

Contents

1	Introduction	3
1.1	What motivated data mining? Why is it important?	3
1.2	So, what is data mining?	6
1.3	Data mining — on what kind of data?	10
1.3.1	Relational databases	10
1.3.2	Data warehouses	12
1.3.3	Transactional databases	15
1.3.4	Advanced database systems and advanced database applications	15
1.4	Data mining functionalities — what kinds of patterns can be mined?	20
1.4.1	Concept/class description: characterization and discrimination	20
1.4.2	Association analysis	22
1.4.3	Classification and prediction	23
1.4.4	Cluster analysis	24
1.4.5	Outlier analysis	24
1.4.6	Evolution analysis	25
1.5	Are all of the patterns interesting?	25
1.6	Classification of data mining systems	27
1.7	Major issues in data mining	29
1.8	Summary	32
2	Data Warehouse and OLAP Technology for Data Mining	37
2.1	What is a data warehouse?	37
2.2	A multidimensional data model	42

2.2.1	From tables and spreadsheets to data cubes	42
2.2.2	Stars, snowflakes, and fact constellations: schemas for multidimensional databases	45
2.2.3	Examples for defining star, snowflake, and fact constellation schemas	48
2.2.4	Measures: their categorization and computation	50
2.2.5	Introducing concept hierarchies	52
2.2.6	OLAP operations in the multidimensional data model	54
2.2.7	A starnet query model for querying multidimensional databases	57
2.3	Data warehouse architecture	58
2.3.1	Steps for the design and construction of data warehouses	58
2.3.2	A three-tier data warehouse architecture	61
2.3.3	Types of OLAP servers: ROLAP vs. MOLAP vs. HOLAP	64
2.4	Data warehouse implementation	66
2.4.1	Efficient computation of data cubes	66
2.4.2	Indexing OLAP data	73
2.4.3	Efficient processing of OLAP queries	75
2.4.4	Meta data repository	77
2.4.5	Data warehouse back-end tools and utilities	78
2.5	Further development of data cube technology	79
2.5.1	Discovery-driven exploration of data cubes	79
2.5.2	Complex aggregation at multiple granularities: Multifeature cubes	83
2.5.3	Other developments	86
2.6	From data warehousing to data mining	86
2.6.1	Data warehouse usage	86
2.6.2	From on-line analytical processing to on-line analytical mining	88
2.7	Summary	90
3	Data Preprocessing	97
3.1	Why preprocess the data?	97
3.2	Data cleaning	100
3.2.1	Missing values	100

3.2.2	Noisy data	101
3.2.3	Inconsistent data	103
3.3	Data integration and transformation	104
3.3.1	Data integration	104
3.3.2	Data transformation	105
3.4	Data reduction	107
3.4.1	Data cube aggregation	108
3.4.2	Dimensionality reduction	110
3.4.3	Data compression	112
3.4.4	Numerosity reduction	115
3.5	Discretization and concept hierarchy generation	120
3.5.1	Discretization and concept hierarchy generation for numeric data	121
3.5.2	Concept hierarchy generation for categorical data	125
3.6	Summary	127
4	Data Mining Primitives, Languages, and System Architectures	135
4.1	Data mining primitives: what defines a data mining task?	136
4.1.1	Task-relevant data	137
4.1.2	The kind of knowledge to be mined	140
4.1.3	Background knowledge: concept hierarchies	140
4.1.4	Interestingness measures	144
4.1.5	Presentation and visualization of discovered patterns	147
4.2	A data mining query language	148
4.2.1	Syntax for task-relevant data specification	150
4.2.2	Syntax for specifying the kind of knowledge to be mined	151
4.2.3	Syntax for concept hierarchy specification	154
4.2.4	Syntax for interestingness measure specification	156
4.2.5	Syntax for pattern presentation and visualization specification	156
4.2.6	Putting it all together — an example of a DMQL query	157
4.2.7	Other data mining languages and the standardization of data mining primitives	158

