Contents

Foreword xxi
Preface xxiii
Acknowledgements xxxi

Chapter 1. Enterprise Data Mining: A Review and Research Directions, by T. W. Liao 1
1. Introduction 2
2. The Basics of Data Mining and Knowledge Discovery 6
   2.1 Data mining and the knowledge discovery process 6
   2.2 Data mining algorithms/methodologies 9
   2.3 Data mining system architectures 12
   2.4 Data mining software programs 14
3. Types and Characteristics of Enterprise Data 17
4. Overview of the Enterprise Data Mining Activities 23
   4.1 Customer related 23
   4.2 Sales related 30
   4.3 Product related 37
   4.4 Production planning and control related 43
   4.5 Logistics related 51
   4.6 Process related 55
      4.6.1 For the semi-conductor industry 55
      4.6.2 For the electronics industry 63
      4.6.3 For the process industry 72
      4.6.4 For other industries 79
   4.7 Others 83
   4.8 Summary 87
      4.8.1 Data type, size, and sources 87
      4.8.2 Data preprocessing 88
5. Discussion 90
Recent Advances in Data Mining of Enterprise Data

6. Research Programs and Directions 91
   6.1 On e-commerce and web mining 91
   6.2 On customer-related mining 92
   6.3 On sales-related mining 93
   6.4 On product-related mining 94
   6.5 On process-related mining 94
   6.6 On the use of text mining in enterprise systems 95

References 96

Author’s Biographical Statement 109


1. Credit Risk and Credit Rating 112
2. Data and Variables 115
3. Classification Techniques 115
   3.1 Logistic regression 116
   3.2 Discriminant analysis 117
   3.3 K-nearest neighbors 119
   3.4 Naïve Bayes 120
   3.5 The TAN technique 121
   3.6 Decision trees 122
   3.7 Associative classification 124
   3.8 Artificial neural networks 126
   3.9 Support vector machines 129
4. An Empirical Study 131
   4.1 Experimental settings 131
   4.2 The ROC curve and the Delong-Pearson method 133
   4.3 Experimental results 135
5. Conclusions and Future Work 139

References 140

Authors’ Biographical Statements 144

Chapter 3. Predictive Classification with Imbalanced Enterprise Data, by S. Daskalaki, I. Kopanas, and N. M. Avouris 147

1. Introduction 148
2. Enterprise Data and Predictive Classification 151
3. The Process of Knowledge Discovery from Enterprise Data 154
   3.1 Definition of the problem and application domain 155
   3.2 Creating a target database 156
   3.3 Data cleaning and preprocessing 157

1. Introduction
   1.1 Background and motives
   1.2 Objectives

2. Literature Review
   2.1 Traditional time series forecasting research
   2.2 Neural network based forecasting methods
   2.3 Hybridizing a genetic algorithm (GA) with a neural network for forecasting
      2.3.1 Using a GA to design the NN architecture
      2.3.2 Using a GA to generate the NN connection weights
   2.4 Review of sales forecasting research

3. Problem Definition
   3.1 Scope of the research data
   3.2 Characteristics of the variables considered
      3.2.1 Macroeconomic domain
      3.2.2 Downstream demand domain
      3.2.3 Industrial production domain
      3.2.4 Time series domain
   3.3 The performance index

4. Methodology
   4.1 Data preprocessing
      4.1.1 Gray relation analysis
      4.1.2 Winter’s exponential smoothing
   4.2 Evolving neural networks (ENN)
      4.2.1 ENN modeling
      4.2.2 ENN parameters design
Chapter 5. Data Mining Applications of Process Platform Formation for High Variety Production,
by J. Jiao and L. Zhang

1. Background
2. Methodology
3. Routing Similarity Measure
   3.1 Node content similarity measure
      3.1.1 Material similarity measure
         3.1.1.1 Procedure for calculating similarities between primitive components
         3.1.1.2 Procedure for calculating similarities between compound components
      3.1.2 Product similarity measure
      3.1.3 Resource similarity measure
      3.1.4 Operation similarity and node content similarity measures
         3.1.5 Normalized node content similarity matrix
   3.2 Tree structure similarity measure
   3.3 ROU similarity measure
4. ROU Clustering
5. ROU Unification
   5.1 Basic routing elements
   5.2 Master and selective routing elements
   5.3 Basic tree structures
   5.4 Tree growing
Chapter 6. A Data Mining Approach to Production Control in Dynamic Manufacturing Systems, by H.-S. Min and Y. Yih

