Package 'UniversalCVI'

January 27, 2025

Type Package

Title Hard and Soft Cluster Validity Indices

Version 1.2.0

Imports e1071, mclust

Description Algorithms for checking the accuracy of a clustering result with known classes, computing cluster validity indices, and generating plots for comparing them.
The package is compatible with K-means, fuzzy C means, EM clustering, and hierarchical clustering (single, average, and complete linkage).
The details of the indices in this package can be found in:
J. C. Bezdek, M. Moshtaghi, T. Runkler, C. Leckie (2016) <doi:10.1109/TFUZZ.2016.2540063>,
T. Calinski, J. Harabasz (1974) <doi:10.1080/03610927408827101>,
C. H. Chou, M. C. Su, E. Lai (2004) <doi:10.1007/s10044-004-0218-1>,
D. L. Davies, D. W. Bouldin (1979) <doi:10.1109/TPAMI.1979.4766909>,
J. C. Dunn (1973) <doi:10.1080/01969727308546046>,
F. Haouas, Z. Ben Dhiaf, A. Hammouda, B. Solaiman (2017) <doi:10.1109/FUZZ-IEEE.2017.8015651>,
M. Kim, R. S. Ramakrishna (2005) <doi:10.1016/j.patrec.2005.04.007>,
S. H. Kwon (1998) <doi:10.1049/EL:19981523>,
S. H. Kwon, J. Kim, S. H. Son (2021) <doi:10.1049/ell2.12249>,
G. W. Miligan (1980) <doi:10.1007/BF02293907>,
M. K. Pakhira, S. Bandyopadhyay, U. Maulik (2004) <doi:10.1016/j.patcog.2003.06.005>,
M. Popescu, J. C. Bezdek, T. C. Havens, J. M. Keller (2013) <doi:10.1109/TSMCB.2012.2205679>,
S. Saitta, B. Raphael, I. Smith (2007) <doi:10.1007/978-3-540-73499-4_14>,
A. Starczewski (2017) <doi:10.1007/s10044-015-0525-8>,
Y. Tang, F. Sun, Z. Sun (2005) <doi:10.1109/ACC.2005.1470111>,
N. Wiroonsri (2024) <doi:10.1016/j.patcog.2023.109910>,
N. Wiroonsri, O. Preedasawakul (2023) <doi:10.48550/arXiv.2308.14785>,
C. H. Wu, C. S. Ouyang, L. W. Chen, L. W. Lu (2015) <doi:10.1109/TFUZZ.2014.2322495>,
X. Xie, G. Beni (1991) <doi:10.1109/34.85677> and
Rousseeuw (1987) and Kaufman and Rousseeuw(2009) <doi:10.1016/0377-0427(87)90125-7> and
<doi:10.1002/9780470316801>
C. Alok. (2010).

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Depends R (>= 2.10)

NeedsCompilation no

Author Nathakhun Wiroonsri [cre, aut]
(<https://orcid.org/0000-0003-2167-9641>),
Onthada Preedasawakul [aut] (<https://orcid.org/0000-0002-4186-3158>)

Maintainer Nathakhun Wiroonsri <nathakhun.wir@kmutt.ac.th>

Repository CRAN

Date/Publication 2025-01-27 16:10:05 UTC

Contents

`AccClust` *Accuracy detection for a clustering result with known classes*

Description

Computes the accuracy of a clustering result of a dataset with known classes from the k-means, fuzzy c-means, or EM algorithm.

Usage

```
AccClust(x, label.names = "label", algorithm = "FCM", fzm = 2,
  scale = TRUE, nstart = 100, iter = 100)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`label.names`	a character string indicating the true label column name. The default is `"label"`
`algorithm`	a character string indicating which clustering methods to be used (`"FCM"`, `"EM"`, `"Kmeans"`). More than one methods may be selected. The default is `"FCM"`.
`fzm`	a number greater than 1 giving the degree of fuzzification for `method = "FCM"`. The default is `2`.
`scale`	logical, if `TRUE` (default), the dataset is normalized before clustering.
`nstart`	a maximum number of initial random sets for FCM for `method = "FCM"` or `"Kmeans"` or `c("Kmeans","FCM")`. The default is `20`.
`iter`	a maximum number of iterations for `method = "FCM"`. The default is `100`.

Value

`kmeans`	Accuracy score from `0` to `1` of the k-means result
`FCM`	Accuracy score from `0` to `1` of the FCM result
`EM`	Accuracy score from `0` to `1` of the EM result

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

R1_data, D1_data, FzzyCVIs, WP.IDX, XB.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data

# Check accuracy of clustering results obtained by kmeans, FCM, and EM clustering
AccClust(x, label.names = "label",algorithm = c("Kmeans","FCM","EM"), fzm = 2,
  scale = TRUE, nstart = 20,iter = 100)

# Check accuracy of a clustering result obtained by the FCM algoritm
AccClust(x, label.names = "label",algorithm = "FCM", fzm = 2,
  scale = TRUE, nstart = 20,iter = 100)
```

CCV.IDX *Correlation Cluster Validity (CCV) index*

Description

Computes the CCVP and CCVS (M. Popescu et al., 2013) indexes for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
CCV.IDX(x, cmax, cmin = 2, indexlist = "all", method = 'FCM', fzm = 2,
  iter = 100, nstart = 20)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
indexlist	a character string indicating which The generalized C index be computed ("all","CCVP","CCVS"). More than one indexes can be selected.

method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.

Details

A new cluster validity framework that compares the structure in the data to the structure of dissimilarity matrices induced by a matrix transformation of the partition being tested. The largest value of $CCV(c)$ indicates a valid optimal partition.

Value

Each of the followings shows the values of each index for c from cmin to cmax in a data frame.

CCVP	the Pearson Correlation Cluster Validity index.
CCVS	the Spearman's (rho) Correlation Cluster Validity index.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. Popescu, J. C. Bezdek, T. C. Havens and J. M. Keller (2013). "A Cluster Validity Framework Based on Induced Partition Dissimilarity." https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6246717&isnumber=6340245

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# Iris data
x = iris[,1:4]

# ---- FCM algorithm ----


# Compute all the indices by CCV.IDX
FCM.ALL.CCV = CCV.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "all",
  method =  'FCM', fzm = 2, iter = 100, nstart = 20)
print(FCM.ALL.CCV)
```

```
# Compute CCVP index
FCM.CCVP = CCV.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "CCVP",
  method =  'FCM', fzm = 2, iter = 100, nstart = 20)
print(FCM.CCVP)


# ---- EM algorithm ----

# Compute all the indices by CCV.IDX
EM.ALL.CCV = CCV.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "all",
  method =  'EM', iter = 100, nstart = 20)
print(EM.ALL.CCV)

