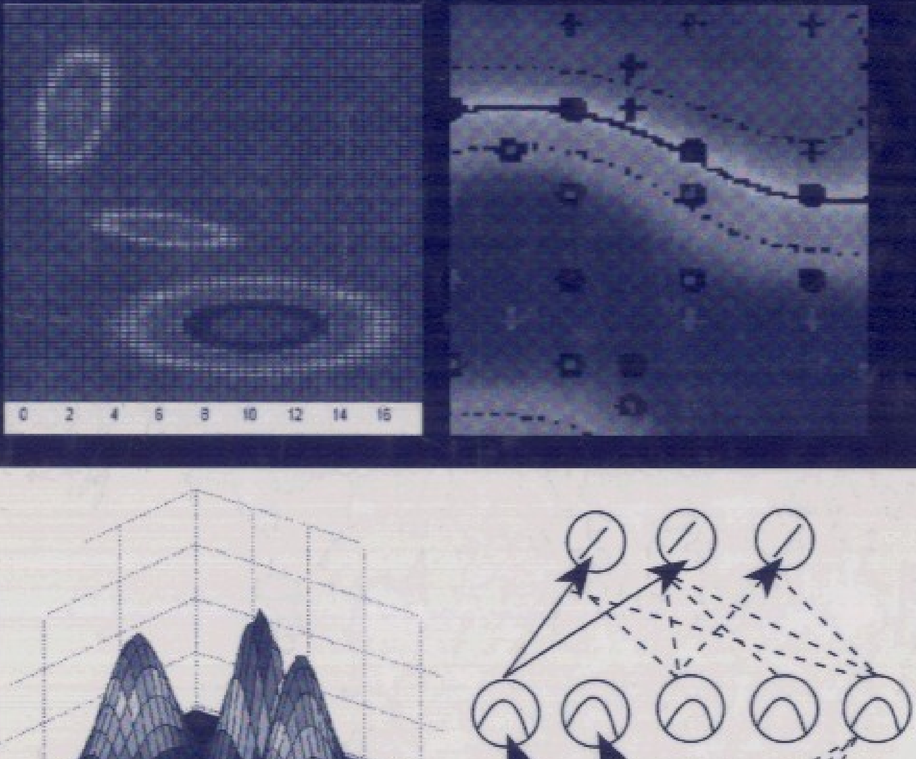
Statistical Data Mining and Knowledge Discovery



Edited by
Hamparsum Bozdogan

CHAPMAN & HALL/CRC

Contents

1 **The Role of Bayesian and Frequentist Multivariate Modeling in Statistical Data Mining** 1

S. James Press University of California, Riverside, USA

1.1 Introduction 1

1.2 Is Data Mining Science? 2

1.3 Genesis of Data Mining 3

1.4 The Data Cube and Databases 3

1.4.1 Some Examples 4

1.5 Structured Query Language 5

1.6 Statistical Problems with Data Mining 6

1.7 Some DM Approaches to Dimension Reduction 7

1.7.1 Graph-Based Multivariate Models 7

1.7.2 Multivariate Bayes Classification and Prediction 8

1.7.3 Markov Independence Models 9

1.8 Prior Distributions in Data Mining 9

1.9 Some New DM Applications 10

2 **Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms** 15

Hamparsum Bozdogan University of Tennessee, Knoxville, USA

2.1 Introduction 15

2.2 What is Information Complexity:ICOMP? 17

2.2.1 The Concept of Complexity and Complexity of a System . . 19

2.2.2 Information Theoretic Measure of Complexity of a Multivariate Distribution 20

2.2.3 Initial Definition of Covariance Complexity 22

2.2.4 Definition of Maximal Covariance Complexity 23

2.2.5 ICOMP as an Approximation to the Sum of Two Kullback-Leibler Distances 27

2.3 Information Criteria for Multiple Regression Models 31

2.3.1 ICOMP Based on Complexity Measures 33

2.3.2 ICOMP Under Misspecification 34

2.3.3 AIC and AIC-Type Criteria 36

2.4 A GA for the Regression Modeling 36

2.5 Numerical Examples 41

2.5.1 Subset Selection of Best Predictors in Multiple Regression: A Simulation Example 41
2.5.2 Subset Selection of Best Predictors in Multiple Regression: A Real Example 43
2.6 Conclusion and Discussion 49

