

# DATA CLUSTERING

## Algorithms and Applications

Edited by

**Charu C. Aggarwal**  
**Chandan K. Reddy**



**CRC Press**

Taylor & Francis Group  
Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business  
A CHAPMAN & HALL BOOK

---

# Contents

<b>Preface</b>	<b>xxi</b>
<b>Editor Biographies</b>	<b>xxiii</b>
<b>Contributors</b>	<b>xxv</b>
<b>1 An Introduction to Cluster Analysis</b>	<b>1</b>
<i>Charu C. Aggarwal</i>	
1.1 Introduction . . . . .	2
1.2 Common Techniques Used in Cluster Analysis . . . . .	3
1.2.1 Feature Selection Methods . . . . .	4
1.2.2 Probabilistic and Generative Models . . . . .	4
1.2.3 Distance-Based Algorithms . . . . .	5
1.2.4 Density- and Grid-Based Methods . . . . .	7
1.2.5 Leveraging Dimensionality Reduction Methods . . . . .	8
1.2.5.1 Generative Models for Dimensionality Reduction . . . . .	8
1.2.5.2 Matrix Factorization and Co-Clustering . . . . .	8
1.2.5.3 Spectral Methods . . . . .	10
1.2.6 The High Dimensional Scenario . . . . .	11
1.2.7 Scalable Techniques for Cluster Analysis . . . . .	13
1.2.7.1 I/O Issues in Database Management . . . . .	13
1.2.7.2 Streaming Algorithms . . . . .	14
1.2.7.3 The Big Data Framework . . . . .	14
1.3 Data Types Studied in Cluster Analysis . . . . .	15
1.3.1 Clustering Categorical Data . . . . .	15
1.3.2 Clustering Text Data . . . . .	16
1.3.3 Clustering Multimedia Data . . . . .	16
1.3.4 Clustering Time-Series Data . . . . .	17
1.3.5 Clustering Discrete Sequences . . . . .	17
1.3.6 Clustering Network Data . . . . .	18
1.3.7 Clustering Uncertain Data . . . . .	19
1.4 Insights Gained from Different Variations of Cluster Analysis . . . . .	19
1.4.1 Visual Insights . . . . .	20
1.4.2 Supervised Insights . . . . .	20
1.4.3 Multiview and Ensemble-Based Insights . . . . .	21
1.4.4 Validation-Based Insights . . . . .	21
1.5 Discussion and Conclusions . . . . .	22

<b>2</b>	<b>Feature Selection for Clustering: A Review</b>	<b>29</b>
	<i>Salem Alelyani, Jiliang Tang, and Huan Liu</i>	
2.1	Introduction . . . . .	30
2.1.1	Data Clustering . . . . .	32
2.1.2	Feature Selection . . . . .	32
2.1.3	Feature Selection for Clustering . . . . .	33
2.1.3.1	Filter Model . . . . .	34
2.1.3.2	Wrapper Model . . . . .	35
2.1.3.3	Hybrid Model . . . . .	35
2.2	Feature Selection for Clustering . . . . .	35
2.2.1	Algorithms for Generic Data . . . . .	36
2.2.1.1	Spectral Feature Selection (SPEC) . . . . .	36
2.2.1.2	Laplacian Score (LS) . . . . .	36
2.2.1.3	Feature Selection for Sparse Clustering . . . . .	37
2.2.1.4	Localized Feature Selection Based on Scatter Separability (LFSBSS) . . . . .	38
2.2.1.5	Multicluster Feature Selection (MCFS) . . . . .	39
2.2.1.6	Feature Weighting $k$ -Means . . . . .	40
2.2.2	Algorithms for Text Data . . . . .	41
2.2.2.1	Term Frequency (TF) . . . . .	41
2.2.2.2	Inverse Document Frequency (IDF) . . . . .	42
2.2.2.3	Term Frequency-Inverse Document Frequency (TF-IDF) . . . . .	42
2.2.2.4	Chi Square Statistic . . . . .	42
2.2.2.5	Frequent Term-Based Text Clustering . . . . .	44
2.2.2.6	Frequent Term Sequence . . . . .	45
2.2.3	Algorithms for Streaming Data . . . . .	47
2.2.3.1	Text Stream Clustering Based on Adaptive Feature Selection (TSC-AFS) . . . . .	47
2.2.3.2	High-Dimensional Projected Stream Clustering (HPStream) . . . . .	48
2.2.4	Algorithms for Linked Data . . . . .	50
2.2.4.1	Challenges and Opportunities . . . . .	50
2.2.4.2	LUFS: An Unsupervised Feature Selection Framework for Linked Data . . . . .	51
2.2.4.3	Conclusion and Future Work for Linked Data . . . . .	52
2.3	Discussions and Challenges . . . . .	53
2.3.1	The Chicken or the Egg Dilemma . . . . .	53
2.3.2	Model Selection: $K$ and $l$ . . . . .	54
2.3.3	Scalability . . . . .	54
2.3.4	Stability . . . . .	55
<b>3</b>	<b>Probabilistic Models for Clustering</b>	<b>61</b>
	<i>Hongbo Deng and Jiawei Han</i>	
3.1	Introduction . . . . .	61
3.2	Mixture Models . . . . .	62
3.2.1	Overview . . . . .	62
3.2.2	Gaussian Mixture Model . . . . .	64
3.2.3	Bernoulli Mixture Model . . . . .	67
3.2.4	Model Selection Criteria . . . . .	68
3.3	EM Algorithm and Its Variations . . . . .	69
3.3.1	The General EM Algorithm . . . . .	69
3.3.2	Mixture Models Revisited . . . . .	73