# Compute CCVP index
EM.CCVP = CCV.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "CCVP",
  method =  'EM', iter = 100, nstart = 20)
print(EM.CCVP)
```

CH.IDX *Calinski–Harabasz (CH) index*

Description

Computes the CH (T. Calinski and J. Harabasz, 1974) index for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
CH.IDX(x, kmax, kmin = 2, method = "kmeans", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The CH index is defined as

$$CH(k) = \frac{n-k}{k-1} \frac{\sum_{i=1}^{k} |C|_i d(v_i, \bar{x})}{\sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, v_i)}$$

The largest value of $CH(k)$ indicates a valid optimal partition.

Value

CH	the CH index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the CH index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

T. Calinski, J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, 3, 1-27 (1974).

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute the CH index
K.CH = CH.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans", nstart = 100)
print(K.CH)

# The optimal number of cluster
K.CH[which.max(K.CH$CH),]

# ---- Hierarchical ----

# Average linkage

# Compute the CH index
H.CH = CH.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average")
print(H.CH)

# The optimal number of cluster
H.CH[which.max(H.CH$CH),]
```

CSL.IDX *Chou-Su-Lai (CSL) index*

Description

Computes the CSL (C. H. Chou et al., 2004) index for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
CSL.IDX(x, kmax, kmin = 2, method = "kmeans", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The CSL index is defined as

$$CSL(k) = \frac{\sum_{i=1}^{k} \left\{ \frac{1}{|C_i|} \sum_{x_j \in C_i} \max_{x_l \in C_i} d(x_j, x_l) \right\}}{\sum_{i=1}^{k} \left\{ \min_{j:j \neq i} d(v_i, v_j) \right\}}.$$

The smallest value of $CSL(k)$ indicates a valid optimal partition.

Value

CSL	the CSL index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the CSL index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. H. Chou, M. C. Su, E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Anal Applic*, 7, 205-220 (2004).

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute the CSL index
K.CSL = CSL.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans", nstart = 100)
print(K.CSL)

# The optimal number of cluster
K.CSL[which.min(K.CSL$CSL),]

# ---- Hierarchical ----

# Average linkage

# Compute the CSL index
H.CSL = CSL.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average")
print(H.CSL)

# The optimal number of cluster
H.CSL[which.min(H.CSL$CSL),]
```

D10_data *D10 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 3 different Gaussian and 2 Uniform distributions labeled as 1-5.

Usage

```
D10_data
```

Format

A data frame with 1250 data points and 3 variables

`x` Numeric values generated from Gaussian and Uniform distributions

`y` Numeric values generated from Gaussian and Uniform distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`D1_data` *D1 Artificial Dataset*

Description

A `2`-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from `6` different Gaussian distributions labeled as `1-6`.

Usage

```
D1_data
```

Format

A data frame with 1500 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

D2_data *D2 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 6 different Gaussian distributions labeled as 1-6.

Usage

```
D2_data
```

Format

A data frame with 1200 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

D3_data *D3 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 4 different Gaussian distributions labeled as 1-4.

Usage

```
D3_data
```

Format

A `data frame` with 1400 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`D4_data` *D4 Artificial Dataset*

Description

A `2`-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from `4` different Gaussian distributions labeled as `1-4`.

Usage

```
D4_data
```

Format

A data frame with 2400 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`D5_data` *D5 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 5 different Gaussian distributions labeled as 1-5.

Usage

```
D5_data
```

Format

A `data frame` with 350 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

D6_data *D6 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 5 different Gaussian distributions labeled as 1-5.

Usage

```
D6_data
```

Format

A data frame with 1100 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

D7_data *D7 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 6 different Gaussian distributions labeled as 1-6.

Usage

```
D7_data
```

Format

A data frame with 1500 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`D8_data` *D8 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 6 different Gaussian distributions labeled as 1-6.

Usage

```
D8_data
```

Format

A data frame with 2000 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`D9_data` *D9 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 3 different Uniform distributions labeled as 1-3.

Usage

```
D9_data
```

Format

A data frame with 1000 data points and 3 variables

`x` Numeric values generated from Uniform distributions

`y` Numeric values generated from Uniform distributions

`label` Categorical labels 1,2,3

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

DB.IDX *Davies–Bouldin (DB) and DB* (DBs) indexes*

Description

Computes the DB (D. L. Davies and D. W. Bouldin, 1979) and DBs (M. Kim and R. S. Ramakrishna, 2005) indexes for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
DB.IDX(x, kmax, kmin = 2, method = "kmeans",
  indexlist = "all", p = 2, q = 2, nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
indexlist	a character string indicating which cluster validity indexes to be computed ("all", "DB", "DBs"). More than one indexes can be selected.
p	the power of the Minkowski distance between centroids of clusters. The default is 2.
q	the power of dispersion measure of a cluster. The default is 2.
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The lowest value of $DB(k), DBs(k)$ indicates a valid optimal partition.

Value

DB	the DB index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the DB index, respectively.
DBs	the DBs index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the DBs index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

D. L. Davies, D. W. Bouldin, "A cluster separation measure," *IEEE Trans Pattern Anal Machine Intell*, 1, 224-227 (1979).

M. Kim, R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognition Letters*, 26, 2353-2363 (2005).

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute all the indices by DB.IDX
K.ALL = DB.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans",
  indexlist = "all", p = 2, q = 2, nstart = 100)
print(K.ALL)

# Compute DB index
K.DB = DB.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans",
  indexlist = "DB", p = 2, q = 2, nstart = 100)
print(K.DB)

# ---- Hierarchical ----

# Average linkage

# Compute all the indices by DB.IDX
H.ALL = DB.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average",
  indexlist = "all", p = 2, q = 2)
print(H.ALL)

# Compute DB index
H.DB = DB.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average",
  indexlist = "DB", p = 2, q = 2)
print(H.DB)
```

DI.IDX *Dunn index*

Description

Computes the DI (J. C. Dunn, 1973) index for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
DI.IDX(x, kmax, kmin = 2, method = "kmeans", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The DI index is defined as

$$DI(k) = \min_{i \neq j \in [k]} \left\{ \frac{\min \{d(x_u, x_v) | x_u \in C_i, x_v \in C_j\}}{\max_{l \in [k]} \max \{d(x_u, x_v) | x_u, x_v \in C_l\}} \right\}.$$

The largest value of $DI(k)$ indicates a valid optimal partition.

Value

DI	the DI index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the DI index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J Cybern*, 3(3), 32-57 (1973).

