Package 'divclust'

June 23, 2015

Type Package

Title Divisive hierarchical clustering

Version 0.3

Date 2015-06-22

Author Marc Fuentes, Marie Chavent

Maintainer <marie.chavent@inria.fr>

Description This package provides monothetic hierarchical clustering for both quantitative, qualitative and mixed data.

License GPL-2

Imports Rcpp (>= 0.11.0), intervals

LinkingTo Rcpp

R topics documented:

`cutreediv` *Cut the tree*

Description

This function cuts the tree into several cluters by specifying the desired number of clusters.

Usage

```
cutreediv(tree, K)
```

Arguments

`tree`	the divclust object
`K`	an integer with the desired number of clusters.

Value

`clusters`	the list of observations in each cluster
`description`	the monothetic description of each cluster
`which_cluster`	a vector of integers indicating the cluster of each observation
`B`	the proportion of inertia explained by the partition (between-cluster inertia/total inertia)
`leaves`	an internal list of **leaves**

Examples

```
data(protein) # pure quantitatives data
tree <- divclust(protein) # full clustering
p_5 <- cutreediv(tree,K=5) # partition in 5 clusters
p_5

data(dogs) # pure qualitative data
tree <- divclust(dogs) # full clustering
p_4 <- cutreediv(tree,K=4) # partition in 4 clusters
data(wine) # mixed data
data <- wine[,1:29]
tree <- divclust(data) # full clustering
p_4 <- cutreediv(tree, 4) #
```

`divclust` *Monothetic divisive hierarchical clustering*

Description

DIVCLUS-T is a divisive hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. It is designed for numerical, categorical (ordered or not) or mixed data. Like the Ward agglomerative hierarchical clustering algorithm and the k-means partitioning algorithm, it is based on the minimization of the inertia criterion. However, it provides a simple and natural monothetic interpretation of the clusters. Indeed, each cluster is decribed by set of binary questions. The inertia criterion is calculated on all the principal components of PCAmix (and then on standardized data in the numerical case).

Usage

```
divclust(data, K = NULL)
```

Arguments

`data`	a data frame with numerical and/or categorical variables. If the variable is ordinal, the column must be of class factor with the argument ordered=TRUE.
`K`	the number of final clusters (leaves of the tree). By default, the complete dendrogram is performed.

Details

The tree has K leaves corresponding to a partition in K clusters if K is specified in input. Otherwise, each final cluster contains one observation and the tree is the complete dendrogram. The between-cluster inertia of the final partition of the leaves is the sum of the heights of the clusters in the tree. The total inertia for the quantitative dataset is equal to p1 (the number of quantitative variables). The total inertia for the qualitative dataset is m-p2 where m is the total number of categories and p2 is the number of qualitative variables. For a mixture of quantitative and qualitative data, the total variance is p1+m-p2. The quality of a partition is the proportion of inertia explained by the patition which is the between-cluster inertia divided by the total inertia. The height of a cluster in the dendrogram of divclust is the inertia variation which is also the aggregation criterion of Ward used in ascendant hierarchical clustering. This can be used, to help in the choice of the number of clusters as for Ward hierarchical clustering. For ordered qualitative variables (class factor with argument ordered =TRUE), this order on the categories is used to reduce the number of possible binary questions.

Value

`tree`	an internal **tree**
`clusters`	the list of observations in each final cluster (the leaves of the tree)
`description`	the monothetic description of each final cluster (the leaves of the tree)
`which_cluster`	a vector of integers indicating the final cluster of each observation
`height`	the height of the clusters in the dendrogram of the tree
`B`	the proportion of inertia explained by the final partition (between-cluster inertia/total inertia)

data_quanti	the quantitative data set
data_quali	the qualitative data set
mod_quali	the list of categories of qualitative variables
vec_quali	number of categories of each qualitative variable
kmax	the number of different observations i.e. the maximal number of leaves
T	The total inertia

