

Package ‘bootcluster’

October 12, 2022

Type Package

Title Bootstrapping Estimates of Clustering Stability

Version 0.3.2

Author Han Yu [aut],
Mingmei Tian [aut],
Tianmou Liu [aut, cre]

Maintainer Tianmou Liu <tianmoul@buffalo.edu>

Description Implementation of the bootstrapping approach for the estimation of clustering stability and its application in estimating the number of clusters, as introduced by Yu et al (2016) <[doi:10.1142/9789814749411_0007](https://doi.org/10.1142/9789814749411_0007)>. Implementation of the non-parametric bootstrap approach to assessing the stability of module detection in a graph, the extension for the selection of a parameter set that defines a graph from data in a way that optimizes stability and the corresponding visualization functions, as introduced by Tian et al (2021) <[doi:10.1002/sam.11495](https://doi.org/10.1002/sam.11495)>.

Depends R (>= 3.5.1)

Imports cluster, mclust, flexclust, fpc, plyr, dplyr, doParallel,
foreach, igraph, compiler, stats, parallel, grid, ggplot2,
gridExtra, intergraph, GGally, network, sna

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2.9000

NeedsCompilation no

Repository CRAN

Date/Publication 2022-01-29 22:50:03 UTC

R topics documented:

k.select	2
k.select_ref	3
network.stability	4

network.stability.output	6
ob.stability	7
stability	8
threshold.select	9
wine	11

Index	12
--------------	-----------

k.select	<i>Estimate number of clusters</i>
----------	------------------------------------

Description

Estimate number of clusters by bootstrapping stability

Usage

```
k.select(x, range = 2:7, B = 20, r = 5, threshold = 0.8, scheme_2 = TRUE)
```

Arguments

x	a data.frame of the data set
range	a vector of integer values, of the possible numbers of clusters k
B	number of bootstrap re-samplings
r	number of runs of k-means
threshold	the threshold for determining k
scheme_2	logical TRUE if scheme 2 is used, FALSE if scheme 1 is used

Details

This function estimates the number of clusters through a bootstrapping approach, and a measure S_{min} , which is based on an observation-wise similarity among clusterings. The number of clusters k is selected as the largest number of clusters, for which the S_{min} is greater than a threshold. The threshold is often selected between 0.8 ~ 0.9. Two schemes are provided. Scheme 1 uses the clustering of the original data as the reference for stability calculations. Scheme 2 searches across the clustering samples that gives the most stable clustering.

Value

profile a vector of S_{min} measures for determining k
 k integer estimated number of clusters

Author(s)

Han Yu

References

Bootstrapping estimates of stability for clusters, observations and model selection. Han Yu, Brian Chapman, Arianna DiFlorio, Ellen Eischen, David Gotz, Matthews Jacob and Rachael Hageman Blair.

Examples

```
set.seed(1)
data(wine)
x0 <- wine[,2:14]
x <- scale(x0)
k.select(x, range = 2:10, B=20, r=5, scheme_2 = TRUE)
```

k.select_ref	<i>Estimate number of clusters</i>
--------------	------------------------------------

Description

Estimate number of clusters by bootstrapping stability

Usage

```
k.select_ref(df, k_range = 2:7, n_ref = 5, B = 100, B_ref = 50, r = 5)
```

Arguments

df	data.frame of the input dataset
k_range	integer valued vector of the numbers of clusters k to be tested upon
n_ref	number of reference distribution to be generated
B	number of bootstrap re-samples
B_ref	number of bootstrap resamples for the reference distributions
r	number of runs of k-means

Details

This function uses the out-of-bag scheme to estimate the number of clusters in a dataset. The function calculate the Smin of the dataset and at the same time, generate a reference dataset with the same range as the original dataset in each dimension and calculate the Smin_ref. The differences between Smin and Smin_ref at each k, Smin_diff(k), is taken into consideration as well as the standard deviation of the differences. We choose the k to be the argmax of (Smin_diff(k) - (Smin_diff(k+1) + (Smin_diff(k+1)))). If Smin_diff(k) less than 0.1 for all k in k_range, we say k = 1

Value

profile vector of $(S_{\text{min_diff}}(k) - (S_{\text{min_diff}}(k+1) + \text{se}(S_{\text{min_diff}}(k+1))))$ measures for researchers's inspection

k estimated number of clusters

Author(s)

Tianmou Liu

References

Bootstrapping estimates of stability for clusters, observations and model selection. Han Yu, Brian Chapman, Arianna DiFlorio, Ellen Eischen, David Gotz, Matthews Jacob and Rachael Hageman Blair.