See Also

Hvalid, Wvalid, DB.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]
```

```
# ---- Kmeans ----

# Compute the DI index
K.DI = DI.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans", nstart = 100)
print(K.DI)

# The optimal number of cluster
K.DI[which.max(K.DI$DI),]

# ---- Hierarchical ----

# Average linkage

# Compute the DI index
H.DI = DI.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average")
print(H.DI)

# The optimal number of cluster
H.DI[which.max(H.DI$DI),]
```

`FzzyCVIs` *Fuzzy cluster validity indexes used in Wiroonsri and Preedasawakul (2023)*

Description

Computes the cluster validity indexes for a result of either FCM or EM clustering from user specified `cmin` to `cmax` used in Wiroonsri and Preedasawakul (2023). It includes the XB (X. L. Xie and G. Beni, 1991) index, KWON (S. H. Kwon, 1998) index, KWON2 (S. H. Kwon et al., 2021) index, TANG (Y. Tang et al., 2005) index , HF (F. Haouas et al., 2017) index, WL (C. H. Wu et al., 2015) index, PBM (M. K. Pakhira et al., 2004) index, KPBM (C. Alok, 2010) index, CCVP and CCVS (M. Popescu et al., 2013) index, GC1, GC2, GC3, and GC4 (J. C. Bezdek et al., 2016) indexes , WPC, WP, WPCI1, and, WPCI2 (N. Wiroonsri and O. Preedasawakul, 2023) indexes.

Usage

```
FzzyCVIs(x, cmax, cmin = 2, indexlist = 'all', corr = 'pearson',
  method = 'FCM', fzm = 2, gamma = (fzm^2*7)/4, sampling = 1,
  iter = 100, nstart = 20, NCstart = TRUE)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`cmax`	a maximum number of clusters to be considered.
`cmin`	a minimum number of clusters to be considered. The default is `2`.

indexlist	a character string indicating which cluster validity indexes to be computed ("all", "WPC", "WP", "WPCI1", "WPCI2", "XB", ""KWON"", ""KWON2"", ""TANG"", ""HF"", "WL", "PBM", "KPBM", "CCVP", "CCVS", "GC1", "GC2", "GC3", "GC4"). More than one indexes can be selected.
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman") for indexlist = ("WP", "WPC", "WPCI1","WPCI2", "CCVP", "CCVS", "GC1", "GC2", "GC3" or "GC4"). The default is "pearson".
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
gamma	adjusted fuzziness parameter for indexlist = ("WP", "WPC", "WPCI1", "WPCI2"). The default is $7fzm^2/4$.
sampling	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is 1.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
NCstart	logical for indexlist includes either of the "WP", "WPC", "WPCI1", and "WPCI2"), if TRUE, the WP correlation at c=1 is defined as the ratio introduced in the reference. Otherwise, it is assigned as 0.

Details

The well-known cluster validity indexes for either FCM or EM clustering. It includes the XB (X. L. Xie and G. Beni., 1991) index, KWON (S. H. Kwon, 1998) index, KWON2 (S. H. Kwon et al., 2021) index, TANG (Y. Tang et al., 2005) index , HF (F. Haouas et al., 2017) index, WL (C. H. Wu et al., 2015) index, PBM (M. K. Pakhira et al., 2004) index, KPBM (C. Alok, 2010) index, CCVP and CCVS (M. Popescu et al., 2013) index, GC1, GC2, GC3, and GC4 (J. C. Bezdek et al., 2016) indexes , WPC, WP, WPCI1, and, WPCI2 (N. Wiroonsri and O. Preedasawakul, 2023) indexes.

The WPC computes the correlation between the actual distance between a pair of data points and the distance between adjusted centroids with respect to the pair. WPCI1 and WPCI2 are the proportion and the subtraction, respectively, of the same two ratios. The first ratio is the WPC improvement from c-1 clusters to c clusters over the entire room for improvement. The second ratio is the WPC improvement from c clusters to c+1 clusters over the entire room for improvement. WP is defined as a combination of WPCI1 and WPCI2.

Value

WPC	the WP correlation from c from cmin-1 to cmax+1 shown in a data frame.

Each of the followings shows the values of each index for c from cmin to cmax in a data frame.

WP	the WP index.
WPCI1	the WPCI1 index.
WPCI2	the WPCI2 index.

XB	the XB index.
KWON	the KWON index.
KWON2	the KWON2 index.
TANG	the TANG index.
HF	the HF index.
WL	the WL index.
PBM	the PBM index
KPBM	the KPBM index
CCVP	the Pearson Correlation Cluster Validity index.
CCVS	the Spearman's (rho) Correlation Cluster Validity index.
GC1	the generalized C index ($\sum \cdot \sim$ Sum-Product).
GC2	the generalized C index ($\sum \wedge \sim$ Sum-Min).
GC3	the generalized C index ($\vee \cdot \sim$ Max-Product).
GC4	the generalized C index ($\vee \wedge \sim$ Max-Min).

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. Alok. (2010). "An investigation of clustering algorithms and soft computing approaches for pattern recognition," Department of Computer Science, Assam University.

J. C. Bezdek, M. Moshtaghi, T. Runkler, C. Leckie, "The generalized c index for internal fuzzy cluster validity," IEEE Transactions on Fuzzy Systems, vol. 24, no. 6, pp. 1500–1512, 2016.

F. Haouas, Z. Ben Dhiaf, A. Hammouda, B. Solaiman, "A new efficient fuzzy cluster validity index: Application to images clustering," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 2017, pp. 1-6.

S. H. Kwon, "Cluster validity index for fuzzy clustering," Electronics letters, vol. 34, no. 22, pp. 2176–2177, 1998.

S. H. Kwon, J. Kim, S. H. Son, "Improved cluster validity index for fuzzy clustering," Electronics Letters, vol. 57, no. 21, pp. 792–794, 2021.

M. K. Pakhira, S. Bandyopadhyay, U. Maulik, "Validity index for crisp and fuzzy clusters," Pattern recognition, vol. 37, no. 3, pp. 487–501, 2004.

M. Popescu, J. C. Bezdek, T. C. Havens, J. M. Keller, "A Cluster Validity Framework Based on Induced Partition Dissimilarity," in IEEE Transactions on Cybernetics, vol. 43, no. 1, pp. 308-320, Feb. 2013.

Y. Tang, F. Sun, Z. Sun, “Improved validation index for fuzzy clustering,” in Proceedings of the 2005, American Control Conference, 2005., pp. 1120–1125 vol. 2, 2005.

N. Wiroonsri, O. Preedasawakul, "A correlation-based fuzzy cluster validity index with secondary options detector," arXiv:2308.14785, 2023

C. H. Wu, C. S. Ouyang, L. W. Chen, L. W. Lu, “A new fuzzy clustering validity index with a median factor for centroid-based clustering,” IEEE Transactions on Fuzzy Systems, vol. 23, no. 3, pp. 701–718, 2015.