3.3.3 Limitations of the EM Algorithm . . . . . 75

3.3.4 Applications of the EM Algorithm . . . . . 76

3.4 Probabilistic Topic Models . . . . . 76

3.4.1 Probabilistic Latent Semantic Analysis . . . . . 77

3.4.2 Latent Dirichlet Allocation . . . . . 79

3.4.3 Variations and Extensions . . . . . 81

3.5 Conclusions and Summary . . . . . 81

**4 A Survey of Partitional and Hierarchical Clustering Algorithms 87**

*Chandan K. Reddy and Bhanukiran Vinzamuri*

4.1 Introduction . . . . . 88

4.2 Partitional Clustering Algorithms . . . . . 89

4.2.1 *K*-Means Clustering . . . . . 89

4.2.2 Minimization of Sum of Squared Errors . . . . . 90

4.2.3 Factors Affecting *K*-Means . . . . . 91

4.2.3.1 Popular Initialization Methods . . . . . 91

4.2.3.2 Estimating the Number of Clusters . . . . . 92

4.2.4 Variations of *K*-Means . . . . . 93

4.2.4.1 *K*-Medoids Clustering . . . . . 93

4.2.4.2 *K*-Medians Clustering . . . . . 94

4.2.4.3 *K*-Modes Clustering . . . . . 94

4.2.4.4 Fuzzy *K*-Means Clustering . . . . . 95

4.2.4.5 *X*-Means Clustering . . . . . 95

4.2.4.6 Intelligent *K*-Means Clustering . . . . . 96

4.2.4.7 Bisecting *K*-Means Clustering . . . . . 97

4.2.4.8 Kernel *K*-Means Clustering . . . . . 97

4.2.4.9 Mean Shift Clustering . . . . . 98

4.2.4.10 Weighted *K*-Means Clustering . . . . . 98

4.2.4.11 Genetic *K*-Means Clustering . . . . . 99

4.2.5 Making *K*-Means Faster . . . . . 100

4.3 Hierarchical Clustering Algorithms . . . . . 100

4.3.1 Agglomerative Clustering . . . . . 101

4.3.1.1 Single and Complete Link . . . . . 101

4.3.1.2 Group Averaged and Centroid Agglomerative Clustering . . . . . 102

4.3.1.3 Ward’s Criterion . . . . . 103

4.3.1.4 Agglomerative Hierarchical Clustering Algorithm . . . . . 103

4.3.1.5 Lance–Williams Dissimilarity Update Formula . . . . . 103

4.3.2 Divisive Clustering . . . . . 104

4.3.2.1 Issues in Divisive Clustering . . . . . 104

4.3.2.2 Divisive Hierarchical Clustering Algorithm . . . . . 105

4.3.2.3 Minimum Spanning Tree-Based Clustering . . . . . 105

4.3.3 Other Hierarchical Clustering Algorithms . . . . . 106

4.4 Discussion and Summary . . . . . 106

**5 Density-Based Clustering 111**

*Martin Ester*

5.1 Introduction . . . . . 111

5.2 DBSCAN . . . . . 113

5.3 DENCLUE . . . . . 115

5.4 OPTICS . . . . . 116

5.5 Other Algorithms . . . . . 116

5.6	Subspace Clustering . . . . .	118
5.7	Clustering Networks . . . . .	120
5.8	Other Directions . . . . .	123
5.9	Conclusion . . . . .	124
<b>6</b>	<b>Grid-Based Clustering</b> . . . . .	<b>127</b>
	<i>Wei Cheng, Wei Wang, and Sandra Batista</i>	
6.1	Introduction . . . . .	128
6.2	The Classical Algorithms . . . . .	131
6.2.1	Earliest Approaches: GRIDCLUS and BANG . . . . .	131
6.2.2	STING and STING+: The Statistical Information Grid Approach . . . . .	132
6.2.3	WaveCluster: Wavelets in Grid-Based Clustering . . . . .	134
6.3	Adaptive Grid-Based Algorithms . . . . .	135
6.3.1	AMR: Adaptive Mesh Refinement Clustering . . . . .	135
6.4	Axis-Shifting Grid-Based Algorithms . . . . .	136
6.4.1	NSGC: New Shifting Grid Clustering Algorithm . . . . .	136
6.4.2	ADCC: Adaptable Deflect and Conquer Clustering . . . . .	137
6.4.3	ASGC: Axis-Shifted Grid-Clustering . . . . .	137
6.4.4	GDILC: Grid-Based Density-IsoLine Clustering Algorithm . . . . .	138
6.5	High-Dimensional Algorithms . . . . .	139
6.5.1	CLIQUE: The Classical High-Dimensional Algorithm . . . . .	139
6.5.2	Variants of CLIQUE . . . . .	140
6.5.2.1	ENCLUS: Entropy-Based Approach . . . . .	140
6.5.2.2	MAFIA: Adaptive Grids in High Dimensions . . . . .	141
6.5.3	OptiGrid: Density-Based Optimal Grid Partitioning . . . . .	141
6.5.4	Variants of the OptiGrid Approach . . . . .	143
6.5.4.1	O-Cluster: A Scalable Approach . . . . .	143
6.5.4.2	CBF: Cell-Based Filtering . . . . .	144
6.6	Conclusions and Summary . . . . .	145
<b>7</b>	<b>Nonnegative Matrix Factorizations for Clustering: A Survey</b> . . . . .	<b>149</b>
	<i>Tao Li and Chris Ding</i>	
7.1	Introduction . . . . .	150
7.1.1	Background . . . . .	150
7.1.2	NMF Formulations . . . . .	151
7.2	NMF for Clustering: Theoretical Foundations . . . . .	151
7.2.1	NMF and $K$ -Means Clustering . . . . .	151
7.2.2	NMF and Probabilistic Latent Semantic Indexing . . . . .	152
7.2.3	NMF and Kernel $K$ -Means and Spectral Clustering . . . . .	152
7.2.4	NMF Boundedness Theorem . . . . .	153
7.3	NMF Clustering Capabilities . . . . .	153
7.3.1	Examples . . . . .	153
7.3.2	Analysis . . . . .	153
7.4	NMF Algorithms . . . . .	155
7.4.1	Introduction . . . . .	155
7.4.2	Algorithm Development . . . . .	155
7.4.3	Practical Issues in NMF Algorithms . . . . .	156
7.4.3.1	Initialization . . . . .	156
7.4.3.2	Stopping Criteria . . . . .	156
7.4.3.3	Objective Function vs. Clustering Performance . . . . .	157
7.4.3.4	Scalability . . . . .	157

