business, privacy, security, and legal aspects of data mining technologyThis text offers guidance on how and when to use a particular software tool (with the companion data sets) from among the hundreds offered when faced with a data set to mine. This allows analysts to create and perform their own data mining experiments using their knowledge of the methodologies and

techniques provided. The book emphasizes the selection of appropriate methodologies and data analysis software, as well as parameter tuning. These critically important, qualitative decisions can only be made with the deeper understanding of parameter meaning and its role in the technique that is offered here.This volume is primarily intended as a data-mining textbook for computer science, computer engineering, and computer information systems majors at the graduate level. Senior students at the undergraduate level and with the appropriate background can also successfully comprehend all topics presented here.