

Charu C. Aggarwal

# Data Mining

The Textbook



Springer

Charu C. Aggarwal

# Data Mining

The Textbook

 Springer

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

A solution manual for this book is available on Springer.com.

ISBN 978-3-319-14141-1                      ISBN 978-3-319-14142-8 (eBook)  
DOI 10.1007/978-3-319-14142-8

Library of Congress Control Number: 2015930833

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

---

---

# Contents

---

---

<b>1</b>	<b>An Introduction to Data Mining</b>	<b>1</b>
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	Nondependency-Oriented Data	7
1.3.1.1	Quantitative Multidimensional Data	7
1.3.1.2	Categorical and Mixed Attribute Data	8
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	14
1.4.1	Association Pattern Mining	15
1.4.2	Data Clustering	16
1.4.3	Outlier Detection	17
1.4.4	Data Classification	18
1.4.5	Impact of Complex Data Types on Problem Definitions	19
1.4.5.1	Pattern Mining with Complex Data Types	20
1.4.5.2	Clustering with Complex Data Types	20
1.4.5.3	Outlier Detection with Complex Data Types	21
1.4.5.4	Classification with Complex Data Types	21
1.5	Scalability Issues and the Streaming Scenario	21
1.6	A Stroll Through Some Application Scenarios	22
1.6.1	Store Product Placement	22
1.6.2	Customer Recommendations	23
1.6.3	Medical Diagnosis	23
1.6.4	Web Log Anomalies	24
1.7	Summary	24

1.8	Bibliographic Notes . . . . .	25
1.9	Exercises . . . . .	25
<b>2</b>	<b>Data Preparation</b>	<b>27</b>
2.1	Introduction . . . . .	27
2.2	Feature Extraction and Portability . . . . .	28
2.2.1	Feature Extraction . . . . .	28
2.2.2	Data Type Portability . . . . .	30
2.2.2.1	Numeric to Categorical Data: Discretization . . . . .	30
2.2.2.2	Categorical to Numeric Data: Binarization . . . . .	31
2.2.2.3	Text to Numeric Data . . . . .	31
2.2.2.4	Time Series to Discrete Sequence Data . . . . .	32
2.2.2.5	Time Series to Numeric Data . . . . .	32
2.2.2.6	Discrete Sequence to Numeric Data . . . . .	33
2.2.2.7	Spatial to Numeric Data . . . . .	33
2.2.2.8	Graphs to Numeric Data . . . . .	33
2.2.2.9	Any Type to Graphs for Similarity-Based Applications . . . . .	33
2.3	Data Cleaning . . . . .	34
2.3.1	Handling Missing Entries . . . . .	35
2.3.2	Handling Incorrect and Inconsistent Entries . . . . .	36
2.3.3	Scaling and Normalization . . . . .	37
2.4	Data Reduction and Transformation . . . . .	37
2.4.1	Sampling . . . . .	38
2.4.1.1	Sampling for Static Data . . . . .	38
2.4.1.2	Reservoir Sampling for Data Streams . . . . .	39
2.4.2	Feature Subset Selection . . . . .	40
2.4.3	Dimensionality Reduction with Axis Rotation . . . . .	41
2.4.3.1	Principal Component Analysis . . . . .	42
2.4.3.2	Singular Value Decomposition . . . . .	44
2.4.3.3	Latent Semantic Analysis . . . . .	47
2.4.3.4	Applications of PCA and SVD . . . . .	48
2.4.4	Dimensionality Reduction with Type Transformation . . . . .	49
2.4.4.1	Haar Wavelet Transform . . . . .	50
2.4.4.2	Multidimensional Scaling . . . . .	55
2.4.4.3	Spectral Transformation and Embedding of Graphs . . . . .	57
2.5	Summary . . . . .	59
2.6	Bibliographic Notes . . . . .	60
2.7	Exercises . . . . .	61
<b>3</b>	<b>Similarity and Distances</b>	<b>63</b>
3.1	Introduction . . . . .	63
3.2	Multidimensional Data . . . . .	64
3.2.1	Quantitative Data . . . . .	64
3.2.1.1	Impact of Domain-Specific Relevance . . . . .	65
3.2.1.2	Impact of High Dimensionality . . . . .	65
3.2.1.3	Impact of Locally Irrelevant Features . . . . .	66
3.2.1.4	Impact of Different $L_p$ -Norms . . . . .	67
3.2.1.5	Match-Based Similarity Computation . . . . .	68
3.2.1.6	Impact of Data Distribution . . . . .	69