1. Introduction 288
2. Previous Approaches to Scheduling of Wafer Fabrication 291
3. Simulation Model and Solution Methodology 294
   3.1 Simulation model 294
   3.2 Development of a scheduler 298
      3.2.1 Decision variables and decision rules 298
      3.2.2 Evaluation criteria: system performance and status 300
      3.2.3 Data collection: a simulation approach 300
      3.2.4 Data classification: a competitive neural network approach 301
      3.2.5 Selection of decision rules for decision variables 306
4. An Experimental Study 306
   4.1 Experimental design 306
   4.2 Results and analyses 309
5. Related Studies 313
6. Conclusions 317
References 319
Authors’ Biographical Statements 321

Chapter 7. Predicting Wine Quality from Agricultural Data with Single-Objective and Multi-Objective Data Mining Algorithms, by M. Last, S. Elnekave, A. Naor, and V. Schoenfeld

1. Introduction 324
2. Problem Description 325
3. Information Networks and the Information Graph 329
   3.1 An extended classification task 329
   3.2 Single-objective information networks 330
   3.3 Multi-objective information networks 336
   3.4 Information graphs 338
4. A Case Study: the Cabernet Sauvignon problem
   4.1 Data selection
   4.2 Data pre-processing
      4.2.1 Ripening data
      4.2.2 Meteorological measurements
   4.3 Design of data mining runs
   4.4 Single-objective models
   4.5 Multi-objective models
   4.6 Comparative evaluation
   4.7 The knowledge discovered and its potential use
5. Related Work
   5.1 Mining of agricultural data
   5.2 Multi-objective classification models and algorithms
6. Conclusions
References
Authors’ Biographical Statements

Chapter 8. Enhancing Competitive Advantages and Operational Excellence for High-Tech Industry through Data Mining and Digital Management, by C.-F. Chien and S.-C. Hsu
1. Introduction
2. Knowledge Discovery in Databases and Data Mining
   2.1 Problem types for data mining in the high-tech industry
   2.2 Data mining methodologies
      2.2.1 Decision trees
         2.2.1.1 Decision tree construction
         2.2.1.2 CART
         2.2.1.3 C4.5
         2.2.1.4 CHAID
      2.2.2 Artificial neural networks
         2.2.2.1 Associate learning networks
         2.2.2.2 Supervised learning networks
         2.2.2.3 Unsupervised learning networks
3. Application of Data Mining in Semiconductor Manufacturing
   3.1 Problem definition
   3.2 Types of data mining applications
      3.2.1 Extracting characteristics from WAT data
      3.2.2 Process failure diagnosis of CP and engineering data
      3.2.3 Process failure diagnosis of WAT and engineering data
      3.2.4 Extracting characteristics from semiconductor manufacturing data
### Contents

3.3 A Hybrid decision tree approach for CP low yield diagnosis 400  
3.4 Key stage screening 402  
3.5 Construction of the decision tree 404  
4. Conclusions 406  
References 407  
Authors’ Biographical Statements 411  

#### Chapter 9. Multivariate Control Charts from a Data Mining Perspective, by G. C. Porzio and G. Ragozini

1. Introduction 414  
2. Control Charts and Statistical Process Control Phases 415  
3. Multivariate Statistical Process Control 419  
   3.1 The sequential quality control setting 419  
   3.2 The hotelling $T^2$ control chart 421  
4. Is the $T^2$ Statistic Really Able to Tackle Data Mining Issues? 424  
   4.1 Many data, many outliers 424  
   4.2 Questioning the assumptions on shape and distribution 430  
5. Designing Nonparametric Charts When Large HDS Are Available: the Data Depth Approach 434  
   5.1 Data depth and control charts 436  
   5.2 Towards a parametric setting for data depth control charts 438  
   5.3 A Shewhart chart for changes in location and increases in scale 442  
   5.4 An illustrative example 443  
   5.5 Average run length functions for data depth control charts 446  
   5.6 A simulation study of chart performance 448  
   5.7 Choosing an empirical depth function 453  
6. Final Remarks 454  
References 456  
Authors’ Biographical Statements 462  

#### Chapter 10. Data Mining of Multi-Dimensional Functional Data for Manufacturing Fault Diagnosis, by M. K. Jeong, S. G. Kong, and O. A. Omitaomu

