Package 'EnsCat'

February 1, 2017

Type Package

Title Clustering of Categorical Data

Version 1.1

Date 2017-1-29

Author Saeid Amiri, Bertrand Clarke and Jennifer Clarke.

Maintainer Saeid Amiri <saeid.amiri1@gmail.com>

Depends stats, utils, graphics, dendextend, ggplot2, ggdendro, seqinr, R (>= 3.3.2)

Description An implementation of the clustering methods of categorical data discussed in Amiri, S., Clarke, B., and Clarke, J. (2015). Clustering categorical data via ensembling dissimilarity matrices. Preprint <arXiv:1506.07930>.

License GPL (>= 2)

URL https://github.com/jlp2duke/EnsCat/wiki/How-To-with-Examples

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2017-02-01 01:39:44

R topics documented:

alphadata 2
Benhc 3
cancer 4
CTN 5
ebola 5
enhcHi 6
EnsCat 7
ggdplot 8
hammingD 9
kmodes 10
lympho 10

mush 11
rhabdodata 12
soybean 13
tangle 13
USFlag 15
zoo 16

Index **18**

`alphadata` *Alphaherpesvirinae virus genome sequence data*

Description

A dataset consisting of whole genome sequences for viruses from the family Alphaherpesvirinae

Usage

```
data("alphadata")
```

Format

This dataset is a matrix of dimension 98x359883 that represents the sequences of 98 viral genomes from the subfamily Alphaherpesvirinae.

Details

This data includes whole genome sequences of viruses belonging to the subfamily Alphaherpesvirinae. Alphaherpesvirinae is a subfamily of the Herpesviridae family of viruses that cause diseases in humans and animals. The data is downloaded from ViPR, http://www.viprbrc.org, and are aligned using "MAFFT", see Katoh et a. (2013), and saved in "alphadata". Alphaherpesvirinae has five genera: Iltovirus (Ilt), Mardivirus (Mar), Scutavirus, Simplexvirus (Sim), and Vari-cellovirus (Var). The viruses were collected from different hosts, namely, human, monkey, chicken, turkey, duck, cow, bat, equidae, boar, cat, amazona oratrix (denoted hum, mon, chi, tur, duc, cow, bat, equ, boa, cat, aor). The codes in the example show the labels.

References

Katoh, K., and D.M. Standley (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

Pickett, BE et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 40: D593-8.

Examples

```
### load Alphaherpesvirinae data
#data("alphadata")
### the following codes define the labels of the data by genera and host.
#xlab1<-NULL
#xlab1[1:8]<-"Var-boa";xlab1[9:13]<-"Var-hum";xlab1[14]<-"Var-cat"
#xlab1[15:32]<-"Var-equ";xlab1[33]<-"Var-mon";xlab1[34:40]<-"Var-cow"
#xlab1[41:45]<-"Sim-mon";xlab1[46:47]<-"Sim-mon";xlab1[48:58]<-"Sim-hum"
#xlab1[59]<-"Sim-bat";xlab1[60]<-"Sim-mon";xlab1[61]<-"Mar-tur"
#xlab1[62:71]<-"Mar-chi";xlab1[72:78]<-"Mar-duc";xlab1[79]<-"Ilt-ora";xlab1[80:98]<-"Ilt-chi"
```

Benhc *Performs bootstrap ensemble hierarchical clustering for categorical data.*

Description

This function performs a bootstrap ensemble hierarchical clustering of categorical data, as described in details below.

Usage

```
Benhc(x, En)
```

Arguments

x	A nxp data matrix or data frame; n is the number of observations and p is the number of dimensions.
En	Number of clusterings to include in the ensemble, i.e., cardinality of the ensemble.

Details

The function 'Benhc' generates a dissimilarity matrix via the bootstrap ensemble. The ensembled dissimilarity matrix is generated using the same procedure as described for the function 'enhc' except that each clustering is based on a bootstrap sample of the data. The number of clusters for each clustering is selected randomly from {2,...,sqrt(n)}.

References

Amiri, S., Clarke, B., and Clarke, J. (2015). Clustering categorical data via ensembling dissimilarity matrices. arXiv preprint arXiv:1506.07930.

