User Guide for CLARA.seq

Louis Tochon

06/11/2021

Contents

Introduction

CLARA.seq is a R-package to cluster big datasets of sequences. It uses the TraMineR (http://cran.r-project.org/web/packages/TraMineR/) package to calculate distances between each sequence but this package is made for 32-bits computers, which implies that the quantity of objects is limited to 46'300 objects (because we need to compute a matrix of size $n * n$). CLARA.seq provides three different functions : CLARA, CLARANS and ROBUST FUZZY C-MEDOIDS.
The package has an index called Davies-Bouldin index to check the validity of the clustering found after the use of a clustering method.

For the functions calls, the details are in the documentation file of the package or in the manual.

Use of the functions

CLARA

CLARA has been implemented with a parameter which can determine which aspect of the clustering is the most important. Indeed the parameter *find_best_method* allows users to choose between :

- **Distance** : finding the best clustering with the mean distance method which aims to minimize the distance between each object and his medoid
- **DB** : finding the best clustering with the Davies-Bouldin index method which aims to minimize the index of the global clustering

Research of the best clutering with the mean distance method

Here is an example of the call of this function :

```
my_cluster <- clara_clust(mvad.seq,nb_cluster = 4, nb_sample = 20,
+   size_sample = 20, with.diss = TRUE)
```

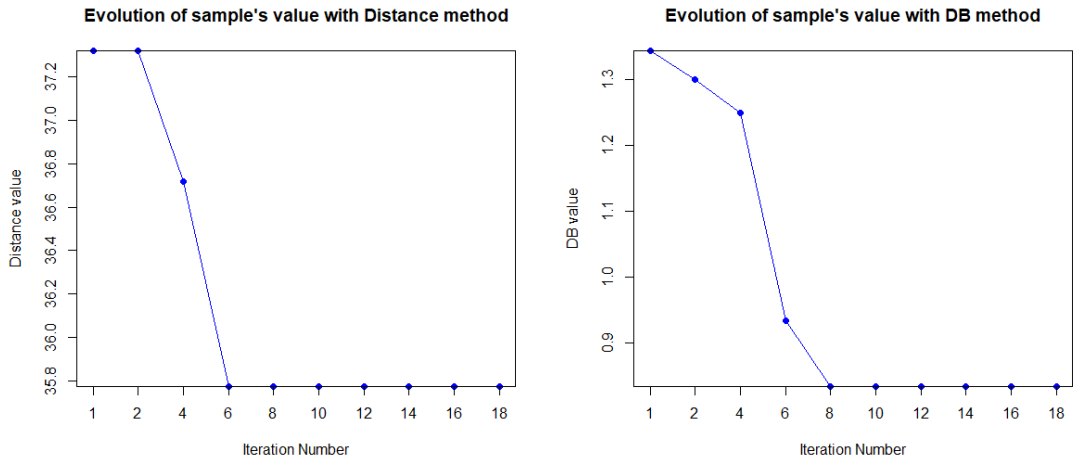
Research of the best clutering with the Davies-Bouldin index method

Here is an example of the call of this function :

```
my_cluster <- clara_clust(mvad.seq,nb_cluster = 4, nb_sample = 20,
+   size_sample = 20, with.diss = TRUE, find_best_method = "DB")
```

Outputs

Plots The graph is the representation of the value evolution for each iteration (depending of the method used) :