7.5	NMF Related Factorizations . . . . .	158
7.6	NMF for Clustering: Extensions . . . . .	161
7.6.1	Co-Clustering . . . . .	161
7.6.2	Semisupervised Clustering . . . . .	162
7.6.3	Semisupervised Co-Clustering . . . . .	162
7.6.4	Consensus Clustering . . . . .	163
7.6.5	Graph Clustering . . . . .	164
7.6.6	Other Clustering Extensions . . . . .	164
7.7	Conclusions . . . . .	165
<b>8</b>	<b>Spectral Clustering</b> . . . . .	<b>177</b>
	<i>Jialu Liu and Jiawei Han</i>	
8.1	Introduction . . . . .	177
8.2	Similarity Graph . . . . .	179
8.3	Unnormalized Spectral Clustering . . . . .	180
8.3.1	Notation . . . . .	180
8.3.2	Unnormalized Graph Laplacian . . . . .	180
8.3.3	Spectrum Analysis . . . . .	181
8.3.4	Unnormalized Spectral Clustering Algorithm . . . . .	182
8.4	Normalized Spectral Clustering . . . . .	182
8.4.1	Normalized Graph Laplacian . . . . .	183
8.4.2	Spectrum Analysis . . . . .	184
8.4.3	Normalized Spectral Clustering Algorithm . . . . .	184
8.5	Graph Cut View . . . . .	185
8.5.1	Ratio Cut Relaxation . . . . .	186
8.5.2	Normalized Cut Relaxation . . . . .	187
8.6	Random Walks View . . . . .	188
8.7	Connection to Laplacian Eigenmap . . . . .	189
8.8	Connection to Kernel $k$ -Means and Nonnegative Matrix Factorization . . . . .	191
8.9	Large Scale Spectral Clustering . . . . .	192
8.10	Further Reading . . . . .	194
<b>9</b>	<b>Clustering High-Dimensional Data</b> . . . . .	<b>201</b>
	<i>Arthur Zimek</i>	
9.1	Introduction . . . . .	201
9.2	The “Curse of Dimensionality” . . . . .	202
9.2.1	Different Aspects of the “Curse” . . . . .	202
9.2.2	Consequences . . . . .	206
9.3	Clustering Tasks in Subspaces of High-Dimensional Data . . . . .	206
9.3.1	Categories of Subspaces . . . . .	206
9.3.1.1	Axis-Parallel Subspaces . . . . .	206
9.3.1.2	Arbitrarily Oriented Subspaces . . . . .	207
9.3.1.3	Special Cases . . . . .	207
9.3.2	Search Spaces for the Clustering Problem . . . . .	207
9.4	Fundamental Algorithmic Ideas . . . . .	208
9.4.1	Clustering in Axis-Parallel Subspaces . . . . .	208
9.4.1.1	Cluster Model . . . . .	208
9.4.1.2	Basic Techniques . . . . .	208
9.4.1.3	Clustering Algorithms . . . . .	210
9.4.2	Clustering in Arbitrarily Oriented Subspaces . . . . .	215
9.4.2.1	Cluster Model . . . . .	215