Examples

```
#data('zoo')
### zoo includes the zoo data downloaded from UCI
### Machine Learning Repository
### Calculate ensemble dissimilarities with 150 ensemble members
#disten<-Benhc(zoo$obs,En=150)
### This function performs a hierarchical cluster analysis using
### dissimilarities obtained by the ensembling procedure in Benhc
#en<-hclust(disten,method='average')
### A plot of the dendrogram can be generated by
#plot(en,label=zoo$lab)
```

---

cancer *Primary tumor domain (cancer) data*

Description

Classification data set from the UCI Machine Learning Repository

Usage

```
data("cancer")
```

Format

The format of the data is a list with components $obs and $lab. "cancer$obs" includes the observations that are stored as numerical values. "cancer$lab" contains the labels of the data.

Details

A simple classification data set containing 16 attributes with 339 observations. Since the true labels are known, this data can be used to evaluate clustering methods.

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/datasets/Primary+Tumor

Examples

```
#data(cancer)
```

CTN *convert genetic data (nucleotides) to numerical values*

Description

This function converts genetic data (nucleotides) to numeric data.

Usage

```
CTN(x)
```

Arguments

x	x should be a dataset in fasta format

Details

R is more efficient with numerical data and storage of data via numerical values takes less memory. Genetic data consists of nucleotide data A,T,C,G and are usually saved in Fasta format. After downloading the data from one of the bioinformatics repositories and importing it to R, this function converts the data to numerical values.

Examples

```
### import fasta data to R.
##x.dna0 <- read.fasta("dna.fasta")
###  convert data to numerical values
##x.dna<-CTN(x.dna0)
```

ebola *Ebolavirus genome sequence data*

Description

A dataset consisting of whole genome sequences for Ebolavirus from the family Filovirdae.

Usage

```
data("ebola")
```

Format

The format of the data is a list with components $obs and $lab. "ebola$obs" includes the observations that are stored as numerical values. "ebola$lab" contains the labels of the data.

This ebola$obs is a matrix of dimension 103x26445 that represents the sequences of 103 viral genomes from the Ebolavirus.

Details

This data includes whole genome sequences of viruses belonging to the Ebolavirus. Ebolavirus is a subfamily of the Filovirdae family of viruses. The data is downloaded from ViPR, http://www.viprbrc.org, and are aligned using "MAFFT", see Katoh et a. (2013), and saved in "ebola". Ebolavirus subdivides into five species: Bundibugyo virus (Bun), Reston ebolavirus (Res), Sudan ebolavirus (Sud), Tai Forest ebolavirus (Tai), and Zaire ebolavirus (Zai). The hosts are human, monkey, swine, guinea pig, mouse, and bat (denoted hum, mon, swi, gpi, mou, bat, respectively, in our dataset). The ebola$lab in the example show the labels, the combination of species and host.

References

Katoh, K., and D.M. Standley (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

Pickett, BE et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 40: D593-8.

Examples

```
### load Ebolavirus data
#data("ebola")
```

`enhcHi` *Performs ensemble hierarchical clustering for high dimensional categorical data*

Description

This function performs an ensemble hierarchical clustering of high dimensional categorical data (p » n).

Usage

```
enhcHi(data, En=100, len=c(2,10), type=2)
```

Arguments

`data`	A nxp data matrix of data frame; n is the number of observations and p is the number of features or dimensions.
`En`	Number of clusterings to include in the ensemble, i.e., cardinality of the ensemble.
`len`	Range of sizes of clusterings (i.e., number of clusters) to run and ensemble.
`type`	Numeric indicator of single bootstrap (type=1) or double bootstrap (type=2) for selecting subsets of variables to include in each clustering within the ensemble. The default is type=2

References

Amiri, S., Clarke, B., and Clarke, J. (2015). Clustering categorical data via ensembling dissimilarity matrices. arXiv preprint arXiv:1506.07930.

Examples

```
#data("rhabdodata")
### The following code generates the dissimilary matrix of sequence data stored in alphadata
### The ensemble has 100 member clusterings, and the number of clusters in each clustering
### is generated randomly from a discrete uniform on (2,10). A double bootstrap procedure is
### used to select a subset of variables for each clustering.
#ens<-enhcHi(rhabdodata$dat,En=100,len=c(2,10), type=2)
### Calculate the hamming distance
#dis0<-hammingD(ens)
### Save as distance format
#REDIST<-as.dist(dis0)
#hc0 <- hclust(REDIST,method = "average")
#plot(hc0,label=rhabdodata$lab,hang =-1)
```

EnsCat — *This package includes several methods that can be used to cluster categorical data.*

Description

EnsCat implements several methods for clustering of categorical data.

