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` *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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L. and Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In D. Barbara and S. Jajodia (Eds): Applications of Data Mining in Computer Security, p. 78-100. Norwell: Kluwer Academic Publishers.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

See Also

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

Examples

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

`evalclust` *Evaluation of the Clustering*

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

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

Arguments

data	data frame or matrix with cases in rows and variables in colums. 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

```
#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)
```

---

good1 *Goodall 1 Measure*

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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Goodall, V.D. (1966). A new similarity index based on probability. Biometrics, 22(4), p. 882.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

See Also

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

Examples

```
#sample data
data(data20)
# Creation of proximity matrix
prox_goodall_1 <- good1(data20)
```

good2 *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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Goodall, V.D. (1966). A new similarity index based on probability. Biometrics, 22(4), p. 882.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

See Also

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

Examples

```
#sample data
data(data20)
# Creation of proximity matrix
prox_goodall_2 <- good2(data20)
```

---

good3 *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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Goodall, V.D. (1966). A new similarity index based on probability. Biometrics, 22(4), p. 882.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

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 ssigns 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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Goodall, V.D. (1966). A new similarity index based on probability. Biometrics. Vol. 22, No.4, p. 882.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

See Also

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

Examples

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

---

iof *Inverse Occurence 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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Spark-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, 28(1), 11-21. Later: Journal of Documentation, 60(5) (2002), 493-502.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

See Also

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

Examples

```
#sample data
data(data20)
# 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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Lin, D. (1998). An information-theoretic definition of similarity. In: ICML '98: Proceedings of the 15th International Conference on Machine Learning. San Francisco, p. 296-304.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

See Also

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

Examples

```
#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). In 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 `-Inf`. Therefore, the following adjustment is applied: `max(prox)+1`, where `prox` is a proximity matrix.

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

Usage

```
lin1(data)
```

Arguments

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

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

See Also

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

Examples

```
#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).

Usage

```
morlini(data)
```

Arguments

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

Value

Function returns a matrix of the size n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Morlini, I., Zani, S. (2012). A new class of weighted similarity indices using polytomous variables. In Journal of Classification, 29(2), p. 199-226.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

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 serie 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(data, measure = iof, clu_low = 2, clu_high = 6, eval = TRUE,
  prox = FALSE, method = "complete")
```

Arguments

data	data frame or a matrix with cases in rows and variables in colums. Cases are characterized by nominal (categorical) variables coded as numbers.
measure	character string defining the similarity measure which wil be used for computation of proximity matrix: "eskin", "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.

Examples

```
#sample data
data(data20)
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
```

---

nomprox *Nominal Clustering based on a Proximity Matrix*

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

```
nomprox(data, 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 colums. 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 serie of cluster solutions and columns clustering evaluation statistics in a following order: `WCM`, `WCE`, `PSTau`, `PSU`, `PSFM`, `PSFE`.

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
```

---

of *Occurence 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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Spark-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, 28(1), p. 11-21. Later: Journal of Documentation, 60(5) (2002), p. 493-502.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

See Also

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

Examples

```
#sample data
data(data20)
# Creation of proximity matrix
prox_of <- of(data20)
```

---

sm *Simple Matching Coefficient*

Description

The simple matching coefficient (Sokal, 1958) represents the simplest way for measuring of similarity. It does not impose any weigts. 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

```
sm(data)
```

Arguments

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

Value

Function returns a matrix of the size n x 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

Boriah, S., Chandola and V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, p. 243-254.

Sokal, R., Michener, C. (1958). A statistical method for evaluating systematic relationships. In: Science bulletin, 38(22), The University of Kansas.

Sulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In Sbornik praci vedeckeho seminare doktorskeho studia FIS VSE. Praha: Oeconomica, 2015, p. 112-118. Available at: http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

Sulc, Z. and Rezankova, H. (2014). Evaluation of recent similarity measures for categorical data. In: AMSE. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, p. 249-258. Available at: http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.

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)
```

ve *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 colums. Cases are characterized by nominal (categorical) variables coded as numbers.

Value

Function returns a matrix of the size `n x 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

Sulc, Z. and Rezankova H. (2015). Novel similarity measures for categorical data based on mutability and entropy. Conference of the International Federation of Classification Societies. Bologna: Ospitalia, p. 209.

See Also

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

Examples

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

---

vm *Variable Mutability measure*

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

```
vm(data)
```

Arguments

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

Value

Function returns a matrix of the size `n x 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

Sulc, Z. and Rezankova H. (2015). Novel similarity measures for categorical data based on mutability and entropy. Conference of the International Federation of Classification Societies. Bologna: Ospitalia, p. 209.

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