

SUBJECT INDEX

- Absorbing state, 188
- Acute lymphoblastic leukemia (ALL),
 identifying, 161–162
- Acute myeloid leukemia (AML),
 identifying, 157–162
- Adaptive fuzzy *c*-shells (AFCS), 91
- Adaptive learning rate strategy, 115
- Adaptive Resonance Theory (ART), 113,
 116–130, 214, 229. *See also* ART
 entries; Fuzzy ART (FA)
 cancer identification using, 157–162
- Adaptive-subspace SOFM, 142. *See also*
 Self-Organizing Feature Maps
 (SOFMs)
- Adenocarcinomas, 56
- Affinity matrix, 251
- Affinity threshold parameter, 83
- Affymetrix GeneChip technology, 50
- Affymetrix oligonucleotide microarray
 technology, versus cDNA
 microarray technology, 51
- Agglomerative clustering, similarity-
 based, 46
- Agglomerative hierarchical clustering
 algorithm, 29, 31, 32–37, 41, 45, 61,
 198, 196
 flowchart for, 33
 parameters for, 34
 sequential patterns in, 186
- Agglomerative (divisive) hierarchical
 clusters, 172
- Agglomerative likelihood tree (ALT), 46
- Akaike’s information criterion (AIC),
 275
- Algorithm plasticity, 117
- Algorithms, for large-scale data sets
 cluster analysis, 215
- ALgorithms Of Pattern EXtraction
 (ALOPEX), 106
- AlignACE program, 100
- AMOEBa algorithm, 45
- Annotations, 108
- Approximate weight of evidence (AWE)
 criterion, 275
- AR(*p*) model, 198
- ARMA models, mixtures of, 198–199

- ART1/2/3 architectures, 117–121. *See also* Adaptive Resonance Theory (ART)
- ART1 flowchart, 119
- ART1 hierarchy, 121, 122
- ART1 macrocircuits, 148
- ART1 modules, 148
- ART1 networks, neural information retrieval system using, 147–148
- ARTMAP system, 120
- ART networks, 127–130, 230, 276
- Arts & Humanities Citation Index⁷ (A&HCI), 9
- ART tree database structure, 120, 149
- Ascending FLVQ, 137. *See also* Fuzzy LVQ (FLVQ)
- Association probability, 97
- Astronomy, use of clustering in, 8
- Asymmetric binary features, 26–27
- ATG codon, 259
- Auto-associative multilayer perceptron, 241
- AutoClass, 80
- Average linkage hierarchical clustering algorithm, 52, 55
- Average partition density, 272, 274
- Backtracking process, 183, 184
- Backward algorithm, 189–191
- Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm, 32, 40–41, 219–220
- general procedures of, 41, 42
- Base pairing rules, 202–203
- Basic Local Alignment Search Tool (BLAST), 204, 205, 207
- Batch LVQ algorithms, 138. *See also* Learning Vector Quantization (LVQ)
- Batch mode learning, 68
- Batch-mode RBF kernel K -means algorithm, 169
- Baum-Welch algorithm, 192–193, 194
- Bayes formula, 74
- Bayesian clustering by dynamics (BCD), 200
- Bayesian Information Criterion (BIC), 79, 275
- Bayesian methods, 77
- Bayesian Ying-Yang harmony learning, 133
- B-cell lymphoma, 56
- bEMADS algorithm, 220
- “Bends” modules, 148
- Between-cluster scatter matrix, 65
- Biased sampling, 217
- Bias term, 131
- Binary data clustering, K -means in, 72–73
- Binary features, 16
- Binary objects, contingency table for, 26
- Binary tree, 31–32
- Binary variables, similarity measures for, 26–27
- Biological and genomic sequence clustering, 206–211
- Biological and genomic sequence comparison, 203–206
- Biological sequence clustering, 201–211
- Biological taxonomy, 1–2
- B-lineage acute lymphoblastic leukemia, 161
- Boltzmann distribution, 95
- Border points, 221
- Bounded support vectors (BSVs), 172
- Box-counting plot, 227
- BUBBLE algorithm, 219–220
- BUBBLE-FM algorithm, 219–220
- Caliński and Harabasz index, 269
- Cancer identification, using SOFM and ART, 157–162
- Cancer research, gene profiling-based, 56
- Cancer samples, molecular clustering of, 56
- CAT codon, 259
- Categorical variables, 80
- Category choice function, 122, 127, 129
- Gaussian-defined, 127
- Category match function, 123, 129
- Category proliferation problem, 127
- Cauchy-Schwartz independence measure, 170
- cDNA clones, 47
- cDNA microarray technology, 49
- versus Affymetrix oligonucleotide microarray technology, 51

- Centroid average hierarchical clustering algorithm, 56
- Centroid linkage algorithm, 35
- Centroids, 35–36, 95
- CF** vector, 41
- Chaining effect, 34
- Chameleon algorithm, 44–45
- Cheeseman-Stutz approximation, 197
- Chernoff bounds, 42, 216
- City-block distance, 23, 24, 26, 28, 30
- Class-conditional probability density, 74
- Classification, role in human development, 1
- Classification systems, 2
- Classification tasks, feature extraction in, 175
- Clindex, 253
- Clinical images, segmentation results for, 234
- Clipping procedure, 198
- Clique graph, 83
- Cluster Affinity Search Technique (CAST) algorithm, 83, 84, 103
- Cluster analysis, x, 1–13, 48–49, 110, 180
 - alternative names of, 12
 - books devoted to, 11–12
 - goals of, 3
 - hypotheses suggestion and, 8
 - issues and research trends related to, 280
 - major challenges in, 13
 - practical issues in, 11
 - procedure for, 5–8
 - steps in, 279
 - subjectivity of, 4, 6, 30
- Cluster-approximating polygons, 223
- Cluster centers, 86
- Cluster dominance factor, 254
- Clustering. *See also* Complete linkage clustering; Data clustering; Fuzzy clustering entries; Hard competitive learning clustering; Model-based entries; Neural network-based clustering; Partitional clustering; PROjected CLUstering (PROCLUS); Sequential clustering; Sequential data clustering; Similarity-Based Agglomerative Clustering (SBAC); Soft competitive learning clustering
 - advances in, 282
 - document, 153–157
 - goal of, 2
 - graph theory-based, 81–83
 - HMM-based, 195
 - importance of, ix–x
 - mixture density-based, 73–81
 - nonpredictive, 3
 - relative hierarchical, 46
 - as a subjective process, 279–280
 - top-down method of, 225
 - trials and repetitions related to, 8
- Clustering algorithms, ix, 3–4, 211, 280.
 - See also* Agglomerative hierarchical clustering algorithm; Divisive hierarchical clustering algorithms; Document clustering algorithms; Hierarchical clustering algorithms; HMM clustering algorithm; Kernel-based clustering algorithms; Novel clustering algorithms; Robust entries; TS-based clustering algorithm; Two-phase entries
 - comparative research on, 11
 - computational complexity of, 213–214
 - design/selection of, 6–7
 - fuzzy *c*-shells, 91–92
 - literature on, 9–12
 - review of, 11, 12
 - search techniques–based, 92–99
 - in spatial database systems, 11
 - using a random sampling approach, 216–218
- Clustering applications, 8–9
- Clustering criteria
 - differing, 3
 - in partitional clustering, 64–67
- Clustering features (CFs), 219–220
- Clustering feature tree, 40–41, 219
- Clustering genes, applications of, 52
- Clustering Identification via Connectivity Kernels (CLICK), 81–82, 103, 104
- CLustering In QUest (CLIQUE), 253–254
- Clustering LARge Applications (CLARA) algorithm, 216–217