X. Xie, G. Beni, “A validity measure for fuzzy clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 841–847, 1991.

See Also

WP.IDX, GC.IDX, CCV.IDX, R1_data

Examples

```
library(UniversalCVI)

# Iris data
x = iris[,1:4]

# ---- FCM algorithm ----


# Compute selected a set of indices ("WPC","WP","XB") using default gamma
F.s = FzzyCVIs(scale(x), cmax = 10, cmin = 2, indexlist = c("WPC","WP","XB"),
  corr = 'pearson', method = 'FCM', fzm = 2, iter = 100, nstart = 20, NCstart = TRUE)

# Plot the computed indexes
plot_idx(F.s)

# ---- EM algorithm ----

# Compute all the indices by FzzyCVIs using default gamma
```

```
E.all = FzzyCVIs(scale(x), cmax = 10, cmin = 2, indexlist = 'all', corr = 'pearson',
  method = 'EM', iter = 100, nstart = 20, NCstart = TRUE)

# Plot the computed indexes
plot_idx(E.all)
```

GC.IDX *The generalized C index*

Description

Computes the GC1 GC2 GC3 and GC4 (J. C. Bezdek et al., 2016) indexes for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
GC.IDX(x, cmax, cmin = 2, indexlist = "all", method = 'FCM', fzm = 2,
  iter = 100, nstart = 20)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
indexlist	a character string indicating which The generalized C index be computed ("all","GC1","GC2","GC3","GC4" More than one indexes can be selected.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.

Details

The GC index is a soft version of the C-index, formulated based on relational transformations of the membership degree matrix μ. It comprises four distinct variants, each with its own definition. The smallest value of $GC(c)$ indicates a valid optimal partition.

Value

Each of the followings shows the values of each index for c from cmin to cmax in a data frame.

GC1	the generalized C index ($\sum \cdot \sim$ Sum-Product).
GC2	the generalized C index ($\sum \wedge \sim$ Sum-Min).
GC3	the generalized C index ($\vee \cdot \sim$ Max-Product).
GC4	the generalized C index ($\vee \wedge \sim$ Max-Min).

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

J. C. Bezdek, M. Moshtaghi, T. Runkler, and C. Leckie, "The generalized c index for internal fuzzy cluster validity," IEEE Transactions on Fuzzy Systems, vol. 24, no. 6, pp. 1500–1512, 2016. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7429723&isnumber=7797168

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# Iris data
x = iris[,1:4]

# ---- FCM algorithm ----

# Compute all the indices by GC.IDX
FCM.all.GC = GC.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "all",
  method =  'FCM', fzm = 2, iter = 100, nstart = 5)
print(FCM.all.GC)

# Compute GC2 index
FCM.GC2 = GC.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "GC2",
  method = 'FCM', fzm = 2, iter = 100, nstart = 5)
print(FCM.GC2)

# ---- EM algorithm ----

# Compute all the indices by GC.IDX
EM.all.GC = GC.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "all",
  method =  'EM', iter = 100, nstart = 5)
print(EM.all.GC)

# Compute GC2 index
EM.GC2 = GC.IDX(scale(x), cmax = 10, cmin = 2, indexlist = "GC2",
```

```
  method =  'EM', iter = 100, nstart = 5)
print(EM.GC2)
```

HF.IDX *HF index*

Description

Computes the HF (F. Haouas et al., 2017) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
HF.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The HF index is defined as

$$HF(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}\|x_i - v_j\|^2 + \frac{1}{c(c-1)}\sum_{j\neq k}\|v_j - v_k\|^2}{\frac{n}{2c}\left(\min_{j\neq k}\{\|v_j - v_k\|^2\} + \text{median}_{j\neq k}\{\|v_j - v_k\|^2\}\right)}.$$

The smallest value of $HF(c)$ indicates a valid optimal partition.

Value

HF	the HF index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the HF index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

F. Haouas, Z. Ben Dhiaf, A. Hammouda and B. Solaiman, "A new efficient fuzzy cluster validity index: Application to images clustering," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 2017, pp. 1-6. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8015651&isnumber=8015374

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the HF index
FCM.HF = HF.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.HF)

# The optimal number of cluster
FCM.HF[which.min(FCM.HF$HF),]

# ---- EM algorithm ----

# Compute the HF index
EM.HF = HF.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.HF)

# The optimal number of cluster
EM.HF[which.min(EM.HF$HF),]
```

Hvalid *Wiroonsri(2024) correlation-based cluster validity indices and other well-known cluster validity indices*

Description

Computes the cluster validity indexes for a result of either kmeans or hierarchical clustering from user specified kmin to kmax used in Wiroonsri(2024). It includes the DI (J. C. Dunn, 1973) index, CH (T. Calinski and J. Harabasz, 1974) index, DB (D. L. Davies and D. W. Bouldin, 1979) index, PB (G. W. Miligan, 1985) index, CSL (C. H. Chou et al., 2004) index, PBM (M. K. Pakhira et al., 2004) index, DBs (M. Kim and R. S. Ramakrishna, 2005), Score function (S. Saitta et al., 2007), STR (A. Starczewski, 2017) index, NC, NCI, NCI1, and, NCI2 (N. Wiroonsri, 2024) indexes.

