

Contents

List of Figures	xiii
List of Tables	xv
List of Algorithms	xvii
Preface to the Second Edition	xix
Preface to the First Edition	xxi
I Clustering, Data, and Similarity Measures	1
1 Data Clustering	3
1.1 Definition of Data Clustering	3
1.2 The Vocabulary of Clustering	5
1.2.1 Records and Attributes	5
1.2.2 Distances and Similarities	5
1.2.3 Clusters, Centers, and Modes	6
1.2.4 Hard Clustering and Fuzzy Clustering	6
1.2.5 Validity Indices	7
1.3 Clustering Processes	8
1.4 Dealing with Missing Values	9
1.5 Resources for Clustering	11
1.5.1 Surveys and Reviews on Clustering	11
1.5.2 Books on Clustering	12
1.5.3 Journals and Conference Proceedings	14
1.5.4 Conferences	16
1.5.5 Data Sets	16
1.6 Summary	17
2 Data Types	19
2.1 Structured Data	19
2.1.1 Categorical Data	20
2.1.2 Binary Data	22
2.2 Unstructured Data	22
2.2.1 Text Data	22
2.2.2 Transaction Data	23
2.2.3 Symbolic Data	23

	2.2.4	Time Series	24
2.3		Summary	24
3		Scale Conversion	25
3.1		Introduction	25
	3.1.1	Interval to Ordinal	25
	3.1.2	Interval to Nominal	27
	3.1.3	Ordinal to Nominal	27
	3.1.4	Nominal to Ordinal	27
	3.1.5	Ordinal to Interval	28
	3.1.6	Other Conversions	29
3.2		Categorization of Numerical Data	29
	3.2.1	Direct Categorization	29
	3.2.2	Cluster-Based Categorization	30
	3.2.3	Automatic Categorization	34
3.3		Summary	39
4		Data Standardization and Transformation	41
4.1		Data Standardization	41
4.2		Data Transformation	44
	4.2.1	Power Transformation	44
	4.2.2	Principal Component Analysis	44
	4.2.3	SVD	45
	4.2.4	The Karhunen–Loève Transformation	47
4.3		Summary	49
5		Data Visualization	51
5.1		Sammon’s Mapping	51
5.2		MDS	52
5.3		SOM	54
5.4		Class-Preserving Projections	56
5.5		Parallel Coordinates	57
5.6		Tree Maps	58
5.7		<i>t</i> -SNE	59
5.8		Categorical Data Visualization	60
5.9		Other Visualization Techniques	62
5.10		Summary	63
6		Similarity and Dissimilarity Measures	65
6.1		Preliminaries	65
	6.1.1	Proximity Matrix	66
	6.1.2	Proximity Graph	67
	6.1.3	Scatter Matrix	67
	6.1.4	Covariance Matrix	67
6.2		Measures for Numerical Data	68
	6.2.1	Euclidean Distance	68
	6.2.2	Manhattan Distance	69
	6.2.3	Maximum Distance	69
	6.2.4	Minkowski Distance	69
	6.2.5	Mahalanobis Distance	69

6.2.6	Average Distance	70
6.2.7	Other Distances	71
6.3	Measures for Categorical Data	71
6.3.1	The Simple Matching Distance	71
6.3.2	Other Matching Coefficients	73
6.4	Measures for Binary Data	73
6.5	Measures for Mixed-Type Data	75
6.5.1	A General Similarity Coefficient	76
6.5.2	A General Distance Coefficient	77
6.5.3	A Generalized Minkowski Distance	77
6.6	Measures for Time Series Data	79
6.6.1	The Minkowski Distance	80
6.6.2	Time Series Preprocessing	81
6.6.3	Dynamic Time Warping	83
6.6.4	Measures Based on Longest Common Subsequences	84
6.6.5	Measures Based on Probabilistic Models	85
6.6.6	Measures Based on Landmark Models	86
6.6.7	Evaluation	87
6.7	Other Measures	87
6.7.1	The Cosine Similarity Measure	88
6.7.2	A Link-Based Similarity Measure	88
6.7.3	Support	89
6.8	Similarity and Dissimilarity Measures between Clusters	89
6.8.1	The Mean-Based Distance	89
6.8.2	The Nearest Neighbor Distance	89
6.8.3	The Farthest Neighbor Distance	90
6.8.4	The Average Neighbor Distance	90
6.8.5	Lance–Williams Formula	91
6.9	Similarity and Dissimilarity between Variables	92
6.9.1	Pearson’s Correlation Coefficients	92
6.9.2	Measures Based on the Chi-Square Statistic	95
6.9.3	Measures Based on Optimal Class Prediction	97
6.9.4	Group-Based Distance	99
6.10	Summary	100
II	Clustering Algorithms	101
7	Hierarchical Clustering Techniques	103
7.1	Representations of Hierarchical Clusterings	104
7.1.1	n -Tree	104
7.1.2	Dendrogram	104
7.1.3	Banner	106
7.1.4	Pointer Representation	106
7.1.5	Packed Representation	107
7.1.6	Icicle Plot	108
7.1.7	Other Representations	109
7.2	Agglomerative Hierarchical Methods	109
7.2.1	The Single-Link Method	110
7.2.2	The Complete Link Method	113