9.4.2.2 Basic Techniques and Example Algorithms . . . . . 216

9.5 Open Questions and Current Research Directions . . . . . 218

9.6 Conclusion . . . . . 219

**10 A Survey of Stream Clustering Algorithms . . . . . 231**

*Charu C. Aggarwal*

10.1 Introduction . . . . . 231

10.2 Methods Based on Partitioning Representatives . . . . . 233

10.2.1 The STREAM Algorithm . . . . . 233

10.2.2 CluStream: The Microclustering Framework . . . . . 235

10.2.2.1 Microcluster Definition . . . . . 235

10.2.2.2 Pyramidal Time Frame . . . . . 236

10.2.2.3 Online Clustering with CluStream . . . . . 237

10.3 Density-Based Stream Clustering . . . . . 239

10.3.1 DenStream: Density-Based Microclustering . . . . . 240

10.3.2 Grid-Based Streaming Algorithms . . . . . 241

10.3.2.1 D-Stream Algorithm . . . . . 241

10.3.2.2 Other Grid-Based Algorithms . . . . . 242

10.4 Probabilistic Streaming Algorithms . . . . . 243

10.5 Clustering High-Dimensional Streams . . . . . 243

10.5.1 The HPSTREAM Method . . . . . 244

10.5.2 Other High-Dimensional Streaming Algorithms . . . . . 244

10.6 Clustering Discrete and Categorical Streams . . . . . 245

10.6.1 Clustering Binary Data Streams with *k*-Means . . . . . 245

10.6.2 The StreamCluCD Algorithm . . . . . 245

10.6.3 Massive-Domain Clustering . . . . . 246

10.7 Text Stream Clustering . . . . . 249

10.8 Other Scenarios for Stream Clustering . . . . . 252

10.8.1 Clustering Uncertain Data Streams . . . . . 253

10.8.2 Clustering Graph Streams . . . . . 253

10.8.3 Distributed Clustering of Data Streams . . . . . 254

10.9 Discussion and Conclusions . . . . . 254

**11 Big Data Clustering . . . . . 259**

*Hanghang Tong and U Kang*

11.1 Introduction . . . . . 259

11.2 One-Pass Clustering Algorithms . . . . . 260

11.2.1 CLARANS: Fighting with Exponential Search Space . . . . . 260

11.2.2 BIRCH: Fighting with Limited Memory . . . . . 261

11.2.3 CURE: Fighting with the Irregular Clusters . . . . . 263

11.3 Randomized Techniques for Clustering Algorithms . . . . . 263

11.3.1 Locality-Preserving Projection . . . . . 264

11.3.2 Global Projection . . . . . 266

11.4 Parallel and Distributed Clustering Algorithms . . . . . 268

11.4.1 General Framework . . . . . 268

11.4.2 DBDC: Density-Based Clustering . . . . . 269

11.4.3 ParMETIS: Graph Partitioning . . . . . 269

11.4.4 PKMeans: *K*-Means with MapReduce . . . . . 270

11.4.5 DisCo: Co-Clustering with MapReduce . . . . . 271

11.4.6 BoW: Subspace Clustering with MapReduce . . . . . 272

11.5 Conclusion . . . . . 274

Downloaded by [107.194.173.43] at 23:22 01 August 2016

<b>12 Clustering Categorical Data</b>	<b>277</b>
<i>Bill Andreopoulos</i>	
12.1 Introduction . . . . .	278
12.2 Goals of Categorical Clustering . . . . .	279
12.2.1 Clustering Road Map . . . . .	280
12.3 Similarity Measures for Categorical Data . . . . .	282
12.3.1 The Hamming Distance in Categorical and Binary Data . . . . .	282
12.3.2 Probabilistic Measures . . . . .	283
12.3.3 Information-Theoretic Measures . . . . .	283
12.3.4 Context-Based Similarity Measures . . . . .	284
12.4 Descriptions of Algorithms . . . . .	284
12.4.1 Partition-Based Clustering . . . . .	284
12.4.1.1 <i>k</i> -Modes . . . . .	284
12.4.1.2 <i>k</i> -Prototypes (Mixed Categorical and Numerical) . . . . .	285
12.4.1.3 Fuzzy <i>k</i> -Modes . . . . .	286
12.4.1.4 Squeezer . . . . .	286
12.4.1.5 COOLCAT . . . . .	286
12.4.2 Hierarchical Clustering . . . . .	287
12.4.2.1 ROCK . . . . .	287
12.4.2.2 COBWEB . . . . .	288
12.4.2.3 LIMBO . . . . .	289
12.4.3 Density-Based Clustering . . . . .	289
12.4.3.1 Projected (Subspace) Clustering . . . . .	290
12.4.3.2 CACTUS . . . . .	290
12.4.3.3 CLICKS . . . . .	291
12.4.3.4 STIRR . . . . .	291
12.4.3.5 CLOPE . . . . .	292
12.4.3.6 HIERDENC: Hierarchical Density-Based Clustering . . . . .	292
12.4.3.7 MULIC: Multiple Layer Incremental Clustering . . . . .	293
12.4.4 Model-Based Clustering . . . . .	296
12.4.4.1 BILCOM Empirical Bayesian (Mixed Categorical and Numerical) . . . . .	296
12.4.4.2 AutoClass (Mixed Categorical and Numerical) . . . . .	296
12.4.4.3 SVM Clustering (Mixed Categorical and Numerical) . . . . .	297
12.5 Conclusion . . . . .	298
<b>13 Document Clustering: The Next Frontier</b>	<b>305</b>
<i>David C. Anastasiu, Andrea Tagarelli, and George Karypis</i>	
13.1 Introduction . . . . .	306
13.2 Modeling a Document . . . . .	306
13.2.1 Preliminaries . . . . .	306
13.2.2 The Vector Space Model . . . . .	307
13.2.3 Alternate Document Models . . . . .	309
13.2.4 Dimensionality Reduction for Text . . . . .	309
13.2.5 Characterizing Extremes . . . . .	310
13.3 General Purpose Document Clustering . . . . .	311
13.3.1 Similarity/Dissimilarity-Based Algorithms . . . . .	311
13.3.2 Density-Based Algorithms . . . . .	312
13.3.3 Adjacency-Based Algorithms . . . . .	313
13.3.4 Generative Algorithms . . . . .	313
13.4 Clustering Long Documents . . . . .	315

Downloaded by [107.194.173.43] at 23:22 01 August 2016



13.4.1	Document Segmentation . . . . .	315
13.4.2	Clustering Segmented Documents . . . . .	317
13.4.3	Simultaneous Segment Identification and Clustering . . . . .	321
13.5	Clustering Short Documents . . . . .	323
13.5.1	General Methods for Short Document Clustering . . . . .	323
13.5.2	Clustering with Knowledge Infusion . . . . .	324
13.5.3	Clustering Web Snippets . . . . .	325
13.5.4	Clustering Microblogs . . . . .	326
13.6	Conclusion . . . . .	328
<b>14</b>	<b>Clustering Multimedia Data</b>	<b>339</b>
	<i>Shen-Fu Tsai, Guo-Jun Qi, Shiyu Chang, Min-Hsuan Tsai, and Thomas S. Huang</i>	
14.1	Introduction . . . . .	340
14.2	Clustering with Image Data . . . . .	340
14.2.1	Visual Words Learning . . . . .	341
14.2.2	Face Clustering and Annotation . . . . .	342
14.2.3	Photo Album Event Recognition . . . . .	343
14.2.4	Image Segmentation . . . . .	344
14.2.5	Large-Scale Image Classification . . . . .	345
14.3	Clustering with Video and Audio Data . . . . .	347
14.3.1	Video Summarization . . . . .	348
14.3.2	Video Event Detection . . . . .	349
14.3.3	Video Story Clustering . . . . .	350
14.3.4	Music Summarization . . . . .	350
14.4	Clustering with Multimodal Data . . . . .	351
14.5	Summary and Future Directions . . . . .	353
<b>15</b>	<b>Time-Series Data Clustering</b>	<b>357</b>
	<i>Dimitrios Kotsakos, Goce Trajcevski, Dimitrios Gunopulos, and Charu C. Aggarwal</i>	
15.1	Introduction . . . . .	358
15.2	The Diverse Formulations for Time-Series Clustering . . . . .	359
15.3	Online Correlation-Based Clustering . . . . .	360
15.3.1	Selective Muscles and Related Methods . . . . .	361
15.3.2	Sensor Selection Algorithms for Correlation Clustering . . . . .	362
15.4	Similarity and Distance Measures . . . . .	363
15.4.1	Univariate Distance Measures . . . . .	363
15.4.1.1	$L_p$ Distance . . . . .	363
15.4.1.2	Dynamic Time Warping Distance . . . . .	364
15.4.1.3	EDIT Distance . . . . .	365
15.4.1.4	Longest Common Subsequence . . . . .	365
15.4.2	Multivariate Distance Measures . . . . .	366
15.4.2.1	Multidimensional $L_p$ Distance . . . . .	366
15.4.2.2	Multidimensional DTW . . . . .	367
15.4.2.3	Multidimensional LCSS . . . . .	368
15.4.2.4	Multidimensional Edit Distance . . . . .	368
15.4.2.5	Multidimensional Subsequence Matching . . . . .	368
15.5	Shape-Based Time-Series Clustering Techniques . . . . .	369
15.5.1	$k$ -Means Clustering . . . . .	370
15.5.2	Hierarchical Clustering . . . . .	371
15.5.3	Density-Based Clustering . . . . .	372