3 Econometric and Statistical Data Mining, Prediction and Policy-Making 57
Arnold Zellner University of Chicago, Chicago, USA
3.1 Introduction 58
3.2 Brief Comments on Scientific Method and Data Mining 59
3.3 The Structural Econometric Modeling, Time Series Analysis (SEMTSA) Approach 61
3.3.1 The SEMTSA Approach 62
3.3.2 Statistical Inference Procedures 64
3.4 Methods Employed in Data Analysis, Modeling and Forecasting . . 67
3.5 Disaggregation and the Marshallian Macroeconomic Model 71
3.6 A Complete Marshallian Macroeconomic Model 74

4 Data Mining Strategies for the Detection of Chemical Warfare Agents 79
Jeffrey. L. Solka,[1,2] Edward J. Wegman,[1] and David J. Marchette[2]
[2]Naval Surface Warfare Center (NSWCDD), Dahlgren, VA, [1]George Mason University, Fairfax, VA, USA
4.1 Introduction 79
4.2 Results 82
4.3 Conclusions 90

5 Disclosure Limitation Methods Based on Bounds for Large Contingency Tables With Applications to Disability 93
Adrian Dobra, Elena A. Erosheva and Stephen E. Fienberg Duke University, Durham, University of Washington, Seattle, and Carnegie-Mellon University, Pittsburgh, USA
5.1 Introduction 94
5.2 Example: National Long Term Care Survey Data 95
5.3 Technical Background on Cell Entry Bounds 96
5.4 Decomposable Frontiers 99
5.4.1 Calculating Decomposable Frontiers 100
5.4.2 Analysis of the 2^{16} NLTCS Example 101
5.5 "Greedy" Frontiers 103
5.6 Bounds 108
5.6.1 Bounds in the Decomposable Case 108
5.6.2 Bounds in the Non-decomposable Case 109
5.7 Discussion 112
References 114

6 Partial Membership Models with Application to Disability Survey Data **117**

Elena A. Erosheva University of Washington, USA

6.1 Motivation 118

6.2 Functional Disability Data 119

6.2.1 Marginal Frequencies and Simple Statistics 120

6.2.2 Frequent Responses 120

6.2.3 Total Number of Disabilities 121

6.3 Full Versus Partial Membership 123

6.3.1 Full Membership: Latent Class Model 123

6.3.2 Partial Membership: GoM Model 124

6.4 Bayesian Estimation of the GoM Model 125

6.5 Analysis and Comparison 127

6.5.1 Latent Class Analysis 127

6.5.2 Grade of Membership Analysis and Comparison 128

6.6 Concluding Remarks 132

7 Automated Scoring of Polygraph Data **135**

Aleksandra B. Slavkovic Department of Statistics, Carnegie-Mellon University, Pittsburgh, USA

7.1 Introduction 135

7.2 Background 136

7.2.1 The Polygraph Examination 136

7.2.2 Instrumentation and Measurements 137

7.2.3 Chart Evaluations 138

7.3 Statistical Models for Classification and Prediction 139

7.4 The Data 141

7.5 Statistical Analysis 144

7.5.1 Signal Processing 144

7.5.2 A Simplified Approach to Feature Extraction 145

7.5.3 Feature Evaluation, Modeling, and Classification 147

7.5.4 Logistic Regression 148

7.5.5 Classification Results 149

7.6 Discussion 150

7.7 Conclusion 152

8 Missing Value Algorithms in Decision Trees **155**

Hyunjoong Kim and Sumer Yates University of Tennessee, Knoxville, USA

8.1 Introduction 155

8.2 The Seven Algorithms 156

8.2.1 Probability Split 156

8.2.2 Alternative Split 156

8.2.3 Proxy Split 157

8.2.4 Surrogate Split 158

8.2.5 Root Node Imputation 159

8.2.6 Nodewise Imputation 159
8.2.7 Separate Node 159
8.3 The Simulation Study 159
8.3.1 Experiment-A 159
8.3.2 Experiment-B 160
8.4 Results 162
8.4.1 Experiment-A 162
8.4.2 Experiment-B 163
8.5 Conclusions 166