See Also

plot.divclust cutreediv

Examples

```
data(protein) # pure quantitatives data
tree <- divclust(protein) # full clustering
plot(tree)
plot(1:(tree$kmax-1),tree$height,xlab="number of cluster",ylab="height",main="Split levels")
c_5 <- divclust(protein, K=5) # stops clustering to 5 clusters
plot(c_5,nqbin=4)
c_5$B*100 #explained inertia
c_5$clusters  # retrieve the list of observations in each cluster
c_5$description # and their monothetic description

data(dogs) # pure qualitative data
tree <- divclust(dogs) # full clustering
plot(tree)
plot(1:(tree$kmax-1),tree$height,xlab="number of cluster",ylab="height",main="Split levels")
c_4 <- divclust(dogs, K=4) # stops clustering to 4 clusters
plot(c_4)
c_4$clusters # retrieve the list of observations in each cluster
c_4$description # and their monothetic description
c_4$which_cluster # return a vector indicating to which cluster belongs each individual
c_4$B*100 #explained variance

dogs2 <- dogs # take the order of categories into account (to reduce the complexity)
levels(dogs$Size)
size2 <- factor(dogs$Size,c("small","large","medium")) #changes the order of the levels
levels(size2)
dogs2$Size <- ordered(size2) #specify argument ordered=TRUE in the class factor
tree <- divclust(dogs2) # full clustering with variable Size considered as ordered.
plot(tree) #the constraint on the order changes the clustering
data(wine) # mixed data
data <- wine[,1:29]
c_tot <- divclust(data) # full clustering
plot(c_tot)
c_4 <- divclust(data, 4) # stops clustering to 4 clusters
plot(c_4)
p2 <- length(c_4$vec_quali)
p1 <- ncol(data)-p2
sum(c_4$height)/(p1+sum(c_4$vec_quali)-p2)*100 #explained variance
c_tot$tree$v #internal contain of the root node
c_tot$tree$r$v #internal contain of the right node of the root node
```

dogs *Breeds of Dogs data*

Description

Data refering to 27 breeds of dogs.

Format

A data frame with 27 rows (the breeds of dogs) and 7 columns: their size, weight and speed with 3 categories (small, medium, large), their intelligence (low, medium, high), their affectivity and aggressiveness with 3 categories (low, high), their function (utility, compagny, hunting).

Source

Originated by A. Brefort (1982) and cited in Saporta G. (2011).

equality_case *Equality case data*

Description

These data illstrate the case where two binary questions give the same bipartition and how the binary question with the most discriminant variable (X1 here) is chosen.

Format

A numerial data frame with 20 rows and 2 columns simulated from two gaussien distributions.

Examples

```
data(equality_case)
plot(equality_case) #X1 discriminates the bipartition better than X2
tree <- divclust(equality_case,K=3)
plot(tree,nqbin=1) # the binary question with X1 is chosen
```

gironde *gironde*

Description

A list of 4 datasets caracterizing conditions of life of 542 cities in Gironde. The four datasets correspond to four thematics relative to conditions of life. Each dataset contains a different number of variables (quantitative and/or qualitative). The first three datasets come from the 2009 population census realized in Gironde by INSEE (Institut National de la Statistique et des Etudes Economiques). The fourth come from an IGN (Institut National de l'Information Geographique et forestiere) database.

Usage

```
data(gironde)
```

Format

A list of 4 data frames.

Value

`gironde$employment`
: This data frame contains the description of 542 cities by 9 quantitative variables. These variables are related to employment conditions like, for instance, the average income (income), the percentage of farmers (farmer).

`gironde$housing`
: This data frame contains the description of 542 cities by 5 variables (2 qualitative variables and 3 quantitative variables). These variables are related to housing conditions like, for instance, the population density (density), the percentage of counsil housing within the cities (council).

`gironde$services`
: This data frame contains the description of 542 cities by 9 qualitative variables. These variables are related to the number of services within the cities, like, for instance, the number of bakeries (baker) or the number of post office (postoffice).

