Package 'ASW'

October 31, 2023

Title Clustering Algorithms for Optimizing the Average Silhouette Width

Version 0.0.1

Author Minh Long Nguyen <edelweiss611428@gmail.com>

Maintainer Minh Long Nguyen <edelweiss611428@gmail.com>

Description This package implements clustering algorithms for optimizing the Average Silhouette Width, including PAMSil, Efficient Optimum Silhouette (effOSil), and Scalable Optimum Silhoutte (scalOSil).

License GPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

LinkingTo Rcpp

Imports Rcpp,
cluster,
stats

R topics documented:

`effOSil` *The Efficient Optimum Silhouette algorithm*

Description

This function implements the Efficient Optimum Silhouette (effOSil) algorithm.

Usage

```
effOSil(dx, K, initMethod, variant)
```

Arguments

dx	A "dist" object, which can be computed using stats::dist().
K	An integer vector (or scalar) specifying the numbers of clusters. By default, K = 2:12.
initMethod	A character vector (or string) specifying initialization methods. By default, initMethod = "average". See ?Init for more details.
variant	A character string specifying a variant. Options include "efficient" and "original". If variant = "original", the original OSil algorithm is used. If variant = "efficient", effOSil is used. By default, variant = "efficient".

Details

This function implements the Efficient Optimum Silhouette (effOSil) algorithm, an O(N) runtime improvement of the original, computationally expensive OSil algorithm proposed by Batool & Hennig (2021) (N is the number of observations). An implementation of the original OSil algorithm is also available for run time comparisions.

Value

best_clustering The effOSil clustering achieving the highest ASW value.

best_asw The highest ASW value.

k The estimated number of clusters.

clusterings The effOSil clustering solutions for all k in K.

asw The ASW values associated with the effOSil clusterings.

nIter The numbers of iterations needed for convergence.

Author(s)

Minh Long Nguyen <edelweiss611428@gmail.com>

References

Batool, F. and Hennig, C., 2021. Clustering with the average silhouette width. Computational Statistics & Data Analysis, 158, p.107190.

Examples

```
dx = dist(faithful)
effC = effOSil(dx, 2:8)
par(mfrow = c(2,1))
plot(faithful, col = effC$best_clustering, pch = 4)
plot(2:8, effC$asw, xlab = "k", ylab = "ASW")
```

Init *Initialization methods for the Optimum Silhouette algorithm*

Description

This function computes an initialization for the Optimum Silhouette algorithm.

Usage

```
Init(dx, k, initMethod)
```

Arguments

dx	dx A "dist" object, which can be computed using stats::dist().
k	An integer scalar specifying the number of clusters.
initMethod	A character vector (or string) specifying initialization methods. Options include any combination of "pam", "average", "single", "complete", "ward.D", "ward.D2", "mcquitty", "median", and "centroid". By default, initMethod = "average".

Details

This function computes an initialization for the Optimum Silhouette algorithm, but it can be used as a stand-alone clustering method.

Value

clustering An initialized clustering.

asw The ASW associated with the initialized clustering.

method The "best" initialization method.

Author(s)

Minh Long Nguyen <edelweiss611428@gmail.com>

References

Batool, F. and Hennig, C., 2021. Clustering with the average silhouette width. Computational Statistics & Data Analysis, 158, p.107190. Batool, F., 2019. Initialization methods for optimum average silhouette width clustering. arXiv preprint arXiv:1910.08644.

Examples

```
x = faithful
dx = dist(x)
Initres = Init(dx, 2, c("pam", "average", "complete"))
plot(x, col = Initres$clustering, pch = 4)
print(paste(Initres$method, "achieves the highest ASW value"))
```

PAMSil *The PAMSil algorithm*

Description

This function implements the PAMSil algorithm.

Usage

```
PAMSil(dx, K)
```

Arguments

dx	A "dist" object, which can be computed using stats::dist().
K	An integer vector (or scalar) specifying the numbers of clusters. By default, K = 2:12.