Usage

```
Hvalid(x, kmax, kmin = 2, indexlist = "all", method = "kmeans",
  p = 2, q = 2, corr = "pearson", nstart = 100, sampling = 1, NCstart = TRUE)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
indexlist	a character string indicating which cluster validity indexes to be computed ("all", "NC", "NCI", "NCI1", "NCI2", "PB", "CSL", "CH", "DB", "DBs", "SF", "DI", "STR", "PBM"). More than one indexes can be selected.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
p	the power of the Minkowski distance between centroids of clusters for indexlist = c("DB","DBs"). The default is 2.
q	the power of dispersion measure of a cluster for indexlist = c("DB","DBs"). The default is 2.
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman"). The default is "pearson".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
sampling	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is 1.
NCstart	logical for indexlist includes the "NC", "NCI", "NCI1", and "NCI2"), if TRUE, the NC correlation at k=1 is defined as the ratio introduced in the reference. Otherwise, it is assigned as 0.

Details

The well-known cluster validity indices used in Wiroonsri(2024). It includes the DI (J. C. Dunn, 1973) index, CH (T. Calinski and J. Harabasz, 1974) index, DB (D. L. Davies and D. W. Bouldin, 1979) index, PB (G. W. Miligan, 1980) index, CSL (C. H. Chou et al., 2004) index, PBM (M. K. Pakhira et al., 2004) index, DBs (M. Kim and R. S. Ramakrishna, 2005), Score function (S. Saitta et al., 2007), STR (A. Starczewski, 2017), NC, NCI, NCI1, and, NCI2 (N. Wiroonsri, 2024) indexes.

The NC correlation computes the correlation between an actual distance between a pair of data points and a centroid distance of clusters that the two points locate in. NCI1 and NCI2 are the proportion and the subtraction, respectively, of the same two ratios. The first ratio is the NC improvement from k-1 clusters to k clusters over the entire room for improvement. The second ratio is the NC improvement from k clusters to k+1 clusters over the entire room for improvement. NCI is a combination of NCI1 and NCI2.

Value

NC	the NC correlations for k from kmin-1 to kmax+1 shown in a data frame where the first and the second columns are k and the NC, respectively.

Each of the followings shows the values of each index for k from kmin to kmax in a data frame.

NCI	the NCI index.
NCI1	the NCI1 index.
NCI2	the NCI2 index.
PB	the PB index.
DI	the DI index.
DB	the DB index.
DBs	the DBs index.
CSL	the CSL index.
CH	the CH index.
SF	the Score function.
STR	the STR index.
PBM	the PBM index.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

J. C. Bezdek, N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics*, Part B, 28, 301-315 (1998).

T. Calinski, J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, 3, 1-27 (1974).

C. H. Chou, M. C. Su, E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Anal Applic*, 7, 205-220 (2004).

D. L. Davies, D. W. Bouldin, "A cluster separation measure," *IEEE Trans Pattern Anal Machine Intell*, 1, 224-227 (1979).

J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J Cybern*, 3(3), 32-57 (1973).

M. Kim, R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognition Letters*, 26, 2353-2363 (2005).

G. W. Miligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, 45, 325-342 (1980).

M. K. Pakhira, S. Bandyopadhyay and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recogn* 37(3):487–501 (2004).

S. Saitta, B. Raphael, I. Smith, "A bounded index for cluster validity," *In Perner, P.: Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science*, 4571, Springer (2007).

A. Starczewski, "A new validity index for crisp clusters," *Pattern Anal Applic* 20, 687–700 (2017).

N. Wiroonsri, "Clustering performance analysis using a new correlation based cluster validity index," Pattern Recognition, 145, 109910, 2024.

See Also

Wvalid, FzzyCVIs, DI.IDX, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]


# ---- Kmeans ----

# Compute all the indices by Hvalid
Hvalid(scale(x), kmax = 15, kmin = 2, indexlist = "all",
  method = "kmeans", p = 2, q = 2, corr = "pearson", nstart = 100, NCstart = TRUE)

# Compute selected a set of indices ("NC","NCI","DI","DB")
Hvalid(scale(x), kmax = 15, kmin = 2, indexlist = c("NC","NCI","DI","DB"),
  method = "kmeans", p = 2, q = 2, corr = "pearson", nstart = 100, NCstart = TRUE)

# ---- Hierarchical ----

# Average linkage

# Compute all the indices by Hvalid
Hvalid(scale(x), kmax = 15, kmin = 2, indexlist = "all",
  method = "hclust_average", p = 2, q = 2, corr = "pearson", nstart = 100, NCstart = TRUE)

# Compute selected a set of indices ("NC","NCI","DI","DB")
Hvalid(scale(x), kmax = 15, kmin = 2, indexlist = c("NC","NCI","DI","DB"),
  method = "hclust_average", p = 2, q = 2, corr = "pearson", nstart = 100, NCstart = TRUE)

#---Plot and compare the indexes---

# Compute six cluster validity indexes of a kmeans clustering result for k from 2 to 15
IDX.list = c("NCI", "DI", "DB", "DBs", "CSL", "CH")

Hvalid.result = Hvalid(scale(x), kmax = 15, kmin = 2, indexlist = IDX.list,
  method = "hclust_average", p = 2, q = 2, corr = "pearson", nstart = 100, NCstart = TRUE)

# Plot the computed indexes
plot_idx(Hvalid.result)
```

KPBM.IDX *Modified Kernel form of Pakhira-Bandyopadhyay-Maulik (KPBM) index*

Description

Computes the KPBM (C. Alok, 2010) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
KPBM.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The KPBM index is defined as

$$KPBM(c) = \left(\frac{\max_{j \neq k} \|v_j - v_k\|}{c \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij} \|x_i - v_j\|} \right)^2 .$$

The largest value of $KPBM(c)$ indicates a valid optimal partition.

Value

KPBM	the KPBM index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the KPBM index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. Alok. (2010). "An investigation of clustering algorithms and soft computing approaches for pattern recognition", Department of Computer Science, Assam University.

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the KPBM index
FCM.KPBM = KPBM.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.KPBM)

# The optimal number of cluster
FCM.KPBM[which.max(FCM.KPBM$KPBM),]

# ---- EM algorithm ----

# Compute the KPBM index
EM.KPBM = KPBM.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.KPBM)

# The optimal number of cluster
EM.KPBM[which.max(EM.KPBM$KPBM),]
```

KWON.IDX *KWON index*

Description

Computes the KWON (S. H. Kwon, 1998) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
KWON.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The KWON index is defined as

$$KWON(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{2}\|x_i - v_j\|^2 + \frac{1}{c}\sum_{j=1}^{c}\|v_j - v_0\|^2}{\min_{i\neq j}\|v_i - v_j\|^2}.$$

The smallest value of $KWON(c)$ indicates a valid optimal partition.

Value

KWON	the KWON index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the KWON index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. H. Kwon, "Cluster validity index for fuzzy clustering," Electronics letters, vol. 34, no. 22, pp. 2176–2177, 1998. doi:10.1049/el:19981523

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]
```

```
# ---- FCM algorithm ----

# Compute the KWON index
FCM.KWON = KWON.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.KWON)
# The optimal number of cluster
FCM.KWON[which.min(FCM.KWON$KWON),]

# ---- EM algorithm ----

# Compute the KWON index
EM.KWON = KWON.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.KWON)
# The optimal number of cluster
EM.KWON[which.min(EM.KWON$KWON),]
```

KWON2.IDX *KWON2 index*

Description

Computes the KWON2 (S. H. Kwon et al., 2021) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
KWON2.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

KWON2 is defined as

$$KWON2(c) = \frac{w_1 \left[w_2 \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{2\sqrt{\frac{m}{2}}} \|x_i - v_j\|^2 + \frac{\sum_{j=1}^{c} \|v_j - v_0\|^2}{\max_j \|v_j - v_0\|^2} + w_3 \right]}{\min_{i \neq j} \|v_i - v_j\|^2 + \frac{1}{c} + \frac{1}{c^m - 1}}.$$

where $w_1 = \frac{n-c+1}{n}$, $w_2 = \left(\frac{c}{c-1}\right)^{\sqrt{2}}$ and $w_3 = \frac{nc}{(n-c+1)^2}$.

The smallest value of $KWON2(c)$ indicates a valid optimal partition.

Value

KWON2	the KWON2 index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the KWON2 index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. H. Kwon, J. Kim, and S. H. Son, "Improved cluster validity index for fuzzy clustering," Electronics Letters, vol. 57, no. 21, pp. 792–794, 2021.

