Hierarchical clustering

From Wikipedia, the free encyclopedia

In data mining, **hierarchical clustering** is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

In the general case, the complexity of agglomerative clustering is $\mathcal{O}(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $\mathcal{O}(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK[1] for single-linkage and CLINK[2] for complete-linkage clustering.

Contents

Cluster dissimilarity

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Metric

Further information: metric (mathematics)

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (1,0) and the origin (0,0) is always 1 according to the usual norms, but the distance between the point (1,1) and the origin (0,0) can be 2, $\sqrt{2}$ or 1 under Manhattan distance, Euclidean distance or maximum distance respectively.

Some commonly used metrics for hierarchical clustering are:[3]

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\|a\|\|b\|}$

For text or other non-numeric data, metrics such as the Hamming distance or Levenshtein distance are often used.

A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance. [*citation needed*]

Linkage criteria

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.

Some commonly used linkage criteria between two sets of observations *A* and *B* are:[4][5]

Names	Formula
Maximum or complete linkage clustering	$\max \{ d(a,b) : a \in A,\, b \in B \}.$
Minimum or single-linkage clustering	$\min \{ d(a,b) : a \in A,\, b \in B \}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b).$

Minimum energy clustering	$\frac{2}{nm}\sum_{i,j=1}^{n,m}\|a_i-b_j\|_2-\frac{1}{n^2}\sum_{i,j=1}^{n}\|a_i-a_j\|_2-\frac{1}{m^2}\sum_{i,j=1}^{m}\|b_i-b_j\|_2$

where *d* is the chosen metric. Other linkage criteria include:

- The sum of all intra-cluster variance.
- The decrease in variance for the cluster being merged (Ward's criterion).[6]
- The probability that candidate clusters spawn from the same distribution function (V-linkage).

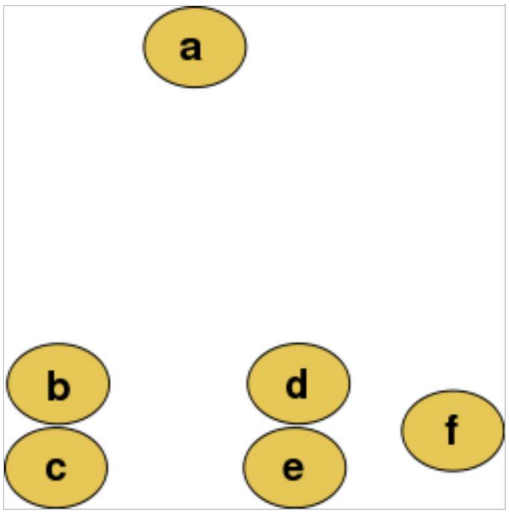
Discussion

Hierarchical clustering has the distinct advantage that any valid measure of distance can be used. In fact, the observations themselves are not required: all that is used is a matrix of distances.

Example for Agglomerative Clustering

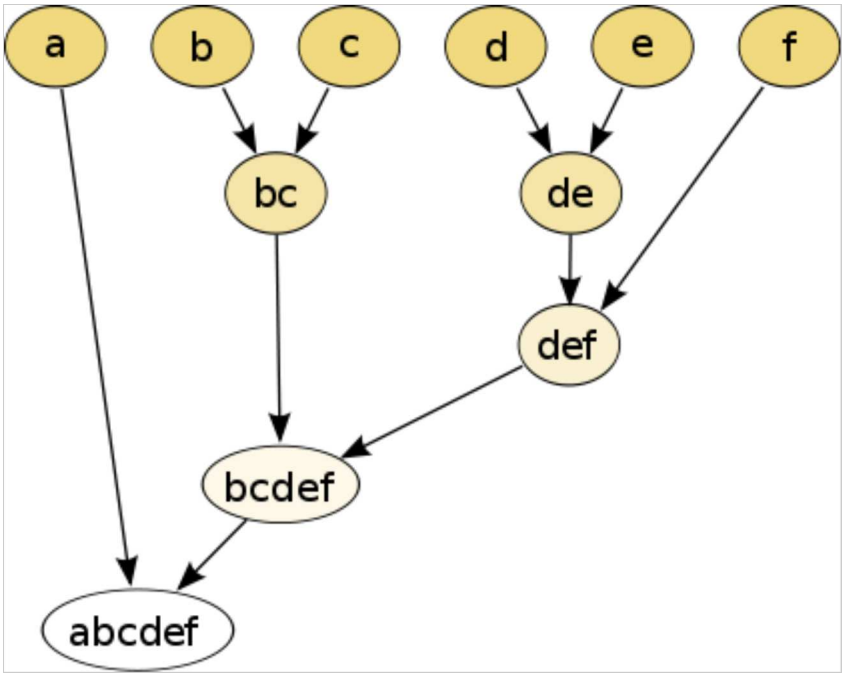
For example, suppose this data is to be clustered, and the Euclidean distance is the distance metric.

Cutting the tree at a given height will give a partitioning clustering at a selected precision. In this example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number of larger clusters.



Raw data

The hierarchical clustering dendrogram would be as such:



Traditional representation

This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements {a} {b} {c} {d} {e} and {f}. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance.

Optionally, one can also construct a distance matrix at this stage, where the number in the *i*-th row *j*-th column is the distance between the *i*-th and *j*-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage (see below).

