

Contents

Chapter 1

Data Mining Techniques in Clustering, Association and Classification	1
Dawn E. Holmes, Jeffrey Tweedale, Lakhmi C. Jain	
1 Introduction	1
1.1 Data.....	1
1.2 Knowledge	2
1.3 Clustering	2
1.4 Association	3
1.5 Classification	3
2 Data Mining	4
2.1 Methods and Algorithms	4
2.2 Applications	4
3 Chapters Included in the Book	5
4 Conclusion	5
References	6

Chapter 2

Clustering Analysis in Large Graphs with Rich Attributes	7
Yang Zhou, Ling Liu	
1 Introduction	8
2 General Issues in Graph Clustering	11
2.1 Graph Partition Techniques	12
2.2 Basic Preparation for Graph Clustering	14
2.3 Graph Clustering with SA-Cluster	15
3 Graph Clustering Based on Structural/Attribute Similarities ...	16
4 The Incremental Algorithm	19
5 Optimization Techniques	21
5.1 The Storage Cost and Optimization	22
5.2 Matrix Computation Optimization	23
5.3 Parallelism	24
6 Conclusion	24
References	25

Chapter 3

Temporal Data Mining: Similarity-Profiled Association Pattern		29
Jin Soung Yoo		
1	Introduction	29
2	Similarity-Profiled Temporal Association Pattern	32
	2.1 Problem Statement	32
	2.2 Interest Measure	34
3	Mining Algorithm	35
	3.1 Envelope of Support Time Sequence	35
	3.2 Lower Bounding Distance	36
	3.3 Monotonicity Property of Upper Lower-Bounding Distance	38
	3.4 SPAMINE Algorithm	39
4	Experimental Evaluation	41
5	Related Work	43
6	Conclusion	45
	References	45

Chapter 4

Bayesian Networks with Imprecise Probabilities: Theory and Application to Classification		49
G. Corani, A. Antonucci, M. Zaffalon		
1	Introduction	49
2	Bayesian Networks	51
3	Credal Sets	52
	3.1 Definition	53
	3.2 Basic Operations with Credal Sets	53
	3.3 Credal Sets from Probability Intervals	55
	3.4 Learning Credal Sets from Data	55
4	Credal Networks	56
	4.1 Credal Network Definition and Strong Extension	56
	4.2 Non-separately Specified Credal Networks	57
5	Computing with Credal Networks	60
	5.1 Credal Networks Updating	60
	5.2 Algorithms for Credal Networks Updating	61
	5.3 Modelling and Updating with Missing Data	62
6	An Application: Assessing Environmental Risk by Credal Networks	64
	6.1 Debris Flows	64
	6.2 The Credal Network	65
7	Credal Classifiers	70
8	Naive Bayes	71
	8.1 Mathematical Derivation	73
9	Naive Credal Classifier (NCC)	74

9.1	Comparing NBC and NCC in Texture Recognition	76
9.2	Treatment of Missing Data	79
10	Metrics for Credal Classifiers	80
11	Tree-Augmented Naive Bayes (TAN)	81
11.1	Variants of the Imprecise Dirichlet Model: Local and Global IDM	82
12	Credal TAN	83
13	Further Credal Classifiers	85
13.1	Lazy NCC (LNCC)	85
13.2	Credal Model Averaging (CMA)	86
14	Open Source Software	88
15	Conclusions	88
	References	88

Chapter 5

Hierarchical Clustering for Finding Symmetries and Other Patterns in Massive, High Dimensional Datasets		95
Fionn Murtagh, Pedro Contreras		
1	Introduction: Hierarchy and Other Symmetries in Data Analysis	95
1.1	About This Article	96
1.2	A Brief Introduction to Hierarchical Clustering	96
1.3	A Brief Introduction to p-Adic Numbers	97
1.4	Brief Discussion of p-Adic and m-Adic Numbers	98
2	Ultrametric Topology	98
2.1	Ultrametric Space for Representing Hierarchy	98
2.2	Some Geometrical Properties of Ultrametric Spaces	100
2.3	Ultrametric Matrices and Their Properties	100
2.4	Clustering through Matrix Row and Column Permutation	101
2.5	Other Miscellaneous Symmetries	103
3	Generalized Ultrametric	103
3.1	Link with Formal Concept Analysis	103
3.2	Applications of Generalized Ultrametrics	104
3.3	Example of Application: Chemical Database Matching	105
4	Hierarchy in a p-Adic Number System	110
4.1	p-Adic Encoding of a Dendrogram	110
4.2	p-Adic Distance on a Dendrogram	113
4.3	Scale-Related Symmetry	114
5	Tree Symmetries through the Wreath Product Group	114
5.1	Wreath Product Group Corresponding to a Hierarchical Clustering	115
5.2	Wreath Product Invariance	115