Examples

```
set.seed(1)
data(iris)
df <- data.frame(iris[,1:4])
df <- scale(df)
k.select_ref(df, k_range = 2:7, n_ref = 5, B=500, B_ref = 500, r=5)
```

network.stability *Estimate of detect module stability*

Description

Estimate of detect module stability

Usage

```
network.stability(
  data.input,
  threshold,
  B = 20,
  cor.method,
  large.size,
  PermuNo,
  scheme_2 = FALSE
)
```

Arguments

<code>data.input</code>	a data.frame of the data set where the rows are observations and columns are covariates
<code>threshold</code>	a numeric number of threshold for correlation matrix
<code>B</code>	number of bootstrap re-samplings
<code>cor.method</code>	the correlation method applied to the data set, three method are available: "pearson", "kendall", "spearman".
<code>large.size</code>	the smallest set of modules, the <code>large.size=0</code> is recommended to use right now.
<code>PermuNo</code>	number of random graphs for null
<code>scheme_2</code>	logical TRUE if scheme 2 is used, FASLE if scheme 1 is used. Right now, only FASLE is recommended.

Details

This function estimates the modules' stability through bootstrapping approach for the given threshold. The approach to stability estimation is to compare the module composition of the reference correlation graph to the various bootstrapped correlation graphs, and to assess the stability at the (1) node-level, (2) module-level, and (3) overall.

Value

`stabilityresult` a list of result for nodes-wise stability
`modularityresult` list of modularity information with the given threshold
`jaccardresult` list estimated unconditional observed stability and the estimates of expected stability under the null
`originalinformation` list information for original data, igraph object and adjacency matrix constructed with the given threshold

Author(s)

Mingmei Tian

References

A framework for stability-based module detection in correlation graphs. Mingmei Tian, Rachael Hageman Blair, Lina Mu, Matthew Bonner, Richard Browne and Han Yu.

Examples

```
set.seed(1)
data(wine)
x0 <- wine[1:50,]

mytest<-network.stability(data.input=x0,threshold=0.7, B=20,
cor.method='pearson',large.size=0,
```

```
PermuNo = 10,  
scheme_2 = FALSE)
```

network.stability.output

Plot method for objects from threshold.select

Description

Plot method for objects from threshold.select

Usage

```
network.stability.output(input, optimal.only = FALSE)
```

Arguments

input	a list of results from function threshold.select
optimal.only	a logical value indicating whether only plot the network with optimal threshold or not. The default is False, generating all network figures with a large number of nodes could take some time.

Details

network.stability.output is used to generate a series of network plots based on the given threshold.seq, where the nodes are colored by the level of stability. The network with optimal threshold value selected by function threshold.select is colored as red.

Value

Plot of network figures

Author(s)

Mingmei Tian

References

A framework for stability-based module detection in correlation graphs. Mingmei Tian, Rachael Hageman Blair, Lina Mu, Matthew Bonner, Richard Browne and Han Yu.

Examples

```

set.seed(1)
data(wine)
x0 <- wine[1:50,]

mytest<-threshold.select(data.input=x0,threshold.seq=seq(0.1,0.5,by=0.05), B=20,
cor.method='pearson',large.size=0,
PermuNo = 10,
no_cores=1,
scheme_2 = FALSE)
network.stability.output(mytest)

```

ob.stability	<i>Estimate the stability of a clustering based on non-parametric bootstrap out-of-bag scheme, with option for subsampling scheme</i>
--------------	---

Description

Estimate the stability of a clustering based on non-parametric bootstrap out-of-bag scheme, with option for subsampling scheme

Usage

```
ob.stability(x, k, B = 500, r = 5, subsample = FALSE, cut_ratio = 0.5)
```

Arguments

x	data.frame of the data set where the rows as observations and columns as dimensions of features
k	number of clusters for which to estimate the stability
B	number of bootstrap re-samples
r	integer parameter in the kmeansCBI() funtion
subsample	logical parameter to use the subsampling scheme option in the resampling process (instead of bootstrap)
cut_ratio	numeric parameter between 0 and 1 for subsampling scheme training set ratio

Details

This function estimates the stability through out-of-bag observations It estimate the stability at the (1) observation level, (2) cluster level, and (3) overall.

