Package 'mdendro'

December 6, 2018

Version 1.0.1

Date 2018-12-06

Title Variable-Group Methods for Agglomerative Hierarchical Clustering

Description A collection of methods for agglomerative hierarchical clustering strategies on a matrix of distances, implemented using the variable-group approach introduced in Fernandez and Gomez (2008) <doi:10.1007/s00357-008-9004-x>. Descriptive measures to analyze the resulting hierarchical trees are also provided. In addition to the usual clustering methods, two parameterized methods are provided to explore an infinite family of hierarchical clustering strategies. When there are ties in proximity values, the hierarchical trees obtained are unique and independent of the order of the elements in the input matrix.

Depends R (>= 3.5.0)

Imports rJava (>= 0.9.8)

SystemRequirements Java (>= 6)

License LGPL-2.1

Encoding UTF-8

RoxygenNote 6.1.1

NeedsCompilation no

Author Alberto Fernandez [aut, cre] (<https://orcid.org/0000-0002-1241-1646>), Sergio Gomez [aut] (<https://orcid.org/0000-0003-1820-0062>)

Maintainer Alberto Fernandez <alberto.fernandez@urv.cat>

Repository CRAN

Date/Publication 2018-12-06 15:20:08 UTC

R topics documented:

dendesc *Dendrogram Descriptive Measures*

Description

Descriptive measures for analyzing objects of class "dendrogram".

Usage

```
ntb(dendro)

ultrametric(dendro)

mae(prox, ultr)

sdr(prox, ultr)
```

Arguments

dendro	Object of class "dendrogram" as produced by linkage() or by as.dendrogram() applied to the hierarchical trees returned by hclust() and agnes().
prox	Object of class "dist" containing the proximity data used to build the dendrogram.
ultr	Object of class "dist" containing the ultrametric distances in the dendrogram, sorted in the same order as the proximity data in prox.

Details

This package allows the calculation of several descriptive measures for dendrograms, such as normalized tree balance, cophenetic correlation coefficient, normalized mean absolute error, and space distortion ratio.

For each node in a dendrogram, its entropy is calculated using the concept of Shannon's entropy, which gives a maximum entropy of 1 to nodes merging subdendrograms with the same number of leaves. The average entropy for all nodes in a dendrogram is called its tree balance. Normalized tree balance is computed by the ntb() function as the ratio between the tree balance of a dendrogram and the minimum tree balance of any dendrogram with the same number of elements. Perfectly balanced dendrograms have a normalized tree balance equal to 1, while binary dendrograms formed chaining one new element at a time have a normalized tree balance equal to 0.

To calculate the cophenetic correlation coefficient, the cor() function in the **stats** package needs that the matrix of ultrametric distances (also known as cophenetic distances) and the matrix of proximity data used to build the corresponding dendrogram, they both have their rows and columns sorted in the same order. When the cophenetic() function is used with objects of class "hclust", it returns ultrametric matrices sorted in appropriate order. However, when the cophenetic() function is used with objects of class "dendrogram", it returns ultrametric matrices sorted in the order of dendrogram leaves. The ultrametric() function in this package returns ultrametric matrices in appropriate order to calculate the cophenetic correlation coefficient using the cor() function.

The space distortion ratio of a dendrogram is computed by the sdr() function as the difference between the maximum and minimum ultrametric distances, divided by the difference between the maximum and minimum original distances used to build the dendrogram. Space dilation occurs when the space distortion ratio is greater than 1.

Functions

- ntb: Returns a number between 0 and 1 representing the normalized tree balance of the input dendrogram.
- ultrametric: Returns an object of class "dist" containing the ultrametric distance matrix sorted in the same order as the proximity matrix used to build the corresponding dendrogram.
- mae: Returns the normalized mean absolute error.
- sdr: Returns the space distortion ratio.

See Also

linkage() in this package, hclust() in the **stats** package, and agnes() in the **cluster** package for building hierarchical trees.

