

Contents

| | |
|--|-------------|
| List of Figures | xiii |
| List of Tables | xv |
| List of Algorithms | xvii |
| Preface | xix |
| 1 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 | 7 |
| 1.2.5 Validity Indices | 8 |
| 1.3 Clustering Processes | 8 |
| 1.4 Dealing with Missing Values | 10 |
| 1.5 Resources for Clustering | 12 |
| 1.5.1 Surveys and Reviews on Clustering | 12 |
| 1.5.2 Books on Clustering | 12 |
| 1.5.3 Journals | 13 |
| 1.5.4 Conference Proceedings | 15 |
| 1.5.5 Data Sets | 17 |
| 1.6 Summary | 17 |
| 2 Data Types | 19 |
| 2.1 Categorical Data | 19 |
| 2.2 Binary Data | 21 |
| 2.3 Transaction Data | 23 |
| 2.4 Symbolic Data | 23 |
| 2.5 Time Series | 24 |
| 2.6 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 | 28 |
| 3.1.4 | Nominal to Ordinal | 28 |
| 3.1.5 | Ordinal to Interval | 29 |
| 3.1.6 | Other Conversions | 29 |
| 3.2 | Categorization of Numerical Data | 30 |
| 3.2.1 | Direct Categorization | 30 |
| 3.2.2 | Cluster-based Categorization | 31 |
| 3.2.3 | Automatic Categorization | 37 |
| 3.3 | Summary | 41 |
| 4 | Data Standardization and Transformation | 43 |
| 4.1 | Data Standardization | 43 |
| 4.2 | Data Transformation | 46 |
| 4.2.1 | Principal Component Analysis | 46 |
| 4.2.2 | SVD | 48 |
| 4.2.3 | The Karhunen-Loève Transformation | 49 |
| 4.3 | Summary | 51 |
| 5 | Data Visualization | 53 |
| 5.1 | Sammon's Mapping | 53 |
| 5.2 | MDS | 54 |
| 5.3 | SOM | 56 |
| 5.4 | Class-preserving Projections | 59 |
| 5.5 | Parallel Coordinates | 60 |
| 5.6 | Tree Maps | 61 |
| 5.7 | Categorical Data Visualization | 62 |
| 5.8 | Other Visualization Techniques | 65 |
| 5.9 | Summary | 65 |
| 6 | Similarity and Dissimilarity Measures | 67 |
| 6.1 | Preliminaries | 67 |
| 6.1.1 | Proximity Matrix | 68 |
| 6.1.2 | Proximity Graph | 69 |
| 6.1.3 | Scatter Matrix | 69 |
| 6.1.4 | Covariance Matrix | 70 |
| 6.2 | Measures for Numerical Data | 71 |
| 6.2.1 | Euclidean Distance | 71 |
| 6.2.2 | Manhattan Distance | 71 |
| 6.2.3 | Maximum Distance | 72 |
| 6.2.4 | Minkowski Distance | 72 |
| 6.2.5 | Mahalanobis Distance | 72 |

