
Contents

Preface	xiii
Contributors	xvii
 Part I KNOWLEDGE DISCOVERY AND DATA MINING IN THEORY	 1
 1 Estimating concept difficulty with cross entropy.	
<i>K. Nazar and M.A. Bramer</i>	3
1.1 Introduction	3
1.1.1 Attributes and examples	4
1.2 The need for bias in machine learning	4
1.2.1 Bias interference	5
1.3 Concept dispersion and feature interaction	7
1.4 Why do we need to estimate concept difficulty?	8
1.5 Data-based difficulty measures	10
1.5.1 μ -ness	10
1.5.2 Variation	11
1.5.3 Blurring	11
1.6 Why use the J measure as a basis for a blurring measure?	14
1.7 Effects of various data characteristics on blurring and variation	15
1.7.1 Irrelevant attributes	17
1.8 Dataset analysis	19
1.9 Problems with information-theoretic-based blurring measures	21
1.10 Estimating attribute relevance with the RELIEF algorithm	22
1.10.1 Experiments with RELIEF	23
1.11 Blurring over disjoint subsets	26
1.12 Results and discussion	27
1.13 Summary and future research	28
1.14 References	30
 2 Analysing outliers by searching for plausible hypotheses.	
<i>X. Liu and G. Cheng</i>	32
2.1 Introduction	32
2.2 Statistical treatment of outliers	33

2.3	An algorithm for outlier analysis	35
2.4	Experimental results	38
2.4.1	Case I: onchocerciasis	38
2.4.2	Case II: glaucoma	39
2.5	Evaluation	41
2.6	Concluding remarks	44
2.7	Acknowledgments	44
2.8	References	44
3	Attribute-value distribution as a technique for increasing the efficiency of data mining. <i>D. McSherry</i>	46
3.1	Introduction	46
3.2	Targeting a restricted class of rules	48
3.3	Discovery effort and yield	50
3.4	Attribute-value distribution	53
3.5	Experimental results	55
3.5.1	The contact-lens data	55
3.5.2	The project-outcome dataset	60
3.6	Discussion and conclusions	62
3.7	References	63
4	Using background knowledge with attribute-oriented data mining. <i>M. Shapcott, S. McClean and B. Scotney</i>	64
4.1	Introduction	64
4.2	Partial value model	66
4.2.1	Definition: partial value	66
4.2.2	Definition: partial-value relation	66
4.2.3	Example	66
4.2.4	Aggregation	67
4.2.5	Definition: simple count operator	67
4.2.6	Example	68
4.2.7	Definition: simple aggregate operator	68
4.2.8	Example	68
4.2.9	Aggregation of partial values	69
4.2.10	Definition: partial value aggregate operator	69
4.2.11	Example	69
4.2.12	Definition: partial-value count operator	70
4.2.13	Theoretical framework	70
4.3	Reengineering the database – the role of background knowledge	71
4.3.1	Concept hierarchies	71
4.3.2	Database-integrity constraints	73
4.3.3	Definition: simple comparison predicate	73
4.3.4	Examples of simple comparison predicates	74
4.3.5	Definition: table-based predicate	74

4.3.6	Example of a table-based predicate	74
4.3.7	Expression of rules as table-based predicates	74
4.3.8	Reengineering the database	74
4.4	Multiattribute count operators	76
4.4.1	Example	77
4.4.2	Example	77
4.4.3	Example	78
4.4.4	Quasi-independence	79
4.4.5	Example	81
4.4.6	Interestingness	81
4.4.7	Example	82
4.5	Related work	82
4.6	Conclusions	83
4.7	Acknowledgments	84
4.8	References	84
5	A development framework for temporal data mining.	
	<i>X. Chen and I. Petrounias</i>	87
5.1	Introduction	87
5.2	Analysis and representation of temporal features	89
5.2.1	Time domain	89
5.2.2	Calendar expression of time	90
5.2.3	Periodicity of time	92
5.2.4	Time dimensions in temporal databases	94
5.3	Potential knowledge and temporal data mining problems	95
5.3.1	Forms of potential temporal knowledge	95
5.3.2	Associating knowledge with temporal features	97
5.3.3	Temporal mining problems	98
5.4	A framework for temporal data mining	100
5.4.1	A temporal mining language	100
5.4.2	System architecture	102
5.5	An example: discovery of temporal association rules	104
5.5.1	Mining problem	104
5.5.2	Description of mining tasks in TQML	106
5.5.3	Search algorithms	107
5.6	Conclusion and future research direction	110
5.7	References	110
6	An integrated architecture for OLAP and data mining.	
	<i>Z. Chen</i>	114
6.1	Introduction	115
6.2	Preliminaries	116
6.2.1	Decision-support queries	116
6.2.2	Data warehousing	116
6.2.3	Basics of OLAP	117
6.2.4	Star schema	118

6.2.5	A materialised view for sales profit	118
6.3	Differences between OLAP and data mining	119
6.3.1	Basic concepts of data mining	119
6.3.2	Different types of query can be answered at different levels	120
6.3.3	Aggregation semantics	120
6.3.4	Sensitivity analysis	122
6.3.5	Different assumptions or heuristics may be needed at different levels	123
6.4	Combining OLAP and data mining: the feedback sandwich model	123
6.4.1	Two different ways of combining OLAP and data mining	124
6.4.2	The feedback sandwich model	125
6.5	Towards integrated architecture for combined OLAP/data mining	126
6.6	Three specific issues	128
6.6.1	On the use and reuse of intensional historical data	128
6.6.2	How data mining can benefit OLAP	131
6.6.3	OLAP-enriched data mining	133
6.7	Conclusion	135
6.8	References	135
Part II	KNOWLEDGE DISCOVERY AND DATA MINING IN PRACTICE	137
7	Empirical studies of the knowledge discovery approach to health-information analysis. <i>M. Lloyd-Williams</i>	139
7.1	Introduction	139
7.2	Knowledge discovery and data mining	140
7.2.1	The knowledge discovery process	140
7.2.2	Artificial neural networks	147
7.3	Empirical studies	149
7.3.1	The 'Health for all' database	149
7.3.2	The 'Babies at risk of intrapartum asphyxia' database	153
7.3.3	Infertility databases	155
7.4	Conclusions	157
7.5	References	158
8	Direct knowledge discovery and interpretation from a multilayer perceptron network which performs low-back-pain classification <i>M.L. Vaughn, S.J. Cavill, S.J. Taylor, M.A. Foy and A.J.B. Fogg</i>	160
8.1	Introduction	160
8.2	The MLP network	161
8.3	The low-back-pain MLP network	163

