

Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data Kampert, M.M.D.

Citation

Kampert, M. M. D. (2019, July 3). *Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data*. Retrieved from https://hdl.handle.net/1887/74690

Version:Not Applicable (or Unknown)License:Leiden University Non-exclusive licenseDownloaded from:https://hdl.handle.net/1887/74690

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <u>http://hdl.handle.net/1887/74690</u> holds various files of this Leiden University dissertation.

Author: Kampert, M.M.D. Title: Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data Issue Date: 2019-07-03

Index

1SE rule, 73

ApoE3 data, 76 attribute dispersion, 56, 100 distance, 10, 26 importance, 18, 174 weighting, 8, 186 weights, 26, 34, 100 attribute dispersion, 56 mean-based, see non-robust median-based, see robust non-robust, 57, 100 robust, 57, 100 attribute weights, 26, 100 for validation, 171 pairwise, 28, 99 zero-valued, 102 pre-specified, 101 cluster analysis, 2 hierarchical. 6 partitional, 4 cluster happy assumption, 28 cluster validation, 169 clustering L-groups, 150 algorithm, 4 enhanced soft subspace, see ESSC entropy weighted K-means, see EWKM hierarchical, 6 partitional, 4

sparse K-means, see SPARCL sparse alternate sum, see SAS subspace, 8, 153, 154 problem, 2 computational costs for COSA, 181 convergence for COSA, 182 COSA, 25 λ NN algorithm, 115 KNN algorithm, 44 KNN criterion, 33, 99 KNN_0 algorithm, 100 KNN_1 algorithm, 103 and missing data, 183 attribute weights, 34 criterion, 29, 67 distance, 13, 27, 28, 103, 186 pseudo algorithm, 34 robust criterion, 69 within-cluster distance, 27, 99 $COSA-\lambda NN, 110, 115$ algorithm, 110 complexity, 181 tuning, 160 COSA-KNN algorithm, 44 complexity, 181 tuning, 84, 160 $COSA-KNN_0$, 100 algorithm, 100 $COSA-KNN_1$, 103 algorithm, 103 tuning, 160

criterion, 29 COSA-KNN, 33 for cluster validation, 171 for COSA-KNN. 67 for MVPIN. 144 for robust COSA-KNN, 69 dendrogram, 14 cutting a, 132 dispersion, 57 dissimilarity, see distance distance, 10 attribute, 26 COSA, see COSA dual target, 188 Euclidean, 11 inverse exponential, 38, 99 majorizing, 25, 28 Manhattan, 11 Minkowski, 11 single target, 187 target, 187 within-cluster, 27, 99 enhanced soft subspace clustering, see ESSC entropy, 32 entropy weighted K-means, see EWKM ESSC, 154 Euclidean distance, 11 EWKM. 153 tuning, 160 folded-normal distribution, 61 Gap statistic procedure, 72 algorithm, 74 golden ratio, 159 golden section search, 159 hierarchical clustering, 6 high-dimensional data, 7 homotopy, 40 parameter, 40 relaxation path, 44

strategy, 43 information retrieval. 189 inverse exponential distance, 38, 99 K-means algorithm, 5 Karush-Kuhn-Tucker conditions, 121 Kullback-Leibler divergence, 40, 101 regularization, 101, 170 L-groups clustering, 150 Lance-Williams update formula, 6, 134 linkage, 7 average, 7 complete, 7 single, 7 Ward, 7 majorizing, 25 distance, 28 property, 28 Manhattan distance, 11 MDS, see multidimensional scaling Minkowski distance, 11 missing data for COSA, 183 multidimensional scaling, 15 classical, 16 configuration, 16 SMACOF, 16 **MVPIN**, 137 criterion, 144 nearest neighbors, 30, 110–114 λNN, 110–114 KNN. 30 shared. 137 similarity measure, 137 negative entropy, 32 omics data, 156 one standard error rule, 73 pairwise attribute weights, see attribute weights

PAM, 128 partitioning, 4 around medoids, see PAM points of view analysis, 188 pre-specified attribute weights, 101 prototype model, 13, 104 Rand Index, 142 rate parameter, 39 relaxation parameter, 44 robust COSA-KNN criterion, 69 SAS, 151 default grid search, 157, 159 golden section search, 157, 159 scale parameter, 35, 46 self-organizing maps, 189 shared nearest neighbors, 137 similarity between two clusters, 137similarity between two objects, 137silhouette, 130 SMACOF, 16 softmax function. 35 SPARCL, 153 tuning, 159

sparse K-means, see SPARCL sparse alternate sum clustering, see SAS STRAIN. 16 STRESS, 15 targeted distance, 187 dual, 188 single, 187 triangular inequality, 10 violation, 29 Type-I error, 87 validation attribute importance, 174 attribute weights, 171 clustering, 169 Ward linkage, 7, 134 minimum variance method, 134 weak spot model, 106 weights attribute, see attribute weights zero-valued attribute weights, 102 zigzag tendency, 69