| | | | |
|---------------------------------|---|--|------------|
| | 6.2.6 | Average Distance | 73 |
| | 6.2.7 | Other Distances | 74 |
| 6.3 | | Measures for Categorical Data | 74 |
| | 6.3.1 | The Simple Matching Distance | 76 |
| | 6.3.2 | Other Matching Coefficients | 76 |
| 6.4 | | Measures for Binary Data | 77 |
| 6.5 | | Measures for Mixed-type Data | 79 |
| | 6.5.1 | A General Similarity Coefficient | 79 |
| | 6.5.2 | A General Distance Coefficient | 80 |
| | 6.5.3 | A Generalized Minkowski Distance | 81 |
| 6.6 | | Measures for Time Series Data | 83 |
| | 6.6.1 | The Minkowski Distance | 84 |
| | 6.6.2 | Time Series Preprocessing | 85 |
| | 6.6.3 | Dynamic Time Warping | 87 |
| | 6.6.4 | Measures Based on Longest Common Subsequences | 88 |
| | 6.6.5 | Measures Based on Probabilistic Models | 90 |
| | 6.6.6 | Measures Based on Landmark Models | 91 |
| | 6.6.7 | Evaluation | 92 |
| 6.7 | | Other Measures | 92 |
| | 6.7.1 | The Cosine Similarity Measure | 93 |
| | 6.7.2 | A Link-based Similarity Measure | 93 |
| | 6.7.3 | Support | 94 |
| 6.8 | | Similarity and Dissimilarity Measures between Clusters | 94 |
| | 6.8.1 | The Mean-based Distance | 94 |
| | 6.8.2 | The Nearest Neighbor Distance | 95 |
| | 6.8.3 | The Farthest Neighbor Distance | 95 |
| | 6.8.4 | The Average Neighbor Distance | 96 |
| | 6.8.5 | Lance-Williams Formula | 96 |
| 6.9 | | Similarity and Dissimilarity between Variables | 98 |
| | 6.9.1 | Pearson's Correlation Coefficients | 98 |
| | 6.9.2 | Measures Based on the Chi-square Statistic | 101 |
| | 6.9.3 | Measures Based on Optimal Class Prediction | 103 |
| | 6.9.4 | Group-based Distance | 105 |
| 6.10 | | Summary | 106 |
| II Clustering Algorithms | | | 107 |
| 7 | Hierarchical Clustering Techniques | | 109 |
| | 7.1 | Representations of Hierarchical Clusterings | 109 |
| | | 7.1.1 n -tree | 110 |
| | | 7.1.2 Dendrogram | 110 |
| | | 7.1.3 Banner | 112 |
| | | 7.1.4 Pointer Representation | 112 |
| | | 7.1.5 Packed Representation | 114 |
| | | 7.1.6 Icicle Plot | 115 |
| | | 7.1.7 Other Representations | 115 |

| | | |
|----------|---|------------|
| 7.2 | Agglomerative Hierarchical Methods | 116 |
| 7.2.1 | The Single-link Method | 118 |
| 7.2.2 | The Complete Link Method | 120 |
| 7.2.3 | The Group Average Method | 122 |
| 7.2.4 | The Weighted Group Average Method | 125 |
| 7.2.5 | The Centroid Method | 126 |
| 7.2.6 | The Median Method | 130 |
| 7.2.7 | Ward's Method | 132 |
| 7.2.8 | Other Agglomerative Methods | 137 |
| 7.3 | Divisive Hierarchical Methods | 137 |
| 7.4 | Several Hierarchical Algorithms | 138 |
| 7.4.1 | SLINK | 138 |
| 7.4.2 | Single-link Algorithms Based on Minimum Spanning Trees | 140 |
| 7.4.3 | CLINK | 141 |
| 7.4.4 | BIRCH | 144 |
| 7.4.5 | CURE | 144 |
| 7.4.6 | DIANA | 145 |
| 7.4.7 | DISMEA | 147 |
| 7.4.8 | Edwards and Cavalli-Sforza Method | 147 |
| 7.5 | Summary | 149 |
| 8 | Fuzzy Clustering Algorithms | 151 |
| 8.1 | Fuzzy Sets | 151 |
| 8.2 | Fuzzy Relations | 153 |
| 8.3 | Fuzzy k -means | 154 |
| 8.4 | Fuzzy k -modes | 156 |
| 8.5 | The c -means Method | 158 |
| 8.6 | Summary | 159 |
| 9 | Center-based Clustering Algorithms | 161 |
| 9.1 | The k -means Algorithm | 161 |
| 9.2 | Variations of the k -means Algorithm | 164 |
| 9.2.1 | The Continuous k -means Algorithm | 165 |
| 9.2.2 | The Compare-means Algorithm | 165 |
| 9.2.3 | The Sort-means Algorithm | 166 |
| 9.2.4 | Acceleration of the k -means Algorithm with the kd -tree | 167 |
| 9.2.5 | Other Acceleration Methods | 168 |
| 9.3 | The Trimmed k -means Algorithm | 169 |
| 9.4 | The x -means Algorithm | 170 |
| 9.5 | The k -harmonic Means Algorithm | 171 |
| 9.6 | The Mean Shift Algorithm | 173 |
| 9.7 | MEC | 175 |
| 9.8 | The k -modes Algorithm (Huang) | 176 |
| 9.8.1 | Initial Modes Selection | 178 |
| 9.9 | The k -modes Algorithm (Chaturvedi et al.) | 178 |

