

# Data Mining

## Concepts and Techniques

Fourth Edition

**Jiawei Han**

**Jian Pei**

**Hanghang Tong**



# Contents

Foreword .....	xvii
Foreword to second edition .....	xix
Preface .....	xxi
Acknowledgments .....	xxvii
About the authors .....	xxix
<b>CHAPTER 1 Introduction .....</b>	<b>1</b>
<b>1.1</b> What is data mining? .....	1
<b>1.2</b> Data mining: an essential step in knowledge discovery .....	2
<b>1.3</b> Diversity of data types for data mining .....	4
<b>1.4</b> Mining various kinds of knowledge .....	5
1.4.1 Multidimensional data summarization .....	6
1.4.2 Mining frequent patterns, associations, and correlations .....	6
1.4.3 Classification and regression for predictive analysis .....	7
1.4.4 Cluster analysis .....	9
1.4.5 Deep learning .....	9
1.4.6 Outlier analysis .....	10
1.4.7 Are all mining results interesting? .....	10
<b>1.5</b> Data mining: confluence of multiple disciplines .....	12
1.5.1 Statistics and data mining .....	12
1.5.2 Machine learning and data mining .....	13
1.5.3 Database technology and data mining .....	15
1.5.4 Data mining and data science .....	15
1.5.5 Data mining and other disciplines .....	16
<b>1.6</b> Data mining and applications .....	17
<b>1.7</b> Data mining and society .....	19
<b>1.8</b> Summary .....	19
<b>1.9</b> Exercises .....	20
<b>1.10</b> Bibliographic notes .....	21
<b>CHAPTER 2 Data, measurements, and data preprocessing .....</b>	<b>23</b>
<b>2.1</b> Data types .....	24
2.1.1 Nominal attributes .....	24
2.1.2 Binary attributes .....	25
2.1.3 Ordinal attributes .....	25
2.1.4 Numeric attributes .....	26
2.1.5 Discrete vs. continuous attributes .....	27
<b>2.2</b> Statistics of data .....	27
2.2.1 Measuring the central tendency .....	28
2.2.2 Measuring the dispersion of data .....	31
	<b>vii</b>

2.2.3	Covariance and correlation analysis . . . . .	34
2.2.4	Graphic displays of basic statistics of data . . . . .	38
<b>2.3</b>	<b>Similarity and distance measures . . . . .</b>	<b>43</b>
2.3.1	Data matrix vs. dissimilarity matrix . . . . .	43
2.3.2	Proximity measures for nominal attributes . . . . .	44
2.3.3	Proximity measures for binary attributes . . . . .	46
2.3.4	Dissimilarity of numeric data: Minkowski distance . . . . .	48
2.3.5	Proximity measures for ordinal attributes . . . . .	49
2.3.6	Dissimilarity for attributes of mixed types . . . . .	50
2.3.7	Cosine similarity . . . . .	52
2.3.8	Measuring similar distributions: the Kullback-Leibler divergence . . . .	53
2.3.9	Capturing hidden semantics in similarity measures . . . . .	55
<b>2.4</b>	<b>Data quality, data cleaning, and data integration . . . . .</b>	<b>55</b>
2.4.1	Data quality measures . . . . .	55
2.4.2	Data cleaning . . . . .	56
2.4.3	Data integration . . . . .	62
<b>2.5</b>	<b>Data transformation . . . . .</b>	<b>63</b>
2.5.1	Normalization . . . . .	64
2.5.2	Discretization . . . . .	65
2.5.3	Data compression . . . . .	68
2.5.4	Sampling . . . . .	70
<b>2.6</b>	<b>Dimensionality reduction . . . . .</b>	<b>71</b>
2.6.1	Principal components analysis . . . . .	71
2.6.2	Attribute subset selection . . . . .	72
2.6.3	Nonlinear dimensionality reduction methods . . . . .	74
<b>2.7</b>	<b>Summary . . . . .</b>	<b>79</b>
<b>2.8</b>	<b>Exercises . . . . .</b>	<b>80</b>
<b>2.9</b>	<b>Bibliographic notes . . . . .</b>	<b>83</b>
<b>CHAPTER 3</b>	<b>Data warehousing and online analytical processing . . . . .</b>	<b>85</b>
<b>3.1</b>	<b>Data warehouse . . . . .</b>	<b>85</b>
3.1.1	Data warehouse: what and why? . . . . .	85
3.1.2	Architecture of data warehouses: enterprise data warehouses and data marts . . . . .	88
3.1.3	Data lakes . . . . .	93
<b>3.2</b>	<b>Data warehouse modeling: schema and measures . . . . .</b>	<b>96</b>
3.2.1	Data cube: a multidimensional data model . . . . .	97
3.2.2	Schemas for multidimensional data models: stars, snowflakes, and fact constellations . . . . .	99
3.2.3	Concept hierarchies . . . . .	103
3.2.4	Measures: categorization and computation . . . . .	105
<b>3.3</b>	<b>OLAP operations . . . . .</b>	<b>106</b>
3.3.1	Typical OLAP operations . . . . .	106
3.3.2	Indexing OLAP data: bitmap index and join index . . . . .	108
3.3.3	Storage implementation: column-based databases . . . . .	111