- Clustering Large Applications based on RANdomized Search (CLARANS) algorithm, 217, 218
- Clustering linkage, 33–37
- Clustering methods, advances in, 40–46
- Clustering problems, describing by means of graphs, 37
- Clustering prototypes, 153
- Clustering results
 - differing, 3
 - justifying, 277–278
- Clustering strategy, designing, 7
- Clustering structures, 263
 - validity paradigm for, 264
- Clustering survey papers, 10–11
- Clustering tendency analysis, 263
- Clustering Using REpresentatives (CURE) algorithm, 32, 41–43, 216
- Cluster labeling, 170–171
- Cluster membership histogram, 107
- Cluster number, estimating, 115–116. *See also* *K* (number of clusters)
- Cluster partition, 167–168
- Cluster proliferation problem, 123
- Cluster prototype matrix, 168
- Cluster prototypes, 87
- Clusters
 - defined, 3–8
 - distance between, 35
 - hyper-rectangular representations of, 124
 - meaningful, 7
- Cluster shapes, 91, 92
- Cluster validation, 11, 13, 57
 - in cluster analysis, 7
- Cluster validation indices, 269–271
- Cluster validity, 263–278
- Code words, 68
- Codons, 258
- Coincident clusters, 89
- Color image segmentation, large-scale
 - data clustering in, 232–234
- Competitive Hebbian rule, 143, 144
- Competitive learning, 113
 - neural network-based clustering and, 162
- Competitive learning network, 112
- Competitive learning scheme, 111
- Competitive neural networks, false alarm rates and intervals of normality for, 153
- Complement coding, 124
- Complete linkage algorithm, 34–35
- Complete linkage clustering, 37
- Component models, 195
- Component similarity, 29
- Compression set (CS) data points, 220
- Computer sciences, use of clustering in, 8
- Concurrent engineering team
 - structuring, hierarchical clustering algorithms in, 60–61
- Condensation-based methods, 219–220
- Conditional probability, 97
- Conscience mechanism, 131–132
- Conservative limit, 122
- Constructive clustering algorithms, 276–277
- Contact printing, 47
- Contingency table, 26
- Continuous dissimilarity measures, 28
- Continuous features, 16
- Continuous observation densities,
 - HMMs with, 193–195
- Continuous variables
 - mixed with discrete variables, 29–30
 - proximity measures for, 22–26
- Cophenetic correlation coefficient (CPC), 267–268
- Core-distance, 222
- Co-regulated gene clusters, 52
- Core points, 221
- Correlation coefficients, 25
- Cosine similarity, 25, 26
- Cost function, 88, 89
 - global, 142–143
- Covariance inflation criterion (CIC), 275
- Covariance matrix, 78
- Cover's theorem, 163
- Criterion functions, 45, 63, 64, 82, 94, 98, 109
 - optimizing, 85
- Cross validation-based information criterion (CVIC), 275
- Cumulative distance, 196–197
- “Curse of dimensionality,” 48, 56, 163, 237

- Cut index, 67
- Cutting hyperplanes, 256
- Data
 - clustered display of, 53
 - feature types and measurement levels of, 15–21
 - importance of, 1
 - missing, 20–21
 - pre-clustered, 43
- Data analysis, gene-expression, 46–57, 157–162
- Data clustering, large-scale, 213–235. *See also* High-dimensional data clustering; Sequential data clustering
- Data compression, 220
- DAta-driven thematic Map quality Assessment (DAMA) strategy, 103–104, 105
- Data mining, 11
- Data objects
 - calculating average distances between, 20–21
 - classifying, 1
 - generating a set of, 74
 - proximity measure between, 12
- Data partition, interpreting, 8
- Data points
 - density of, 220–221
 - geodesic distances between, 247
 - link between, 29
- Data point separation, high dimensionality and, 238
- Data samples, labeling, 3
- Data standardization, 23
- Data visualization, 268–269
 - technologies for, 238–239
- Davies-Bouldin index, 269–270
- DBLP database, 60
- d -dimensional objects, 20
- Dead neuron problem, 130–131
 - conscience mechanism and, 132
- Dead neurons, 162
- Delaunay triangulation, 144
- Delaunay triangulation graph (DTG), 45
- De-learning, 132–133
- Dendrogram, 31–32, 52, 61
- Density attractors, 224
- DENSity-based CLUstEring (DENCLUE) algorithm, 215–216, 223–225
- Density-based methods, 220–225
- Density-based Optimal projective Clustering (DOC), 257–258, 260
- Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, 221, 222–224, 233
- Density-connectivity concept, 221
- Density-reachability concept, 221
- Deoxyribonucleic acid (DNA), 202–203. *See also* DNA entries
- Depth-first binary divisive clustering, 197
- Descending FLVQ, 137. *See also* Fuzzy LVQ (FLVQ)
- Deterministic annealing (DA), 92, 96–97
- Dichotomous features, 16
- Differential entropy, 243–244
- Diffuse large B-cell lymphoma, subtypes of, 56
- Diffusion maps, 251
- Dimension reduction, 238
- Discard set (DS) data points, 220
- Discrete component densities, 80
- Discrete features, 16
 - similarity measures for, 26
- Discrete hidden Markov model, 187
- Discrete variables
 - mixed with continuous variables, 29–30
 - with more than two values, 27–29
 - proximity measures for, 26–29
- Discrete wavelet transform, 226
- Disease diagnosis, cluster analysis for, 2
- Dissimilarity function, 21
- Dissimilarity matrix, 37
- Dissimilarity measure, 29
- Distance-based methods, 211
- Distance definition, modifying, 131–132
- Distance function, 34, 25, 112, 173
- Distance measure, 91
- Distance-sensitive RPCL, 133. *See also* Rival penalized competitive learning (RPCL)
- Distortion function, 69