15.5.4 Trajectory Clustering . . . . . 372  
 15.6 Time-Series Clustering Applications . . . . . 374  
 15.7 Conclusions . . . . . 375

**16 Clustering Biological Data 381**

*Chandan K. Reddy, Mohammad Al Hasan, and Mohammed J. Zaki*

16.1 Introduction . . . . . 382  
 16.2 Clustering Microarray Data . . . . . 383  
     16.2.1 Proximity Measures . . . . . 383  
     16.2.2 Categorization of Algorithms . . . . . 384  
     16.2.3 Standard Clustering Algorithms . . . . . 385  
         16.2.3.1 Hierarchical Clustering . . . . . 385  
         16.2.3.2 Probabilistic Clustering . . . . . 386  
         16.2.3.3 Graph-Theoretic Clustering . . . . . 386  
         16.2.3.4 Self-Organizing Maps . . . . . 387  
         16.2.3.5 Other Clustering Methods . . . . . 387  
     16.2.4 Biclustering . . . . . 388  
         16.2.4.1 Types and Structures of Biclusters . . . . . 389  
         16.2.4.2 Biclustering Algorithms . . . . . 390  
         16.2.4.3 Recent Developments . . . . . 391  
     16.2.5 Triclustering . . . . . 391  
     16.2.6 Time-Series Gene Expression Data Clustering . . . . . 392  
     16.2.7 Cluster Validation . . . . . 393  
 16.3 Clustering Biological Networks . . . . . 394  
     16.3.1 Characteristics of PPI Network Data . . . . . 394  
     16.3.2 Network Clustering Algorithms . . . . . 394  
         16.3.2.1 Molecular Complex Detection . . . . . 394  
         16.3.2.2 Markov Clustering . . . . . 395  
         16.3.2.3 Neighborhood Search Methods . . . . . 395  
         16.3.2.4 Clique Percolation Method . . . . . 395  
         16.3.2.5 Ensemble Clustering . . . . . 396  
         16.3.2.6 Other Clustering Methods . . . . . 396  
     16.3.3 Cluster Validation and Challenges . . . . . 397  
 16.4 Biological Sequence Clustering . . . . . 397  
     16.4.1 Sequence Similarity Metrics . . . . . 397  
         16.4.1.1 Alignment-Based Similarity . . . . . 398  
         16.4.1.2 Keyword-Based Similarity . . . . . 398  
         16.4.1.3 Kernel-Based Similarity . . . . . 399  
         16.4.1.4 Model-Based Similarity . . . . . 399  
     16.4.2 Sequence Clustering Algorithms . . . . . 399  
         16.4.2.1 Subsequence-Based Clustering . . . . . 399  
         16.4.2.2 Graph-Based Clustering . . . . . 400  
         16.4.2.3 Probabilistic Models . . . . . 402  
         16.4.2.4 Suffix Tree and Suffix Array-Based Method . . . . . 403  
 16.5 Software Packages . . . . . 403  
 16.6 Discussion and Summary . . . . . 405

Downloaded by [107.194.173.43] at 23:22 01 August 2016

<b>17 Network Clustering</b>	<b>415</b>
<i>Srinivasan Parthasarathy and S M Faisal</i>	
17.1 Introduction . . . . .	416
17.2 Background and Nomenclature . . . . .	417
17.3 Problem Definition . . . . .	417
17.4 Common Evaluation Criteria . . . . .	418
17.5 Partitioning with Geometric Information . . . . .	419
17.5.1 Coordinate Bisection . . . . .	419
17.5.2 Inertial Bisection . . . . .	419
17.5.3 Geometric Partitioning . . . . .	420
17.6 Graph Growing and Greedy Algorithms . . . . .	421
17.6.1 Kernighan-Lin Algorithm . . . . .	422
17.7 Agglomerative and Divisive Clustering . . . . .	423
17.8 Spectral Clustering . . . . .	424
17.8.1 Similarity Graphs . . . . .	425
17.8.2 Types of Similarity Graphs . . . . .	425
17.8.3 Graph Laplacians . . . . .	426
17.8.3.1 Unnormalized Graph Laplacian . . . . .	426
17.8.3.2 Normalized Graph Laplacians . . . . .	427
17.8.4 Spectral Clustering Algorithms . . . . .	427
17.9 Markov Clustering . . . . .	428
17.9.1 Regularized MCL (RMCL): Improvement over MCL . . . . .	429
17.10 Multilevel Partitioning . . . . .	430
17.11 Local Partitioning Algorithms . . . . .	432
17.12 Hypergraph Partitioning . . . . .	433
17.13 Emerging Methods for Partitioning Special Graphs . . . . .	435
17.13.1 Bipartite Graphs . . . . .	435
17.13.2 Dynamic Graphs . . . . .	436
17.13.3 Heterogeneous Networks . . . . .	437
17.13.4 Directed Networks . . . . .	438
17.13.5 Combining Content and Relationship Information . . . . .	439
17.13.6 Networks with Overlapping Communities . . . . .	440
17.13.7 Probabilistic Methods . . . . .	442
17.14 Conclusion . . . . .	443
<b>18 A Survey of Uncertain Data Clustering Algorithms</b>	<b>457</b>
<i>Charu C. Aggarwal</i>	
18.1 Introduction . . . . .	457
18.2 Mixture Model Clustering of Uncertain Data . . . . .	459
18.3 Density-Based Clustering Algorithms . . . . .	460
18.3.1 FDBSCAN Algorithm . . . . .	460
18.3.2 FOPTICS Algorithm . . . . .	461
18.4 Partitional Clustering Algorithms . . . . .	462
18.4.1 The UK-Means Algorithm . . . . .	462
18.4.2 The CK-Means Algorithm . . . . .	463
18.4.3 Clustering Uncertain Data with Voronoi Diagrams . . . . .	464
18.4.4 Approximation Algorithms for Clustering Uncertain Data . . . . .	464
18.4.5 Speeding Up Distance Computations . . . . .	465
18.5 Clustering Uncertain Data Streams . . . . .	466
18.5.1 The UMicro Algorithm . . . . .	466
18.5.2 The LuMicro Algorithm . . . . .	471

