Package 'NbClust'

April 13, 2015

Type Package

Title Determining the Best Number of Clusters in a Data Set

Version 3.0

Depends R (>= 3.1.0)

Date 2015-04-13

Author Malika Charrad and Nadia Ghazzali and Veronique Boiteau and Azam Niknafs

Maintainer Malika Charrad <malika.charrad.1@ulaval.ca>

Description It provides 30 indexes for determining the optimal number of clusters in a data set and offers the best clustering scheme from different results to the user.

URL https://sites.google.com/site/malikacharrad/research/nbclust-package

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2015-04-13 22:54:43

R topics documented:

`NbClust` *NbClust Package for determining the best number of clusters*

Description

`NbClust` package provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.

Usage

```
NbClust(data = NULL, diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 15,

method = NULL, index = "all", alphaBeale = 0.1)
```

Arguments

data	matrix or dataset.
diss	dissimilarity matrix to be used. By default, diss=NULL, but if it is replaced by a dissimilarity matrix, distance should be "NULL".
distance	the distance measure to be used to compute the dissimilarity matrix. This must be one of: "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski" or "NULL". By default, distance="euclidean". If the distance is "NULL", the dissimilarity matrix (diss) should be given by the user. If distance is not "NULL", the dissimilarity matrix should be "NULL".
min.nc	minimal number of clusters, between 1 and (number of objects - 1)
max.nc	maximal number of clusters, between 2 and (number of objects - 1), greater or equal to min.nc. By default, max.nc=15.
method	the cluster analysis method to be used. This should be one of: "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", "centroid", "kmeans".
index	the index to be calculated. This should be one of : "kl", "ch", "hartigan", "ccc", "scott", "marriot", "trcovw", "tracew", "friedman", "rubin", "cindex", "db", "silhouette", "duda", "pseudot2", "beale", "ratkowsky", "ball", "ptbiserial", "gap", "frey", "mcclain", "gamma", "gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw", "all" (all indices except GAP, Gamma, Gplus and Tau), "alllong" (all indices with Gap, Gamma, Gplus and Tau included).
alphaBeale	significance value for Beale's index.

Details

1. **Notes on the "Distance" argument**
   The following distance measures are written for two vectors **x** and **y**. They are used when the data is a **d**-dimensional vector arising from measuring **d** characteristics on each of **n** objects or individuals.

   - **Euclidean distance** : Usual square distance between the two vectors (2 norm).

$$d(x,y) = \left(\sum_{j=1}^{d} (x_j - y_j)^2 \right)^{\frac{1}{2}}$$

   - **Maximum distance**: Maximum distance between two components of **x** and **y** (supremum norm).

$$d(x,y) = \sup_{1 \le j \le d} |x_j - y_j|$$

- **Manhattan distance** : Absolute distance between the two vectors (1 norm).

$$d(x,y) = \sum_{j=1}^{d} |x_j - y_j|$$

- **Canberra distance** : Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing.

$$d(x,y) = \sum_{j=1}^{d} \frac{|x_j - y_j|}{|x_j| + |y_j|}$$

- **Binary distance** : The vectors are regarded as binary bits, so non-zero elements are "on" and zero elements are "off". The distance is the proportion of bits in which only one is on amongst those in which at least one is on.
- **Minkowski distance** : The **p** norm, the p^{th} root of the sum of the p^{th} powers of the differences of the components.

$$d(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{\frac{1}{p}}$$

2. **Notes on the "method" argument**
   The following aggregation methods are available in this package.
   - **Ward** : Ward method minimizes the total within-cluster variance. At each step the pair of clusters with minimum cluster distance are merged. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. Two different algorithms are found in the literature for Ward clustering. The one used by option "ward.D" (equivalent to the only Ward option "ward" in R versions <= 3.0.3) does not implement Ward's (1963) clustering criterion, whereas option "ward.D2" implements that criterion (Murtagh and Legendre 2013). With the latter, the dissimilarities are squared before cluster updating.
   - **Single** : The distance D_{ij} between two clusters C_i and C_j is the minimum distance between two points x and y, with $x \in C_i, y \in C_j$.

$$D_{ij} = min_{x \in C_i, y \in C_j} d(x,y)$$

     A drawback of this method is the so-called chaining phenomenon: clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other.
   - **Complete** : The distance D_{ij} between two clusters C_i and C_j is the maximum distance between two points x and y, with $x \in C_i, y \in C_j$.

$$D_{ij} = max_{x \in C_i, y \in C_j} d(x,y)$$

   - **Average** : The distance D_{ij} between two clusters C_i and C_j is the mean of the distances between the pair of points x and y, where $x \in C_i, y \in C_j$.

$$D_{ij} = sum_{x \in C_i, y \in C_j} \frac{d(x,y)}{n_i \times n_j}$$

where n_i and n_j are respectively the number of elements in clusters C_i and C_j. This method has the tendency to form clusters with the same variance and, in particular, small variance.

