Package ‘GIC’

December 15, 2021

Type Package
Title A General Iterative Clustering Algorithm
Version 1.0.0
Description An iterative algorithm that improves the proximity matrix (PM) from a random forest (RF) and the resulting clusters as measured by the silhouette score.
License GPL-3
Encoding UTF-8
Imports randomForest, cluster, ggplot2
RoxygenNote 7.1.2
NeedsCompilation no
Author Ziqiang Lin [aut, cre], Eugene Laska [aut], Carole Siegel [aut]
Maintainer Ziqiang Lin <linziqiang0314@gmail.com>
Repository CRAN
Date/Publication 2021-12-15 08:30:02 UTC

R topics documented:

GIC ................................. 1
iteration .......................... 3

Index 5

GIC A General Iterative Clustering Algorithm
Description

An algorithm improves the proximity matrix (PM) from a random forest (RF) and the resulting clusters from an arbitrary cluster algorithm, such as PAM, as measured by the silhouette_score. The first PM that uses unlabeled data is produced by one of many ways to provide pseudo labels for a RF. After running a cluster program on the resulting initial PM, cluster labels are obtained. These are used as labels with the same feature data to grow a new RF yielding an updated proximity matrix. This is entered into the clustering program and the process is repeated until convergence.

Usage

GIC(data,cluster,initial="breiman",ntree=500,
   label=sample(1:cluster,nrow(data),replace = TRUE))

Arguments

data an input dataframe without label
cluster The number of clusters in the solution
initial A method to calculate initial clusters to begin the iteration (default breiman).
   breiman: using Breimans' unsupervised method to find initial clusters, or purpose: using Siegel and her colleagues’ purposeful clustering method to find initial clusters
ntree the number of trees (default 500).
label A truth set of labels, only required if purpose is used as the method to find the initial PM

Details

This code include Breimans’ unsupervised method and Siegel and her colleagues’ purposeful clustering method to calculate initial labels To imput user specified initial labels, please use the function initial

Value

An object of class GIC, which is a list with the following components:

PAM output final PAM information
randomforest output final randomforest information
clustering A vector of integers indicating the cluster to which each point is allocated.
silhouette_score A value of mean silhouette score for clusters
plot A scatter plot which X-axis, y-axis, and color are first important feature, second important feature, and final clusters, respectively.
References


Examples

```r
data(iris)
##Using breiman's method
rs=GIC(iris[,1:4],3,ntree=100)
print(rs$clustering)
```

---

**iteration**

*A General Iterative Clustering Algorithm*

**Description**

An algorithm that improves the proximity matrix (PM) from a random forest (RF) and the resulting clusters from an arbitrary cluster algorithm as measured by the silhouette score. The initial PM, that uses unlabeled data, is produced by one of many ways to provide pseudo labels for a RF. After running a cluster program on the resulting initial PM, cluster labels are obtained. These are used as labels with the same feature data to grow a new RF yielding an updated proximity matrix. This is entered into the clustering program and the process is repeated until convergence.

**Usage**

```r
iteration(data,initiallabel,ntree=500)
```

**Arguments**

- **data**: an input dataframe without label
- **initiallabel**: a vector of label to begin with
- **ntree**: the number of trees (default 500).

**Details**

This code requires initial labels as input, which can be obtained by any method of the users choice. As an alternative, Breimans\’ unsupervised method or Siegel and her colleagues’ purposeful clustering method to obtain initial labels, use the function GIC.
Value

An object of class `iteration`, which is a list with the following components:

- **PAM**: output final PAM information
- **randomforest**: output final randomforest information
- **clustering**: A vector of integers indicating the cluster to which each point is allocated.
- **silhouette_score**: A value of mean silhouette score for clusters
- **plot**: A scatter plot which X-axis, y-axis, and color are first important feature, second important feature, and final clusters, respectively.

References


Examples

```r
data(iris)
##Using KMEANS to find initial label
cl=kmeans(iris[,1:4],3)
###Doing GIC to find final clustering
rs=iteration(iris[,1:4],cl$cluster,ntree=100)
print(rs$clustering)
```
Index

GIC, 1

iteration, 3