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the KWON2 index
FCM.KWON2 = KWON2.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.KWON2)

# The optimal number of cluster
FCM.KWON2[which.min(FCM.KWON2$KWON2),]

# ---- EM algorithm ----

# Compute the KWON2 index
```

```
EM.KWON2 = KWON2.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.KWON2)

# The optimal number of cluster
EM.KWON2[which.min(EM.KWON2$KWON2),]
```

PB.IDX *Point biserial correlation (PB)*

Description

Computes the PB (G. W. Miligan, 1980) index for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
PB.IDX(x, kmax, kmin = 2, method = "kmeans", corr = "pearson", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman"). The default is "pearson".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The largest value of $PB(k)$ indicates a valid optimal partition.

Value

PB	the PB index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the PB index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

G. W. Miligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, 45, 325-342 (1980).

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute PB index
K.PB = PB.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans",
  corr = "pearson", nstart = 100)
print(K.PB)

# The optimal number of cluster
K.PB[which.max(K.PB$PB),]

# ---- Hierarchical ----

# Average linkage

# Compute PB index
H.PB = PB.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average",
  corr = "pearson")
print(H.PB)

# The optimal number of cluster
H.PB[which.max(H.PB$PB),]
```

PBM.IDX *Pakhira-Bandyopadhyay-Maulik (PBM) index*

Description

Computes the PBM (M. K. Pakhira et al., 2004) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
PBM.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The PBM index is defined as

$$PBM(c) = \left(\frac{\sum_{i=1}^{n} \|x_i - v_0\| \cdot \max_{j \neq k} \|v_j - v_k\|}{c \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij} \|x_i - v_j\|} \right)^2.$$

The largest value of $PBM(c)$ indicates a valid optimal partition.

Value

PBM	the PBM index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the PBM index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," Pattern recognition, vol. 37, no. 3, pp. 487–501, 2004.

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]
```

```
# ---- FCM algorithm ----

# Compute the PBM index
FCM.PBM = PBM.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.PBM)

# The optimal number of cluster
FCM.PBM[which.max(FCM.PBM$PBM),]

# ---- EM algorithm ----

# Compute the PBM index
EM.PBM = PBM.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.PBM)

# The optimal number of cluster
EM.PBM[which.max(EM.PBM$PBM),]
```

`plot_idx` *Plots for visualizing CVIs*

Description

Plot and compare upto 8 indices computed by the algorithms in this package.

Usage

```
plot_idx(idxresult,selected.idx = NULL)
```

Arguments

`idxresult`	a result from one of the algorithms `FzzyCVIs`, `WP.IDX`, `GC.IDX`, `CCV.IDX`, `XB.IDX`, `WL.IDX`, `TANG.IDX`, `PBM.IDX`, `KWON.IDX`, `KWON2.IDX`, `KPBM.IDX`, `HF.IDX`, `Hvalid`, `Wvalid`, `SF.IDX`, `PB.IDX`, `DI.IDX`, `DB.IDX`, `CSL.IDX`, `CH.IDX` or `STRPBM.IDX`.
`selected.idx`	a numeric vector indicates a part of the indexes from the `idxresult` in respective order selected by a user. For instance, `selected.idx = 3` or `selected.idx = c(1,3,5)` may be selected. If not specified, the full idxresult will be considered.

Value

Plots of upto 8 cluster validity indices computed from `FzzyCVIs`, `WP.IDX`, `GC.IDX`, `CCV.IDX`, `XB.IDX`, `WL.IDX`, `TANG.IDX`, `PBM.IDX`, `KWON.IDX`, `KWON2.IDX`, `KPBM.IDX`, `HF.IDX`, `Hvalid`, `Wvalid`, `SF.IDX`, `PB.IDX`, `DI.IDX`, `DB.IDX`, `CSL.IDX`, `CH.IDX` or `STRPBM.IDX`. When using the isolated index algorithm, all the plots computed by that algorithm will be shown. When using `FzzyCVIs` or `Hvalid` with more than 8 selected indices, the first 8 indices will be plotted.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, "A correlation-based fuzzy cluster validity index with secondary options detector," arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, XB.IDX, Hvalid

Examples

```
library(UniversalCVI)

# Iris data
x = iris[,1:4]

# ----Compute all the indices by FzzyCVIs ----
FCVIs = FzzyCVIs(scale(x), cmax = 10, cmin = 2, indexlist = 'all', corr = 'pearson',
                 method = 'FCM', fzm = 2, iter = 100, nstart = 20, NCstart = TRUE)

# plots of the eight indices by default
plot_idx(idxresult = FCVIs)

# plots of a specific selected.idx
plot_idx(idxresult = FCVIs, selected.idx = c(2,5,7))

# ----Compute all the indices by Wvalid ----
FCM.NC = Wvalid(scale(x), kmax = 10, kmin=2, method = 'kmeans',
  corr='pearson', nstart=100, NCstart = TRUE)

# plots of the four indices by default
plot_idx(idxresult = FCM.NC)

# ----Compute all the indices by XB.IDX ----

FCM.XB = XB.IDX(scale(x), cmax = 10, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
plot_idx(idxresult = FCM.XB)
```

R1_data *R1 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 9 different Gaussian distributions labeled as 1-9.

Usage

```
R1_data
```

Format

A data frame with 450 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6,7,8,9

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`R2_data` *R2 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 7 different Gaussian distributions labeled as 1-7.

Usage

```
R2_data
```

Format

A data frame with 1750 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6,7

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`R3_data` *R3 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 16 different Gaussian distributions labeled as 1-16.

Usage

```
R3_data
```

Format

A data frame with 1600 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,...,16

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

R4_data *R4 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 5 different Gaussian distributions labeled as 1-5.

Usage

```
R4_data
```

Format

A data frame with 1250 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

R5_data *R5 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 6 different Gaussian distributions labeled as 1-6.

Usage

```
R5_data
```

Format

A data frame with 1200 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`R6_data` *R6 Artificial Dataset*

Description

A `2`-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from `6` different Gaussian distributions labeled as `1-6`.

Usage

```
R6_data
```

Format

A data frame with 1500 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

`R7_data` *R7 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2023) generated from 6 different Gaussian and 3 Uniform distributions labeled as 1-3.

Usage

```
R7_data
```

Format

A data frame with 1200 data points and 3 variables

`x` Numeric values generated from Gaussian and Uniform distributions

`y` Numeric values generated from Gaussian and Uniform distributions

`label` Categorical labels 1,2,3

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A correlation-based fuzzy cluster validity index with secondary options detector, arXiv:2308.14785, 2023

See Also

FzzyCVIs, WP.IDX, D1_data, Hvalid, DI.IDX

SF.IDX *The score function*

Description

Computes the SF (S. Saitta et al., 2007) index for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
SF.IDX(x, kmax, kmin = 2, method = "kmeans", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

The smallest value of $SF(k)$ indicates a valid optimal partition.

Value

SF	the Score function index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the SF index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. Saitta, B. Raphael, I. Smith, "A bounded index for cluster validity," *In Perner, P.: Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science*, 4571, Springer (2007).

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute the SF index
K.SF = SF.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans", nstart = 100)
print(K.SF)

# The optimal number of cluster
K.SF[which.min(K.SF$SF),]

# ---- Hierarchical ----

# Average linkage

# Compute the SF index
H.SF = SF.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average")
print(H.SF)

# The optimal number of cluster
H.SF[which.min(H.SF$SF),]
```

SH.IDX *Silhouette index*

Description

Computes the SH (Rousseeuw, 1987; Kaufman and Rousseeuw, 2009) index for a result either kmeans or hierarchical clustering from user specified `kmin` to `kmax`.