3.2.1.7	Nonlinear Distributions: ISOMAP . . . . .	70
3.2.1.8	Impact of Local Data Distribution . . . . .	72
3.2.1.9	Computational Considerations . . . . .	73
3.2.2	Categorical Data . . . . .	74
3.2.3	Mixed Quantitative and Categorical Data . . . . .	75
3.3	Text Similarity Measures . . . . .	75
3.3.1	Binary and Set Data . . . . .	77
3.4	Temporal Similarity Measures . . . . .	77
3.4.1	Time-Series Similarity Measures . . . . .	77
3.4.1.1	Impact of Behavioral Attribute Normalization . . . . .	78
3.4.1.2	$L_p$ -Norm . . . . .	79
3.4.1.3	Dynamic Time Warping Distance . . . . .	79
3.4.1.4	Window-Based Methods . . . . .	82
3.4.2	Discrete Sequence Similarity Measures . . . . .	82
3.4.2.1	Edit Distance . . . . .	82
3.4.2.2	Longest Common Subsequence . . . . .	84
3.5	Graph Similarity Measures . . . . .	85
3.5.1	Similarity between Two Nodes in a Single Graph . . . . .	85
3.5.1.1	Structural Distance-Based Measure . . . . .	85
3.5.1.2	Random Walk-Based Similarity . . . . .	86
3.5.2	Similarity Between Two Graphs . . . . .	86
3.6	Supervised Similarity Functions . . . . .	87
3.7	Summary . . . . .	88
3.8	Bibliographic Notes . . . . .	89
3.9	Exercises . . . . .	90
<b>4</b>	<b>Association Pattern Mining</b>	<b>93</b>
4.1	Introduction . . . . .	93
4.2	The Frequent Pattern Mining Model . . . . .	94
4.3	Association Rule Generation Framework . . . . .	97
4.4	Frequent Itemset Mining Algorithms . . . . .	99
4.4.1	Brute Force Algorithms . . . . .	99
4.4.2	The Apriori Algorithm . . . . .	100
4.4.2.1	Efficient Support Counting . . . . .	102
4.4.3	Enumeration-Tree Algorithms . . . . .	103
4.4.3.1	Enumeration-Tree-Based Interpretation of Apriori . . . . .	105
4.4.3.2	TreeProjection and DepthProject . . . . .	106
4.4.3.3	Vertical Counting Methods . . . . .	110
4.4.4	Recursive Suffix-Based Pattern Growth Methods . . . . .	112
4.4.4.1	Implementation with Arrays but No Pointers . . . . .	114
4.4.4.2	Implementation with Pointers but No FP-Tree . . . . .	114
4.4.4.3	Implementation with Pointers and FP-Tree . . . . .	116
4.4.4.4	Trade-offs with Different Data Structures . . . . .	118
4.4.4.5	Relationship Between FP-Growth and Enumeration-Tree Methods . . . . .	119
4.5	Alternative Models: Interesting Patterns . . . . .	122
4.5.1	Statistical Coefficient of Correlation . . . . .	123
4.5.2	$\chi^2$ Measure . . . . .	123
4.5.3	Interest Ratio . . . . .	124

4.5.4	Symmetric Confidence Measures . . . . .	124
4.5.5	Cosine Coefficient on Columns . . . . .	125
4.5.6	Jaccard Coefficient and the Min-hash Trick . . . . .	125
4.5.7	Collective Strength . . . . .	126
4.5.8	Relationship to Negative Pattern Mining . . . . .	127
4.6	Useful Meta-algorithms . . . . .	127
4.6.1	Sampling Methods . . . . .	128
4.6.2	Data Partitioned Ensembles . . . . .	128
4.6.3	Generalization to Other Data Types . . . . .	129
4.6.3.1	Quantitative Data . . . . .	129
4.6.3.2	Categorical Data . . . . .	129
4.7	Summary . . . . .	129
4.8	Bibliographic Notes . . . . .	130
4.9	Exercises . . . . .	132
<b>5</b>	<b>Association Pattern Mining: Advanced Concepts</b>	<b>135</b>
5.1	Introduction . . . . .	135
5.2	Pattern Summarization . . . . .	136
5.2.1	Maximal Patterns . . . . .	136
5.2.2	Closed Patterns . . . . .	137
5.2.3	Approximate Frequent Patterns . . . . .	139
5.2.3.1	Approximation in Terms of Transactions . . . . .	139
5.2.3.2	Approximation in Terms of Itemsets . . . . .	140
5.3	Pattern Querying . . . . .	141
5.3.1	Preprocess-once Query-many Paradigm . . . . .	141
5.3.1.1	Leveraging the Itemset Lattice . . . . .	142
5.3.1.2	Leveraging Data Structures for Querying . . . . .	143
5.3.2	Pushing Constraints into Pattern Mining . . . . .	146
5.4	Putting Associations to Work: Applications . . . . .	147
5.4.1	Relationship to Other Data Mining Problems . . . . .	147
5.4.1.1	Application to Classification . . . . .	147
5.4.1.2	Application to Clustering . . . . .	148
5.4.1.3	Applications to Outlier Detection . . . . .	148
5.4.2	Market Basket Analysis . . . . .	148
5.4.3	Demographic and Profile Analysis . . . . .	148
5.4.4	Recommendations and Collaborative Filtering . . . . .	149
5.4.5	Web Log Analysis . . . . .	149
5.4.6	Bioinformatics . . . . .	149
5.4.7	Other Applications for Complex Data Types . . . . .	150
5.5	Summary . . . . .	150
5.6	Bibliographic Notes . . . . .	151
5.7	Exercises . . . . .	152
<b>6</b>	<b>Cluster Analysis</b>	<b>153</b>
6.1	Introduction . . . . .	153
6.2	Feature Selection for Clustering . . . . .	154
6.2.1	Filter Models . . . . .	155
6.2.1.1	Term Strength . . . . .	155
6.2.1.2	Predictive Attribute Dependence . . . . .	155

