Package 'HDoutliers'

February 11, 2018

Version 1.0

Date 2018-02-09

Title Leland Wilkinson's Algorithm for Detecting Multidimensional Outliers

Author Chris Fraley [aut, cre], Leland Wilkinson [ctb]

Maintainer Chris Fraley <cfraley@tableau.com>

Depends R (>= 3.1.0), FNN, FactoMineR, mclust

Description

An implementation of an algorithm for outlier detection that can handle a) data with a mixed categorical and continuous variables, b) many columns of data, c) many rows of data, d) outliers that mask other outliers, and e) both unidimensional and multidimensional datasets. Unlike ad hoc methods found in many machine learning papers, HDoutliers is based on a distributional model that uses probabilities to determine outliers. See <https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf>.

License MIT + file LICENSE

URL https://www.r-project.org, https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf

NeedsCompilation no

Repository CRAN

Date/Publication 2018-02-11 14:41:07 UTC

R topics documented:

dataTrans *Data Transformation for Leland Wilkinson's* hdoutliers *Algorithm*

Description

Transforms the data according to the specifications in Wilkinson's *hdoutliers* algorithm.

Usage

```
dataTrans(data)
```

Arguments

data	A vector, matrix, or data frame consisting of numeric and/or categorical variables.

Details

Replaces each categorical variables with a numeric variable corresponding to its first component in multiple correspondence analysis, then maps the data to the unit square. There is no porvision for handling missing data. Functions HDoutliers and getHDoutliers apply this transformation to their input data.

Value

The transformed data, according to Wilkinson's specifications for the *hdoutliers* algorithm.

References

Wilkinson, L. (2016). Visualizing Outliers.

See Also

HDoutliers, getHDoutliers

Examples

```
require(FactoMineR)
data(tea)
head(tea)
dataTrans(tea[,-1])
```

dots *One dimensional dots dataset — outlier detection example*

Description

A matrix whose columns are the `Z` and `W` `dots` datasets from Wilkinson (2016).

Usage

```
data(dots)
```

References

L. Wilkinson. 2016. Vizualizing Outliers. <https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf>.

ex2D *Two dimensional dataset — outlier detection example*

Description

A dataset with 510 rows and 2 columns comprised of 500 normally-distributed samples and 10 uniformly distributed outliers.

Usage

```
data(ex2D)
```

getHDmembers *Partitioning Stage of the* hdoutliers *Algorithm*

Description

Implements the first stage of the *hdoutliers* Algorithm, in which the data is partitioned according to *exemplars* and their associated lists of *members*.

Usage

```
getHDmembers(data, maxrows = 10000, radius = NULL)
```

Arguments

data	A vector, matrix, or data frame consisting of numeric and/or categorical variables.
maxrows	If the number of observations is greater than maxrows, HDoutliers reduces the number used in nearest-neighbor computations to a set of *exemplars*. The default value is 10000.
radius	Threshold for determining membership in the *exemplars*'s lists (used only when the number of observations is greater than maxrows). An observation is added to an *exemplars*' list if its distance to that *exemplar* is less than radius. The default value is $.1/(logn)^(1/p)$, where n is the number of observations and p is the dimension of the data.

Details

If the number of observations exceeds maxrows, the data is partitioned into lists corresponding to *exemplars* and their *members* within radius of each *exemplar*, to reduce the number of nearest-neighbor computations required for outlier detection.
When there are fewer observations, the result is a list whose elements are the individual observations (each observation is an exemplar, with no other members).

Value

A list in which each component is a vector of observation indexes. The first index in each list is the index of the *exemplar* defining that list, and any remaining indexes are the associated *members*, within radius of the *exemplar*.

References

Wilkinson, L. (2016). Visualizing Outliers. <https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf>.

See Also

HDoutliers, getHDoutliers

Examples

```
data(dots)
mem.W <- getHDmembers(dots$W)
out.W <- getHDoutliers(dots$W,mem.W)

data(ex2D)
mem.ex2D <- getHDmembers(ex2D)
out.ex2D <- getHDoutliers(ex2D,mem.ex2D)

## Not run:
n <- 100000 # number of observations
set.seed(3)
x <- matrix(rnorm(2*n),n,2)
nout <- 10 # number of outliers
```

```
x[sample(1:n,size=nout),] <- 10*runif(2*nout,min=-1,max=1)

mem.x <- getHDmembers(x)
out.x <- getHDoutliers(x,mem.x)
## End(Not run)
```

`getHDoutliers` *Outlier Detection Stage of Wilkinson's* hdoutliers *Algorithm*

Description

Detects outliers based on a probability model.

Usage

```
getHDoutliers(data, memberLists, alpha = 0.05, transform = TRUE)
```

Arguments

`data`	A vector, matrix, or data frame consisting of numeric and/or categorical variables.
`memberLists`	A list following the structure of the output to `getHDmembers`, in which each component is a vector of observation indexes. The first index in each list is the index of the *exemplar* representing that list, and any remaining indexes are the associated *members*, considered 'close to' the *exemplar*.
`alpha`	Threshold for determining the cutoff for outliers. Observations are considered outliers outliers if they fall in the $(1 - alpha)$ tail of the distribution of the nearest-neighbor distances between *exemplars*.
`transform`	A logical variable indicating whether or not the data needs to be transformed to conform to Wilkinson's specifications before outlier detection. The default is to transform the data using function `dataTrans`. In Wilksinson's algorithm, `memberLists` would have been created with transformed data.