7.2.3	The Group Average Method	115
7.2.4	The Weighted Group Average Method	117
7.2.5	The Centroid Method	118
7.2.6	The Median Method	122
7.2.7	Ward's Method	123
7.2.8	Other Agglomerative Methods	128
7.3	Divisive Hierarchical Methods	128
7.4	Several Hierarchical Algorithms	129
7.4.1	SLINK	129
7.4.2	Single-Link Algorithms Based on Minimum Spanning Trees	130
7.4.3	CLINK	132
7.4.4	BIRCH	134
7.4.5	CURE	134
7.4.6	DIANA	135
7.4.7	DISMEA	136
7.4.8	Edwards and Cavalli-Sforza Method	137
7.5	Summary	138
8	Fuzzy Clustering Algorithms	139
8.1	Fuzzy Sets	139
8.2	Fuzzy Relations	141
8.3	Fuzzy k -Means	142
8.4	Fuzzy k -Modes	143
8.5	The c -Means Method	145
8.6	Summary	146
9	Center-Based Clustering Algorithms	147
9.1	The k -Means Algorithm	147
9.2	Variations of the k -Means Algorithm	150
9.2.1	The Continuous k -Means Algorithm	150
9.2.2	The Compare-Means Algorithm	151
9.2.3	The Sort-Means Algorithm	151
9.2.4	Acceleration of the k -Means Algorithm with the kd -Tree	152
9.2.5	Other Acceleration Methods	153
9.3	The Trimmed k -Means Algorithm	154
9.4	The x -Means Algorithm	155
9.5	The k -Harmonic Means Algorithm	156
9.6	The Mean Shift Algorithm	157
9.7	MEC	159
9.8	The k -Modes Algorithm (Huang)	160
9.8.1	Initial Modes Selection	162
9.9	The k -Modes Algorithm (Chaturvedi et al.)	162
9.10	The k -Probabilities Algorithm	163
9.11	The k -Prototypes Algorithm	165
9.12	Summary	166
10	Search-Based Clustering Algorithms	167
10.1	Genetic Algorithms	167
10.2	The Tabu Search Method	169
10.3	Variable Neighborhood Search for Clustering	170

10.4	Al-Sultan's Method	171
10.5	Tabu Search-Based Categorical Clustering Algorithm	172
10.6	<i>J</i> -means	173
10.7	GKA	175
10.8	The Global <i>k</i> -Means Algorithm	177
10.9	The Genetic <i>k</i> -Modes Algorithm	178
	10.9.1 The Selection Operator	178
	10.9.2 The Mutation Operator	179
	10.9.3 The <i>k</i> -Modes Operator	179
10.10	The Genetic Fuzzy <i>k</i> -Modes Algorithm	180
	10.10.1 String Representation	180
	10.10.2 Initialization Process	181
	10.10.3 Selection Process	181
	10.10.4 Crossover Process	181
	10.10.5 Mutation Process	182
	10.10.6 Termination Criterion	182
10.11	SARS	182
10.12	Summary	184
11	Graph-Based Clustering Algorithms	185
11.1	Chameleon	185
11.2	CACTUS	186
11.3	A Dynamic System-Based Approach	187
11.4	ROCK	188
11.5	Summary	189
12	Grid-Based Clustering Algorithms	191
12.1	STING	191
12.2	OptiGrid	192
12.3	GRIDCLUS	194
12.4	GDILC	196
12.5	WaveCluster	197
12.6	Summary	198
13	Density-Based Clustering Algorithms	199
13.1	DBSCAN	199
13.2	BRIDGE	200
13.3	DBCLASD	201
13.4	DENCLUE	203
13.5	CUBN	204
13.6	Summary	205
14	Model-Based Clustering Algorithms	207
14.1	Introduction	207
14.2	Gaussian Clustering Models	209
14.3	Model-Based Agglomerative Hierarchical Clustering	210
14.4	The EM Algorithm	214
14.5	Model-Based Clustering	216
14.6	COOLCAT	218
14.7	STUCCO	219

14.8	Summary	220
15	Subspace Clustering	221
15.1	CLIQUE	222
15.2	PROCLUS	224
15.3	ORCLUS	226
15.4	ENCLUS	230
15.5	FINDIT	232
15.6	MAFIA	234
15.7	DOC	235
15.8	CLTree	237
15.9	PART	238
15.10	SUBCAD	239
15.11	Fuzzy Subspace Clustering	245
15.12	Mean Shift for Subspace Clustering	249
15.13	Summary	258
16	Scalable Clustering Algorithms	261
16.1	Overview	261
16.2	WAND- k -Means	262
16.3	TFCM	264
16.4	Summary	265
17	Miscellaneous Algorithms	267
17.1	Time Series Clustering Algorithms	267
17.2	Streaming Algorithms	269
17.2.1	LSEARCH	269
17.2.2	Other Streaming Algorithms	272
17.3	Transaction Data Clustering Algorithms	272
17.3.1	LargeItem	273
17.3.2	CLOPE	274
17.3.3	OAK	275
17.4	Summary	276
18	Evaluation of Clustering Algorithms	277
18.1	Introduction	277
18.1.1	Hypothesis Testing	277
18.1.2	External Criteria	279
18.1.3	Internal Criteria	280
18.1.4	Relative Criteria	281
18.2	Evaluation of Partitional Clustering	281
18.2.1	Modified Hubert's Γ Statistic	281
18.2.2	The Davies–Bouldin Index	282
18.2.3	Dunn's Index	283
18.2.4	The SD Validity Index	283
18.2.5	The S_Dbw Validity Index	284
18.2.6	The RMSSTD Index	285
18.2.7	The RS Index	286
18.2.8	The Calinski–Harabasz Index	286
18.2.9	Rand's Index	287