18.5.3	Enhancements to Stream Clustering . . . . .	471
18.6	Clustering Uncertain Data in High Dimensionality . . . . .	472
18.6.1	Subspace Clustering of Uncertain Data . . . . .	473
18.6.2	UPStream: Projected Clustering of Uncertain Data Streams . . . . .	474
18.7	Clustering with the Possible Worlds Model . . . . .	477
18.8	Clustering Uncertain Graphs . . . . .	478
18.9	Conclusions and Summary . . . . .	478
<b>19</b>	<b>Concepts of Visual and Interactive Clustering</b>	<b>483</b>
	<i>Alexander Hinneburg</i>	
19.1	Introduction . . . . .	483
19.2	Direct Visual and Interactive Clustering . . . . .	484
19.2.1	Scatterplots . . . . .	485
19.2.2	Parallel Coordinates . . . . .	488
19.2.3	Discussion . . . . .	491
19.3	Visual Interactive Steering of Clustering . . . . .	491
19.3.1	Visual Assessment of Convergence of Clustering Algorithm . . . . .	491
19.3.2	Interactive Hierarchical Clustering . . . . .	492
19.3.3	Visual Clustering with SOMs . . . . .	494
19.3.4	Discussion . . . . .	494
19.4	Interactive Comparison and Combination of Clusterings . . . . .	495
19.4.1	Space of Clusterings . . . . .	495
19.4.2	Visualization . . . . .	497
19.4.3	Discussion . . . . .	497
19.5	Visualization of Clusters for Sense-Making . . . . .	497
19.6	Summary . . . . .	500
<b>20</b>	<b>Semisupervised Clustering</b>	<b>505</b>
	<i>Amrudin Agovic and Arindam Banerjee</i>	
20.1	Introduction . . . . .	506
20.2	Clustering with Pointwise and Pairwise Semisupervision . . . . .	507
20.2.1	Semisupervised Clustering Based on Seeding . . . . .	507
20.2.2	Semisupervised Clustering Based on Pairwise Constraints . . . . .	508
20.2.3	Active Learning for Semisupervised Clustering . . . . .	511
20.2.4	Semisupervised Clustering Based on User Feedback . . . . .	512
20.2.5	Semisupervised Clustering Based on Nonnegative Matrix Factorization . . . . .	513
20.3	Semisupervised Graph Cuts . . . . .	513
20.3.1	Semisupervised Unnormalized Cut . . . . .	515
20.3.2	Semisupervised Ratio Cut . . . . .	515
20.3.3	Semisupervised Normalized Cut . . . . .	516
20.4	A Unified View of Label Propagation . . . . .	517
20.4.1	Generalized Label Propagation . . . . .	517
20.4.2	Gaussian Fields . . . . .	517
20.4.3	Tikhonov Regularization (TIKREG) . . . . .	518
20.4.4	Local and Global Consistency . . . . .	518
20.4.5	Related Methods . . . . .	519
20.4.5.1	Cluster Kernels . . . . .	519
20.4.5.2	Gaussian Random Walks EM (GWEM) . . . . .	519
20.4.5.3	Linear Neighborhood Propagation . . . . .	520
20.4.6	Label Propagation and Green's Function . . . . .	521
20.4.7	Label Propagation and Semisupervised Graph Cuts . . . . .	521

Downloaded by [107.194.173.43] at 23:22 01 August 2016

20.5	Semisupervised Embedding . . . . .	521
20.5.1	Nonlinear Manifold Embedding . . . . .	522
20.5.2	Semisupervised Embedding . . . . .	522
20.5.2.1	Unconstrained Semisupervised Embedding . . . . .	523
20.5.2.2	Constrained Semisupervised Embedding . . . . .	523
20.6	Comparative Experimental Analysis . . . . .	524
20.6.1	Experimental Results . . . . .	524
20.6.2	Semisupervised Embedding Methods . . . . .	529
20.7	Conclusions . . . . .	530
<b>21</b>	<b>Alternative Clustering Analysis: A Review</b>	<b>535</b>
	<i>James Bailey</i>	
21.1	Introduction . . . . .	535
21.2	Technical Preliminaries . . . . .	537
21.3	Multiple Clustering Analysis Using Alternative Clusterings . . . . .	538
21.3.1	Alternative Clustering Algorithms: A Taxonomy . . . . .	538
21.3.2	Unguided Generation . . . . .	539
21.3.2.1	Naive . . . . .	539
21.3.2.2	Meta Clustering . . . . .	539
21.3.2.3	Eigenvectors of the Laplacian Matrix . . . . .	540
21.3.2.4	Decorrelated $k$ -Means and Convolutional EM . . . . .	540
21.3.2.5	CAMI . . . . .	540
21.3.3	Guided Generation with Constraints . . . . .	541
21.3.3.1	COALA . . . . .	541
21.3.3.2	Constrained Optimization Approach . . . . .	541
21.3.3.3	MAXIMUS . . . . .	542
21.3.4	Orthogonal Transformation Approaches . . . . .	543
21.3.4.1	Orthogonal Views . . . . .	543
21.3.4.2	ADFT . . . . .	543
21.3.5	Information Theoretic . . . . .	544
21.3.5.1	Conditional Information Bottleneck (CIB) . . . . .	544
21.3.5.2	Conditional Ensemble Clustering . . . . .	544
21.3.5.3	NACI . . . . .	544
21.3.5.4	mSC . . . . .	545
21.4	Connections to Multiview Clustering and Subspace Clustering . . . . .	545
21.5	Future Research Issues . . . . .	547
21.6	Summary . . . . .	547
<b>22</b>	<b>Cluster Ensembles: Theory and Applications</b>	<b>551</b>
	<i>Joydeep Ghosh and Ayan Acharya</i>	
22.1	Introduction . . . . .	551
22.2	The Cluster Ensemble Problem . . . . .	554
22.3	Measuring Similarity Between Clustering Solutions . . . . .	555
22.4	Cluster Ensemble Algorithms . . . . .	558
22.4.1	Probabilistic Approaches to Cluster Ensembles . . . . .	558
22.4.1.1	A Mixture Model for Cluster Ensembles (MMCE) . . . . .	558
22.4.1.2	Bayesian Cluster Ensembles (BCE) . . . . .	558
22.4.1.3	Nonparametric Bayesian Cluster Ensembles (NPBCE) . . . . .	559
22.4.2	Pairwise Similarity-Based Approaches . . . . .	560
22.4.2.1	Methods Based on Ensemble Co-Association Matrix . . . . .	560

