Contents

Foreword **xviii**
Gavriel Salvendy

Preface **xix**
Nong Ye

About the Editor **xxiii**

Advisory Board **xxv**

Contributors **xxvii**

I: METHODOLOGIES OF DATA MINING

1 Decision Trees **3**
Johannes Gehrke
Introduction **3**
Problem Definition **4**
Classification Tree Construction **7**
Split Selection **7**
Data Access **8**
Tree Pruning **15**
Missing Values **17**
A Short Introduction to Regression Trees **20**
Problem Definition **20**
Split Selection **20**
Data Access **21**
Applications and Available Software **22**
Cataloging Sky Objects **22**
Decision Trees in Today's Data Mining Tools **22**
Summary **22**
References **23**

2 Association Rules **25**
Geoffrey I. Webb
Introduction **26**
Market Basket Analysis **26**
Association Rule Discovery **27**
The Apriori Algorithm **28**
The Power of the Frequent Item Set Strategy **29**
Measures of Interestingness **31**

Lift 31
Leverage 32
Item Set Discovery 32
Techniques for Frequent Item Set Discovery 33
Closed Item Set Strategies 33
Long Item Sets 35
Sampling 35
Techniques for Discovering Association Rules without Item Set Discovery 35
Associations with Numeric Values 36
Applications of Association Rule Discovery 36
Summary 37
References 38

3 Artificial Neural Network Models for Data Mining 41
Jennie Si, Benjamin J. Nelson, and George C. Runger
Introduction to Multilayer Feedforward Networks 42
Gradient Based Training Methods for MFN 43
The Partial Derivatives 44
Nonlinear Least Squares Methods 45
Batch versus Incremental Learning 47
Comparison of MFN and Other Classification Methods 47
Decision Tree Methods 47
Discriminant Analysis Methods 48
Multiple Partition Decision Tree 49
A Growing MFN 50
Case Study 1—Classifying Surface Texture 52
Experimental Conditions 52
Quantitative Comparison Results of Classification Methods 53
Closing Discussions on Case 1 55
Introduction to SOM 55
The SOM Algorithm 56
SOM Building Blocks 57
Implementation of the SOM Algorithm 58
Case Study 2—Decoding Monkey's Movement Directions from Its Cortical Activities 59
Trajectory Computation from Motor Cortical Discharge Rates 60
Using Data from Spiral Tasks to Train the SOM 62
Using Data from Spiral and Center→Out Tasks to Train the SOM 62
Average Testing Result Using the Leave-K-Out Method 63
Closing Discussions on Case 2 64
Final Conclusions and Discussions 65
References 65

4 Statistical Analysis of Normal and Abnormal Data 67
Connie M. Borror
Introduction 67
Univariate Control Charts 68
Variables Control Charts 68
Attributes Control Charts 81

Cumulative Sum Control Charts **89**
Exponentially Weighted Moving Average Control Charts **93**
Choice of Control Charting Techniques **95**
Average Run Length **96**
Multivariate Control Charts **98**
Data Description **98**
Hotelling T^2 Control Chart **98**
Multivariate EWMA Control Charts **101**
Summary **102**
References **102**

5 Bayesian Data Analysis **103**
David Madigan and Greg Ridgeway
Introduction **104**
Fundamentals of Bayesian Inference **104**
A Simple Example **104**
A More Complicated Example **106**
Hierarchical Models and Exchangeability **109**
Prior Distributions in Practice **111**
Bayesian Model Selection and Model Averaging **113**
Model Selection **113**
Model Averaging **114**
Model Assessment **114**
Bayesian Computation **115**
Importance Sampling **115**
Markov Chain Monte Carlo (MCMC) **116**
An Example **117**
Application to Massive Data **118**
Importance Sampling for Analysis of Massive Data Sets **118**
Variational Methods **120**
Bayesian Modeling **121**
BUGS and Models of Realistic Complexity via MCMC **121**
Bayesian Predictive Modeling **125**
Bayesian Descriptive Modeling **127**
Available Software **128**
Discussion and Future Directions **128**
Summary **128**
Acknowledgments **129**
References **129**

6 Hidden Markov Processes and Sequential Pattern Mining **133**
Steven L. Scott
Introduction to Hidden Markov Models **134**
Parameter Estimation in the Presence of Missing Data **136**
The EM Algorithm **136**
MCMC Data Augmentation **138**
Missing Data Summary **140**
Local Computation **140**
The Likelihood Recursion **140**

The Forward-Backward Recursions 141
The Viterbi Algorithm 142
Understanding the Recursions 143
A Numerical Example Illustrating the Recursions 143
Illustrative Examples and Applications 144
Fetal Lamb Movements 144
The Business Cycle 150
HMM Stationary and Predictive Distributions 153
Stationary Distribution of d_t 153
Predictive Distributions 154
Posterior Covariance of h 154
Available Software 154
Summary 154
References 155