Examples

```
## distances between 21 cities in Europe
data(eurodist)

## comparison of dendrograms in terms of the following descriptive mesures:
## - normalized tree balance
## - cophenetic correlation coefficient
## - normalized mean absolute error
## - space distortion ratio

## single linkage (call to the mdendro package)
dendro1 <- linkage(eurodist, method="single")
ntb(dendro1)          # 0.2500664
ultr1 <- ultrametric(dendro1)
cor(eurodist, ultr1)  # 0.7842797
mae(eurodist, ultr1)  # 0.6352011
sdr(eurodist, ultr1)  # 0.150663

## complete linkage (call to the stats package)
dendro2 <- as.dendrogram(hclust(eurodist, method="complete"))
ntb(dendro2)          # 0.8112646
ultr2 <- ultrametric(dendro2)
cor(eurodist, ultr2)  # 0.735041
mae(eurodist, ultr2)  # 0.8469728
sdr(eurodist, ultr2)  # 1

## unweighted arithmetic linkage (UPGMA)
dendro3 <- linkage(eurodist, method="arithmetic", weighted=FALSE)
ntb(dendro3)          # 0.802202
ultr3 <- ultrametric(dendro3)
cor(eurodist, ultr3)  # 0.7279432
```

```
mae(eurodist, ultr3)  # 0.294578
sdr(eurodist, ultr3)  # 0.5066903

## unweighted geometric linkage
dendro4 <- linkage(eurodist, method="geometric", weighted=FALSE)
ntb(dendro4)          # 0.7531278
ultr4 <- ultrametric(dendro4)
cor(eurodist, ultr4)  # 0.7419569
mae(eurodist, ultr4)  # 0.2891692
sdr(eurodist, ultr4)  # 0.4548112
```

`linkage` *Linkage Methods for Hierarchical Clustering*

Description

Agglomerative hierarchical clustering of a matrix of dissimilarities.

Usage

```
linkage(prox, method = "arithmetic", weighted = FALSE,
  par.method = 0, digits = NULL)
```

Arguments

`prox`	Object of class `"dist"` containing the lower triangle of a proximity matrix in the form of distances.
`method`	Character string specifying the linkage method to be used. This should be one of: `"versatile"`, `"single"`, `"complete"`, `"arithmetic"` (default), `"geometric"`, `"harmonic"`, `"ward"`, `"centroid"` or `"flexible"`. See the *Details* section.
`weighted`	Logical to choose between the weighted and the unweighted (default) versions of some clustering methods. Weighted clustering gives merging branches in a hierarchical tree equal weight regardless of the number of individuals carried on each branch. Such a procedure weights the individuals unequally, contrasting with unweighted clustering that gives equal weight to each individual in the clusters. This parameter has no effect on the `"single"`, `"complete"` and `"ward"` linkages.
`par.method`	A real value in the range [-1, 1] required as parameter for the methods `"versatile"` and `"flexible"`. See the *Details* section.
`digits`	Integer specifying the precision, i.e. the number of significant decimal digits of the data and for the calculations. This is a very important parameter, since equal proximity values at a certain precision may become different by increasing its value. Thus, it may be responsible of the existence of tied distances. The rule should be not to use a precision larger than the resolution given by the experimental setup that has generated the data. If `digits=NULL` (default), then the precision is set to that of the data value with the largest number of significant decimal digits.

Details

Starting from a matrix of dissimilarities, `linkage()` calculates its dendrogram with the most commonly used agglomerative hierarchical clustering methods, e.g. single linkage, complete linkage, arithmetic linkage (also known as average linkage) and Ward's method. You can also choose between the weighted and the unweighted versions of some clustering methods, e.g. weighted centroid (WPGMC) and unweighted centroid (UPGMC). Importantly, it contains a new parameterized method named versatile linkage, which includes single linkage, complete linkage and average linkage as particular cases, and which naturally defines two new methods, geometric linkage and harmonic linkage (hence the convenience to rename average linkage as arithmetic linkage, to emphasize the existence of different types of averages).