Usage

```
SH.IDX(x, kmax, kmin = 2, method = "kmeans", nstart = 100)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`kmax`	a maximum number of clusters to be considered.
`kmin`	a minimum number of clusters to be considered. The default is `2`.
`method`	a character string indicating which clustering method to be used (`"kmeans"`, `"hclust_complete"`, `"hclust_average"`, `"hclust_single"`). The default is `"kmeans"`.
`nstart`	a maximum number of initial random sets for kmeans for `method = "kmeans"`. The default is `100`.

Details

For $i \in [n]$, $l \in [k]$, and $x_i \in C_l$, let

$$a(i) = \frac{1}{|C_l| - 1} \sum_{y \in C_l} \|x_i - y\| \; and$$

$$b(i) = \min_{r \neq l} \frac{1}{|C_r|} \sum_{y \in C_r} \|x_i - y\| .$$

The silhouette value of one data point x_j is defined as:

$$s(j) = \begin{cases} \dfrac{b(j) - a(j)}{\max\{a(j), b(i)\}} & \text{if } |C_j| > 1 \\ 0 & \text{if } |C_j| = 1 \end{cases} .$$

The silhouette index is defined as
$SH(k) = \frac{1}{n} \sum_{i=1}^{n} s(i)$.
The largest value of $SH(k)$ indicates a valid optimal partition.

Value

SH	the SH index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the SH index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Kaufman, L. and Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

See Also

Hvalid, Wvalid, DI.IDX, FzzyCVIs, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Hierarchical ----

# Average linkage
```

```
# Compute the SH index
H.SH = SH.IDX(scale(x), kmax = 10, kmin = 2, method = "hclust_average", nstart = 1)
print(H.SH)

# The optimal number of cluster
H.SH[which.max(H.SH$SH),]
```

STRPBM.IDX *Starczewski and Pakhira-Bandyopadhyay-Maulik for crisp clustering indexes*

Description

Computes the STR (A. Starczewski, 2017) and PBM (M. K. Pakhira et al., 2004) indexes for a result either kmeans or hierarchical clustering from user specified kmin to kmax.

Usage

```
STRPBM.IDX(x, kmax, kmin = 2, method = "kmeans", indexlist = "all", nstart = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
indexlist	a character string indicating which cluster validity indexes to be computed ("all", "STR", "PBM"). More than one indexes can be selected.
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.

Details

PBM index can be used with both crisp and fuzzy clustering algorithms.
The largest value of $STR(k)$ indicates a valid optimal partition.
The largest value of $PBM(k)$ indicates a valid optimal partition.

Value

STR	the STR index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the STR index, respectively.
PBM	the PBM index for k from kmin to kmax shown in a data frame where the first and the second columns are k and the PBM index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. K. Pakhira, S. Bandyopadhyay and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recogn* 37(3):487–501 (2004).

A. Starczewski, "A new validity index for crisp clusters," *Pattern Anal Applic* 20, 687–700 (2017).

See Also

Wvalid, FzzyCVIs, DI.IDX, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute all the indices by STRPBM.IDX
K.ALL = STRPBM.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans",
  indexlist = "all", nstart = 100)
print(K.ALL)

# Compute STR index
K.STR = STRPBM.IDX(scale(x), kmax = 15, kmin = 2, method = "kmeans",
  indexlist = "STR", nstart = 100)
print(K.STR)

# ---- Hierarchical ----

# Average linkage

# Compute all the indices by STRPBM.IDX
H.ALL = STRPBM.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average",
  indexlist = "all")
print(H.ALL)

# Compute STR index
H.STR = STRPBM.IDX(scale(x), kmax = 15, kmin = 2, method = "hclust_average",
  indexlist = "STR")
print(H.STR)
```

TANG.IDX *Tang index*

Description

Computes the TANG (Y. Tang et al., 2005) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
TANG.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The Tang index is defined as

$$TANG(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{2}\|x_i - v_j\|^2 + \frac{1}{c(c-1)}\sum_{j\neq k}\|v_j - v_k\|^2}{\min_{j\neq k}\{\|v_j - v_k\|^2\} + \frac{1}{c}}.$$

The smallest value of $TANG(c)$ indicates a valid optimal partition.

Value

TANG	the TANG index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the TANG index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

Y. Tang, F. Sun, and Z. Sun, "Improved validation index for fuzzy clustering," in Proceedings of the 2005, American Control Conference, 2005., pp. 1120–1125 vol. 2, 2005. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1470111&isnumber=31519

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the TANG index
FCM.TANG = TANG.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.TANG)

# The optimal number of cluster
FCM.TANG[which.min(FCM.TANG$TANG),]

# ---- EM algorithm ----

# Compute the TANG index
EM.TANG = TANG.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.TANG)

# The optimal number of cluster
EM.TANG[which.min(EM.TANG$TANG),]
```

WL.IDX *Wu and Li (WL) index*

Description

Computes the WL (C. H. Wu et al., 2015) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
WL.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.
cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The WL index is defined as

$$WL(c) = \frac{\sum_{j=1}^{c}\left(\frac{\sum_{i=1}^{n}\mu_{ij}^{2}\|x_i - v_j\|^2}{\sum_{i=1}^{n}\mu_{ij}}\right)}{min_{j\neq k}\{\|v_j - v_k\|^2\} + median_{j\neq k}\{\|v_j - v_k\|^2\}}.$$

The smallest value of $WL(c)$ indicates a valid optimal partition.

Value

WL	the WL index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the WL index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. H. Wu, C. S. Ouyang, L. W. Chen, and L. W. Lu, "A new fuzzy clustering validity index with a median factor for centroid-based clustering," IEEE Transactions on Fuzzy Systems, vol. 23, no. 3, pp. 701–718, 2015.https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6811211&isnumber=7115244

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the WL index
FCM.WL = WL.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.WL)

# The optimal number of cluster
FCM.WL[which.min(FCM.WL$WL),]

# ---- EM algorithm ----

# Compute the WL index
EM.WL = WL.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.WL)

# The optimal number of cluster
EM.WL[which.min(EM.WL$WL),]
```

WP.IDX *Wiroonsri and Preedasawakul (WP) index*

Description

Computes the WPC (WP correlation), WP, WPCI1 and WPCI2 (N. Wiroonsri and O. Preedasawakul, 2023) indexes for a result of either FCM or EM clustering from user specified `cmin` to `cmax`.