9 Unsupervised Learning from Incomplete Data Using a Mixture Model Approach 173

Lynette Hunt and Murray Jorgensen University of Waikato, Hamilton, New Zealand

9.1 Introduction 173
9.2 Clustering by Mixture Models 175
9.2.1 Latent Class Models and Local Independence 177
9.2.2 Generalizing Local Independence: the Multimix Model . . . 177
9.2.3 Missing Data 178
9.3 Applications 182
9.3.1 Example 1: The Iris Data 183
9.3.2 Example 2: The Cancer Data 186
9.4 Discussion 188

10 Improving the Performance of Radial Basis Function (RBF) Classification Using Information Criteria 193

Zhenqiu Liu and Hamparsum Bozdogan University of Tennessee, Knoxville, USA

10.1 Introduction 193
10.1.1 Binary Classifiers 194
10.1.2 Logistic Regression and Mixture Model 195
10.1.3 Multi-Class Classifier 197
10.2 Regression Trees 197
10.2.1 Information Criteria 199
10.2.2 Parameter Estimation 200
10.3 New Kernel Functions 201
10.4 The EM Algorithm 204
10.4.1 Constrained Maximization of a Convex Function 204
10.4.2 The EM Algorithm 205
10.4.3 EM-Algorithm for PE Mixture Models 207
10.5 Hybrid Training 208
10.6 Computational Results 210
10.7 Conclusions 212

11 Use of Kernel Based Techniques for Sensor Validation in Nuclear Power Plants 217
Andrei V. Gribok, Aleksey M. Urmanov, J. Wesley Hines, and Robert E. Uhrig University of Tennessee, Knoxville, USA
11.1 Introduction 217
11.2 Collinear, Ill-Posed Problems, Regularization 218
11.3 Kernel Regression and MSET 222
11.4 Support Vector Machines 223
11.5 Data Description and Results 225
11.6 Conclusions 228

12 Data Mining and Traditional Regression 233
Christopher M. Hill, Linda C. Malone, and Linda Trocine University of Central Florida, Orlando, FL, USA
12.1 Introduction 234
12.2 Military Manpower Application 234
12.3 Data Mining and Traditional Regression 236
12.4 General Problems 237
12.5 Attempted Solutions 239
12.5.1 Algorithmic Improvements 239
12.5.2 Data Complexity Reduction 240
12.6 Regression Specific Issues 240
12.6.1 Linear Regression 241
12.6.2 Logistic Regression 243
12.7 Conclusion 246

13 An Extended Sliced Inverse Regression 251
Masahiro Mizuta Hokkaido University, Sapporo, Japan
13.1 Introduction 251
13.2 Algorithms for SIR Model 252
13.3 Relative Projection Pursuit 254
13.4 SIRrpp 254
13.5 Concluding Remarks 256

14 Using Genetic Programming to Improve the Group Method of Data Handling in Time Series Prediction 257
M. Hiassat, M.F. Abbod, and N. Mort University of Sheffield, Sheffield, UK
14.1 Introduction 257
14.2 The Data 258
14.3 Financial Data 259
14.4 Weather Data 259
14.5 Processing of Data 260
14.6 The Group Method of Data Handling (GMDH) 261
14.7 Genetic Programming (GP) 262
14.8 GP-GMDH 263

14.9 Results and Discussion 264
14.10Conclusion and Further Work 267

15 Data Mining for Monitoring Plant Devices Using GMDH and Pattern Classification 269
B.R. Upadhyaya and B. Lu University of Tennessee, Knoxville
15.1 Introduction 269
15.2 Description of the Method 273
15.2.1 Group Method of Data Handling (GMDH) 273
15.2.2 Analysis of Residuals for Fault Diagnosis 276
15.3 Analysis and Results 277
15.4 Concluding Remarks 278

