Package ‘nomclust’

June 6, 2017

Title Hierarchical Nominal Clustering Package
Author Zdenek Sulc [aut, cre],
Hana Rezankova [aut]
Maintainer Zdenek Sulc <zdenek.sulc@vse.cz>
Version 1.1.1106
Date 2017-6-6
Description Package for hierarchical clustering of objects characterized by
nominal variables.
Imports cluster, dummies
License GPL (>= 2)
RoxygenNote 6.0.1
NeedsCompilation no
Repository CRAN
Date/Publication 2017-06-06 14:51:41 UTC

R topics documented:

data20 ........................................ 2
eskin ...................................... 2
evalclust .................................. 3
good1 .................................... 4
good2 .................................... 6
good3 .................................... 7
good4 .................................... 8
iof ........................................ 9
lin ......................................... 11
lin1 ....................................... 12
morlini ................................... 13
nomclust ................................ 15
nomprox .................................. 16
of ......................................... 17
sm ......................................... 18


**data20**

*Artificial nominal dataset*

---

### Description

This dataset consists of 5 nominal variables and 20 cases. Its main aim is to demonstrate the desired entry data structure for the nomclust package.

### Usage

```
data(data20)
```

### Format

A data frame containing 5 variables and 20 cases.

### Source

created by the authors of the nomclust package

---

**eskin**

*Eskin Measure*

---

### Description

The Eskin similarity measure was proposed by Eskin et al. (2002). It is constructed to assign higher weights to mismatches on variables with more categories, see (Boriah et al., 2008). Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $D$ is computed from similarity $S$ according the equation $1/S-1$.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc and Rezankova, 2014).

### Usage

```
eskin(data)
```

### Arguments

- **data**
  
  data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.
Value

Function returns a matrix of the size $n \times n$, where $n$ is a number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in `agnes`.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

`good1, good2, good3, good4, iof, lin, lin1, morlini, of, sm, ve, vm`.

Examples

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_eskin <- eskin(data20)
```

description

The function evaluates clustering results no matter which clustering method they were obtained by. The clusters are evaluated from a point of view of the within-cluster variability by the following indices: Within-cluster mutability coefficient (WCM), Within-cluster entropy coefficient (WCE), Pseudo tau coefficient (PSTau), Pseudo uncertainty coefficient (PSU) and Pseudo F, Indices based on the mutability (PSFM) and the entropy (PSFE).
Usage

```r
evalclust(data, num_var, clu_low = 2, clu_high = 6)
```

Arguments

data: data frame or matrix with cases in rows and variables in columns. First `m1` variables are the original data used for clustering, the next `m2` variables express the cluster memberships in an increasing way (e.g. from `clu_2` to `clu_6`).

num_var: numeric value which determines how many variables in a dataset were used for the clustering.

clu_low: numeric value expressing the lower bound for number of cluster solutions.

clu_high: numeric value expressing the higher bound for number of cluster solutions.

Value

Function returns a data frame, where the rows express a serie of cluster solutions and columns clustering evaluation statistics in a following order: `wcm`, `wce`, `pstau`, `psu`, `psfm`, `psfe`.

See Also

`nomclust`.

Examples

```r
# sample data
data(data20)
# creation of a dataset with cluster memberships
data_clu <- nomclust(data20, iof, clu_high = 7)
# binding an original dataset to cluster memberships variables
data_clu2 <- cbind(data20, data_clu$mem)
# evaluation of created clusters
evaluation <- evalclust(data_clu2, 5, clu_high = 7)
```

Description

The Goodall 1 similarity measure was mentioned e.g. in (Boriah et al., 2008). It is a simple modification of the original Goodall measure (Goodall, 1966). The measure assigns higher similarity to infrequent matches. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity \( D \) is computed from similarity \( S \) according the equation \( 1/S - 1 \).

The use and evaluation of clustering with this measure can be found e.g. in (Sulc, 2015).
### Usage

`good1(data)`

### Arguments

- **data**: data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

### Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in `agnes`.

### Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

### References


### See Also

`eskin`, `good2`, `good3`, `good4`, `iof`, `lin`, `lin1`, `morlini`, `of`, `sm`, `ve`, `vm`.

### Examples

