

Contents

Preface	xi
Part One Introduction	1
1 Introduction	3
1.1 Basic Data Mining Tasks	5
1.1.1 Classification	5
1.1.2 Regression	6
1.1.3 Time Series Analysis	6
1.1.4 Prediction	7
1.1.5 Clustering	7
1.1.6 Summarization	8
1.1.7 Association Rules	8
1.1.8 Sequence Discovery	9
1.2 Data Mining Versus Knowledge Discovery in Databases	9
1.2.1 The Development of Data Mining	12
1.3 Data Mining Issues	14
1.4 Data Mining Metrics	15
1.5 Social Implications of Data Mining	16
1.6 Data Mining from a Database Perspective	16
1.7 The Future	17
1.8 Exercises	19
1.9 Bibliographic Notes	19
2 Related Concepts	21
2.1 Database/OLTP Systems	21
2.2 Fuzzy Sets and Fuzzy Logic	23
2.3 Information Retrieval	26
2.4 Decision Support Systems	28
2.5 Dimensional Modeling	29
2.5.1 Multidimensional Schemas	31
2.5.2 Indexing	34
2.6 Data Warehousing	35
2.7 OLAP	39
2.8 Web Search Engines	41
2.9 Statistics	41
2.10 Machine Learning	42
2.11 Pattern Matching	44
2.12 Summary	44
2.13 Exercises	45
2.14 Bibliographic Notes	45

3	Data Mining Techniques	46
3.1	Introduction	46
3.2	A Statistical Perspective on Data Mining	47
3.2.1	Point Estimation	47
3.2.2	Models Based on Summarization	51
3.2.3	Bayes Theorem	52
3.2.4	Hypothesis Testing	54
3.2.5	Regression and Correlation	55
3.3	Similarity Measures	57
3.4	Decision Trees	58
3.5	Neural Networks	61
3.5.1	Activation Functions	64
3.6	Genetic Algorithms	67
3.7	Exercises	70
3.8	Bibliographic Notes	71

Part Two Core Topics 73

4	Classification	75
4.1	Introduction	75
4.1.1	Issues in Classification	77
4.2	Statistical-Based Algorithms	80
4.2.1	Regression	80
4.2.2	Bayesian Classification	86
4.3	Distance-Based Algorithms	89
4.3.1	Simple Approach	89
4.3.2	K Nearest Neighbors	90
4.4	Decision Tree–Based Algorithms	92
4.4.1	ID3	97
4.4.2	C4.5 and C5.0	100
4.4.3	CART	102
4.4.4	Scalable DT Techniques	103
4.5	Neural Network–Based Algorithms	103
4.5.1	Propagation	105
4.5.2	NN Supervised Learning	106
4.5.3	Radial Basis Function Networks	112
4.5.4	Perceptrons	112
4.6	Rule-Based Algorithms	114
4.6.1	Generating Rules from a DT	114
4.6.2	Generating Rules from a Neural Net	115
4.6.3	Generating Rules Without a DT or NN	116
4.7	Combining Techniques	119
4.8	Summary	121
4.9	Exercises	121
4.10	Bibliographic Notes	122

5	Clustering	125
5.1	Introduction	125
5.2	Similarity and Distance Measures	129
5.3	Outliers	130
5.4	Hierarchical Algorithms	131
5.4.1	Agglomerative Algorithms	132
5.4.2	Divisive Clustering	138
5.5	Partitional Algorithms	138
5.5.1	Minimum Spanning Tree	138
5.5.2	Squared Error Clustering Algorithm	139
5.5.3	<i>K</i> -Means Clustering	140
5.5.4	Nearest Neighbor Algorithm	142
5.5.5	PAM Algorithm	142
5.5.6	Bond Energy Algorithm	145
5.5.7	Clustering with Genetic Algorithms	146
5.5.8	Clustering with Neural Networks	147
5.6	Clustering Large Databases	149
5.6.1	BIRCH	150
5.6.2	DBSCAN	152
5.6.3	CURE Algorithm	154
5.7	Clustering with Categorical Attributes	157
5.8	Comparison	159
5.9	Exercises	161
5.10	Bibliographic Notes	161
6	Association Rules	164
6.1	Introduction	164
6.2	Large Itemsets	167
6.3	Basic Algorithms	169
6.3.1	Apriori Algorithm	169
6.3.2	Sampling Algorithm	173
6.3.3	Partitioning	177
6.4	Parallel and Distributed Algorithms	178
6.4.1	Data Parallelism	179
6.4.2	Task Parallelism	180
6.5	Comparing Approaches	181
6.6	Incremental Rules	184
6.7	Advanced Association Rule Techniques	184
6.7.1	Generalized Association Rules	184
6.7.2	Multiple-Level Association Rules	185
6.7.3	Quantitative Association Rules	185
6.7.4	Using Multiple Minimum Supports	186
6.7.5	Correlation Rules	187
6.8	Measuring the Quality of Rules	188
6.9	Exercises	190
6.10	Bibliographic Notes	191

Part Three Advanced Topics 193

7	Web Mining	195
7.1	Introduction	195
7.2	Web Content Mining	197
7.2.1	Crawlers	198
7.2.2	Harvest System	201
7.2.3	Virtual Web View	201
7.2.4	Personalization	202
7.3	Web Structure Mining	204
7.3.1	PageRank	205
7.3.2	Clever	205
7.4	Web Usage Mining	206
7.4.1	Preprocessing	208
7.4.2	Data Structures	209
7.4.3	Pattern Discovery	211
7.4.4	Pattern Analysis	218
7.5	Exercises	218
7.6	Bibliographic Notes	219
8	Spatial Mining	221
8.1	Introduction	221
8.2	Spatial Data Overview	222
8.2.1	Spatial Queries	222
8.2.2	Spatial Data Structures	223
8.2.3	Thematic Maps	226
8.2.4	Image Databases	226
8.3	Spatial Data Mining Primitives	227
8.4	Generalization and Specialization	228
8.4.1	Progressive Refinement	228
8.4.2	Generalization	229
8.4.3	Nearest Neighbor	231
8.4.4	STING	231
8.5	Spatial Rules	233
8.5.1	Spatial Association Rules	234
8.6	Spatial Classification Algorithm	236
8.6.1	ID3 Extension	236
8.6.2	Spatial Decision Tree	236
8.7	Spatial Clustering Algorithms	237
8.7.1	CLARANS Extensions	238
8.7.2	SD(CLARANS)	239
8.7.3	DBCLASD	240
8.7.4	BANG	241
8.7.5	WaveCluster	241
8.7.6	Approximation	241
8.8	Exercises	243
8.9	Bibliographic Notes	243

9	Temporal Mining	245
9.1	Introduction	245
9.2	Modeling Temporal Events	248
9.3	Time Series	252
9.3.1	Time Series Analysis	252
9.3.2	Trend Analysis	253
9.3.3	Transformation	255
9.3.4	Similarity	255
9.3.5	Prediction	256
9.4	Pattern Detection	257
9.4.1	String Matching	257
9.5	Sequences	260
9.5.1	AprioriAll	262
9.5.2	SPADE	262
9.5.3	Generalization	264
9.5.4	Feature Extraction	266
9.6	Temporal Association Rules	266
9.6.1	Intertransaction Rules	267
9.6.2	Episode Rules	267
9.6.3	Trend Dependencies	268
9.6.4	Sequence Association Rules	270
9.6.5	Calendric Association Rules	271
9.7	Exercises	272
9.8	Bibliographic Notes	272

APPENDICES

A	Data Mining Products	274
A.1	Bibliographic Notes	289
B	Bibliography	290
	Index	305
	About the Author	315