
Contents

Foreword	
<i>Arkadij Zakrevskij</i>	vii
Preface	xi
Acknowledgments	xvii
List of Figures	xxvii
List of Tables	xxxii

Part I Algorithmic Issues

1 Introduction	3
1.1 What Is Data Mining and Knowledge Discovery?	3
1.2 Some Potential Application Areas for Data Mining and Knowledge Discovery	4
1.2.1 Applications in Engineering	5
1.2.2 Applications in Medical Sciences	5
1.2.3 Applications in the Basic Sciences	6
1.2.4 Applications in Business	6
1.2.5 Applications in the Political and Social Sciences	7
1.3 The Data Mining and Knowledge Discovery Process	7
1.3.1 Problem Definition	7
1.3.2 Collecting the Data	9
1.3.3 Data Preprocessing	10
1.3.4 Application of the Main Data Mining and Knowledge Discovery Algorithms	11
1.3.5 Interpretation of the Results of the Data Mining and Knowledge Discovery Process	12

1.4	Four Key Research Challenges in Data Mining and Knowledge Discovery	12
1.4.1	Collecting Observations about the Behavior of the System	13
1.4.2	Identifying Patterns from Collections of Data	14
1.4.3	Which Data to Consider for Evaluation Next?	17
1.4.4	Do Patterns Always Exist in Data?	19
1.5	Concluding Remarks	20
2	Inferring a Boolean Function from Positive and Negative Examples	21
2.1	An Introduction	21
2.2	Some Background Information	22
2.3	Data Binarization	26
2.4	Definitions and Terminology	29
2.5	Generating Clauses from Negative Examples Only	32
2.6	Clause Inference as a Satisfiability Problem	33
2.7	An SAT Approach for Inferring CNF Clauses	34
2.8	The One Clause At a Time (OCAT) Concept	35
2.9	A Branch-and-Bound Approach for Inferring a Single Clause	38
2.10	A Heuristic for Problem Preprocessing	45
2.11	Some Computational Results	47
2.12	Concluding Remarks	50
	Appendix	52
3	A Revised Branch-and-Bound Approach for Inferring a Boolean Function from Examples	57
3.1	Some Background Information	57
3.2	The Revised Branch-and-Bound Algorithm	57
3.2.1	Generating a Single CNF Clause	58
3.2.2	Generating a Single DNF Clause	62
3.2.3	Some Computational Results	64
3.3	Concluding Remarks	69
4	Some Fast Heuristics for Inferring a Boolean Function from Examples	73
4.1	Some Background Information	73
4.2	A Fast Heuristic for Inferring a Boolean Function from Complete Data	75
4.3	A Fast Heuristic for Inferring a Boolean Function from Incomplete Data	80
4.4	Some Computational Results	84
4.4.1	Results for the RA1 Algorithm on the Wisconsin Cancer Data	86
4.4.2	Results for the RA2 Heuristic on the Wisconsin Cancer Data with Some Missing Values	91

4.4.3	Comparison of the RA1 Algorithm and the B&B Method Using Large Random Data Sets	92
4.5	Concluding Remarks	98
5	An Approach to Guided Learning of Boolean Functions	101
5.1	Some Background Information	101
5.2	Problem Description	104
5.3	The Proposed Approach	105
5.4	On the Number of Candidate Solutions	110
5.5	An Illustrative Example	111
5.6	Some Computational Results	113
5.7	Concluding Remarks	122
6	An Incremental Learning Algorithm for Inferring Boolean Functions	125
6.1	Some Background Information	125
6.2	Problem Description	126
6.3	Some Related Developments	127
6.4	The Proposed Incremental Algorithm	130
6.4.1	Repairing a Boolean Function that Incorrectly Rejects a Positive Example	131
6.4.2	Repairing of a Boolean Function that Incorrectly Accepts a Negative Example	133
6.4.3	Computational Complexity of the Algorithms for the ILE Approach	134
6.5	Experimental Data	134
6.6	Analysis of the Computational Results	135
6.6.1	Results on the Classification Accuracy	136
6.6.2	Results on the Number of Clauses	139
6.6.3	Results on the CPU Times	141
6.7	Concluding Remarks	144
7	A Duality Relationship Between Boolean Functions in CNF and DNF Derivable from the Same Training Examples	147
7.1	Introduction	147
7.2	Generating Boolean Functions in CNF and DNF Form	147
7.3	An Illustrative Example of Deriving Boolean Functions in CNF and DNF	148
7.4	Some Computational Results	149
7.5	Concluding Remarks	150
8	The Rejectability Graph of Two Sets of Examples	151
8.1	Introduction	151
8.2	The Definition of the Rejectability Graph	152

8.2.1	Properties of the Rejectability Graph	153
8.2.2	On the Minimum Clique Cover of the Rejectability Graph	155
8.3	Problem Decomposition	156
8.3.1	Connected Components	156
8.3.2	Clique Cover	157
8.4	An Example of Using the Rejectability Graph	158
8.5	Some Computational Results	160
8.6	Concluding Remarks	170

