Contents

1 Data Mining and Information Systems: Quo Vadis? .................. 1
Robert Stahlbock, Stefan Lessmann, and Sven F. Crone
1.1 Introduction .................................................. 1
1.2 Special Issues in Data Mining .................................. 3
   1.2.1 Confirmatory Data Analysis ............................. 3
   1.2.2 Knowledge Discovery from Supervised Learning ........ 4
   1.2.3 Classification Analysis .................................. 6
   1.2.4 Hybrid Data Mining Procedures .......................... 8
   1.2.5 Web Mining ............................................. 10
   1.2.6 Privacy-Preserving Data Mining ......................... 11
1.3 Conclusion and Outlook ..................................... 12
References ..................................................... 13

Part I Confirmatory Data Analysis

2 Response-Based Segmentation Using Finite Mixture Partial Least
   Squares ....................................................... 19
Christian M. Ringle, Marko Sarstedt, and Erik A. Mooi
2.1 Introduction .................................................. 20
   2.1.1 On the Use of PLS Path Modeling ....................... 20
   2.1.2 Problem Statement ...................................... 22
   2.1.3 Objectives and Organization ............................ 23
2.2 Partial Least Squares Path Modeling .......................... 24
2.3 Finite Mixture Partial Least Squares Segmentation ............ 26
   2.3.1 Foundations ............................................. 26
   2.3.2 Methodology ............................................ 28
   2.3.3 Systematic Application of FIMIX-PLS ................... 31
2.4 Application of FIMIX-PLS .................................... 34
   2.4.1 On Measuring Customer Satisfaction .................... 34
   2.4.2 Data and Measures ..................................... 34
   2.4.3 Data Analysis and Results ............................. 36
2.5 Summary and Conclusion .................................. 44
References .................................................... 45

Part II Knowledge Discovery from Supervised Learning

3 Building Acceptable Classification Models ...................... 53
David Martens and Bart Baesens
3.1 Introduction ............................................. 54
3.2 Comprehensibility of Classification Models ................ 55
  3.2.1 Measuring Comprehensibility ........................ 57
  3.2.2 Obtaining Comprehensible Classification Models ... 58
3.3 Justifiability of Classification Models ...................... 59
  3.3.1 Taxonomy of Constraints ............................ 60
  3.3.2 Monotonicity Constraint ............................ 62
  3.3.3 Measuring Justifiability ............................. 63
  3.3.4 Obtaining Justifiable Classification Models ......... 68
3.4 Conclusion ............................................. 70
References .................................................... 71

4 Mining Interesting Rules Without Support Requirement: A General Universal Existential Upward Closure Property .... 75
Yannick Le Bras, Philippe Lenca, and Stéphane Lallich
4.1 Introduction ............................................. 76
4.2 State of the Art ........................................ 77
4.3 An Algorithmic Property of Confidence ...................... 80
  4.3.1 On UEUC Framework ................................ 80
  4.3.2 The UEUC Property .................................. 80
  4.3.3 An Efficient Pruning Algorithm ..................... 81
  4.3.4 Generalizing the UEUC Property .................... 82
4.4 A Framework for the Study of Measures ....................... 84
  4.4.1 Adapted Functions of Measure ....................... 84
  4.4.2 Expression of a Set of Measures of $D_{d_{conf}}$ .... 87
4.5 Conditions for GUEUC ................................... 90
  4.5.1 A Sufficient Condition .............................. 90
  4.5.2 A Necessary Condition .............................. 91
  4.5.3 Classification of the Measures ..................... 92
4.6 Conclusion ............................................. 94
References .................................................... 95

5 Classification Techniques and Error Control in Logic Mining .... 99
Giovanni Felici, Bruno Simeone, and Vincenzo Spinelli
5.1 Introduction ............................................. 100
5.2 Brief Introduction to Box Clustering ......................... 102
5.3 $BC$-Based Classifier .................................. 104
5.4 Best Choice of a Box System ................................ 108
5.5 Bi-criterion Procedure for $BC$-Based Classifier .......... 111
5.6 Examples ................................................................. 112
5.6.1 The Data Sets ...................................................... 112
5.6.2 Experimental Results with BC .............................. 113
5.6.3 Comparison with Decision Trees ......................... 115
5.7 Conclusions ............................................................ 117
References ................................................................. 117