7 Strategies and Methods for Prediction 159
Greg Ridgeway
Introduction to the Prediction Problem 160
Guiding Examples 160
Prediction Model Components 161
Loss Functions—What We are Trying to Accomplish 162
Common Regression Loss Functions 162
Common Classification Loss Functions 163
Cox Loss Function for Survival Data 166
Linear Models 167
Linear Regression 168
Classification 169
Generalized Linear Model 172
Nonlinear Models 174
Nearest Neighbor and Kernel Methods 174
Tree Models 177
Smoothing, Basis Expansions, and Additive Models 179
Neural Networks 182
Support Vector Machines 183
Boosting 185
Availability of Software 188
Summary 189
References 190

8 Principal Components and Factor Analysis 193
Daniel W. Apley
Introduction 194
Examples of Variation Patterns in Correlated Multivariate Data 194
Overview of Methods for Identifying Variation Patterns 197
Representation and Illustration of Variation Patterns in Multivariate Data 197
Principal Components Analysis 198
Definition of Principal Components 199
Using Principal Components as Estimates of the Variation Patterns 199

Factor Rotation 202
Capabilities and Limitations of PCA 202
Methods for Factor Rotation 203
Blind Source Separation 205
The Classic Blind Source Separation Problem 205
Blind Separation Principles 206
Fourth-Order Blind Separation Methods 208
Additional Manufacturing Applications 211
Available Software 211
Summary 212
References 212

9 Psychometric Methods of Latent Variable Modeling 215
Edward Ip, Igor Cadez, and Padhraic Smyth
Introduction 216
Basic Latent Variable Models 217
The Basic Latent Class Model 217
The Basic Finite Mixture Model 221
The Basic Latent Trait Model 224
The Basic Factor Analytic Model 226
Common Structure 229
Extension for Data Mining 229
Extending the Basic Latent Class Model 229
Extending the Basic Mixture Model 232
Extending the Latent Trait Model 233
Extending the Factor Analytic Model 234
An Illustrative Example 236
Hierarchical Structure in Transaction Data 236
Individualized Mixture Models 237
Data Sets 238
Experimental Results 238
References and Tools 241
References 241
Tools 243
Summary 244
References 244

10 Scalable Clustering 247
Joydeep Ghosh
Introduction 248
Clustering Techniques: A Brief Survey 249
Partitional Methods 250
Hierarchical Methods 255
Discriminative versus Generative Models 256
Assessment of Results 256
Visualization of Results 258
Clustering Challenges in Data Mining 259
Transactional Data Analysis 259

Next Generation Clickstream Clustering **260**
Clustering Coupled Sequences **261**
Large Scale Remote Sensing **261**
Scalable Clustering for Data Mining **262**
Scalability to Large Number of Records or Patterns, N **262**
Scalability to Large Number of Attributes or Dimensions, d **264**
Balanced Clustering **266**
Sequence Clustering Techniques **266**
Case Study: Similarity Based Clustering of Market Baskets and Web Logs **267**
Case Study: Impact of Similarity Measures on Web Document Clustering **270**
Similarity Measures: A Sampler **270**
Clustering Algorithms and Text Data Sets **272**
Comparative Results **273**
Clustering Software **274**
Summary **274**
Acknowledgments **274**
References **275**

11 Time Series Similarity and Indexing **279**
Gautam Das and Dimitrios Gunopulos
Introduction **279**
Time Series Similarity Measures **281**
Euclidean Distances and L_p Norms **281**
Normalization Transformations **282**
General Transformations **282**
Dynamic Time Warping **283**
Longest Common Subsequence Similarity **284**
Piecewise Linear Representations **287**
Probabilistic Methods **288**
Other Similarity Measures **288**
Indexing Techniques for Time Series **289**
Indexing Time Series When the Distance Function Is a Metric **290**
A Survey of Dimensionality Reduction Techniques **292**
Similar Time-Series Retrieval When the Distance Function Is Not a Metric **299**
Subsequence Retrieval **301**
Summary **302**
References **302**

12 Nonlinear Time Series Analysis **305**
Ying-Cheng Lai, Zonghua Liu, Nong Ye, and Tolga Yalcinkaya
Introduction **305**
Embedding Method for Chaotic Time Series Analysis **307**
Reconstruction of Phase Space **307**
Computation of Dimension **309**
Detection of Unstable Periodic Orbits **311**
Computing Lyapunov Exponents from Time Series **317**
Time-Frequency Analysis of Time Series **323**
Analytic Signals and Hilbert Transform **324**
Method of EMD **331**

