

Index

Note: Figures and Tables are indicated by *italic page numbers*, footnotes by suffix ‘n’

- Abalone database 47, 57–8, 65–6
- Abalone.xml file 65, 66, 435–6
- Agglomerative clustering 51
- Aggregation functions 126
- Aggregation index/measure 241
- Argmin function 183
- Aristotle, definition of objects 22, 36
- Artificial data sets 250, 338
 - discriminant analysis 337–9
 - number-of-clusters determination 250–3
- Assertions 27, 45, 124
- ASSO files 62, 125, 146
 - abalone.xml 65, 66, 435–6
 - enviro.xml 136, 147
 - see also* Data sets; SODAS2 files
- ASSO project 82, 429
- ASSO Workbench 124, 125, 429
 - see also* SODAS2 Software
- Association rules 320
- Average rectangle for vertex-type distance 222

- Background knowledge, retaining after generalization process 11–13
- Bar diagrams 215
- Basque country, time use survey 36, 421–8
- Bayesian decision trees 35, 333–40
- Bayesian rule 335
- Beale test 237–8
 - in examples 251, 254, 257, 258, 259
- Binary tree, construction process 334, 335
- Boolean symbolic data, partitioning of 186–90
 - Boolean symbolic descriptions 125
 - dissimilarity measures for 126–30
 - Boolean symbolic objects (BSOs) 27, 46, 62, 125
 - matching functions for 140–3
 - Bound-based coding 376
 - Breakdown process 85, 114–19
 - Bumps (in kernel method) 152, 247, 336
 - Buneman’s inequality 135

- C-index 237
 - in examples 251, 252, 254, 256, 257, 258, 259
- Calinski–Harabasz index 236
 - in examples 251, 252, 254, 256, 257, 258, 259
- Canonical matching 31, 64, 124
 - between boolean symbolic objects 140–1
- ‘Capacities’ 17
- Car models data set 323–4
 - SGCA applied to 322–30
- Categorical multi-valued variables
 - factorial analysis of 320
 - generalization of 88, 161–2, 166, 376
 - recoding of 376, 377, 379
- Categorical single-valued variables
 - generalization of 88, 161–2, 166, 376
 - recoding of 375, 377, 378–9
- Categorical variables 3, 7, 238
 - dissimilarity measures for 127, 239–40
 - generality degree for 163
 - under hierarchical dependence rules 166–7

- Categorical variables (*Continued*)
 generalization of 88, 161–2
 under hierarchical dependence rules 166
 inclusion and 116
 number-of-clusters determination, examples 256–7
 rules between 166–7
 in time use survey 422
 with high numbers of categories 381
- Categories 3, 8
- Center principal component analysis (CPCA)
 280, 283, 309
 compared with SPCA 284
 example (oils and fats data set) 298–9, 300
 visualization of results on factor planes 294
- Chernoff's distance 131, 132
 χ^2 divergence 131, 132, 145, 183
- China
 climatic variations 385
 meteorological stations data set 199–202, 273–7, 383
- Circle of correlation, *see* Correlation circles
- Class prototypes
 initial configuration of 217–18
 visualizing 214–16, 413, 414
- Classes 8
- Classical data
 classical analysis of 19
 partitioning by dynamic clustering 186
 symbolic analysis of 19
- Classification rules 341, 345–7
- CLINT module (in SODAS2) 442
- Cluster analysis 235
- Cluster cohesion 34
- Cluster isolation 34
- Cluster prototypes 220–3
- Cluster stability 263–78
 measuring w.r.t. cohesion criterion 266
 measuring w.r.t. global validity criterion 266–7
 measuring w.r.t. isolation criterion 265–6
- Cluster validation 235–78
- Clustering 3
 agglomerative 51
 dynamic 33, 182–202
 hierarchical 33, 50–1, 157–79
 pyramidal 33, 157–79
- Clustering algorithm on dissimilarity data tables 33
see also DCLUST algorithm
- Clustering algorithm on distance tables 181, 191–3
see also DCLUST algorithm
- Clustering criterion 160–3
- Clustering interpretative aids 33, 193–5
- Clustering methods
 aims 149, 158
 divisive classification 32, 149–56
 hierarchical clustering 33, 50–1, 157–79
 pyramidal clustering 33, 157–79
see also Divisive classification/clustering, algorithm; Hierarchical clustering; Pyramidal clustering
- Clustering modules (in SODAS2) 441–2
- Clustering problem 149
- Clustering tree method 150–4
 application to interval data 154
 bumps and multi-modalities 152
 example 154–6
 gap test 153
 non-homogeneous Poisson process 150–1
 kernel method used to estimate intensity 151–2
 output data 154
 pruning method 153–4
 splitting criteria 152–3
 stopping rule 153
- Clusters
 meaning of term 149
 number of 235–62
- Cognitive psychology, visual data mining explained using 32, 109
- Cohesion criterion, cluster stability measured w.r.t. 266
- Complete symbolic object 29–30
- Complex data 3, 13–14
 mining of 45–59
- Componentwise dissimilarities 126
- Computer methodologies 429–44
see also SODAS2 software
- Concepts 8–9, 20, 37, 160
 extent of 8–9, 26
 modelling by symbolic description 23–9
- Conceptual clustering 157
- Conceptual lattice 29–30
- Conceptual variables 10, 31
- Conjunctive data 15
- Constrained boolean symbolic descriptions 126n2
 dissimilarity measures for 126
- Contingencies, retention after generalization 12–13
- Contribution measure 194–5
 listed in worked example 200
- Convex hulls, visualization of PCA results as 294–5
- Cooling parameter, in SYKSOM algorithms 226–7
- Correlation circles 293, 296
 in example PCAs 297, 300, 302, 307, 309