3.4	Data cube computation	113
3.4.1	Terminology of data cube computation	113
3.4.2	Data cube materialization: ideas	115
3.4.3	OLAP server architectures: ROLAP vs. MOLAP vs. HOLAP	117
3.4.4	General strategies for data cube computation	119
3.5	Data cube computation methods	120
3.5.1	Multiway array aggregation for full cube computation	121
3.5.2	BUC: computing iceberg cubes from the apex cuboid downward	125
3.5.3	Precomputing shell fragments for fast high-dimensional OLAP	129
3.5.4	Efficient processing of OLAP queries using cuboids	132
3.6	Summary	133
3.7	Exercises	135
3.8	Bibliographic notes	142
<b>CHAPTER 4</b>	<b>Pattern mining: basic concepts and methods</b>	<b>145</b>
4.1	Basic concepts	145
4.1.1	Market basket analysis: a motivating example	145
4.1.2	Frequent itemsets, closed itemsets, and association rules	147
4.2	Frequent itemset mining methods	149
4.2.1	Apriori algorithm: finding frequent itemsets by confined candidate generation	150
4.2.2	Generating association rules from frequent itemsets	153
4.2.3	Improving the efficiency of Apriori	155
4.2.4	A pattern-growth approach for mining frequent itemsets	157
4.2.5	Mining frequent itemsets using the vertical data format	160
4.2.6	Mining closed and max patterns	162
4.3	Which patterns are interesting?—Pattern evaluation methods	163
4.3.1	Strong rules are not necessarily interesting	163
4.3.2	From association analysis to correlation analysis	164
4.3.3	A comparison of pattern evaluation measures	165
4.4	Summary	169
4.5	Exercises	170
4.6	Bibliographic notes	173
<b>CHAPTER 5</b>	<b>Pattern mining: advanced methods</b>	<b>175</b>
5.1	Mining various kinds of patterns	175
5.1.1	Mining multilevel associations	175
5.1.2	Mining multidimensional associations	179
5.1.3	Mining quantitative association rules	180
5.1.4	Mining high-dimensional data	183
5.1.5	Mining rare patterns and negative patterns	185
5.2	Mining compressed or approximate patterns	187
5.2.1	Mining compressed patterns by pattern clustering	187
5.2.2	Extracting redundancy-aware top- <i>k</i> patterns	189

5.3	Constraint-based pattern mining . . . . .	191
5.3.1	Pruning pattern space with pattern pruning constraints . . . . .	193
5.3.2	Pruning data space with data pruning constraints . . . . .	196
5.3.3	Mining space pruning with succinctness constraints . . . . .	197
5.4	Mining sequential patterns . . . . .	198
5.4.1	Sequential pattern mining: concepts and primitives . . . . .	198
5.4.2	Scalable methods for mining sequential patterns . . . . .	200
5.4.3	Constraint-based mining of sequential patterns . . . . .	210
5.5	Mining subgraph patterns . . . . .	211
5.5.1	Methods for mining frequent subgraphs . . . . .	212
5.5.2	Mining variant and constrained substructure patterns . . . . .	219
5.6	Pattern mining: application examples . . . . .	223
5.6.1	Phrase mining in massive text data . . . . .	223
5.6.2	Mining copy and paste bugs in software programs . . . . .	230
5.7	Summary . . . . .	232
5.8	Exercises . . . . .	233
5.9	Bibliographic notes . . . . .	235
<b>CHAPTER 6</b>	<b>Classification: basic concepts and methods . . . . .</b>	<b>239</b>
6.1	Basic concepts . . . . .	239
6.1.1	What is classification? . . . . .	239
6.1.2	General approach to classification . . . . .	240
6.2	Decision tree induction . . . . .	243
6.2.1	Decision tree induction . . . . .	244
6.2.2	Attribute selection measures . . . . .	248
6.2.3	Tree pruning . . . . .	257
6.3	Bayes classification methods . . . . .	259
6.3.1	Bayes' theorem . . . . .	260
6.3.2	Naïve Bayesian classification . . . . .	262
6.4	Lazy learners (or learning from your neighbors) . . . . .	266
6.4.1	<i>k</i> -nearest-neighbor classifiers . . . . .	266
6.4.2	Case-based reasoning . . . . .	269
6.5	Linear classifiers . . . . .	269
6.5.1	Linear regression . . . . .	270
6.5.2	Perceptron: turning linear regression to classification . . . . .	272
6.5.3	Logistic regression . . . . .	274
6.6	Model evaluation and selection . . . . .	278
6.6.1	Metrics for evaluating classifier performance . . . . .	278
6.6.2	Holdout method and random subsampling . . . . .	283
6.6.3	Cross-validation . . . . .	283
6.6.4	Bootstrap . . . . .	284
6.6.5	Model selection using statistical tests of significance . . . . .	285
6.6.6	Comparing classifiers based on cost–benefit and ROC curves . . . . .	286
6.7	Techniques to improve classification accuracy . . . . .	290
6.7.1	Introducing ensemble methods . . . . .	290

