Package 'outlierensembles'

June 15, 2021

Type Package

Title A Collection of Outlier Ensemble Algorithms

Version 0.1.0

Maintainer Sevvandi Kandanaarachchi <sevvandik@gmail.com>

Description Ensemble functions for outlier/anomaly detection. There is a new ensemble method proposed using
Item Response Theory. Existing outlier ensemble methods from Schubert et al (2012) <doi:10.1137/1.9781611972825.90>,
Chiang et al (2017) <doi:10.1016/j.jal.2016.12.002> and Aggarwal and Sathe (2015) <doi:10.1145/2830544.2830549>
are also included.

License GPL (>= 3)

Encoding UTF-8

Depends R (>= 3.4.0)

Imports airt, EstCRM, psych, apcluster

RoxygenNote 7.1.1

Suggests DDoutlier, knitr, rmarkdown, ggplot2

VignetteBuilder knitr

URL https://sevvandi.github.io/outlierensembles/

NeedsCompilation no

Author Sevvandi Kandanaarachchi [aut, cre]
(<https://orcid.org/0000-0002-0337-0395>)

Repository CRAN

Date/Publication 2021-06-15 07:30:02 UTC

R topics documented:

`average_ensemble` *Uses the mean as the ensemble score*

Description

This function uses the mean as the ensemble score.

Usage

```
average_ensemble(X)
```

Arguments

`X`	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.

Value

The ensemble scores.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- average_ensemble(Y)
ens
```

greedy_ensemble *Computes an ensemble score using the greedy algorithm proposed by Schubert et al (2012)*

Description

This function computes an ensemble score using the greedy algorithm in the paper titled Evaluation of Outlier Rankings and Outlier Scores by Schubert et al (2012) <doi:10.1137/1.9781611972825.90>. The greedy ensemble is detailed in Section 4.3.

Usage

```
greedy_ensemble(X, kk = 5)
```

Arguments

X	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.
kk	The number of estimated outliers.

Value

A list with the components:

scores	The ensemble scores.
methods	The methods that are chosen for the ensemble.
chosen	The chosen subset of original anomaly scores.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- greedy_ensemble(Y, kk=5)
ens$scores
```

`icwa_ensemble` *Computes an ensemble score using inverse cluster weighted averaging method by Chiang et al (2017)*

Description

This function computes an ensemble score using inverse cluster weighted averaging in the paper titled A Study on Anomaly Detection Ensembles by Chiang et al (2017) <doi:10.1016/j.jal.2016.12.002>. The ensemble is detailed in Algorithm 2.

Usage

```
icwa_ensemble(X)
```

Arguments

`X`	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.

Value

The ensemble scores.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- icwa_ensemble(Y)
ens
```

`irt_ensemble` *Computes an ensemble score using Item Response Theory*

Description

This function computes an ensemble score using Item Response Theory (IRT). This was proposed as an ensemble method for anomaly/outlier detection in Kandanaarachchi (2021) <doi:10.13140/RG.2.2.18355.96801>.

Usage

```
irt_ensemble(X)
```

Arguments

`X`	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.

Details

For outlier detection, higher ensemble scores indicate higher levels of anomalousness. This ensemble uses IRT's latent trait to uncover the hidden ground truth, which is used as the ensemble score. It uses the R packages airt and EstCRM to fit the IRT models. It can also be used for other ensembling tasks.

Value

A list with the components:

`scores`	The ensemble scores.
`model`	The IRT model.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- irt_ensemble(Y)
ens$scores
```

`max_ensemble` *Computes an ensemble score using the maximum score of each observation*

Description

This function computes an ensemble score using the maximum score for each observation as detailed in Aggarwal and Sathe (2015) <doi:10.1145/2830544.2830549>.

Usage

```
max_ensemble(X)
```

Arguments

`X`	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.

Value

The ensemble scores.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- max_ensemble(Y)
ens
```

threshold_ensemble *Computes an ensemble score by aggregating values above the mean*

Description

This function computes an ensemble score by aggregating values above the mean as detailed in Aggarwal and Sathe (2015) <doi:10.1145/2830544.2830549>.

Usage

```
threshold_ensemble(X)
```

Arguments

X	The input data containing the outlier scores in a dataframe, matrix or tibble format. Rows contain observations and columns contain outlier detection methods.

Value

The ensemble scores.

Examples

```
set.seed(123)
X <- data.frame(x1 = rnorm(200), x2 = rnorm(200))
X[199, ] <- c(4, 4)
X[200, ] <- c(-3, 5)
y1 <- DDoutlier::KNN_AGG(X)
y2 <- DDoutlier::LOF(X)
y3 <- DDoutlier::COF(X)
y4 <- DDoutlier::INFLO(X)
y5 <- DDoutlier::KDEOS(X)
y6 <- DDoutlier::LDF(X)
y7 <- DDoutlier::LDOF(X)
Y <- cbind.data.frame(y1, y2, y3, y4, y5, y6, y7)
ens <- threshold_ensemble(Y)
ens
```

Index