
Contents

List of Figures

List of Tables

Preface

Acknowledgments

About the Authors

List of Symbols

Foreword

1	Introduction	1
1.1	History	8
1.2	Bioinformatics and Drug Discovery	13
1.3	Statistical Learning Theory and Exploratory Data Analysis	16
1.4	Clustering Algorithms	17
1.5	Computational Complexity	18
1.5.1	Data Structures	21
1.5.2	Parallel Algorithms	23
1.6	Glossary	24
1.7	Exercises	24
2	Data	27
2.1	Types	27
2.1.1	Binary Data	28
2.1.2	Count Data	31
2.1.3	Continuous Data	32
2.1.4	Categorical Data	32
2.1.5	Mixed Type Data	32
2.2	Normalization and Scaling	33
2.3	Transformations	33
2.4	Formats	36
2.5	Data Matrices	36
2.6	Measures of Similarity	37

2.6.1	Binary Data Measures	38
2.6.1.1	The Tanimoto or Jaccard Measure	39
2.6.1.2	The Baroni-Urbani/Buser Coefficient	41
2.6.1.3	The Simple Matching Coefficient	42
2.6.1.4	The Binary Euclidean Measure	46
2.6.1.5	The Binary Cosine or Ochai Measure	46
2.6.1.6	The Hamann Measure	47
2.6.1.7	Other Binary Measures	47
2.6.2	Count Data Measures	49
2.6.2.1	The Tanimoto Count Measure	49
2.6.2.2	The Cosine Count Measure	50
2.6.3	Continuous Data Measures	50
2.6.3.1	Continuous and Weighted Forms of Euclidean Distance	50
2.6.3.2	Manhattan Distance	51
2.6.3.3	L_∞ or Supremum Norm	51
2.6.3.4	Cosine	51
2.6.3.5	Pearson Correlation Coefficient	52
2.6.3.6	Mahalanobis Distance	52
2.6.4	Mixed Type Data	53
2.7	Proximity Matrices	53
2.8	Symmetric Matrices	53
2.8.1	Asymmetric Matrices	55
2.8.2	Hadamard Product of Two Matrices	56
2.8.3	Ultrametricity	57
2.8.4	Positive Semidefinite Matrices	57
2.9	Dimensionality, Components, Discriminants	57
2.9.1	Principal Component Analysis (PCA)	58
2.9.1.1	Covariance Matrix	58
2.9.1.2	Singular Value Decomposition	59
2.9.2	Non-Negative Matrix Factorization	59
2.9.3	Multidimensional Scaling	59
2.9.4	Discriminants	60
2.9.4.1	Fisher's Linear Discriminant	60
2.10	Graph Theory	60
2.11	Glossary	65
2.12	Exercises	66
3	Clustering Forms	69
3.1	Partitional	71
3.2	Hierarchical	73
3.2.1	Dendrograms and Heatmaps	73
3.3	Mixture Models	80
3.4	Sampling	81
3.5	Overlapping	82

3.6	Fuzzy	83
3.7	Self-Organizing	84
3.8	Hybrids	85
3.9	Glossary	85
3.10	Exercises	86
4	Partitional Algorithms	91
4.1	<i>K</i> -Means	91
4.1.1	<i>K</i> -Medoid	93
4.1.2	<i>K</i> -Modes	94
4.1.3	Online <i>K</i> -Means	94
4.2	Jarvis-Patrick	95
4.3	Spectral Clustering	98
4.4	Self-Organizing Maps	99
4.5	Glossary	101
4.6	Exercises	101
5	Cluster Sampling Algorithms	103
5.1	Leader Algorithms	103
5.2	Taylor-Butina Algorithm	107
5.3	Glossary	109
5.4	Exercises	112
6	Hierarchical Algorithms	113
6.1	Agglomerative	113
6.1.1	Reciprocal Nearest Neighbors Class of Algorithms	114
6.1.1.1	Complete Link	114
6.1.1.2	Group Average	114
6.1.1.3	Wards	115
6.1.2	Others	116
6.2	Divisive	119
6.3	Glossary	119
6.4	Exercises	120
7	Hybrid Algorithms	121
7.1	Self-Organizing Tree Algorithm	121
7.2	Divisive Hierarchical <i>K</i> -Means	123
7.3	Exclusion Region Hierarchies	124
7.4	Biclustering	126
7.5	Glossary	126
7.6	Exercises	127

8	Asymmetry	129
8.1	Measures	132
8.1.1	Tversky	132
8.1.2	Edit Distance	133
8.2	Algorithms	134
8.2.1	Strongly Connected Components Agglomerative Hierarchical Algorithm	134
8.2.2	Modified Taylor-Butina Algorithm	134
8.3	Glossary	135
8.4	Exercises	136
9	Ambiguity	137
9.1	Discrete Valued Data Types	138
9.2	Precision	138
9.3	Ties in Proximity	139
9.4	Measure Probability and Distributions	140
9.4.1	Number Theory in Measures	140
9.5	Algorithm Decision Ambiguity	145
9.6	Overlapping Clustering Algorithms Based on Ambiguity	148
9.6.1	Modified Taylor-Butina Algorithm	148
9.6.2	Jarvis-Patrick	150
9.6.3	K-Means	151
9.6.4	Hierarchical or Pyramidal Algorithms	152
9.7	Glossary	152
9.8	Exercises	152
10	Validation	155
10.1	Validation Measures	156
10.1.1	Internal Cluster Measures	156
10.1.2	External Cluster Measures	157
10.1.3	Uses of External Cluster Measures	158
10.1.3.1	Tendency to Cluster	159
10.1.3.2	Application Dependence and Domain Knowledge	159
10.1.3.3	Monte Carlo Analysis	160
10.2	Visualization	161
10.3	Example	162
10.4	Glossary	166
10.5	Exercises	166
11	Large Scale and Parallel Algorithms	169
11.1	Leader and Leader-Follower Algorithms	170
11.2	Taylor-Butina	171
11.3	K-Means and Variants	172
11.3.1	Divisive K-Means	173

11.4 Examples	174
11.4.1 Compound Diversity	174
11.4.2 Compound Acquisition	176
11.5 Glossary	179
11.6 Exercises	179
12 Appendices	181
12.1 Primer on Matrix Algebra	181
12.2 Primer on Probability Theory	188
12.3 Primer on Number Theory	189
Bibliography	193
Index	205