4.3	Designing graphical user interfaces based on a data mining query language	160
4.4	Architectures of data mining systems	161
4.5	Summary	163
5	Concept Description: Characterization and Comparison	169
5.1	What is concept description?	169
5.2	Data generalization and summarization-based characterization	171
5.2.1	Attribute-oriented induction	171
5.2.2	Efficient implementation of attribute-oriented induction	177
5.2.3	Presentation of the derived generalization	179
5.3	Analytical characterization: Analysis of attribute relevance	183
5.3.1	Why perform attribute relevance analysis?	183
5.3.2	Methods of attribute relevance analysis	184
5.3.3	Analytical characterization: An example	186
5.4	Mining class comparisons: Discriminating between different classes	188
5.4.1	Class comparison methods and implementations	189
5.4.2	Presentation of class comparison descriptions	192
5.4.3	Class description: Presentation of both characterization and comparison	193
5.5	Mining descriptive statistical measures in large databases	196
5.5.1	Measuring the central tendency	196
5.5.2	Measuring the dispersion of data	197
5.5.3	Graph displays of basic statistical class descriptions	200
5.6	Discussion	204
5.6.1	Concept description: A comparison with typical machine learning methods	204
5.6.2	Incremental and parallel mining of concept description	206
5.7	Summary	207
6	Mining Association Rules in Large Databases	211
6.1	Association rule mining	211
6.1.1	Market basket analysis: A motivating example for association rule mining	212
6.1.2	Basic concepts	213

6.1.3	Association rule mining: A road map	214
6.2	Mining single-dimensional Boolean association rules from transactional databases	215
6.2.1	The Apriori algorithm: Finding frequent itemsets using candidate generation	216
6.2.2	Generating association rules from frequent itemsets	219
6.2.3	Improving the efficiency of Apriori	221
6.2.4	Mining frequent itemsets without candidate generation	223
6.2.5	Iceberg queries	227
6.3	Mining multilevel association rules from transaction databases	229
6.3.1	Multilevel association rules	229
6.3.2	Approaches to mining multilevel association rules	230
6.3.3	Checking for redundant multilevel association rules	234
6.4	Mining multidimensional association rules from relational databases and data warehouses	235
6.4.1	Multidimensional association rules	235
6.4.2	Mining multidimensional association rules using static discretization of quantitative attributes	237
6.4.3	Mining quantitative association rules	238
6.4.4	Mining distance-based association rules	240
6.5	From association mining to correlation analysis	242
6.5.1	Strong rules are not necessarily interesting: An example	242
6.5.2	From association analysis to correlation analysis	243
6.6	Constraint-based association mining	245
6.6.1	Metarule-guided mining of association rules	246
6.6.2	Mining guided by additional rule constraints	247
6.7	Summary	251
7	Classification and Prediction	259
7.1	What is classification? What is prediction?	259
7.2	Issues regarding classification and prediction	262
7.3	Classification by decision tree induction	263
7.3.1	Decision tree induction	264
7.3.2	Tree pruning	268

7.3.3	Extracting classification rules from decision trees	269
7.3.4	Enhancements to basic decision tree induction	270
7.3.5	Scalability and decision tree induction	271
7.3.6	Integrating data warehousing techniques and decision tree induction	273
7.4	Bayesian classification	274
7.4.1	Bayes theorem	275
7.4.2	Naive Bayesian classification	276
7.4.3	Bayesian belief networks	278
7.4.4	Training Bayesian belief networks	280
7.5	Classification by backpropagation	281
7.5.1	A multilayer feed-forward neural network	282
7.5.2	Defining a network topology	283
7.5.3	Backpropagation	283
7.5.4	Backpropagation and interpretability	288
7.6	Classification based on concepts from association rule mining	289
7.7	Other classification methods	291
7.7.1	k -nearest neighbor classifiers	291
7.7.2	Case-based reasoning	292
7.7.3	Genetic algorithms	293
7.7.4	Rough set approach	293
7.7.5	Fuzzy set approaches	294
7.8	Prediction	295
7.8.1	Linear and multiple regression	296
7.8.2	Nonlinear regression	297
7.8.3	Other regression models	298
7.9	Classifier accuracy	299
7.9.1	Estimating classifier accuracy	299
7.9.2	Increasing classifier accuracy	300
7.9.3	Is accuracy enough to judge a classifier?	301
7.10	Summary	303

