

Charu C. Aggarwal

Data Mining

The Textbook

About the Book

This textbook explores the different aspects of data mining from the fundamentals to the complex data types and their applications, capturing the wide diversity of problem domains for data mining issues. It goes beyond the traditional focus on data mining problems to introduce advanced data types such as text, time series, discrete sequences, spatial data, graph data, and social networks. Until now, no single book has addressed all these topics in a comprehensive and integrated way. The chapters of this book fall into one of three categories:

- **Fundamental chapters:** Data mining has four main problems, which correspond to clustering, classification, association pattern mining, and outlier analysis. These chapters comprehensively discuss a wide variety of methods for these problems.
- **Domain chapters:** These chapters discuss the specific methods used for different domains of data such as text data, time-series data, sequence data, graph data, and spatial data.
- **Application chapters:** These chapters study important applications such as stream mining, Web mining, ranking, recommendations, social networks, and privacy preservation. The domain chapters also have an applied flavor.

Appropriate for both introductory and advanced data mining courses, *Data Mining: The Textbook* balances mathematical details and intuition. It contains the necessary mathematical details for professors and researchers, but it is presented in a simple and intuitive style to improve accessibility for students and industrial practitioners (including those with a limited mathematical background). Numerous illustrations, examples, and exercises are included, with an emphasis on semantically interpretable examples.

About the Author

Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T.J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996. He has published more than 250 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 14 books, including the first comprehensive book on outlier analysis, which



is written from a computer science point of view. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He has won multiple awards, chaired conferences, and currently serves on several journal editorial boards. He is a fellow of the ACM and the IEEE, for “contributions to knowledge discovery and data mining algorithms.”

Computer Science

ISBN 978-3-319-14141-1



► springer.com

Aggarwal



Data Mining

Charu C. Aggarwal

Data Mining

The Textbook

 Springer

**Computers connected to subscribing institutions can
download book from the following clickable URL:
<http://rd.springer.com/book/10.1007/978-3-319-14142-8>**

Data Mining: The Textbook

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, New York

March 8, 2015

**Computers connected to subscribing institutions can
download book from the following clickable URL:
<http://rd.springer.com/book/10.1007/978-3-319-14142-8>**

**Computers connected to subscribing institutions can
download book from the following clickable URL:
<http://rd.springer.com/book/10.1007/978-3-319-14142-8>**

To my wife Lata,
and my daughter Sayani

**Computers connected to subscribing institutions can
download book from the following clickable URL:
<http://rd.springer.com/book/10.1007/978-3-319-14142-8>**

Contents

1	An Introduction to Data Mining	1
1.1	Introduction	1
1.2	The Data Mining Process	3
1.2.1	The Data Preprocessing Phase	5
1.2.2	The Analytical Phase	6
1.3	The Basic Data Types	6
1.3.1	Non-dependency Oriented Data	6
1.3.1.1	Quantitative Multidimensional Data	7
1.3.1.2	Categorical and Mixed Attribute Data	7
1.3.1.3	Binary and Set Data	8
1.3.1.4	Text Data	8
1.3.2	Dependency Oriented Data	9
1.3.2.1	Time-Series Data	9
1.3.2.2	Discrete Sequences and Strings	10
1.3.2.3	Spatial Data	11
1.3.2.4	Network and Graph Data	12
1.4	The Major Building Blocks: A Bird's Eye View	13
1.4.1	Association Pattern Mining	14
1.4.2	Data Clustering	15
1.4.3	Outlier Detection	16
1.4.4	Data Classification	17
1.4.5	Impact of Complex Data Types on Problem Definitions	18
1.4.5.1	Pattern Mining with Complex Data Types	18
1.4.5.2	Clustering with Complex Data Types	19
1.4.5.3	Outlier Detection with Complex Data Types	19
1.4.5.4	Classification with Complex Data Types	20
1.5	Scalability Issues and the Streaming Scenario	20
1.6	A Stroll through some Application Scenarios	21
1.6.1	Store Product Placement	21
1.6.2	Customer Recommendations	21
1.6.3	Medical Diagnosis	22
1.6.4	Web Log Anomalies	22
1.7	Summary	23
1.8	Bibliographic Notes	23
1.9	Exercises	24

