Contents

Part I Domain Driven KDD Methodology

1 Introduction to Domain Driven Data Mining 3
Longbing Cao
1.1 Why Domain Driven Data Mining 3
1.2 What Is Domain Driven Data Mining 5
1.2.1 Basic Ideas 5
1.2.2 D^3M for Actionable Knowledge Discovery 6
1.3 Open Issues and Prospects 9
1.4 Conclusions 9
References 10

2 Post-processing Data Mining Models for Actionability 11
Qiang Yang
2.1 Introduction 11
2.2 Plan Mining for Class Transformation 12
2.2.1 Overview of Plan Mining 12
2.2.2 Problem Formulation 14
2.2.3 From Association Rules to State Spaces 14
2.2.4 Algorithm for Plan Mining 17
2.2.5 Summary 19
2.3 Extracting Actions from Decision Trees 20
2.3.1 Overview 20
2.3.2 Generating Actions from Decision Trees 22
2.3.3 The Limited Resources Case 23
2.4 Learning Relational Action Models from Frequent Action Sequences 25
2.4.1 Overview 25
2.4.2 ARMS Algorithm: From Association Rules to Actions . . 26
2.4.3 Summary of ARMS 28
2.5 Conclusions and Future Work 29

References 29

3 On Mining Maximal Pattern-Based Clusters 31
Jian Pei, Xiaoling Zhang, Moonjung Cho, Haixun Wang, and Philip S.Yu
3.1 Introduction 32
3.2 Problem Definition and Related Work 34
3.2.1 Pattern-Based Clustering 34
3.2.2 Maximal Pattern-Based Clustering 35
3.2.3 Related Work 35
3.3 Algorithms *MaPle* and *MaPle+* 36
3.3.1 An Overview of *MaPle* 37
3.3.2 Computing and Pruning MDS's 38
3.3.3 Progressively Refining, Depth-first Search of Maximal pClusters 40
3.3.4 *MaPle+*: Further Improvements 44
3.4 Empirical Evaluation 46
3.4.1 The Data Sets 46
3.4.2 Results on Yeast Data Set 47
3.4.3 Results on Synthetic Data Sets 48
3.5 Conclusions 50
References 50

4 Role of Human Intelligence in Domain Driven Data Mining 53
Sumana Sharma and Kweku-Muata Osei-Bryson
4.1 Introduction 53
4.2 DDDM Tasks Requiring Human Intelligence 54
4.2.1 Formulating Business Objectives 54
4.2.2 Setting up Business Success Criteria 55
4.2.3 Translating Business Objective to Data Mining Objectives 56
4.2.4 Setting up of Data Mining Success Criteria 56
4.2.5 Assessing Similarity Between Business Objectives of New and Past Projects 57
4.2.6 Formulating Business, Legal and Financial Requirements 57
4.2.7 Narrowing down Data and Creating Derived Attributes . . 58
4.2.8 Estimating Cost of Data Collection, Implementation and Operating Costs 58
4.2.9 Selection of Modeling Techniques 59
4.2.10 Setting up Model Parameters 59
4.2.11 Assessing Modeling Results 59
4.2.12 Developing a Project Plan 60
4.3 Directions for Future Research 60
4.4 Summary 61
References 61

5 Ontology Mining for Personalized Search 63
Yuefeng Li and Xiaohui Tao
5.1 Introduction 63
5.2 Related Work 64
5.3 Architecture 65
5.4 Background Definitions 66
5.4.1 World Knowledge Ontology 66
5.4.2 Local Instance Repository 67
5.5 Specifying Knowledge in an Ontology 68
5.6 Discovery of Useful Knowledge in LIRs 70
5.7 Experiments 71
5.7.1 Experiment Design 71
5.7.2 Other Experiment Settings 74
5.8 Results and Discussions 75
5.9 Conclusions 77
References 77

Part II Novel KDD Domains & Techniques

6 Data Mining Applications in Social Security 81
Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Hans Bohlscheid, Yuming Ou, and Chengqi Zhang
6.1 Introduction and Background 81
6.2 Case Study I: Discovering Debtor Demographic Patterns with Decision Tree and Association Rules 83
6.2.1 Business Problem and Data 83
6.2.2 Discovering Demographic Patterns of Debtors 83
6.3 Case Study II: Sequential Pattern Mining to Find Activity Sequences of Debt Occurrence 85
6.3.1 Impact-Targeted Activity Sequences 86
6.3.2 Experimental Results 87
6.4 Case Study III: Combining Association Rules from Heterogeneous Data Sources to Discover Repayment Patterns 89
6.4.1 Business Problem and Data 89
6.4.2 Mining Combined Association Rules 89
6.4.3 Experimental Results 90
6.5 Case Study IV: Using Clustering and Analysis of Variance to Verify the Effectiveness of a New Policy 92
6.5.1 Clustering Declarations with Contour and Clustering 92
6.5.2 Analysis of Variance 94
6.6 Conclusions and Discussion 94
References 95