Details

Package:	EnsCat
Type:	Package
Version:	1.1
Date:	2017-01-29
License:	>=GPL-2
URL:	https://github.com/jlp2duke/EnsCat/wiki/How-To-with-Examples

Author(s)

Maintainer: Saeid Amiri <saeid.amiri1@gmail.com>

References

Amiri, S., Clarke, B., and Clarke, J. (2015). Clustering categorical data via ensembling dissimilarity matrices. arXiv preprint arXiv:1506.07930.

ggdplot *Nice plots of hierarchical clustering results via ggdendrogram*

Description

This function provides two different plotting options for either a dendro object or an object that can be coerced to class dendro, using the function ggdendrogram from the package ggdendro

Usage

```
ggdplot(hc, lab = NULL, ptype = 1, title = NULL, ...)
```

Arguments

hc	Either a dendro object or an object that can be coerced to class dendro using the dendro_data function, i.e. objects of class dendrogram, hclust or tree
lab	A character vector of labels for the leaves of the tree. By default labels in hc are used.
ptype	A numeric indicator of the type of plot desired. If ptype==1, the plot is rotated by 90 degrees, the theme is blank, and the title is plotted. If ptype!=1, the plot is rotated by 90 degrees, and the background is a grayscale grid.
title	A character label for the title of the plot. Only used if ptype!=1.
...	other parameters passed to geom_text

Details

Given either a dendro object or an object that can be coerced to class dendro, this is a convenience function for plotting. For an object of type dendro, if ptype==1, the function executes the equivalent of

```
ggdendrogram(hcdata, rotate=TRUE, size=2) + labs(title="Dendrogram in ggplot2")
```

If ptype!=1, the function executes the equivalent of

```
ggdendrogram(hcdata, rotate = TRUE, theme_dendro = FALSE)
```

Objects that are not of class dendro are coerced to class dendro prior to plotting.

Value

A ggplot object

Examples

```
library(ggplot2)
hc <- hclust(dist(USArrests), "ave")
p<-ggdplot(hc, ptype=2)
```

hammingD *Calculate the hamming distance between data points.*

Description

Hamming distance is defined on categorical vectors. It counts the number of times the coordinates in two data vectors differ, or the number of substitutions required to convert one data vector into the other. Here the Hamming distance is normalized, so the result is the number of coordinates that differ divided by the vector length.

Usage

```
hammingD(dat)
```

Arguments

dat	dat should be a matrix or data frame of data. n is the number of observations (rows) and p is the number of dimensions (columns).

Details

This function calculates the Hamming distance (normalized) between rows of the input data.

Value

The result is a nxn matrix whose (i,j) element is the Hamming distance between rows i and j

See Also

See Also as alphadata,

Examples

```
### The running is time consuming
### Run hamming distance
#dis0<-hammingD(alphadata)
### Save as distance format
#REDIST<-as.dist(dis0)
### Run a hierarchical clustering using average linkage
#hc0 <- hclust(REDIST,method = "average")
### plot the dendrogram
#plot(hc0,label=xlab1,hang =-1)
```

kmodes *Run Kmodes*

Description

This function runs Kmodes. The user must choose the number of clusters and the initial modes.

Usage

```
kmodes(data, k, k2)
```

Arguments

data	data should be a matrix or data frame, columns include the variables.
k	number of clusters
k2	set of initial modes; indices of data points

Details

This function clusters the rows of the data.

References

Huang, Z. (1998). Extensions to the v-means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery, 2, 283-304.

Examples

```
data("zoo")
### Run Kmodes on zoo data with 7 clusters and the first seventh observations as initial modes
kmodes(zoo$obs,k=7,1:7)
### Run Kmodes with seven random initial modes selected from data points
kmodes(zoo$obs,k=7,sort(sample(dim(zoo$obs)[1],7)))
```

lympho *Lymphography domian (lympho) data*

Description

Classification data set from the UCI Machine Learning Repository

Usage

```
data("cancer")
```

Format

The format of the data is a list with components $obs and $lab. "lympho$obs" includes the observations that are stored as numerical values. "lympho$lab" contains the labels of the data.

Details

A simple classification data set containing 18 attributes with 148 observations. Since the true labels are known, this data can be used to evaluate clustering methods.