2	Data Preparation	25
2.1	Introduction	25
2.2	Feature Extraction and Portability	26
2.2.1	Feature Extraction	26
2.2.2	Data Type Portability	27
2.2.2.1	Numeric to Categorical Data: Discretization	28
2.2.2.2	Categorical to Numeric Data: Binarization	29
2.2.2.3	Text to Numeric Data	29
2.2.2.4	Time Series to Discrete Sequence Data	30
2.2.2.5	Time Series to Numeric Data	30
2.2.2.6	Discrete Sequence to Numeric Data	30
2.2.2.7	Spatial to Numeric Data	31
2.2.2.8	Graphs to Numeric Data	31
2.2.2.9	Any Type to Graphs for Similarity-based Applications	31
2.3	Data Cleaning	32
2.3.1	Handling Missing Entries	33
2.3.2	Handling Incorrect and Inconsistent Entries	33
2.3.3	Scaling and Normalization	34
2.4	Data Reduction and Transformation	35
2.4.1	Sampling	35
2.4.1.1	Sampling for Static Data	36
2.4.1.2	Reservoir Sampling for Data Streams	36
2.4.2	Feature Subset Selection	38
2.4.3	Dimensionality Reduction with Axis Rotation	38
2.4.3.1	Principal Component Analysis	39
2.4.3.2	Singular Value Decomposition	41
2.4.3.3	Latent Semantic Analysis	45
2.4.3.4	Applications of PCA and SVD	45
2.4.4	Dimensionality Reduction with Type Transformation	46
2.4.4.1	Haar Wavelet Transform	47
2.4.4.2	Multidimensional Scaling	52
2.4.4.3	Spectral Transformation and Embedding of Graphs	54
2.5	Summary	56
2.6	Bibliographic Notes	57
2.7	Exercises	57
3	Similarity and Distances	59
3.1	Introduction	59
3.2	Multidimensional Data	60
3.2.1	Quantitative Data	60
3.2.1.1	Impact of Domain-specific Relevance	61
3.2.1.2	Impact of High Dimensionality	61
3.2.1.3	Impact of Locally Irrelevant Features	62
3.2.1.4	Impact of Different L_p -norms	63
3.2.1.5	Match-based Similarity Computation	64
3.2.1.6	Impact of Data Distribution	65
3.2.1.7	Nonlinear Distributions: ISOMAP	66
3.2.1.8	Impact of Local Data Distribution	67
3.2.1.9	Computational Considerations	69
3.2.2	Categorical Data	69
3.2.3	Mixed Quantitative and Categorical Data	70

3.3	Text Similarity Measures	71
3.3.1	Binary and Set Data	72
3.4	Temporal Similarity Measures	72
3.4.1	Time-Series Similarity Measures	73
3.4.1.1	Impact of Behavioral Attribute Normalization	74
3.4.1.2	L_p -norm	74
3.4.1.3	Dynamic Time Warping Distance	74
3.4.1.4	Window-based Methods	77
3.4.2	Discrete Sequence Similarity Measures	77
3.4.2.1	Edit Distance	77
3.4.2.2	Longest Common Subsequence	79
3.5	Graph Similarity Measures	80
3.5.1	Similarity between Two Nodes in a Single Graph	80
3.5.1.1	Structural Distance-based Measure	80
3.5.1.2	Random Walk-based Similarity	81
3.5.2	Similarity between Two Graphs	81
3.6	Supervised Similarity Functions	82
3.7	Summary	83
3.8	Bibliographic Notes	84
3.9	Exercises	85
4	Association Pattern Mining	87
4.1	Introduction	87
4.2	The Frequent Pattern Mining Model	88
4.3	Association Rule Generation Framework	91
4.4	Frequent Itemset Mining Algorithms	92
4.4.1	Brute Force Algorithms	93
4.4.2	The Apriori Algorithm	94
4.4.2.1	Efficient Support Counting	95
4.4.3	Enumeration-Tree Algorithms	96
4.4.3.1	Enumeration-Tree-based Interpretation of Apriori	99
4.4.3.2	TreeProjection and DepthProject	99
4.4.3.3	Vertical Counting Methods	104
4.4.4	Recursive Suffix-based Pattern Growth Methods	106
4.4.4.1	Implementation with Arrays but no Pointers	107
4.4.4.2	Implementation with Pointers but no FP-Tree	108
4.4.4.3	Implementation with Pointers and FP-Tree	109
4.4.4.4	Trade-offs with Different Data Structures	112
4.4.4.5	Relationship between FP-growth and Enumeration-Tree Methods	113
4.5	Alternative Models: Interesting Patterns	115
4.5.1	Statistical Coefficient of Correlation	116
4.5.2	χ^2 Measure	116
4.5.3	Interest Ratio	117
4.5.4	Symmetric Confidence Measures	117
4.5.5	Cosine Coefficient on Columns	118
4.5.6	Jaccard Coefficient and the Min-hash Trick	118
4.5.7	Collective Strength	119
4.5.8	Relationship to Negative Pattern Mining	120
4.6	Useful Meta-Algorithms	120
4.6.1	Sampling Methods	120

4.6.2	Data Partitioned Ensembles	121
4.6.3	Generalization to Other Data Types	121
4.6.3.1	Quantitative Data	122
4.6.3.2	Categorical Data	122
4.7	Summary	122
4.8	Bibliographic Notes	123
4.9	Exercises	124
5	Association Pattern Mining: Advanced Concepts	127
5.1	Introduction	127
5.2	Pattern Summarization	128
5.2.1	Maximal Patterns	128
5.2.2	Closed Patterns	129
5.2.3	Approximate Frequent Patterns	131
5.2.3.1	Approximation in Terms of Transactions	131
5.2.3.2	Approximation in Terms of Itemsets	132
5.3	Pattern Querying	133
5.3.1	Preprocess-once Query-many Paradigm	133
5.3.1.1	Leveraging the Itemset Lattice	133
5.3.1.2	Leveraging Data Structures for Querying	134
5.3.2	Pushing Constraints into Pattern Mining	138
5.4	Putting Associations to Work: Applications	139
5.4.1	Relationship to Other Data Mining Problems	139
5.4.1.1	Application to Classification	139
5.4.1.2	Application to Clustering	139
5.4.1.3	Applications to Outlier Detection	139
5.4.2	Market Basket Analysis	140
5.4.3	Demographic and Profile Analysis	140
5.4.4	Recommendations and Collaborative Filtering	140
5.4.5	Web Log Analysis	140
5.4.6	Bioinformatics	141
5.4.7	Other Applications for Complex Data Types	141
5.5	Summary	141
5.6	Bibliographic Notes	142
5.7	Exercises	143
6	Cluster Analysis	145
6.1	Introduction	145
6.2	Feature Selection for Clustering	146
6.2.1	Filter Models	147
6.2.1.1	Term Strength	147
6.2.1.2	Predictive Attribute Dependence	147
6.2.1.3	Entropy	148
6.2.1.4	Hopkins Statistic	149
6.2.2	Wrapper Models	150
6.3	Representative-based Algorithms	151
6.3.1	The k -Means Algorithm	154
6.3.2	The Kernel k -Means Algorithm	155
6.3.3	The k -Medians Algorithm	155
6.3.4	The k -Medoids Algorithm	155
6.4	Hierarchical Clustering Algorithms	158