```r
#sample data
data(data20)
# Creation of proximity matrix
prox_goodall_1 <- good1(data20)
```
Goodall 2 Measure

Description

The Goodall 2 similarity measure was firstly introduced in (Boriah et al., 2008). The measure assigns higher similarity to infrequent matches under condition that there are also other categories, which are even less frequent than the examined one. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $d$ is computed from similarity $s$ according the equation $1/s - 1$.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc, 2015).

Usage

good2(data)

Arguments

data data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also
eskin, good1, good3, good4, iof, lin, lin1, morlini, of, sm, ve, vm.

Examples

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_goodall_2 <- good3(data20)
```

## Goodall 3 Measure

### Description

The Goodall 3 similarity measure was firstly introduced in (Boriah et al., 2008). The measure assigns higher similarity if the infrequent categories match regardless on frequencies of other categories. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $d$ is computed from similarity $s$ according the equation $1/s - 1$.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc, 2015).

### Usage

good3(data)

### Arguments

data  data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

### Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

### Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>
References


See Also

eskin, good1, good2, good4, iof, lin, lin1, morlini, of, sm, ve, vm.

Examples

# sample data
data(data20)
# Creation of proximity matrix
prox_goodall_3 <- good3(data20)

good4 Goodall 4 Measure

Description

The Goodall 4 similarity measure was firstly introduced in (Boriah et al., 2008). The measure assigns higher similarity if the frequent categories match. When measuring similarity between two variables, this measure provides complement results of Goodall 3 to one. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $D$ is computed from similarity $S$ according the equation $1/S - 1$.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc, 2015).

Usage

good4(data)

Arguments

data data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.
**Value**

Function returns a matrix of the size \( n \times n \), where \( n \) is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in `agnes`.

**Author(s)**

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

**References**


**See Also**

`eskin, good1, good2, good3, iof, lin, lin1, morlini of, sm, ve, vm`.

**Examples**

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_goodall_4 <- good4(data20)
```

---

### iof

**Inverse Occurrence Frequency (IOF) Measure**

**Description**

The IOF (Inverse Occurrence Frequency) measure was originally constructed for the text mining, see (Sparck-Jones, 1972), later, it was adjusted for categorical variables. The measure assigns higher similarity to mismatches on less frequent values and vice versa. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity \( d \) is computed from similarity \( S \) according the equation \( 1/S - 1 \).
The use and evaluation of clustering with this measure can be found e.g. in (Sulc and Rezankova, 2014).

Usage

iof(data)

Arguments

data data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

eskin, good1, good2, good3, good4, lin, lin1, morlini, of, sm, ve, vm.

Examples

# sample data
data(data28)
# Creation of proximity matrix
prox_iof <- iof(data20)

lin

Lin Measure

Description

The Lin measure was introduced by Lin (1998). The measure assigns higher weights to more frequent categories in case of matches and lower weights to less frequent categories in case of mismatches. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity \( d \) is computed from similarity \( s \) according the equation \( 1/s = 1 \).

The use and evaluation of clustering with this measure can be found e.g. in (Sulc and Rezankova, 2014).

Usage

lin(data)

Arguments

data data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size \( n \times n \), where \( n \) is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in \texttt{agnes}.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

eskin, good1, good2, good3, good4, iof, lin1, morlini, of, sm, ve, vm.

Examples

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_lin <- lin(data20)
```

---

**lin1**  
*Lin 1 Measure*

**Description**

The Lin 1 similarity measure was firstly introduced in (Boriah et al., 2008). It has a complex system of weights. In case of mismatch, lower similarity is assigned if either the mismatching values are very frequent or their relative frequency is in between the relative frequencies of mismatching values. Higher similarity is assigned if the mismatched categories are infrequent and there are a few other infrequent categories. In case of match, lower similarity is given for matches on frequent categories or matches on categories that have many other values of the same frequency. Higher similarity is given to matches on infrequent categories.

Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $d$ is computed from similarity $s$ according the equation $1/s = 1$. After this transformation, it may happen that some values in a proximity matrix get the value $\text{Inf}$. Therefore, the following adjustment is applied: $\max(\text{prox}) + 1$, where $\text{prox}$ is a proximity matrix.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc, 2015).

**Usage**

```r
lin1(data)
```

**Arguments**

- `data`: data frame with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.
Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

eskin, good1, good2, good3, good4, iof, lin, morlini, of, sm, ve, vm.