Console outputs The function print for the users the details of the run and the calculation time needed :

```
CLARA ALGORITHM Improved


Table of Sample's Values with Distance Method

N°1  (Med: 609, 186, 167, 309) : 37.32022471911  |  N°2  (Med: 684, 143, 40, 200)  : 40.20224719101
N°3  (Med: 639, 54, 479, 63)   : 36.82303378652  |  N°4  (Med: 259, 212, 288, 177) : 36.71629213415
N°5  (Med: 599, 691, 262, 27)  : 35.77528088764  |  N°6  (Med: 208, 509, 251, 219) : 41.72471911236
N°7  (Med: 134, 686, 21, 465)  : 40.33988404494  |  N°8  (Med: 323, 350, 2, 125)   : 40.08707516854
N°9  (Med: 48, 640, 131, 56)   : 42.2528988764   |  N°10 (Med: 525, 473, 125, 390) : 39.15443820225
N°11 (Med: 314, 464, 50, 366)  : 38.69213483146  |  N°12 (Med: 269, 628, 273, 200) : 39.68314606742
N°13 (Med: 592, 435, 560, 470) : 37.19629213483  |  N°14 (Med: 280, 506, 355, 246) : 39.38483106742
N°15 (Med: 299, 125, 604, 618) : 37.1404494382   |  N°16 (Med: 579, 186, 320, 514) : 47.43252696629
N°17 (Med: 282, 321, 354, 679) : 39.23370786517  |  N°18 (Med: 88, 247, 166, 376)  : 40.45561797753
N°19 (Med: 126, 376, 307, 19)  : 44.57078651685  |  N°20 (Med: 252, 582, 25, 549)  : 38.02247191011

[>] Minimum Index Value for Sample N°5

Calculation time : 14.4400000000001 sec.
```

```
CLARA ALGORITHM Improved


Table of Sample's Values with DB Method

N°1  (Med: 182, 71, 98, 118)  : 1.342827221118   |  N°2  (Med: 25, 126, 200, 153) : 1.299852116316
N°3  (Med: 77, 174, 189, 179) : 1.40897064430    |  N°4  (Med: 3, 96, 36, 69)     : 1.248435482041
N°5  (Med: 68, 171, 111, 170) : 1.0031981757334  |  N°6  (Med: 142, 127, 119, 66) : 0.9345574879117
N°7  (Med: 171, 116, 118, 33) : 0.834863495085   |  N°8  (Med: 127, 90, 101, 31)  : 1.479710535507
N°9  (Med: 74, 189, 92, 32)   : 1.168584125339   |  N°10 (Med: 26, 146, 89, 118)  : 1.50465609725
N°11 (Med: 22, 157, 65, 142)  : 1.445766706488   |  N°12 (Med: 119, 33, 149, 28)  : 1.264641148222
N°13 (Med: 126, 91, 179, 74)  : 0.9903531541278  |  N°14 (Med: 146, 66, 94, 111)  : 1.31802845133
N°15 (Med: 1, 22, 111, 119)   : 1.970945630127   |  N°16 (Med: 155, 139, 65, 201) : 1.655903988049
N°17 (Med: 63, 20, 93, 182)   : 1.546096302749   |  N°18 (Med: 137, 119, 33, 92)  : 1.388637698447
N°19 (Med: 179, 115, 95, 3)   : 1.508608204084   |  N°20 (Med: 173, 66, 95, 142)  : 1.186216041016

[>] Minimum Index Value for Sample N°7

Calculation time : 11.9299999999998 sec.
```

These informations can be hidden with the function *suppressMessages()*.

Object The function gives as output an object with the characteristic values of clustering.To get the dissimilarity matrix, the parameter *with_diss* must be TRUE.
The following parameters are returned :

```
$ seq #the original sequences dataset
$ id.med #the identifier of the medoids
$ clusters #the resulting clustering
$ diss #the matrix of distance of size k*n
$ evol.diss #the value of the best value found for each iteration
$ plot #function to compute the plot again
```

CLARANS

CLARANS function has been implemented but is less efficient than CLARA. Indeed CLARANS cannot be parallelized and the calculation time is more important than CLARA's. If we compare CLARA and CLARANS for the same amount of time, we get several better results for CLARA.

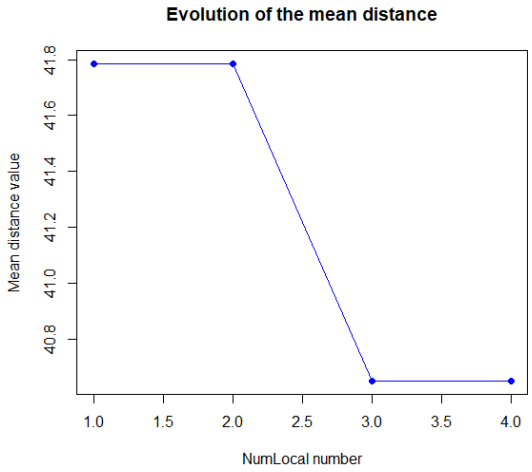
Call

This function can still be used as follow :

```
my_cluster <- clarans_clust(mvad.seq,nb_cluster = 4, maxneighbours = 20, numlocal = 4,
+   plot = TRUE)
```

Outputs

Plots The graph is the representation of the mean distance evolution :



Console outputs The function print for the users the details of the run and the calculation time needed :

```
CLARANS ALGORITHM

Table of Iteration's Distance

N°1 (Med: 163, 176, 232, 282) : 41.78651685393   N°2 (Med: 570, 306, 662, 435) : 51.21348314606
N°3 (Med: 211, 269, 479, 299) : 40.64887640449   N°4 (Med: 7, 469, 89, 562)    : 43.56741573033

[>] Minimum Distance for iteration N°3

Calculation time : 16.6700000000001 sec.
```

These informations can be hidden with the function *suppressMessages()*.