6.4.1	Bottom-up Agglomerative Methods	159
6.4.1.1	Group-based Statistics	160
6.4.2	Top-down Divisive Methods	163
6.4.2.1	Bisecting k -Means	164
6.5	Probabilistic Model-based Algorithms	164
6.5.1	Relationship of EM to k -means and other Representative Methods	167
6.6	Grid-based and Density-based Algorithms	169
6.6.1	Grid-based Methods	170
6.6.2	DBSCAN	172
6.6.3	DENCLUE	174
6.7	Graph-based Algorithms	177
6.7.1	Properties of Graph-based Algorithms	180
6.8	Nonnegative Matrix Factorization	181
6.8.1	Comparison with Singular Value Decomposition	185
6.9	Cluster Validation	186
6.9.1	Internal Validation Criteria	186
6.9.1.1	Parameter Tuning with Internal Measures	188
6.9.2	External Validation Criteria	189
6.9.3	General Comments	191
6.10	Summary	191
6.11	Bibliographic Notes	192
6.12	Exercises	192
7	Cluster Analysis: Advanced Concepts	195
7.1	Introduction	195
7.2	Clustering Categorical Data	196
7.2.1	Representative-based Algorithms	197
7.2.1.1	k -Modes Clustering	198
7.2.1.2	k -Medoids Clustering	198
7.2.2	Hierarchical Algorithms	199
7.2.2.1	ROCK	199
7.2.3	Probabilistic Algorithms	200
7.2.4	Graph-based Algorithms	202
7.3	Scalable Data Clustering	202
7.3.1	CLARANS	202
7.3.2	BIRCH	203
7.3.3	CURE	205
7.4	High-Dimensional Clustering	207
7.4.1	CLIQUE	208
7.4.2	PROCLUS	209
7.4.3	ORCLUS	212
7.5	Semisupervised Clustering	214
7.5.1	Pointwise Supervision	214
7.5.2	Pairwise Supervision	215
7.6	Human and Visually Supervised Clustering	216
7.6.1	Modifications of Existing Clustering Algorithms	217
7.6.2	Visual Clustering	217
7.7	Cluster Ensembles	220
7.7.1	Selecting Different Ensemble Components	221
7.7.2	Combining Different Ensemble Components	221
7.7.2.1	Hypergraph Partitioning Algorithm	221

7.7.2.2	Meta-clustering Algorithm	222
7.8	Putting Clustering to Work: Applications	222
7.8.1	Applications to Other Data Mining Problems	222
7.8.1.1	Data Summarization	222
7.8.1.2	Outlier Analysis	222
7.8.1.3	Classification	223
7.8.1.4	Dimensionality Reduction	223
7.8.1.5	Similarity Search and Indexing	223
7.8.2	Customer Segmentation and Collaborative Filtering	223
7.8.3	Text Applications	223
7.8.4	Multimedia Applications	224
7.8.5	Temporal and Sequence Applications	224
7.8.6	Social Network Analysis	224
7.9	Summary	224
7.10	Bibliographic Notes	224
7.11	Exercises	225
8	Outlier Analysis	227
8.1	Introduction	227
8.2	Extreme Value Analysis	229
8.2.1	Univariate Extreme Value Analysis	230
8.2.2	Multivariate Extreme Values	231
8.2.3	Depth-based Methods	233
8.3	Probabilistic Models	234
8.4	Clustering for Outlier Detection	236
8.5	Distance-based Outlier Detection	238
8.5.1	Pruning Methods	239
8.5.1.1	Sampling Methods	239
8.5.1.2	Early Termination Trick with Nested Loops	239
8.5.2	Local Distance Correction Methods	240
8.5.2.1	Local Outlier Factor (LOF)	242
8.5.2.2	Instance-specific Mahalanobis Distance	243
8.6	Density-based Methods	244
8.6.1	Histogram- and Grid-based Techniques	244
8.6.2	Kernel Density Estimation	245
8.7	Information-Theoretic Models	246
8.8	Outlier Validity	247
8.8.1	Methodological Challenges	248
8.8.2	Receiver Operating Characteristic	249
8.8.3	Common Mistakes	250
8.9	Summary	250
8.10	Bibliographic Notes	251
8.11	Exercises	251
9	Outlier Analysis: Advanced Concepts	253
9.1	Introduction	253
9.2	Outlier Detection with Categorical Data	254
9.2.1	Probabilistic Models	254
9.2.2	Clustering and Distance-based Methods	255
9.2.3	Binary and Set-Valued Data	256
9.3	High-Dimensional Outlier Detection	256