16 Statistical Modeling and Data Mining to Identify Consumer Preferences 281
Francois Boussu[1] *and Jean Jacques Denimal*[2] [1]Ecole Nationale Superieure des Arts et Industries Textiles, Roubaix, and [2]University of Sciences and Technologies of Lille, France
16.1 Introduction 282
16.1.1 The Garment Market 282
16.1.2 Interest of the Proposed Method 282
16.2 Data Mining Method 283
16.2.1 From Data to Analyses 283
16.2.2 Proposed Methodology 284
16.2.3 Data Visualization Tools 288
16.3 Application to a Textile Data Set 288
16.3.1 Classical and Circular Representations of the Hierarchy . . 288
16.3.2 Graphic Plane Representations Associated with Nodes . . . 289
16.3.3 Shaded Tables Visualizing Associations Between Items and Periods 291
16.4 Conclusion 292

17 Testing for Structural Change Over Time of Brand Attribute Perceptions in Market Segments 297
Sara Dolničar and Friedrich Leisch University of Wollongong, and Vienna University of Technology, Austria
17.1 Introduction 297
17.2 The Managerial Problem 298
17.3 Results from Traditional Analysis 299
17.3.1 Brand Image Analysis 299
17.3.2 Competition Analysis 300
17.4 The PBMS and DynPBMS Approaches 300
17.4.1 General Principles 300
17.4.2 Complexity Reduction 301

17.4.3 Tests on Heterogeneity 303
17.4.4 Competition Analysis 305
17.5 Summary 306

18 Kernel PCA for Feature Extraction with Information Complexity 309
Zhenqiu Liu and Hamparsum Bozdogan University of Tennessee, Knoxville, USA
18.1 Introduction 310
18.2 Kernel Functions 312
18.3 Kernel PCA 314
18.4 EM for Kernel PCA and On-line PCA 318
18.5 Choosing the Number of Components with Information Complexity 319
18.6 Computational Results 320
18.7 Conclusions 321

19 Global Principal Component Analysis for Dimensionality Reduction in Distributed Data Mining 323
Hairong Qi, Tse-Wei Wang, and J. Douglas Birdwell University of Tennessee, Knoxville, USA
19.1 Introduction 324
19.2 Principal Component Analysis 326
19.3 Global PCA for Distributed Homogeneous Databases 327
19.4 Global PCA for Distributed Heterogeneous Databases 330
19.5 Experiments and Results 331
19.5.1 Global PCA for Distributed Homogeneous Databases 332
19.5.2 Global PCA for Distributed Heterogeneous Databases . . . 334
19.6 Conclusion 336

20 A New Metric for Categorical Data 339
S. H. Al-Harbi, G. P. McKeown and V. J. Rayward-Smith University of East Anglia, Norwich, UK
20.1 Introduction 339
20.2 Dissimilarity Measure 340
20.2.1 Mahalanobis Distance 341
20.3 D_{CV} Metric 343
20.3.1 Weights of Fields 344
20.4 Synthetic Examples 345
20.5 Exploiting the D_{CV} Metric 348
20.5.1 k-Means Algorithm 348
20.5.2 Case-Based Reasoning 348
20.5.3 Supervised Clustering 349
20.6 Conclusions and Future Work 349

21 **Ordinal Logistic Modeling Using ICOMP as a Goodness-of-Fit Criterion** **353**
J. Michael Lanning and Hamparsum Bozdogan University of Tennessee, Knoxville, USA
21.1 Introduction 353
21.2 Model Selection Criteria 356
21.2.1 Computational Software 358
21.3 Ordinal Logistic Regression 359
21.3.1 Cumulative Link Models 360
21.3.2 Cumulative Logit (Proportional Odds) Model 361
21.3.3 Proportional Hazard Model 363
21.3.4 Continuation-Ratio Model 363
21.3.5 Category Boundary Cutoffs (Correlated Data) 363
21.3.6 Model Selection 365
21.4 Example Problem: Diabetes Severity 367
21.4.1 Proportional Odds Model 367
21.4.2 Residual (Outlier) Analysis 367
21.4.3 Nested Dichotomous Models 368
21.4.4 Nested Cumulative Link Dichotomous Models 368
21.4.5 Nested Continuation-Ratio Link Dichotomous Models . . . 368
21.5 Conclusions 369