7 Security Data Mining: A Survey Introducing Tamper-Resistance ... 97
Clifton Phua and Mafruz Ashrafi
7.1 Introduction ... 97
7.2 Security Data Mining ... 98
7.2.1 Definitions ... 98
7.2.2 Specific Issues ... 99
7.2.3 General Issues ... 101
7.3 Tamper-Resistance ... 102
7.3.1 Reliable Data ... 102
7.3.2 Anomaly Detection Algorithms ... 104
7.3.3 Privacy and Confidentiality Preserving Results ... 105
7.4 Conclusion ... 108
References ... 108

8 A Domain Driven Mining Algorithm on Gene Sequence Clustering ... 111
Yun Xiong, Ming Chen, and Yangyong Zhu
8.1 Introduction ... 111
8.2 Related Work ... 112
8.3 The Similarity Based on Biological Domain Knowledge ... 114
8.4 Problem Statement ... 114
8.5 A Domain-Driven Gene Sequence Clustering Algorithm ... 117
8.6 Experiments and Performance Study ... 121
8.7 Conclusion and Future Work ... 124
References ... 125

9 Domain Driven Tree Mining of Semi-structured Mental Health Information ... 127
Maja Hadzic, Fedja Hadzic, and Tharam S. Dillon
9.1 Introduction ... 127
9.2 Information Use and Management within Mental Health Domain ... 128
9.3 Tree Mining - General Considerations ... 130
9.4 Basic Tree Mining Concepts ... 131
9.5 Tree Mining of Medical Data ... 135
9.6 Illustration of the Approach ... 139
9.7 Conclusion and Future Work ... 139
References ... 140

10 Text Mining for Real-time Ontology Evolution ... 143
Jackei H.K. Wong, Tharam S. Dillon, Allan K.Y. Wong, and Wilfred W.K. Lin
10.1 Introduction ... 144
10.2 Related Text Mining Work ... 145
10.3 Terminology and Multi-representations ... 145
10.4 Master Aliases Table and OCOE Data Structures ... 149
10.5 Experimental Results ... 152
10.5.1 CAV Construction and Information Ranking ... 153

10.5.2 Real-Time CAV Expansion Supported by Text Mining . . 154
10.6 Conclusion 155
10.7 Acknowledgement 156
References 156

11 Microarray Data Mining: Selecting Trustworthy Genes with Gene Feature Ranking 159
Franco A. Ubaudi, Paul J. Kennedy, Daniel R. Catchpoole, Dachuan Guo, and Simeon J. Simoff
11.1 Introduction 159
11.2 Gene Feature Ranking 161
11.2.1 Use of Attributes and Data Samples in Gene Feature Ranking 162
11.2.2 Gene Feature Ranking: Feature Selection Phase 1 163
11.2.3 Gene Feature Ranking: Feature Selection Phase 2 163
11.3 Application of Gene Feature Ranking to Acute Lymphoblastic Leukemia data 164
11.4 Conclusion 166
References 167

12 Blog Data Mining for Cyber Security Threats 169
Flora S. Tsai and Kap Luk Chan
12.1 Introduction 169
12.2 Review of Related Work 170
12.2.1 Intelligence Analysis 171
12.2.2 Information Extraction from Blogs 171
12.3 Probabilistic Techniques for Blog Data Mining 172
12.3.1 Attributes of Blog Documents 172
12.3.2 Latent Dirichlet Allocation 173
12.3.3 Isometric Feature Mapping (Isomap) 174
12.4 Experiments and Results 175
12.4.1 Data Corpus 175
12.4.2 Results for Blog Topic Analysis 176
12.4.3 Blog Content Visualization 178
12.4.4 Blog Time Visualization 179
12.5 Conclusions 180
References 181

13 Blog Data Mining: The Predictive Power of Sentiments 183
Yang Liu, Xiaohui Yu, Xiangji Huang, and Aijun An
13.1 Introduction 183
13.2 Related Work 185
13.3 Characteristics of Online Discussions 186
13.3.1 Blog Mentions 186
13.3.2 Box Office Data and User Rating 187
13.3.3 Discussion 187

