Package ‘SONO’

March 22, 2025

Title Scores of Nominal Outlyingness (SONO)

Version 1.1

Description

Computes scores of outlyingness for data sets consisting of nominal variables and includes various evaluation metrics for assessing performance of outlier identification algorithms producing scores of outlyingness. The scores of nominal outlyingness are computed based on the framework of Costa and Papatsouma (2025) <doi:10.48550/arXiv.2408.07463>.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.2

Imports data.table (>= 1.14.0), DescTools (>= 0.99.0), ggplot2 (>= 3.3.5), Rdpack (>= 2.0), rje (>= 0.9)

RdMacros Rdpack

NeedsCompilation no

Author Efthymios Costa [aut, cre]

Maintainer Efthymios Costa <efthymios.costa17@imperial.ac.uk>

Repository CRAN

Date/Publication 2025-03-22 11:50:07 UTC

Contents

`avg_rank_outs` *Average Rank Of Outlier*

Description

Function computing the average rank of the outliers given a vector with scores of outlyingness.

Usage

```
avg_rank_outs(scores, outs, ties = ”min”)
```

Arguments

`scores` Scores of (nominal) outlyingness. A higher score here implies an observation is more likely to be an outlier.

`outs` Vector of outlier indices.

`ties` A character string specifying how ties in scores are treated; can be one of "average", "first", "last", "random", "max" or "min".

Value

Average rank of outliers.

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
sono_out <- sono(data = dt, probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)),
alpha = 0.01, r = 2, MAXLEN = 0, frequent = FALSE)
# Suppose observations 1 up to 5 are outliers
avg_rank_outs(scores = sono_out[[2]][, 2], outs = c(1:5), ties = ”min”)
```

`MAXLEN_est` *Estimate MAXLEN*

Description

Function estimating the value of MAXLEN (stopping criterion) prior to running the SONO algorithm. The estimation is done using the ideas described in Costa and Papatsouma (2025), using simultaneous confidence intervals for Multinomial proportions, as done by Sison and Glaz (1995).

Usage

```
MAXLEN_est(data, probs, alpha = 0.01, frequent = FALSE)
```

Arguments

`data` Dataset; needs to be of class data.frame and consist of factor variables only.

`probs` List of probability vectors for each variable. Each element of the list must include as many probabilities as the number of levels associated with it in the dataset.

`alpha` Significance level for the simultaneous Multinomial confidence intervals constructed, determining what the frequency thresholds should be for itemsets of different length, used for outlier detection for discrete features. Must be a positive real, at most equal to 0.50. A greater value leads to a much more conservative algorithm. Default value is 0.01.

`frequent` Logical determining whether highly frequent or highly infrequent itemsets are considered as outliers. Defaults to FALSE, treating highly infrequent itemsets as outlying.

Value

Estimated MAXLEN value.

References

Costa E, Papatsouma I (2025). “A novel framework for quantifying nominal outlyingness.” doi:10.48550/arXiv.2408.07463, arXiv:2408.07463, `http://arxiv.org/abs/2408.07463`.

Sison CP, Glaz J (1995). “Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions.” *Journal of the American Statistical Association*, **90**(429), 366–369. ISSN 0162-1459, doi:10.2307/2291162.

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
MAXLEN_est(data = dt, probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)), alpha = 0.01, frequent = FALSE)
```

recall_at_k *Recall@K*

Description

Function computing the recall based on the top K% scores of outlyingness.

Usage

```
recall_at_k(scores, outs, grid)
```

Arguments

scores	Scores of (nominal) outlyingness. A higher score here implies an observation is more likely to be an outlier.
outs	Vector of outlier indices.
grid	Grid of K values over which the recall is computed. Must be between 0 and 1.

Value

Recall values at the points of the provided grid.

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
sono_out <- sono(data = dt,
probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)),
alpha = 0.01,
r = 2,
MAXLEN = 0,
frequent = FALSE)
# Suppose observations 1 up to 5 are outliers
recall_at_k(scores = sono_out[[2]][, 2],
outs = c(1:5),
grid = c(1, 2.5, seq(5, 50, by = 5))/100)
```

`roc_auc` *ROC AUC function*

Description

Function computing the ROC AUC given a vector with scores of outlyingness. The computation for this is based on Hanley and McNeil (1982).

Usage

