Data Mining and Data Visualization

Edited by

C.R. Rao

Center for Multivariate Analysis
Department of Statistics, The Pennsylvania State University
University Park, PA, USA

E.J. Wegman

Center for Computational Statistics
George Mason University
Fairfax, VA, USA

J.L. Solka

Naval Surface Warfare Center, DD
Dahlgren, VA, USA

ELSEVIER
NORTH-HOLLAND

2005

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Table of contents

Preface v

Contributors xiii

Ch. 1. Statistical Data Mining 1
Edward J. Wegman and Jeffrey L. Solka

1. Introduction 1
2. Computational complexity 2
3. The computer science roots of data mining 9
4. Data preparation 14
5. Databases 19
6. Statistical methods for data mining 21
7. Visual data mining 29
8. Streaming data 37
9. A final word 44
Acknowledgements 44
References 44

Ch. 2. From Data Mining to Knowledge Mining 47
Kenneth A. Kaufman and Ryszard S. Michalski

1. Introduction 47
2. Knowledge generation operators 49
3. Strong patterns vs. complete and consistent rules 60
4. Ruleset visualization via concept association graphs 62
5. Integration of knowledge generation operators 66
6. Summary 69
Acknowledgements 70
References 71

Ch. 3. Mining Computer Security Data 77
David J. Marchette

1. Introduction 77
2. Basic TCP/IP 78

3. The threat 84
4. Network monitoring 92
5. TCP sessions 97
6. Signatures versus anomalies 101
7. User profiling 102
8. Program profiling 104
9. Conclusions 107
References 107

Ch. 4. Data Mining of Text Files 109
Angel R. Martinez

1. Introduction and background 109
2. Natural language processing at the word and sentence level 110
3. Approaches beyond the word and sentence level 119
4. Summary 129
References 130

Ch. 5. Text Data Mining with Minimal Spanning Trees 133
Jeffrey L. Solka, Avory C. Bryant and Edward J. Wegman

Introduction 133
1. Approach 133
2. Results 140
3. Conclusions 167
Acknowledgements 168
References 169

Ch. 6. Information Hiding: Steganography and Steganalysis 171
Zoran Duric, Michael Jacobs and Sushil Jajodia

1. Introduction 171
2. Image formats 172
3. Steganography 174
4. Steganalysis 179
5. Relationship of steganography to watermarking 181
6. Literature survey 184
7. Conclusions 186
References 186

Ch. 7. Canonical Variate Analysis and Related Methods for Reduction of Dimensionality and Graphical Representation 189
C. Radhakrishna Rao

1. Introduction 189
2. Canonical coordinates 190
3. Principal component analysis 197

4. Two-way contingency tables (correspondence analysis) 201
5. Discussion 209
References 210

Ch. 8. **Pattern Recognition 213**
David J. Hand

1. Background 213
2. Basics 214
3. Practical classification rules 216
4. Other issues 226
5. Further reading 227
References 227

Ch. 9. **Multidimensional Density Estimation 229**
David W. Scott and Stephan R. Sain

1. Introduction 229
2. Classical density estimators 230
3. Kernel estimators 239
4. Mixture density estimation 248
5. Visualization of densities 252
6. Discussion 258
References 258

Ch. 10. **Multivariate Outlier Detection and Robustness 263**
Mia Hubert, Peter J. Rousseeuw and Stefan Van Aelst

1. Introduction 263
2. Multivariate location and scatter 264
3. Multiple regression 272
4. Multivariate regression 278
5. Classification 282
6. Principal component analysis 283
7. Principal component regression 292
8. Partial Least Squares Regression 296
9. Some other multivariate frameworks 297
10. Availability 297
Acknowledgements 300
References 300

Ch. 11. **Classification and Regression Trees, Bagging, and Boosting 303**
Clifton D. Sutton

1. Introduction 303
2. Using CART to create a classification tree 306
3. Using CART to create a regression tree 315
4. Other issues pertaining to CART 316

5. Bagging 317
6. Boosting 323
References 327

Ch. 12. Fast Algorithms for Classification Using Class Cover Catch Digraphs 331
David J. Marchette, Edward J. Wegman and Carey E. Priebe

1. Introduction 331
2. Class cover catch digraphs 332
3. CCCD for classification 334
4. Cluster catch digraph 338
5. Fast algorithms 340
6. Further enhancements 343
7. Streaming data 344
8. Examples using the fast algorithms 346
9. Sloan Digital Sky Survey 351
10. Text processing 355
11. Discussion 357
Acknowledgements 357
References 358

Ch. 13. On Genetic Algorithms and their Applications 359
Yasmin H. Said

1. Introduction 359
2. History 360
3. Genetic algorithms 361
4. Generalized penalty methods 372
5. Mathematical underpinnings 378
6. Techniques for attaining optimization 381
7. Closing remarks 386
Acknowledgements 387
References 387

Ch. 14. Computational Methods for High-Dimensional Rotations in Data Visualization 391
Andreas Buja, Dianne Cook, Daniel Asimov and Catherine Hurley

1. Introduction 391
2. Tools for constructing plane and frame interpolations: orthonormal frames and planar rotations 397
3. Interpolating paths of planes 403
4. Interpolating paths of frames 406
5. Conclusions 411
References 412

Ch. 15. Some Recent Graphics Templates and Software for Showing Statistical Summaries 415
Daniel B. Carr

1. Introduction 415
2. Background for quantitative graphics design 417

3. The template for linked micromap (LM) plots 420
4. Dynamically conditioned choropleth maps 426
5. Self-similar coordinates plots 431
6. Closing remarks 434
Acknowledgements 435
References 435

Ch. 16. Interactive Statistical Graphics: the Paradigm of Linked Views 437
Adalbert Wilhelm

1. Graphics, statistics and the computer 437
2. The interactive paradigm 444
3. Data displays 446
4. Direct object manipulation 467
5. Selection 469
6. Interaction at the frame level 475
7. Interaction at the type level 476
8. Interactions at the model level 483
9. Interaction at sample population level 486
10. Indirect object manipulation 487
11. Internal linking structures 488
12. Querying 494
13. External linking structure 496
14. Linking frames 498
15. Linking types 499
16. Linking models 499
17. Linking sample populations 503
18. Visualization of linked highlighting 505
19. Visualization of grouping 508
20. Linking interrogation 508
21. Bi-directional linking in the trace plot 509
22. Multivariate graphical data analysis using linked low-dimensional views 509
23. Conditional probabilities 511
24. Detecting outliers 518
25. Clustering and classification 519
26. Geometric structure 520
27. Relationships 522
28. Conclusion 532
Future work 533
References 534

Ch. 17. Data Visualization and Virtual Reality 539
Jim X. Chen

1. Introduction 539
2. Computer graphics 539
3. Graphics software tools 543
4. Data visualization 547
5. Virtual reality 556

6. Some examples of visualization using VR 560
References 561

Colour Figures 565

Subject Index 609

Contents of Previous Volumes 619