- Correlation interval matrix 293
 - computation of 310
- Correlations, recapture after generalization 12
- CPCA, *see* Centres principal component analysis (CPCA)
- Criterion-based divisive clustering algorithm 50–5
- Cross-country comparisons 35–6, 405–19
- Cross-entropy distance 378, 379, 380
- Cut rule (for binary trees) 334, 336
- Cut rule(s) for decision trees 154, 334, 336, 337
 - examples 155, 255, 339
- Data management modules (in) 439, 441
- Data mining 19–20
 - extension to second-order objects 4, 22, 37
 - visualization used in 109–10, 322
- Data sets
 - abalone 47, 57–8, 65–6
 - artificial 250, 338
 - car models 323–4
 - e-Fashion stores data set 257–60
 - Merovingian buckles data set 256, 271–2
 - meteorological stations in China 199–202, 273–7
 - micro-organisms 195–9
 - oils and fats 155, 253, 297
 - Portuguese people gender/age/employment 348–50
 - work/demographic/medical 365–6, 370–1
- Database catalogues 68
- Database management system (DBMS) 68
- Database relations 96–7
 - compared with symbolic descriptions 97
- Databases
 - extracting knowledge from 5, 45–59
 - symbolic objects exported to 61–6
 - see also* Relational databases
- DB2SO module (in SODAS2) 10, 22, 30, 45, 48, 97, 136, 439
- DCLUS algorithm 192–3
- DCLUS module (in SODAS2) 33, 181, 191–3, 442
 - generality of method 192
 - input data 192
 - output data 192–3
- De Carvalho distance 130, 240
 - in examples 258
- Decision trees 4, 20, 35
 - binary questions at each node division step 153, 335–6
 - construction process 334, 335
- Definiteness property (of dissimilarity matrix) 134
- Description potential 97–8, 284–5, 345
 - computation of 98–9
 - under normal symbolic form 100–1
 - dissimilarity measures based on 126–7, 190
 - linear 285
- Description space 24
- Descriptive statistics modules (in SODAS2) 441
- Descriptor potential increase (dpi)
 - index 346
 - compared with other dissimilarity measures 345–7
 - in example 355–7
- Dimensionality, curse of 343
- Discrimination analysis
 - by Bayesian decision tree 333–40
 - SODAS modules implementing 443
- DISS module (in SODAS2) 32, 124, 125, 126, 127, 441
 - applications 136–9, 416, 417
 - input data 125
 - output 134, 171, 191, 243
- Dissimilarity data tables, clustering on 33
- Dissimilarity matrix 134
 - properties 134–5
 - representation/visualization of 134–5, 138–9, 416, 417
- Dissimilarity measures 124, 125–34, 238–40
 - for boolean symbolic descriptions 126–30, 188
 - comparison of 127, 183, 239, 345–7, 356–7
 - for probabilistic symbolic descriptions 130–4
 - SODAS2 module implementing 441
 - for symbolic objects 238–40
- DIV module (in SODAS2) 241, 442
 - application in examples 255–6, 400–1, 410–11, 411, 412, 414–16
 - compared with SCLASS 410–11, 412
- Divergence coefficients 131–2
- Divisive classification/clustering 149–56
 - algorithm 32, 50–5, 149
 - application to interval data 154, 400
 - clustering tree method 150–4
 - input data 150
 - output data 154
 - pruning method 153–4
 - splitting criteria for 152–3
 - stopping rule for 153
 - examples 154–6, 400–2, 410–11, 412, 414–16
- Drilldown 85, 118–19
- DSTAT module 441

