Package ‘densityClust’

February 6, 2016

Type Package

Title Clustering by Fast Search and Find of Density Peaks

Version 0.2.1

Date 2016-02-06

Maintainer Thomas Lin Pedersen <thomasp85@gmail.com>

Description An implementation of the clustering algorithm described by Alex Rodriguez and Alessandro Laio (Science, 2014 vol. 344), along with tools to inspect and visualize the results.

License GPL (>= 2)

Suggests testthat

LinkingTo Rcpp

Imports Rcpp

RoxygenNote 5.0.1

NeedsCompilation yes

Author Thomas Lin Pedersen [aut, cre],
Sean Hughes [aut]

Repository CRAN

Date/Publication 2016-02-06 13:47:07

R topics documented:

densityClust-package *Clustering by fast search and find of density peaks*

Description

This package implement the clustering algorithm described by Alex Rodriguez and Alessandro Laio (2014). It provides the user with tools for generating the initial rho and delta values for each observation as well as using these to assign observations to clusters. This is done in two passes so the user is free to reassign observations to clusters using a new set of rho and delta thresholds, without needing to recalculate everything.

Details

Plotting

Two types of plots are supported by this package, and both mimics the types of plots used in the publication for the algorithm. The standard plot function produces a decision plot, with optional colouring of cluster peaks if these are assigned. Furthermore `plotMDS` performs a multidimensional scaling of the distance matrix and plots this as a scatterplot. If clusters are assigned observations are coloured according to their assignment.

Cluster detection

The two main functions for this package are `densityClust` and `findClusters`. The former takes a distance matrix and optionally a distance cutoff and calculates rho and delta for each observation. The latter takes the output of `densityClust` and make cluster assignment for each observation based on a user defined rho and delta threshold. If the thresholds are not specified the user is able to supply them interactively by clicking on a decision plot.

References

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492-1496. doi:10.1126/science.1242072

See Also

`densityClust`, `findClusters`, `plotMDS`

Examples

```
irisDist <- dist(iris[,1:4])
irisClust <- densityClust(irisDist, gaussian=TRUE)
plot(irisClust) # Inspect clustering attributes to define thresholds

irisClust <- findClusters(irisClust, rho=2, delta=2)
plotMDS(irisClust)
split(iris[,5], irisClust$clusters)
```

clustered *Check whether a densityCluster object have been clustered*

Description

This function checks whether findClusters has been performed on the given object and returns a boolean depending on the outcome

Usage

```
clustered(x)

## S3 method for class 'densityCluster'
clustered(x)
```

Arguments

x	A densityCluster object

Value

TRUE if findClusters have been performed, otherwise FALSE

clusters *Extract cluster membership from a densityCluster object*

Description

This function allows the user to extract the cluster membership of all the observations in the given densityCluster object. The output can be formatted in two ways as described below. Halo observations can be chosen to be removed from the output.

Usage

```
clusters(x, ...)

## S3 method for class 'densityCluster'
clusters(x, as.list = FALSE, halo.rm = TRUE, ...)
```

Arguments

x	The densityCluster object. findClusters must have been performed prior to this call to avoid throwing an error.
...	Currently ignored
as.list	Should the output be in the list format. Defaults to FALSE
halo.rm	Logical. should halo observations be removed. Defaults to TRUE

Details

Two formats for the output are available. Either a vector of integers denoting for each observation, which cluster the observation belongs to. If halo observations are removed, these are set to NA. The second format is a list with a vector for each group containing the index for the member observations in the group. If halo observations are removed their indexes are omitted. The list format correspond to the following transform of the vector format `split(1:length(clusters), clusters)`, where `clusters` are the cluster information in vector format.

Value

A vector or list with cluster memberships for the observations in the initial distance matrix

---

`densityClust` *Calculate clustering attributes based on the densityClust algorithm*

Description

This function takes a distance matrix and optionally a distance cutoff and calculates the values necessary for clustering based on the algorithm proposed by Alex Rodrigues and Alessandro Laio (see references). The actual assignment to clusters are done in a later step, based on user defined threshold values.

Usage

```
densityClust(distance, dc, gaussian = FALSE)
```

Arguments

`distance`	A distance matrix
`dc`	A distance cutoff for calculating the local density. If missing it will be estimated with estimateDc(distance)
`gaussian`	Logical. Should a gaussian kernel be used to estimate the density (defaults to FALSE)

Details

The function calculates rho and delta for the observations in the provided distance matrix. If a distance cutoff is not provided this is first estimated using `estimateDc` with default values.