18.2.10	Average of Compactness	287
18.2.11	Distances between Partitions	288
18.3	Evaluation of Hierarchical Clustering	289
18.3.1	Testing Absence of Structure	289
18.3.2	Testing Hierarchical Structures	290
18.4	Validity Indices for Fuzzy Clustering	290
18.4.1	The Partition Coefficient Index	291
18.4.2	The Partition Entropy Index	291
18.4.3	The Fukuyama–Sugeno Index	291
18.4.4	Validity Based on Fuzzy Similarity	292
18.4.5	A Compact and Separate Fuzzy Validity Criterion	293
18.4.6	A Partition Separation Index	293
18.4.7	An Index Based on the Mini-Max Filter Concept and Fuzzy Theory	294
18.5	Summary	295
III	Clustering Software	297
19	Open-Source Clustering Software	299
19.1	R Software	299
19.2	Python Software	307
19.3	Java Software	308
19.4	C++ Software	309
19.5	Summary	309
20	A Lightweight Java Clustering Framework	311
20.1	Overview of the Framework	311
20.2	Data sets	312
20.3	Clusters	313
20.4	Clustering Algorithms	316
20.5	Distances	319
20.6	Clustering Validation	320
20.7	Initialization	322
20.8	Utilities	323
20.9	Usage	323
20.10	Summary	326
IV	Applications of Clustering	327
21	Clustering Gene Expression Data	329
21.1	Background	329
21.2	Applications of Gene Expression Data Clustering	330
21.3	Types of Gene Expression Data Clustering	330
21.4	Some Guidelines for Gene Expression Clustering	331
21.5	Similarity Measures for Gene Expression Data	331
21.5.1	Euclidean Distance	332
21.5.2	Pearson’s Correlation Coefficient	332
21.6	A Case Study	333

21.6.1	Java Code	333
21.6.2	Results	336
21.7	Summary	339
22	Clustering Variable Annuity Policies	341
22.1	Background	341
22.2	Similarity Measures for Variable Annuity Data	342
22.3	A Case Study	343
22.3.1	Java Code	344
22.3.2	Results	346
22.4	Summary	348
A	Some Clustering Algorithms	349
B	The kd-Tree Data Structure	351
	Bibliography	353
	Index	403

List of Figures

1.1	Data-mining tasks.	4
1.2	Three well-separated center-based clusters in a two-dimensional space.	6
1.3	Two chained clusters in a two-dimensional space.	7
1.4	Processes of data clustering.	8
1.5	Diagram of clustering algorithms.	9
2.1	Diagram of data types.	19
2.2	Diagram of data scales.	20
3.1	An example two-dimensional data set with 60 points.	31
3.2	Examples of direct categorization when $N = 5$	31
3.3	Examples of direct categorization when $N = 2$	31
3.4	Examples of k -means-based categorization when $N = 5$	32
3.5	Examples of k -means-based categorization when $N = 2$	33
3.6	Examples of cluster-based categorization based on the least squares partition when $N = 5$	35
3.7	Examples of cluster-based categorization based on the least squares partition when $N = 2$	35
3.8	Examples of automatic categorization using the k -means algorithm and the compactness-separation criterion	36
3.9	Plots of the compactness-separation validity measure against the number of clusters.	37
3.10	Examples of automatic categorization based on the least squares partition and the SSC	38
3.11	Plots of the sum of squares validity measure against the number of clusters.	38
5.1	The architecture of the SOM.	54
5.2	The axes of the parallel coordinates system.	57
5.3	A two-dimensional data set containing five points.	57
5.4	The parallel coordinates plot of the five points in Figure 5.3.	58
5.5	The dendrogram of the single-linkage hierarchical clustering of the five points in Figure 5.3.	59
5.6	The tree maps of the dendrogram in Figure 5.5.	59
5.7	Plot of the two clusters in Table 5.1.	62
6.1	Nearest neighbor distance between two clusters.	90
6.2	Farthest neighbor distance between two clusters.	90
7.1	Agglomerative hierarchical clustering and divisive hierarchical clustering.	103

7.2	A 5-tree.	104
7.3	A dendrogram of five data points.	105
7.4	A banner constructed from the dendrogram given in Figure 7.3.	106
7.5	The dendrogram determined by the packed representation given in Table 7.3.	108
7.6	An icicle plot corresponding to the dendrogram given in Figure 7.3.	108
7.7	A loop plot corresponding to the dendrogram given in Figure 7.3.	109
7.8	Some commonly used hierarchical methods.	109
7.9	A two-dimensional data set with five data points.	112
7.10	The dendrogram produced by applying the single-link method to the data set given in Figure 7.9.	113
7.11	The dendrogram produced by applying the complete link method to the data set given in Figure 7.9.	115
7.12	The dendrogram produced by applying the group average method to the data set given in Figure 7.9.	117
7.13	The dendrogram produced by applying the weighted group average method to the data set given in Figure 7.9.	118
7.14	The dendrogram produced by applying the centroid method to the data set given in Figure 7.9.	122
7.15	The dendrogram produced by applying the median method to the data set given in Figure 7.9.	123
7.16	The dendrogram produced by applying Ward's method to the data set given in Figure 7.9.	128
14.1	The flowchart of the model-based clustering procedure.	209
15.1	The relationship between the mean shift algorithm and its derivatives.	250
16.1	Techniques for clustering big data.	262
18.1	Diagram of the cluster validity indices.	278
20.1	Directory structure of the <code>jclust</code> framework.	312
20.2	A screenshot of the Eclipse IDE with the <code>jclust</code> project.	313
20.3	A class diagram of the dataset package.	314
20.4	A class diagram of the cluster package.	315
20.5	A class diagram of the algorithm package.	317
20.6	A class diagram of the distance classes.	320
20.7	A class diagram of the validation classes.	321
20.8	A class diagram of the initialization classes.	322
20.9	A class diagram of some utility classes.	324
21.1	The first six clusters.	339
21.2	The last four clusters.	340