Summary 338
Acknowledgment 338
References 338

13 Distributed Data Mining 341
Byung-Hoon Park and Hillol Kargupta
Introduction 342
Related Research 343
Data Distribution and Preprocessing 344
Homogeneous/Heterogeneous Data Scenarios 345
Data Preprocessing 345
Distributed Data Mining Algorithms 346
Distributed Classifier Learning 346
Collective Data Mining 349
Distributed Association Rule Mining 350
Distributed Clustering 351
Privacy Preserving Distributed Data Mining 352
Other DDM Algorithms 353
Distributed Data Mining Systems 353
Architectural Issues 354
Communication Models in DDM Systems 356
Components Maintenance 356
Future Directions 357
References 358

II: MANAGEMENT OF DATA MINING

14 Data Collection, Preparation, Quality, and Visualization 365
Dorian Pyle
Introduction 366
How Data Relates to Data Mining 366
The "10 Commandments" of Data Mining 368
What You Need to Know about Algorithms Before Preparing Data 369
Why Data Needs to be Prepared Before Mining It 370
Data Collection 370
Choosing the Right Data 370
Assembling the Data Set 371
Assaying the Data Set 372
Assessing the Effect of Missing Values 373
Data Preparation 374
Why Data Needs Preparing: The Business Case 374
Missing Values 375
Representing Time: Absolute, Relative, and Cyclic 376
Outliers and Distribution Normalization 377
Ranges and Normalization 378
Numbers and Categories 379
Data Quality 380
What Is Quality? 382
Enforcing Quality: Advantages and Disadvantages 384
Data Quality and Model Quality 384

Data Visualization 384
Seeing Is Believing 385
Absolute Versus Relative Visualization 388
Visualizing Multiple Interactions 391
Summary 391

15 Data Storage and Management 393
Tong (Teresa) Wu and Xiangyang (Sean) Li
Introduction 393
Text Files and Spreadsheets 395
Text Files for Data 395
Spreadsheet Files 395
Database Systems 397
Historical Databases 397
Relational Database 398
Object-Oriented Database 399
Advanced Topics in Data Storage and Management 402
OLAP 402
Data Warehouse 403
Distributed Databases 404
Available Software 406
Summary 406
Acknowledgments 407
References 407

16 Feature Extraction, Selection, and Construction 409
Huan Liu, Lei Yu, and Hiroshi Motoda
Introduction 410
Feature Extraction 411
Concepts 411
Algorithms 412
An Example 413
Summary 413
Feature Selection 414
Concepts 414
Algorithm 415
An Example 416
Summary 417
Feature Construction 417
Concepts 417
Algorithms and Examples 418
Summary 419
Some Applications 420
Summary 421
References 422

17 Performance Analysis and Evaluation 425
Sholom M. Weiss and Tong Zhang
Overview of Evaluation 426
Training versus Testing 426

Measuring Error 427
Error Measurement 427
Error from Regression 428
Error from Classification 429
Error from Conditional Density Estimation 429
Accuracy 430
False Positives and Negatives 430
Precision, Recall, and the *F* Measure 430
Sensitivity and Specificity 431
Confusion Tables 431
ROC Curves 432
Lift Curves 432
Clustering Performance: Unlabeled Data 432
Estimating Error 433
Independent Test Cases 433
Significance Testing 433
Resampling and Cross-Validation 435
Bootstrap 436
Time Series 437
Estimating Cost and Risk 437
Other Attributes of Performance 438
Training Time 438
Application Time 438
Interpretability 438
Expert Evaluation 439
Field Testing 439
Cost of Obtaining Labeled Data 439
References 439

18 Security and Privacy 441
Chris Clifton
Introduction: Why There Are Security and Privacy Issues with Data Mining 441
Detailed Problem Analysis, Solutions, and Ongoing Research 442
Privacy of Individual Data 442
Fear of What Others May Find in Otherwise Releasable Data 448
Summary 451
References 451

19 Emerging Standards and Interfaces 453
Robert Grossman, Mark Hornick, and Gregor Meyer
Introduction 453
XML Standards 454
XML for Data Mining Models 454
XML for Data Mining Metadata 456
APIs 456
SQL APIs 456
Java APIs 457
OLE DB APIs 457

Web Standards 457
Semantic Web 457
Data Web 458
Other Web Services 458
Process Standards 458
Relationships 458
Summary 459
References 459

III: APPLICATIONS OF DATA MINING

20 Mining Human Performance Data 463
David A. Nembhard
Introduction and Overview 463
Mining for Organizational Learning 464
Methods 464
Individual Learning 467
Data on Individual Learning 468
Methods 468
Individual Forgetting 474
Distributions and Patterns of Individual Performance 474
Other Areas 476
Privacy Issues for Human Performance Data 477
References 477

