Package ‘hetcorFS’

October 28, 2025

Type Package

Title Unsupervised Feature Selection using the Heterogeneous Correlation Matrix

Version 1.0

Description Unsupervised multivariate filter feature selection using the UFS-rHCM or UFS-cHCM algorithms based on the heterogeneous correlation matrix (HCM). The HCM consists of Pearson's correlations between numerical features, polyserial correlations between numerical and ordinal features, and polychoric correlations between ordinal features. Tortora C., Madhvani S., Punzo A. (2025). ``Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix.'' International Statistical Review. Forthcoming.

License GPL-2

Imports polycor, dplyr, cluster, graphics,psych

Depends R (>= 3.5.0)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

NeedsCompilation no

Author Cristina Tortora [aut, cre, fnd],
Antonio Punzo [aut],
Shaam Madhvani [aut]

Maintainer Cristina Tortora <grikris1@gmail.com>

Repository CRAN

Date/Publication 2025-10-28 08:40:02 UTC

Contents

ESI 2
FS_barplot 3
HCPM 4
JaccardRate 4
RedRate 5
UFS 6

Index 8

`ESI` *Employee Satisfaction Index (ESI) Data Set*

Description

The Employee Satisfaction Index (ESI) data set, from Kaggle (Harris, 2023), is a fictional data set that measures employee satisfaction

Usage

```
data(ESI)
```

Format

A data frame with 500 rows and 10 features.

emp_id label.

age continuous from 23 to 45.

Dept categorical.

location binary.

education binary.

recruitment_type categorical.

job_level ordinal from 1 to 5.

rating ordinal from 1 to 5.

onsite binary.

awards number of awards 0-9.

certifications binary.

salary continuous from 24.1 to 86.8.

satisfied binary.

Source

Harris, M. (2023). Employee Satisfaction Index Dataset. Evanston, Illinois: Kaggle. Version 1

FS_barplot *Feature importance bar plot*

Description

Displays retained features for different values of alpha in a bar plot.

Usage

```
FS_barplot(
  data = NULL,
  grid.alpha = seq(0.01, 0.99, by = 0.01),
  missing = FALSE,
  pv_adj = "none",
  smooth.tol = 10^-12,
  method = "c"
)
```

Arguments

`data` A data frame. Values of type 'numeric' or 'integer' are treated as numerical.

`grid.alpha` A vector of alpha values to be plotted, default = seq(0.01,0.99,by=0.01).

`missing` Pairwise complete by default, set to TRUE for complete deletion.

`pv_adj` Correction method for p-value, "none" by default. For options see p.adjust.

`smooth.tol` Minimum acceptable eigenvalue for the smoothing, default 10^-12.

`method` Algorithm used. c (cell-wise) by default, r (row-wise) as the alternative.

Value

Displays a bar plot depicting which features are selected at each value of alpha (multiplied by 100) and a list with elements:

`survivors` Vector depicting how many alphas a variable is selected for

`data_names` Vector depicting the corresponding names of the features

References

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. International Statistical Review. https://doi.org/10.1111/insr.70016

Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)]##removing categorical features
FS_barplot(data, pv_adj='BH') #using BH adkustment for the p-values
```

HCPM *Heterogeneous correlation and p-value matrices*

Description

Extends the traditional correlation matrix (between numerical data) to also include binary and ordinal categorical data and computes the p-values for the tests of uncorrelation.

Usage

```
HCPM(data = NULL)
```

Arguments

`data`	A data frame. Values of type 'numeric' or 'integer' are treated as numerical.

Value

A list with with elements:

`cor_mat`	An p by p heterogeneous correlation matrix
`p_value`	An p by p heterogeneous p-values matrix

References

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. International Statistical Review. https://doi.org/10.1111/insr.70016

Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)]##removing categorical features
HCPM(data)
```

JaccardRate *Jaccard Rate*

Description

Computes the Jaccard index using Gower's dissimilarity.

Usage

```
JaccardRate(
  data,
  data_red,
  k=6
)
```

Arguments

data	A data frame. Values of type 'numeric' or 'integer' are treated as numerical.
data_red	A data frame. A subset of data with the selected features.
k	number of neighbors

Value

Jaccard Index	numeric

References

Zhao, Z., L. Wang, and H. Liu (2010). Efficient spectral feature selection with minimum redundancy. In Proceedings of the AAAI conference on artificial intelligence, Volume 24, pp. 673–678.

Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)] ##removing categorical features
out=UFS(data,alpha=0.01,method='c',pv_adj='BH')
JR=JaccardRate(data,out$selected.features)
JR #visualize the index
```

RedRate *Redundancy Rate*

Description

Computes the Redundancy Rate using heterogeneous correlation matrix.

Usage

```
RedRate(
  data_red
)
```

Arguments

data_red	A data frame. A subset of data with the selected features.

Value

`Redundancy Rate` numeric

References

Zhao, Z., L. Wang, and H. Liu (2010). Efficient spectral feature selection with minimum redundancy. In Proceedings of the AAAI conference on artificial intelligence, Volume 24, pp. 673–678.

Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)] ##removing categorical features
out=UFS(data,alpha=0.01,method='c',pv_adj='BH')
RR=RedRate(out$selected.features)
RR #visualize the index
```

UFS *Unsupervised Feature Selection*

Description

Performs unsupervised feature selection for mixed type data. Both algorithms are based on the heterogeneous correlation matrix.

Usage

```
UFS(
  data = NULL,
  alpha = 0.05,
  missing = FALSE,
  pv_adj = ”none”,
  smooth.tol = 10^-12,
  method = ”c”
)
```

Arguments

`data` A data frame. Values of type ’numeric’ or ’integer’ are treated as numerical, factors as ordinal categorical.

`alpha` Significance level to be used for testing, default = 0.05.

`missing` Pairwise complete by default, set to TRUE for complete deletion.

`pv_adj` Correction method for p-value, "none" by default. For options see p.adjust.

`smooth.tol` Minimum acceptable eigenvalue for the smoothing, default = 10^-12.

`method` Algorithm used. c (cell-wise) by default, r (row-wise) as the alternative.

Value

An list of elements:

`rearranged.data.set`
: Original data frame with with numerical features first

`selected.features`
: A data frame of the selected features

`feature.indices`
: The indices of the selected features from the original data frame

`original.corr.matrix`
: The p by p extended correlation matrix of all the inputted features

`corr.matrix`
: The d by d extended correlation matrix of the selected features

`original.p.value.matrix`
: The p by p p-values matrix of all the inputted features

`p.value.matrix`
: The d by d p-values matrix of the selected features

References

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. International Statistical Review. https://doi.org/10.1111/insr.70016

Examples

```
data(ESI)#Loading the data
data = ESI[,-c(1,3,4,6,9)]##removing categorical features
res = UFS(data)

### visualize selected features
colnames(res$selected.features)
```

Index

∗ **datasets**
ESI, 2

ESI, 2

FS_barplot, 3

HCPM, 4

JaccardRate, 4

RedRate, 5

UFS, 6