5.3	Example of Wreath Product Invariance: Haar Wavelet Transform of a Dendrogram	116
6	Remarkable Symmetries in Very High Dimensional Spaces	118
6.1	Application to Very High Frequency Data Analysis: Segmenting a Financial Signal	119
7	Conclusions	126
	References	126

Chapter 6

Randomized Algorithm of Finding the True Number of Clusters Based on Chebychev Polynomial Approximation		131
R. Avros, O. Granichin, D. Shalymov, Z. Volkovich, G.-W. Weber		
1	Introduction	131
2	Clustering	135
2.1	Clustering Methods	135
2.2	Stability Based Methods	138
2.3	Geometrical Cluster Validation Criteria	141
3	Randomized Algorithm	144
4	Examples	147
5	Conclusion	152
	References	152

Chapter 7

Bregman Bubble Clustering: A Robust Framework for Mining Dense Clusters		157
Joydeep Ghosh, Gunjan Gupta		
1	Introduction	157
2	Background	161
2.1	Partitional Clustering Using Bregman Divergences	161
2.2	Density-Based and Mode Seeking Approaches to Clustering	162
2.3	Iterative Relocation Algorithms for Finding a Single Dense Region	164
2.4	Clustering a Subset of Data into Multiple Overlapping Clusters	165
3	Bregman Bubble Clustering	165
3.1	Cost Function	165
3.2	Problem Definition	166
3.3	Bregmanian Balls and Bregman Bubbles	166
3.4	BBC-S: Bregman Bubble Clustering with Fixed Clustering Size	167
3.5	BBC-Q: Dual Formulation of Bregman Bubble Clustering with Fixed Cost	169

4	Soft Bregman Bubble Clustering (Soft BBC)	169
4.1	Bregman Soft Clustering	169
4.2	Motivations for Developing Soft BBC	170
4.3	Generative Model	171
4.4	Soft BBC EM Algorithm	171
4.5	Choosing an Appropriate p_0	173
5	Improving Local Search: Pressurization	174
5.1	Bregman Bubble Pressure	174
5.2	Motivation	175
5.3	BBC-Press	176
5.4	Soft BBC-Press	177
5.5	Pressurization vs. Deterministic Annealing	177
6	A Unified Framework	177
6.1	Unifying Soft Bregman Bubble and Bregman Bubble Clustering	177
6.2	Other Unifications	178
7	Example: Bregman Bubble Clustering with Gaussians	180
7.1	σ^2 Is Fixed	180
7.2	σ^2 Is Optimized	181
7.3	“Flavors” of BBC for Gaussians	182
7.4	Mixture-6: An Alternative to BBC Using a Gaussian Background	182
8	Extending BBOCC & BBC to Pearson Distance and Cosine Similarity	183
8.1	Pearson Correlation and Pearson Distance	183
8.2	Extension to Cosine Similarity	185
8.3	Pearson Distance vs. (1-Cosine Similarity) vs. Other Bregman Divergences – Which One to Use Where?	185
9	Seeding BBC and Determining k Using Density Gradient Enumeration (DGRADE)	185
9.1	Background	186
9.2	DGRADE Algorithm	186
9.3	Selecting s_{one} : The Smoothing Parameter for DGRADE	188
10	Experiments	190
10.1	Overview	190
10.2	Datasets	190
10.3	Evaluation Methodology	192
10.4	Results for BBC with Pressurization	194
10.5	Results on BBC with DGRADE	198
11	Concluding Remarks	202
	References	204