6.2.1.3	Entropy . . . . .	156
6.2.1.4	Hopkins Statistic . . . . .	157
6.2.2	Wrapper Models . . . . .	158
6.3	Representative-Based Algorithms . . . . .	159
6.3.1	The $k$ -Means Algorithm . . . . .	162
6.3.2	The Kernel $k$ -Means Algorithm . . . . .	163
6.3.3	The $k$ -Medians Algorithm . . . . .	164
6.3.4	The $k$ -Medoids Algorithm . . . . .	164
6.4	Hierarchical Clustering Algorithms . . . . .	166
6.4.1	Bottom-Up Agglomerative Methods . . . . .	167
6.4.1.1	Group-Based Statistics . . . . .	169
6.4.2	Top-Down Divisive Methods . . . . .	172
6.4.2.1	Bisecting $k$ -Means . . . . .	173
6.5	Probabilistic Model-Based Algorithms . . . . .	173
6.5.1	Relationship of EM to $k$ -means and Other Representative Methods . . . . .	176
6.6	Grid-Based and Density-Based Algorithms . . . . .	178
6.6.1	Grid-Based Methods . . . . .	179
6.6.2	DBSCAN . . . . .	181
6.6.3	DENCLUE . . . . .	184
6.7	Graph-Based Algorithms . . . . .	187
6.7.1	Properties of Graph-Based Algorithms . . . . .	189
6.8	Non-negative Matrix Factorization . . . . .	191
6.8.1	Comparison with Singular Value Decomposition . . . . .	194
6.9	Cluster Validation . . . . .	195
6.9.1	Internal Validation Criteria . . . . .	196
6.9.1.1	Parameter Tuning with Internal Measures . . . . .	198
6.9.2	External Validation Criteria . . . . .	198
6.9.3	General Comments . . . . .	201
6.10	Summary . . . . .	201
6.11	Bibliographic Notes . . . . .	201
6.12	Exercises . . . . .	202
<b>7</b>	<b>Cluster Analysis: Advanced Concepts</b>	<b>205</b>
7.1	Introduction . . . . .	205
7.2	Clustering Categorical Data . . . . .	206
7.2.1	Representative-Based Algorithms . . . . .	207
7.2.1.1	$k$ -Modes Clustering . . . . .	208
7.2.1.2	$k$ -Medoids Clustering . . . . .	209
7.2.2	Hierarchical Algorithms . . . . .	209
7.2.2.1	ROCK . . . . .	209
7.2.3	Probabilistic Algorithms . . . . .	211
7.2.4	Graph-Based Algorithms . . . . .	212
7.3	Scalable Data Clustering . . . . .	212
7.3.1	CLARANS . . . . .	213
7.3.2	BIRCH . . . . .	214
7.3.3	CURE . . . . .	216
7.4	High-Dimensional Clustering . . . . .	217
7.4.1	CLIQUE . . . . .	219
7.4.2	PROCLUS . . . . .	220



7.4.3	ORCLUS . . . . .	222
7.5	Semisupervised Clustering . . . . .	224
7.5.1	Pointwise Supervision . . . . .	225
7.5.2	Pairwise Supervision . . . . .	226
7.6	Human and Visually Supervised Clustering . . . . .	227
7.6.1	Modifications of Existing Clustering Algorithms . . . . .	228
7.6.2	Visual Clustering . . . . .	228
7.7	Cluster Ensembles . . . . .	231
7.7.1	Selecting Different Ensemble Components . . . . .	231
7.7.2	Combining Different Ensemble Components . . . . .	232
7.7.2.1	Hypergraph Partitioning Algorithm . . . . .	232
7.7.2.2	Meta-clustering Algorithm . . . . .	232
7.8	Putting Clustering to Work: Applications . . . . .	233
7.8.1	Applications to Other Data Mining Problems . . . . .	233
7.8.1.1	Data Summarization . . . . .	233
7.8.1.2	Outlier Analysis . . . . .	233
7.8.1.3	Classification . . . . .	233
7.8.1.4	Dimensionality Reduction . . . . .	234
7.8.1.5	Similarity Search and Indexing . . . . .	234
7.8.2	Customer Segmentation and Collaborative Filtering . . . . .	234
7.8.3	Text Applications . . . . .	234
7.8.4	Multimedia Applications . . . . .	234
7.8.5	Temporal and Sequence Applications . . . . .	234
7.8.6	Social Network Analysis . . . . .	235
7.9	Summary . . . . .	235
7.10	Bibliographic Notes . . . . .	235
7.11	Exercises . . . . .	236
<b>8</b>	<b>Outlier Analysis . . . . .</b>	<b>237</b>
8.1	Introduction . . . . .	237
8.2	Extreme Value Analysis . . . . .	239
8.2.1	Univariate Extreme Value Analysis . . . . .	240
8.2.2	Multivariate Extreme Values . . . . .	242
8.2.3	Depth-Based Methods . . . . .	243
8.3	Probabilistic Models . . . . .	244
8.4	Clustering for Outlier Detection . . . . .	246
8.5	Distance-Based Outlier Detection . . . . .	248
8.5.1	Pruning Methods . . . . .	249
8.5.1.1	Sampling Methods . . . . .	249
8.5.1.2	Early Termination Trick with Nested Loops . . . . .	250
8.5.2	Local Distance Correction Methods . . . . .	251
8.5.2.1	Local Outlier Factor (LOF) . . . . .	252
8.5.2.2	Instance-Specific Mahalanobis Distance . . . . .	254
8.6	Density-Based Methods . . . . .	255
8.6.1	Histogram- and Grid-Based Techniques . . . . .	255
8.6.2	Kernel Density Estimation . . . . .	256
8.7	Information-Theoretic Models . . . . .	256
8.8	Outlier Validity . . . . .	258
8.8.1	Methodological Challenges . . . . .	258