- **McQuitty** : The distance between clusters C_i and C_j is the weighted mean of the between-cluster dissimilarities:

$$D_{ij} = (D_{ik} + D_{il})/2$$

where cluster C_j is formed from the aggregation of clusters C_k and C_l.

- **Median** : The distance D_{ij} between two clusters C_i and C_j is given by the following formula:

$$D_{ij} = \frac{(D_{ik} + D_{il})}{2} - \frac{D_{kl}}{4}$$

where cluster C_j is formed by the aggregation of clusters C_k and C_l.

- **Centroid** : The distance D_{ij} between two clusters C_i and C_j is the squared euclidean distance between the gravity centers of the two clusters, i.e. between the mean vectors of the two clusters, $\bar{x_i}$ and $\bar{x_j}$ respectively.

$$D_{ij} = \|\bar{x_i} - \bar{x_j}\|^2$$

This method is more robust than others in terms of isolated points.

- **Kmeans** : This method is said to be a reallocation method. Here is the general principle:
  (a) Select as many points as the number of desired clusters to create initial centers.
  (b) Each observation is then associated with the nearest center to create temporary clusters.
  (c) The gravity centers of each temporary cluster is calculated and these become the new clusters centers.
  (d) Each observation is reallocated to the cluster which has the closest center.
  (e) This procedure is iterated until convergence.

3. **Notes on the "Index" argument**

The table below summarizes indices implemented in NbClust and the criteria used to select the optimal number of clusters.

Index in NbClust	**Optimal number of clusters**
1. "kl" or "all" or "alllong" (Krzanowski and Lai 1988)	Maximum value of the index
2. "ch" or "all" or "alllong" (Calinski and Harabasz 1974)	Maximum value of the index
3. "hartigan" or "all" or "alllong" (Hartigan 1975)	Maximum difference between hierarchy levels of the index
4. "ccc" or "all" or "alllong" (Sarle 1983)	Maximum value of the index
5. "scott" or "all" or "alllong" (Scott and Symons 1971)	Maximum difference between hierarchy levels of the index
6. "marriot" or "all" or "alllong"	Max. value of second differences

(Marriot 1971)	between levels of the index
7. "trcovw" or "all" or "alllong"	Maximum difference between
(Milligan and Cooper 1985)	hierarchy levels of the index
8. "tracew" or "all" or "alllong"	Maximum value of absolute second
(Milligan and Cooper 1985)	differences between levels of the index
9. "friedman" or "all" or "alllong"	Maximum difference between
(Friedman and Rubin 1967)	hierarchy levels of the index
10. "rubin" or "all" or "alllong"	Minimum value of second differences
(Friedman and Rubin 1967)	between levels of the index
11. "cindex" or "all" or "alllong"	Minimum value of the index
(Hubert and Levin 1976)	
12. "db" or "all" or "alllong"	Minimum value of the index
(Davies and Bouldin 1979)	
13. "silhouette" or "all" or "alllong"	Maximum value of the index
(Rousseeuw 1987)	
14. "duda" or "all" or "alllong"	Smallest n_c such that index > criticalValue
(Duda and Hart 1973)	
15. "pseudot2" or "all" or "alllong"	Smallest n_c such that index < criticalValue
(Duda and Hart 1973)	
16. "beale" or "all" or "alllong"	n_c such that critical value of the index >= alpha
(Beale 1969)	
17. "ratkowsky" or "all" or "alllong"	Maximum value of the index
(Ratkowsky and Lance 1978)	
18. "ball" or "all" or "alllong"	Maximum difference between hierarchy
(Ball and Hall 1965)	levels of the index
19. "ptbiserial" or "all" or "alllong"	Maximum value of the index
(Milligan 1980, 1981)	
20. "gap" or "alllong"	Smallest n_c such that criticalValue >= 0
(Tibshirani et al. 2001)	
21. "frey" or "all" or "alllong"	the cluster level before that index value < 1.00
(Frey and Van Groenewoud 1972)	
22. "mcclain" or "all" or "alllong"	Minimum value of the index
(McClain and Rao 1975)	
23. "gamma" or "alllong"	Maximum value of the index
(Baker and Hubert 1975)	
24. "gplus" or "alllong"	Minimum value of the index
(Rohlf 1974) (Milligan 1981)	
25. "tau" or "alllong"	Maximum value of the index
(Rohlf 1974) (Milligan 1981)	
26. "dunn" or "all" or "alllong"	Maximum value of the index
(Dunn 1974)	
27. "hubert" or "all" or "alllong"	Graphical method
(Hubert and Arabie 1985)	
28. "sdindex" or "all" or "alllong"	Minimum value of the index
(Halkidi et al. 2000)	
29. "dindex" or "all" or "alllong"	Graphical method
(Lebart et al. 2000)	
30. "sdbw" or "all" or "alllong"	Minimum value of the index