9.3.1	Grid-based Rare Subspace Exploration	258
9.3.1.1	Modeling Abnormal Lower Dimensional Projections	258
9.3.1.2	Grid Search for Subspace Outliers	259
9.3.2	Random Subspace Sampling	261
9.4	Outlier Ensembles	262
9.4.1	Categorization by Component Independence	263
9.4.1.1	Sequential Ensembles	263
9.4.1.2	Independent Ensembles	264
9.4.2	Categorization by Constituent Components	264
9.4.2.1	Model-centered Ensembles	265
9.4.2.2	Data-centered Ensembles	265
9.4.3	Normalization and Combination	265
9.5	Putting Outliers to Work: Applications	267
9.5.1	Quality Control and Fault Detection	267
9.5.2	Financial Fraud and Anomalous Events	267
9.5.3	Web Log Analytics	268
9.5.4	Intrusion Detection Applications	268
9.5.5	Biological and Medical Applications	268
9.5.6	Earth Science Applications	268
9.6	Summary	269
9.7	Bibliographic Notes	269
9.8	Exercises	270
10	Data Classification	271
10.1	Introduction	271
10.2	Feature Selection for Classification	273
10.2.1	Filter Models	274
10.2.1.1	Gini Index	274
10.2.1.2	Entropy	275
10.2.1.3	Fisher Score	275
10.2.1.4	Fisher's Linear Discriminant	276
10.2.2	Wrapper Models	277
10.2.3	Embedded Models	278
10.3	Decision Trees	278
10.3.1	Split Criteria	281
10.3.2	Stopping Criterion and Pruning	283
10.3.3	Practical Issues	284
10.4	Rule-based Classifiers	284
10.4.1	Rule Generation from Decision Trees	286
10.4.2	Sequential Covering Algorithms	287
10.4.2.1	Learn-One-Rule	287
10.4.3	Rule Pruning	290
10.4.4	Associative Classifiers	290
10.5	Probabilistic Classifiers	291
10.5.1	Naive Bayes Classifier	291
10.5.1.1	The Ranking Model for Classification	294
10.5.1.2	Discussion of the Naive Assumption	295
10.5.2	Logistic Regression	295
10.5.2.1	Training a Logistic Regression Classifier	297
10.5.2.2	Relationship with Other Linear Models	298
10.6	Support Vector Machines	298

10.6.1	Support Vector Machines for Linearly Separable Data	298
10.6.1.1	Solving the Lagrangian Dual	303
10.6.2	Support Vector Machines with Soft Margin for Nonseparable Data	304
10.6.2.1	Comparison with other Linear Models	306
10.6.3	Nonlinear Support Vector Machines	306
10.6.4	The Kernel Trick	307
10.6.4.1	Other Applications of Kernel Methods	309
10.7	Neural Networks	311
10.7.1	Single-Layer Neural Network: The Perceptron	311
10.7.2	Multilayer Neural Networks	313
10.7.3	Comparing Various Linear Models	315
10.8	Instance-based Learning	316
10.8.1	Design Variations of Nearest Neighbor Classifiers	316
10.8.1.1	Unsupervised Mahalanobis Metric	317
10.8.1.2	Nearest Neighbors with Linear Discriminant Analysis	317
10.9	Classifier Evaluation	319
10.9.1	Methodological Issues	319
10.9.1.1	Holdout	320
10.9.1.2	Cross-Validation	320
10.9.1.3	Bootstrap	321
10.9.2	Quantification Issues	322
10.9.2.1	Output as Class Labels	322
10.9.2.2	Output as Numerical Score	323
10.10	Summary	326
10.11	Bibliographic Notes	326
10.12	Exercises	327
11	Data Classification: Advanced Concepts	331
11.1	Introduction	331
11.2	Multiclass Learning	332
11.3	Rare Class Learning	333
11.3.1	Example Re-weighting	334
11.3.2	Sampling Methods	335
11.3.2.1	Relationship between Weighting and Sampling	336
11.3.2.2	Synthetic Over-sampling: SMOTE	336
11.4	Scalable Classification	336
11.4.1	Scalable Decision Trees	337
11.4.1.1	RainForest	337
11.4.1.2	BOAT	337
11.4.2	Scalable Support Vector Machines	337
11.5	Regression Modeling with Numeric Classes	339
11.5.1	Linear Regression	339
11.5.1.1	Relationship with Fisher's Linear Discriminant	341
11.5.2	Principal Component Regression	342
11.5.3	Generalized Linear Models	343
11.5.4	Nonlinear and Polynomial Regression	344
11.5.5	From Decision Trees to Regression Trees	345
11.5.6	Assessing Model Effectiveness	346
11.6	Semisupervised Learning	346
11.6.1	Generic Meta-Algorithms	348
11.6.1.1	Self-Training	348