The difference between the available hierarchical clustering methods rests in the way the distance between clusters is defined. During the agglomeration process, the data items are iteratively joined to form clusters, merging first the clusters that are at the minimum distance. However, given two clusters, each one formed by several data observations, there exist many ways of defining the distance between the clusters from the dissimilarities between their constituent individuals. Among these linkage methods, we have the following ones:

- `"single"`: the distance between clusters equals the minimum distance between individuals.
- `"complete"`: the distance between clusters equals the maximum distance between individuals.
- `"arithmetic"`: the distance between clusters equals the arithmetic mean distance between individuals. Also known as average linkage, WPGMA (weighted version) or UPGMA (unweighted version).
- `"geometric"`: the distance between clusters equals the geometric mean distance between individuals.
- `"harmonic"`: the distance between clusters equals the harmonic mean distance between individuals.
- `"versatile"`: the distance between clusters equals the generalized power mean distance between individuals. It depends on the value of `par.method`, with the following linkage methods as particular cases: `"complete"` (`par.method=+1`), `"arithmetic"` (`par.method=+0.1`), `"geometric"` (`par.method=0`), `"harmonic"` (`par.method=-0.1`) and `"single"` (`par.method=-1`).
- `"ward"`: the distance between clusters is a weighted squared Euclidean distance between the centroids of each cluster (Ward, 1963).
- `"centroid"`: the distance between clusters equals the square of the Euclidean distance between the centroids of each cluster. Also known as WPGMC (weighted version) or UPGMC (unweighted version).
- `"flexible"`: the distance between clusters is a weighted sum of the distances between clusters in the previous iteration (Lance and Williams, 1967; Belbin *et al.*, 1992). It depends on the value of `par.method`, and it is equivalent to `"arithmetic"` linkage when `par.method=0`.

Except for the cases containing ties in proximity values as described in the next paragraph, the following equivalences hold between the `linkage()` function in this package, the `hclust()` function in the **stats** package, and the `agnes()` function in the **cluster** package. When relevant, weighted (`W`) or unweighted (`U`) versions of the linkage methods and the values for `par.method` (β) are indicated:

linkage	hclust	agnes

"single"	"single"	"single"
"complete"	"complete"	"complete"
"arithmetic", U	"average"	"average"
"arithmetic", W	"mcquitty"	"weighted"
"ward"	"ward.D2"	"ward"
"centroid", U	"centroid"	--------
"centroid", W	"median"	--------
"flexible", U, β	--------	"gaverage", β
"flexible", W, β	--------	"flexible", $(1-\beta)/2$

linkage() implements the variable-group approach introduced in Fernandez and Gomez (2008) to solve the non-uniqueness problem found in the pair-group implementations. This problem arises when two or more minimum distances between different clusters are equal during the amalgamation process. The pair-group approach consists in choosing a pair, breaking the ties between distances, and proceeds in the same way until the final hierarchical classification is obtained. However, different dendrograms are possible depending on the criterion used to break the ties (usually a pair is just chosen at random). The variable-group approach groups more than two clusters at the same time when ties occur, what always produces a uniquely determined solution. When there are no ties, the variable-group approach gives the same results as the pair-group one.

Value

Returns an object of class "dendrogram".

References

L. Belbin, D.P. Faith and G.W. Milligan (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research*, 27(3):417-433.

A. Fernández and S. Gómez (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, 25(1):43-65.

G.N. Lance and W.T. Williams (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal*, 9(4):373-380.

J.H. Ward (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236-244.

See Also

dendesc for descriptive measures to analyze dendrograms.

Examples

```
## distances between 10 cities in the US
data(UScitiesD)

## unweighted arithmetic linkage (UPGMA)
lnk1 <- linkage(UScitiesD, method="arithmetic", weighted=FALSE)
plot(lnk1, main="linkage(arithmetic, U)")
```

```
## weighted arithmetic linkage (WPGMA)
lnk2 <- linkage(UScitiesD, method="arithmetic", weighted=TRUE)

## equivalence with hclust, except for the ordering of the leaves
hcl2 <- as.dendrogram(hclust(UScitiesD, method="mcquitty"))
sum(abs(ultrametric(lnk2) - ultrametric(hcl2)))  # 0
opar <- par(mfrow=c(1, 2))
plot(lnk2, main="linkage(arithmetic, W)")
plot(hcl2, main="hclust(mcquitty)")
par(opar)

## unweighted versatile linkage, with par.method=-0.6
lnk3 <- linkage(UScitiesD, method="versatile", weighted=FALSE,
                par.method=-0.6)
plot(lnk3, main="linkage(versatile, -0.6, U)")

## cophenetic correlation coefficient
cor(UScitiesD, ultrametric(lnk1))  # 0.8101937
cor(UScitiesD, ultrametric(lnk2))  # 0.8076422
cor(UScitiesD, ultrametric(lnk3))  # 0.8163286
```

Index