List of Tables

1.1	A list of methods for dealing with missing values.	10
2.1	A sample categorical data set.	20
2.2	One of the symbol tables of the data set in Table 2.1.	21
2.3	Another symbol table of the data set in Table 2.1.	21
2.4	The frequency table computed from the symbol table in Table 2.2.	21
2.5	The frequency table computed from the symbol table in Table 2.3.	21
4.1	Some data standardization methods, where \bar{x}_j^* , R_j^* , and σ_j^* are defined in equation (4.3).	42
5.1	The coordinate system for the two clusters of the data set in Table 2.1.	61
5.2	Coordinates of the attribute values of the two clusters in Table 5.1.	61
6.1	Some other dissimilarity measures for numerical data.	72
6.2	Some matching coefficients for nominal data.	74
6.3	Similarity measures for binary vectors.	75
6.4	Some symmetrical coefficients for binary feature vectors.	75
6.5	Some asymmetrical coefficients for binary feature vectors.	76
6.6	Some commonly used values for the parameters in the Lance–Williams formula, where $n_i = C_i $ is the number of data points in C_i , and $\Sigma_{ijk} = n_i + n_j + n_k$	91
6.7	Some common parameters for the general recurrence formula proposed by Jambu (1978). Adapted from Gordon (1996).	93
6.8	The contingency table of variables u and v	96
6.9	Measures of association based on the chi-square statistic.	97
7.1	The pointer representation corresponding to the dendrogram given in Figure 7.3.	107
7.2	The packed representation corresponding to the pointer representation given in Table 7.1.	107
7.3	A packed representation of six objects.	108
7.4	The cluster centers agglomerated from two clusters and the dissimilarities between two cluster centers for geometric hierarchical methods, where $\mu(C)$ denotes the center of cluster C	110
7.5	The dissimilarity matrix of the data set given in Figure 7.9. The entry (i, j) in the matrix is the Euclidean distance between \mathbf{x}_i and \mathbf{x}_j	112
7.6	The dissimilarity matrix of the data set given in Figure 7.9.	126

11.1	Description of the chameleon algorithm, where n is the number of data in the database and m is the number of initial subclusters.	186
11.2	The properties of the ROCK algorithm, where n is the number of data points in the data set, m_m is the maximum number of neighbors for a point, and m_a is the average number of neighbors.	189
14.1	Description of Gaussian mixture models in the general family.	211
14.2	Description of Gaussian mixture models in the diagonal family. \mathbf{B} is a diagonal matrix.	211
14.3	Description of Gaussian mixture models in the diagonal family. \mathbf{I} is an identity matrix.	212
14.4	Four parameterizations of the covariance matrix in the Gaussian model and their corresponding criteria to be minimized.	213
15.1	List of some subspace clustering algorithms.	222
15.2	Description of the MAFIA algorithm.	235
18.1	Some indices that measure the degree of similarity between C and P based on the external criteria.	279
19.1	A list of R packages related to clustering.	299
19.2	Some Python packages related to data clustering.	307
19.3	Some Java software related to data clustering.	308
19.4	Some C++ software related to data clustering.	309

List of Algorithms

Algorithm 5.1	Nonmetric MDS	53
Algorithm 5.2	Pseudocode of the SOM algorithm	56
Algorithm 7.1	The SLINK algorithm	130
Algorithm 7.2	Pseudocode of the CLINK algorithm	132
Algorithm 8.1	The fuzzy k -means algorithm	142
Algorithm 8.2	The fuzzy k -modes algorithm	144
Algorithm 9.1	The conventional k -means algorithm	148
Algorithm 9.2	The k -means algorithm treated as an optimization problem	149
Algorithm 9.3	The compare-means algorithm	151
Algorithm 9.4	An iteration of the sort-means algorithm	152
Algorithm 9.5	The k -modes algorithm	161
Algorithm 9.6	The k -probabilities algorithm	164
Algorithm 9.7	The k -prototypes algorithm	165
Algorithm 10.1	The VNS heuristic	170
Algorithm 10.2	Al-Sultan’s tabu search–based clustering algorithm	171
Algorithm 10.3	The J -means algorithm	174
Algorithm 10.4	Mutation (s_W)	176
Algorithm 10.5	The pseudocode of GKA	177
Algorithm 10.6	Mutation (s_W) in GKMODE	179
Algorithm 10.7	The SARS algorithm	184
Algorithm 11.1	The procedure of the chameleon algorithm	186
Algorithm 11.2	The CACTUS algorithm	187
Algorithm 11.3	The dynamic system–based clustering algorithm	188
Algorithm 11.4	The ROCK algorithm	189
Algorithm 12.1	The STING algorithm	192
Algorithm 12.2	The OptiGrid algorithm	193
Algorithm 12.3	The GRIDCLUS algorithm	195
Algorithm 12.4	Procedure $NEIGHBOR_SEARCH(B, C)$	195
Algorithm 12.5	The GDILC algorithm	196
Algorithm 13.1	The BRIDGE algorithm	201
Algorithm 14.1	Model-based clustering procedure	217
Algorithm 14.2	The COOLCAT clustering algorithm	219
Algorithm 14.3	The STUCCO clustering algorithm procedure	220
Algorithm 15.1	The PROCLUS algorithm	224
Algorithm 15.2	The pseudocode of the ORCLUS algorithm	226
Algorithm 15.3	$Assign(s_1, \dots, s_{k_c}, P_1, \dots, P_{k_c})$	228
Algorithm 15.4	$Merge(C_1, \dots, C_{k_c}, K_{new}, l_{new})$	228
Algorithm 15.5	$FindVectors(C, q)$	229
Algorithm 15.6	ENCLUS procedure for mining significant subspaces	231