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/datasets/Lymphography

Examples

```
data(lympho)
```

---

mush *Mushroom data*

Description

Classification data set from the UCI Machine Learning Repository

Usage

```
data("mush")
```

Format

The format of the data is a list with components $obs and $lab. "mush$obs" includes the observations that are stored as numerical values. "mush$lab" contains the labels of the data.

Details

A simple classification data set containing 22 attributes with 8124 observations, because the dataset is large, we only used the last 400 observations in our analysis. Since the true labels are known, this data can be used to evaluate clustering methods.

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/datasets/Mushroom

Examples

```
data(mush)
```

`rhabdodata` *Rhabdoviridae virus genome sequence data*

Description

This dataset consists of whole genome sequences of viruses from the Family Rhabdoviridae.

Usage

```
data("rhabdodata")
```

Format

The format of the data is a list with components $dat and $lab. The components dat includes a matrix of dimension 53x26035 that represents the sequences of 53 viral genomes from the family Rhabdoviridae. The component lab includes labels for each sample that include abbreviations of the relevant genus and viral host.

Details

This data includes whole genome sequences of viruses belonging to the subfamily Rhabdoviridae. Rhabdoviridae is a family of viruses with single-stranded RNA genomes that are able to infect a wide variety of hosts, both plants and animals. cause diseases in humans and animals. The data is downloaded from ViPR, http://www.viprbrc.org, and are aligned using "MAFFT", see Katoh et a. (2013), and saved in "rhabdodata". Rhabdoviridae has twelve genera of which nine are repesented here: Cytorhabdovirus, Ephemerovirus, Novirhabdovirus, Nucleorhabdovirus, Perhabdovirus, Sigmavirus, Sprivivirus, Tibrovirus, and Tupavirus. The viruses were collected from different hosts, namely, Alfalfa, Cattle, Drosophila, Eel, Fish, Garlic, Midge, Mosquito, Eggplant, Taro, Trout, and Unknown.

References

Katoh, K., and D.M. Standley (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

Pickett, BE et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 40: D593-8.

Examples

```
### load Rhabdoviridae data
data("rhabdodata")
### the following codes define the labels of the data by genera and host.
dim(rhabdodata$dat)
#[1] 53 26035
```

soybean *Soybean (small) data*

Description

Classification data set from the UCI Machine Learning Repository

Usage

```
data("soybean")
```

Format

The format of the data is a list with components $obs and $lab. "soybean$obs" includes the observations that are stored as numerical values. "soybean$lab" contains the labels of the data.

Details

A simple classification data set containing 35 attributes with 47 observations. Since the true labels are known, this data can be used to evaluate clustering methods.

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/datasets/Soybean+(Small)

Examples

```
data(soybean)
```

tangle *Generate a tanglegram from two hierarchical clusterings of a data set*

Description

This function generates a tanglegram of two different hierarchical clusterings of the same dataset. This is essentially a convenience wrapper for the function `tanglegram` in the package `dendextend`; see Galili (2015).

Usage

```
tangle(hc0, hc1)
```

Arguments

hc0	An object that can be coerced to a dendrogram, e.g., an object from hclust. This object will be plotted on the left side of the tanglegram.
hc1	An object that can be coerced to a dendrogram, e.g., an object from hclust. This object will be plotted on the right side of the tanglegram.

Details

This function is a convenience wrapper for the function tanglegram in the R package dendextend; see http://cran.at.r-project.org/web/packages/dendextend/. A tanglegram is used to visualize the similarities and differences between two different hierarchical clusterings of the same dataset.

Value

An invisible dendlist, with two trees after being modified during the creation of the tanglegram

References

https://cran.r-project.org/package=dendextend, https://github.com/talgalili/dendextend/, http://www.r-statistics.com/tag/dendextend/, http://bioinformatics.oxfordjournals.org/content/31/22/3718 al Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. doi:10.1093/bioinformatics/btv428

See Also

tanglegram

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==>  Define data, use random,
##--or do  help(data=index)  for the standard data sets.