| | | |
|-----------|--|------------|
| 9.10 | The k -probabilities Algorithm | 179 |
| 9.11 | The k -prototypes Algorithm | 181 |
| 9.12 | Summary | 182 |
| 10 | Search-based Clustering Algorithms | 183 |
| 10.1 | Genetic Algorithms | 184 |
| 10.2 | The Tabu Search Method | 185 |
| 10.3 | Variable Neighborhood Search for Clustering | 186 |
| 10.4 | Al-Sultan's Method | 187 |
| 10.5 | Tabu Search-based Categorical Clustering Algorithm | 189 |
| 10.6 | J -means | 190 |
| 10.7 | GKA | 192 |
| 10.8 | The Global k -means Algorithm | 195 |
| 10.9 | The Genetic k -modes Algorithm | 195 |
| | 10.9.1 The Selection Operator | 196 |
| | 10.9.2 The Mutation Operator | 196 |
| | 10.9.3 The k -modes Operator | 197 |
| 10.10 | The Genetic Fuzzy k -modes Algorithm | 197 |
| | 10.10.1 String Representation | 198 |
| | 10.10.2 Initialization Process | 198 |
| | 10.10.3 Selection Process | 199 |
| | 10.10.4 Crossover Process | 199 |
| | 10.10.5 Mutation Process | 200 |
| | 10.10.6 Termination Criterion | 200 |
| 10.11 | SARS | 200 |
| 10.12 | Summary | 202 |
| 11 | Graph-based Clustering Algorithms | 203 |
| 11.1 | Chameleon | 203 |
| 11.2 | CACTUS | 204 |
| 11.3 | A Dynamic System-based Approach | 205 |
| 11.4 | ROCK | 207 |
| 11.5 | Summary | 208 |
| 12 | Grid-based Clustering Algorithms | 209 |
| 12.1 | STING | 209 |
| 12.2 | OptiGrid | 210 |
| 12.3 | GRIDCLUS | 212 |
| 12.4 | GDILC | 214 |
| 12.5 | WaveCluster | 216 |
| 12.6 | Summary | 217 |
| 13 | Density-based Clustering Algorithms | 219 |
| 13.1 | DBSCAN | 219 |
| 13.2 | BRIDGE | 221 |
| 13.3 | DBCLASD | 222 |

| | | |
|-----------|---|------------|
| 13.4 | DENCLUE | 223 |
| 13.5 | CUBN | 225 |
| 13.6 | Summary | 226 |
| 14 | Model-based Clustering Algorithms | 227 |
| 14.1 | Introduction | 227 |
| 14.2 | Gaussian Clustering Models | 230 |
| 14.3 | Model-based Agglomerative Hierarchical Clustering | 232 |
| 14.4 | The EM Algorithm | 235 |
| 14.5 | Model-based Clustering | 237 |
| 14.6 | COOLCAT | 240 |
| 14.7 | STUCCO | 241 |
| 14.8 | Summary | 242 |
| 15 | Subspace Clustering | 243 |
| 15.1 | CLIQUE | 244 |
| 15.2 | PROCLUS | 246 |
| 15.3 | ORCLUS | 249 |
| 15.4 | ENCLUS | 253 |
| 15.5 | FINDIT | 255 |
| 15.6 | MAFIA | 258 |
| 15.7 | DOC | 259 |
| 15.8 | CLTree | 261 |
| 15.9 | PART | 262 |
| 15.10 | SUBCAD | 264 |
| 15.11 | Fuzzy Subspace Clustering | 270 |
| 15.12 | Mean Shift for Subspace Clustering | 275 |
| 15.13 | Summary | 285 |
| 16 | Miscellaneous Algorithms | 287 |
| 16.1 | Time Series Clustering Algorithms | 287 |
| 16.2 | Streaming Algorithms | 289 |
| 16.2.1 | LSEARCH | 290 |
| 16.2.2 | Other Streaming Algorithms | 293 |
| 16.3 | Transaction Data Clustering Algorithms | 293 |
| 16.3.1 | LargeItem | 294 |
| 16.3.2 | CLOPE | 295 |
| 16.3.3 | OAK | 296 |
| 16.4 | Summary | 297 |
| 17 | Evaluation of Clustering Algorithms | 299 |
| 17.1 | Introduction | 299 |
| 17.1.1 | Hypothesis Testing | 301 |
| 17.1.2 | External Criteria | 302 |
| 17.1.3 | Internal Criteria | 303 |
| 17.1.4 | Relative Criteria | 304 |

