hclust {stats} R Documentation

Hierarchical Clustering

Description

Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.

Usage

```
hclust(d, method = "complete", members = NULL)

## S3 method for class 'hclust'
plot(x, labels = NULL, hang = 0.1, check = TRUE,
     axes = TRUE, frame.plot = FALSE, ann = TRUE,
     main = "Cluster Dendrogram",
     sub = NULL, xlab = NULL, ylab = "Height", ...)
```

Arguments

d

a dissimilarity structure as produced by dist.

method

the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC).

members

NULL or a vector with length size of d. See the ‘Details’ section.

x

an object of the type produced by hclust.

hang

The fraction of the plot height by which labels should hang below the rest of the plot. A negative value will cause the labels to hang down from 0.

check

logical indicating if the x object should be checked for validity. This check is not necessary when x is known to be valid such as when it is the direct result of hclust(). The default is check=TRUE, as invalid inputs may crash R due to memory violation in the internal C plotting code.

labels

A character vector of labels for the leaves of the tree. By default the row names or row numbers of the original data are used. If labels = FALSE no labels at all are plotted.

axes, frame.plot, ann

logical flags as in plot.default.

main, sub, xlab, ylab

character strings for title. sub and xlab have a non-NULL default when there's a tree$call.

...

Further graphical arguments. E.g., cex controls the size of the labels (if plotted) in the same way as text.

Details

This function performs a hierarchical cluster analysis using a set of dissimilarities for the n objects being clustered. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance–Williams dissimilarity update formula according to the particular clustering method being used.

A number of different clustering methods are provided. Ward's minimum variance method aims at finding compact, spherical clusters. The complete linkage method finds similar clusters. The single linkage method (which is closely related to the minimal spanning tree) adopts a ‘friends of friends’ clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods. Note however, that methods "median" and "centroid" are not leading to a monotone distance measure, or equivalently the resulting dendrograms can have so called inversions or reversals which are hard to interpret, but note the trichotomies in Legendre and Legendre (2012).

Two different algorithms are found in the literature for Ward clustering. The one used by option "ward.D" (equivalent to the only Ward option "ward" in R versions <= 3.0.3) does not implement Ward's (1963) clustering criterion, whereas option "ward.D2" implements that criterion (Murtagh and Legendre 2014). With the latter, the dissimilarities are squared before cluster updating. Note that agnes(*, method="ward") corresponds to hclust(*, "ward.D2").

If members != NULL, then d is taken to be a dissimilarity matrix between clusters instead of dissimilarities between singletons and members gives the number of observations per cluster. This way the hierarchical cluster algorithm can be ‘started in the middle of the dendrogram’, e.g., in order to reconstruct the part of the tree above a cut (see examples). Dissimilarities between clusters can be efficiently computed (i.e., without hclust itself) only for a limited number of distance/linkage combinations, the simplest one being squared Euclidean distance and centroid linkage. In this case the dissimilarities between the clusters are the squared Euclidean distances between cluster means.

In hierarchical cluster displays, a decision is needed at each merge to specify which subtree should go on the left and which on the right. Since, for n observations there are n-1 merges, there are 2^{(n-1)} possible orderings for the leaves in a cluster tree, or dendrogram. The algorithm used in hclust is to order the subtree so that the tighter cluster is on the left (the last, i.e., most recent, merge of the left

subtree is at a lower value than the last merge of the right subtree). Single observations are the tightest clusters possible, and merges involving two observations place them in order by their observation sequence number.

Value

An object of class hclust which describes the tree produced by the clustering process. The object is a list with components:

merge
an n-1 by 2 matrix. Row i of merge describes the merging of clusters at step i of the clustering. If an element j in the row is negative, then observation -j was merged at this stage. If j is positive then the merge was with the cluster formed at the (earlier) stage j of the algorithm. Thus negative entries in merge indicate agglomerations of singletons, and positive entries indicate agglomerations of non-singletons.
height
a set of n-1 real values (non-decreasing for ultrametric trees). The clustering height: that is, the value of the criterion associated with the clustering method for the particular agglomeration.
order
a vector giving the permutation of the original observations suitable for plotting, in the sense that a cluster plot using this ordering and matrix merge will not have crossings of the branches.
labels
labels for each of the objects being clustered.
call
the call which produced the result.
method
the cluster method that has been used.
dist.method
the distance that has been used to create d (only returned if the distance object has a "method" attribute).
There are print, plot and identify (see identify.hclust) methods and the rect.hclust() function for hclust objects.

Note

Method "centroid" is typically meant to be used with squared Euclidean distances.

Author(s)

The hclust function is based on Fortran code contributed to STATLIB by F. Murtagh.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. (S version.)

Everitt, B. (1974). Cluster Analysis. London: Heinemann Educ. Books.

Hartigan, J.A. (1975). Clustering Algorithms. New York: Wiley.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. San Francisco: Freeman.

Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press: New York.

Gordon, A. D. (1999). Classification. Second Edition. London: Chapman and Hall / CRC

Murtagh, F. (1985). “Multidimensional Clustering Algorithms”, in COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used).

McQuitty, L.L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. Educational and Psychological Measurement, 26, 825–831.

Legendre, P. and L. Legendre (2012). Numerical Ecology, 3rd English ed. Amsterdam: Elsevier Science BV.

Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification 31 (forthcoming).

See Also

identify.hclust, rect.hclust, cutree, dendrogram, kmeans.

For the Lance–Williams formula and methods that apply it generally, see agnes from package cluster.

Examples

```
require(graphics)

### Example 1: Violent crime rates by US state

hc <- hclust(dist(USArrests), "ave")
plot(hc)
plot(hc, hang = -1)

## Do the same with centroid clustering and *squared* Euclidean distance,
## cut the tree into ten clusters and reconstruct the upper part of the
## tree from the cluster centers.
hc <- hclust(dist(USArrests)^2, "cen")
memb <- cutree(hc, k = 10)
cent <- NULL
for(k in 1:10){
  cent <- rbind(cent, colMeans(USArrests[memb == k, , drop = FALSE]))
}
hc1 <- hclust(dist(cent)^2, method = "cen", members = table(memb))
opar <- par(mfrow = c(1, 2))
plot(hc,    labels = FALSE, hang = -1, main = "Original Tree")
plot(hc1, labels = FALSE, hang = -1, main = "Re-start from 10 clusters")
```

```
par(opar)

### Example 2: Straight-line distances among 10 US cities
##    Compare the results of algorithms "ward.D" and "ward.D2"

data(UScitiesD)

mds2 <- -cmdscale(UScitiesD)
plot(mds2, type="n", axes=FALSE, ann=FALSE)
text(mds2, labels=rownames(mds2), xpd = NA)

hcity.D    <- hclust(UScitiesD, "ward.D") # "wrong"
hcity.D2 <- hclust(UScitiesD, "ward.D2")
opar <- par(mfrow = c(1, 2))
plot(hcity.D,    hang=-1)
plot(hcity.D2, hang=-1)
par(opar)
```