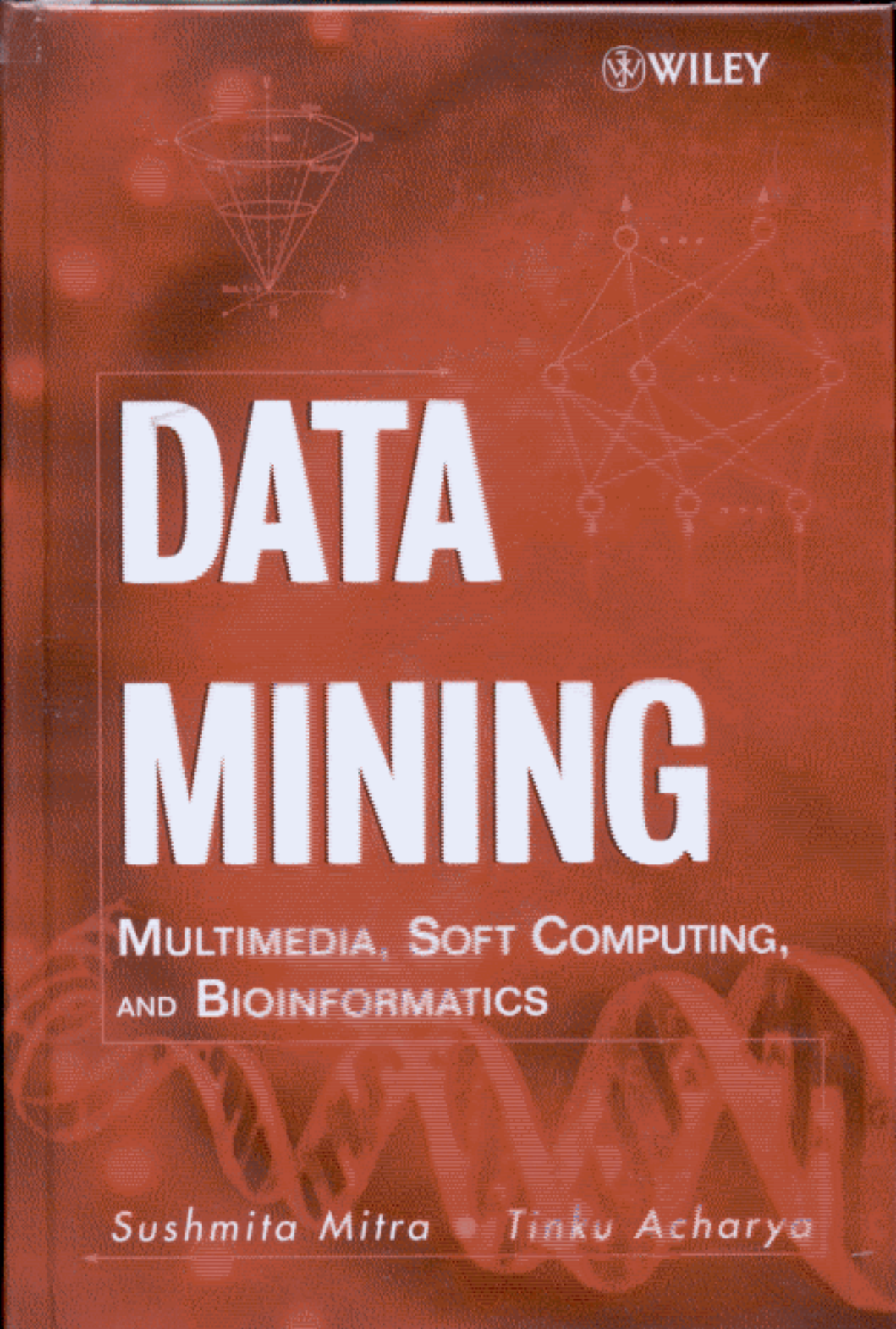


 WILEY

DATA MINING

MULTIMEDIA, SOFT COMPUTING,
AND BIOINFORMATICS



Sushmita Mitra • Tinku Acharya

Contents

Preface	xv
1 Introduction to Data Mining	1
1.1 Introduction	1
1.2 Knowledge Discovery and Data Mining	5
1.3 Data Compression	10
1.4 Information Retrieval	12
1.5 Text Mining	14
1.6 Web Mining	15
1.7 Image Mining	16
1.8 Classification	18
1.9 Clustering	19
1.10 Rule Mining	20
1.11 String Matching	21
1.12 Bioinformatics	23
1.13 Data Warehousing	24
1.14 Applications and Challenges	25
1.15 Conclusions and Discussion	28
References	30