- Distribution Based Clustering of LArge Spatial Databases (DBCLASD) algorithm, 222–223, 232
- Divide-and-conquer algorithm, 227–228, 229, 230–232
- DIVISive ANALYSIS (DIANA) algorithm, 38, 39
- Divisive hierarchical clustering algorithms, 31, 32, 37–40, 46, 61
- DNA libraries, 47. *See also* Deoxyribonucleic acid (DNA)
- DNA microarray technologies, 47
- DNA sequence alignment, using the Needleman-Wunsch algorithm, 182–183
- DNA sequences, clustering applications for, 202
- Document clustering algorithms, 153–157
experimental evaluations of, 11
XML, 59–60
- Document map, 153
- Drug sensitivity, gene expression patterns and, 56
- Dunn index, 270
- Dynamic NG algorithm, 143
- Dynamic programming algorithms, 185, 191, 203–204
- Dynamic programming-based sequence alignment algorithm, 181–182
- Dynamic programming matrix, 183
- Dynamic time warping (DTW) technique, 196
- Earth sciences, use of clustering in, 8
- Economics, use of clustering in, 9
- Edge aging scheme, 144
- Edit distance, 181
- Edit operations, 181
substitution, insertion, and deletion in, 210
- Effective measures, selecting, 30
- Efficient Projective Clustering by Histograms (EPCH), 257
- Eigenfunctions, 251
- Eigenvalue decomposition, 276
- Eigenvalues, 78, 165, 240–241
- Eigenvectors, 165, 240–241, 251–252
- Elastic maps, 249–250, 258
- Elastic nets, 249, 250
- Ellipsoid ART (EA), 128. *See also* Adaptive resonance theory (ART)
partition of leukemia samples with, 161–162
- Ellipsoid ARTMAP (EAM), 120
- EM Algorithm for Data Summaries (EMADS), 220. *See also* Expectation-maximization (EM) algorithm
- Emission distribution, 193
- Emission probability, 190
estimation of, 192
- EMMIX algorithm, 80
- Energy function, 249–250
- Engineering, use of clustering in, 8
- Enhanced LBG (ELBG) algorithm, 69, 70, 95. *See also* Linde-Buzo-Gray (LBG) algorithm
- ENTropy-based CLUStering (ENCLUS), 255
- Epanechnikov kernel function, 218
- Eps* parameter, 222
- Error criterion function, 240
- Euclidean distance, 22, 24, 26, 28
- Euclidean distance matrix, 247
- Evaluation standards, 7
- Event-related potentials (ERPs), 175
- Evolutionary algorithms (EAs), 92
- Evolutionary-learning SOFM, 142. *See also* Self-Organizing Feature Maps (SOFMs)
- Evolutionary programming (EP), 93
- Evolution strategies (ESs), 93
- Expectation-maximization (EM) algorithm, 77–78, 128, 192, 199, 274
a posteriori-based, 201
in parameter estimation, 198–199
- Exploratory data analysis, 2
- Expressed sequence tag (EST), 47
- Extended cluster feature (ECF) vector, 256
- Extended FLVQ family (EFLVQ-F) cost function, 138. *See also* Learning Vector Quantization (LVQ)
- Extended Xie-Beni index, 273, 274
- eXtensible Markup Language (XML) document clustering, 59–60
- External indices, 7, 13
- External testing criteria, 263–267

- Face recognition, 259–260
- FASTA algorithm, 204–205, 206
- Fast algorithms, comparing, 11
- Fast-commit slow-recode method, 123
- FASTDOC, 258. *See also* Density-based
Optimal projective Clustering
(DOC)
- Fast learning rule, 125
- FastMap algorithm, 220
- Feature dependent parameters, 225
- Feature independent parameter, 225
- Feature selection/extraction, 185, 211,
280
in cluster analysis, 5–6
in regression problems, 175
- Feature types, of data objects, 15–21
- Feedback pathway, 6, 8
- Finite mixture density based clustering
algorithms, 78–81
- Finite mixture modeling, 133
- Finite order ARMA models, 198. *See also* ARMA models
- Fisher Discriminant Analysis (FDA), 109
- Fitness function, 93, 95, 99
- Floyd's algorithm, 247
- fMRI time series clustering, 186
- Forward-backward algorithm, 189–191
- Four-state hidden Markov model, 188
- Fowlkes and Mallows index, 266
- Fractal clustering (FC) algorithm, 215,
216, 226–227
- Fractal dimension, 226–227
- Frequency Sensitive Competitive
Learning (FSCL), 131
fuzzy extension of, 132
- Frequent-Pattern-based Clustering
(FPC), 258
- Fukuyama-Sugeno index, 274
- Fully Automatic Clustering System
(FACS), 69, 95
- Fully self-Organizing SART (FOSART)
network, 128. *See also* Adaptive
Resonance Theory (ART)
- Fuzzification parameter, 272, 274
- Fuzzy algorithms for LVQ (FALVQ),
138. *See also* Learning Vector
Quantization (LVQ)
application in MR images, 150
- Fuzzy AND operator, 122
- Fuzzy ART (FA), 117, 121–127. *See also* Adaptive Resonance Theory
(ART)
in gene expression data analysis, 157,
160
properties of, 126–127
weight vector of clusters generated by,
160
- Fuzzy ARTMAP, 120, 121
- Fuzzy clustering, 5, 11, 83–92, 93, 110,
113
advances in, 92
robustness of, 89–90
- Fuzzy clustering partition, 12
- Fuzzy clustering validity, 272
- Fuzzy *c*-means (FCM) algorithm, 83–91,
106, 137
application to yeast data, 103
problems associated with, 86
steps of, 85–86
- Fuzzy *c*-quadratic shells (FCQS), 92
- Fuzzy *c*-rectangular shells (FCRS), 92
- Fuzzy *c*-rings (FCR), 92
- Fuzzy *c*-shells clustering algorithms,
91–92
- Fuzzy *c*-spherical shells (FCSS)
algorithm, 92
- Fuzzy hypervolume, 272, 274
- Fuzzy Kohonen clustering network
(FKCN), 136
- Fuzzy logic, 83
- Fuzzy LVQ (FLVQ), 134, 136–138. *See also* Learning Vector Quantization
(LVQ)
- Fuzzy medians, 87
- Fuzzy membership, 143
- Fuzzy membership coefficient, 137
- Fuzzy membership function, 102–103,
173
- Fuzzy min-max clustering neural
networks, 124
- Fuzzy OR operator, 125
- Fuzzy partition matrix, 84
- Fuzzy validation indices, 271–274
- GA clustering algorithm, tabu search in,
99. *See also* Genetic algorithms
(GAs)
- Γ statistics, 267