- Duda–Hart number-of-clusters test 236–7
- Dynamic clustering method 33, 181–204
 allocation step 182
 existence and uniqueness conditions 182–3
 generality 182
 representation step 182
- E-Fashion stores data set
 modal variables 258
 number-of-clusters determination 257–60
- Editing of symbolic data 31, 81–92
- Eight-point neighbourhood 224, 225
- Ellipsoidal null model 269
- Entity–relationship model 71
- Envelope-type prototype 220–1
- Euclidean distance 131, 184
- European Social Survey (ESS) 35–6, 395–6
 background variables 396
 Finnish/Spanish/Portuguese data 396
 divisive classification 400–2
 hierarchical and pyramidal clustering 402–3
 zoom star visualization 399–400
 ‘political’ variables 396, 397
 ‘trust’ variables 396, 405, 406
- Eurostat 4
- Evenness property (of dissimilarity matrix) 134
- Exploratory data analysis, extension to
 second-order objects 4, 37
- Exponential distribution kernel 212, 225, 227
- Extent, meaning of term 8, 21–2, 61, 160
- Factor analysis methods
 extension to symbolic objects 34, 279–330, 341–357
see also Generalized canonical analysis;
 Principal component analysis (PCA)
- Factor discriminant analysis 35, 341
 on symbolic objects 341–357
see also SFDA module
- Factorial techniques SODAS2 modules
 implementing 442
- Finnish people
 compared with Portuguese and Spanish
 ‘life-value’ variables 409
 political opinions 35–6, 399–403
 ‘trust’ variables 409
- First-level objects 20, 36
- First-order units 8
 examples 5, 6, 7, 8
- Fission rule (for hierarchies and pyramids) 170
- Flexible matching 31, 64, 124–5
 between Boolean symbolic objects 141, 142–3
 between probabilistic symbolic objects 143–5
- Form recognition 110
- Fusion rule (for pyramids) 170
- Fuzzy coding 315–16, 342, 347, 355
- Fuzzy data 14
- Galois lattice 29
- Gap test 34, 153–4, 247–8
 application to interval-valued data 248–50
 in examples 251–2
- Gaussian distribution kernel 212, 225, 227
- Generality degree criterion 160, 163–4, 427
 for categorical variables 163
 under hierarchical dependence
 rules 166–7
 for modal variables 163–4
 under hierarchical dependence
 rules 168–9
- Generalization by intersection 87, 88
- Generalization by maximum (for modal variables) 89
- Generalization by minimum (for modal variables) 89–90
- Generalization process 9–10, 47, 86–91, 161–3
 background knowledge retained
 after 11–13
 for categorical variables 88, 161–2
 under hierarchical dependence rules 166
 improvement by decomposition 55–6
 for interval variables 87, 161
 for modal variables 88–90, 162, 163–4
 under hierarchical dependence rules 167–8
 for ordinal variables 163
 supervised approach 45–6
 for taxonomic variables 90, 163
- Generalization by union 87, 88
- Generalization when size is known (for modal variables) 88–9
- Generalized canonical analysis 34, 35, 313, 314
 of symbolic objects 313–30
- Generalized hypervolumes clustering
 criterion 246
- Generalized hypervolumes clustering method 245–7
- Global growth factor 105
- Global validity criterion, cluster stability
 measured w.r.t 266–7
- Goodman–Kruskal index/indices 321, 343
- Graphical representation
 symbolic hierarchy 174, 176
 symbolic pyramid 176–7

- Hamming distance 131
- Hausdorff distance 20, 187, 188, 219–20, 288, 346–7, 410
 - compared with other dissimilarity measures 127, 183, 239, 345–7
 - in example 355–7
- Hausdorff-type distance, in SYKSOM algorithm 220
- Hausdorff-type L_1 -distance 220
 - median prototype for 222–3
- Hellinger coefficient 131, 145
- Hierarchical clustering 33, 50–1, 157–79
 - algorithm 164–5
 - classical methods 241–3
 - centroid method 242–3
 - complete linkage method 242
 - single linkage method 241
 - symbolic extensions 243
 - Ward method 243
 - examples 175–7, 402–3
 - pruning 169
 - SODAS software used 171–5
 - see also HIPYR algorithm
- Hierarchical dependence
 - between categorical variables 166–7
 - between modal variables 167–9
- Hierarchical dependence (HD) rule 317
- Hierarchical dependencies 94, 126n2
 - memory growth under 104–5
- Hierarchically dependent variables 10, 90–1, 94, 166, 362, 363
 - generalization of 91
 - linear regression for 367–70
 - example 368–70
 - input data 367
 - methodology 367–8
- Hierarchies of symbolic objects 117
- Hierarchy
 - definition 117, 158–9, 159
 - graphical representation of 174–5, 174
 - rule generation for 170
- Hierarchy tree 174, 176, 363
 - pruning of 169
- Higher-level units 3
 - examples 5, 6, 7, 8
- HIPYR algorithm 173
- HIPYR module (in SODAS2) 33, 171–5, 442
 - applications 412, 413, 426–7
 - objectives 171–2
 - options and parameters 172–3
 - output
 - graphical representation 174–5, 413
 - as listing 173–4
- Histogram-valued observations, linear regression of 361–2
- Hoards 27, 59
- Homogeneous data set, cluster stability measures for 268–70
- Homogeneous groups 149
- Homogeneous Poisson process 151, 244
 - conditional uniformity property 244
- Hybrid numbers theory 288
- Hypervolumes clustering criterion 34, 245
- Hypervolumes clustering method 244–5
- Hypervolumes test 34, 247
 - application to interval-valued data 248–50
 - in examples 251–2, 253, 254
- Ichino–de Carvalho dissimilarity index 345–6
 - compared with other dissimilarity measures 345–7
 - in example 355–7
- Icon-based representation/visualization methods 112
- Imprecise data 15
- Inclusion between concepts or symbolic objects 115–16
 - definition by extension 115
 - definition by intension 115
- Individuals
 - meaning of term 8, 20, 81
 - retrieving 63–4
- Inertia criterion 54
- Inputs of symbolic data analysis 10–11
- Intent–extent duality 160
- Intent, meaning of term 8, 160
- Inter-country comparisons 35–6, 405–19
- Internal variation 3
- Interpretative aids, clustering 193–5
- Interval algebra 280, 288
- Interval algebra based methods 288–93
 - hybrid approach 288–91
 - IPCA 291–3
 - MRPCA 288–9
 - spaghetti PCA 290–1
- Interval arithmetic 381
- Interval correlation, computation of 310
- Interval principal component analysis (IPCA) 280, 288, 291–3, 309
 - example (oils and fats data set) 306–8, 308–9
 - standardization of data for 310
 - visualization of results on factor planes 296–7
- Interval-valued data
 - dissimilarity measures for 188, 239
 - linear regression of 360–1
 - principal component analysis extended to 281–2, 291–3