2	Soft Computing	35
2.1	Introduction	35
2.2	What is Soft Computing?	37
2.2.1	Relevance	37
2.2.2	Fuzzy sets	39
2.2.3	Neural networks	44
2.2.4	Neuro-fuzzy computing	53
2.2.5	Genetic algorithms	55
2.2.6	Rough sets	59
2.2.7	Wavelets	61
2.3	Role of Fuzzy Sets in Data Mining	62
2.3.1	Clustering	63
2.3.2	Granular computing	63
2.3.3	Association rules	64
2.3.4	Functional dependencies	65
2.3.5	Data summarization	65
2.3.6	Image mining	66
2.4	Role of Neural Networks in Data Mining	67
2.4.1	Rule extraction	67
2.4.2	Rule evaluation	67
2.4.3	Clustering and self-organization	69
2.4.4	Regression	69
2.4.5	Information retrieval	69
2.5	Role of Genetic Algorithms in Data Mining	70
2.5.1	Regression	71
2.5.2	Association rules	71
2.6	Role of Rough Sets in Data Mining	72
2.7	Role of Wavelets in Data Mining	73
2.8	Role of Hybridizations in Data Mining	74
2.9	Conclusions and Discussion	77
	References	78
3	Multimedia Data Compression	89
3.1	Introduction	89
3.2	Information Theory Concepts	91
3.2.1	Discrete memoryless model and entropy	91
3.2.2	Noiseless Source Coding Theorem	92
3.3	Classification of Compression Algorithms	94

3.4	A Data Compression Model	95
3.5	Measures of Compression Performance	96
3.5.1	Compression ratio and bits per sample	97
3.5.2	Quality metric	97
3.5.3	Coding complexity	99
3.6	Source Coding Algorithms	99
3.6.1	Run-length coding	99
3.6.2	Huffman coding	100
3.7	Principal Component Analysis for Data Compression	103
3.8	Principles of Still Image Compression	105
3.8.1	Predictive coding	105
3.8.2	Transform coding	107
3.8.3	Wavelet coding	109
3.9	Image Compression Standard: JPEG	112
3.10	The JPEG Lossless Coding Algorithm	113
3.11	Baseline JPEG Compression	116
3.11.1	Color space conversion	116
3.11.2	Source image data arrangement	118
3.11.3	The baseline compression algorithm	119
3.11.4	Decompression process in baseline JPEG	126
3.11.5	JPEG2000: Next generation still picture coding standard	129
3.12	Text Compression	131
3.12.1	The LZ77 algorithm	132
3.12.2	The LZ78 algorithm	133
3.12.3	The LZW algorithm	136
3.12.4	Other applications of Lempel–Ziv coding	139
3.13	Conclusions and Discussion	140
	References	140
4	String Matching	143
4.1	Introduction	143
4.1.1	Some definitions and preliminaries	144
4.1.2	String matching problem	146
4.1.3	Brute force string matching	148
4.2	Linear-Order String Matching Algorithms	150
4.2.1	String matching with finite automata	150
4.2.2	Knuth–Morris–Pratt algorithm	152
4.2.3	Boyer–Moore algorithm	158

4.2.4	Boyer–Moore–Horspool algorithm	161
4.2.5	Karp–Rabin algorithm	165
4.3	String Matching in Bioinformatics	169
4.4	Approximate String Matching	171
4.4.1	Basic definitions	172
4.4.2	Wagner–Fischer algorithm for computation of string distance	173
4.4.3	Text search with k -differences	176
4.5	Compressed Pattern Matching	177
4.6	Conclusions and Discussion	179
	References	179
5	Classification in Data Mining	181
5.1	Introduction	181
5.2	Decision Tree Classifiers	184
5.2.1	ID3	187
5.2.2	IBM IntelligentMiner	189
5.2.3	Serial PaRallelizable INtroduction of decision Trees (SPRINT)	189
5.2.4	RainForest	192
5.2.5	Overfitting	192
5.2.6	PrUning and BuiLding Integrated in Classification (PUBLIC)	194
5.2.7	Extracting classification rules from trees	194
5.2.8	Fusion with neural networks	195
5.3	Bayesian Classifiers	196
5.3.1	Bayesian rule for minimum risk	196
5.3.2	Naive Bayesian classifier	196
5.3.3	Bayesian belief network	198
5.4	Instance-Based Learners	199
5.4.1	Minimum distance classifiers	199
5.4.2	k -nearest neighbor (k -NN) classifier	201
5.4.3	Locally weighted regression	201
5.4.4	Radial basis functions (RBFs)	202
5.4.5	Case-based reasoning (CBR)	203
5.4.6	Granular computing and CBR	203
5.5	Support Vector Machines	204
5.6	Fuzzy Decision Trees	205
5.6.1	Classification	207