- GA-NMM algorithm, 109. *See also*
Normal Mixture Models-based
(NMM) algorithm
- Gapped BLAST, 204, 205, 207
- Gaussian ART, 127–128. *See also*
Adaptive Resonance Theory
(ART)
- Gaussian ARTMAP (GAM), 120
- Gaussian-defined category choice and
match functions, 127
- Gaussian distributions, 80
- Gaussian influence function, 224
- Gaussian kernel, 172, 251, 276
- Gaussian mixture density decomposition
(GMDD), 79–80
- Gaussian radial basis function (RBF)
kernels, 164–165
- gEMADS algorithm, 220
- GenBank database, 180
- GeneChip, 47
- Gene clustering, 49–52
- GENECLUSTER software, 157
- Gene expression data analysis, 46–57,
157–162
clustering algorithms for, 11
partitioned clustering in, 99–103
- Gene expression levels, measuring, 49
- Gene expression profiling, 56
- Gene expression time series data
clustering, 201
- Gene functions/interactions,
investigating, 46–47
- Gene profiles, as fingerprints for
diseases, 52
- General agglomerative hierarchical
clustering, 32–33
- Generalized Lloyd algorithm, 68
- Generalized LVQ (GLVQ), 134, 136. *See
also* Learning Vector Quantization
(LVQ)
- Generative Topographic Mapping
(GTM) algorithm, 252–253
- Genes, encoding proteins from, 203
- Genetic algorithms (GAs), 93–95. *See
also* GA clustering algorithm
- Genetically guided algorithm (GGA),
93–94
- Genetic *K*-means algorithm (GKA),
70
- Genetic programming (GP), 93
- Genome sequencing projects, 46
advances in, 201–202
- Genomic and biological sequence
clustering, 201–211
- Genomic projects, sequential data from,
180
- Genomic research, 47
- Geodesic distances, calculating, 247
- Geometric methods, 37
- Gibbs distribution, 97
- Global cost function optimization,
142–143
- Global feature selection, 185–186
- Global *K*-means algorithm, 70–71
- Globally optimal clustering, 224
- Global normality profiles, 151
- GLVQ-F algorithm, 136. *See also*
Generalized LVQ (GLVQ)
- Goodall similarity measure, 46
- Goodman-Kruskal γ statistic, 268
- Gradient descent algorithm, 246
- Graph methods, 37
- Graph-partitioning algorithm, 44
- Graph theory, 37
- Graph theory-based clustering
algorithms, 81–83, 103, 208
- Greedy algorithm, 186
- Greedy search process, 257
- Greedy strategy, 255
- Grid-based methods, 225–227
- Group average linkage algorithm, 35
- Group technology
neural information retrieval system
for, 146–149
system architecture for, 147
- Growing Neural Gas (GNG) algorithm,
144–146
- Grow When Required (GWR) network,
146
- Haar wavelet decomposition, 186
- Hamiltonian cycle, 229
- Hamming distance, 27
- Handwritten digit patterns, prototype
identification within, 175–176
- Hard clustering, 83, 93
- Hard competitive learning clustering,
113–130

- Hard competitive learning paradigm,
113, 114, 134, 139, 162
adding a conscience to, 131
- Hausdorff dimension, 227
- Heat map, 52
- Hebb's rule, 113
- Hebbian type learning rules, 241
- Hematopoietic differentiation SOFM,
157, 159. *See also* Self-Organizing
Feature Maps (SOFMs)
- Hessian eigenmaps, 248
- Heterogeneous behaviors, modeling, 201
- Heuristic algorithms, 204
- Heuristic alignment algorithms, 185
- Heuristic approaches/methods, 38, 276
- Hidden Markov model (HMM),
186–189, 206. *See also* HMM entries
with continuous observation densities,
193–195
- Hierarchical abstraction tree, 148
- Hierarchical approach with Automatic
Relevant dimension selection for
Projected clustering (HARP), 257
- Hierarchical ART with joining
(HART-J), 121. *See also* Adaptive
Resonance Theory (ART)
- Hierarchical ART with splitting
(HART-S), 121
- Hierarchical clustering algorithms, 5, 7,
12, 31–62, 213–214
advances in, 11
alternative methods of, 45–46
applications of, 46–61
computational complexity of, 32
criticisms of, 40
GAs in, 95
use of ART in, 120–121
- Hierarchical partitions, 79
- Hierarchical unsupervised fuzzy
clustering (HUFC) algorithm, 92
- High-dimensional data clustering,
237–261
applications of, 258–260
linear projection algorithms, 239–244
nonlinear projection algorithms,
244–253
- High-dimensional data visualization, 13
- Highly connected sub-graphs (HCS),
81
- High scoring segment pairs (HSPs), 204
- Hill-climbing algorithms/methods/
procedures, 68, 69, 92, 110
- hMetis graph-partitioning algorithm, 44
- HMM-based clustering, 195–197. *See also*
Hidden Markov model (HMM)
- HMM clustering algorithm, 196, 208–211
- HMM training, 189
- “Holes” modules, 148
- Homogeneity-based indices, 64–67
- Homology, transitivity of, 207–208
- Horse Colic data set, 16–18
features of, 19
- Human visual system research,
hierarchical clustering based on,
45–46
- Hybridization, 48
- Hybrid niching genetic algorithm
(HNGA), 94
- Hyperellipsoidal clusters, 24
- Hypergraph, 45
- Hyper-quadtrees data structure, 95
- Hyper-sphere ART (HA), 128–129. *See
also* Adaptive Resonance Theory
(ART)
- Hypersphere construction, 170
- Hyperspherical clusters, 22
- Image analysis, 48
- Incremental clustering, 229
- Incremental kernel- K -means algorithm,
169–170
- Independent Component Analysis
(ICA), 242–243
- Independent features, 185–186
- Indirect sequence clustering, 185–186
- Individual clusters, 7
- Influence functions, 223–224
- Information retrieval systems, 146–147.
See also neural information retrieval
system
- Initial probability, estimation of, 192
- Initial state distribution, 187
- Inkjet printing technology, 47
- Input patterns, high-dimensional, 138
- Input space density approximation, 141
- In-situ oligonucleotide arrays, 47
- Instar learning rule, 113, 138
- Integrated squared error, 170