11.6.1.2	Co-Training	348
11.6.2	Specific Variations of Classification Algorithms	349
11.6.2.1	Semisupervised Bayes Classification with EM	349
11.6.2.2	Transductive Support Vector Machines	351
11.6.3	Graph-based Semisupervised Learning	352
11.6.4	Discussion of Semisupervised Learning	353
11.7	Active Learning	353
11.7.1	Heterogeneity-based Models	355
11.7.1.1	Uncertainty Sampling	355
11.7.1.2	Query-by-Committee	356
11.7.1.3	Expected Model Change	356
11.7.2	Performance-based Models	357
11.7.2.1	Expected Error Reduction	357
11.7.2.2	Expected Variance Reduction	358
11.7.3	Representativeness-based Models	358
11.8	Ensemble Methods	358
11.8.1	Why does Ensemble Analysis Work?	360
11.8.2	Formal Statement of Bias-Variance Trade-off	362
11.8.3	Specific Instantiations of Ensemble Learning	364
11.8.3.1	Bagging	364
11.8.3.2	Random Forests	365
11.8.3.3	Boosting	366
11.8.3.4	Bucket of Models	368
11.8.3.5	Stacking	368
11.9	Summary	369
11.10	Bibliographic Notes	370
11.11	Exercises	370
12	Mining Data Streams	373
12.1	Introduction	373
12.2	Synopsis Data Structures for Streams	375
12.2.1	Reservoir Sampling	375
12.2.1.1	Handling Concept Drift	376
12.2.1.2	Useful Theoretical Bounds for Sampling	377
12.2.2	Synopsis Structures for the Massive-Domain Scenario	381
12.2.2.1	Bloom Filter	382
12.2.2.2	Count-Min Sketch	386
12.2.2.3	AMS Sketch	389
12.2.2.4	Flajolet-Martin Algorithm for Distinct Element Counting	391
12.3	Frequent Pattern Mining in Data Streams	392
12.3.1	Leveraging Synopsis Structures	392
12.3.1.1	Reservoir Sampling	393
12.3.1.2	Sketches	393
12.3.2	Lossy Counting Algorithm	393
12.4	Clustering Data Streams	394
12.4.1	STREAM Algorithm	394
12.4.2	CluStream Algorithm	396
12.4.2.1	Micro-cluster Definition	396
12.4.2.2	Micro-clustering Algorithm	397
12.4.2.3	Pyramidal Time Frame	398
12.4.3	Massive-Domain Stream Clustering	400

12.5 Streaming Outlier Detection	400
12.5.1 Individual Data Points as Outliers	401
12.5.2 Aggregate Change Points as Outliers	402
12.6 Streaming Classification	404
12.6.1 VFDT Family	404
12.6.2 Supervised Micro-cluster Approach	406
12.6.3 Ensemble Method	407
12.6.4 Massive-Domain Streaming Classification	407
12.7 Summary	408
12.8 Bibliographic Notes	408
12.9 Exercises	408
13 Mining Text Data	411
13.1 Introduction	411
13.2 Document Preparation and Similarity Computation	413
13.2.1 Document Normalization and Similarity Computation	413
13.2.2 Specialized Preprocessing for Web Documents	415
13.3 Specialized Clustering Methods for Text	416
13.3.1 Representative-based Algorithms	416
13.3.1.1 Scatter/Gather Approach	416
13.3.2 Probabilistic Algorithms	418
13.3.3 Simultaneous Document and Word Cluster Discovery	419
13.3.3.1 Co-clustering	420
13.4 Topic Modeling	422
13.4.1 Use in Dimensionality Reduction and Comparison with Latent Semantic Analysis	425
13.4.2 Use in Clustering and Comparison with Probabilistic Clustering	427
13.4.3 Limitations of PLSA	427
13.5 Specialized Classification Methods for Text	428
13.5.1 Instance-based Classifiers	428
13.5.1.1 Leveraging Latent Semantic Analysis	428
13.5.1.2 Centroid-based Classification	429
13.5.1.3 Rocchio Classification	429
13.5.2 Bayes Classifiers	430
13.5.2.1 Multinomial Bayes Model	430
13.5.3 SVM Classifiers for High-dimensional and Sparse Data	432
13.6 Novelty and First-Story Detection	434
13.6.1 Micro-clustering Method	435
13.7 Summary	435
13.8 Bibliographic Notes	436
13.9 Exercises	436
14 Mining Time-Series Data	439
14.1 Introduction	439
14.2 Time-Series Preparation and Similarity	441
14.2.1 Handling Missing Values	441
14.2.2 Noise Removal	441
14.2.3 Normalization	443
14.2.4 Data Transformation and Reduction	444
14.2.4.1 Discrete Wavelet Transform	444
14.2.4.2 Discrete Fourier Transform	444