- 22.4.2.2 Relating Consensus Clustering to Other Optimization Formulations . . . . . 562
- 22.4.3 Direct Approaches Using Cluster Labels . . . . . 562
  - 22.4.3.1 Graph Partitioning . . . . . 562
  - 22.4.3.2 Cumulative Voting . . . . . 563
- 22.5 Applications of Consensus Clustering . . . . . 564
  - 22.5.1 Gene Expression Data Analysis . . . . . 564
  - 22.5.2 Image Segmentation . . . . . 564
- 22.6 Concluding Remarks . . . . . 566

**23 Clustering Validation Measures 571**

*Hui Xiong and Zhongmou Li*

- 23.1 Introduction . . . . . 572
- 23.2 External Clustering Validation Measures . . . . . 573
  - 23.2.1 An Overview of External Clustering Validation Measures . . . . . 574
  - 23.2.2 Defective Validation Measures . . . . . 575
    - 23.2.2.1 *K*-Means: The Uniform Effect . . . . . 575
    - 23.2.2.2 A Necessary Selection Criterion . . . . . 576
    - 23.2.2.3 The Cluster Validation Results . . . . . 576
    - 23.2.2.4 The Issues with the Defective Measures . . . . . 577
    - 23.2.2.5 Improving the Defective Measures . . . . . 577
  - 23.2.3 Measure Normalization . . . . . 577
    - 23.2.3.1 Normalizing the Measures . . . . . 578
    - 23.2.3.2 The *DCV* Criterion . . . . . 581
    - 23.2.3.3 The Effect of Normalization . . . . . 583
  - 23.2.4 Measure Properties . . . . . 584
    - 23.2.4.1 The Consistency Between Measures . . . . . 584
    - 23.2.4.2 Properties of Measures . . . . . 586
    - 23.2.4.3 Discussions . . . . . 589
- 23.3 Internal Clustering Validation Measures . . . . . 589
  - 23.3.1 An Overview of Internal Clustering Validation Measures . . . . . 589
  - 23.3.2 Understanding of Internal Clustering Validation Measures . . . . . 592
    - 23.3.2.1 The Impact of Monotonicity . . . . . 592
    - 23.3.2.2 The Impact of Noise . . . . . 593
    - 23.3.2.3 The Impact of Density . . . . . 594
    - 23.3.2.4 The Impact of Subclusters . . . . . 595
    - 23.3.2.5 The Impact of Skewed Distributions . . . . . 596
    - 23.3.2.6 The Impact of Arbitrary Shapes . . . . . 598
  - 23.3.3 Properties of Measures . . . . . 600
- 23.4 Summary . . . . . 601

**24 Educational and Software Resources for Data Clustering 607**

*Charu C. Aggarwal and Chandan K. Reddy*

- 24.1 Introduction . . . . . 607
- 24.2 Educational Resources . . . . . 608
  - 24.2.1 Books on Data Clustering . . . . . 608
  - 24.2.2 Popular Survey Papers on Data Clustering . . . . . 608
- 24.3 Software for Data Clustering . . . . . 610
  - 24.3.1 Free and Open-Source Software . . . . . 610
    - 24.3.1.1 General Clustering Software . . . . . 610
    - 24.3.1.2 Specialized Clustering Software . . . . . 610

Downloaded by [107.194.173.43] at 23:22 01 August 2016

24.3.2	Commercial Packages . . . . .	611
24.3.3	Data Benchmarks for Software and Research . . . . .	611
24.4	Summary . . . . .	612
<b>Index</b>		<b>617</b>

Downloaded by [107.194.173.43] at 23:22 01 August 2016

---

# Preface

The problem of clustering is perhaps one of the most widely studied in the data mining and machine learning communities. This problem has been studied by researchers from several disciplines over five decades. Applications of clustering include a wide variety of problem domains such as text, multimedia, social networks, and biological data. Furthermore, the problem may be encountered in a number of different scenarios such as streaming or uncertain data. Clustering is a rather diverse topic, and the underlying algorithms depend greatly on the data domain and problem scenario.

Therefore, this book will focus on three primary aspects of data clustering. The first set of chapters will focus on the core methods for data clustering. These include methods such as probabilistic clustering, density-based clustering, grid-based clustering, and spectral clustering. The second set of chapters will focus on different problem domains and scenarios such as multimedia data, text data, biological data, categorical data, network data, data streams and uncertain data. The third set of chapters will focus on different detailed insights from the clustering process, because of the subjectivity of the clustering process, and the many different ways in which the same data set can be clustered. How do we know that a particular clustering is good or that it solves the needs of the application? There are numerous ways in which these issues can be explored. The exploration could be through interactive visualization and human interaction, external knowledge-based supervision, explicitly examining the multiple solutions in order to evaluate different possibilities, combining the multiple solutions in order to create more robust ensembles, or trying to judge the quality of different solutions with the use of different validation criteria.

The clustering problem has been addressed by a number of different communities such as pattern recognition, databases, data mining and machine learning. In some cases, the work by the different communities tends to be fragmented and has not been addressed in a unified way. This book will make a conscious effort to address the work of the different communities in a unified way. The book will start off with an overview of the basic methods in data clustering, and then discuss progressively more refined and complex methods for data clustering. Special attention will also be paid to more recent problem domains such as graphs and social networks.