Object The function gives as output an object with the characteristic values of clustering.
The following parameters are returned :

```
 $ seq #the original sequences dataset
 $ id.med #the identifier of the medoids
 $ clusters #the resulting clustering
 $ diss #the matrix of distance of size k*n
```

ROBUST FUZZY C-MEDOIDS

This function comes from the fuzzy algorithm family and provide a fuzzy clustering based on samples from the big dataset. This variant applies a fuzzy clustering on each subset and takes the most popular objects as medoids (determined with the membership matrix). For this algorithm, the objective function to minimise is :

$$\sum_{x_j \in S} u_{i,j}^m d(med, x_j)$$

where x_j is an object in the dataset without outliers S, i is the cluster for which a medoid is being sought, and med is the medoid tested.

Warning this function has not been fully tested.

Call

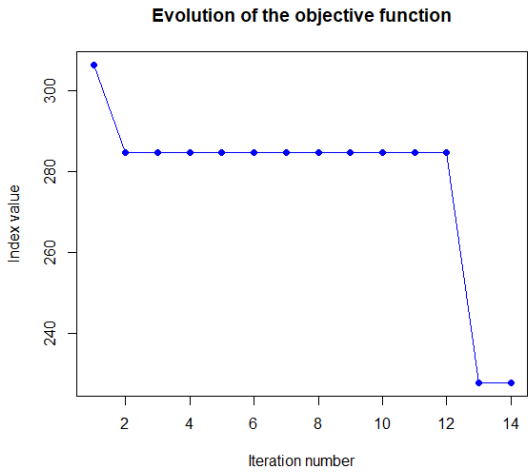
This function can be used as follow :

```
my_cluster <- fuzzy_clust(mvad.seq,nb_sample = 14, size_sample = 50, plot = TRUE,
+   threshold = 7, max_iter = 10, p=5)
```

Outputs

Plots

- The first graph is the representation of the objective function evolution :



Console outputs The function print for the users the details of the run and the calculation time needed :

```
ROBUST FUZZY C-MEDOIDS ALGORITHM

Table of objective function values

N°1  (Med: 166, 171, 1, 672)   : 306.3782076078  |  N°2  (Med: 478, 54, 446, 489)  : 284.7549317274
N°3  (Med: 345, 430, 140, 355) : 316.5071933888  |  N°4  (Med: 228, 184, 685, 504) : 351.6212640614
N°5  (Med: 48, 505, 455, 588)  : 367.0773016113  |  N°6  (Med: 481, 425, 338, 638) : 286.4788921731
N°7  (Med: 412, 120, 77, 639)  : 298.6053701533  |  N°8  (Med: 502, 158, 190, 51)  : 389.8523768593
N°9  (Med: 192, 646, 666, 363) : 387.4922466893  |  N°10 (Med: 253, 467, 214, 223) : 319.6099482256
N°11 (Med: 173, 554, 585, 495) : 334.3587832990  |  N°12 (Med: 453, 2, 144, 109)   : 331.5818583005
N°13 (Med: 497, 543, 602, 605) : 227.8510767599  |  N°14 (Med: 425, 350, 599, 396) : 239.3543957804

[>] Minimum Index for Sample N°13

Calcul time : 14.27 sec.
```

These informations can be hidden with the function *suppressMessages()*.

Object The function gives as output an object with the characteristic values of clustering.
The following parameters are returned :

```
$ seq #the original sequences dataset
$ id.med #the identifier of the medoids
$ clusters #the resulting clustering
$ diss #distance matrix
$ harm_value #harm value to adjust the threshold
$ nb_iter #number of iterations needed to find the medoids
$ membership #membership matrix
$ q #value of the index for each iteration
```