5.6.2	Rule generation and evaluation	212
5.6.3	Mapping of rules to fuzzy neural network	214
5.6.4	Results	216
5.7	Conclusions and Discussion	220
	References	221
6	Clustering in Data Mining	227
6.1	Introduction	227
6.2	Distance Measures and Symbolic Objects	229
6.2.1	Numeric objects	229
6.2.2	Binary objects	229
6.2.3	Categorical objects	231
6.2.4	Symbolic objects	231
6.3	Clustering Categories	232
6.3.1	Partitional clustering	232
6.3.2	Hierarchical clustering	235
6.3.3	Leader clustering	237
6.4	Scalable Clustering Algorithms	237
6.4	Clustering large applications	238
6.4	Density-based clustering	239
6.4.3	Hierarchical clustering	241
6.4.4	Grid-based methods	243
6.4.5	Other variants	244
6.5	Soft Computing-Based Approaches	244
6.5.1	Fuzzy sets	244
6.5.2	Neural networks	246
6.5.3	Wavelets	248
6.5.4	Rough sets	249
6.5.5	Evolutionary algorithms	250
6.6	Clustering with Categorical Attributes	251
6.6.1	Sieving Through Iterated Relational Reinforcements (STIRR)	252
6.6.2	Robust Hierarchical Clustering with Links (ROCK)	252
6.6.3	c-modes algorithm	253
6.7	Hierarchical Symbolic Clustering	255
6.7.1	Conceptual clustering	255
6.7.2	Agglomerative symbolic clustering	256
6.7.3	Cluster validity indices	257
6.7.4	Results	259

6.8	Conclusions and Discussion	261
	References	262
7	Association Rules	267
7.1	Introduction	267
7.2	Candidate Generation and Test Methods	269
7.2.1	A priori algorithm	269
7.2.2	Partition algorithm	272
7.2.3	Some extensions	272
7.3	Depth-First Search Methods	273
7.4	Interesting Rules	275
7.5	Multilevel Rules	276
7.6	Online Generation of Rules	277
7.7	Generalized Rules	278
7.8	Scalable Mining of Rules	280
7.9	Other Variants	281
7.9.1	Quantitative association rules	281
7.9.2	Temporal association rules	281
7.9.3	Correlation rules	282
7.9.4	Localized associations	282
7.9.5	Optimized association rules	283
7.10	Fuzzy Association Rules	283
7.11	Conclusions and Discussion	288
	References	289
8	Rule Mining with Soft Computing	293
8.1	Introduction	293
8.2	Connectionist Rule Generation	294
8.2.1	Neural models	295
8.2.2	Neuro-fuzzy models	296
8.2.3	Using knowledge-based networks	297
8.3	Modular Hybridization	302
8.3.1	Rough fuzzy MLP	302
8.3.2	Modular knowledge-based network	305
8.3.3	Evolutionary design	308
8.3.4	Rule extraction	310
8.3.5	Results	311
8.4	Conclusions and Discussion	315

References	315
9 Multimedia Data Mining	319
9.1 Introduction	319
9.2 Text Mining	320
9.2.1 Keyword-based search and mining	321
9.2.2 Text analysis and retrieval	322
9.2.3 Mathematical modeling of documents	323
9.2.4 Similarity-based matching for documents and queries	325
9.2.5 Latent semantic analysis	326
9.2.6 Soft computing approaches	328
9.3 Image Mining	329
9.3.1 Content-Based Image Retrieval	330
9.3.2 Color features	332
9.3.3 Texture features	337
9.3.4 Shape features	338
9.3.5 Topology	340
9.3.6 Multidimensional indexing	342
9.3.7 Results of a simple CBIR system	343
9.4 Video Mining	345
9.4.1 MPEG-7: Multimedia content description interface	347
9.4.2 Content-based video retrieval system	348
9.5 Web Mining	350
9.5.1 Search engines	351
9.5.2 Soft computing approaches	353
9.6 Conclusions and Discussion	357
References	357
10 Bioinformatics: An Application	365
10.1 Introduction	365
10.2 Preliminaries from Biology	367
10.2.1 Deoxyribonucleic acid	367
10.2.2 Amino acids	368
10.2.3 Proteins	369
10.2.4 Microarray and gene expression	371
10.3 Information Science Aspects	371
10.3.1 Protein folding	372
10.3.2 Protein structure modeling	373

10.3.3 Genomic sequence analysis	374
10.3.4 Homology search	374
10.4 Clustering of Microarray Data	378
10.4.1 First-generation algorithms	379
10.4.2 Second-generation algorithms	380
10.5 Association Rules	381
10.6 Role of Soft Computing	381
10.6.1 Predicting protein secondary structure	382
10.6.2 Predicting protein tertiary structure	382
10.6.3 Determining binding sites	385
10.6.4 Classifying gene expression data	385
10.7 Conclusions and Discussion	386
References	387
Index	392
About the Authors	399