Algorithm 15.7	ENCLUS procedure for mining interesting subspaces	231
Algorithm 15.8	The FINDIT algorithm	233
Algorithm 15.9	Procedure of adaptive grids computation in the MAFIA algorithm . .	234
Algorithm 15.10	The DOC algorithm for approximating an optimal projective cluster .	235
Algorithm 15.11	The SUBCAD algorithm	241
Algorithm 15.12	The pseudocode of the FSC algorithm	249
Algorithm 15.13	The pseudocode of the MSSC algorithm	255
Algorithm 15.14	The postprocessing procedure to get the final subspace clusters . . .	256
Algorithm 16.1	Pseudocode of the WAND- k -means algorithm	263
Algorithm 16.2	Pseudocode of the TFCM algorithm	265
Algorithm 17.1	The InitialSolution algorithm	270
Algorithm 17.2	The LSEARCH algorithm	271
Algorithm 17.3	The $FL(D, d(\cdot, \cdot), z, \epsilon, (I, a))$ function	271
Algorithm 17.4	The CLOPE algorithm	275
Algorithm 17.5	A sketch of the OAK algorithm	276
Algorithm 18.1	The Monte Carlo technique for computing the probability density function of the indices	278

Preface to the Second Edition

The monograph *Data Clustering: Theory, Algorithms, and Applications* was published in 2007. Starting with the common ground and knowledge for data clustering, the monograph focuses on several popular clustering algorithms and groups them according to some specific baseline methodologies, such as hierarchical, center-based, and search-based methods. Since the publication of this monograph, development in the subject area has exploded in many different directions, especially in clustering algorithms for big data and open-source software for cluster analysis.

This second edition aims to reflect some of these new developments. In addition to expanding some existing chapters, we added a few new chapters (e.g., Chapters 16, 19, 20, 21, and 22). Chapter 16 introduces scalable clustering algorithms for big data. Chapter 19 discusses open-source software that can be used to perform data clustering. Chapter 20 introduces a lightweight Java framework that can be used by researchers and practitioners to develop and test clustering algorithms. Chapters 21 and 22 demonstrate additional applications of data clustering in different areas. We also removed Chapters 19 and 20 from the first edition. We removed Chapter 19, which is about data clustering in MATLAB, for two reasons: first, MATLAB is a scripting language and is not efficient for developing clustering algorithms; second, the R programming language has become popular since the first edition. We removed Chapter 20, which is about data clustering in C/C++ because there is a whole book (i.e., Gan, 2011) devoted to this topic.

As in the first edition, we made no attempt to provide a comprehensive coverage of the subject area. The second edition follows the same structure as the first edition and covers a few clustering algorithms of different types. The criteria for selecting these algorithms are personal rather than scientific, and it is highly possible that some of those famous algorithms deserving an entry are absent. We apologize sincerely for any such unintentional omissions and for having to leave out many contributions. In this second edition, we also removed the author index of the first edition because the names of authors after the first or the second author in multi-author articles do not appear in the referenced context.

Java code and relevant data sets used in this new edition are available online from <https://bookstore.siam.org/mn05/bonus>. We hope that this new edition continues to provide researchers, students, and practitioners from a variety of disciplines with a useful reference to theory, algorithms, and applications of data clustering.

We would like to acknowledge Paula M. Callaghan from the Society for Industrial and Applied Mathematics (SIAM) for encouraging us to work on the second edition. We are deeply grateful for the researchers and readers who have published reviews for the first edition of this book: Dick Burkhart, David J. Hand, Fatih Kurugollu, Kenneth Joseph Ryan, Jin-Hong Park, and Hao Helen Zhang. We apologize again if we missed someone here. We thank the anonymous reviewers who provided valuable feedback and suggestions for the second edition.

We would also like to acknowledge the following financial support: the CAE (Centers of Actuarial Excellence) grant awarded to the University of Connecticut by the Society of Actuaries, the Key Projects of the National Natural Science Foundation of China (grant 71431008, the

Modeling of Financial Data with High Dimensional, Nonlinear, Nonstationary, Time-Varying and Their Application), the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada's Discovery Grant Program and Collaborative Research Development Program, and Mathematics for Information Technology and Complex Systems.

May 31, 2020

Guojun, Chaoqun, and Jianhong
Storrs, CT, USA, Changsha, Hunan, China, and Toronto, ON, Canada

Preface to the First Edition

Cluster analysis is an unsupervised process that divides a set of objects into homogeneous groups. There have been many clustering algorithms scattered in publications in very diversified areas such as pattern recognition, artificial intelligence, information technology, image processing, biology, psychology, and marketing. As such, readers and users often find it very difficult to identify an appropriate algorithm for their applications and/or to compare novel ideas with existing results.

In this monograph, we shall focus on a small number of popular clustering algorithms and group them according to some specific baseline methodologies, such as hierarchical, center-based, and search-based methods. We shall, of course, start with the common ground and knowledge for cluster analysis, including the classification of data and the corresponding similarity measures, and we shall also provide examples of clustering applications to illustrate the advantages and shortcomings of different clustering architectures and algorithms.

This monograph is intended not only for statistics, applied mathematics, and computer science senior undergraduates and graduates, but also for research scientists who need cluster analysis to deal with data. It may be used as a textbook for introductory courses in cluster analysis or as source material for an introductory course in data mining at the graduate level. We assume that the reader is familiar with elementary linear algebra, calculus, and basic statistical concepts and methods.

The book is divided into four parts: basic concepts (clustering, data, and similarity measures), algorithms, applications, and programming languages. We now briefly describe the content of each chapter.

Chapter 1. Data clustering. In this chapter, we introduce the basic concepts of clustering. Cluster analysis is defined as a way to create groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct. Some working definitions of clusters are discussed, and several popular books relevant to cluster analysis are introduced.