Examples

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_lin1 <- lin1(data20)
```

morlini Morlini and Zani’s Measure S2

Description

The S2 measure was proposed by Morlini and Zani (2012) and it is based on a transformed dataset, which contains only binary variables (dummy coding). Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $D$ is computed from similarity $S$ according the equation $1/S - 1$.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc and Rezankova, 2014) or (Sulc, 2015).
morlini

Usage

morlini(data)

Arguments

data data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size \( n \times n \), where \( n \) is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

eskin, good1, good2, good3, good4, iof, lin, lin1, of, sm, ve, vm.

Examples

#sample data
data(data20)
# Creation of proximity matrix
prox_morlini <- morlini(data20)
nomclust

Nominal Clustering

Description

The Nominal Clustering (nomclust) performs hierarchical cluster analysis (HCA) with objects characterized by nominal (categorical) variables. It performs a series of cluster solutions, usually from two-cluster solution till six-cluster solution. It allows to choose one from 11 different similarity measures and one from 3 linkage methods. The function also contains an evaluation part. The created clusters are evaluated from a point of view of the within-cluster variability by the following indices: Within-cluster Mutability coefficient ($wcm$), Within-cluster entropy coefficient ($wce$), Pseudo tau coefficient ($pstau$), Pseudo uncertainty coefficient ($psu$) and Pseudo F Indices based on the mutability ($psfm$) and the entropy ($psfe$).

Usage

nomclust(dataL measure = iofL clu_low = 2L clu_high = 6L eval = TRUEL prox = FALSEL method = "complete")

Arguments

data data frame or a matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

measure character string defining the similarity measure which will be used for computation of proximity matrix: "beskin", "good1", "good2", "good3", "good4", "iof", "lin", "lin1", "morlini", "of", "sm".

clu_low numeric value expressing the lower bound for number of cluster solutions.

clu_high numeric value expressing the higher bound for number of cluster solutions.

eval logical operator; if TRUE, there is performed an evaluation of clustering results

prox logical operator; if TRUE, the proximity matrix is a part of the output

method character string defining the clustering method. The following methods can be used: "average", "complete", "single".

Value

Function returns a list following components:

mem data frame consisting of cluster membership variables

eval data frame containing clustering evaluation statistics

prox matrix containing proximities between all combination of pairs of objects (voluntary)

See Also

evalclust, agnes.
nomprox

Examples

```r
# sample data
data(data20)

# Creating an object for nominal clustering
hca <- nomclust(data20, iof, method = "average", clu_high = 5, prox = TRUE)

# Getting evaluation statistics
eval <- hca$eval

# Getting cluster membership variables
mem <- hca$mem

# Getting a proximity matrix
prox <- hca$prox
```

Description

Based on the original dataset and the proximity matrix, the function computes cluster membership variables for a user-defined number of cluster solutions. Optionally, it evaluates clustering results using six evaluation criteria based on the within-cluster variability: Within-cluster mutability coefficient (WCM), Within-cluster entropy coefficient (WCE), Pseudo tau coefficient (PSTau), Pseudo uncertainty coefficient (PSU) and Pseudo F, Indices based on the mutability (PSFM) and the entropy (PSFE).

Usage

```r
nomprox(data = data, prox_matrix = prox_matrix, clu_low = 2, clu_high = 6, eval = TRUE, method = "complete")
```

Arguments

- `data`: data frame or a matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.
- `prox_matrix`: full proximity matrix computed using any similarity measure from the data analyzed.
- `clu_low`: numeric value expressing the lower bound for number of cluster solutions.
- `clu_high`: numeric value expressing the higher bound for number of cluster solutions.
- `eval`: logical operator; if TRUE, there is performed an evaluation of clustering results.
- `method`: character string defining the clustering method. The following methods can be used: "average", "complete", "single".

Value

Function returns a data frame, where the rows express a series of cluster solutions and columns clustering evaluation statistics in a following order: WCM, WCE, PSTau, PSU, PSFM, PSFE.
of Occurrence Frequency (OF) Measure

Description
The OF (Occurrence Frequency) measure was originally constructed for the text mining, see (Sparck-Jones, 1972), later, it was adjusted for categorical variables. It assigns higher similarity to mismatches on less frequent values and otherwise. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $D$ is computed from similarity $S$ according the equation $1/S - 1$.

Usage
of(data)

Arguments
data data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value
Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

See Also
nomclust, evalclust.

Examples
#sample data
data(data20)
#computation of a proximity matrix using the iof similarity measure
matrix <- iof(data20)
#creation of a dataset with cluster memberships
hca <- nomprox(data20, matrix, clu_high = 5, method = "complete")
#getting evaluation statistics
eval <- hca$eval
#getting cluster membership variables
mem <- hca$mem
Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also
eskin, good1, good2, good3, good4, iof, lin, lin1, morlini, sm, ve, vm.

Examples

