Package 'OutlierDetection'

February 7, 2019

Type Package

Title Outlier Detection

Version 0.1.0

Author Vinay Tiwari, Akanksha Kashikar

Maintainer Vinay Tiwari <vinaystiwari786@gmail.com>

Description To detect outliers using different methods namely model based outlier detection (Barnett, V. 1978 <https://www.jstor.org/stable/2347159>), distance based outlier detection (Hautamaki, V., Karkkainen, I., and Franti, P. 2004 <http://cs.uef.fi/~franti/papers.html>), dispersion based outlier detection (Jin, W., Tung, A., and Han, J. 2001 <https://link.springer.com/chapter/10.1007/0-387-25465-X_7>), depth based outlier detection (Johnson, T., Kwok, I., and Ng, R.T. 1998 <http://www.aaai.org/Library/KDD/1998/kdd98-038.php>) and density based outlier detection (Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996 <https://dl.acm.org/citation.cfm?id=3001507>). This package provides labelling of observations as outliers and outlierliness of each outlier. For univariate and bivariate data, visualization is also provided.

Imports ggplot2, DDoutlier, depth, depthTools,ldbod

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-02-07 17:43:36 UTC

R topics documented:

`dens` *Outlier detection using Robust Kernal-based Outlier Factor(RKOF) algorithm*

Description

Takes a dataset and find its outliers using Robust Kernal-based Outlier Factor(RKOF) algorithm

Usage

```
dens(x, k = 0.05 * nrow(x), C = 1, alpha = 1, sigma2 = 1,
  cutoff = 0.95, rnames = F, boottimes = 100)
```

Arguments

`x`	dataset for which outliers are to be found
`k`	No. of nearest neighbours to be used, default value is 0.05*nrow(x)
`C`	Multiplication parameter for k-distance of neighboring observations. Act as bandwidth increaser. Default is 1 such that k-distance is used for the gaussian kernel
`alpha`	Sensivity parameter for k-distance/bandwidth. Small alpha creates small variance in RKOF and vice versa. Default is 1
`sigma2`	Variance parameter for weighting of neighboring observations
`cutoff`	Percentile threshold used for distance, default value is 0.95
`rnames`	Logical value indicating whether the dataset has rownames, default value is False
`boottimes`	Number of bootsrap samples to find the cutoff, default is 100 samples

Details

dens computes outlier score of an observation using DDoutlier package(based on RKOF algorithm) and based on the bootstrapped cutoff, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

References

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
dens(X,k=4,C=1)
```

`depthout` *Outlier detection using depth based method*

Description

Takes a dataset and find its outliers using depth-based method

Usage

```
depthout(x, rnames = FALSE, cutoff = 0.05, boottimes = 100)
```

Arguments

`x`	dataset for which outliers are to be found
`rnames`	Logical value indicating whether the dataset has rownames, default value is False
`cutoff`	Percentile threshold used for depth, default value is 0.05
`boottimes`	Number of bootsrap samples to find the cutoff, default is 100 samples

Details

depthout computes depth of an observation using depthTools package and based on the bootstrapped cutoff, label an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

References

Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY. Kno

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
depthout(X,cutoff=0.05)
```

`disp` *Outlier detection using genralised dispersion*

Description

Takes a dataset and find its outliers using dispersion-based method

Usage

```
disp(x, cutoff = 0.95, rnames = FALSE, boottimes = 100)
```

Arguments

`x`	dataset for which outliers are to be found
`cutoff`	Percentile threshold used for distance, default value is 0.95
`rnames`	Logical value indicating whether the dataset has rownames, default value is False
`boottimes`	Number of bootsrap samples to find the cutoff, default is 100 samples

Details

disp computes LOO dispersion matrix for each observation(dispersion matrix without cosidering the current observation) and based on the bootstrapped cutoff for score(difference between determinant of LOO dispersion matrix and det of actual dispersion matrix), labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

References

Jin, W., Tung, A., and Han, J. 2001. Mining top-n local outliers in large databases. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA.

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
disp(X,cutoff=0.99)
```

`maha` *Outlier detection using Mahalanobis Distance*

Description

Takes a dataset and find its outliers using modelbased method

Usage

```
maha(x, cutoff = 0.95, rnames = FALSE)
```

Arguments

`x`	dataset for which outliers are to be found
`cutoff`	Percentile threshold used for distance, default value is 0.95
`rnames`	Logical value indicating whether the dataset has rownames, default value is False

Details

maha computes Mahalanibis distance an observation and based on the Chi square cutoff, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported based on its p vlaues. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: vector of Sr. no. of outliers

Outlier probability: vector of (1-p value) of outlier observations

References

Barnett, V. 1978. The study of outliers: purpose and model. Applied Statistics, 27(3), 242–250.

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
maha(X,cutoff=0.9)
```