Chapter 2. Data types. The type of data is directly associated with data clustering, and it is a major factor to consider in choosing an appropriate clustering algorithm. Five data types are discussed in this chapter: categorical, binary, transaction, symbolic, and time series. They share a common feature that nonnumerical similarity measures must be used. There are many other data types, such as image data, that are not discussed here, though we believe that once readers get familiar with these basic types of data, they should be able to adjust the algorithms accordingly.

Chapter 3. Scale conversion. Scale conversion is concerned with the transformation between different types of variables. For example, one may convert a continuous measured variable to an interval variable. In this chapter, we first review several scale conversion techniques and then discuss several approaches for categorizing numerical data.

Chapter 4. Data standardization and transformation. In many situations, raw data should be normalized and/or transformed before a cluster analysis. One reason to do this is that objects

in raw data may be described by variables measured with different scales; another reason is to reduce the size of the data to improve the effectiveness of clustering algorithms. Therefore, we present several data standardization and transformation techniques in this chapter.

Chapter 5. Data visualization. Data visualization is vital in the final step of data-mining applications. This chapter introduces various techniques of visualization with an emphasis on visualization of clustered data. Some dimension reduction techniques, such as multidimensional scaling (MDS) and self-organizing maps (SOMs), are discussed.

Chapter 6. Similarity and dissimilarity measures. In the literature of data clustering, a similarity measure or distance (dissimilarity measure) is used to quantitatively describe the similarity or dissimilarity of two data points or two clusters. Similarity and distance measures are basic elements of a clustering algorithm, without which no meaningful cluster analysis is possible. Due to the important role of similarity and distance measures in cluster analysis, we present a comprehensive discussion of different measures for various types of data in this chapter. We also introduce measures between points and measures between clusters.

Chapter 7. Hierarchical clustering techniques. Hierarchical clustering algorithms and partitioning algorithms are two major clustering algorithms. Unlike partitioning algorithms, which divide a data set into a single partition, hierarchical algorithms divide a data set into a sequence of nested partitions. There are two major hierarchical algorithms: agglomerative algorithms and divisive algorithms. Agglomerative algorithms start with every single object in a single cluster, while divisive ones start with all objects in one cluster and repeat splitting large clusters into small pieces. In this chapter, we present representations of hierarchical clustering and several popular hierarchical clustering algorithms.

Chapter 8. Fuzzy clustering algorithms. Clustering algorithms can be classified into two categories: hard clustering algorithms and fuzzy clustering algorithms. Unlike hard clustering algorithms, which require that each data point of the data set belong to one and only one cluster, fuzzy clustering algorithms allow a data point to belong to two or more clusters with different probabilities. There is also a huge number of published works related to fuzzy clustering. In this chapter, we review some basic concepts of fuzzy logic and present three well-known fuzzy clustering algorithms: fuzzy k -means, fuzzy k -modes, and c -means.

Chapter 9. Center-based clustering algorithms. Compared to other types of clustering algorithms, center-based clustering algorithms are more suitable for clustering large data sets and high-dimensional data sets. Several well-known center-based clustering algorithms (e.g., k -means, k -modes) are presented and discussed in this chapter.

Chapter 10. Search-based clustering algorithms. A well-known problem associated with most of the clustering algorithms is that they may not be able to find the globally optimal clustering that fits the data set, since these algorithms will stop if they find a local optimal partition of the data set. This problem led to the invention of search-based clustering algorithms to search the solution space and find a globally optimal clustering that fits the data set. In this chapter, we present several clustering algorithms based on genetic algorithms, tabu search algorithms, and simulated annealing algorithms.

Chapter 11. Graph-based clustering algorithms. Graph-based clustering algorithms cluster a data set by clustering the graph or hypergraph constructed from the data set. The construction of a graph or hypergraph is usually based on the dissimilarity matrix of the data set under consideration. In this chapter, we present several graph-based clustering algorithms that do not use the spectral graph partition techniques, although we also list a few references related to spectral graph partition techniques.

Chapter 12. Grid-based clustering algorithms. In general, a grid-based clustering algorithm consists of the following five basic steps: partitioning the data space into a finite number of cells (or creating grid structure), estimating the cell density for each cell, sorting the cells according to their densities, identifying cluster centers, and traversal of neighbor cells. A major

advantage of grid-based clustering is that it significantly reduces the computational complexity. Some recent works on grid-based clustering are presented in this chapter.

Chapter 13. Density-based clustering algorithms. The density-based clustering approach is capable of finding arbitrarily shaped clusters, where clusters are defined as dense regions separated by low-density regions. Usually, density-based clustering algorithms are not suitable for high-dimensional data sets, since data points are sparse in high-dimensional spaces. Five density-based clustering algorithms (DBSCAN, BRIDGE, DBCLASD, DENCLUE, and CUBN) are presented in this chapter.

Chapter 14. Model-based clustering algorithms. In the framework of model-based clustering algorithms, the data are assumed to come from a mixture of probability distributions, each of which represents a different cluster. There is a huge number of published works related to model-based clustering algorithms. In particular, there are more than 400 articles devoted to the development and discussion of the expectation-maximization (EM) algorithm. In this chapter, we introduce model-based clustering and present two model-based clustering algorithms: COOLCAT and STUCCO.

Chapter 15. Subspace clustering. Subspace clustering is a relatively new concept. After the first subspace clustering algorithm, CLIQUE, was published by the IBM group, many subspace clustering algorithms were developed and studied. One feature of the subspace clustering algorithms is that they are capable of identifying different clusters embedded in different subspaces of the high-dimensional data. Several subspace clustering algorithms are presented in this chapter, including the neural network–inspired algorithm PART.

Chapter 16. Miscellaneous algorithms. This chapter introduces some clustering algorithms for clustering time series, data streams, and transaction data. Proximity measures for these data and several related clustering algorithms are presented.