6.7.2	Bagging	291
6.7.3	Boosting	292
6.7.4	Random forests	296
6.7.5	Improving classification accuracy of class-imbalanced data	297
<b>6.8</b>	Summary	298
<b>6.9</b>	Exercises	299
<b>6.10</b>	Bibliographic notes	302
<b>CHAPTER 7</b>	<b>Classification: advanced methods</b>	<b>307</b>
<b>7.1</b>	Feature selection and engineering	307
7.1.1	Filter methods	308
7.1.2	Wrapper methods	311
7.1.3	Embedded methods	312
<b>7.2</b>	Bayesian belief networks	315
7.2.1	Concepts and mechanisms	315
7.2.2	Training Bayesian belief networks	317
<b>7.3</b>	Support vector machines	318
7.3.1	Linear support vector machines	319
7.3.2	Nonlinear support vector machines	324
<b>7.4</b>	Rule-based and pattern-based classification	327
7.4.1	Using IF-THEN rules for classification	328
7.4.2	Rule extraction from a decision tree	330
7.4.3	Rule induction using a sequential covering algorithm	331
7.4.4	Associative classification	335
7.4.5	Discriminative frequent pattern-based classification	338
<b>7.5</b>	Classification with weak supervision	342
7.5.1	Semisupervised classification	343
7.5.2	Active learning	345
7.5.3	Transfer learning	346
7.5.4	Distant supervision	348
7.5.5	Zero-shot learning	349
<b>7.6</b>	Classification with rich data type	351
7.6.1	Stream data classification	352
7.6.2	Sequence classification	354
7.6.3	Graph data classification	355
<b>7.7</b>	Potpourri: other related techniques	359
7.7.1	Multiclass classification	359
7.7.2	Distance metric learning	362
7.7.3	Interpretability of classification	364
7.7.4	Genetic algorithms	367
7.7.5	Reinforcement learning	367
<b>7.8</b>	Summary	369
<b>7.9</b>	Exercises	370
<b>7.10</b>	Bibliographic notes	374

<b>CHAPTER 8</b>	<b>Cluster analysis: basic concepts and methods</b>	<b>379</b>
8.1	Cluster analysis	379
8.1.1	What is cluster analysis?	380
8.1.2	Requirements for cluster analysis	381
8.1.3	Overview of basic clustering methods	383
8.2	Partitioning methods	385
8.2.1	$k$ -Means: a centroid-based technique	386
8.2.2	Variations of $k$ -means	388
8.3	Hierarchical methods	394
8.3.1	Basic concepts of hierarchical clustering	394
8.3.2	Agglomerative hierarchical clustering	397
8.3.3	Divisive hierarchical clustering	400
8.3.4	BIRCH: scalable hierarchical clustering using clustering feature trees	402
8.3.5	Probabilistic hierarchical clustering	404
8.4	Density-based and grid-based methods	407
8.4.1	DBSCAN: density-based clustering based on connected regions with high density	408
8.4.2	DENCLUE: clustering based on density distribution functions	411
8.4.3	Grid-based methods	414
8.5	Evaluation of clustering	417
8.5.1	Assessing clustering tendency	417
8.5.2	Determining the number of clusters	419
8.5.3	Measuring clustering quality: extrinsic methods	420
8.5.4	Intrinsic methods	424
8.6	Summary	425
8.7	Exercises	427
8.8	Bibliographic notes	429
<b>CHAPTER 9</b>	<b>Cluster analysis: advanced methods</b>	<b>431</b>
9.1	Probabilistic model-based clustering	431
9.1.1	Fuzzy clusters	433
9.1.2	Probabilistic model-based clusters	435
9.1.3	Expectation-maximization algorithm	438
9.2	Clustering high-dimensional data	441
9.2.1	Why is clustering high-dimensional data challenging?	441
9.2.2	Axis-parallel subspace approaches	445
9.2.3	Arbitrarily oriented subspace approaches	447
9.3	Biclustering	447
9.3.1	Why and where is biclustering useful?	448
9.3.2	Types of biclusters	450
9.3.3	Biclustering methods	452
9.3.4	Enumerating all biclusters using MaPle	453
9.4	Dimensionality reduction for clustering	454
9.4.1	Linear dimensionality reduction methods for clustering	455
9.4.2	Nonnegative matrix factorization (NMF)	458