Details

An exponential distribution is fitted to the upper tail of the nearest-neighbor distances between *exemplars* (the observations considered representatives of each component of `memberLists`). Observations are considered outliers if they fall in the $(1 - alpha)$ tail of the fitted CDF.

Value

The indexes of the observations determined to be outliers.

References

Wilkinson, L. (2016). Visualizing Outliers. <https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf>.

Note

A call to `getHDoutliers` in which `membersLists` result from a call to `getHDmembers` is equivalent to calling `HDoutliers`.

See Also

`HDoutliers`, `getHDmembers`, `dataTrans`

Examples

```
data(dots)
mem.W <- getHDmembers(dots$W)
out.W <- getHDoutliers(dots$W,mem.W)
## Not run:
plotHDoutliers( dots.W, out.W)
## End(Not run)

data(ex2D)
mem.ex2D <- getHDmembers(ex2D)
out.ex2D <- getHDoutliers( ex2D, mem.ex2D)
## Not run:
plotHDoutliers( ex2D, out.ex2D)
## End(Not run)

## Not run:
n <- 100000 # number of observations
set.seed(3)
x <- matrix(rnorm(2*n),n,2)
nout <- 10 # number of outliers
x[sample(1:n,size=nout),] <- 10*runif(2*nout,min=-1,max=1)

mem.x <- getHDmembers(x)
out.x <- getHDoutliers(x)
## End(Not run)
```

`HDoutliers` *Leland Wilkinson's* hdoutliers *Algorithm for Outlier Detection*

Description

Detects outliers based on a probability model.

Usage

```
HDoutliers(data, maxrows=10000, radius=NULL, alpha=0.05, transform=TRUE)
```

Arguments

`data`	A vector, matrix, or data frame consisting of numeric and/or categorical variables.
`maxrows`	If the number of observations is greater than `maxrows`, `HDoutliers` reduces the number used in nearest-neighbor computations to a set of *exemplars*. The default value is 10000.
`radius`	Threshold for determining membership in the *exemplars*'s lists (used only when the number of observations is greater than $maxrows$). An observation is added to an *exemplars*' lists if its distance to that *exemplar* is less than `radius`. The default value is $.1/(logn)^(1/p)$, where n is the number of observations and p is the dimension of the data.
`alpha`	Threshold for determining the cutoff for outliers. Observations are considered outliers outliers if they fall in the $(1 - alpha)$ tail of the distribution of the nearest-neighbor distances between *exemplars*.
`transform`	A logical variable indicating whether or not the data needs to be transformed to conform to Wilkinson's specifications before outlier detection. The default is to transform the data using function `dataTrans`.

Details

Wilkinson replaces categorical variables with the leading component from correspondence analysis, and maps the data to the unit square. This is done as a preprocessing step if `transform = TRUE` (the default).

If the number of observations exceeds `maxrows`, the data is first partitioned into lists associated with *exemplars* and their *members* within `radius` of each *exemplar*, to reduce the number of nearest-neighbor computations required for outlier detection.

An exponential distribution is then fitted to the upper tail of the nearest-neighbor distances between *exemplars*. Observations are considered outliers if they fall in the $(1 - alpha)$ tail of the fitted CDF.

Value

The indexes of the observations determined to be outliers.

References

Wilkinson, L. (2016). Visualizing Outliers.

See Also

`getHDmembers`, `getHDoutliers`, `dataTrans`

Examples

```
data(dots)
out.W <- HDoutliers(dots$W)
## Not run:
plotHDoutliers(dots$W,out.W)
```

```
## End(Not run)

data(ex2D)
out.ex2D <- HDoutliers(ex2D)
## Not run:
plotHDoutliers(ex2D,out.ex2D)
## End(Not run)

## Not run:
n <- 100000 # number of observations
set.seed(3)
x <- matrix(rnorm(2*n),n,2)
nout <- 10 # number of outliers
x[sample(1:n,size=nout),] <- 10*runif(2*nout,min=-1,max=1)

out.x <- HDoutliers(x)
## End(Not run)
```

`plotHDoutliers` *Display Outlier Detection Results*

Description

Plotting function showing observations determined to be outliers.

Usage

```
plotHDoutliers(data, indexes = NULL, transform = TRUE, ...)
```

Arguments

`data`	A vector, matrix, or data frame consisting of numeric and/or categorical variables.
`indexes`	The (row) indexes of the outliers in `data`.
`transform`	A logical variable indicating whether or not the data needs to be transformed to conform to Wilkinson's specifications before outlier detection. The default is to transform the data using function `dataTrans`. In Wilksinson's algorithm, `indexes` would have been derived from transformed data.
`...`	Additional plotting arguments.

Details

Produces a plot of the data (transformed according to the Wilkinson's specifications) showing the outliers. If the data has more than two dimensions, it is plotted onto the principal components of the data that remains after removing outliers.

Value

The indexes of the observations determined to be outliers.

References

Wilkinson, L. (2016). Visualizing Outliers.

See Also

HDoutliers, dataTrans

Examples

```
data(dots)
out.W <- HDoutliers(dots$W)
## Not run:
plotHDoutliers(dots$W,out.W)
## End(Not run)

data(ex2D)
out.ex2D <- HDoutliers(ex2D)
## Not run:
plotHDoutliers(ex2D,out.ex2D)
## End(Not run)
```

Index