8.8.2	Receiver Operating Characteristic . . . . .	259
8.8.3	Common Mistakes . . . . .	261
8.9	Summary . . . . .	261
8.10	Bibliographic Notes . . . . .	262
8.11	Exercises . . . . .	262
<b>9</b>	<b>Outlier Analysis: Advanced Concepts</b>	<b>265</b>
9.1	Introduction . . . . .	265
9.2	Outlier Detection with Categorical Data . . . . .	266
9.2.1	Probabilistic Models . . . . .	266
9.2.2	Clustering and Distance-Based Methods . . . . .	267
9.2.3	Binary and Set-Valued Data . . . . .	268
9.3	High-Dimensional Outlier Detection . . . . .	268
9.3.1	Grid-Based Rare Subspace Exploration . . . . .	270
9.3.1.1	Modeling Abnormal Lower Dimensional Projections . . . . .	271
9.3.1.2	Grid Search for Subspace Outliers . . . . .	271
9.3.2	Random Subspace Sampling . . . . .	273
9.4	Outlier Ensembles . . . . .	274
9.4.1	Categorization by Component Independence . . . . .	275
9.4.1.1	Sequential Ensembles . . . . .	275
9.4.1.2	Independent Ensembles . . . . .	276
9.4.2	Categorization by Constituent Components . . . . .	277
9.4.2.1	Model-Centered Ensembles . . . . .	277
9.4.2.2	Data-Centered Ensembles . . . . .	278
9.4.3	Normalization and Combination . . . . .	278
9.5	Putting Outliers to Work: Applications . . . . .	279
9.5.1	Quality Control and Fault Detection . . . . .	279
9.5.2	Financial Fraud and Anomalous Events . . . . .	280
9.5.3	Web Log Analytics . . . . .	280
9.5.4	Intrusion Detection Applications . . . . .	280
9.5.5	Biological and Medical Applications . . . . .	281
9.5.6	Earth Science Applications . . . . .	281
9.6	Summary . . . . .	281
9.7	Bibliographic Notes . . . . .	281
9.8	Exercises . . . . .	283
<b>10</b>	<b>Data Classification</b>	<b>285</b>
10.1	Introduction . . . . .	285
10.2	Feature Selection for Classification . . . . .	287
10.2.1	Filter Models . . . . .	288
10.2.1.1	Gini Index . . . . .	288
10.2.1.2	Entropy . . . . .	289
10.2.1.3	Fisher Score . . . . .	290
10.2.1.4	Fisher's Linear Discriminant . . . . .	290
10.2.2	Wrapper Models . . . . .	292
10.2.3	Embedded Models . . . . .	292
10.3	Decision Trees . . . . .	293
10.3.1	Split Criteria . . . . .	294
10.3.2	Stopping Criterion and Pruning . . . . .	297

10.3.3	Practical Issues . . . . .	298
10.4	Rule-Based Classifiers . . . . .	298
10.4.1	Rule Generation from Decision Trees . . . . .	300
10.4.2	Sequential Covering Algorithms . . . . .	301
10.4.2.1	Learn-One-Rule . . . . .	302
10.4.3	Rule Pruning . . . . .	304
10.4.4	Associative Classifiers . . . . .	305
10.5	Probabilistic Classifiers . . . . .	306
10.5.1	Naive Bayes Classifier . . . . .	306
10.5.1.1	The Ranking Model for Classification . . . . .	309
10.5.1.2	Discussion of the Naive Assumption . . . . .	310
10.5.2	Logistic Regression . . . . .	310
10.5.2.1	Training a Logistic Regression Classifier . . . . .	311
10.5.2.2	Relationship with Other Linear Models . . . . .	312
10.6	Support Vector Machines . . . . .	313
10.6.1	Support Vector Machines for Linearly Separable Data . . . . .	313
10.6.1.1	Solving the Lagrangian Dual . . . . .	318
10.6.2	Support Vector Machines with Soft Margin for Nonseparable Data . . . . .	319
10.6.2.1	Comparison with Other Linear Models . . . . .	321
10.6.3	Nonlinear Support Vector Machines . . . . .	321
10.6.4	The Kernel Trick . . . . .	323
10.6.4.1	Other Applications of Kernel Methods . . . . .	325
10.7	Neural Networks . . . . .	326
10.7.1	Single-Layer Neural Network: The Perceptron . . . . .	326
10.7.2	Multilayer Neural Networks . . . . .	328
10.7.3	Comparing Various Linear Models . . . . .	330
10.8	Instance-Based Learning . . . . .	331
10.8.1	Design Variations of Nearest Neighbor Classifiers . . . . .	332
10.8.1.1	Unsupervised Mahalanobis Metric . . . . .	332
10.8.1.2	Nearest Neighbors with Linear Discriminant Analysis . . . . .	332
10.9	Classifier Evaluation . . . . .	334
10.9.1	Methodological Issues . . . . .	335
10.9.1.1	Holdout . . . . .	336
10.9.1.2	Cross-Validation . . . . .	336
10.9.1.3	Bootstrap . . . . .	337
10.9.2	Quantification Issues . . . . .	337
10.9.2.1	Output as Class Labels . . . . .	338
10.9.2.2	Output as Numerical Score . . . . .	339
10.10	Summary . . . . .	342
10.11	Bibliographic Notes . . . . .	342
10.12	Exercises . . . . .	343
<b>11</b>	<b>Data Classification: Advanced Concepts</b>	<b>345</b>
11.1	Introduction . . . . .	345
11.2	Multiclass Learning . . . . .	346
11.3	Rare Class Learning . . . . .	347
11.3.1	Example Reweighting . . . . .	348
11.3.2	Sampling Methods . . . . .	349