9.4.3	Spectral clustering . . . . .	460
<b>9.5</b>	<b>Clustering graph and network data . . . . .</b>	<b>463</b>
9.5.1	Applications and challenges . . . . .	463
9.5.2	Similarity measures . . . . .	465
9.5.3	Graph clustering methods . . . . .	470
<b>9.6</b>	<b>Semisupervised clustering . . . . .</b>	<b>475</b>
9.6.1	Semisupervised clustering on partially labeled data . . . . .	475
9.6.2	Semisupervised clustering on pairwise constraints . . . . .	476
9.6.3	Other types of background knowledge for semisupervised clustering . . . . .	477
<b>9.7</b>	<b>Summary . . . . .</b>	<b>479</b>
<b>9.8</b>	<b>Exercises . . . . .</b>	<b>480</b>
<b>9.9</b>	<b>Bibliographic notes . . . . .</b>	<b>482</b>
<b>CHAPTER 10</b>	<b>Deep learning . . . . .</b>	<b>485</b>
<b>10.1</b>	<b>Basic concepts . . . . .</b>	<b>485</b>
10.1.1	What is deep learning? . . . . .	485
10.1.2	Backpropagation algorithm . . . . .	489
10.1.3	Key challenges for training deep learning models . . . . .	498
10.1.4	Overview of deep learning architecture . . . . .	499
<b>10.2</b>	<b>Improve training of deep learning models . . . . .</b>	<b>500</b>
10.2.1	Responsive activation functions . . . . .	500
10.2.2	Adaptive learning rate . . . . .	501
10.2.3	Dropout . . . . .	504
10.2.4	Pretraining . . . . .	507
10.2.5	Cross-entropy . . . . .	509
10.2.6	Autoencoder: unsupervised deep learning . . . . .	511
10.2.7	Other techniques . . . . .	514
<b>10.3</b>	<b>Convolutional neural networks . . . . .</b>	<b>517</b>
10.3.1	Introducing convolution operation . . . . .	517
10.3.2	Multidimensional convolution . . . . .	519
10.3.3	Convolutional layer . . . . .	523
<b>10.4</b>	<b>Recurrent neural networks . . . . .</b>	<b>526</b>
10.4.1	Basic RNN models and applications . . . . .	526
10.4.2	Gated RNNs . . . . .	532
10.4.3	Other techniques for addressing long-term dependence . . . . .	536
<b>10.5</b>	<b>Graph neural networks . . . . .</b>	<b>539</b>
10.5.1	Basic concepts . . . . .	540
10.5.2	Graph convolutional networks . . . . .	541
10.5.3	Other types of GNNs . . . . .	545
<b>10.6</b>	<b>Summary . . . . .</b>	<b>547</b>
<b>10.7</b>	<b>Exercises . . . . .</b>	<b>548</b>
<b>10.8</b>	<b>Bibliographic notes . . . . .</b>	<b>552</b>
<b>CHAPTER 11</b>	<b>Outlier detection . . . . .</b>	<b>557</b>
<b>11.1</b>	<b>Basic concepts . . . . .</b>	<b>557</b>