14.2.4.3	Symbolic Aggregate Approximation (SAX)	445
14.2.5	Time-Series Similarity Measures	446
14.3	Time-Series Forecasting	446
14.3.1	Autoregressive Models	448
14.3.2	Autoregressive Moving Average Models	450
14.3.3	Multivariate Forecasting with Hidden Variables	451
14.4	Time-Series Motifs	453
14.4.1	Distance-based Motifs	454
14.4.2	Transformation to Sequential Pattern Mining	456
14.4.3	Periodic Patterns	457
14.5	Time-Series Clustering	458
14.5.1	Online Clustering of Co-evolving Series	458
14.5.2	Shape-based Clustering	460
14.5.2.1	k -Means	461
14.5.2.2	k -Medoids	462
14.5.2.3	Hierarchical Methods	462
14.5.2.4	Graph-based Methods	462
14.6	Time-Series Outlier Detection	462
14.6.1	Point Outliers	463
14.6.2	Shape Outliers	464
14.7	Time-Series Classification	465
14.7.1	Supervised Event Detection	466
14.7.2	Whole-Series Classification	468
14.7.2.1	Wavelet-based Rules	469
14.7.2.2	Nearest Neighbor Classifier	469
14.7.2.3	Graph-based Methods	470
14.8	Summary	470
14.9	Bibliographic Notes	470
14.10	Exercises	471
15	Mining Discrete Sequences	473
15.1	Introduction	473
15.2	Sequential Pattern Mining	474
15.2.1	Frequent Patterns to Frequent Sequences	477
15.2.2	Constrained Sequential Pattern Mining	480
15.3	Sequence Clustering	481
15.3.1	Distance-based Methods	482
15.3.2	Graph-based Methods	482
15.3.3	Subsequence-based Clustering	482
15.3.4	Probabilistic Clustering	483
15.3.4.1	Markovian Similarity-based Algorithm: CLUSEQ	483
15.3.4.2	Mixture of Hidden Markov Models	486
15.4	Outlier Detection in Sequences	487
15.4.1	Position Outliers	487
15.4.1.1	Efficiency Issues: Probabilistic Suffix Trees	490
15.4.2	Combination Outliers	491
15.4.2.1	Distance-based Models	492
15.4.2.2	Frequency-based Models	493
15.5	Hidden Markov Models	494
15.5.1	Formal Definition and Techniques for HMMs	496
15.5.2	Evaluation: Computing the Fit Probability for Observed Sequence	497

15.5.3	Explanation: Determining the Most Likely State Sequence for Observed Sequence	498
15.5.4	Training: Baum-Welch Algorithm	499
15.5.5	Applications	500
15.6	Sequence Classification	501
15.6.1	Nearest Neighbor Classifier	501
15.6.2	Graph-based Methods	501
15.6.3	Rule-based Methods	502
15.6.4	Kernel Support Vector Machines	503
15.6.4.1	Bag-of-Words Kernel	503
15.6.4.2	Spectrum Kernel	503
15.6.4.3	Weighted Degree Kernel	504
15.6.5	Probabilistic Methods: Hidden Markov Models	504
15.7	Summary	505
15.8	Bibliographic Notes	505
15.9	Exercises	506
16	Mining Spatial Data	509
16.1	Introduction	509
16.2	Mining with Contextual Spatial Attributes	510
16.2.1	Shape to Time-Series Transformation	512
16.2.2	Spatial to Multidimensional Transformation with Wavelets	515
16.2.3	Spatial Co-location Patterns	516
16.2.4	Clustering Shapes	517
16.2.5	Outlier Detection	518
16.2.5.1	Point Outliers	518
16.2.5.2	Shape Outliers	520
16.2.6	Classification of Shapes	521
16.3	Trajectory Mining	522
16.3.1	Equivalence of Trajectories and Multivariate Time Series	522
16.3.2	Converting Trajectories to Multidimensional Data	523
16.3.3	Trajectory Pattern Mining	524
16.3.3.1	Frequent Trajectory Paths	524
16.3.3.2	Co-location Patterns	526
16.3.4	Trajectory Clustering	526
16.3.4.1	Computing Similarity between Trajectories	526
16.3.4.2	Similarity-based Clustering Methods	527
16.3.4.3	Trajectory Clustering as a Sequence Clustering Problem	528
16.3.5	Trajectory Outlier Detection	528
16.3.5.1	Distance-based Methods	529
16.3.5.2	Sequence-based Methods	529
16.3.6	Trajectory Classification	530
16.3.6.1	Distance-based Methods	530
16.3.6.2	Sequence-based Methods	530
16.4	Summary	531
16.5	Bibliographic Notes	531
16.6	Exercises	532

17 Mining Graph Data	533
17.1 Introduction	533
17.2 Matching and Distance Computation in Graphs	535
17.2.1 Ullman's Algorithm for Subgraph Isomorphism	537
17.2.1.1 Algorithm Variations and Refinements	539
17.2.2 Maximum Common Subgraph Problem	540
17.2.3 Graph Matching Methods for Distance Computation	541
17.2.3.1 Maximum Common Subgraph-based Distances	541
17.2.3.2 Graph Edit Distance	542
17.3 Transformation-based Distance Computation	546
17.3.1 Frequent Substructure-based Transformation and Distance Computation	546
17.3.2 Topological Descriptors	547
17.3.3 Kernel-based Transformations and Computation	549
17.3.3.1 Random-Walk Kernels	549
17.3.3.2 Shortest-Path Kernels	550
17.4 Frequent Substructure Mining in Graphs	550
17.4.1 Node-based Join Growth	552
17.4.2 Edge-based Join Growth	554
17.4.3 Frequent Pattern Mining to Graph Pattern Mining	554
17.5 Graph Clustering	554
17.5.1 Distance-based Methods	555
17.5.2 Frequent Substructure-based Methods	555
17.5.2.1 Generic Transformational Approach	556
17.5.2.2 XProj: Direct Clustering with Frequent Subgraph Discovery	556
17.6 Graph Classification	558
17.6.1 Distance-based Methods	558
17.6.2 Frequent Substructure-based Methods	558
17.6.2.1 Generic Transformational Approach	559
17.6.2.2 XRules: A Rule-based Approach	559
17.6.3 Kernel Support Vector Machines	560
17.7 Summary	560
17.8 Bibliographic Notes	561
17.9 Exercises	562
18 Mining Web Data	565
18.1 Introduction	565
18.2 Web Crawling and Resource Discovery	567
18.2.1 A Basic Crawler Algorithm	567
18.2.2 Preferential Crawlers	569
18.2.3 Multiple Threads	569
18.2.4 Combatting Spider Traps	569
18.2.5 Shingling for Near Duplicate Detection	570
18.3 Search Engine Indexing and Query Processing	570
18.4 Ranking Algorithms	573
18.4.1 PageRank	573
18.4.1.1 Topic-Sensitive PageRank	576
18.4.1.2 SimRank	577
18.4.2 HITS	578
18.5 Recommender Systems	579
18.5.1 Content-based Recommendations	581