- Interactive Projected CLUStering (IPCLUS), 257
- Inter-cluster distance, calculating, 36
- Inter-cluster distance measures, 37
- Internal indices, 7, 13
- Internal testing criteria, 263–264, 267–268
- Interval measurement scale, 16
- Intrinsic dimensionality, 238
- Invariant similarity measures, 27
- ISometric feature MAPPING (ISOMAP) algorithm, 247–248, 260
- Iterative optimization procedure, 68, 110, 169
- Iterative Self-Organizing Data (ISODATA) analysis technique, 71, 72, 277
- ITRI database, 176
- Jaccard coefficient, 27, 28, 266–267
- Joint density function, 75
- Journals, clustering papers in, 10
- Junk, 258
- K (number of clusters), 71–72. *See also* Cluster number estimation of, 268
- Kaplan-Meier curves, 56, 57
- Karhunen-Loève transformation, 239
- k -dimensional subspace, 253
- k -d tree, 87
- kd -tree data structure, 214
- Kendall's τ statistic, 268
- Kernel-based clustering algorithms, 13, 163–178
 applications of, 175–176
 properties of, 178
- Kernel eigenmap methods, 251
- Kernel functions, 164–165, 166, 168, 178
 for density estimation, 217–218
 squared-error-based clustering with, 167–170
- Kernel independent component analysis, 170
- Kernel K -means algorithm, 169
 interpreting in terms of information theoretic quantities, 170
- Kernel matrix, 165, 248–249
- Kernel MaxEnt, 167
- Kernel Principal Component Analysis (PCA) algorithms, 165–167, 176, 244–245
 in feature extraction, 175
 flowchart for, 166
- Kernel Principal Component Regression (KPCR), 175
- Kernel trick, 163–164, 165
- Key Performance Indicators (KPIs), for cellular networks, 150
- K-GA-NMM algorithm, 109. *See also* Normal Mixture Models-based (NMM) algorithm
- K -means, 11
 as a case of expectation-maximization, 78
- K -means algorithm, 6, 7, 12, 64, 67–73, 104, 186, 208
 convergence and initial partition of, 69–71
 disadvantages of, 69–73
 GAs and, 95
 in gene partitioning, 99–100
 implementation of, 69
 online version of, 68
 robustness of, 72
 time and space complexity of, 213
 versus leader-follower algorithm, 116
- K -means performance, improving, 214–215
- K -medoids-based algorithm, 72
- K -medoids-based projected clustering method, 255
- k -nearest-neighbor graph, 44, 45
- k -nearest-neighbor graph-based algorithm, 37
- Kullback-Leibler distance, 200
- L_1 norm, 23, 24
 in fuzzy clustering, 87
- L_2 norm, 22, 24
- L_4 norm, 23, 24
- Lagrange multipliers, 171
- Lagrange optimization, 240
- Laplacian eigenmaps, 248
- Large-scale data clustering, 13, 213–235
 applications of, 229–234
 condensation-based methods, 219–220
 density-based methods, 220–225

- divide-and-conquer algorithm, 227–228
- grid-based methods, 225–227
- incremental, 229
- Large-scale data sets
 - in cluster analysis, 215
 - clustering, 69
 - proliferation of, 235
- Large-scale genomic DNA sequences, aligning, 206
- Latent data structures, displaying, 138
- Laterally Primed Adaptive Resonance Theory (LAPART), 120. *See also* Adaptive Resonance Theory (ART)
- Leader algorithm, 72
- Leader-follower clustering algorithm, 115–116
- Leaky learning model, 131
- Learning neural networks, in network behavior prediction, 150–153
- Learning rate, 114, 115
- Learning rate series, 136
- Learning rules, 113, 136, 137–138
 - SOFM, 138
- Learning Vector Quantization (LVQ), 113, 133–138. *See also* LVQ entries
- Learning without a teacher, 12
- Leukemias, identifying, 157–162
- Leukemia samples, clustering of, 161
- Levenshtein distance, 181
- Life sciences
 - computational methods for exploring, 202
 - use of clustering in, 8
- Likelihood evaluation, 189–191
- Linde-Buzo-Gray (LBG) algorithm, 68. *See also* Enhanced LBG (ELBG) algorithm
- Linear discrimination analysis, 259
- Linear PCA, 175. *See also* Principal Component Analysis (PCA)
- Linear projection algorithms, 239–244
- Linear transformations, 22
- Lin-Kernighan (LK) local optimization algorithm, 230
- Lloyd algorithm, 68
- Local feature selection, 186
- Locality analysis, 255
- Locally linear embedding (LLE), 248–249, 260, 261
- Local sequence alignment, 184
- Log-likelihood equations, 75, 76, 77
- Log-likelihood function, 199
- Long-term memory (LTM), 117
- Lung tumors, cluster analysis of, 57–58
- LVQ1/2/3, 134. *See also* Learning Vector Quantization (LVQ)
- LVQ algorithms, application in MR images, 150
- Machine learning, 163
- Machine Learning Repository, 70
- Macrocircuit modules, hierarchical abstraction tree of, 148–149
- Magnetic Resonance Imaging (MRI) segmentation, 149–150, 152
- Mahalanobis distance, 24, 26, 220
- Mahalanobis radius, 129–130
- Manhattan distance, 23
- Manhattan segmental distance, 255
- Markov Chain Monte Carlo (MCMC)-based method, 46
- Markov chains, mixtures of, 199–200
- Markov matrices, 251
- Match function, Gaussian-defined, 127
- Mathematical programming scheme, 11
- Maximum accumulated error, 145
- Maximum *a posteriori* (MAP) estimation, 80
- Maximum entropy preservation, 167
- Maximum information compression index, 6
- Maximum likelihood (ML) estimation, 75–77, 128
- Maximum-minimum normalization approach, 23
- MCLUST software, 78
- Means, extended definition of, 72–73. *See also* *K*-means entries
- Measurement levels, of data objects, 16–21
- Median linkage algorithm, 36
- Medical sciences, use of clustering in, 8
- Medoids, 72, 255
- Membership equation, 89
- Mercer kernels, 178, 244, 245
- Mercer's theorem, 163, 165