Part III Classification Analysis

6 An Extended Study of the Discriminant Random Forest .............. 123
   Tracy D. Lemmond, Barry Y. Chen, Andrew O. Hatch,
   and William G. Hanley
   6.1 Introduction ....................................................... 123
   6.2 Random Forests .................................................. 124
   6.3 Discriminant Random Forests ................................. 125
      6.3.1 Linear Discriminant Analysis ............................ 126
      6.3.2 The Discriminant Random Forest Methodology ........ 127
   6.4 DRF and RF: An Empirical Study ............................. 128
      6.4.1 Hidden Signal Detection ................................. 129
      6.4.2 Radiation Detection ...................................... 132
      6.4.3 Significance of Empirical Results ...................... 136
      6.4.4 Small Samples and Early Stopping ...................... 137
      6.4.5 Expected Cost ............................................. 143
   6.5 Conclusions ....................................................... 143
References ................................................................. 145

7 Prediction with the SVM Using Test Point Margins ................. 147
   Süreyya Özöğür-Akyüz, Zakria Hussain, and John Shawe-Taylor
   7.1 Introduction ....................................................... 147
   7.2 Methods ......................................................... 151
   7.3 Data Set Description .......................................... 154
   7.4 Results .......................................................... 154
   7.5 Discussion and Future Work .................................. 155
References ................................................................. 157

8 Effects of Oversampling Versus Cost-Sensitive Learning for
   Bayesian and SVM Classifiers .................................... 159
   Alexander Liu, Cheryl Martin, Brian La Cour, and Joydeep Ghosh
   8.1 Introduction ....................................................... 159
   8.2 Resampling ....................................................... 161
      8.2.1 Random Oversampling .................................... 161
      8.2.2 Generative Oversampling ................................. 161
   8.3 Cost-Sensitive Learning ...................................... 162
   8.4 Related Work .................................................... 163
   8.5 A Theoretical Analysis of Oversampling Versus Cost-Sensitive
         Learning ......................................................... 164
8.5.1 Bayesian Classification .......................................................... 164
8.5.2 Resampling Versus Cost-Sensitive Learning in Bayesian Classifiers .......................................................... 165
8.5.3 Effect of Oversampling on Gaussian Naive Bayes ............. 166
8.5.4 Effects of Oversampling for Multinomial Naive Bayes ........ 168
8.6 Empirical Comparison of Resampling and Cost-Sensitive Learning ........................................................................... 170
8.6.1 Explaining Empirical Differences Between Resampling and Cost-Sensitive Learning ........................................ 170
8.6.2 Naive Bayes Comparisons on Low-Dimensional Gaussian Data .................................................................................. 171
8.6.3 Multinomial Naive Bayes .......................................................... 176
8.6.4 SVMs ...................................................................................... 178
8.6.5 Discussion ............................................................................. 181
8.7 Conclusion ................................................................................ 182
Appendix ...................................................................................... 183
References .................................................................................. 190

9 The Impact of Small Disjuncts on Classifier Learning ............. 193
Gary M. Weiss
9.1 Introduction ............................................................................. 193
9.2 An Example: The Vote Data Set ............................................. 195
9.3 Description of Experiments .................................................... 197
9.4 The Problem with Small Disjuncts ........................................ 198
9.5 The Effect of Pruning on Small Disjuncts ............................... 202
9.6 The Effect of Training Set Size on Small Disjuncts ................. 210
9.7 The Effect of Noise on Small Disjuncts .................................. 213
9.8 The Effect of Class Imbalance on Small Disjuncts ................. 217
9.9 Related Work .......................................................................... 220
9.10 Conclusion ............................................................................ 223
References .................................................................................. 225