(Halkidi and Vazirgiannis 2001)

Value

`All.index`	Values of indices for each partition of the dataset obtained with a number of clusters between min.nc and max.nc.
`All.CriticalValues`	Critical values of some indices for each partition obtained with a number of clusters between min.nc and max.nc.
`Best.nc`	Best number of clusters proposed by each index and the corresponding index value.
`Best.partition`	Partition that corresponds to the best number of clusters

Author(s)

Malika Charrad, Nadia Ghazzali, Veronique Boiteau and Azam Niknafs

References

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.", "Journal of Statistical Software, 61(6), 1-36.", "URL http://www.jstatsoft.org/v61/i06/".

Examples

```
## DATA MATRIX IS GIVEN

## A 2-dimensional example

set.seed(1)
x<-rbind(matrix(rnorm(100,sd=0.1),ncol=2),
         matrix(rnorm(100,mean=1,sd=0.2),ncol=2),
         matrix(rnorm(100,mean=5,sd=0.1),ncol=2),
         matrix(rnorm(100,mean=7,sd=0.2),ncol=2))

res<-NbClust(x, distance = "euclidean", min.nc=2, max.nc=8,
            method = "complete", index = "ch")

res$All.index
res$Best.nc
res$Best.partition

## A 5-dimensional example

set.seed(1)
x<-rbind(matrix(rnorm(150,sd=0.3),ncol=5),
          matrix(rnorm(150,mean=3,sd=0.2),ncol=5),
```

```
          matrix(rnorm(150,mean=1,sd=0.1),ncol=5),
          matrix(rnorm(150,mean=6,sd=0.3),ncol=5),
          matrix(rnorm(150,mean=9,sd=0.3),ncol=5))

res<-NbClust(x, distance = "euclidean", min.nc=2, max.nc=10,
            method = "ward.D", index = "all")

res$All.index
res$Best.nc
res$All.CriticalValues
res$Best.partition

## A real data example

data<-iris[,-c(5)]
res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,
            method = "ward.D2", index = "kl")
res$All.index
res$Best.nc
res$Best.partition

res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,
            method = "kmeans", index = "hubert")
res$All.index


res<-NbClust(data, diss=NULL, distance = "manhattan", min.nc=2, max.nc=6,
            method = "complete", index = "all")
res$All.index
res$Best.nc
res$All.CriticalValues
res$Best.partition

## Examples with a dissimilarity matrix

## Data matrix is given

set.seed(1)
x<-rbind(matrix(rnorm(150,sd=0.3),ncol=3),
          matrix(rnorm(150,mean=3,sd=0.2),ncol=3),
          matrix(rnorm(150,mean=5,sd=0.3),ncol=3))
diss_matrix<- dist(x, method = "euclidean", diag=FALSE)
res<-NbClust(x, diss=diss_matrix, distance = NULL, min.nc=2, max.nc=6,
            method = "ward.D", index = "ch")
res$All.index
res$Best.nc
res$Best.partition

data<-iris[,-c(5)]
diss_matrix<- dist(data, method = "euclidean", diag=FALSE)
NbClust(data, diss=diss_matrix, distance = NULL, min.nc=2, max.nc=6,
            method = "ward.D2", index = "all")
res$All.index
```

```
res$Best.nc
res$All.CriticalValues
res$Best.partition

set.seed(1)
x<-rbind(matrix(rnorm(20,sd=0.1),ncol=2),
         matrix(rnorm(20,mean=1,sd=0.2),ncol=2),
         matrix(rnorm(20,mean=5,sd=0.1),ncol=2),
         matrix(rnorm(20,mean=7,sd=0.2),ncol=2))
diss_matrix<- dist(x, method = "euclidean", diag=FALSE)
res<-NbClust(x, diss=diss_matrix, distance = NULL, min.nc=2, max.nc=6,
            method = "ward.D2", index = "alllong")
res$All.index
res$Best.nc
res$All.CriticalValues
res$Best.partition

## Data matrix is not available. Only the dissimilarity matrix is given
## In this case, only these indices can be computed : frey, mcclain, cindex, silhouette and dunn

res<-NbClust(diss=diss_matrix, distance = NULL, min.nc=2, max.nc=6,
            method = "ward.D2", index = "silhouette")
res$All.index
res$Best.nc
res$All.CriticalValues
res$Best.partition
```

Index