	11.3.2.1	Relationship Between Weighting and Sampling . . . . .	350
	11.3.2.2	Synthetic Oversampling: SMOTE . . . . .	350
11.4		Scalable Classification . . . . .	350
	11.4.1	Scalable Decision Trees . . . . .	351
	11.4.1.1	RainForest . . . . .	351
	11.4.1.2	BOAT . . . . .	351
	11.4.2	Scalable Support Vector Machines . . . . .	352
11.5		Regression Modeling with Numeric Classes . . . . .	353
	11.5.1	Linear Regression . . . . .	353
	11.5.1.1	Relationship with Fisher's Linear Discriminant . . . . .	356
	11.5.2	Principal Component Regression . . . . .	356
	11.5.3	Generalized Linear Models . . . . .	357
	11.5.4	Nonlinear and Polynomial Regression . . . . .	359
	11.5.5	From Decision Trees to Regression Trees . . . . .	360
	11.5.6	Assessing Model Effectiveness . . . . .	361
11.6		Semisupervised Learning . . . . .	361
	11.6.1	Generic Meta-algorithms . . . . .	363
	11.6.1.1	Self-Training . . . . .	363
	11.6.1.2	Co-training . . . . .	363
	11.6.2	Specific Variations of Classification Algorithms . . . . .	364
	11.6.2.1	Semisupervised Bayes Classification with EM . . . . .	364
	11.6.2.2	Transductive Support Vector Machines . . . . .	366
	11.6.3	Graph-Based Semisupervised Learning . . . . .	367
	11.6.4	Discussion of Semisupervised Learning . . . . .	367
11.7		Active Learning . . . . .	368
	11.7.1	Heterogeneity-Based Models . . . . .	370
	11.7.1.1	Uncertainty Sampling . . . . .	370
	11.7.1.2	Query-by-Committee . . . . .	371
	11.7.1.3	Expected Model Change . . . . .	371
	11.7.2	Performance-Based Models . . . . .	372
	11.7.2.1	Expected Error Reduction . . . . .	372
	11.7.2.2	Expected Variance Reduction . . . . .	373
	11.7.3	Representativeness-Based Models . . . . .	373
11.8		Ensemble Methods . . . . .	373
	11.8.1	Why Does Ensemble Analysis Work? . . . . .	375
	11.8.2	Formal Statement of Bias-Variance Trade-off . . . . .	377
	11.8.3	Specific Instantiations of Ensemble Learning . . . . .	379
	11.8.3.1	Bagging . . . . .	379
	11.8.3.2	Random Forests . . . . .	380
	11.8.3.3	Boosting . . . . .	381
	11.8.3.4	Bucket of Models . . . . .	383
	11.8.3.5	Stacking . . . . .	384
11.9		Summary . . . . .	384
11.10		Bibliographic Notes . . . . .	385
11.11		Exercises . . . . .	386

<b>12 Mining Data Streams</b>	<b>389</b>
12.1 Introduction . . . . .	389
12.2 Synopsis Data Structures for Streams . . . . .	391
12.2.1 Reservoir Sampling . . . . .	391
12.2.1.1 Handling Concept Drift . . . . .	393
12.2.1.2 Useful Theoretical Bounds for Sampling . . . . .	394
12.2.2 Synopsis Structures for the Massive-Domain Scenario . . . . .	398
12.2.2.1 Bloom Filter . . . . .	399
12.2.2.2 Count-Min Sketch . . . . .	403
12.2.2.3 AMS Sketch . . . . .	406
12.2.2.4 Flajolet–Martin Algorithm for Distinct Element Counting . . . . .	408
12.3 Frequent Pattern Mining in Data Streams . . . . .	409
12.3.1 Leveraging Synopsis Structures . . . . .	409
12.3.1.1 Reservoir Sampling . . . . .	410
12.3.1.2 Sketches . . . . .	410
12.3.2 Lossy Counting Algorithm . . . . .	410
12.4 Clustering Data Streams . . . . .	411
12.4.1 STREAM Algorithm . . . . .	411
12.4.2 CluStream Algorithm . . . . .	413
12.4.2.1 Microcluster Definition . . . . .	413
12.4.2.2 Microclustering Algorithm . . . . .	414
12.4.2.3 Pyramidal Time Frame . . . . .	415
12.4.3 Massive-Domain Stream Clustering . . . . .	417
12.5 Streaming Outlier Detection . . . . .	417
12.5.1 Individual Data Points as Outliers . . . . .	418
12.5.2 Aggregate Change Points as Outliers . . . . .	419
12.6 Streaming Classification . . . . .	421
12.6.1 VFDT Family . . . . .	421
12.6.2 Supervised Microcluster Approach . . . . .	424
12.6.3 Ensemble Method . . . . .	424
12.6.4 Massive-Domain Streaming Classification . . . . .	425
12.7 Summary . . . . .	425
12.8 Bibliographic Notes . . . . .	425
12.9 Exercises . . . . .	426
<b>13 Mining Text Data</b>	<b>429</b>
13.1 Introduction . . . . .	429
13.2 Document Preparation and Similarity Computation . . . . .	431
13.2.1 Document Normalization and Similarity Computation . . . . .	432
13.2.2 Specialized Preprocessing for Web Documents . . . . .	433
13.3 Specialized Clustering Methods for Text . . . . .	434
13.3.1 Representative-Based Algorithms . . . . .	434
13.3.1.1 Scatter/Gather Approach . . . . .	434
13.3.2 Probabilistic Algorithms . . . . .	436
13.3.3 Simultaneous Document and Word Cluster Discovery . . . . .	438
13.3.3.1 Co-clustering . . . . .	438
13.4 Topic Modeling . . . . .	440

