Package 'bootcluster'

November 13, 2017

Type Package

Title Bootstrapping Estimates of Clustering Stability

Version 0.1.0

Author Han Yu

Maintainer Han Yu <hyu9@buffalo.edu>

Description Implementation of the bootstrapping approach for the estimation of clustering stability on observation and cluster level, as well as its application in estimating the number of clusters.

Depends R (>= 3.3.1)

Imports cluster, mclust, flexclust, sets, fpc, plyr

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-11-13 15:38:38 UTC

R topics documented:

k.select *Estimate number of clusters*

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, FASLE if scheme 1 is used

Details

This function estimates the number of clusters through a bootstrapping approach, and a measure Smin, 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 Smin 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 acrossthe clustering samples that gives the most stable clustering.

Value

profile a vector of Smin 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)
```

`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, `FASLE` 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 acrossthe 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)
```

---

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