Part IV Hybrid Data Mining Procedures

10 Predicting Customer Loyalty Labels in a Large Retail Database: A Case Study in Chile ........................................... 229
Cristián J. Figueroa
10.1 Introduction .......................................................................... 229
10.2 Related Work ....................................................................... 231
10.3 Objectives of the Study ......................................................... 233
10.3.1 Supervised and Unsupervised Learning ............................ 234
10.3.2 Unsupervised Algorithms ............................................... 234
10.3.3 Variables for Segmentation .............................................. 238
10.3.4 Exploratory Data Analysis .............................................. 239
10.3.5 Results of the Segmentation ............................................ 240
10.4 Results of the Classifier ....................................................... 241
## 11 PCA-Based Time Series Similarity Search

Leonidas Karamitopoulos, Georgios Evangelidis, and Dimitris Dervos

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1 Introduction</td>
<td>256</td>
</tr>
<tr>
<td>11.2 Background</td>
<td>258</td>
</tr>
<tr>
<td>11.2.1 Review of PCA</td>
<td>258</td>
</tr>
<tr>
<td>11.2.2 Implications of PCA in Similarity Search</td>
<td>259</td>
</tr>
<tr>
<td>11.2.3 Related Work</td>
<td>261</td>
</tr>
<tr>
<td>11.3 Proposed Approach</td>
<td>263</td>
</tr>
<tr>
<td>11.4 Experimental Methodology</td>
<td>265</td>
</tr>
<tr>
<td>11.4.1 Data Sets</td>
<td>265</td>
</tr>
<tr>
<td>11.4.2 Evaluation Methods</td>
<td>266</td>
</tr>
<tr>
<td>11.4.3 Rival Measures</td>
<td>267</td>
</tr>
<tr>
<td>11.5 Results</td>
<td>268</td>
</tr>
<tr>
<td>11.5.1 1-NN Classification</td>
<td>268</td>
</tr>
<tr>
<td>11.5.2 k-NN Similarity Search</td>
<td>271</td>
</tr>
<tr>
<td>11.5.3 Speeding Up the Calculation of APEdist</td>
<td>272</td>
</tr>
<tr>
<td>11.6 Conclusion</td>
<td>274</td>
</tr>
</tbody>
</table>

## 12 Evolutionary Optimization of Least-Squares Support Vector Machines

Arjan Gijsbers, Giorgio Metta, and Léon Rothkrantz

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.1 Introduction</td>
<td>278</td>
</tr>
<tr>
<td>12.2 Kernel Machines</td>
<td>278</td>
</tr>
<tr>
<td>12.2.1 Least-Squares Support Vector Machines</td>
<td>279</td>
</tr>
<tr>
<td>12.2.2 Kernel Functions</td>
<td>280</td>
</tr>
<tr>
<td>12.3 Evolutionary Computation</td>
<td>281</td>
</tr>
<tr>
<td>12.3.1 Genetic Algorithms</td>
<td>281</td>
</tr>
<tr>
<td>12.3.2 Evolution Strategies</td>
<td>282</td>
</tr>
<tr>
<td>12.3.3 Genetic Programming</td>
<td>283</td>
</tr>
<tr>
<td>12.4 Related Work</td>
<td>283</td>
</tr>
<tr>
<td>12.4.1 Hyperparameter Optimization</td>
<td>284</td>
</tr>
<tr>
<td>12.4.2 Combined Kernel Functions</td>
<td>284</td>
</tr>
<tr>
<td>12.5 Evolutionary Optimization of Kernel Machines</td>
<td>286</td>
</tr>
<tr>
<td>12.5.1 Hyperparameter Optimization</td>
<td>286</td>
</tr>
<tr>
<td>12.5.2 Kernel Construction</td>
<td>287</td>
</tr>
<tr>
<td>12.5.3 Objective Function</td>
<td>288</td>
</tr>
<tr>
<td>12.6 Results</td>
<td>289</td>
</tr>
<tr>
<td>12.6.1 Data Sets</td>
<td>289</td>
</tr>
</tbody>
</table>
12.6.2 Results for Hyperparameter Optimization 290
12.6.3 Results for EvoKM\textsuperscript{GP} 293
12.7 Conclusions and Future Work 294
References 295