Part II Application Issues

9	The Reliability Issue in Data Mining: The Case of Computer-Aided Breast Cancer Diagnosis	173
9.1	Introduction	173
9.2	Some Background Information on Computer-Aided Breast Cancer Diagnosis	173
9.3	Reliability Criteria	175
9.4	The Representation/Narrow Vicinity Hypothesis	178
9.5	Some Computational Results	181
9.6	Concluding Remarks	183
	Appendix I: Definitions of the Key Attributes	185
	Appendix II: Technical Procedures	187
	9.A.1 The Interactive Approach	187
	9.A.2 The Hierarchical Approach	188
	9.A.3 The Monotonicity Property	188
	9.A.4 Logical Discriminant Functions	189
10	Data Mining and Knowledge Discovery by Means of Monotone Boolean Functions	191
10.1	Introduction	191
10.2	Background Information	193
10.2.1	Problem Descriptions	193
10.2.2	Hierarchical Decomposition of Attributes	196
10.2.3	Some Key Properties of Monotone Boolean Functions	197
10.2.4	Existing Approaches to Problem 1	201
10.2.5	An Existing Approach to Problem 2	203
10.2.6	Existing Approaches to Problem 3	204
10.2.7	Stochastic Models for Problem 3	204
10.3	Inference Objectives and Methodology	206
10.3.1	The Inference Objective for Problem 1	206
10.3.2	The Inference Objective for Problem 2	207
10.3.3	The Inference Objective for Problem 3	208
10.3.4	Incremental Updates for the Fixed Misclassification Probability Model	208

10.3.5	Selection Criteria for Problem 1	209
10.3.6	Selection Criteria for Problems 2.1, 2.2, and 2.3	210
10.3.7	Selection Criterion for Problem 3	210
10.4	Experimental Results	215
10.4.1	Experimental Results for Problem 1	215
10.4.2	Experimental Results for Problem 2	217
10.4.3	Experimental Results for Problem 3	219
10.5	Summary and Discussion	223
10.5.1	Summary of the Research Findings	223
10.5.2	Significance of the Research Findings	225
10.5.3	Future Research Directions	226
10.6	Concluding Remarks	227
11	Some Application Issues of Monotone Boolean Functions	229
11.1	Some Background Information	229
11.2	Expressing Any Boolean Function in Terms of Monotone Ones	229
11.3	Formulations of Diagnostic Problems as the Inference of Nested Monotone Boolean Functions	231
11.3.1	An Application to a Reliability Engineering Problem	231
11.3.2	An Application to the Breast Cancer Diagnosis Problem	232
11.4	Design Problems	233
11.5	Process Diagnosis Problems	234
11.6	Three Major Illusions in the Evaluation of the Accuracy of Data Mining Models	234
11.6.1	First Illusion: The Single Index Accuracy Rate	235
11.6.2	Second Illusion: Accurate Diagnosis without Hard Cases	235
11.6.3	Third Illusion: High Accuracy on Random Test Data Only	236
11.7	Identification of the Monotonicity Property	236
11.8	Concluding Remarks	239
12	Mining of Association Rules	241
12.1	Some Background Information	241
12.2	Problem Description	243
12.3	Methodology	244
12.3.1	Some Related Algorithmic Developments	244
12.3.2	Alterations to the RA1 Algorithm	245
12.4	Computational Experiments	247
12.5	Concluding Remarks	255
13	Data Mining of Text Documents	257
13.1	Some Background Information	257
13.2	A Brief Description of the Document Clustering Process	259
13.3	Using the OACT Approach to Classify Text Documents	260

13.4	An Overview of the Vector Space Model	262
13.5	A Guided Learning Approach for the Classification of Text Documents	264
13.6	Experimental Data	265
13.7	Testing Methodology	267
13.7.1	The Leave-One-Out Cross Validation	267
13.7.2	The 30/30 Cross Validation	267
13.7.3	Statistical Performance of Both Algorithms	267
13.7.4	Experimental Setting for the Guided Learning Approach	268
13.8	Results for the Leave-One-Out and the 30/30 Cross Validations ..	269
13.9	Results for the Guided Learning Approach	272
13.10	Concluding Remarks	275
14	First Case Study: Predicting Muscle Fatigue from EMG Signals	277
14.1	Introduction	277
14.2	General Problem Description	277
14.3	Experimental Data	279
14.4	Analysis of the EMG Data	280
14.4.1	The Effects of Load and Electrode Orientation	280
14.4.2	The Effects of Muscle Condition, Load, and Electrode Orientation	280
14.5	A Comparative Analysis of the EMG Data	281
14.5.1	Results by the OCAT/RA1 Approach	282
14.5.2	Results by Fisher's Linear Discriminant Analysis	283
14.5.3	Results by Logistic Regression	284
14.5.4	A Neural Network Approach	285
14.6	Concluding Remarks	287
15	Second Case Study: Inference of Diagnostic Rules for Breast Cancer ..	289
15.1	Introduction	289
15.2	Description of the Data Set	289
15.3	Description of the Inferred Rules	292
15.4	Concluding Remarks	296
16	A Fuzzy Logic Approach to Attribute Formalization: Analysis of Lobulation for Breast Cancer Diagnosis	297
16.1	Introduction	297
16.2	Some Background Information on Digital Mammography	297
16.3	Some Background Information on Fuzzy Sets	299
16.4	Formalization with Fuzzy Logic	300
16.5	Degrees of Lobularity and Microlobularity	306
16.6	Concluding Remarks	308

17	Conclusions	309
17.1	General Concluding Remarks	309
17.2	Twelve Key Areas of Potential Future Research on Data Mining and Knowledge Discovery from Databases	310
17.2.1	Overfitting and Overgeneralization	310
17.2.2	Guided Learning	311
17.2.3	Stochasticity	311
17.2.4	More on Monotonicity	311
17.2.5	Visualization	311
17.2.6	Systems for Distributed Computing Environments	312
17.2.7	Developing Better Exact Algorithms and Heuristics	312
17.2.8	Hybridization and Other Algorithmic Issues	312
17.2.9	Systems with Self-Explanatory Capabilities	313
17.2.10	New Systems for Image Analysis	313
17.2.11	Systems for Web Applications	313
17.2.12	Developing More Applications	314
17.3	Epilogue	314
	References	317
	Subject Index	335
	Author Index	345
	About the Author	349