Details

This function implements the PAMSil algorithm proposed by Van der Laan et al. (2003). It is a k-medoids clustering algorithm whose objective function is the Average Silhouette Width.

Value

best_clustering The PAMSil clustering achieving the highest ASW value.

best_asw The highest ASW value.

best_medoids The medoids associated with the clustering maximize the ASW.

k The estimated number of clusters.

clusterings The PAMSil clustering solutions for all k in K.

asw The ASW values associated with the PAMSil clusterings.

medoids The medoids associated with the clustering solutions.

nIter The numbers of iterations needed for convergence.

Author(s)

Minh Long Nguyen <edelweiss611428@gmail.com>

References

Van der Laan, M., Pollard, K. and Bryan, J., 2003. A new partitioning around medoids algorithm. Journal of Statistical Computation and Simulation, 73(8), pp.575-584.

Examples

```
dx = dist(faithful)
pamsilC = PAMSil(dx, 2:8)
par(mfrow = c(2,1))
plot(faithful, col = pamsilC$best_clustering, pch = 4)
plot(2:8, pamsilC$asw, xlab = "k", ylab = "ASW")
```

scalOSil *The Scalable Optimum Silhouette algorithm*

Description

This function implements the Scalable Optimum Silhouette algorithm.

Usage

```
scalOSil(dx, K, n, ns, rep, initMethod, variant)
```

Arguments

dx	A "dist" object, which can be computed using stats::dist().
K	An integer vector (or scalar) specifying the numbers of clusters. By default, K = 2:12.
n	An integer specifying the sample size. If not specified (NULL), n is set to 0.2*N where N is the number of observations.
ns	An integer specifying the number of random samples used in each instance. By default, ns = 1.
rep	An integer specifying the number of scalOSil instances. By default, rep = 10.
initMethod	A character vector (or string) specifying initialization methods. By default, initMethod = "average". See ?Init for more details.
variant	A character string specifying a variant. Options include "scalable" and "original". If variant = "original", the original FOSil algorithm is used. If variant = "scalable", scalOSil is used. By default, variant = "scalable".

Details

This function implements the Scalable Optimum Silhouette (scalOSil) algorithm, an O(n) runtime improvement of the original, computationally expensive Fast OSil (FOSil) algorithm proposed by Batool & Hennig (2021) (n is the sample size). An implementation of the original FOSil algorithm is also available for run time comparisions.

Value

best_clustering The scalOSil clustering achieving the highest ASW value.

best_asw The highest ASW value.

k The estimated number of clusters.

clusterings The scalOSil clustering solutions for all k in K.

asw The ASW values associated with the scalOSil clusterings.

Author(s)

Minh Long Nguyen <edelweiss611428@gmail.com>

References

Batool, F. and Hennig, C., 2021. Clustering with the average silhouette width. Computational Statistics & Data Analysis, 158, p.107190.

Examples

```
dx = dist(faithful)
scalC = scalOSil(dx, 2:8)
par(mfrow = c(2,1))
plot(faithful, col = scalC$best_clustering, pch = 4)
plot(2:8, scalC$asw, xlab = "k", ylab = "ASW")
```

`Silhouette` *Silhouette Width*

Description

This function computes the Silhouette Widths for all data points in the dataset.

Usage

```
Silhouette(C, dx)
```

Arguments

`C`	An integer vector specifying a k-partition of the dataset. min(C) must be 1 and max(C) must be k.
`dx`	A "dist" object, which can be computed using stats::dist().

Value

A numeric matrix of class "silhouette" containing three columns

cluster A clustering of the dataset.

neighbor The clustering labels of the nearest clusters for all data points.

sil_width The silhouette widths of data points.

Author(s)

Minh Long Nguyen <edelweiss611428@gmail.com>

References

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53–65.

Examples

```
library("cluster")
dx = dist(faithful)
C = pam(dx, 2)$clustering
plot(Silhouette(C,dx))
```

Index