13 Genetically Evolved kNN Ensembles 299
Ulf Johansson, Rikard König, and Lars Niklasson
13.1 Introduction 299
13.2 Background and Related Work 301
13.3 Method 302
13.3.1 Data sets 305
13.4 Results 307
13.5 Conclusions 312
References 313

Part V Web-Mining

14 Behaviorally Founded Recommendation Algorithm for Browsing Assistance Systems 317
Peter Géczy, Noriaki Izumi, Shotaro Akaho, and Kōiti Hasida
14.1 Introduction 317
14.1.1 Related Works 318
14.1.2 Our Contribution and Approach 319
14.2 Concept Formalization 319
14.3 System Design 323
14.3.1 A Priori Knowledge of Human–System Interactions 323
14.3.2 Strategic Design Factors 323
14.3.3 Recommendation Algorithm Derivation 325
14.4 Practical Evaluation 327
14.4.1 Intranet Portal 328
14.4.2 System Evaluation 330
14.4.3 Practical Implications and Limitations 331
14.5 Conclusions and Future Work 332
References 333

15 Using Web Text Mining to Predict Future Events: A Test of the Wisdom of Crowds Hypothesis 335
Scott Ryan and Lutz Hamel
15.1 Introduction 335
15.2 Method 337
15.2.1 Hypotheses and Goals 337
15.2.2 General Methodology 339
15.2.3 The 2006 Congressional and Gubernatorial Elections 339
15.2.4 Sporting Events and Reality Television Programs 340
15.2.5 Movie Box Office Receipts and Music Sales 341
15.2.6 Replication 342
15.3 Results and Discussion .................................................. 343
  15.3.1 The 2006 Congressional and Gubernatorial Elections ...... 343
  15.3.2 Sporting Events and Reality Television Programs .......... 345
  15.3.3 Movie and Music Album Results ............................... 347
15.4 Conclusion .............................................................. 348
References ........................................................................... 349

Part VI Privacy-Preserving Data Mining

16 Avoiding Attribute Disclosure with the (Extended) $p$-Sensitive $k$-Anonymity Model ................................................. 353
  Traian Marius Truta and Alina Campan
  16.1 Introduction ............................................................. 353
  16.2 Privacy Models and Algorithms .................................... 354
    16.2.1 The $p$-Sensitive $k$-Anonymity Model and Its Extension . 354
    16.2.2 Algorithms for the $p$-Sensitive $k$-Anonymity Model .... 357
  16.3 Experimental Results .................................................. 360
    16.3.1 Experiments for $p$-Sensitive $k$-Anonymity ............... 360
    16.3.2 Experiments for Extended $p$-Sensitive $k$-Anonymity .... 362
  16.4 New Enhanced Models Based on $p$-Sensitive $k$-Anonymity .... 366
    16.4.1 Constrained $p$-Sensitive $k$-Anonymity .................... 366
    16.4.2 $p$-Sensitive $k$-Anonymity in Social Networks .......... 370
  16.5 Conclusions and Future Work ........................................ 372
References ........................................................................... 372

17 Privacy-Preserving Random Kernel Classification of Checkerboard
  Partitioned Data ................................................................ 375
  Olvi L. Mangasarian and Edward W. Wild
  17.1 Introduction ............................................................. 375
  17.2 Privacy-Preserving Linear Classifier for Checkerboard
    Partitioned Data ............................................................. 379
  17.3 Privacy-Preserving Nonlinear Classifier for Checkerboard
    Partitioned Data ............................................................. 381
  17.4 Computational Results .................................................. 382
  17.5 Conclusion and Outlook ................................................ 384
References ........................................................................... 386
Data Mining
Special Issue in Annals of Information Systems
Stahlbock, R.; Crone, S.F.; Lessmann, S. (Eds.)
2010, XIII, 387 p., Softcover
ISBN: 978-1-4419-1279-4