22 **Comparing Latent Class Factor Analysis with the Traditional Approach in Data Mining** **373**
Jay Magidson and Jeroen Vermunt Statistical Innovations Inc., USA, and Tilburg University, The Netherlands
22.1 Introduction 373
22.2 The Basic LC Factor Model 375
22.3 Examples 376
22.3.1 Rater Agreement 376
22.3.2 MBTI Personality Items 379
22.4 Conclusion 381

23 **On Cluster Effects in Mining Complex Econometric Data** **385**
M. Ishaq Bhatti Sultan Qaboos University, Muscat, OMAN
23.1 Introduction 386
23.2 The Model 387
23.3 An Algorithm for Full Maximum Likelihood Estimation 389
23.4 Application of the Model 392
23.5 Fixed Coefficient Regression Models 394
23.6 Concluding Remarks 395

24 **Neural Network-Based Data Mining Techniques for Steel Making** **401**
Ravindra K. Sarma, Amar Gupta, and Sanjeev Vadhavkar Massachusetts Institute of Technology, Cambridge, USA

24.1 Introduction 402
24.2 Productivity from Information Technology (PROFIT) Initiative . . . 403
24.3 Description of Predictive Model 406
24.4 NNRUN – ANN Training Suite 407
24.4.1 Data Manipulation 407
24.4.2 Automated Search For Best Network 408
24.5 Results and Analysis 409
24.6 Conclusions 411

25 Solving Data Clustering Problem as a String Search Problem 415
V. Olman, D. Xu, and Y. Xu Oak Ridge National Laboratory, Oak Ridge, TN, USA
25.1 Introduction 415
25.2 Mathematical Framework 417
25.2.1 Definition of Cluster 418
25.2.2 MST-Representation of High-Dimensional Data 418
25.2.3 Relationship between MSTs and Clusters 418
25.3 Stability of MST Structure Under Noise 421
25.4 Statistical Assessment of Identified Clusters 422
25.5 Applications 423
25.5.1 Partitioning of Simulated Data 423
25.5.2 Regulatory Binding Site Identification 424
25.5.3 Cluster Identification in Gene Expression Profiles 426
25.6 Discussion 428

26 Behavior-Based Recommender Systems as Value-Added Services for Scientific Libraries 433
Andreas Geyer-Schulz 1, Michael Hahsler 2, Andreas Neumann 1, and Anke Thede 1 1Universität Karlsruhe (TH), Germany, 2WU-Wien, Austria
26.1 Introduction 433
26.2 Recommender Services for Legacy Library Systems 435
26.3 Ehrenberg's Repeat-Buying Theory for Libraries 439
26.4 A Recommender System for the Library of the Universität Karlsruhe (TH) 448
26.5 Conclusion 451

27 GTP (General Text Parser) Software for Text Mining 455
Justin T. Giles, Ling Wo, and Michael W. Berry University of Tennessee, Knoxville, USA
27.1 Introduction 455
27.2 Model Facilitated by GTP 456
27.3 GTP Usage and Files Generated 457
27.4 Overview of GTP Options 458
27.5 Query Processing with GTPQUERY 464
27.6 Example 464

27.7 Versions of GTP and GTPQUERY 469
27.8 Code Evolution 470
27.9 Future Work 470

28 Implication Intensity: From the Basic Statistical Definition to the Entropic Version **473**
Julien Blanchard, Pascale Kuntz, Fabrice Guillet, and Regis Gras Ecole Polytechnique de l'Universite de Nantes, France
28.1 Introduction 473
28.2 First Definitions 475
28.3 Entropic Version 476
28.4 Experimental Results 478
28.4.1 Experiments with Synthetic Data 478
28.4.2 Experiments with Real Data 479
28.5 Conclusion 483