11.1.1	What are outliers? . . . . .	558
11.1.2	Types of outliers . . . . .	559
11.1.3	Challenges of outlier detection . . . . .	561
11.1.4	An overview of outlier detection methods . . . . .	562
<b>11.2</b>	<b>Statistical approaches . . . . .</b>	<b>565</b>
11.2.1	Parametric methods . . . . .	565
11.2.2	Nonparametric methods . . . . .	569
<b>11.3</b>	<b>Proximity-based approaches . . . . .</b>	<b>572</b>
11.3.1	Distance-based outlier detection . . . . .	572
11.3.2	Density-based outlier detection . . . . .	573
<b>11.4</b>	<b>Reconstruction-based approaches . . . . .</b>	<b>576</b>
11.4.1	Matrix factorization–based methods for numerical data . . . . .	577
11.4.2	Pattern-based compression methods for categorical data . . . . .	582
<b>11.5</b>	<b>Clustering- vs. classification-based approaches . . . . .</b>	<b>585</b>
11.5.1	Clustering-based approaches . . . . .	585
11.5.2	Classification-based approaches . . . . .	588
<b>11.6</b>	<b>Mining contextual and collective outliers . . . . .</b>	<b>590</b>
11.6.1	Transforming contextual outlier detection to conventional outlier detection . . . . .	591
11.6.2	Modeling normal behavior with respect to contexts . . . . .	591
11.6.3	Mining collective outliers . . . . .	592
<b>11.7</b>	<b>Outlier detection in high-dimensional data . . . . .</b>	<b>593</b>
11.7.1	Extending conventional outlier detection . . . . .	594
11.7.2	Finding outliers in subspaces . . . . .	595
11.7.3	Outlier detection ensemble . . . . .	596
11.7.4	Taming high dimensionality by deep learning . . . . .	597
11.7.5	Modeling high-dimensional outliers . . . . .	599
<b>11.8</b>	<b>Summary . . . . .</b>	<b>600</b>
<b>11.9</b>	<b>Exercises . . . . .</b>	<b>601</b>
<b>11.10</b>	<b>Bibliographic notes . . . . .</b>	<b>602</b>
<b>CHAPTER 12</b>	<b>Data mining trends and research frontiers . . . . .</b>	<b>605</b>
<b>12.1</b>	<b>Mining rich data types . . . . .</b>	<b>605</b>
12.1.1	Mining text data . . . . .	605
12.1.2	Spatial-temporal data . . . . .	610
12.1.3	Graph and networks . . . . .	612
<b>12.2</b>	<b>Data mining applications . . . . .</b>	<b>617</b>
12.2.1	Data mining for sentiment and opinion . . . . .	617
12.2.2	Truth discovery and misinformation identification . . . . .	620
12.2.3	Information and disease propagation . . . . .	623
12.2.4	Productivity and team science . . . . .	626
<b>12.3</b>	<b>Data mining methodologies and systems . . . . .</b>	<b>629</b>
12.3.1	Structuring unstructured data for knowledge mining: a data-driven approach . . . . .	629
12.3.2	Data augmentation . . . . .	632

12.3.3	From correlation to causality	635
12.3.4	Network as a context	637
12.3.5	Auto-ML: methods and systems	640
<b>12.4</b>	<b>Data mining, people, and society</b>	<b>642</b>
12.4.1	Privacy-preserving data mining	642
12.4.2	Human-algorithm interaction	646
12.4.3	Mining beyond maximizing accuracy: fairness, interpretability, and robustness	648
12.4.4	Data mining for social good	652
<b>APPENDIX A</b>	<b>Mathematical background</b>	<b>655</b>
<b>A.1</b>	<b>Probability and statistics</b>	<b>655</b>
A.1.1	PDF of typical distributions	655
A.1.2	MLE and MAP	656
A.1.3	Significance test	657
A.1.4	Density estimation	658
A.1.5	Bias-variance tradeoff	659
A.1.6	Cross-validation and Jackknife	660
<b>A.2</b>	<b>Numerical optimization</b>	<b>661</b>
A.2.1	Gradient descent	661
A.2.2	Variants of gradient descent	662
A.2.3	Newton's method	664
A.2.4	Coordinate descent	666
A.2.5	Quadratic programming	666
<b>A.3</b>	<b>Matrix and linear algebra</b>	<b>668</b>
A.3.1	Linear system $\mathbf{Ax} = \mathbf{b}$	668
A.3.2	Norms of vectors and matrices	669
A.3.3	Matrix decompositions	669
A.3.4	Subspace	671
A.3.5	Orthogonality	672
<b>A.4</b>	<b>Concepts and tools from signal processing</b>	<b>673</b>
A.4.1	Entropy	673
A.4.2	Kullback-Leibler divergence (KL-divergence)	674
A.4.3	Mutual information	675
A.4.4	Discrete Fourier transform (DFT) and fast Fourier transform (FFT)	676
<b>A.5</b>	<b>Bibliographic notes</b>	<b>678</b>
	Bibliography	681
	Index	735