- Interval variables 7, 123, 238
 - divisive clustering algorithm used 150, 154
 - in European Social Survey 396, 407
 - fuzzy coding of 315
 - generalization of 87, 161
 - inclusion and 116
 - number-of-clusters determination, examples 250–6
 - recoding of 375–6, 377–8
- IPCA, *see* Interval principal component analysis (IPCA)
- J-coefficient (J-divergence) 132
- J-index 236–7
 - in examples 251, 254, 257, 258, 259
- Joint distribution, modelling description by 23
- K-criterion 231
- K-nearest-neighbour algorithm 382
- Kernel functions 212, 225
- Kernel method, estimation of intensity of non-homogeneous Poisson process using 151–2, 246–7, 336
- Kernel, properties 152, 247, 336
- Knowledge base 97
- Knowledge discovery 45–59
- Knowledge mining 22
- Kohonen maps 33, 205–33
 - meaning of term 206
 - reason for use in data analysis 206, 213
 - visualizing SYKSOM output by means of 213–16
 - see also* SYKSOM algorithms
- KT estimate 132
- Kullback divergence 131
- Kullback–Leibler (KL) divergence 131, 145
- L_1 distance 127, 130, 239
 - in example 258
- L_2 distance 127, 130, 183, 239
 - in example 259
- Lattice 207
 - Cartesian coordinates 225
- Lattice structure of symbolic objects 29–30
 - see also* Conceptual lattice
- Learning factors 223–4
- Lebesgue measure 151, 248
- ‘Life-value’ variables 406
 - divisive clustering 414–16
 - inter-country comparisons 407–9, 414–16
- Line representation, of dissimilarity matrix 416, 417
- Linear description potential (LDP) 285
- Linear projection 206–7
- Linear regression 360
 - with histograms 361–2
 - with interval-valued data 360–1
- Local growth factor 104
- Loevinger’s index 265, 266, 267, 270
- Logical dependence (LD) rule 317
- Logical dependencies 10, 94, 126n2
- Logistic activation function 379
- Low-quality data, in multi-layer perceptron methods 386–90
- MacQueen algorithms 212, 230–3
 - compared with StochApprox algorithm 233
- MacQueen’s clustering method for data points 229
- MacQueen’s clustering method for interval data, symbolic version 230
- Manhattan distance 131
- Margins, modelling description by 23
- MATCH module (in SODAS2) 27, 31, 32, 124, 125, 140, 145, 441
 - application 146–7
 - input data 125
 - output 145–6
- Matching functions 31, 32, 124–5, 139–45
 - for Boolean symbolic objects 140–3
 - for probabilistic symbolic objects 143–5
- Matching operators SODAS2 module implementing 441
- Maximum covering area rectangles 284
- Maximum likelihood estimation 379
 - principle 378, 380
 - rule 335
- Mean and length coding 375
- Median prototype 223
- Membership functions 15, 26, 31, 61
 - resemblance index 269–70
 - scores 269
- Memory growth, under normal symbolic form transformation 103–5
- Merging of symbolic data tables 32, 91–2
- Merovingian buckles data set 256, 272
 - cluster validation 271–3
 - number-of-clusters determination 256–7
- Metadata 31, 68, 438
 - in SODAS2 438–9
 - for symbolic data table 71, 125
 - symbolic descriptions enriched by 30–1
 - for symbolic object 70
 - for symbolic variables 70–1
 - for variables 70
- Metadata representation 76, 77
- Metainformation, *see* Metadata
- Meteorological analysis 381, 383–90
- Meteorological stations data set 199–202, 273–7