8.4	The interpretation and knowledge-discovery method	163
8.4.1	Discovery of the feature-detector neurons	163
8.4.2	Discovery of the significant inputs	164
8.4.3	Discovery of the negated significant inputs	164
8.4.4	Knowledge learned by the MLP from the training data	166
8.4.5	MLP network validation and verification	167
8.5	Knowledge discovery from LBP example training cases	168
8.5.1	Discovery of the feature detectors for example training cases	168
8.5.2	Discovery of the significant inputs for example training cases	168
8.5.3	Data relationships and explanations for example training cases	168
8.5.4	Discussion of the training-example data relationships	170
8.5.5	Induced rules from training-example cases	170
8.6	Knowledge discovery from all LBP MLP training cases	173
8.6.1	Discussion of the class key input rankings	173
8.7	Validation of the LBP LMP network	176
8.7.1	Validation of the training cases	176
8.7.2	Validation of the test cases	176
8.8	Conclusions	177
8.9	Future work	177
8.10	Acknowledgments	178
8.11	References	178
9	Discovering knowledge from low-quality meteorological databases.	
	<i>C.M. Howard and V.J. Rayward-Smith</i>	180
9.1	Introduction	180
9.1.1	The meteorological domain	180
9.2	The preprocessing stage	182
9.2.1	Visualisation	182
9.2.2	Missing values	182
9.2.3	Unreliable data	185
9.2.4	Discretisation	186
9.2.5	Feature selection	187
9.2.6	Feature construction	188
9.3	The data-mining stage	188
9.3.1	Simulated annealing	189
9.3.2	SA with missing and unreliable data	190
9.4	A toolkit for knowledge discovery	195
9.5	Results and analysis	196
9.5.1	Results from simulated annealing	198
9.5.2	Results from C5.0	198
9.5.3	Comparison and evaluation of results	199
9.6	Summary	199

9.7	Discussion and further work	200
9.8	References	201
10	A meteorological knowledge-discovery environment. <i>A.G. Büchner, J.C.L. Chan, S.L. Hung and J.G. Hughes</i>	204
10.1	Introduction	204
10.2	Some meteorological background	205
10.2.1	Available data sources	206
10.2.2	Related work	208
10.3	MADAME's architecture	211
10.4	Building a meteorological data warehouse	212
10.4.1	The design	212
10.4.2	Information extraction	213
10.4.3	Data cleansing	214
10.4.4	Data processing	215
10.4.5	Data loading and refreshing	216
10.5	The knowledge-discovery components	217
10.5.1	Knowledge modelling	217
10.5.2	Domain knowledge	221
10.6	Prediction trial runs	222
10.6.1	Nowcasting of heavy rainfall	222
10.6.2	Landslide nowcasting	223
10.7	Conclusions and further work	224
10.8	Acknowledgments	224
10.9	References	225
11	Mining the organic compound jungle – a functional programming approach. <i>K.E. Burn-Thornton and J. Bradshaw</i>	227
11.1	Introduction	227
11.2	Decision-support requirements in the pharmaceutical industry	227
11.2.1	Graphical comparison	228
11.2.2	Structural keys	228
11.2.3	Fingerprints	229
11.2.4	Variable-sized fingerprints	230
11.3	Functional programming language Gofer	230
11.3.1	Functional programming	231
11.4	Design of prototype tool and main functions	234
11.4.1	Design of tool	234
11.4.2	Main functions	235
11.5	Methodology	236
11.6	Results	236
11.6.1	Sets A, B and C (256, 512 and 1024 bytes)	237
11.6.2	Set D (2048 bytes)	237
11.7	Conclusions	238

11.8	Future work	239
11.9	References	239
12	Data mining with neural networks – an applied example in understanding electricity consumption patterns.	
	<i>P. Brierley and B. Batty</i>	240
12.1	Neural networks	241
12.1.1	What are neural networks?	241
12.1.2	Why use neural networks?	242
12.1.3	How do neural networks process information?	242
12.1.4	Things to be aware of . . .	244
12.2	Electric load modelling	250
12.2.1	The data being mined	250
12.2.2	Why forecast electricity demand?	252
12.2.3	Previous work	253
12.2.4	Network used	254
12.3	Total daily load model	255
12.3.1	Overfitting and generalisation	264
12.4	Rule extraction	266
12.4.1	Day of the week	267
12.4.2	Time of year	268
12.4.3	Growth	269
12.4.4	Weather factors	271
12.4.5	Holidays	272
12.5	Model comparisons	274
12.6	Half-hourly model	276
12.6.1	Initial input data	276
12.6.2	Results	277
12.6.3	Past loads	286
12.6.4	How the model is working	286
12.6.5	Extracting the growth	287
12.6.6	Populations of models	288
12.7	Summary	288
12.8	References	289
12.9	Appendixes	292
12.9.1	Backpropagation weight update rule	292
12.9.2	Fortran 90 code for a multilayer perceptron	299
Index		304