13.4.1	Use in Dimensionality Reduction and Comparison with Latent Semantic Analysis . . . . .	443
13.4.2	Use in Clustering and Comparison with Probabilistic Clustering . . . . .	445
13.4.3	Limitations of PLSA . . . . .	446
13.5	Specialized Classification Methods for Text . . . . .	446
13.5.1	Instance-Based Classifiers . . . . .	447
13.5.1.1	Leveraging Latent Semantic Analysis . . . . .	447
13.5.1.2	Centroid-Based Classification . . . . .	447
13.5.1.3	Rocchio Classification . . . . .	448
13.5.2	Bayes Classifiers . . . . .	448
13.5.2.1	Multinomial Bayes Model . . . . .	449
13.5.3	SVM Classifiers for High-Dimensional and Sparse Data . . . . .	451
13.6	Novelty and First Story Detection . . . . .	453
13.6.1	Micro-clustering Method . . . . .	453
13.7	Summary . . . . .	454
13.8	Bibliographic Notes . . . . .	454
13.9	Exercises . . . . .	455
<b>14</b>	<b>Mining Time Series Data</b>	<b>457</b>
14.1	Introduction . . . . .	457
14.2	Time Series Preparation and Similarity . . . . .	459
14.2.1	Handling Missing Values . . . . .	459
14.2.2	Noise Removal . . . . .	460
14.2.3	Normalization . . . . .	461
14.2.4	Data Transformation and Reduction . . . . .	462
14.2.4.1	Discrete Wavelet Transform . . . . .	462
14.2.4.2	Discrete Fourier Transform . . . . .	462
14.2.4.3	Symbolic Aggregate Approximation (SAX) . . . . .	464
14.2.5	Time Series Similarity Measures . . . . .	464
14.3	Time Series Forecasting . . . . .	464
14.3.1	Autoregressive Models . . . . .	467
14.3.2	Autoregressive Moving Average Models . . . . .	468
14.3.3	Multivariate Forecasting with Hidden Variables . . . . .	470
14.4	Time Series Motifs . . . . .	472
14.4.1	Distance-Based Motifs . . . . .	473
14.4.2	Transformation to Sequential Pattern Mining . . . . .	475
14.4.3	Periodic Patterns . . . . .	476
14.5	Time Series Clustering . . . . .	476
14.5.1	Online Clustering of Coevolving Series . . . . .	477
14.5.2	Shape-Based Clustering . . . . .	479
14.5.2.1	$k$ -Means . . . . .	480
14.5.2.2	$k$ -Medoids . . . . .	480
14.5.2.3	Hierarchical Methods . . . . .	481
14.5.2.4	Graph-Based Methods . . . . .	481
14.6	Time Series Outlier Detection . . . . .	481
14.6.1	Point Outliers . . . . .	482
14.6.2	Shape Outliers . . . . .	483
14.7	Time Series Classification . . . . .	485

14.7.1	Supervised Event Detection . . . . .	485
14.7.2	Whole Series Classification . . . . .	488
14.7.2.1	Wavelet-Based Rules . . . . .	488
14.7.2.2	Nearest Neighbor Classifier . . . . .	489
14.7.2.3	Graph-Based Methods . . . . .	489
14.8	Summary . . . . .	489
14.9	Bibliographic Notes . . . . .	490
14.10	Exercises . . . . .	490
<b>15</b>	<b>Mining Discrete Sequences</b>	<b>493</b>
15.1	Introduction . . . . .	493
15.2	Sequential Pattern Mining . . . . .	494
15.2.1	Frequent Patterns to Frequent Sequences . . . . .	497
15.2.2	Constrained Sequential Pattern Mining . . . . .	500
15.3	Sequence Clustering . . . . .	501
15.3.1	Distance-Based Methods . . . . .	502
15.3.2	Graph-Based Methods . . . . .	502
15.3.3	Subsequence-Based Clustering . . . . .	503
15.3.4	Probabilistic Clustering . . . . .	504
15.3.4.1	Markovian Similarity-Based Algorithm: CLUSEQ . . .	504
15.3.4.2	Mixture of Hidden Markov Models . . . . .	506
15.4	Outlier Detection in Sequences . . . . .	507
15.4.1	Position Outliers . . . . .	508
15.4.1.1	Efficiency Issues: Probabilistic Suffix Trees . . . . .	510
15.4.2	Combination Outliers . . . . .	512
15.4.2.1	Distance-Based Models . . . . .	513
15.4.2.2	Frequency-Based Models . . . . .	514
15.5	Hidden Markov Models . . . . .	514
15.5.1	Formal Definition and Techniques for HMMs . . . . .	517
15.5.2	Evaluation: Computing the Fit Probability for Observed Sequence . . . . .	518
15.5.3	Explanation: Determining the Most Likely State Sequence for Observed Sequence . . . . .	519
15.5.4	Training: Baum–Welch Algorithm . . . . .	520
15.5.5	Applications . . . . .	521
15.6	Sequence Classification . . . . .	521
15.6.1	Nearest Neighbor Classifier . . . . .	522
15.6.2	Graph-Based Methods . . . . .	522
15.6.3	Rule-Based Methods . . . . .	523
15.6.4	Kernel Support Vector Machines . . . . .	524
15.6.4.1	Bag-of-Words Kernel . . . . .	524
15.6.4.2	Spectrum Kernel . . . . .	524
15.6.4.3	Weighted Degree Kernel . . . . .	525
15.6.5	Probabilistic Methods: Hidden Markov Models . . . . .	525
15.7	Summary . . . . .	526
15.8	Bibliographic Notes . . . . .	527
15.9	Exercises . . . . .	528

