

Package ‘bootcluster’

June 10, 2021

Type Package

Title Bootstrapping Estimates of Clustering Stability

Version 0.2.5

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

Maintainer Mingmei Tian <mingmei@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 (>= 3.3.1)

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 7.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2021-06-10 08:20:04 UTC

R topics documented:

k.select	2
network.stability	3
network.stability.output	4
stability	5
threshold.select	7
wine	8

Index**9**

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)

```

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

data.input	a data.frame of the data set where the rows are observations and columns are covariates
threshold	a numeric number of threshold for correlation matrix
B	number of bootstrap re-samplings
cor.method	the correlation method applied to the data set,three method are available: "pearson", "kendall", "spearman"
large.size	the smallest set of modules, the large.size=0 is recommended to use right now.
PermuNo	number of random graphs for null
scheme_2	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)
```

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"
large.size	the smallest set of modules, the large.size=0 is recommended to use right now.
PermuNo	number of random graphs for the estimation of expected stability
no_cores	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](#)

network.stability, [3](#)

network.stability.output, [4](#)

stability, [5](#)

threshold.select, [7](#)

wine, [8](#)