The information kept in the densityCluster object is:

rho A vector of local density values

delta A vector of minimum distances to observations of higher density

distance The initial distance matrix

dc The distance cutoff used to calculate rho

threshold A named vector specifying the threshold values for rho and delta used for cluster detection

peaks A vector of indexes specifying the cluster center for each cluster

clusters A vector of cluster affiliations for each observation. The clusters are referenced as indexes in the peaks vector

halo A logical vector specifying for each observation if it is considered part of the halo

Before running findClusters the threshold, peaks, clusters and halo data is NA.

Value

A densityCluster object. See details for a description.

References

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492-1496. doi:10.1126/science.1242072

See Also

estimateDc, findClusters

Examples

```
irisDist <- dist(iris[,1:4])
irisClust <- densityClust(irisDist, gaussian=TRUE)
plot(irisClust) # Inspect clustering attributes to define thresholds

irisClust <- findClusters(irisClust, rho=2, delta=2)
plotMDS(irisClust)
split(iris[,5], irisClust$clusters)
```

estimateDc *Estimate the distance cutoff for a specified neighbor rate*

Description

This function calculates a distance cutoff value for a specific distance matrix that makes the average neighbor rate (number of points within the distance cutoff value) fall between the provided range. The authors of the algorithm suggests aiming for a neighbor rate between 1 and 2 percent, but also states that the algorithm is quite robust with regards to more extreme cases.

Usage

```
estimateDc(distance, neighborRateLow = 0.01, neighborRateHigh = 0.02)
```

Arguments

`distance`	A distance matrix
`neighborRateLow`	The lower bound of the neighbor rate
`neighborRateHigh`	The upper bound of the neighbor rate

Value

A numeric value giving the estimated distance cutoff value

References

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492-1496. doi:10.1126/science.1242072

Examples

```
irisDist <- dist(iris[,1:4])
estimateDc(irisDist)
```

findClusters *Detect clusters in a densityCluster obejct*

Description

This function uses the supplied rho and delta thresholds to detect cluster peaks and assign the rest of the observations to one of these clusters. Furthermore core/halo status is calculated. If either rho or delta threshold is missing the user is presented with a decision plot where they are able to click on the plot area to set the treshold. If either rho or delta is set, this takes presedence over the value found by clicking.

Usage

```
findClusters(x, ...)

## S3 method for class 'densityCluster'
findClusters(x, rho, delta, plot = FALSE, ...)
```

Arguments

`x`	A densityCluster object as produced by `densityClust`
`...`	Additional parameters passed on to `findClusters.densityCluster`
`rho`	The threshold for local density when detecting cluster peaks
`delta`	The threshold for minimum distance to higher density when detecting cluster peaks
`plot`	Logical. Should a decision plot be shown after cluster detection

Value

A densityCluster object with clusters assigned to all observations

References

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492-1496. doi:10.1126/science.1242072

See Also

findClusters

Examples

```
irisDist <- dist(iris[,1:4])
irisClust <- densityClust(irisDist, gaussian=TRUE)
plot(irisClust) # Inspect clustering attributes to define thresholds

irisClust <- findClusters(irisClust, rho=2, delta=2)
plotMDS(irisClust)
split(iris[,5], irisClust$clusters)
```

---

plotMDS *Plot observations using multidimensional scaling and colour by cluster*

Description

This function produces an MDS scatterplot based on the distance matrix of the densityCluster object, and, if clusters are defined, colours each observation according to cluster affiliation. Observations belonging to a cluster core is plotted with filled circles and observations belonging to the halo with hollow circles.

Usage

```
plotMDS(x, ...)
```

Arguments

x	A densityCluster object as produced by densityClust
...	Additional parameters. Currently ignored

See Also

densityClust

Examples

```
irisDist <- dist(iris[,1:4])
irisClust <- densityClust(irisDist, gaussian=TRUE)
plot(irisClust) # Inspect clustering attributes to define thresholds

irisClust <- findClusters(irisClust, rho=2, delta=2)
plotMDS(irisClust)
split(iris[,5], irisClust$clusters)
```

Index