- Micro-data 62
 - see also* Individuals
- Micro-organism data set 196
 - dynamic clustering application 195–9
- Midpoints radii principal component analysis (MRPCA) 280, 288–9, 309
 - example (oils and fats data set) 302–5
 - visualization of results on factor planes 295–6
- Mini-clusters 208
- Minimal cluster inertia 263
- Minimum covering area rectangles (MCARs) 293, 294
 - in example (oils and fats data set) 297, 299, 301, 304, 307, 309
- Minkowski L_p distance 131, 132
- Minkowski metric 126, 132, 189
- Missing data, ways of handling 54, 382
- Mixed symbolic data 314, 347, 381
 - partitioning of 190–1
- Mixed symbolic descriptions, dissimilarity measures for 126
- Modal symbolic data, partitioning of 190
- Modal symbolic objects 27, 46
- Modal variables 123, 238
 - generality degree for 163–4
 - under hierarchical dependence rules 168–9
 - generalization of 88–90, 162
 - under hierarchical dependence rules 167–8
 - inclusion and 116
 - number-of-clusters determination, examples 257–60
 - in political opinions survey 396, 397
 - recoding of 376, 377, 380
 - rules between 167–9
 - with high numbers of categories 381
- Mode, distinguished from ‘bump’ 152, 247
- Monte Carlo simulations 264, 268
- Mother–daughter variables 10, 94, 362
 - see also* Hierarchically dependent variables
- MRPCA, *see* Midpoints radii principal component analysis (MRPCA)
- Multi-layer perceptrons (MLPs) 35, 373–91
 - construction of 374
 - effect of low-quality data 386–90
 - examples 382–90
 - model selection for 374–5
 - numerical coding approach 375–82
 - in example 384
 - recoding a symbolic variable 375–6
 - recoding the inputs 376–7
 - recoding the outputs 377–80
 - problems
 - high number of categories 381
 - missing data 382
 - multiple outputs 381–2
 - taxonomies 382
 - symbolic approaches 35, 375–81
 - benefits compared with standard approaches 382–3, 386–7, 389–90
 - training 374
 - factors affecting 381
- Multi-valued variables 238
 - dissimilarity measures for 127, 239–40
 - generalization of 88, 161–2
 - number-of-clusters determination, examples 256–7
- Multidimensional data, visualization methods for representing 111, 112
- Multiple correspondence analysis on symbolic data (SMCA) 314
- Multivariate analysis 419
- Native data file 31, 82
 - importation from 82–3, 398
- Natural symbolic objects 117
- NBCLUST module (in SODAS2) 127, 133, 243
 - application in examples 251, 253, 254, 256, 257, 258, 259, 272, 274
- ND2SO module (in SODAS2) 83, 398, 439
- Neural net models 373–4, 382
 - see also* Multi-layer perceptrons (MLPs)
- Neurons 33, 373
- Non-applicable values (nulls) 94, 126n2, 166
- Non-homogeneous Poisson process 151, 245
 - estimation of intensity 151–2, 246–7
 - transformation to homogeneous Poisson process 248
- Non-linear operation of Kohonen approach 206
- Non-linear regression, *see* Multi-layer perceptrons (MLPs)
- Normal kernel 152, 247
- Normal symbolic form (NSF) 32, 93–107
 - advantages 102
 - applications 105–6, 317
 - computation of description potential under 100–1
 - computation time reduction using 106
 - definition 101–3
 - meaning of term 99–100
- Number of clusters
 - criteria for 236–8
 - determination of 235–62
 - examples 250–60
 - statistical tests 247–50
 - in partition 272

- Numerical recoding approach for multi-layer perceptrons 375–81
 choice in real-world examples 383, 384, 390
- Object-oriented paradigm 71–2
- Objects
 first-level 20, 36
 second-level 20, 36–7
- Oils and fats data set 155, 253, 297
 divisive clustering 154–6, 255–6
 number-of-clusters determination 253–6
 principal component analysis 297–309
 CPCA 298–9, 299–300
 IPCA 306–8, 308–9
 MRPCA 302–5
 ‘spaghetti’ PCA 305–6, 307–8
 SPCA 299, 301–2
 VPCA 297–8, 298–9
- Ordinal variables, generalization of 163
- Overygeneralization 12, 56, 294
 avoidance of 30, 56–7
- Parallel coordinate representation/visualization
 methods 112
- Parallel Edges Connected Shape (PECS) 295
- Partial membership 269
- Partition, stability measures of 267–8
- Perceptor model 109
- Pie chart representation, of dissimilarity matrix
 416, 417
- Poisson process 151, 244
see also Homogeneous Poisson process;
 Non-homogeneous Poisson process
- Political opinions, inter-country comparison
 (Finland/Portugal/Spain) 35–6, 399–403
- Portuguese people
 compared with Finnish and Spanish
 ‘life-value’ variables 409
 political opinions 35–6, 399–403
 ‘trust’ variables 409
 cultural survey data 175–7
 gender/age/employment data set 348–50
 factor discriminant analysis 347–57
- ‘Possibility’ 17
- Power of discrimination 194
- Premise (conclusion) variable 94
- Principal component analysis (PCA) 4, 20, 34, 206, 279–311
 extension to interval data 291–3
 extension to symbolic data 207, 283–8
 applications 299–302, 418
 visualization of results on factor planes
 293–7, 298, 300, 301, 304, 307, 309
see also SPCA module
- Principal component analysis w.r.t. reference
 subspace (PCAR) 284
- Prior probabilities 335–6
- Probabilistic symbolic descriptions 125
 dissimilarity measures for 130–5
- Probabilistic symbolic objects (PSOs)
 27, 62, 125
 matching functions for 143–5
- Probability distributions, comparison
 functions for 130–2, 145
- Propagation on database 62
- Proportionate sampling 264–5, 272
- Proximit initial configuration 217–18
- Pruning
 in clustering tree method 153–4
 decision trees 337
 in hierarchical or pyramidal clustering
 169, 177
- Pseudo-metric (of dissimilarity matrix) 135
- Pyramid
 definition 159
 graphical representations 174–5, 176–7,
 403, 426
 pruning of 169
 rule generation for 170
- Pyramidal clustering 33, 157–79
 algorithm 164–5
 examples 175–7, 402–3, 426–8
 pruning 169
 SODAS software used 171–5
see also HIPYR algorithm
- Quality index
 of cluster 194
 listed in worked example 200
 of partition 194
- Quality, metadata model 80
- Quantitative variables, inclusion and 115
- Quartile range intervals 397
- Radial coordinate representation/visualization
 methods 112
- Radius rotation matrix 289
- Random data table 17
- Range transformation, principal component
 analysis (RTPCA) 284–6
 combined with VPCA 287–8
 compared with VPCA 287
- Reconstruction process 103
- Reference partitions 265
- Reference variables 102
- Regression analysis
 SODAS module implementing 443
 on symbolic data 35, 359–72
 applications 370–1, 418