18.5.2	Neighborhood-based Methods for Collaborative Filtering	582
18.5.2.1	User-based Similarity with Ratings	582
18.5.2.2	Item-based Similarity with Ratings	583
18.5.3	Graph-based Methods	583
18.5.4	Clustering Methods	585
18.5.4.1	Adapting k -Means Clustering	585
18.5.4.2	Adapting Co-Clustering	585
18.5.5	Latent Factor Models	586
18.5.5.1	Singular Value Decomposition	587
18.5.5.2	Matrix Factorization	587
18.6	Web Usage Mining	588
18.6.1	Data Preprocessing	589
18.6.2	Applications	589
18.7	Summary	590
18.8	Bibliographic Notes	591
18.9	Exercises	591
19	Social Network Analysis	593
19.1	Introduction	593
19.2	Social Networks: Preliminaries and Properties	594
19.2.1	Homophily	595
19.2.2	Triadic Closure and Clustering Coefficient	595
19.2.3	Dynamics of Network Formation	595
19.2.4	Power-Law Degree Distributions	597
19.2.5	Measures of Centrality and Prestige	597
19.2.5.1	Degree Centrality and Prestige	597
19.2.5.2	Closeness Centrality and Proximity Prestige	598
19.2.5.3	Betweenness Centrality	600
19.2.5.4	Rank Centrality and Prestige	600
19.3	Community Detection	601
19.3.1	Kernighan-Lin Algorithm	602
19.3.1.1	Speeding up Kernighan-Lin	604
19.3.2	Girvan-Newman Algorithm	604
19.3.3	Multilevel Graph Partitioning: METIS	607
19.3.4	Spectral Clustering	610
19.3.4.1	Important Observations and Intuitions	613
19.4	Collective Classification	614
19.4.1	Iterative Classification Algorithm	615
19.4.2	Label Propagation with Random Walks	616
19.4.2.1	Iterative Label Propagation: The Spectral Interpretation	619
19.4.3	Supervised Spectral Methods	619
19.4.3.1	Supervised Feature Generation with Spectral Embedding	620
19.4.3.2	Graph Regularization Approach	620
19.4.3.3	Connections with Random-Walk Methods	622
19.5	Link Prediction	623
19.5.1	Neighborhood-based Measures	623
19.5.2	Katz Measure	624
19.5.3	Random Walk-based Measures	625
19.5.4	Link Prediction as a Classification Problem	626
19.5.5	Link Prediction as a Missing Value Estimation Problem	627
19.5.6	Discussion	627

19.6 Social Influence Analysis	627
19.6.1 Linear Threshold Model	629
19.6.2 Independent Cascade Model	629
19.6.3 Influence Function Evaluation	630
19.7 Summary	630
19.8 Bibliographic Notes	631
19.9 Exercises	632
20 Privacy-Preserving Data Mining	635
20.1 Introduction	635
20.2 Privacy during Data Collection	636
20.2.1 Reconstructing Aggregate Distributions	637
20.2.2 Leveraging Aggregate Distributions for Data Mining	639
20.3 Privacy-Preserving Data Publishing	639
20.3.1 The k -anonymity Model	641
20.3.1.1 Samarati's Algorithm	645
20.3.1.2 Incognito	646
20.3.1.3 Mondrian Multidimensional k -Anonymity	649
20.3.1.4 Synthetic Data Generation: Condensation-based Approach	651
20.3.2 The ℓ -diversity Model	653
20.3.3 The t -closeness Model	655
20.3.4 The Curse of Dimensionality	658
20.4 Output Privacy	658
20.5 Distributed Privacy	659
20.6 Summary	661
20.7 Bibliographic Notes	661
20.8 Exercises	663

**Computers connected to subscribing institutions can
download book from the following clickable URL:
<http://rd.springer.com/book/10.1007/978-3-319-14142-8>**

Preface

“Data is the new oil.” – Clive Humby

The field of data mining has seen rapid strides over the past two decades, especially from the perspective of the computer science community. While data analysis has been studied extensively in the conventional field of probability and statistics, *data mining* is a term coined by the computer science-oriented community. For computer scientists, issues such as scalability, usability, and computational implementation are extremely important.

