Hierarchical clustering algorithms in data mining

SLINK redirects here. For the online magazine, see Slink. Statistical method of analysis, who seeks to build a hierarchy of clusters Part of the series ofMachine learninganddata mining problems Classification Reduce Regression Anomalies Detection AutoML Rules Stop Rules Strengthening Training Structured Forecasting Function Training Online Training Semi-Addressed Training Stationary Training to rank Gram Induction Learning (classification • regression) Solution trees ensembles Grip Enhancing Random Forest k-NN Linear Regression Naïve Bays Artificial Neural Networks Logistic Regression Perceptor Vector Vector Vector Machine (RVM) Support Vector Machine (SVM) Cluster Brilliant BRIV TREATMENT Hierarchical k-means Expected-maximization (EM) DBSCAN OPTICS Medium displacement Size reduction Factor reduction factor CCA ICA LDA NMF PCA PCA PGD T-SNE Structural forecast Graphical models Bayes net Conditional random field Hidden mark anomalies detection K-NN Local out-of-gabarist factor Artificial neural network Autocoded learning deepening 2000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 0 000 0 000 000 000 000 000 000 00 00 000 000 000 000 000 00 00 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 Occam 2000000000000000000000 The study of statistical theory of BV Machine Learning locations NeurIPS ICML ML JMLR ArXiv:cs. LG Dictionary of Artificial Intelligence Dictionary of Artificial Intelligence Related Articles List of datasets for machine learning Research Structure of machine learning in data mining and statistics , hierarchical grouping (also called hierarchical cluster analysis or HCA) is a cluster analysis method that seeks to build a hierarchy of clusters. Hierarchical grouping strategies usually fall into two types: [1] agglomerative: This is a bottom-up approach: each observation starts in its own cluster, and cluster pairs merge, with one moving up the hierarchy. Business: This is a top-down approach: all observations start in one cluster and divisions are recursive, with one moving down the hierarchy. In general, mergers and splits are determined in a greedy way. The results of the hierarchical cluster[2] are usually presented in a dendropgram. The standard hierarchical agglomerative cluster (HAC) algorithm has a time complexity of O (n 3) {\displaystyle {\mathcal {O}}(n^{3})} and requires Ω (n 2) {\displaystyle \(n^{2})}} memory, making it too slow for even average data. However, for some special cases, optimal effective agglomeration methods (of complexity O (n 2) {\displaystyle {\mathcal {O}}(n^{2})} are known: SLINK[3] for single binding and CLINK[4] for full cluster connectivity. (2 journals n) {\displaystyle {\mathcal {O}}(n^{2}\log n)} , O (n 3) limit improvement {\displaystyle {\mathcal {O}{3}}, In many cases, the memory above the head of this approach is too large to make it practically usable. O (2 n) {\displaystyle {\mathcal {O}}(2^{n})} , but it is customary to use faster heuristic methods to select tabs, such as k-means. Cluster non-similarity In order to decide which clusters should be combined (for agglomerations), or where a cluster should be divided (for dividing), a measure of dissities between observation groups is required. In most methods of hierarchical grouping, this is achieved by using an appropriate metric (distance measure between pairs of observations) and a connection criterion that defines the non-similarity of sets as a function of the pairs of observation distances in the sets. Indicator Additional information: Metric (mathematics) Choosing an appropriate indicator will affect the shape of clusters, as some elements may be relatively closer to each other under one indicator than another. For example, in two dimensions, below the metric of the Distance from Manhattan, the distance between origin (0,0) and (.5, .5) is the same as the