- Merge criterion, 78
- Merging of Adaptive Finite Intervals (MAFIA), 254
- Metric, 21
- Metropolis algorithm, 95
- Microarray data, computational analysis of, 48
- Microarray experiments, steps in, 47–48
- Microarray preparation, 47–48
- Microarray scanning, 48
- Microarray technologies, 47
computational challenges of, 48
- Minimal description length (MDL)
principle, 254
- Minimal distance, 33
- Minimum description length (MDL), 275
- Minimum message length (MML)
criterion, 78, 275
- Minimum spanning tree (MST), 81, 208
- Minimum variance method, 36
- Minimum variance partition, 64
- Minimum weight cut procedure, 81
- Minkowski distance/metric, 23, 24, 26, 87
- Minpts* parameter, 221, 222
- Missing feature values, 20–21
- Mixed variables, proximity measures for, 29–30
- Mixture densities, forms of, 73–74
- Mixture density-based clustering, 73–81, 186
- Mixture model, of HMMs, 195–196
- Mixture-model-based clustering scheme, 78–81
- Mixture probability density, 74–75, 195
- Mixtures
of ARMA models, 198–199
of Markov chains, 199–200
of polynomial models, 200–201
- M*-level divide-and-conquer approach, 228
- Model-based agglomerative hierarchical clustering, 79
- Model-based clustering, 211–212
- Model-based sequence clustering, 186–201
- Model selection criteria, 274–275
- Modes, 73
- Modified squared Euclidean distance, 131, 132
- Molecular clustering, of cancer samples, 56
- Monothetic algorithms, 38–40
- MONothetic Analysis (MONA), 38–40
- Monotonically decreasing learning rate, 139
- Monte Carlo Cross-Validation (MCCV)
method, 196, 274
- Motifs, 53, 100–102
- Mountain function, 86
- Mountain method (MM), 86
- Multidimensional scaling (MDS), 2, 246
- Multilayer perceptrons, ICA realization via, 243. *See also* Independent Component Analysis (ICA)
- Multilayer support vector regression (MLSVR), 175
- MULTIMIX, 81
- Multiple sclerosis, detecting using visual evoked potentials, 104–108
- Multiplesphere support vector clustering (MSVC), 173
handwritten digit prototype identification with, 177
steps in, 174
- Multi-sets, modeling, 133
- Multistage random sampling FCM algorithm, 91
- Multivariate Gaussian densities, 74, 76, 79
- Multivariate Gaussian mixtures, 80
- Nearest neighbor method, 34
- Nearest-neighbor rule, 68
- Needleman-Wunsch algorithm, 182–183
- Neighborhood function, 139–140
- Neighbor object, 29
- Nested searches scheme, 197
- Net activation value, 112
- Neural Gas (NG) algorithm, 142–143
major process of, 143
- Neural information retrieval system, for group technology, 146–149
- Neural network-based clustering, 111–162
algorithms, 113
applications of, 146–162
- Neural networks, 13, 111
- Neuron removal rule, 144

- Neurons, 138, 139
 updating equation for, 139
- Noise-free ICA model, 242. *See also*
 Independent Component Analysis
 (ICA)
- Noise/possibilistic clustering (N/PC1)
 approach, 90
- Noisy data, category proliferation in, 127
- Noisy ICA model, 242. *See also*
 Independent Component Analysis
 (ICA)
- Nominal measurement scale, 16
- Non-coding information theoretic
 criterion (ICOMP), 275
- Nonhierarchical clusters, graph theory
 for, 81
- Non-invariant similarity measures, 27
- Nonlinear PCA, 244–246. *See also*
 Principal Component Analysis
 (PCA)
- Nonlinear PCA neural networks, 245
- Nonlinear projection algorithms, 244–253
- “Nonperiodic” clusters, 102
- Nonpredictive clustering, 3
- Nonsupport vectors, 172–173
- Normal Mixture Models-based (NMM)
 algorithm, 108–109
- Novel clustering algorithms, important
 properties of, 281–282
- NP-complete problems, 229
- Null hypotheses, 264–265
- Numerical taxonomy, 12
- Oligonucleotides, 47
- One-mode matrix, 22
- Online clustering approach, 229
- Online *K*-means algorithm, 113–115
- Online learning, 223
- On-line mode *K*-means, 68
- OptiGrid, 256–257
- Optimal alignment problem, 181
- Order dependent algorithms, 229
- Ordering Points To Identify the
 Clustering Structure (OPTICS)
 algorithm, 222
- Ordinal features, 28
- Ordinal measurement scale, 16
- ORiented projected CLUster
 (ORCLUS), 256
- Outlier resistant strategy, 146
- Outliers, 42, 43, 44
K-means and, 72
- Overall density function, 224
- PAM algorithm, 216, 217. *See also*
 Partitioning Around Medoids
 (PAM)
- Parallel algorithms, 235
 for hierarchical clustering, 46
- Parallelized batch map algorithm, 154
- Parameter estimation, 192–193
- Parameter selection, 110
- Partial contrast model, 131
- Particle Swarm Optimization (PSO), 98
- Partitional clustering algorithms, 12
- Partitional clustering, 4–5, 7, 31, 63–110
 algorithms, 63
 applications of, 99–109
 clustering criteria for, 64–67
 in detection of multiple sclerosis,
 104–108
 in gene expression data analysis,
 99–103
 in quality assessment of thematic
 maps, 103–104
 in vision-guided robot navigation,
 108–109
- Partition analysis, 12
- Partition coefficient (PC), 271–272
- Partition criterion function, 67
- Partition density, 273, 274
- Partition entropy (PE), 271–272
- Partitioning Around Medoids (PAM), 72.
See also PAM algorithm
- Partition Mutual Information (PMI)
 measure, 197
- Patient survival analysis, 56
- Pattern recognition, 163
- P-AutoClass, 80
- pCluster method, 257
- Pearson correlation, 26
- Pearson correlation coefficient, 25, 52
- Periodic clusters, 101
- Photolithography technology, 47
- Plastic NG algorithm, 143
- Point symmetry distance, 24–25, 26
- Polynomial kernels, 164–165, 172
- Polynomial models, mixtures of, 200–201