## The function is currently defined as
function (hc0, hc1)
{
    hcd0 <- as.dendrogram(hc0)
    hcd1 <- as.dendrogram(hc1)
    hcd0 <- match_order_by_labels(hcd0, hcd1)
    dends_0_1 <- dendlist(hcd0, hcd1)
    t <- tanglegram(dends_0_1)
    t
  }
```

USFlag *United States Flag Privately-Owned Merchant Fleet Data*

Description

This dataset includes 10 categorical variables that describe U.S. flag privately owned merchant fleet vessels, based on data provided by the United Stated Department of Transportation Maritime Administration (MARAD).

Usage

```
data("USFlag")
```

Format

The format of the data is a list with components $lab and $obs. The component $lab contains a categorical indicator of Ship Type (see below). The component $obs includes a matrix of dimension 170x10 that contains categorical data on 170 United States flag privately owned merchant fleet vessels. The columns are as follows:

Ship.Type a categorical variable with 5 levels, Containership [1], Dry Bulk [2], General Cargo [3], Ro-Ro [4], and Tanker [5], indicating the ship type. This is identical to $lab.

Gross.Tonnage a categorical variable with 6 levels, <20000GT [1], 20000-40000GT [2], 40000-60000GT [3], 60000-80000GT [4],80000-100000GT [5], >100000GT [6], indicating the ship gross tonnage

Deadweight a categorical variable with 6 levels, <20000DWT [1], 20000-40000DWT [2], 40000-60000DWT [3], 60000-100000DWT [4], 100000-140000DWT [5], >140000DWT [6], indicating the ship deadweight

Year.Built a categorical variable with 6 levels, <1960 [1], 1961-1980 [2], 1981-1990 [3], 1991-2000 [4], 2001-2010 [5], and >2010 [6], indicating the year of completion of ship construction

Operator a categorical variable with 49 levels indicating the operator of the ship

MSP a binary variable indicating whether the ship is [1] or is not [0] part of the maritime security program

VISA a binary variable indicating whether the ship is [1] or is not [0] part of the Voluntary Intermodal Sealift Agreement

VTA a binary variable indicating whether the ship is [1] or is not [0] part of the Voluntary Tanker Agreement

Jones.Act.Eligible a binary variable indicating whether the ship is [1] or is not [0] Jones Act Eligible. These vessels are eligible to participate in domestic trade. Jones Act eligible vessels are built in the United States, owned by United States citizens and crewed by U.S. Mariners

Militarily.Useful a binary variable indicating whether the ship is [1] or is not [0] considered a militarily useful sealift vessel

For more information on these definitions please see IHS Maritime, Sea-Web. www.sea-web.com

Details

This data includes categorical variables that describe U.S. flag privately owned merchant fleet vessels. Information is provided only for oceangoing, self-propelled, cargo-carrying vessels of 1,000 gross tons and above. These data are based on information from the U.S. Department of Transportation Maritime Administration (MARAD) as of 3/3/2015, obtained from the MARAD Open Data Portal (http://www.marad.dot.gov/resources/data-statistics/).

Source

United States Maritime Administration (MARAD) Open Data Portal. http://www.marad.dot.gov/resources/data-statistics/

Examples

```
### load USFlag maritime data
data("USFlag")
### the following codes define the labels of the data by genera and host.
dim(USFlag$obs)
#[1] 170  10
length(USFlag$lab)
#[1] 170
```

---

zoo *zoo data*

Description

Classification data set from the UCI Machine Learning Repository

Usage

```
data("zoo")
```

Format

The format of the data is a list with components $obs and $lab. "zoo$obs" includes the observations that are stored as numerical values. "zoo$lab" contains the labels of the data.

Details

A simple classification data set containing 17 Boolean-valued attributes with 101 observations. Since the true labels are known, this data can be used to evaluate clustering methods.

Source

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/datasets/Zoo

Examples

```
data(zoo)
```

Index

∗Topic **datasets**
alphadata, 2
cancer, 4
ebola, 5
lympho, 10
mush, 11
rhabdodata, 12
soybean, 13
USFlag, 15
zoo, 16
∗Topic **dendrogram**
tangle, 13
∗Topic **highdimension**
enhcHi, 6
∗Topic **lowdimension**
Benhc, 3
kmodes, 10
∗Topic **package**
EnsCat, 7
∗Topic **tanglegram**
tangle, 13

alphadata, 2, *9*

Benhc, 3

cancer, 4
CTN, 5

dendlist, *14*
dendro_data, *8*

ebola, 5
enhcHi, 6
EnsCat, 7

geom_text, *8*
ggdplot, 8
ggplot, *8*

hammingD, 9
hclust, *14*

kmodes, 10

lympho, 10

mush, 11

rhabdodata, 12

soybean, 13

tangle, 13
tanglegram, *13, 14*

USFlag, 15

zoo, 16