distance between origin and (0, 1), while in euclid distance metric the latter is strictly greater. Some common indicators for hierarchical grouping are: [5] Names Formula Euclid distance ‖ − b ‖ 2 = ∑ i (i – b i) 2 {\displaystyle \|| b-|_{2}={\sqrt {\sum _{i}(a_{i}-b_{i})}{2} Square Distance ‖ 2 - b ‖ = ∑ (i - b i) 2 {\displaystyle \. |a-b-b+{2}^{2}=\sum_i}(a_{i}-b_{i}) ^ {2}} Distance from Manhattan ‖ − ‖ 1 = ∑ i | a i – b i | {\displaystyle\|| b\|| _{1}=\sum _{i}|a_{i}-b_{i}|} Maximum distance ‖ a − b ‖ ∞ = max i | a i – b i | {\displaystyle\|| b\|____{_infty}==max_max _{i}|a_{i}-b_{i}|} Distance from mahalanobis (a - b) ⊤ S − 1 (a − b) {\displaystyle {\sqrt {(a-b)^\top}S^{-1}(a-b)}}} where S is the insidious matrix for text or other non-numeric data, metrics such as the distance of Blackmings or the Distance of Levenchtine are often used. A review of cluster analysis in health psychology studies found that the most common distance measure in published studies in this area of study was euclid distance or square euclide distance. [Reference required] Connection criteria The connection criterion determines the distance between observation groups as a function of the distances between the two. Some commonly used criteria for linking between observation sets A and B are: [6] Names Formula Maximum or full-linkage clustering max { d (a , b): ∈ A , b ∈ B } . {\displaystyle \max\, \{\, d(a, b): a\ in A, \ in B\, \}}} Minimum or single connection clustering min { d (a , b) : ∈ A, b ∈ B } . {\displaystyle \min\, \{\, d(a, b):a\in A,\, b\in B\, \}}} Weighted Average Link (or UPGMA) 1 | A 2000 ─ | B | ∑ ∈ ∑ b ∈ B d (b) {\displaystyle {\frac {1}{| A|\cdot | B|}} \sum _{a\in A}\sum _{b\in B}d(a,b)} Weighted average connection (or WPGMA) d (i ∪ j , k) = d (i , k) + d (j , k) 2 . {\displaystyle d(i\cup j,k)={d(i,k)+d(j,k)}{2}}.} Centroid or UPGMC connection clusters ‖ c s − c t t t ‖ {\displaystyle \|| c_{s}-c_{t}\|}} where {\displaystyle c_{s}} and c t {\ displaystyle c_{t}} are the centrifugal groups of s and t, respectively. Minimum energy grouping 2 n m ∑ i , j = 1 n , m ‖ — b j ‖ 2 − 1 n 2 ∑ i , j = 1 n ‖ a — a i — a — a j ‖ 2 — 1 m 2 ∑ ‖ <7> i , j = 1 m ‖ b i - b j j ‖ 2 {\displaystyle {\frac {2}{nm}}\sum _{i,j=1}^{n,m}\|a_{i}-b_{j}|{ 2}-{\frac{1}{n^{2}}}}\sum _{i} j=1}^{n}|a_{i}-a_{j}|| _{2}-{\frac {1}{m^{2}}}}\sum _{i=1}^{m}|b_{i}-b_{j}|__{2}}, where d is the selected metric. Other connection criteria include: The sum of all dispersions within the cluster. The increase in dispersion for the cluster that merges (area criterion). Candidates are more likely to load with the same distribution function (V-linkage). The product from a degree of and outside degrees of k-near-adjacent graph (graph link). [9] The growth of some cluster descriptor (i.e. quantity determined to measure cluster quality) after the merger of two clusters. [10] [11] [12] The discussion has the clear advantage that any valid distance measure can be used. In fact, the observations themselves are not necessary: all that is used is a matrix of distances. Agglomeration grouping, for example Raw data, for example, assume that this data will be grouped, and the Euclide distance is an indicator of distance. Hierarchical clustering dendrogram will be like this: Traditional representation Of the tree at a given height will give splitting clustering at selected accuracy. In this example, the cutting after the second row (top) of dendrogram will receive clusters {a} {b c} {d e} {f}. Cutting after the third row will receive clusters {a} {b c} {d e f}, which is a rougher cluster, with a smaller number but larger clusters. This method builds the hierarchy of individual elements by gradually merging clusters. In our example, we have six items {a} {b} {c} {d} {e} and {f}. The first step is to determine which items to merge into a cluster. Usually we want to take the two closest elements, according