- Polythetic algorithms, 38
- Position-Specific Iterated BLAST (PSI-BLAST), 204, 205. *See also* Basic Local Alignment Search Tool (BLAST)
- Positivity condition, 21, 22
- Possibilistic c -means (PCM) clustering algorithm, 87–90
- Possibilistic fuzzy c -means (PFCM) model, 89
- Pre-clustered data, 43
- Pre-clustering phase, 87, 219
- Preprocessing technologies, 62
- Principal Component Analysis (PCA), 13, 109, 165–167, 239–241. *See also* Linear PCA; Nonlinear PCA entries
- Principal curves, 252
- Prior probabilities, 76
- Probabilistic clustering, 85
- Probabilistic Hough Alignment Tool (PHAT) algorithm, 205
- Probabilistic mixture model-based clustering, 274
- Probabilistic models, 186
- Probability density function estimators, 128
- Probability distributions, 80
- PROjected CLUstering (PROCLUS), 253, 255–256
- Projection pursuit, 243–244
- Projection vectors, 165
- Protein sequences
 - clustering applications for, 202
 - clustering with a feature-based algorithm, 209
- ProtoNet 4.0, 207
- Prototype identification, within
 - handwritten digit patterns, 175–176
- Prototype learning, 144
- Prototype vectors, 114–115, 116, 134, 162
 - neighborhood ranking of, 142
- Proximity, for d -dimensional objects, 20
- Proximity-based clustering algorithms, 181
- Proximity function, selection of, 30
- Proximity graph strategy, 172
- Proximity matrix, 22, 33
- Proximity measures, 6, 15–30
 - applications of, 26
 - for continuous variables, 22–26
 - defined, 21–22
 - differing, 30
 - for discrete variables, 26–29
 - for mixed variables, 29–30
 - linkage metrics for, 15
- q -order regression equation, 200–201
- Qualitative features, 20
- Quality function deployment (QFD) process, 61
- Quantitative features, 20
- Radial basis function (RBF) kernels, 164–165, 169
- Rand index, 266–267
- Random effects regression mixtures, 201
- Random graph hypothesis, 265
- Random label hypothesis, 265
- Random position hypothesis, 265
- Random projection method, 154
- Random sampling methods, clustering algorithms using, 216–218
- Rapid Analysis of Pre-Indexed Datastructures (RAPID) algorithm, 205
- Ratio measurement scale, 20
- Reachability-distance, 222
- Recurrence formula, 33–34
- Reference map generation by multiple clustering (RMC), 103
- Reflexivity condition, 21
- Regression problems, feature extraction in, 175
- Relational data, 87
- Relative closeness, 45
- Relative hierarchical clustering (RHC), 46
- Relative indices, 7, 13
- Relative interconnectivity, 45
- Relative testing criteria, 263–264, 268–277
- Renyi quadratic entropy, 170
- Resonance procedure, 118
- Result interpretation, in cluster analysis, 8
- Retained set (RS) data points, 220
- Reversible Jump Markov Chain Monte Carlo (RJCMCMC) algorithm, 96

- Rival of the winner, 132
- Rival penalized competitive learning (RPCL), 132–133
- Robot navigation, vision-guided, 108–109
- Robust clustering algorithms, 89–90
- RObust Clustering using linKs (ROCK), 29, 43–44, 59–60
- Robust competitive clustering algorithm (RCA), 277
- Rough set concept, 173–174
- R^* -tree structure, 222, 223
- SA-based clustering, 96. *See also* Simulated Annealing (SA)
- Saccharomyces cerevisiae*, 52, 99–103
gene expression data analysis for, 157
- SA-Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm, 96
- Scalability, importance of, 213
- Scalable clustering framework, 214–215
- Scale-independent criterion functions, 66
- Scale-space theory, 45
- Scale-space-theory-based method, 276
- Scatter matrices, 65
- Science Citation Index Expanded™, 9
- Search techniques, 110
- Search techniques-based clustering algorithms, 92–99
- Seismic fault detection, large-scale data clustering in, 232
- Self-Organizing Feature Maps (SOFMs), 113, 137, 138–142, 206–207, 239
cancer identification using, 157–162
in gene expression space representation, 157
information retrieval system based on, 153–157
steps of, 140–141
- Self-organizing network, 144
- Self-organizing neural networks, 142
- Self-organizing semantic maps, 153–154
- Self-splitting competitive learning network, 277
- Semimetric, 21
- Semi-supervised clustering, 91
- Sequence alignment, 181
with the Needleman-Wunsch algorithm, 182–183
- Sequence Alignment and Modeling (SAM), 208
- Sequence similarity, 181–185
- Sequential clustering, versus conventional clustering, 180
- Sequential data, properties of, 179, 211
- Sequential data clustering, 13, 179–212
alternative models for, 197–201
genomic and biological, 201–211
indirect, 185–186
model-based, 186–201
sequence similarity and, 181
- Sequential pattern, defined, 185
- S-GRACE algorithm, 59, 60
- “Shape” ART1 module, 148. *See also* Adaptive Resonance Theory (ART)
- Short-term memory (STM), 117
- Shrinking-based method, 252
- Sigmoid kernels, 164–165
- Signal processing techniques, 226
- Similarity-Based Agglomerative Clustering (SBAC), 46
- Similarity function, 21–22
- Similarity measures
for binary variables, 26–27
for comparing continuous-variable data objects, 25
for d -dimensional mixed data objects, 29
for discrete variables with more than two values, 27–29
- Similarity metric, 22
- Simple matching criterion, 28
- Simplified ART (SART), 128. *See also* Adaptive Resonance Theory (ART)
- Simulated Annealing (SA), 92, 95–96.
See also SA- entries
- Single linkage clustering, 34, 37
- Single pass K -means algorithm, 220
- Smart Probabilistic Local Alignment Tool (SPLAT) algorithm, 205
- SMEM algorithm, 78
- Smith-Waterman algorithm, 183–184, 204, 205
- Smith-Waterman alignment score, 208
- SNOB program, 80–81
- Social sciences, use of clustering in, 9

