

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Cluster Analysis</b>	<b>9</b>
2.1	Formalising the Problem	13
2.2	Measures of Similarity/Dissimilarity	16
2.2.1	Comparing the Objects Having Quantitative Features	18
2.2.2	Comparing the Objects Having Qualitative Features	26
2.3	Hierarchical Methods of Cluster Analysis	29
2.4	Partitional Clustering	34
2.4.1	Criteria of Grouping Based on Dissimilarity	35
2.4.2	The Task of Cluster Analysis in Euclidean Space	36
2.4.3	Grouping According to Cluster Volume	45
2.4.4	Generalisations of the Task of Grouping	46
2.4.5	Relationship Between Partitional and Hierarchical Clustering	47
2.5	Other Methods of Cluster Analysis	48
2.5.1	Relational Methods	48
2.5.2	Graph and Spectral Methods	49
2.5.3	Relationship Between Clustering for Embedded and Relational Data Representations	50
2.5.4	Density-Based Methods	51
2.5.5	Grid-Based Clustering Algorithms	55
2.5.6	Model-Based Clustering	56
2.5.7	Potential (Kernel) Function Methods	56
2.5.8	Cluster Ensembles	59
2.6	Whether and When Grouping Is Difficult?	64

<b>3 Algorithms of Combinatorial Cluster Analysis</b>	67
3.1 $k$ -means Algorithm	68
3.1.1 The Batch Variant of the $k$ -means Algorithm	72
3.1.2 The Incremental Variant of the $k$ -means Algorithm	72
3.1.3 Initialisation Methods for the $k$ -means Algorithm	73
3.1.4 Enhancing the Efficiency of the $k$ -means Algorithm	79
3.1.5 Variants of the $k$ -means Algorithm	81
3.2 EM Algorithm	96
3.3 FCM: Fuzzy $c$ -means Algorithm	100
3.3.1 Basic Formulation	100
3.3.2 Basic FCM Algorithm	103
3.3.3 Measures of Quality of Fuzzy Partition	106
3.3.4 An Alternative Formulation	110
3.3.5 Modifications of the FCM Algorithm	111
3.4 Affinity Propagation	128
3.5 Higher Dimensional Cluster “Centres” for $k$ -means	130
3.6 Clustering in Subspaces via $k$ -means	132
3.7 Clustering of Subsets— $k$ -Bregman Bubble Clustering	135
3.8 Projective Clustering with $k$ -means	136
3.9 Random Projection	137
3.10 Subsampling	140
3.11 Clustering Evolving Over Time	142
3.11.1 Evolutionary Clustering	142
3.11.2 Streaming Clustering	143
3.11.3 Incremental Clustering	145
3.12 Co-clustering	145
3.13 Tensor Clustering	147
3.14 Manifold Clustering	149
3.15 Semisupervised Clustering	151
3.15.1 Similarity-Adapting Methods	152
3.15.2 Search-Adapting Methods	153
3.15.3 Target Variable Driven Methods	155
3.15.4 Weakened Classification Methods	156
3.15.5 Information Spreading Algorithms	157
3.15.6 Further Considerations	159
3.15.7 Evolutionary Clustering	161
<b>4 Cluster Quality Versus Choice of Parameters</b>	163
4.1 Preparing the Data	163
4.2 Setting the Number of Clusters	165
4.2.1 Simple Heuristics	167
4.2.2 Methods Consisting in the Use of Information Criteria	168