to the selected distance. Optionally, a distance matrix can also be built stage where the number in the 1th row j-column is the distance between the ith and jth elements. Then, when clustering is performed, rows and columns are merged clusters are merged and distances updated. This is a common way to apply this type of clustering and benefits from caching distances between clusters. A simple agglomerative cluster algorithm is described in a single-linkage clusters page; can be easily adapted to different types of connections (see below). Supposing we have merged the two closest items b and c, now we have the following groups {a}, {b, c}, {d}, {e} and {f}, and you want to merge them further. To do this, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Typically, the distance between two clusters A {\displaystyle {\mathcal {A}} and B {\displaystyle {\mathcal {B}} is one of the following: The maximum distance between the elements of each cluster (called full connection of connections): max { d (x , y) : x ∈ A, y ∈ B } . {\displaystyle\max\{,d(x,y):x\in {\},\, y\ in {\} mathematical {B}}}}}}}} The minimum distance between the elements of each cluster (also called a single cluster connection): min { d (x , y) : x ∈ A, y ∈ B } . {\displaystyle \min\{, d(x,y):x\in {\} in {\} mathematics {A}},\,y\ in {\mathcal {B}}}}}}}} The average distance between the elements of each cluster (also called the average connection of connections used, for example, in UPGMA): 1 | A 2000 ─ | B | ∑ x ∈ ∑ ∈ B d (x , y) . {\displaystyle {1\over {|{ *mathematical {A}}|\cdot | {\mathcal {B}}|}} \sum _{x\in {\mathcal {A}}}\sum _{y\in {\mathcal {B}}}d(x,y)}} The sum of all intra-cluster dispersion. Increasing the dispersion for the cluster merges (area method[8]) Probability candidates clusters throw from the same distribution function (V-linkage). In the case of connected minimum distances, the pair is selected arbitrarily, thus being able to generate several structurally different dendrograms. Alternatively, all connected pairs can be joined simultaneously, generating a unique dendrogramgram. [13] Grouping can always be stopped when there is a sufficiently small number of clusters (numbering criterion). Some connections can ensure that agglomeration occurs at a greater distance between clusters than the previous agglomeration, and then can stop grouping when the clusters are too remote to be merged (distance). However, this is not the case with the centrifugal connection in which so-called reversals can occur[14] (inversions, deviations from ultrametry). Cluster splitting The basic principle of a dividing cluster is published as the DIANA algorithm (DIvisive ANAlysis clustering). [15] Initially, all data is in the same cluster, and the largest cluster is split until each object is separated. As O (2 n) exists, O(2^{n})} partitioning methods Heuristics. DIANA selects the object with the maximum average dissexuality and then moves all the objects in this cluster that are more similar to the new cluster than the others. Open source application software Hierarchical clustering dendrogram of Iris dataset (using R). SOURCE Hierarchical clustering and interactive visualization of dendrogram in Orange data mining suite. ALGLIB performs several hierarchical clustering algorithms (single connection, full-connection, region) in C++ and C# with O (n^2) memory and O(n3) run time. ERC includes multiple hierarchical clustering algorithms, different connectivity strategies, as well as effective SLINK, [4] CLINK algorithms and Anderberg algorithms, flexible cluster extraction from dendrograms and various other cluster analysis algorithms. Octaves, the analog GNU to MATLAB performs the hierarchical grouping in the connection function. Orange, a data retrieval software package, includes a hierarchical grouping with an interactive dendrogram preview. R has many packages that