29 Use of a Secondary Splitting Criterion in Classification Forest Construction **487**
Chang-Yung Yu and Heping Zhang[1] Yale University, New Haven, USA
29.1 Introduction 487
29.2 A Secondary Node-Splitting Criterion 488
29.3 The Formation of a Deterministic Forest 488
29.4 Comparison Data 489
29.4.1 Comparison through Cross-Validation 490
29.4.2 Understanding the Deterministic Forest 490
29.5 Discussion 491

30 A Method Integrating Self-Organizing Maps to Predict the Probability of Barrier Removal **497**
Zhicheng Zhang and Frédéric Vanderhaegen University of Valenciennes, Le Mont Houy, France.
30.1 Introduction 498
30.2 A Method Integrating Self-Organizing Maps Algorithm 498
30.2.1 Problematic in BR Data Analysis 498
30.2.2 A Hierarchical SOM (HSOM) Algorithm-Based Method . . 500
30.3 Experimental Results 503
30.4 Discussions 507
30.5 Conclusions 509

31 Cluster Analysis of Imputed Financial Data Using an Augmentation-Based Algorithm **513**
H. Bensmail and R. P. DeGennaro University of Tennessee, Knoxville, TN, USA
31.1 Introduction 513
31.2 Data and Preliminary Tests 514

31.3 Clustering and Bayesian Data Augmentation 518
  31.3.1 Imputation 520
  31.3.2 Posterior Estimation 521
  31.3.3 Algorithm 522
31.4 Bayesian Model Selection for Choosing the Number of Clusters . . 523
31.5 Analysis of Financial Data 523
31.6 Discussion 525

32 Data Mining in Federal Agencies **529**
David L. Banks and Robert T. Olszewski U.S. Food and Drug Administration, Rockville, MD, and University of Pittsburgh, Pittsburgh, PA, USA
32.1 Data Quality 529
  32.1.1 First Steps 530
  32.1.2 Using Poor Quality Data 533
  32.1.3 Estimating the Probability of Bad Data 534
32.2 Indexing Data 534
  32.2.1 A First-Pass Method 535
  32.2.2 Estimating Distance Functions 536
  32.2.3 Computation 536
32.3 Screening for Structure with Locally Low Dimension 537
  32.3.1 Screening Strategy 539
  32.3.2 The Designed Experiment 540
  32.3.3 Regression 544
32.4 Estimating Exposure 545
  32.4.1 Categorical Data 545

33 STING: Evaluation of Scientific & Technological Innovation and Progress **549**
S. Sirmakessis[1], K. Markellos[2], P. Markellou[3], G. Mayritsakis[4], K. Perdikouri[5], A. Tsakalidis[6], and Georgia Panagopoulou[7] [1-6]Computer Technology Institute, and[7]National Statistical Services of Greece, IT Division, Greece
33.1 Introduction 549
33.2 Methodology for the Analysis of Patents 550
  33.2.1 Textual Analysis Techniques 552
  33.2.2 Correspondence and Cluster Analysis Techniques 553
  33.2.3 Pre-Processing Steps 558
33.3 System Description 559
  33.3.1 Database Manager Module 560
  33.3.2 Statistical Analysis Module 561
  33.3.3 Results Presentation Module 562
  33.3.4 User Interface 562
33.4 Technology Indicators 563
  33.4.1 Indicators Based on the Technological Sector 564

33.4.2 Indicators in Level of Continents/Countries/Designated States 565
33.4.3 Indicators for Inventors/Assignees 566
33.4.4 Indicators Over Time 567
33.5 Conclusion 568

34 The Semantic Conference Organizer 571

Kevin Heinrich, Michael W. Berry, Jack J. Dongarra, and Sathish Vadhiyar

University of Tennessee, Knoxville, USA

34.1 Background 571
34.2 Latent Semantic Indexing 572
34.3 Software Issues 573
34.4 Creating a Conference 575
34.4.1 A Simple Example 575
34.4.2 Benchmarks 578
34.5 Future Extensions 579

Index 583