Chapter 17. Evaluation of clustering algorithms. Clustering is an unsupervised process and there are no predefined classes and no examples to show that the clusters found by the clustering algorithms are valid. Usually one or more validity criteria, presented in this chapter, are required to verify the clustering result of one algorithm or to compare the clustering results of different algorithms.

Chapter 18. Clustering gene expression data. As an application of cluster analysis, gene expression data clustering is introduced in this chapter. The background and similarity measures for gene expression data are introduced. Clustering a real set of gene expression data with the fuzzy subspace clustering (FSC) algorithm is presented.

Chapter 19. Data clustering in MATLAB. In this chapter, we show how to perform clustering in MATLAB in the following three aspects. Firstly, we introduce some MATLAB commands related to file operations, since the first thing to do about clustering is to load data into MATLAB, and data are usually stored in a text file. Secondly, we introduce MATLAB M-files, MEX-files, and MAT-files in order to demonstrate how to code algorithms and save current work. Finally, we present several MATLAB codes, which can be found in Appendix C.

Chapter 20. Clustering in C/C++. C++ is an object-oriented programming language built on the C language. In this last chapter of the book, we introduce the Standard Template Library (STL) in C++ and C/C++ program compilation. C++ data structure for data clustering is introduced. This chapter assumes that readers have basic knowledge of the C/C++ language.

This monograph has grown and evolved from a few collaborative projects for industrial applications undertaken by the Laboratory for Industrial and Applied Mathematics at York University, some of which are in collaboration with Generation 5 Mathematical Technologies, Inc. We would like to thank the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada's Discovery Grant Program and Collaborative Research Development Program, and Mathematics for Information Technology and Complex Systems for their support.

Index

- absolute-valued data, 331
- agglomerative algorithm, 9
- agglomerative hierarchical, 89
- agglomerative hierarchical algorithm, 103
- Andrew's wave estimate, 42
- asymmetric binary, 22
- attribute, 5, 19
- automatic categorization, *see*
 - categorization
- average distance, *see* distance

- BANG, 195
- banner, 106
- Bayesian information criterion (BIC), 217
- BD*-tree, 351
- binarization, 25
- binary attribute, 19
- BIRCH, 30, 134, 349
- Box-Cox transformation, 44
- BRIDGE, 200
- BSP*-tree, 351

- CACTUS, 186
- Canberra metric, 72
- cases by variables, 5
- categorical attribute, 20
- categorical variable, 29
- categorization, 29
 - automatic, 35
 - cluster-based, 30
 - direct, 29
- central-extreme principle, 27
- centroid, 91
- centroid method, 109
- chameleon, 185
- Chernoff bound, 134
- chord distance, *see* distance
- city block distance, *see* distance

- CLARANS, 30
- class, 6
- class-preserving projection, 56
- CLINK, 132
- CLIQUE, 222, 349
- CLTree, 222, 237, 349
- cluster, 3, 6
- cluster analysis, 3
- cluster-based categorization, *see*
 - categorization
- clustering, 3
- c*-means, 145
- coefficient of divergence, 72
- commutativity, 65
- compact cluster, 6
- compare-means, 151
- complete link, 91, 109
- contiguous, 33
- contingency table, 95
- continuous *k*-means, 150
- COOLCAT, 218
- COSA, 222
- cosine similarity measure, 88
- covariance matrix, 68
- crisp clustering, 8, 139
- CUBN, 204
- CURE, 30, 134, 349
- cutting index, 34
- Czekanowski coefficient, 72

- data clustering, 3
- data matrix, 41
- data mining, 4
 - direct, 4
 - indirect, 4
- data point, 5
- data scale, 19
- data transformation, 44

- DBCLASD, 201
- DBCluC, 200
- DBSCAN, 199
- DENCLUE, 30, 203
- dendrogram, 58, 103, 104
- density-based clustering, 199
- DIANA, 349
- dice coefficient, 75
- dichotomization, 25
- dichotomous tree, 104
- dimensionality, 5
- dimensionless, 41
- direct categorization, *see* categorization
- direct data mining, *see* data mining
- discrete attribute, 19
- DISMEA, 136, 349
- dissimilarity, 5
- dissimilarity function, 65
- dissimilarity measure, 65
- distance, 5
 - average, 70
 - chord, 71
 - city block, 69
 - Euclidean, 5, 68
 - generalized Mahalanobis, 70
 - generalized Minkowski, 77
 - geodesic, 71
 - intracluster, 36
 - Mahalanobis, 69
 - Manhattan, 69
 - Manhattan segmental, 69
 - maximum, 69
 - Minkowski, 69
 - simple matching, 71
 - statistical, 90
- distance function, 65
- distance matrix, 66
- divisive algorithm, 9
- divisive hierarchical, 89, 128
- divisive hierarchical algorithm, 103
- DOC, 222, 235, 349
- Dunn's index, 283
- dynamic programming, 33
- dynamic time warping, 83

- EM algorithm, 214
- ENCLUS, 222, 230
- entropy, 62
- error sum-of-squares (ESS), 123
- Euclidean distance, *see* distance

- exponential-family, 215
- external criteria, 277

- feature, 5
- FINDIT, 222, 232
- FLOC, 222
- Folkes and Mallows index, 279
- frequency table, 20
- fuzzy clustering, 7, 167
- fuzzy k -means, 142
- fuzzy k -partition, 7
- fuzzy set, 139

- gap statistic, 37
- Gaussian kernel, 255
- Gaussian mixture model, 209
- GDBSCAN, 200
- GDILC, 196
- gene expression data, 3
- general similarity coefficient, 76
- generalized Mahalanobis distance, *see* distance
- genetic algorithm, 167
- genetic k -means, 175
- genetic k -modes, 178
- geodesic distance, *see* distance
- global k -means, 177
- global standardization, 41
- graph-based clustering, 185
- greedy k -means, 178
- grid-based clustering, 191
- GRIDCLUS, 194
- group, 6
- group average, 91