- see also* Linear regression; Multi-layer perceptrons (MLPs); SREG module
- Reification process 20–1, 31
- Relational databases 21, 46
construction of symbolic objects from 21–2, 46–50
- Relations in databases 96–7
compared with symbolic descriptions 97
- Rényi's divergence 131, 132
- Resemblance measure 32, 140
- Retrieving individuals 63–4
- Robinsonian property (of dissimilarity matrix) 135
- Root (of binary tree) 334
- RTPCA, *see* Range transformation, principal component analysis (RTPCA)
- Rule discovery 13
- Rule generation, in hierarchical or pyramidal clustering 170
- Rules, recapture after generalization 13
- SBTREE module (in SODAS2) 35, 443
- SCLASS module (in SODAS2) 32, 241, 442
application in examples 254–5, 410, 411, 412
compared with DIV 410–11, 412
- SCLUST module (in SODAS2) 20, 33, 34, 127, 133, 181, 191, 241, 244, 249, 442
application in examples 251, 252, 253, 253, 254, 256–7, 258, 259, 272, 274, 411, 412, 413
- SDD module (in SODAS2) 443
- SDT module 443
- Second-level objects 20, 36–7
- Second-order units 8
examples 5, 7, 8
- Semi-distance (of dissimilarity matrix) 135
- Set-valued variables 123, 238
- SFDA module (in SODAS2) 35, 443
- SFDR module (in SODAS2) 443
- SGCA, *see* Symbolic generalized canonical analysis (SGCA)
- SGCA module (in SODAS2) 35, 442
- SHICLUST module (in SODAS2) 241, 243, 248
application in examples 251, 252, 254, 257, 258, 259
- Short-term memory, in data mining 109–10
- Simultaneous component analysis with invariant pattern (SCAP) 289
- SMCA, *see* Multiple correspondence analysis on symbolic data (SMCA)
- SMLP module (in SODAS2) 443
- Smoothing parameter 152, 247, 336
- SO2DB module (in SODAS2) 31, 61–6, 441
application 65–6
input data 62–3
output 64–5
- SODAS2 software 4, 429–44
architecture/structure 82, 433–4
chaining window 430–1, 431, 433
CLINT module 442
clustering modules (listed) 441–2
data management modules (listed) 439, 441
data management procedure 431
data types used 434
DB2SO module 10, 22, 30, 45, 48, 97, 136, 431, 439
DCLUST module 33, 181, 191–3, 442
definitions used 20
descriptive statistics modules (listed) 441
discrimination and regression modules (listed) 443
DISS module 32, 124, 125, 126, 127, 441
applications 136–9, 416, 417
input data 125
output 134, 171, 191, 243
DIV module 241, 442
application in examples 255–6, 400–1, 410–11, 411, 412, 414–16
compared with SCLASS 410–11, 412
DSTAT module 441
execution of analysis 432–3
factorial modules (listed) 442
features 430
file format 434–7
graphical user interface 430
HIPYR module 33, 170–5, 442
applications 412, 413, 426–7
objectives 171–2
options and parameters 172–3
output 173–5
main windows 430–1
MATCH module 27, 31, 32, 124, 125, 140, 145, 441
application 146–7
input data 125
output 145–6
metadata 438–9
methodology 433–9
modules 433–4, 439–44
data management modules 439, 441
treatment modules 441–3
visualization modules 443–4
NBCLUST module 127, 133, 243
application in examples 251, 253, 254, 256, 257, 258, 259, 272, 274