13.4 S-PLSA: A Probabilistic Approach to Sentiment Mining 188
13.4.1 Feature Selection 188
13.4.2 Sentiment PLSA 188
13.5 ARSA: A Sentiment-Aware Model 189
13.5.1 The Autoregressive Model 190
13.5.2 Incorporating Sentiments 191
13.6 Experiments 192
13.6.1 Experiment Settings 192
13.6.2 Parameter Selection 193
13.7 Conclusions and Future Work 194
References 194

14 Web Mining: Extracting Knowledge from the World Wide Web 197
Zhongzhi Shi, Huifang Ma, and Qing He
14.1 Overview of Web Mining Techniques 197
14.2 Web Content Mining 199
14.2.1 Classification: Multi-hierarchy Text Classification 199
14.2.2 Clustering Analysis: Clustering Algorithm Based on Swarm Intelligence and k-Means 200
14.2.3 Semantic Text Analysis: Conceptual Semantic Space 202
14.3 Web Structure Mining: PageRank vs. HITS 203
14.4 Web Event Mining 204
14.4.1 Preprocessing for Web Event Mining 205
14.4.2 Multi-document Summarization: A Way to Demonstrate Event's Cause and Effect 206
14.5 Conclusions and Future Works 206
References 207

15 DAG Mining for Code Compaction 209
T. Werth, M. Wörlein, A. Dreweke, I. Fischer, and M. Philippsen
15.1 Introduction 209
15.2 Related Work 211
15.3 Graph and DAG Mining Basics 211
15.3.1 Graph–based versus Embedding–based Mining 212
15.3.2 Embedded versus Induced Fragments 213
15.3.3 DAG Mining Is *NP*–complete 213
15.4 Algorithmic Details of DAGMA 214
15.4.1 A Canonical Form for DAG enumeration 214
15.4.2 Basic Structure of the DAG Mining Algorithm 215
15.4.3 Expansion Rules 216
15.4.4 Application to Procedural Abstraction 219
15.5 Evaluation 220
15.6 Conclusion and Future Work 222
References 223

16 A Framework for Context-Aware Trajectory Data Mining 225
Vania Bogorny and Monica Wachowicz
16.1 Introduction 225
16.2 Basic Concepts 227
16.3 A Domain-driven Framework for Trajectory Data Mining 229
16.4 Case Study 232
16.4.1 The Selected Mobile Movement-aware Outdoor Game . . 233
16.4.2 Transportation Application 234
16.5 Conclusions and Future Trends 238
References 239

17 Census Data Mining for Land Use Classification 241
E. Roma Neto and D. S. Hamburger
17.1 Content Structure 241
17.2 Key Research Issues 242
17.3 Land Use and Remote Sensing 242
17.4 Census Data and Land Use Distribution 243
17.5 Census Data Warehouse and Spatial Data Mining 243
17.5.1 Concerning about Data Quality 243
17.5.2 Concerning about Domain Driven 244
17.5.3 Applying Machine Learning Tools 246
17.6 Data Integration 247
17.6.1 Area of Study and Data 247
17.6.2 Supported Digital Image Processing 248
17.6.3 Putting All Steps Together 248
17.7 Results and Analysis 249
References 251

18 Visual Data Mining for Developing Competitive Strategies in Higher Education 253
Gürdal Ertek
18.1 Introduction 253
18.2 Square Tiles Visualization 255
18.3 Related Work 256
18.4 Mathematical Model 257
18.5 Framework and Case Study 260
18.5.1 General Insights and Observations 261
18.5.2 Benchmarking 262
18.5.3 High School Relationship Management (HSRM) 263
18.6 Future Work 264
18.7 Conclusions 264
References 265

19 Data Mining For Robust Flight Scheduling 267
Ira Assent, Ralph Krieger, Petra Welter, Jörg Herbers, and Thomas Seidl
19.1 Introduction 267
19.2 Flight Scheduling in the Presence of Delays 268
19.3 Related Work 270
19.4 Classification of Flights 272
19.4.1 Subspaces for Locally Varying Relevance 272
19.4.2 Integrating Subspace Information for Robust Flight Classification 272
19.5 Algorithmic Concept 274
19.5.1 Monotonicity Properties of Relevant Attribute Subspaces 274
19.5.2 Top-down Class Entropy Algorithm: Lossless Pruning Theorem 275
19.5.3 Algorithm: Subspaces, Clusters, Subspace Classification . 276
19.6 Evaluation of Flight Delay Classification in Practice 278
19.7 Conclusion 280
References 280

20 Data Mining for Algorithmic Asset Management 283
Giovanni Montana and Francesco Parrella
20.1 Introduction 283
20.2 Backbone of the Asset Management System 285
20.3 Expert-based Incremental Learning 286
20.4 An Application to the iShare Index Fund 290
References 294

Reviewer List 297

Index 299