The emergence of data science as a discipline requires the development of a book that goes beyond the traditional focus of books on only the fundamental data mining courses. Recent years have seen the emergence of the job description of “data scientists,” who try to glean knowledge from vast amounts of data. In typical applications, the data types are so heterogeneous and diverse that the fundamental methods discussed for a multidimensional data type may not be effective. Therefore, more emphasis needs to be placed on the different data types and the applications which arise in the context of these different data types. A comprehensive data mining book must explore the different aspects of data mining, starting from the fundamentals, and then explore the complex data types, and their relationships with the fundamental techniques. While fundamental techniques form an excellent basis for the further study of data mining, they do not provide a complete picture of the true complexity of data analysis. This book studies these advanced topics without compromising the presentation of fundamental methods. Therefore, this book may be used for both introductory and advanced data mining courses. Until now, no single book has addressed all these topics in a comprehensive and integrated way.

The textbook assumes a basic knowledge of probability, statistics, and linear algebra, which is taught in most undergraduate curricula of science and engineering disciplines. Therefore, the book can also be used by industrial practitioners, who have a working knowledge of these basic skills. While stronger mathematical background is helpful for the more advanced chapters, it is not a pre-requisite. Special chapters are also devoted to different aspects of data mining, such as the text data, time-series data, discrete sequences, and graphs. This kind of specialized treatment is intended to capture the wide diversity of problem domains in which a data mining problem might arise.

The chapters of this book fall into one of three categories:

- **The fundamental chapters:** Data mining has four main “super-problems,” which correspond to clustering, classification, association pattern mining, and outlier analysis. These problems are so important because they are used repeatedly as building blocks in the context of a wide variety of data mining applications. As a result, a large

amount of emphasis has been placed by data mining researchers and practitioners to design effective and efficient methods for these problems. These chapters comprehensively discuss the vast diversity of methods used by the data mining community in the context of these super-problems.

- **Domain chapters:** These chapters discuss the specific methods used for different *domains* of data such as text data, time-series data, sequence data, graph data, and spatial data. Many of these chapters can also be considered application chapters, because they explore the specific characteristics of the problem in a particular domain.
- **Application chapters:** Advancements in hardware technology and software platforms, have lead to a number of data-intensive applications such as streaming systems, Web mining, social networks, and privacy-preservation. These topics are studied in detail in these chapters. The domain chapters are also focused on many different kinds of applications that arise in the context of those data types.

Suggestions for the Instructor

The book was specifically written to enable the teaching of both the basic data mining and advanced data mining courses from a single book. It can be used to offer various types of data mining courses with different emphases. Specifically, the courses that could be offered with various chapters are as follows:

- **Basic data mining course and fundamentals:** The basic data mining course should focus on the fundamentals of data mining. Chapters 1, 2, 3, 4, 6, 8, and 10 can be covered. In fact, the material in these chapters is more than what is possible to teach in a single course. Therefore, instructors may need to select topics of their interest from these chapters. Some portions of Chapters 5, 7, 9, and 11 can also be covered, although these chapters are really meant for an advanced course.
- **Advanced course (fundamentals):** Such a course would cover advanced topics on the fundamentals of data mining and assume that the student is already familiar with Chapters 1 through 3, and parts of Chapters 4, 6, 8, and 10. The course can then focus on Chapters 5, 7, 9, and 11. Topics such as ensemble analysis are useful for the advanced course. Furthermore, some topics from Chapters 4, 6, 8, and 10, which were not covered in the basic course, can be used. In addition, Chapter 20 on privacy can be offered.
- **Advanced course (data types):** Advanced topics such as text mining, time series, sequences, graphs, and spatial data may be covered. The material should focus on Chapters 13, 14, 15, 16, and 17. Some parts of Chapter 19 (e.g., graph clustering) and Chapter 12 (data streaming) can also be used.
- **Advanced course (applications):** An application course overlaps with a data type course but has a different focus. For example, the focus in an application-centered course would be more on the modeling aspect than the algorithmic aspect. Therefore, the same materials in Chapters 13, 14, 15, 16, and 17 can be used while skipping specific details of algorithms. With less focus on specific algorithms, these chapters can be covered fairly quickly. The remaining time should be allocated to three very important chapters on data streams (Chapter 12), Web mining (Chapter 18), and social network analysis (Chapter 19).

The book is written in a simple style to make it accessible to undergraduate students and industrial practitioners with a limited mathematical background. Thus, the book will serve

both as an introductory text and as an advanced text for students, industrial practitioners, and researchers.

Throughout this book, a vector or a multidimensional data point (including categorical attributes), is annotated with a bar, such as \bar{X} or \bar{y} . A vector or multidimensional point may be denoted by either small letters or capital letters, as long as it has a bar. Vector dot products are denoted by centered dots, such as $\bar{X} \cdot \bar{Y}$. A matrix is denoted in capital letters without a bar, such as R . Throughout the book, the $n \times d$ data matrix is denoted by D , with n points and d dimensions. The individual data points in D are therefore d -dimensional row vectors. On the other hand, vectors with one component for each data point are usually n -dimensional column vectors. An example is the n -dimensional column vector \bar{y} of class variables of n data points.