- SODAS2 software (*Continued*)
- ND2SO module 83, 398, 431, 439
 - overview 36, 430–3, 440
 - Parameters windows for data file 437
 - preparing an analysis 431–2
 - SBTREE module 35, 443
 - SCLASS module (in SODAS2) 32, 241, 442
 - application in examples 254–5, 410, 411, 412
 - compared with DIV 410–11, 412
 - SCLUST module 20, 33, 34, 127, 133, 181, 191, 241, 244, 249
 - application in examples 251, 252, 253, 253, 254, 256–7, 258, 259, 272, 274
 - SDD module 443
 - SDT module 443
 - SFDA module 35, 443
 - SFDR module 443
 - SGCA module 35, 322, 442
 - SHICLUST module 241, 243, 248
 - application in examples 251, 252, 254, 257, 258, 259
 - SMLP module 443
 - SO2DB module 31, 61–6, 431, 441
 - application 65–6
 - input 62–3
 - output 64–5
 - SOEDIT module 31, 81, 85–6, 91, 431, 437, 441
 - SPCA module 280, 418, 442
 - dialog box for defining parameters 432
 - SREG module 35, 370, 418, 443
 - starting 431–3
 - STAT module 17
 - SYKSOM module 205–6, 210–13, 442
 - basic steps 210–13
 - example (European Social Survey) 411, 412, 413, 414
 - MacQueen algorithms 229–33
 - StochApprox algorithm 227, 228, 233
 - technical definitions and methodological options 217–27
 - treatment modules (listed) 441–3
 - TREE module 443
 - VDISS module 32, 125, 134, 443
 - VIEW module 213, 214–16, 441
 - visualization modules (listed) 443–4
 - VMAP module 213, 214, 444
 - VPLOT module 191, 213, 216, 443
 - VPYR module 174, 444
 - VSTAR module 191, 193, 443
 - VSTAT module 443
 - VTREE module 411, 444
- Softmax activation function 378, 380
- Software 171, 191, 393, 429
- SOML files 83
- ‘Spaghetti’ principal component analysis 280, 290–1, 309
 - example (oils and fats data set) 305–6, 307–8
 - visualization of results on factor planes 296
- Spanish people
 - compared with Finnish and Portuguese ‘life-value’ variables 409
 - political opinions 35–6, 399–403
 - ‘trust’ variables 409
- SPCA, *see* Symbolic principal component analysis (SPCA)
- SPCA module 280, 418, 442
 - dialog box for defining parameters 432
- Splitting criteria, in tree growing 152–3, 337
- SREG module 35, 370, 418, 443
- Stability-based validation method 263–78
 - applications
 - Chinese meteorological data set 273–7
 - Merovingian buckles data set 271–3
- Stability measures 264–8
 - of clusters 264–7
 - interpretation of 270–1
 - of partitions 267–8
- Standard data tables, extraction into symbolic data tables 4, 5–8
- Star representation, *see* Zoom star representation/visualization
- Statistical metadata 68
- Statistical metadata model(s) 31, 67–80
 - background to 69
 - case study 78–9
 - general requirements for development of 69–71
 - metadata to be included 69–71
 - properties for metadata items selected to be modelled 71
 - selection of modelling technique 71–2
 - step-by-step development of 72–6
 - metadata submodel for original data 72–3
 - metadata submodel for symbolic data 73–6
- Statistical templates 68–9
- StochApprox algorithm 212, 227, 228
 - compared with MacQueen algorithms 233
- Stochastic approximation (for Kohonen maps) 212, 228
- Stopping rule(s)
 - divisive clustering algorithm 153
- SYKSOM algorithms 213
 - symbolic dynamic clustering algorithm 185, 186, 199

- Structured data 17
- Superimposition of symbolic objects 114, 114, 119
- Supervised methods 35, 331–91
- Survey metadata 70
- SYKSOM algorithms 205–6, 210–13, 442
 - basic steps 210–13
 - construction of Kohonen map 212
 - initialization 210
 - iteration cycles 212–13
 - iteration step 210–12
 - stopping rule 213
 - distance measures 218–20
 - example (European Social Survey) 411, 412, 413, 414
 - MacQueen algorithms 212, 230–3
 - StochApprox algorithm 227, 228
 - technical definitions and methodological options 217–27
 - cluster prototypes 220–3, 413
 - cooling 226–7
 - initial configuration of class prototypes 217–18
 - kernel function 224–5
 - learning factors 223–4
 - metrics and distances for hypercubes 218–20
 - visualizing output by means of Kohonen maps 213–17
 - see also* Kohonen maps
- Symbolic clustering 33, 157–79
 - basic method 158–65
 - example 175–7
 - postprocessing 169–70
 - in presence of rules 165–9
 - SODAS software used 171–5, 241
- Symbolic clustering procedures 241–3
- Symbolic data 9, 359
 - classical analysis of 19–20
 - creation of 81–5
 - editing of 31, 81–92
 - representation/visualization methods 111–12
 - symbolic analysis of 20
 - visualization by Kohonen approach 207–10
 - implementation by SYKSOM algorithms 210–33
- Symbolic data analysis
 - basis 22, 36–7
 - early papers 4
 - future developments 37
 - general applications 13–20
 - general theory 23
 - inputs 10–11
 - literature survey for 4
 - philosophical aspects 20–1
- Symbolic data analysis (SDA)
 - aims 4, 123
 - principles 22
 - steps 21–2
- Symbolic data tables (SDTs) 69, 123
 - creation of 4, 5–8, 31, 67, 83–5
 - extraction from standard data tables 4, 5–8
 - interactive creation of 83–5
 - merging of 32, 91–2
 - metadata for 71
 - metadata representation on 76, 77
 - transformations for 76, 85–6
- Symbolic descriptions 9, 74, 81, 123
 - coherence within 95–7
 - constraints on 94
 - enrichment of 30–1
 - generalization of 86–91
 - production of 9, 86–91
 - in time use survey 423
- Symbolic dynamic clustering algorithm 184–91
 - allocation step 185, 186, 187, 197
 - applications
 - meteorological stations in China 199–202
 - micro-organism data 195–9
 - initialization 185, 197
 - partitioning of Boolean symbolic data 186–90
 - partitioning of classical data 186
 - partitioning of mixed symbolic data 190–1
 - partitioning of modal symbolic data 190
 - representation step 185, 186, 187, 197
 - stopping rule 185, 186, 199
 - see also* SCLUST module
- Symbolic dynamic clustering method 183–4
 - input data 183–4
 - symbolic interpretation 184
- Symbolic factor discriminant analysis (SFDA) 35, 341–58
 - example (gender/age/employment data set of people in Portugal) 347–57
 - principles 342
 - steps
 - analysis on transformed predictors 344
 - coding of symbolic descriptors 342
 - definition of classification rule 345–7
 - quantification of class predictors 342–3
 - selection of variables 343–4
 - symbolic interpretation of results 344–5
 - see also* SFDA module