- Social Sciences Citation Index® (SSCI), 9
- SOFM flowchart, 141. *See also* Self-Organizing Feature Maps (SOFMs)
- Soft competition scheme (SCS), 137
- Soft competitive learning, 113, 134, 162
- Soft competitive learning clustering, 130–146
- Soft competitive paradigm, 11
- “Soft-to-hard competitive model transition,” 128
- Spherical *K*-means (*spkmeans*), 69
- Spin-density image, 150, 151
- Split criterion, 78. *See also* Splitting criterion
- Split index, 67
- Splitting criterion, 71, 72
- Spotted cDNA microarray, 47
- Spotted oligonucleotide microarray, 47
- Squared-error-based clustering, 165, 167–170
- Squared-error criterion function, 168
- Squared Euclidean distance function, 273
- Squared Mahalanobis distance, 24
- Square wave influence function, 223–224
- Stability, of an incremental clustering algorithm, 116–117
- Stability-plasticity dilemma, 117, 162
- Stable category learning theorem, 126
- Standard score, 23
- Star index, 67
- State interpretation, 191–192
- State transition probability, estimation of, 192
- State transition probability distribution, 187
- State vector, 152–153
- Static vectors, 211
- STatistical INformation Grid (STING) algorithm, 225–226
- STING+ algorithm, 226
- Stochastic optimal search techniques, 70
- Stochastic optimization methods, 92
- Stopping rules, 269–271
- STRE motif, 100, 101
- Stress functions, 246
- Strongly connected components (SCCs), 208
- Structure-graph (s-graph), 59
- Sub-clusters, 41
- Sub-graphs, 37
- Subspace clustering, 253
- Successive *K*-means initialization, 71
- Sum-of-squared clustering algorithm, 176
- Sum-of-squared-error criterion, 64–65, 66, 109–110
- SUNY CDROM-1*, 176
- Sup distance, 23, 24, 26
- Supervised classification, mixture densities in, 74
- Supervised classification systems, 2
- Supervised learning, 134
- feature selection techniques for, 6
- Support vector clustering (SVC), 165, 170–174, 178
- Support vector regression (SVR), 175
- Support vectors (SVs), 172
- SVM training, 170–171
- SWISS-PROT sequences, 204, 205, 207
- Symbol emission probability distribution, 187
- Symmetric binary features, 26
- Symmetric fuzzy ART (SFART) network, 128. *See also* Adaptive Resonance Theory (ART)
- Symmetry condition, 21, 22
- T1-weighted image, 150, 151
- T2-weighted image, 150, 151
- Tabu search (TS), 92, 99. *See also* TS-based clustering algorithm
- t*-distributions, 80
- Testing criteria, 7
- categories of, 263–264
- Thematic maps, quality assessment of, 103–104
- 3G cellular networks, condition monitoring of, 150–153
- Threshold, 71–72

- Threshold graph, 37
- Time series clustering, 11, 179
- T-lineage acute lymphoblastic leukemia, 161
- Top-down hierarchical clustering, 46
- Topology-preserving maps, 143
- Total mean vector, 65
- Total scatter matrix, 65
- Trace criterion, 65
- Transition probability, 190
- Traveling Salesman Problem (TSP), 229–232
- TreeView software, 52
- Triangle inequality condition, 21
- Triggers, 226
- Triplet distribution visualization, 259
- TS-based clustering algorithm, 99, 100.
See also Tabu search (TS)
- Two-layer clustering scheme, 87
- Two-layer feedforward neural network, 111–112
- Two-mode matrix, 22
- Two-phase agglomerative hierarchical clustering algorithm, 44–45
- Two-phase clustering, 44
- Two-phase cluster labeling algorithm, 172
- Typological analysis, 12

- Ultrametric, 21
- Uniform random sampling, 217
- Universal probabilistic framework, 199
- Unsupervised classification systems, 2–3
- Unsupervised clustering process, 150
- Unsupervised decision tree algorithm, 46
- Unsupervised learning, 12
- Unsupervised LVQ, 134, 135. *See also* Learning Vector Quantization (LVQ)
- Unsupervised pattern recognition system, 106
- Unweighted pair group method average (UPGMA), 35, 46
- Unweighted pair group method centroid (UPGMC), 35–36

- Validation indices, 269–271
- Validity criteria, 277
- Variable-threshold model, 131
- Vector quantization (VQ), 68–69
- Vector space-based clustering algorithms, 185
- Very-large-scale integration (VLSI) design, 117
- Very large-scale integrated (VLSI) circuit clustering, 231–232
- Vigilance criterion, 118, 123, 125
- Vigilance parameter, 125–126, 230
- Vigilance test, 127
- Vision-guided robot navigation, partitional clustering in, 108–109
- Visual evoked potentials (VEPs), multiple sclerosis detection using, 104–108
- Viterbi algorithm, 191
- Voronoi diagram, 45
- Voronoi partition, 68

- Ward's method, 36
- Watson-Crick base pairing rules, 202–203
- WaveCluster⁺, 226, 227, 253
- Wavelet transform, 226, 253
- WebCANVAS, 199
- Web of Science[®], 9
- WEBSOM2 system, 154 construction and operation of, 155
- WEBSOM architecture, 154
- WEBSOM information retrieval system, 153–157, 158
- Weighted images, T1, T2, and SD, 150, 151
- Weighted Pair Group Method Average (WPGMA), 35
- Weighted Pair Group Method Centroid (WPGMC), 36
- Weight function, 251
- Weight matrix, 248
- Weight vectors, 113 redundant, 133
- Winner-take-all (WTA) learning, 113
- Winner-take-all rule, 117, 118, 123

Winner-take-most (WTM) learning, 113
Winning neuron, 144–145
Within-class scatter matrix, 168
Within-cluster scatter matrix, 65
Wobble Aware Bulk Aligner (WABA)
 algorithm, 205–206
Wolfe dual form, 174
Word category maps, 153–154
WTM-based algorithms, 113

Xie-Beni index, 96, 273, 274
XML format, 59. *See also* eXtensible
 Markup Language (XML)
 document clustering
Yeast data sets, cluster analysis of, 54
Yeast genes, major clusters of, 55
z-score, 23