provide hierarchical clustering features. SciPy performs hierarchical grouping in Python, including the effective SLINK algorithm. scikit-learn also performs the hierarchical grouping in Python. Weka includes hierarchical cluster analysis. MATLAB's commercial implementations include hierarchical cluster analysis. Sas includes hierarchical cluster analysis in proc cluster. Mathematics includes a hierarchical cluster package. NCSS includes hierarchical cluster analysis. SPSS includes hierarchical cluster analysis. Qlucore Omics Explorer includes hierarchical cluster analysis. Stata includes hierarchical cluster analysis. CrimeStat includes the closest neighbor hierarchical cluster algorithm with graphical output for Geographic Information System. See also Binary Space Partitioning Boundary Volume Hierarchy Brown Clusters Clusters Cluster Analysis Computer Philology CURE Data Clustering Algorithm Dasgupta Target Dendrogram Determining the Number of Clusters in Data Set Hierarchical Cluster of Networks Locally Sensitive Hashing Adjacent Search Closest Neighbor Chain Algorithm No. Taxonomy Taxonomy Algorithm Stagnation Manuals Fees ^ Rokach, Lior, and Oded Maimon. Methods of grouping. A guide to data mining and knowledge discovery. 2005 321-352. Hierarchical grouping. Introduction to CPI with MPI for Data Science. Springer. ^ Our T. 1 (1973). SLINK: optimally effective algorithm for the single connection cluster method (PDF). Computer log. The British Computer Society. 16 (1): 30–34. Doy:10.1093/comjnl/16.1.30. ^ Atari Todd (1977). Effective algorithm for a full connection method. Computer log. The British Computer Society. 20 (4): 364–366. Doy:10.1093/comjnl/20.4.364. 2014 procedure: proximity measures. SAS/STAT 9.2 User Manual. Institute of the SAS. Retrieved 2009-04-26. At the end of 2015, sas/STAT 9.2 User Manual. Institute of the SAS. Retrieved 2009-04-26. Skelly, G. Y., 1999 Riso, M. L. (2005). Hierarchical grouping by joint interdisplation: expanding Ward's minimum dispersion method. Taxonomy Journal. 22 (2): 151–183. Doy:10.1007/s00357-005-0012-9. Our Earl, 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. 58 (301): 236–244. Doy:10.2307/2282967. 2282967. 0148188. They are 100000000000000000 Wang, Xiaogang; Tao, Deli; Tang, Xiao (2012). Fitzgibbon, Andrew; Lazebnik, Svetlana; Perona, Pietro; Syto, Yoichi; Schmidt, Cordelia (eds.). Graph degree Linkage: agglomerative clustering of directional graphics. Computer vision – ECCV 2012. Lecture notes in computer science. Springer Berlin Heidelberg. 7572: 428–441. 1208.5092. 2012 Xiv1208.5092Z. Doy:10.1007/978-3-642-33718-5_31. 9783642337185. See also: ^ Zhang, et al. Agglomerative grouping by maximum incremental path integral. recognition (2013). They are 1000000000000000000000000 Cyclists, through the zetas of the graph. Advances in neural information processing systems. 2008. ^ Ma, et al. Segmentation of multivariant mixed data by encoding data with loss and compression. IEEE Transactions for Model Analysis and Machine Intelligence, 29(9) (2007): 1546-1562. Fernandez, Alberto 2008. Solve non-uniqueness in agglomerative hierarchical clustering using multidendrograms. Taxonomy Journal. 25 (1): 43–65. 0608049. Doi:10.1007/s00357-008-9004-x. 2014 Legends, L. (2003). Numerical ecology. BV Science Enters. ^1000000000000000000000000000000000000000000000000000000000 Find data groups - Introduction to cluster analysis. Science for John Wylie and Wiley's sons. Further reading Kaufman, L.; 1990 Find Data Groups: Introduction to Cluster Analysis (1 ed.). New York: John Wylie. 0-471-87876-6. Hastie, Trevor; Taboo, Robert; Friedman, Jérôme (2009). 14.3.12 Hierarchical grouping. Elements of statistical training (2nd edition). New York: Springer. 520–528. 1000000000000000000000000000000000000000000000000000000000000000000000000 Archive of original (PDF) of 2009-11-10. Retrieved 2009-10-20. Visited by