- Symbolic generalized canonical analysis (SGCA) 34, 35, 205
 example (car models data set) 322–30
 input data 314
 strategy 314–22
 coding of symbolic descriptors 314–17
 GCA on matrix \mathbf{Z} under cohesion
 constraint 321–2
 logical rules in symbolic object description
 space 317–20
 taxonomical structure on categories of
 symbolic categorical multi-valued
 descriptors 320–1
see also SGCA
- Symbolic linear regression methodology
 359–72
 applications 370–1
see also SREG module
- Symbolic–numerical–symbolic techniques
 factor discriminant analysis (SFDA) 341–58
 generalized canonical analysis (SGCA)
 313–30
 principal component analysis (SPCA)
 282–8
- Symbolic objects
 advantages 28
 attributes 74
 definition 26, 61, 74, 81, 123–4
 exporting to databases 61–6
 extent of 21–2, 61
 factor discriminant analysis on 341–58
 generalized canonical analysis on 313–30
 generation from relational databases 45–59
 hierarchies 117
 kinds 27
 metadata for 70
 modelling concepts by 23–9
 basic spaces for 24–6
 principal component analysis on 34, 205,
 206, 280, 283–8, 309
 quality 28–9
 refinement of 30, 48–50
 visualization of effect 49–50
 reliability 29
 robustness 28–9
 star representation 32
 superimposition of, in zoom star
 representation 114, 114, 119
 syntax in case of assertions and hoards 27
 zoom star representation 32
- Symbolic principal component analysis (SPCA)
 34, 205, 206, 280, 283–8, 309
 compared with CPCA 284
 example (oils and fats data set) 299, 301–2
 visualization of results on factor planes 294
see also SPCA
- Symbolic sequential clustering approach 230
- Symbolic variables 9, 69, 81
 definitions 123, 238
 metadata for 70–1
- T-conorms 9
- T-norms 9
- Taxonomic variables 362
 generalization of 90, 163
 linear regression for 363–7
 example 365–7
 input data 363
 method of regression 363–5
- Taxonomies 314, 320–1, 362
- Taxonomy tree 10, 11, 90, 362
- Terminal nodes (of binary tree) 334
- Threshold kernel 212, 225, 227
- Time use survey(s) 421
 Basque country 36, 421–8
 socio-demographic variables 422
 time-use variables 422
- Topological correctness 207
- Transformations
 symbolic data table 76, 85–6
 symbolic object 75
- Treatment modules (in SODAS2) 441–3
- Tree-growing algorithms
 Bayesian decision trees 35, 334, 335
 clustering tree method 150
- TREE module (in SODAS2) 443
- ‘Trust’ variables 396, 406
 inter-country comparisons 409
- Tucker congruence coefficient 289, 303
- Two-dimensional projected parallelogram
 convex hull (2DPP-CH) 295
- Typicality measure 221–2
- Ultrametric property (of dissimilarity matrix)
 135
- Uncertainty, and symbolic data 16–17
- Unsupervised divisive classification 149–56
- Unsupervised methods 32–5, 121–330
- Vague point 269
- Validation of clustering structure 235–62
- Variation distance 131
- VDISS module (in SODAS2) 32, 125, 134,
 443
- Vertex-type distance 219
 average rectangle for 222
- Vertices data matrix 281
- Vertices principal component analysis (VPCA)
 280, 282–3, 309
 compared with RTPCA 287
 combined with RTPCA 287–8

- example (oils and fats data set) 297–8, 298–9
- visualization of results on factor planes 293–4
- VIEW module (in SODAS2) 213, 214–16, 434, 435, 441
- Visual breakdown 115, 118–19
- Visual data mining 32, 109
- Visual perception 109
- Visualization 32, 109–20
 - applications 259–60, 399–400, 408, 409, 417, 424–6
 - in data analysis 110–11
 - of dissimilarity matrix 134–5, 138–9, 416, 417
 - as exploratory tool 110
 - multidimensional data 112
 - SODAS modules (listed) 443–4
- VMAP display 213, 214, 444
- VPCA, *see* Vertices principal component analysis (VPCA)
- VPLOT display 191, 213, 216, 443
- VPYR module (in SODAS2) 174, 444
- VSTAR module (in SODAS2) 191, 193, 443
- VSTAT module (in SODAS2) 443
- VTREE module (in SODAS2) 411, 444
- Ward criterion 54
- Weight decay parameter 376
- Winsorization 396–7
- Work/demographic/medical data set 365–6, 370–1
- Wrapping effect 294, 308, 309
- XML files 65, 66, 136, 147, 435–6
- Zoom star representation/visualization 32, 112–13, 215–16
 - applications 259–60, 399–400, 408, 409, 424–5
 - metadata 76, 77
 - superimposition of 114, 409
 - symbolic hierarchy 176
 - three-dimensional plots 112, 113, 176, 259–60, 399–400
- Γ -index 237
 - in examples 251, 252, 254, 256, 257, 258, 259