21 Mining Text Data 481
Ronen Feldman
Introduction 482
Architecture of Text Mining Systems 483
Statistical Tagging 485
Text Categorization 485
Term Extraction 489
Semantic Tagging 489
DIAL 491
Development of IE Rules 493
Auditing Environment 499
Structural Tagging 500
Given 500
Find 500
Taxonomy Construction 501
Implementation Issues of Text Mining 505
Soft Matching 505
Temporal Resolution 506
Anaphora Resolution 506
To Parse or Not to Parse? 507
Database Connectivity 507
Visualizations and Analytics for Text Mining 508
Definitions and Notations 508
Category Connection Maps 509

Relationship Maps 510
Trend Graphs 516
Summary 516
References 517

22 Mining Geospatial Data 519
Shashi Shekhar and Ranga Raju Vatsavai
Introduction 520
Spatial Outlier Detection Techniques 521
Illustrative Examples and Application Domains 521
Tests for Detecting Spatial Outliers 522
Solution Procedures 525
Spatial Colocation Rules 525
Illustrative Application Domains 526
Colocation Rule Approaches 527
Solution Procedures 530
Location Prediction 530
An Illustrative Application Domain 530
Problem Formulation 532
Modeling Spatial Dependencies Using the SAR and MRF Models 533
Logistic SAR 534
MRF Based Bayesian Classifiers 535
Clustering 537
Categories of Clustering Algorithms 539
K-Medoid: An Algorithm for Clustering 540
Clustering, Mixture Analysis, and the EM Algorithm 541
Summary 544
Acknowledgments 545
References 545

23 Mining Science and Engineering Data 549
Chandrika Kamath
Introduction 550
Motivation for Mining Scientific Data 551
Data Mining Examples in Science and Engineering 552
Data Mining in Astronomy 552
Data Mining in Earth Sciences 555
Data Mining in Medical Imaging 557
Data Mining in Nondestructive Testing 557
Data Mining in Security and Surveillance 558
Data Mining in Simulation Data 558
Other Applications of Scientific Data Mining 561
Common Challenges in Mining Scientific Data 561
Potential Solutions to Some Common Problems 562
Data Registration 564
De-Noising Data 565
Object Identification 566
Dimensionality Reduction 567
Generating a Good Training Set 568
Software for Scientific Data Mining 568

Summary 569
References 569

24 Mining Data in Bioinformatics 573
Mohammed J. Zaki
Introduction 574
Background 574
Basic Molecular Biology 574
Mining Methods in Protein Structure Prediction 575
Mining Protein Contact Maps 577
Classifying Contacts Versus Noncontacts 578
Mining Methodology 578
How Much Information Is There in Amino Acids Alone? 581
Using Local Structures for Contact Prediction 582
Characterizing Physical, Protein-Like Contact Maps 587
Generating a Database of Protein-Like Structures 588
Mining Dense Patterns in Contact Maps 589
Pruning and Integration 590
Experimental Results 591
Future Directions for Contact Map Mining 593
Heuristic Rules for "Physicality" 593
Rules for Pathways in Contact Map Space 594
Summary 595
References 596

25 Mining Customer Relationship Management (CRM) Data 597
Robert Cooley
Introduction 597
Data Sources 599
Data Types 599
E-Commerce Data 601
Data Preparation 604
Data Aggregation 605
Feature Preparation 607
Pattern Discovery 608
Pattern Analysis and Deployment 610
Robustness 610
Interestingness 611
Deployment 611
Sample Business Problems 612
Strategic Questions 612
Operational Questions 613
Summary 615
References 616

26 Mining Computer and Network Security Data 617
Nong Ye
Introduction 618
Intrusive Activities and System Activity Data 618

Phases of Intrusions **619**
Data of System Activities **620**
Extraction and Representation of Activity Features for Intrusion Detection **623**
Features of System Activities **624**
Feature Representation **625**
Existing Intrusion Detection Techniques **628**
Application of Statistical Anomaly Detection Techniques to Intrusion Detection **629**
Hotelling's T^2 Test and Chi-Square Distance Test **629**
Data Source and Representation **631**
Application of Hotelling's T^2 Test and Chi-Square Distance Test **633**
Testing Performance **633**
Summary **634**
References **635**

27 Mining Image Data **637**
Chabane Djeraba and Gregory Fernandez
Introduction **637**
Related Works **639**
Method **641**
How to Discover the Number of Clusters: k **641**
K-Automatic Discovery Algorithm **644**
Clustering Algorithm **646**
Experimental Results **646**
Data Sets **647**
Data Item Representation **648**
Evaluation Method **649**
Results and Analysis **650**
Summary **654**
References **655**

28 Mining Manufacturing Quality Data **657**
Murat C. Testik and George C. Runger
Introduction **657**
Multivariate Control Charts **658**
Hotelling T^2 Control Charts **658**
MEWMA Charts **660**
Nonparametric Properties of the MEWMA Control Charts **663**
Summary **667**
References **668**

Author Index **669**

Subject Index **681**