Suppose we have merged the two closest elements *b* and *c*, we now have the following clusters {*a*}, {*b*, *c*}, {*d*}, {*e*} and {*f*}, and want to merge them further. To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters $\mathcal{A}$ and $\mathcal{B}$ is one of the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):

$$\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$$

- The minimum distance between elements of each cluster (also called single-linkage clustering):

$$\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}.$$

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$$

- The sum of all intra-cluster variance.

- The increase in variance for the cluster being merged (Ward's method[6])
- The probability that candidate clusters spawn from the same distribution function (V-linkage).

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Software

Free

- R has several functions for hierarchical clustering: see CRAN Task View: Cluster Analysis & Finite Mixture Models (http://cran.r-project.org/web/views/Cluster.html) for more information.
- Orange, a free data mining software suite, module orngClustering (http://www.ailab.si/orange/doc/modules/orngClustering.htm) for scripting in Python, or cluster analysis through visual programming (http://www.ailab.si/orange/screenshots.psp) .
- hcluster (http://code.google.com/p/scipy-cluster/) is Python software, based on NumPy, which supports hierarchical clustering and plotting.
- Cluster 3.0 provides a nice Graphical User Interface to access to different clustering routines and is available for Windows, Mac OS X, Linux, Unix. See: [1] (http://bonsai.hgc.jp/~mdehoon/software/cluster/)
- ELKI includes multiple hierarchical clustering algorithms.
- figue (http://code.google.com/p/figue/) A JavaScript package that implements some agglomerative clustering functions (single-linkage, complete-linkage, average-linkage) and functions to visualize clustering output (e.g. dendograms) (Online demo (http://web.science.mq.edu.au/~jydelort/figue/demo.html)).
- MultiDendrograms (http://deim.urv.cat/~sgomez/multidendrograms.php) An open source Java application for variable-group agglomerative hierarchical clustering,[7] with graphical user interface.
- CrimeStat implements two hierarchical clustering routines, a nearest neighbor (Nnh) and a risk-adjusted(Rnnh).
- Complete C# DEMO (http://www.semanticsearchart.com/researchHAC.html) implemented as visual studio project that includes real text files processing, building of document-term matrix with stop words filtering and stemming. Same site offers comparison with other algorithms.
- HAC C# (http://www.snip-me.de/hierarchical-agglomerative-clustering-c-sharp.aspx) is an implementation of an agglomerative clustering algorithm using single-linkage.

Commercial

- Software for analyzing multivariate data with instant response using Hierarchical clustering (http://www.qlucore.com)
- SAS CLUSTER (http://support.sas.com/documentation/cdl/en/statugclustering/61759/PDF/default/statugclustering.pdf)

See also

- Cluster analysis
- CURE data clustering algorithm
- Dendrogram
- Determining the number of clusters in a data set

- Hierarchical clustering of networks
- Nearest-neighbor chain algorithm
- Numerical taxonomy
- OPTICS algorithm

Notes

1. ^ R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" (http://www.cs.gsu.edu/~wkim/index_files/papers/sibson.pdf) . *The Computer Journal* (British Computer Society) **16** (1): 30–34. http://www.cs.gsu.edu/~wkim/index_files/papers/sibson.pdf.
2. ^ D. Defays (1977). "An efficient algorithm for a complete link method" (http://comjnl.oxfordjournals.org/content/20/4/364.abstract) . *The Computer Journal* (British Computer Society) **20** (4): 364–366. http://comjnl.oxfordjournals.org/content/20/4/364.abstract.
3. ^ "The DISTANCE Procedure: Proximity Measures" (http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_distance_sect016.htm) . *SAS/STAT 9.2 Users Guide*. SAS Institute. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_distance_sect016.htm. Retrieved 2009-04-26.
4. ^ "The CLUSTER Procedure: Clustering Methods" (http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_cluster_sect012.htm) . *SAS/STAT 9.2 Users Guide*. SAS Institute. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_cluster_sect012.htm. Retrieved 2009-04-26.
5. ^ Székely, G. J. and Rizzo, M. L. (2005) Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method, Journal of Classification 22, 151-183.
6. ^ ***a b*** Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* **58** (301): 236–244. doi:10.2307/2282967 (http://dx.doi.org/10.2307%2F2282967) . JSTOR 2282967 (http://www.jstor.org/stable/2282967) . MR 0148188 (http://www.ams.org/mathscinet-getitem?mr=0148188) .
7. ^ Fernández, Alberto; Gómez, Sergio (2008). "Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms" (http://www.springerlink.com/content/c8795u6232184423/) . *Journal of Classification* **25**: 43–65. doi:10.1007/s00357-008-9004-x (http://dx.doi.org/10.1007%2Fs00357-008-9004-x) . http://www.springerlink.com/content/c8795u6232184423/.

References and further reading

- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering" (http://www-stat.stanford.edu/~tibs/ElemStatLearn/) (PDF). *The Elements of Statistical Learning* (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6. http://www-stat.stanford.edu/~tibs/ElemStatLearn/. Retrieved 2009-10-20.
- Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 16.4. Hierarchical Clustering by Phylogenetic Trees" (http://apps.nrbook.com/empanel/index.html#pg=868) . *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8. http://apps.nrbook.com/empanel/index.html#pg=868.