1. Introduction 464  
2. Data Mining of Functional Data 465  
   2.1 Dimensionality reduction techniques for functional data 465  
   2.2 Multi-scale fault diagnosis 468  
      2.2.1 A case study: data mining of functional data 469  
   2.3 Motor shaft misalignment prediction based on functional data 472  
      2.3.1 Techniques for predicting with high number of predictors 474  
      2.3.2 A case study: motor shaft misalignment prediction 477  
References 478  
Authors’ Biographical Statements 484
3. Data Mining in Hyperspectral Imaging 481
   3.1 A hyperspectral fluorescence imaging system 483
   3.2 Hyperspectral image dimensionality reduction 485
   3.3 Spectral band selection 490
   3.4 A case study: data mining in hyperspectral imaging 494
4. Conclusions 496
References 497
Authors’ Biographical Statements 503

Chapter 11. Maintenance Planning Using Enterprise Data Mining,
   by L. P. Khoo, Z. W. Zhong, and H. Y. Lim 505
1. Introduction 506
2. Rough Sets, Genetic Algorithms, and Tabu Search 508
   2.1 Rough sets 508
      2.1.1 Overview 508
      2.1.2 Rough sets and fuzzy sets 509
      2.1.3 Applications 510
      2.1.4 The strengths of the theory of rough sets 511
      2.1.5 Enterprise information and the information system 512
   2.2 Genetic algorithms 516
   2.3 Tabu search 520
3. The Proposed Hybrid Approach 521
   3.1 Background 521
   3.2 The rough set engine 521
   3.3 The tabu-enhanced GA engine 523
   3.4 Rule organizer 528
4. A Case Study 528
   4.1 Background 528
      4.1.1 Mounting bracket failures 531
      4.1.2 The alignment problem 532
      4.1.3 Sea/land inner/outer guide roller failures 532
   4.2 Analysis using the proposed hybrid approach 532
   4.3 Discussion 537
      4.3.1 Validity of the extracted rules 537
      4.3.2 A comparative analysis of the results 538
5. Conclusions 540
References 541
Authors’ Biographical Statements 544
Chapter 12. Data Mining Techniques for Improving Workflow Model, by D. Gunopulos and S. Subramaniam

1. Introduction 546
2. Workflow Models 549
3. Discovery of Models from Workflow Logs 552
4. Managing Flexible Workflow Systems 555
5. Workflow Optimization Through Mining of Workflow Logs 557
   5.1 Repositioning decision points 557
   5.2 Prediction of execution paths 560
6. Capturing the Evolution of Workflow Models 565
7. Applications in Software Engineering 566
   7.1 Discovering reasons for bugs in software processes 567
   7.2 Predicting the control flow of a software process for efficient resource management 568
8. Conclusions 569
References 569
Authors’ Biographical Statements 576

Chapter 13. Mining Images of Cell-Based Assays, by P. Perner

1. Introduction 578
2. The Application Used for the Demonstration of the System Capability 580
3. Challenges and Requirements for the Systems 582
4. The Cell-Interpreter’s Architecture 582
5. Case-Based Image Segmentation 584
   5.1 The case-based reasoning unit 585
   5.2 Management of case bases 587
6. Feature Extraction 588
   6.1 Our flexible texture descriptor 589
7. The Decision Tree Induction Unit 591
   7.1 The basic principle 591
   7.2 Terminology of the decision tree 592
   7.3 Subtasks and design criteria for decision tree induction 594
   7.4 Attribute selection criteria 597
      7.4.1 Information gain criteria and the gain ratio 598
      7.4.2 The Gini function 600
   7.5 Discretization of attribute values 601
      7.5.1 Binary discretization 603
         7.5.1.1 Binary discretization based on entropy 603
         7.5.1.2 Discretization based on inter- and intra-class variance 604
7.5.2 Multi-interval discretization
  7.5.2.1 The basic (Search strategies) algorithm
  7.5.2.2 Determination of the number of intervals
  7.5.2.3 Cluster utility criteria
  7.5.2.4 MLD-based criteria
  7.5.2.5 LVQ-based discretization
  7.5.2.6 Histogram-based discretization
  7.5.2.7 Chi-Merge discretization

7.5.3 The influence of discretization methods on the resulting decision tree

7.5.4 Discretization of categorical or symbolic attributes
  7.5.4.1 Manual abstraction of attribute values
  7.5.4.2 Automatic aggregation