8 Cluster Analysis	311
8.1 What is cluster analysis?	311
8.2 Types of data in clustering analysis	314
8.2.1 Interval-scaled variables	315
8.2.2 Binary variables	317
8.2.3 Nominal, ordinal, and ratio-scaled variables	318
8.2.4 Variables of mixed types	321
8.3 A categorization of major clustering methods	321
8.4 Partitioning methods	324
8.4.1 Classical partitioning methods: k -means and k -medoids	324
8.4.2 Partitioning methods in large databases: from k -medoids to CLARANS	328
8.5 Hierarchical methods	329
8.5.1 Agglomerative and divisive hierarchical clustering	330
8.5.2 BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies	331
8.5.3 CURE: Clustering Using REpresentatives	333
8.5.4 CHAMELEON: A hierarchical clustering algorithm using dynamic modeling	335
8.6 Density-based methods	337
8.6.1 DBSCAN: A density-based clustering method based on connected regions with sufficiently high density	337
8.6.2 OPTICS: Ordering Points To Identify the Clustering Structure	339
8.6.3 DENCLUE: Clustering based on density distribution functions	341
8.7 Grid-based methods	343
8.7.1 STING: A Statistical Information Grid approach	343
8.7.2 WaveCluster: Clustering using wavelet transformation	345
8.7.3 CLIQUE: Clustering high-dimensional space	347
8.8 Model-based clustering methods	348
8.8.1 Statistical approach	348
8.8.2 Neural network approach	350
8.9 Outlier analysis	351
8.9.1 Statistical-based outlier detection	353

8.9.2	Distance-based outlier detection	354
8.9.3	Deviation-based outlier detection	356
8.10	Summary	358
9	Mining Complex Types of Data	367
9.1	Multidimensional analysis and descriptive mining of complex data objects	367
9.1.1	Generalization of structured data	368
9.1.2	Aggregation and approximation in spatial and multimedia data generalization	369
9.1.3	Generalization of object identifiers and class/subclass hierarchies	370
9.1.4	Generalization of class composition hierarchies	371
9.1.5	Construction and mining of object cubes	371
9.1.6	Generalization-based mining of plan databases by divide-and-conquer	372
9.2	Mining Spatial Databases	375
9.2.1	Spatial data cube construction and spatial OLAP	376
9.2.2	Spatial association analysis	380
9.2.3	Spatial clustering methods	381
9.2.4	Spatial classification and spatial trend analysis	381
9.2.5	Mining raster databases	382
9.3	Mining Multimedia Databases	382
9.3.1	Similarity search in multimedia data	383
9.3.2	Multidimensional analysis of multimedia data	384
9.3.3	Classification and prediction analysis of multimedia data	386
9.3.4	Mining associations in multimedia data	387
9.4	Mining Time-Series and Sequence Data	388
9.4.1	Trend analysis	388
9.4.2	Similarity search in time-series analysis	391
9.4.3	Sequential pattern mining	394
9.4.4	Periodicity analysis	396
9.5	Mining Text Databases	397
9.5.1	Text data analysis and information retrieval	397

9.5.2	Text mining: keyword-based association and document classification	402
9.6	Mining the World-Wide Web	404
9.6.1	Mining the Web's link structures to identify authoritative Web pages	406
9.6.2	Automatic classification of Web documents	408
9.6.3	Construction of a multilayered Web information base	409
9.6.4	Web usage mining	410
9.7	Summary	411
10	Data Mining Applications and Trends in Data Mining	417
10.1	Data mining applications	417
10.1.1	Data mining for biomedical and DNA data analysis	417
10.1.2	Data mining for financial data analysis	419
10.1.3	Data mining for the retail industry	420
10.1.4	Data mining for the telecommunication industry	422
10.2	Data mining system products and research prototypes	423
10.2.1	How to choose a data mining system	423
10.2.2	Examples of commercial data mining systems	426
10.3	Additional themes on data mining	427
10.3.1	Visual and audio data mining	427
10.3.2	Scientific and statistical data mining	433
10.3.3	Theoretical foundations of data mining	434
10.3.4	Data mining and intelligent query answering	435
10.4	Social impacts of data mining	437
10.4.1	Is data mining a hype or a persistent, steadily growing business?	437
10.4.2	Is data mining merely managers' business or everyone's business?	439
10.4.3	Is data mining a threat to privacy and data security?	440
10.5	Trends in data mining	442
10.6	Summary	444
A	Appendix A: An Introduction to Microsoft's OLE DB for Data Mining	449
B	Appendix B: An Introduction to DBMiner	457