- Hamann's coefficient, 75
- hard clustering, 6, 103
- hard k -partition, 7
- harmonic average, 156
- harmonic mean, 156
- Hessian matrix, 53
- hierarchical algorithm, 9, 103
- H -means, 174
- Hk -means, 174
- Huber's estimate, 42
- Hubert's Γ statistic, 279, 281

- icicle plot, 108
- index of association, 72
- indirect data mining, *see* data mining
- intercluster density, 284

- internal criteria, 277
- interval scale, 25
- intracluster distance, *see* distance
- intracluster variance, 284
- item, 5

- Jaccard coefficient, 75, 279
- jar file, 337
- J*-means, 173

- Karhunen–Loève transformation, 47
- kd*-tree, 152, 351
- kernel, 157
- k*-harmonic means, 156
- k*-means, 147
- k*-modes, 160
- k*-probabilities, 163
- k*-prototypes, 165
- Kuhn–Tucker problem, 254
- Kullback–Leibler divergence, 60
- Kulzinsky coefficient, 75, 76

- Lagrange multiplier, 250
- Lance–Williams formula, 91, 110
- lateral distance, 55
- LCS, 84
- least squares, 30
- level, 313
- link-based clustering, 185
- link-based similarity measure, 88
- location measure, 42
- log ratio data, 331
- longest common subsequence, 84
- loop plot, 109
- LSEARCH, 269

- MAFIA, 222, 234
- Manhattan distance, *see* distance
- mapping error, 51
- market segmentation, 4
- maximum-entropy clustering (MEC), 159
- mean character difference, 72
- mean shift, 157
- mean-square contingency, 96
- mean standardization, 42
- median, 42, 91
- median method, 109
- median standardization, 42
- metric, 65
- minimum sum of squares clustering (MSSC), 250

- missing value, 9
- modal variable, 23
- model-based clustering, 207
- monothetic, 128
- monotonic hierarchy, 110
- multidimensional scaling (MDS), 52

- n*-tree, 104
- nominal attribute, 20
- nominal scale, 25
- nonranked tree, 104
- normalization, 19, 41
- normalized Γ statistic, 279
- NP-hard, 129
- numerical variable, 29

- OAK, 275
- object, 5
- objective evaluation, 87
- observation, 5
- Ochiai coefficient, 76
- OptiGrid, 192
- ORCLUS, 222, 226
- ordinal scale, 25

- packed representation, 107
- parallel coordinates, 57
- PART, 222, 238
- partition entropy index, 291
- partitioning algorithm, 9, 103
- pattern, 5
- Pearson’s coefficient, 75
- pointer representation, 106
- polythetic, 128
- power transformation, 44
- principal component analysis (PCA), 44
- PROCLUS, 222, 224
- projected cluster, 221
- proximity, 19
- proximity graph, 67
- proximity index, 66
- proximity matrix, 66
- proximity relation, 52
- Python, 307

- quadtree, 352
- qualitative scale, 19
- quantitative scale, 19
- quick sort, 32

- R, 299

- Rand statistic, 279
range standardization, 42
rank, 44
ratio scale, 25
reflection, 323
reflexivity, 65
relative criteria, 277
RMSSDT index, 285
ROCK, 188
Rogers–Tanimoto coefficient, 75
RS index, 286
Russell–Rao coefficient, 75, 76
- Sammon’s mapping, 51
scale, 19, 25
scale conversion, 25
scale measure, 42
scatter matrix, 67
Schwarz criterion, 155
SD validity, 283
segmentation analysis, 3
self-organizing map (SOM), 54
set-valued, 23
silhouette plots, 109
similarity, 5
similarity coefficient, 5, 65
similarity dichotomy, 66
similarity function, 65
similarity matrix, 66
similarity measure, 5, 65
similarity trichotomy, 66
simple matching coefficient, 75
single-link, 91, 109
singular value decomposition (SVD), 45
skyline plot, 109
SLINK, 129
soft clustering, 8
Sokal–Michener coefficient, 75
Sokal–Sneath coefficient, 75
Sørensen coefficient, 76
sort-means, 151
spanning tree, 131
Spearman’s rank correlation, 28
standard deviation, 42
standardization, 41
state, 20
statistical distance, *see* distance
STING, 191
STUCCO, 219
SUBCAD, 239, 350
subjective evaluation, 87
subspace clustering, 221
substitution, 25
sum of squared distance (SSD), 32
supervised learning, 4
support, 89
symbol table, 20
symbolic data, 23
symmetric binary, 22
- tabu, 169
tabu search, 169
taxonomy analysis, 3
temporal data, 24
term-document matrix, 23
transaction data, 23
transformation, 41
tree map, 58
triangle inequality, 65
trimmed k -means, 154
truncated fuzzy c -means, 264
truncated fuzzy partition matrix, 264
 t -SNE, 59
Tukey’s biweight estimate, 42
tuple, 5
- ultrametric condition, 105
ultrametric relation, 132
unsupervised classification, 3
unsupervised learning, 4
UPGMA, 115
- valued tree, 104
variable, 5
variable annuity, 341
variable neighborhood search (VNS), 170
- WAND- k -means, 262
Ward’s method, 91, 109
WaveCluster, 197
weighted centroid, 91
weighted distance, 335
weighted group average, 91
WEKA, 308
within-cluster standardization, 41
within-group sum of squares (WGSS), 145
- x -means, 155
- Yule coefficient, 75
- z -score, 42