Usage

```
WP.IDX(x, cmax, cmin = 2, corr = 'pearson', method = 'FCM', fzm = 2,
  gamma = (fzm^2*7)/4, sampling = 1, iter = 100, nstart = 20, NCstart = TRUE)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`cmax`	a maximum number of clusters to be considered.
`cmin`	a minimum number of clusters to be considered. The default is `2`.
`corr`	a character string indicating which correlation coefficient is to be computed (`"pearson"`, `"kendall"` or `"spearman"`). The default is `"pearson"`.

method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
gamma	adjusted fuzziness parameter for indexlist = ("WP", "WPC", "WPCI1", "WPCI2"). The default is computed from $7fzm^2/4$.
sampling	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is 1.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
NCstart	logical for indexlist = ("WP", "WPC", "WPCI1","WPCI2"), if TRUE, the WP correlation at c=1 is defined as an adjusted sd of the distances between all data points and their mean. Otherwise, the WP correlation at c=1 is defined as 0.

Details

The newly introduced index was inspired by the recently introduced Wiroonsri index which is only compatible with hard clustering methods.

The WPC computes the correlation between the actual distance between a pair of data points and the distance between adjusted centroids with respect to the pair. WPCI1 and WPCI2 are the proportion and the subtraction, respectively, of the same two ratios. The first ratio is the WPC improvement from c-1 clusters to c clusters over the entire room for improvement. The second ratio is the WPC improvement from c clusters to c+1 clusters over the entire room for improvement. WP is defined as a combination of WPCI1 and WPCI2.

The largest value of WP(c) indicates a valid optimal partition.

Value

WPC	the WP correlations for c from cmin-1 to cmax+1 shown in a data frame where the first and the second columns are c and the WPC, respectively.

Each of the followings show the value of each index for c from cmin to cmax in a data frame.

WP	the WP index.
WPCI1	the WPCI1 index.
WPCI2	the WPCI2 index.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, "A correlation-based fuzzy cluster validity index with secondary options detector," arXiv:2308.14785, 2023

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute all the indices by WP.IDX using default gamma
FCM.WP = WP.IDX(scale(x), cmax = 10, cmin = 2, corr = 'pearson', method = 'FCM', fzm = 2,
  iter = 100, nstart = 20, NCstart = TRUE)
print(FCM.WP$WP)

# The optimal number of cluster
FCM.WP$WP[which.max(FCM.WP$WP$WPI),]


# ---- EM algorithm ----

# Compute all the indices by WP.IDX using default gamma
EM.WP = WP.IDX(scale(x), cmax = 10, cmin = 2, corr = 'pearson', method = 'EM',
  iter = 100, nstart = 20, NCstart = TRUE)
print(EM.WP$WP)

# The optimal number of cluster
EM.WP$WP[which.max(EM.WP$WP$WPI),]
```

Wvalid *Wiroonsri(2024) correlation-based cluster validity indices*

Description

Computes the NC correlation, NCI, NCI1 and NCI2 cluster validity indices for the number of clusters from user specified kmin to kmax obtained from either K-means or hierarchical clustering based on the recent paper by Wiroonsri(2024).

Usage

```
Wvalid(x, kmax, kmin = 2, method = "kmeans",
  corr = "pearson", nstart = 100, sampling = 1, NCstart = TRUE)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
kmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman"). The default is "pearson".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
sampling	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is 1.
NCstart	logical for indexlist includes the "NC", "NCI", "NCI1", and "NCI2"), if TRUE, the NC correlation at k=1 is defined as the ratio introduced in the reference. Otherwise, it is assigned as 0.

Details

The NC correlation computes the correlation between an actual distance between a pair of data points and a centroid distance of clusters that the two points locate in. NCI1 and NCI2 are the proportion and the subtraction, respectively, of the same two ratios. The first ratio is the NC improvement from k-1 clusters to k clusters over the entire room for improvement. The second ratio is the NC improvement from k clusters to k+1 clusters over the entire room for improvement. NCI is a combination of NCI1 and NCI2.

Value

NC	the NC correlations for k from kmin-1 to kmax+1 shown in a data frame where the first and the second columns are k and the NC, respectively.

Each of the followings shows the values of each index for k from kmin to kmax in a data frame.

NCI	the NCI index.
NCI1	the NCI1 index.
NCI2	the NCI2 index.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, "Clustering performance analysis using a new correlation based cluster validity index," Pattern Recognition, 145, 109910, 2024. doi:10.1016/j.patcog.2023.109910

See Also

Hvalid, FzzyCVIs, DB.IDX, R1_data

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- Kmeans ----

# Compute all the indices by Wvalid
K.NC = Wvalid(scale(x), kmax = 15, kmin=2, method = 'kmeans',
  corr='pearson', nstart=100, NCstart = TRUE)
print(K.NC)

# The optimal number of cluster
K.NC$NCI[which.max(K.NC$NCI$NCI),]

# ---- Hierarchical ----

# Average linkage

# Compute all the indices by Wvalid
H.NC = Wvalid(scale(x), kmax = 15, kmin=2, method = 'hclust_average',
  corr='pearson', nstart=100, NCstart = TRUE)
print(H.NC)

# The optimal number of cluster
H.NC$NCI[which.max(H.NC$NCI$NCI),]
```

XB.IDX *Xie and Beni (XB) index*

Description

Computes the XB (X. L. Xie and G. Beni, 1991) index for a result of either FCM or EM clustering from user specified cmin to cmax.

Usage

```
XB.IDX(x, cmax, cmin = 2, method = "FCM", fzm = 2, nstart = 20, iter = 100)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
cmax	a maximum number of clusters to be considered.

cmin	a minimum number of clusters to be considered. The default is 2.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

Details

The XB index is defined as

$$XB(c) = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{2}\|x_i - v_j\|^2}{n \cdot \min_{j\neq k}\{\|v_j - v_k\|^2\}}.$$

The lowest value of $XB(c)$ indicates a valid optimal partition.

Value

XB	the XB index for c from cmin to cmax shown in a data frame where the first and the second columns are c and the XB index, respectively.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

X. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 841–847, 1991.

See Also

R1_data, TANG.IDX, FzzyCVIs, WP.IDX, Hvalid

Examples

```
library(UniversalCVI)

# The data is from Wiroonsri (2024).
x = R1_data[,1:2]

# ---- FCM algorithm ----

# Compute the XB index
FCM.XB = XB.IDX(scale(x), cmax = 15, cmin = 2, method = "FCM",
  fzm = 2, nstart = 20, iter = 100)
print(FCM.XB)
```

```
# The optimal number of cluster
FCM.XB[which.min(FCM.XB$XB),]

# ---- EM algorithm ----

# Compute the XB index
EM.XB = XB.IDX(scale(x), cmax = 15, cmin = 2, method = "EM",
  nstart = 20, iter = 100)
print(EM.XB)

# The optimal number of cluster
EM.XB[which.min(EM.XB$XB),]
```

Index