```
roc_auc(scores, outs, grid)
```

Arguments

`scores` Scores of (nominal) outlyingness. A higher score here implies an observation is more likely to be an outlier.

`outs` Vector of outlier indices.

`grid` Grid of Top K values over which the ROC AUC is computed. Must be between 0 and 1.

Value

ROC AUC at the points of the provided grid.

References

Hanley JA, McNeil BJ (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” *Radiology*, **143**(1), 29–36. ISSN 0033-8419, doi:10.1148/radiology.143.1.7063747.

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
sono_out <- sono(data = dt,
probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)),
alpha = 0.01,
r = 2,
MAXLEN = 0,
frequent = FALSE)
# Suppose observations 1 up to 5 are outliers
roc_auc(scores = sono_out[[2]][, 2], outs = c(1:5),
grid = c(1, 2.5, seq(5, 50, by = 5))/100)
```

sono *SONO (Scores Of Nominal Outlyingness)*

Description

Function used to compute scores of nominal outlyingness for datasets consisting of nominal features. The computation is done using the score of Costa and Papatsouma (2025), defined as follows for an observation $\boldsymbol{x}_i$:

$$s(\boldsymbol{x}_i) = \sum_{\substack{d \subseteq \boldsymbol{x}_i: \\ \text{supp}(d) \notin (\sigma_d, n], \\ |d| \leq \text{MAXLEN}}} \frac{\sigma_d}{\text{supp}(d) \times |d|^r}, r > 0,\ i = 1, \ldots, n,$$

for highly infrequent itemsets and:

$$s(\boldsymbol{x}_i) = \sum_{\substack{d \subseteq \boldsymbol{x}_i: \\ \text{supp}(d) \notin [0, \sigma_d), \\ |d| \leq \text{MAXLEN}}} \frac{\text{supp}(d)}{\sigma_d \times (\text{MAXLEN} - |d| + 1)^r}, r > 0,\ i = 1, \ldots, n,$$

for highly frequent itemsets. In the above, $\text{supp}(d)$ is the support of itemset d, σ_d is the the maximum/minimum support threshold and MAXLEN is the maximum length of sequences considered, while r is an exponent term to be determined by the user.

Usage

```
sono(
  data,
  probs,
  alpha = 0.01,
  r = 2,
  MAXLEN = 0,
  frequent = FALSE,
  verbose = TRUE
)
```

Arguments

`data` Dataset; needs to be of class data.frame and consist of factor variables only.

`probs` List of probability vectors for each variable. Each element of the list must include as many probabilities as the number of levels associated with it in the dataset.

`alpha` Significance level for the simultaneous Multinomial confidence intervals constructed, determining what the frequency thresholds should be for itemsets of different length, used for outlier detection for discrete features. Must be a positive real, at most equal to 0.50. A greater value leads to a much more conservative algorithm. Default value is 0.01.

`r` Exponent term in the computation of scores. Must be a non-negative number. The greater its value, the less contribution itemsets of greater length will have in the overall score. It is suggested that this is not much larger than 3. Default value is 2.

`MAXLEN` Maximum itemset sequence length to be considered. Default value is 0 which calculates MAXLEN according to a criterion on the sparsity caused by the total combinations that can be encountered as sequences of greater length are taken into account. Otherwise, MAXLEN can take any value from 1 up to the total number of discrete variables included in the data set. If user-given MAXLEN is larger than the estimated value, MAXLEN will default to the latter and a warning message will be displayed, so that redunand computations are avoided.

`frequent` Logical determining whether highly frequent or highly infrequent itemsets are considered as outliers. Defaults to FALSE, treating highly infrequent itemsets are outlying.

`verbose` Defaults to TRUE to print progress messages. Change to FALSE to suppress.

Value

A list with 4 elements. The first element is the value of MAXLEN. The second element corresponds to a data frame with 2 columns; one for the observation numbers and one with the final score of outlyingness. The third and fourth elements are the matrix of variable contributions and the nominal outlyingness depths vector, respectively.

References

Costa E, Papatsouma I (2025). "A novel framework for quantifying nominal outlyingness." doi:10.48550/arXiv.2408.07463, arXiv:2408.07463, `http://arxiv.org/abs/2408.07463`.

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
sono(data = dt,
probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)),
alpha = 0.01,
r = 2,
MAXLEN = 0,
frequent = FALSE)
```

`vis_contribs` *Visualise Contribution Matrix*

Description

Function producing a visualisation of the matrix of variable contributions. The user can choose to plot just a subset of the data (for instance the outliers), as well as scale the scores if needed.

Arguments

`contribs_mat` Matrix of variable contributions. Must be of class `data.frame` with as many columns as the number of variables and rows representing the observations.

`subset` Subset of observations for which the variable contribution matrix will be plotted.

`scale` Optional scaling parameter; defaults to "none" for no scaling. Possible options are "row", with each row being divided by its sum or "max", where each element of the matrix is divided by the maximum element.

Value

Plot

Examples

```
dt <- as.data.frame(sample(c(1:2), 100, replace = TRUE, prob = c(0.5, 0.5)))
dt <- cbind(dt, sample(c(1:3), 100, replace = TRUE, prob = c(0.5, 0.3, 0.2)))
dt[, 1] <- as.factor(dt[, 1])
dt[, 2] <- as.factor(dt[, 2])
colnames(dt) <- c('V1', 'V2')
sono_out <- sono(data = dt,
probs = list(c(0.5, 0.5), c(1/3, 1/3, 1/3)),
alpha = 0.01,
r = 2,
MAXLEN = 0,
frequent = FALSE)
vis_contribs(contribs_mat = sono_out[[3]], subset = c(1:50), scale = ”row”)
```

Index