Chapter 8

DepMiner: A Method and a System for the Extraction of Significant Dependencies	209
Rosa Meo, Leonardo D'Ambrosi	
1 Introduction	209
2 Related Work	211
3 Estimation of the Referential Probability	213
4 Setting a Threshold for Δ	213
5 Embedding Δ_n in Algorithms	215
6 Determination of the Itemsets Minimum Support Threshold ...	216
7 System Description	218
8 Experimental Evaluation	220
9 Conclusions	221
References	221

Chapter 9

Integration of Dataset Scans in Processing Sets of Frequent Itemset Queries	223
Marek Wojciechowski, Maciej Zakrzewicz, Pawel Boinski	
1 Introduction	223
2 Frequent Itemset Mining and Apriori Algorithm	225
2.1 Basic Definitions and Problem Statement	225
2.2 Algorithm Apriori	226
3 Frequent Itemset Queries – State of the Art	227
3.1 Frequent Itemset Queries	227
3.2 Constraint-Based Frequent Itemset Mining	229
3.3 Reusing Results of Previous Frequent Itemset Queries ...	230
4 Optimizing Sets of Frequent Itemset Queries	231
4.1 Basic Definitions	232
4.2 Problem Formulation	233
4.3 Related Work on Multi-query Optimization	234
5 Common Counting	234
5.1 Basic Algorithm	234
5.2 Motivation for Query Set Partitioning	237
5.3 Key Issues Regarding Query Set Partitioning	237
6 Frequent Itemset Query Set Partitioning by Hypergraph Partitioning	238
6.1 Data Sharing Hypergraph	239
6.2 Hypergraph Partitioning Problem Formulation	239
6.3 Computation Complexity of the Problem	241
6.4 Related Work on Hypergraph Partitioning	241
7 Query Set Partitioning Algorithms	241
7.1 CCRRecursive	242
7.2 CCFull	243
7.3 CCCoarsening	246

7.4	CCAglomerative	247
7.5	CCAglomerativeNoise.....	248
7.6	CCGreedy.....	249
7.7	CCSemiGreedy	250
8	Experimental Results	251
8.1	Comparison of Basic Dedicated Algorithms	252
8.2	Comparison of Greedy Approaches with the Best Dedicated Algorithms	257
9	Review of Other Methods of Processing Sets of Frequent Itemset Queries	260
10	Conclusions.....	261
	References	262

Chapter 10

Text Clustering with Named Entities: A Model, Experimentation and Realization		267
Tru H. Cao, Thao M. Tang, Cuong K. Chau		
1	Introduction	267
2	An Entity-Keyword Multi-Vector Space Model	269
3	Measures of Clustering Quality	271
4	Hard Clustering Experiments	273
5	Fuzzy Clustering Experiments.....	277
6	Text Clustering in VN-KIM Search	282
7	Conclusion	285
	References	286

Chapter 11

Regional Association Rule Mining and Scoping from Spatial Data		289
Wei Ding, Christoph F. Eick		
1	Introduction	289
2	Related Work	291
2.1	Hot-Spot Discovery	291
2.2	Spatial Association Rule Mining.....	292
3	The Framework for Regional Association Rule Mining and Scoping	293
3.1	Region Discovery	293
3.2	Problem Formulation	294
3.3	Measure of Interestingness.....	295
4	Algorithms	298
4.1	Region Discovery	298
4.2	Generation of Regional Association Rules	301

- 5 Arsenic Regional Association Rule Mining and Scoping in the
 Texas Water Supply 302
 - 5.1 Data Collection and Data Preprocessing..... 302
 - 5.2 Region Discovery for Arsenic Hot/Cold Spots 304
 - 5.3 Regional Association Rule Mining 305
 - 5.4 Region Discovery for Regional Association Rule
 Scoping 307
- 6 Summary..... 310
- References 311

Chapter 12

- Learning from Imbalanced Data: Evaluation Matters..... 315**
Troy Raeder, George Forman, Nitesh V. Chawla
 - 1 Motivation and Significance..... 315
 - 2 Prior Work and Limitations 317
 - 3 Experiments 318
 - 3.1 Datasets 321
 - 3.2 Empirical Analysis 321
 - 4 Discussion and Recommendations 325
 - 4.1 Comparisons of Classifiers 325
 - 4.2 Towards Parts-Per-Million 328
 - 4.3 Recommendations 329
 - 5 Summary..... 329
 - References 330
- Author Index..... 333**