The chapters in the book will be divided into three types:

- **Method Chapters:** These chapters discuss the *key techniques* which are commonly used for clustering such as feature selection, agglomerative clustering, partitional clustering, density-based clustering, probabilistic clustering, grid-based clustering, spectral clustering, and non-negative matrix factorization.
- **Domain Chapters:** These chapters discuss the specific methods used for different *domains* of data such as categorical data, text data, multimedia data, graph data, biological data, stream data, uncertain data, time series clustering, high-dimensional clustering, and big data. Many of these chapters can also be considered application chapters, because they explore the specific characteristics of the problem in a particular domain.
- **Variations and Insights:** These chapters discuss the *key variations* on the clustering process such as semi-supervised clustering, interactive clustering, multi-view clustering, cluster ensembles, and cluster validation. Such methods are typically used in order to obtain detailed insights from the clustering process, and also to explore different possibilities on the clustering process through either supervision, human intervention, or through automated generation



of alternative clusters. The methods for cluster validation also provide an idea of the quality of the underlying clusters.

This book is designed to be comprehensive in its coverage of the entire area of clustering, and it is hoped that it will serve as a knowledgeable compendium to students and researchers.

---

## **Editor Biographies**

**Charu C. Aggarwal** is a Research Scientist at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from Massachusetts Institute of Technology in 1996. His research interest during his Ph.D. years was in combinatorial optimization (network flow algorithms), and his thesis advisor was Professor James B. Orlin. He has since worked in the field of performance analysis, databases, and data mining. He has published over 200 papers in refereed conferences and journals, and has applied for or been granted over 80 patents. He is author or editor of nine books, including this one. Because of the commercial value of the above-mentioned patents, he has received several invention achievement awards and has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bioterrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of an IBM Research Division Award (2008) for his scientific contributions to data stream research. He has served on the program committees of most major database/data mining conferences, and served as program vice-chairs of the SIAM Conference on Data Mining (2007), the IEEE ICDM Conference (2007), the WWW Conference (2009), and the IEEE ICDM Conference (2009). He served as an associate editor of the *IEEE Transactions on Knowledge and Data Engineering Journal* from 2004 to 2008. He is an associate editor of the *ACM TKDD Journal*, an action editor of the *Data Mining and Knowledge Discovery Journal*, an associate editor of *ACM SIGKDD Explorations*, and an associate editor of the *Knowledge and Information Systems Journal*. He is a fellow of the IEEE for “contributions to knowledge discovery and data mining techniques,” and a life-member of the ACM.

**Chandan K. Reddy** is an Assistant Professor in the Department of Computer Science at Wayne State University. He received his Ph.D. from Cornell University and M.S. from Michigan State University. His primary research interests are in the areas of data mining and machine learning with applications to healthcare, bioinformatics, and social network analysis. His research is funded by the National Science Foundation, the National Institutes of Health, Department of Transportation, and the Susan G. Komen for the Cure Foundation. He has published over 40 peer-reviewed articles in leading conferences and journals. He received the Best Application Paper Award at the ACM SIGKDD conference in 2010 and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a member of IEEE, ACM, and SIAM.

---

## **Contributors**

### **Ayan Acharya**

University of Texas  
Austin, Texas

### **Charu C. Aggarwal**

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

### **Amrudin Agovic**

Reliancy, LLC  
Saint Louis Park, Minnesota

### **Mohammad Al Hasan**

Indiana University - Purdue University  
Indianapolis, Indiana

### **Salem Alelyani**

Arizona State University  
Tempe, Arizona

### **David C. Anastasiu**

University of Minnesota  
Minneapolis, Minnesota

### **Bill Andreopoulos**

Lawrence Berkeley National Laboratory  
Berkeley, California

### **James Bailey**

The University of Melbourne  
Melbourne, Australia

### **Arindam Banerjee**

University of Minnesota  
Minneapolis, Minnesota

### **Sandra Batista**

Duke University  
Durham, North Carolina

### **Shiyu Chang**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

### **Wei Cheng**

University of North Carolina  
Chapel Hill, North Carolina

### **Hongbo Deng**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

### **Cha-charis Ding**

University of Texas  
Arlington, Texas

### **Martin Ester**

Simon Fraser University  
British Columbia, Canada

### **S M Faisal**

The Ohio State University  
Columbus, Ohio

### **Joydeep Ghosh**

University of Texas  
Austin, Texas

### **Dimitrios Gunopulos**

University of Athens  
Athens, Greece

### **Jiawei Han**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

### **Alexander Hinneburg**

Martin-Luther University  
Halle/Saale, Germany

### **Thomas S. Huang**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

### **U Kang**

KAIST  
Seoul, Korea

**George Karypis**

University of Minnesota  
Minneapolis, Minnesota

**Dimitrios Kotsakos**

University of Athens  
Athens, Greece

**Tao Li**

Florida International University  
Miami, Florida

**Zhongmou Li**

Rutgers University  
New Brunswick, New Jersey

**Huan Liu**

Arizona State University  
Tempe, Arizona

**Jialu Liu**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

**Srinivasan Parthasarathy**

The Ohio State University  
Columbus, Ohio

**Guo-Jun Qi**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

**Chandan K. Reddy**

Wayne State University  
Detroit, Michigan

**Andrea Tagarelli**

University of Calabria  
Arcavacata di Rende, Italy

**Jiliang Tang**

Arizona State University  
Tempe, Arizona

**Hanghang Tong**

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

**Goce Trajcevski**

Northwestern University  
Evanston, Illinois

**Min-Hsuan Tsai**

University of Illinois at Urbana-Champaign  
Urbana, Illinois

**Shen-Fu Tsai**

Microsoft Inc.  
Redmond, Washington

**Bhanukiran Vinzamuri**

Wayne State University  
Detroit, Michigan

**Wei Wang**

University of California  
Los Angeles, California

**Hui Xiong**

Rutgers University  
New Brunswick, New Jersey

**Mohammed J. Zaki**

Rensselaer Polytechnic Institute  
Troy, New York

**Arthur Zimek**

University of Alberta  
Edmonton, Canada