`gironde$environment`
: This data frame contains the description of 542 cities by 4 quantitative variables. These variables are related to the natural environment of the cities, like, for instance the percentage of agricultural land (agricul) or the percentage of buildings (building).

Source

www.INSEE.fr

www.ign.fr

http://siddt.grenoble.cemagref.fr/

Multivariate analysis of mixed data: The PCAmixdata R package, M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco, arXiv:1411.4911 [stat.CO]

---

gsvd *Generalized Singular Value Decomposition*

Description

Performs the generalized singular value decomposition $A = UDV$ with weights M on the columns and N on the rows.

Usage

```
gsvd(A, M, N)
```

Arguments

A	a n times p numerical matrix.
M	a vector of size p with diag(M) the metric on R^p.
N	a vector of size n with diag(N) the metric on R^n.

Value

a 3-tuple (U, V, d) with the left singular vectors, right singular vectors and singular values.

---

plot.divclust *Dendrogram with binary questions*

Description

Plot the dendrogram produced by divclust algorithm with the binary questions associated with each split. The number of binary questions drawn on the dendrogram can be changed and if the text of the question is to long, is is printed on the console.

Usage

```
## S3 method for class divclust
plot(x, nqbin = 3, label = TRUE, ...)
```

Arguments

x	object of class divclust
nqbin	an integer between 0 and K-1 (K is the number of leaves in the tree). Indicates the number of binary questions drawn on plot. This parameter is used to plot the binary questions only of the top levels of the dendrogram. The default value is 3.
label	If TRUE, the labels of the observations are drawn.
...	further arguments passed from other methods

See Also

divclust

Examples

```
data(protein) # pure quantitatives data
c_tot <- divclust(protein) # full clustering
plot(c_tot)
plot(c_tot,nqbin=4) # the text of the 4th question is printed in the console
c_4 <- divclust(protein, K=4) # stops the clustering to 4 clusters
plot(c_4)

data(dogs) # pure qualitative data
c_tot <- divclust(dogs) # full clustering
plot(c_tot,nqbin=4) # the text of the 4th question is printed in the console

data(wine) # mixed data
```

```
data <- wine[,1:29]
c_tot <- divclust(data) # full clustering
plot(c_tot)
```

`print.cutreediv` *Prints a 'cutdiv' object*

Description

This is a method for the function print for objects of the class `cutdiv`.

Usage

```
## S3 method for class cutreediv
print(x, ...)
```

Arguments

`x`	object of class cutdiv
`...`	further arguments passed from other methods. They are ignored in this function.

`print.divclust` *Prints a 'divclust' object*

Description

This is a method for the function print for objects of the class `divclust`.

Usage

```
## S3 method for class divclust
print(x, ...)
```

Arguments

`x`	object of class divclust
`...`	further arguments passed from other methods. They are ignored in this function.

protein *Protein data*

Description

The data measure the amount of protein consumed for nine food groups in 25 European countries. The nine food groups are red meat (RedMeat), white meat (WhiteMeat), eggs (Eggs), milk (Milk), fish (Fish), cereal (Cereal), starch (Starch), nuts (Nuts), and fruits and vegetables (FruitVeg).

Format

A data frame with 25 rows (the European countries) and 9 columns (the food groups)

Source

Originated by A. Weber and cited in Hand et al., A Handbook of Small Data Sets, (1994, p. 297).

split_mix *Splits in quantitative and qualitative datasets*

Description

Splits a dataframe in two data frames. The first one contains the numerical columns and the second one contains the categorical columns.

Usage

```
split_mix(base)
```

Arguments

base	the data

Value

data_quanti	the numerical data frame
data_quali	the categorical data frame

Examples

```
data(wine)
split_mix(wine)$data_quanti
split_mix(wine)$data_quali
```

`wine` *Wines of Val de Loire data*

Description

data refering to 21 wines of Val de Loire.

Format

A data frame with 21 rows (the number of wines) and 31 columns: the first column corresponds to the label of origin, the second column corresponds to the soil, and the others correspond to sensory descriptors.

Source

Centre de recherche INRA d'Angers

Index