```r
# sample data
data(dataRPI)
# Creation of proximity matrix
prox_of <- of(dataRPI)
```

---

sm

**Simple Matching Coefficient**

Description

The simple matching coefficient (Sokal, 1958) represents the simplest way for measuring of similarity. It does not impose any weights. By a given variable, it assigns value 1 in case of match and value 0 otherwise. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity D is computed from similarity S according the equation 1/S-1.

The use and evaluation of clustering with this measure can be found e.g. in (Sulc and Rezankova, 2014) or (Sulc, 2015).

Usage

```r
sm(data)
```
Arguments

data       data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References


See Also

eskin, good1, good2, good3, good4, iof, lin, lin1, morlini, of, ve, vm.

Examples

# sample data
data(data20)
# Creation of proximity matrix
prox_sm <- sm(data20)
Variable Entropy measure

Description
The Variable Entropy similarity measure was introduced in (Sulc and Rezankova, 2015). It treats similarity between two categories according to within-cluster variability expressed by the entropy. The novel similarity measures praise more the match of two categories in a variable with high variability, because it is rarer, than the match in a low-variability variable. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $d$ is computed from similarity $s$ according the equation $1/s - 1$.

Usage
ve(data)

Arguments
data data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

Value
Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

Author(s)
Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

References

See Also
eskin, good1, good2, good3, good4, iof, lin, lin1, morlini, of, sm, vm.
Examples

```r
# sample data
data(data20)
# Creation of proximity matrix
prox_ve <- ve(data20)
```

**Description**

The Variable Mutability similarity measure was introduced in (Sulc and Rezankova, 2015). It treats similarity between two categories according to within-cluster variability expressed by the Gini coefficient (mutability). The novel similarity measures praise more the match of two categories in a variable with high variability, because it is rarer, than the match in a low-variability variable. Hierarchical clustering methods require a proximity (dissimilarity) matrix instead of a similarity matrix as an entry for the analysis; therefore, dissimilarity $d$ is computed from similarity $S$ according the equation $1/S + 1$.

**Usage**

```r
vm(data)
```

**Arguments**

data frame or matrix with cases in rows and variables in columns. Cases are characterized by nominal (categorical) variables coded as numbers.

**Value**

Function returns a matrix of the size $n \times n$, where $n$ is the number of objects in original data. The matrix contains proximities between all pairs of objects. It can be used in hierarchical cluster analyses (HCA), e.g. in agnes.

**Author(s)**

Zdenek Sulc.
Contact: <zdenek.sulc@vse.cz>

**References**

See Also

eskin, good1, good2, good3, good4, iof, lin, lin1, morlini, of, sm, ve.

Examples

#sample data
data(data20)
# Creation of proximity matrix
prox_vm <- vm(data20)
Index

*Topic **datasets**

data20, 2

agnes, 3, 5–7, 9–11, 13–15, 17, 19–21

data20, 2

eskin, 2, 5, 7–10, 12–14, 18–20, 22
evalclust, 3, 15, 17

good1, 3, 4, 7–10, 12–14, 18–20, 22
good2, 3, 5, 6, 8–10, 12–14, 18–20, 22
good3, 3, 5, 7, 7, 9, 10, 12–14, 18–20, 22
good4, 3, 5, 7, 8, 8, 10, 12–14, 18–20, 22

iof, 3, 5, 7–9, 9, 12–14, 18–20, 22

lin, 3, 5, 7–10, 11, 13, 14, 18–20, 22
lin1, 3, 5, 7–10, 12, 12, 14, 18–20, 22

morlini, 3, 5, 7–10, 12, 13, 13, 18–20, 22

nomclust, 4, 15, 17
nomprox, 16

of, 3, 5, 7–10, 12–14, 17, 19, 20, 22

sm, 3, 5, 7–10, 12–14, 18, 18, 20, 22

ve, 3, 5, 7–10, 12–14, 18, 19, 20, 22
vm, 3, 5, 7–10, 12–14, 18–20, 21