7.6 Pruning
  7.6.1 Overview of pruning methods
  7.6.2 Cost-complexity pruning

7.7 Some general remarks

8. The Case-Based Reasoning Unit

9. Concept Clustering as Knowledge Discovery

10. The Overall Image Mining Procedure
  10.1 A case study
  10.2 Brainstorming and image catalogue
  10.3 The interviewing process
  10.4 Collection of image descriptions into the database
  10.5 The image mining experiment
  10.6 Review
  10.7 Lessons learned

11. Conclusions and Future Work

References

Author’s Biographical Statement

Chapter 14. Support Vector Machines and Applications,
by T. B. Trafalis and O. O. Oladunni

1. Introduction

2. Fundamentals of Support Vector Machines
  2.1 Linear separability
  2.2 Linear inseparability
  2.3 Nonlinear separability
  2.4 Numerical testing
    2.4.1 The AND problem
    2.4.2 The XOR problem
3. Least Squares Support Vector Machines 657
4. Multi-Classification Support Vector Machines 662
   4.1 The one-against-all (OAA) method 662
   4.2 The one-against-one (OAO) method 664
   4.3 Pairwise multi-classification support vector machines 665
   4.4 Further techniques based on central representation of the version space 672
5. Some Applications 674
   5.1 Enterprise modeling (novelty detection) 674
   5.2 Non-enterprise modeling application (multiphase flow) 679
6. Conclusions 681
References 682
Authors’ Biographical Statements 689

Chapter 15. A Survey of Manifold-Based Learning Methods, by X. Huo, X. Ni, and A. K. Smith 691
1. Introduction 692
2. Survey of Existing Methods 694
   2.1 Group 1: Principal component analysis (PCA) 695
   2.2 Group 2: Semi-classical methods: multidimensional scaling (MDS) 697
      2.2.1 Solving MDS as an eigenvalue problem 698
   2.3 Group 3: Manifold searching methods 699
      2.3.1 Generative topographic mapping (GTM) 699
      2.3.2 Locally linear embedding (LLE) 701
      2.3.3 ISOMAP 703
   2.4 Group 4: Methods from spectral theory 704
      2.4.1 Laplacian eigenmaps 704
      2.4.2 Hessian eigenmaps 706
   2.5 Group 5: Methods based on global alignment 707
3. Unification via the Null-Space Method 708
   3.1 LLE as a null-space based method 709
   3.2 LTSA as a null-space based method 711
   3.3 Comparison between LTSA and LLE 712
4. Principles Guiding the Methodological Developments 713
   4.1 Sufficient dimension reduction 713
   4.2 Desired statistical properties 714
      4.2.1 Consistency 714
      4.2.2 Rate of convergence 715
      4.2.3 Exhaustiveness 715
      4.2.4 Robustness 716
4.3 Initial results  716
  4.3.1 Formulation and related open questions  716
  4.3.2 Consistency of LTSA  718
5. Examples and Potential Applications  722
  5.1 Successes of manifold based methods on synthetic data  722
    5.1.1 Examples of LTSA recovering implicit parameterization  722
    5.1.2 Examples of Locally Linear Projection (LLP) in denoising  724
  5.2 Curve clustering  725
  5.3 Image detection  728
    5.3.1 Formulation  731
    5.3.2 Distance to manifold  732
    5.3.3 SRA: the significance run algorithm  733
    5.3.4 Parameter estimation  734
      5.3.4.1 Number of nearest neighbors  734
      5.3.4.2 Local dimension  734
    5.3.5 Simulations  736
    5.3.6 Discussion  738
  5.4 Application on the localization of sensor networks  738
6. Conclusions  740
References  741
Authors’ Biographical Statements  745

Chapter 16. Predictive Regression Modeling for Small Enterprise Data Sets with Bootstrap, Clustering, and Bagging
by C. J. Feng and K. Erla  747
1. Introduction  748
2. Literature Review  750
  2.1 Tree-based classifiers and the bootstrap 0.632 rule  750
  2.2 Bagging  751
3. Methodology  753
  3.1 The data modeling procedure  753
  3.2 Bootstrap sampling  753
  3.3 Selecting the best subset regression model  756
  3.4 Evaluation of prediction errors  758
    3.4.1 Prediction error evaluation  758
    3.4.2 The 0.632 prediction error  759
  3.5 Cluster analysis  760
  3.6 Bagging  760
4. A Computational Study  761
  4.1 The experimental data  761
  4.2 Computational results  761
Contents

5. Conclusions 770
References 771
Authors’ Biographic Statements 774

Subject Index 775
List of Contributors 779
About the Editors 785