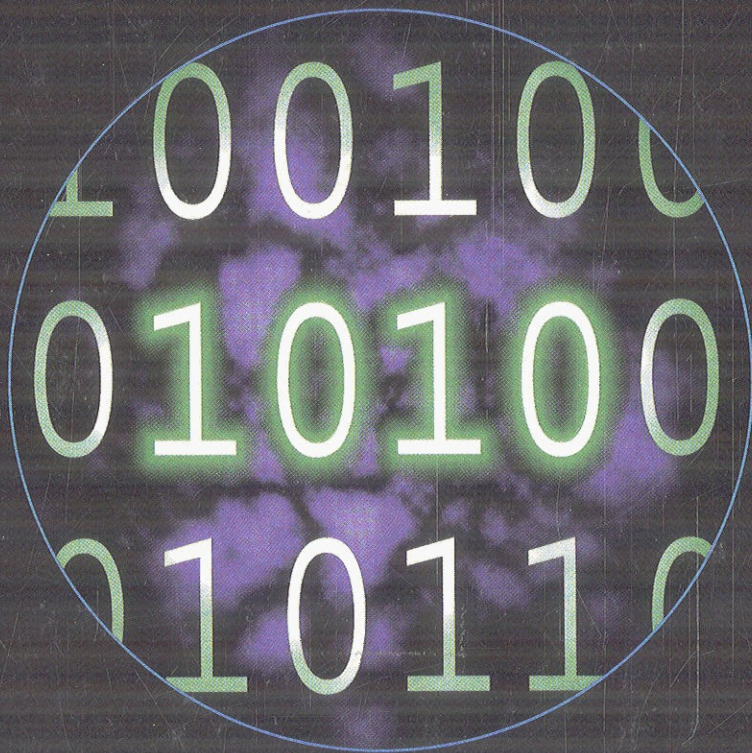


DATA MINING

CONCEPTS, MODELS,
METHODS, AND ALGORITHMS



MEHMED KANTARDZIC

Contents

PREFACE	xi
1 Data Mining Concepts	1
1.1 Introduction	1
1.2 Data-mining roots	4
1.3 Data-mining process	5
1.4 Large data sets	9
1.5 Data warehouses	13
1.6 Organization of this book	16
1.7 Review questions and problems	17
1.8 <i>References for further study</i>	18
2 Preparing the Data	19
2.1 Representation of raw data	19
2.2 Characteristics of raw data	23
2.3 Transformation of raw data	24
2.4 Missing data	27
2.5 Time-dependent data	28
2.6 Outlier analysis	33
2.7 Review questions and problems	36
2.8 <i>References for further study</i>	38
3 Data Reduction	39
3.1 Dimensions of large data sets	39
3.2 Features reduction	41
3.3 Entropy measure for ranking features	46
3.4 Principal component analysis	48
3.5 Values reduction	51
3.6 Feature discretization: ChiMerge technique	54
3.7 Cases reduction	58
3.8 Review questions and problems	61
3.9 <i>References for further study</i>	62
4 Learning from Data	65
4.1 Learning machine	66
4.2 Statistical learning theory	71
4.3 Types of learning methods	76
4.4 Common learning tasks	78

4.5	Model estimation	83
4.6	Review questions and problems	87
4.7	References for further study	88
5	Statistical Methods	91
5.1	Statistical inference	91
5.2	Assessing differences in data sets	93
5.3	Bayesian inference	95
5.4	Predictive regression	98
5.5	Analysis of variance	104
5.6	Logistic regression	106
5.7	Log-linear models	107
5.8	Linear discriminant analysis	111
5.9	Review questions and problems	113
5.10	References for further study	114
6	Cluster Analysis	117
6.1	Clustering concepts	117
6.2	Similarity measures	120
6.3	Agglomerative hierarchical clustering	125
6.4	Partitional clustering	129
6.5	Incremental clustering	132
6.6	Review questions and problems	136
6.7	References for further study	137
7	Decision Trees and Decision Rules	139
7.1	Decision trees	140
7.2	C4.5 Algorithm: generating a decision tree	142
7.3	Unknown attribute values	149
7.4	Pruning decision tree	153
7.5	C4.5 Algorithm: generating decision rules	154
7.6	Limitations of decision trees and decision rules	157
7.7	Associative-classification method	159
7.8	Review questions and problems	161
7.9	References for further study	164
8	Association Rules	165
8.1	Market-Basket Analysis	165
8.2	Algorithm <i>Apriori</i>	167
8.3	From frequent itemsets to association rules	169
8.4	Improving the efficiency of the <i>Apriori</i> algorithm	170
8.5	Frequent pattern-growth method	172
8.6	Multidimensional association-rules mining	174
8.7	Web mining	176
8.8	IIITS and LOGSOM algorithms	178
8.9	Mining path-traversal patterns	184

8.10	Text mining	187
8.11	Review questions and problems	191
8.12	References for further study	193
9	Artificial Neural Networks	195
9.1	Model of an artificial neuron	197
9.2	Architectures of artificial neural networks	200
9.3	Learning process	201
9.4	Learning tasks	205
9.5	Multilayer perceptrons	208
9.6	Competitive networks and competitive learning	214
9.7	Review questions and problems	218
9.8	References for further study	220
10	Genetic Algorithms	221
10.1	Fundamentals of genetic algorithms	222
10.2	Optimization using genetic algorithms	224
10.3	A simple illustration of a genetic algorithm	229
10.4	Schemata	234
10.5	Traveling salesman problem	237
10.6	Machine learning using genetic algorithms	239
10.7	Review questions and problems	243
10.8	References for further study	245
11	Fuzzy Sets and Fuzzy Logic	247
11.1	Fuzzy sets	247
11.2	Fuzzy set operations	253
11.3	Extension principle and fuzzy relations	257
11.4	Fuzzy logic and fuzzy inference systems	261
11.5	Multifactorial evaluation	266
11.6	Extracting fuzzy models from data	268
11.7	Review questions and problems	272
11.8	References for further study	274
12	Visualization Methods	277
12.1	Perception and visualization	277
12.2	Scientific visualization and information visualization	278
12.3	Parallel coordinates	284
12.4	Radial visualization	286
12.5	Kohonen self-organized maps	289
12.6	Visualization systems for data mining	290
12.7	Review questions and problems	294
12.8	References for further study	295

13	References	297
	APPENDIX A: Data-Mining Tools	309
A1	Commercially and publicly available tools	309
A2	Web site links	317
	APPENDIX B: Data-Mining Applications	327
B1	Data mining for financial data analysis	327
B2	Data mining for the telecommunications industry	329
B3	Data mining for the retail industry	331
B4	Data mining in healthcare and biomedical research	333
B5	Data mining in science and engineering	335
B6	Pitfalls of data mining	337
	INDEX	339
	ABOUT THE AUTHOR	345