| | | |
|---|---|------------|
| 17.2 | Evaluation of Partitional Clustering | 305 |
| 17.2.1 | Modified Hubert's Γ Statistic | 305 |
| 17.2.2 | The Davies-Bouldin Index | 305 |
| 17.2.3 | Dunn's Index | 307 |
| 17.2.4 | The SD Validity Index | 307 |
| 17.2.5 | The S_Dbw Validity Index | 308 |
| 17.2.6 | The RMSSTD Index | 309 |
| 17.2.7 | The RS Index | 310 |
| 17.2.8 | The Calinski-Harabasz Index | 310 |
| 17.2.9 | Rand's Index | 311 |
| 17.2.10 | Average of Compactness | 312 |
| 17.2.11 | Distances between Partitions | 312 |
| 17.3 | Evaluation of Hierarchical Clustering | 314 |
| 17.3.1 | Testing Absence of Structure | 314 |
| 17.3.2 | Testing Hierarchical Structures | 315 |
| 17.4 | Validity Indices for Fuzzy Clustering | 315 |
| 17.4.1 | The Partition Coefficient Index | 315 |
| 17.4.2 | The Partition Entropy Index | 316 |
| 17.4.3 | The Fukuyama-Sugeno Index | 316 |
| 17.4.4 | Validity Based on Fuzzy Similarity | 317 |
| 17.4.5 | A Compact and Separate Fuzzy Validity Criterion | 318 |
| 17.4.6 | A Partition Separation Index | 319 |
| 17.4.7 | An Index Based on the Mini-max Filter Concept and Fuzzy Theory | 319 |
| 17.5 | Summary | 320 |
| III Applications of Clustering | | 321 |
| 18 | Clustering Gene Expression Data | 323 |
| 18.1 | Background | 323 |
| 18.2 | Applications of Gene Expression Data Clustering | 324 |
| 18.3 | Types of Gene Expression Data Clustering | 325 |
| 18.4 | Some Guidelines for Gene Expression Clustering | 325 |
| 18.5 | Similarity Measures for Gene Expression Data | 326 |
| 18.5.1 | Euclidean Distance | 326 |
| 18.5.2 | Pearson's Correlation Coefficient | 326 |
| 18.6 | A Case Study | 328 |
| 18.6.1 | C++ Code | 328 |
| 18.6.2 | Results | 334 |
| 18.7 | Summary | 334 |
| IV MATLAB and C++ for Clustering | | 341 |
| 19 | Data Clustering in MATLAB | 343 |
| 19.1 | Read and Write Data Files | 343 |
| 19.2 | Handle Categorical Data | 347 |

| | | |
|-----------|--|------------|
| 19.3 | M-files, MEX-files, and MAT-files | 349 |
| 19.3.1 | M-files | 349 |
| 19.3.2 | MEX-files | 351 |
| 19.3.3 | MAT-files | 354 |
| 19.4 | Speed up MATLAB | 354 |
| 19.5 | Some Clustering Functions | 355 |
| 19.5.1 | Hierarchical Clustering | 355 |
| 19.5.2 | k -means Clustering | 359 |
| 19.6 | Summary | 362 |
| 20 | Clustering in C/C++ | 363 |
| 20.1 | The STL | 363 |
| 20.1.1 | The <i>vector</i> Class | 363 |
| 20.1.2 | The <i>list</i> Class | 364 |
| 20.2 | C/C++ Program Compilation | 366 |
| 20.3 | Data Structure and Implementation | 367 |
| 20.3.1 | Data Matrices and Centers | 367 |
| 20.3.2 | Clustering Results | 368 |
| 20.3.3 | The Quick Sort Algorithm | 369 |
| 20.4 | Summary | 369 |
| A | Some Clustering Algorithms | 371 |
| B | The kd-tree Data Structure | 375 |
| C | MATLAB Codes | 377 |
| C.1 | The MATLAB Code for Generating Subspace Clusters | 377 |
| C.2 | The MATLAB Code for the k -modes Algorithm | 379 |
| C.3 | The MATLAB Code for the MSSC Algorithm | 381 |
| D | C++ Codes | 385 |
| D.1 | The C++ Code for Converting Categorical Values to Integers | 385 |
| D.2 | The C++ Code for the FSC Algorithm | 388 |
| | Bibliography | 397 |
| | Subject Index | 443 |
| | Author Index | 455 |