Package 'flashClust'

Version 1.01-2

Date 2012-08-21

Title Implementation of optimal hierarchical clustering

Author code by Fionn Murtagh and R development team, modifications and packaging by Peter Langfelder

Maintainer Peter Langfelder <Peter.Langfelder@gmail.com>

Depends R (>= 2.3.0)

ZipData no

License GPL (>= 2)

Description Fast implementation of hierarchical clustering

Repository CRAN

Date/Publication 2012-08-21 18:23:43

NeedsCompilation yes

R topics documented:

---

flashClust   *Faster alternative to hclust*

Description

This function implements optimal hierarchical clustering with the same interface as hclust.

Usage

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

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"`, `"single"`, `"complete"`, `"average"`, `"mcquitty"`, `"median"` or `"centroid"`.
members	NULL or a vector with length size of d. See the 'Details' section.

Details

See the description of [hclust](#) for details on available clustering methods.

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.

flashClust is a wrapper for compatibility with older code.

Value

Returned value is the same as that of [hclust](#): 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 $n - 1$ non-decreasing real values. 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).

Author(s)

Fionn Murtagh, adapted and packaged by Peter Langfelder

References

This implementation is mentioned in

Peter Langfelder, Steve Horvath (2012) Fast R Functions for Robust Correlations and Hierarchical Clustering. Journal of Statistical Software, 46(11), 1-17. http://www.jstatsoft.org/v46/i11/

F.Murtagh's software web site: http://www.classification-society.org/csna/mda-sw/ , section 6

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.

See Also

hclust

Examples

```
# generate some data to cluster
set.seed(1);
nNodes = 2000;

# Random "distance" matrix
dst = matrix(runif(n = nNodes^2, min = 0, max = 1), nNodes, nNodes);

# Time the flashClust clustering
system.time( {
  h1 = hclust(as.dist(dst), method= "average");
  } );

# Time the standard R clustering
system.time( {
  h2 = stats::hclust(as.dist(dst), method = "average");
  } );

all.equal(h1, h2)
# What is different:

h1[[6]]
h2[[6]]
```

```
# Everything but the 'call' component is the same; in particular, the trees are exactly equal.
```

Index