4.2.3	Clustergrams . . . . .	168
4.2.4	Minimal Spanning Trees . . . . .	169
4.3	Partition Quality Indexes . . . . .	170
4.4	Comparing Partitions . . . . .	173
4.4.1	Simple Methods of Comparing Partitions . . . . .	175
4.4.2	Methods Measuring Common Parts of Partitions . . . . .	176
4.4.3	Methods Using Mutual Information . . . . .	177
4.5	Cover Quality Measures . . . . .	179
<b>5</b>	<b>Spectral Clustering . . . . .</b>	<b>181</b>
5.1	Introduction . . . . .	181
5.2	Basic Notions . . . . .	184
5.2.1	Similarity Graphs . . . . .	185
5.2.2	Graph Laplacian . . . . .	187
5.2.3	Eigenvalues and Eigenvectors of Graph Laplacian . . . . .	195
5.2.4	Variational Characterization of Eigenvalues . . . . .	198
5.2.5	Random Walk on Graphs . . . . .	203
5.3	Spectral Partitioning . . . . .	209
5.3.1	Graph Bi-Partitioning . . . . .	210
5.3.2	$k$ -way Partitioning . . . . .	214
5.3.3	Isoperimetric Inequalities . . . . .	224
5.3.4	Clustering Using Random Walk . . . . .	225
5.3.5	Total Variation Methods . . . . .	229
5.3.6	Out of Sample Spectral Clustering . . . . .	233
5.3.7	Incremental Spectral Clustering . . . . .	238
5.3.8	Nodal Sets and Nodal Domains . . . . .	239
5.4	Local Methods . . . . .	241
5.4.1	The Nibble Algorithm . . . . .	242
5.4.2	The PageRank-Nibble Algorithm . . . . .	244
5.5	Large Datasets . . . . .	247
5.5.1	Using a Sampling Technique . . . . .	248
5.5.2	Landmark-Based Spectral Clustering . . . . .	250
5.5.3	Randomized SVD . . . . .	253
5.5.4	Incomplete Cholesky Decomposition . . . . .	254
5.5.5	Compressive Spectral Clustering . . . . .	256
<b>6</b>	<b>Community Discovery and Identification in Empirical Graphs . . . . .</b>	<b>261</b>
6.1	The Concept of the Community . . . . .	263
6.1.1	Local Definitions . . . . .	264
6.1.2	Global Definitions . . . . .	265
6.1.3	Node Similarity Based Definitions . . . . .	265
6.1.4	Probabilistic Labelling Based Definitions . . . . .	265
6.2	Structure-based Similarity in Complex Networks . . . . .	266
6.2.1	Local Measures . . . . .	266

6.2.2	Global Measures . . . . .	268
6.2.3	Quasi-Local Indices . . . . .	270
6.3	Modularity—A Quality Measure of Division into Communities . . . . .	270
6.3.1	Generalisations of the Concept of Modularity . . . . .	274
6.3.2	Organised Modularity . . . . .	275
6.3.3	Scaled Modularity . . . . .	276
6.3.4	Community Score . . . . .	276
6.4	Community Discovery in Undirected Graphs . . . . .	277
6.4.1	Clique Based Communities . . . . .	277
6.4.2	Optimisation of Modularity . . . . .	277
6.4.3	Greedy Algorithms Computing Modularity . . . . .	278
6.4.4	Hierarchical Clustering . . . . .	280
6.4.5	Spectral Methods . . . . .	282
6.4.6	Bayesian Methods . . . . .	295
6.5	Discovering Communities in Oriented Graphs . . . . .	296
6.5.1	Newman Spectral Method . . . . .	297
6.5.2	Zhou/Huang/ Schölkopf Method . . . . .	297
6.5.3	Other Random Walk Approaches . . . . .	298
6.6	Communities in Large Empirical Graphs . . . . .	299
6.7	Heuristics and Metaheuristics Applied for Optimization of Modularity . . . . .	300
6.8	Overlapping Communities . . . . .	304
6.8.1	Detecting Overlapping Communities via Edge Clustering . . . . .	306
6.9	Quality of Communities Detection Algorithms . . . . .	307
6.10	Communities in Multi-Layered Graphs . . . . .	311
6.11	Software and Data . . . . .	313
<b>7</b>	<b>Data Sets . . . . .</b>	<b>315</b>
	<b>Appendix A: Justification of the FCM Algorithm . . . . .</b>	<b>319</b>
	<b>Appendix B: Matrix Calculus . . . . .</b>	<b>321</b>
	<b>Appendix C: Personalized PageRank Vector . . . . .</b>	<b>339</b>
	<b>Appendix D: Axiomatic Systems for Clustering . . . . .</b>	<b>347</b>
	<b>Appendix E: Justification for the <math>k</math>-means++ Algorithm . . . . .</b>	<b>381</b>
	<b>References . . . . .</b>	<b>391</b>
	<b>Index . . . . .</b>	<b>417</b>

Modern Algorithms of Cluster Analysis

Wierzchoń, S.; Kłopotek, M.

2018, XX, 421 p. 51 illus., Hardcover

ISBN: 978-3-319-69307-1