Value

membership vector of membership for each observation from the reference clustering
 obs_wise vector of estimated observation-wise stability
 clust_wise vector of estimated cluster-wise stability
 overall numeric estimated overall stability
 Smin numeric estimated Smin through out-of-bag scheme

Author(s)

Tianmou Liu

References

Bootstrapping estimates of stability for clusters, observations and model selection. Han Yu, Brian Chapman, Arianna DiFlorio, Ellen Eischen, David Gotz, Matthews Jacob and Rachael Hageman Blair.

Examples

```
set.seed(123)
data(iris)
df <- data.frame(iris[,1:4])
# You can choose to scale df before clustering by
# df <- scale(df)
ob.stability(df, k = 2, B=500, r=5)
```

stability

Estimate clustering stability of k-means

Description

Estimate of k-means bootstrapping stability

Usage

```
stability(x, k, B = 20, r = 5, scheme_2 = TRUE)
```

Arguments

x	a data.frame of the data set
k	a integer number of clusters
B	number of bootstrap re-samplings
r	number of runs of k-means
scheme_2	logical TRUE if scheme 2 is used, FALSE if scheme 1 is used

Details

This function estimates the clustering stability through bootstrapping approach. Two schemes are provided. Scheme 1 uses the clustering of the original data as the reference for stability calculations. Scheme 2 searches across the clustering samples that gives the most stable clustering.

Value

membership a vector of membership for each observation from the reference clustering

obs_wise vector of estimated observation-wise stability

overall numeric estimated overall stability

Author(s)

Han Yu

References

Bootstrapping estimates of stability for clusters, observations and model selection. Han Yu, Brian Chapman, Arianna DiFlorio, Ellen Eischen, David Gotz, Matthews Jacob and Rachael Hageman Blair.

Examples

```
set.seed(1)
data(wine)
x0 <- wine[,2:14]
x <- scale(x0)
stability(x, k = 3, B=20, r=5, scheme_2 = TRUE)
```

threshold.select	<i>Estimate of the overall Jaccard stability</i>
------------------	--

Description

Estimate of the overall Jaccard stability

Arguments

data.input	a data.frame of the data set where the rows are observations and columns are covariates
threshold.seq	a numeric sequence of candidate threshold
B	number of bootstrap re-samplings
cor.method	the correlation method applied to the data set, three method are available: "pearson", "kendall", "spearman".

<code>large.size</code>	the smallest set of modules, the <code>large.size=0</code> is recommended to use right now.
<code>PermuNo</code>	number of random graphs for the estimation of expected stability
<code>no_cores</code>	a interger number of CPU cores on the current host (This function can't not be used yet).

Details

`threshold.select` is used to estimate of the overall Jaccard stability from a sequence of given threshold candidates, `threshold.seq`.

Value

`stabilityresult` a list of result for nodes-wise stability
`modularityresult` a list of modularity information with each candidate threshold
`jaccardresult` a list estimated unconditional observed stability and the estimates of expected stability under the nul
`originalinformation` a list information for original data, `igraph` object and adjacency matrix constructed with each candidate threshold
`threshold.seq` a list of candicate threshold given to the function

Author(s)

Mingmei Tian

References

A framework for stability-based module detection in correlation graphs. Mingmei Tian,Rachael Hageman Blair,Lina Mu, Matthew Bonner, Richard Browne and Han Yu.

Examples

```
set.seed(1)
data(wine)
x0 <- wine[1:50,]

mytest<-threshold.select(data.input=x0,threshold.seq=seq(0.5,0.8,by=0.05), B=20,
cor.method='pearson',large.size=0,
PermuNo = 10,
no_cores=1,
scheme_2 = FALSE)
```

wine

Wine Data Set

Description

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Usage

```
data(wine)
```

Format

The data set wine contains a data.frame of 14 variables. The first variable is the types of wines. The other 13 variables are quantities of the constituents.

References

<https://archive.ics.uci.edu/ml/datasets/wine>

Index

k.select, [2](#)

k.select_ref, [3](#)

network.stability, [4](#)

network.stability.output, [6](#)

ob.stability, [7](#)

stability, [8](#)

threshold.select, [9](#)

wine, [11](#)