<b>16 Mining Spatial Data</b>	<b>531</b>
16.1 Introduction . . . . .	531
16.2 Mining with Contextual Spatial Attributes . . . . .	532
16.2.1 Shape to Time Series Transformation . . . . .	533
16.2.2 Spatial to Multidimensional Transformation with Wavelets . . . . .	537
16.2.3 Spatial Colocation Patterns . . . . .	538
16.2.4 Clustering Shapes . . . . .	539
16.2.5 Outlier Detection . . . . .	540
16.2.5.1 Point Outliers . . . . .	541
16.2.5.2 Shape Outliers . . . . .	543
16.2.6 Classification of Shapes . . . . .	544
16.3 Trajectory Mining . . . . .	544
16.3.1 Equivalence of Trajectories and Multivariate Time Series . . . . .	545
16.3.2 Converting Trajectories to Multidimensional Data . . . . .	545
16.3.3 Trajectory Pattern Mining . . . . .	546
16.3.3.1 Frequent Trajectory Paths . . . . .	546
16.3.3.2 Colocation Patterns . . . . .	548
16.3.4 Trajectory Clustering . . . . .	549
16.3.4.1 Computing Similarity Between Trajectories . . . . .	549
16.3.4.2 Similarity-Based Clustering Methods . . . . .	550
16.3.4.3 Trajectory Clustering as a Sequence Clustering Problem . . . . .	551
16.3.5 Trajectory Outlier Detection . . . . .	551
16.3.5.1 Distance-Based Methods . . . . .	551
16.3.5.2 Sequence-Based Methods . . . . .	552
16.3.6 Trajectory Classification . . . . .	553
16.3.6.1 Distance-Based Methods . . . . .	553
16.3.6.2 Sequence-Based Methods . . . . .	553
16.4 Summary . . . . .	554
16.5 Bibliographic Notes . . . . .	554
16.6 Exercises . . . . .	555
<b>17 Mining Graph Data</b>	<b>557</b>
17.1 Introduction . . . . .	557
17.2 Matching and Distance Computation in Graphs . . . . .	559
17.2.1 Ullman's Algorithm for Subgraph Isomorphism . . . . .	562
17.2.1.1 Algorithm Variations and Refinements . . . . .	563
17.2.2 Maximum Common Subgraph (MCG) Problem . . . . .	564
17.2.3 Graph Matching Methods for Distance Computation . . . . .	565
17.2.3.1 MCG-based Distances . . . . .	565
17.2.3.2 Graph Edit Distance . . . . .	567
17.3 Transformation-Based Distance Computation . . . . .	570
17.3.1 Frequent Substructure-Based Transformation and Distance Computation . . . . .	570
17.3.2 Topological Descriptors . . . . .	571
17.3.3 Kernel-Based Transformations and Computation . . . . .	573
17.3.3.1 Random Walk Kernels . . . . .	573
17.3.3.2 Shortest-Path Kernels . . . . .	575
17.4 Frequent Substructure Mining in Graphs . . . . .	575
17.4.1 Node-Based Join Growth . . . . .	578



17.4.2	Edge-Based Join Growth . . . . .	578
17.4.3	Frequent Pattern Mining to Graph Pattern Mining . . . . .	578
17.5	Graph Clustering . . . . .	579
17.5.1	Distance-Based Methods . . . . .	579
17.5.2	Frequent Substructure-Based Methods . . . . .	580
17.5.2.1	Generic Transformational Approach . . . . .	580
17.5.2.2	XProj: Direct Clustering with Frequent Subgraph Discovery . . . . .	581
17.6	Graph Classification . . . . .	582
17.6.1	Distance-Based Methods . . . . .	583
17.6.2	Frequent Substructure-Based Methods . . . . .	583
17.6.2.1	Generic Transformational Approach . . . . .	583
17.6.2.2	XRULES: A Rule-Based Approach . . . . .	584
17.6.3	Kernel SVMs . . . . .	585
17.7	Summary . . . . .	585
17.8	Bibliographic Notes . . . . .	586
17.9	Exercises . . . . .	586
<b>18</b>	<b>Mining Web Data . . . . .</b>	<b>589</b>
18.1	Introduction . . . . .	589
18.2	Web Crawling and Resource Discovery . . . . .	591
18.2.1	A Basic Crawler Algorithm . . . . .	591
18.2.2	Preferential Crawlers . . . . .	593
18.2.3	Multiple Threads . . . . .	593
18.2.4	Combating Spider Traps . . . . .	593
18.2.5	Shingling for Near Duplicate Detection . . . . .	594
18.3	Search Engine Indexing and Query Processing . . . . .	594
18.4	Ranking Algorithms . . . . .	597
18.4.1	PageRank . . . . .	598
18.4.1.1	Topic-Sensitive PageRank . . . . .	601
18.4.1.2	SimRank . . . . .	601
18.4.2	HITS . . . . .	602
18.5	Recommender Systems . . . . .	604
18.5.1	Content-Based Recommendations . . . . .	606
18.5.2	Neighborhood-Based Methods for Collaborative Filtering . . . . .	607
18.5.2.1	User-Based Similarity with Ratings . . . . .	607
18.5.2.2	Item-Based Similarity with Ratings . . . . .	608
18.5.3	Graph-Based Methods . . . . .	608
18.5.4	Clustering Methods . . . . .	609
18.5.4.1	Adapting $k$ -Means Clustering . . . . .	610
18.5.4.2	Adapting Co-Clustering . . . . .	610
18.5.5	Latent Factor Models . . . . .	611
18.5.5.1	Singular Value Decomposition . . . . .	612
18.5.5.2	Matrix Factorization . . . . .	612
18.6	Web Usage Mining . . . . .	613
18.6.1	Data Preprocessing . . . . .	614
18.6.2	Applications . . . . .	614
18.7	Summary . . . . .	615
18.8	Bibliographic Notes . . . . .	616
18.9	Exercises . . . . .	616