---

nn *Outlier detection using k Nearest Neighbours Distance method*

Description

Takes a dataset and find its outliers using distance-based method

Usage

```
nn(x, k = 0.05 * nrow(x), cutoff = 0.95, Method = "euclidean",
  rnames = FALSE, boottimes = 100)
```

Arguments

x	dataset for which outliers are to be found
k	No. of nearest neighbours to be used, default value is 0.05*nrow(x)
cutoff	Percentile threshold used for distance, default value is 0.95
Method	Distance method, default is Euclidean
rnames	Logical value indicating whether the dataset has rownames, default value is False
boottimes	Number of bootsrap samples to find the cutoff, default is 100 samples

Details

nn computes average knn distance of observation and based on the bootstrapped cutoff, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

References

Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK.

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
nn(X,k=4)
```

---

nnk *Outlier detection using kth Nearest Neighbour Distance method*

Description

Takes a dataset and find its outliers using distance-based method

Usage

```
nnk(x, k = 0.05 * nrow(x), cutoff = 0.95, Method = "euclidean",
  rnames = FALSE, boottimes = 100)
```

Arguments

x	dataset for which outliers are to be found
k	No. of nearest neighbours to be used, default value is 0.05*nrow(x)
cutoff	Percentile threshold used for distance, default value is 0.95
Method	Distance method, default is Euclidean
rnames	Logical value indicating whether the dataset has rownames, default value is False
boottimes	Number of bootsrap samples to find the cutoff, default is 100 samples

Details

nnk computes kth nearest neighbour distance of an observation and based on the bootstrapped cut-off, labels an observation as outlier. Outlierliness of the labelled 'Outlier' is also reported and it is the bootstrap estimate of probability of the observation being an outlier. For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Outlier probability: Vector of proportion of times an outlier exceeds local bootstrap cutoff

References

Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK.

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
nnk(X,k=4)
```

`OutlierDetection` *Outlier Detection(Intersection of all the methods)*

Description

Takes a dataset and find its outliers using combination of different method

Usage

```
OutlierDetection(x, k = 0.05 * nrow(x), cutoff = 0.95,
  Method = "euclidean", rnames = FALSE, depth = FALSE,
  dense = FALSE, distance = FALSE, dispersion = FALSE)
```

Arguments

`x`	dataset for which outliers are to be found
`k`	No. of nearest neighbours to be used for for outlier detection using bootstrapping, default value is 0.05*nrow(x)
`cutoff`	Percentile threshold used for distance, default value is 0.95
`Method`	Distance method, default is Euclidean
`rnames`	Logical value indicating whether the dataset has rownames, default value is False
`depth`	Logical value indicating whether depth based method should be used or not, default is False
`dense`	Logical value indicating whether density based method should be used or not, default is False
`distance`	Logical value indicating whether distance based methods should be used or not, default is False
`dispersion`	Logical value indicating whether dispersion based methods should be used or not, default is False

Details

OutlierDetection finds outlier observations for the data using different methods and based on all the methods considered, labels an observation as outlier(intersection of all the methods). For bivariate data, it also shows the scatterplot of the data with labelled outliers.

Value

Outlier Observations: A matrix of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Examples

```
OutlierDetection(iris[,-5])
```

`UnivariateOutlierDetection` *Univariate Outlier Detection(Intersection of all the methods)*

Description

Takes a vector and find its outliers using combination of different methods

Usage

```
UnivariateOutlierDetection(x, k = 0.05 * length(x), cutoff = 0.95,
  dist = FALSE, dens = FALSE, depth = FALSE, Method = "euclidean",
  rnames = FALSE)
```

Arguments

`x`	vector for which outliers are to be found
`k`	No. of nearest neighbours to be used for distance methods, default value is 0.05*nrow(x)
`cutoff`	Percentile threshold used for outlier detection using bootstrapping, default value is 0.95
`dist`	Logical value indicating whether distance based methods should be used or not, default is False
`dens`	Logical value indicating whether density based method should be used or not, default is False
`depth`	Logical value indicating whether depth based method should be used or not, default is False
`Method`	Distance method, default is euclidean
`rnames`	Logical value indicating whether the dataset has rownames, default value is False

Details

UnivariateOutlierDetection finds outlier observations for an univariate data using different methods and based on all the methods, labels an observation as outlier(intersection of all the methods). It also shows the scatterplot of the data with labelled outliers with observation no. as x-axis.

Value

Outlier Observations: A vector of outlier observations

Location of Outlier: Vector of Sr. no. of outliers

Examples

```
#Create dataset
X=iris[,1:4]
#Outlier detection
depthout(X,cutoff=0.05)
UnivariateOutlierDetection(iris[,1],cutoff=.95,Method="euclidean",rnames=FALSE)
```

Index