<b>19 Social Network Analysis</b>	<b>619</b>
19.1 Introduction . . . . .	619
19.2 Social Networks: Preliminaries and Properties . . . . .	620
19.2.1 Homophily . . . . .	621
19.2.2 Triadic Closure and Clustering Coefficient . . . . .	621
19.2.3 Dynamics of Network Formation . . . . .	622
19.2.4 Power-Law Degree Distributions . . . . .	623
19.2.5 Measures of Centrality and Prestige . . . . .	623
19.2.5.1 Degree Centrality and Prestige . . . . .	624
19.2.5.2 Closeness Centrality and Proximity Prestige . . . . .	624
19.2.5.3 Betweenness Centrality . . . . .	626
19.2.5.4 Rank Centrality and Prestige . . . . .	627
19.3 Community Detection . . . . .	627
19.3.1 Kernighan–Lin Algorithm . . . . .	629
19.3.1.1 Speeding Up Kernighan–Lin . . . . .	630
19.3.2 Girvan–Newman Algorithm . . . . .	631
19.3.3 Multilevel Graph Partitioning: METIS . . . . .	634
19.3.4 Spectral Clustering . . . . .	637
19.3.4.1 Important Observations and Intuitions . . . . .	640
19.4 Collective Classification . . . . .	641
19.4.1 Iterative Classification Algorithm . . . . .	641
19.4.2 Label Propagation with Random Walks . . . . .	643
19.4.2.1 Iterative Label Propagation: The Spectral Interpretation . . . . .	646
19.4.3 Supervised Spectral Methods . . . . .	646
19.4.3.1 Supervised Feature Generation with Spectral Embedding . . . . .	647
19.4.3.2 Graph Regularization Approach . . . . .	647
19.4.3.3 Connections with Random Walk Methods . . . . .	649
19.5 Link Prediction . . . . .	650
19.5.1 Neighborhood-Based Measures . . . . .	650
19.5.2 Katz Measure . . . . .	652
19.5.3 Random Walk-Based Measures . . . . .	653
19.5.4 Link Prediction as a Classification Problem . . . . .	653
19.5.5 Link Prediction as a Missing-Value Estimation Problem . . . . .	654
19.5.6 Discussion . . . . .	654
19.6 Social Influence Analysis . . . . .	655
19.6.1 Linear Threshold Model . . . . .	656
19.6.2 Independent Cascade Model . . . . .	657
19.6.3 Influence Function Evaluation . . . . .	657
19.7 Summary . . . . .	658
19.8 Bibliographic Notes . . . . .	659
19.9 Exercises . . . . .	660
<b>20 Privacy-Preserving Data Mining</b>	<b>663</b>
20.1 Introduction . . . . .	663
20.2 Privacy During Data Collection . . . . .	664
20.2.1 Reconstructing Aggregate Distributions . . . . .	665
20.2.2 Leveraging Aggregate Distributions for Data Mining . . . . .	667
20.3 Privacy-Preserving Data Publishing . . . . .	667
20.3.1 The $k$ -Anonymity Model . . . . .	670

20.3.1.1	Samarati's Algorithm . . . . .	673
20.3.1.2	Incognito . . . . .	675
20.3.1.3	Mondrian Multidimensional $k$ -Anonymity . . . . .	678
20.3.1.4	Synthetic Data Generation: Condensation-Based Approach . . . . .	680
20.3.2	The $\ell$ -Diversity Model . . . . .	682
20.3.3	The $t$ -closeness Model . . . . .	684
20.3.4	The Curse of Dimensionality . . . . .	687
20.4	Output Privacy . . . . .	688
20.5	Distributed Privacy . . . . .	689
20.6	Summary . . . . .	690
20.7	Bibliographic Notes . . . . .	691
20.8	Exercises . . . . .	692
<b>Bibliography</b>		<b>695</b>
<b>Index</b>		<b>727</b>

---

---

# 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 that 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 prerequisite. Special chapters are also devoted to different aspects of data mining, such as 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 anal-

ysis. 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 Chaps. 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 Chaps. 1–3, and parts of Chaps. 4, 6, 8, and 10. The course can then focus on Chaps. 5, 7, 9, and 11. Topics such as ensemble analysis are useful for the advanced course. Furthermore, some topics from Chaps. 4, 6, 8, and 10, which were not covered in the basic course, can be used. In addition, Chap. 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 Chaps. 13, 14, 15, 16, and 17. Some parts of Chap. 19 (e.g., graph clustering) and Chap. 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 Chaps. 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 (Chap. 12), Web mining (Chap. 18), and social network analysis (Chap. 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.