Package ‘Clustering’

June 22, 2022

Type Package
Title Techniques for Evaluating Clustering
Version 1.7.7
Date 2022-06-06
Author Luis Alfonso Perez Martos [aut, cre]
Maintainer Luis Alfonso Perez Martos <lapm0001@gmail.com>
Description The design of this package allows us to run different clustering packages and compare the results between them, to determine which algorithm behaves best from the data provided.
URL https://github.com/laperez/clustering
Depends R (>= 3.5.0)
License GPL (>= 2)
Encoding UTF-8
LazyData true
RoxygenNote 7.1.1
Repository CRAN
Imports amap, apcluster, cluster, ClusterR, data.table, doParallel, dplyr, foreach, future, ggplot2, gmp, methods, pracma, pvclust, shiny, sqldf, stats, tools, utils, xtable, toOrdinal
Suggests DT, shinyalert, shinyFiles, shinyjs, shinythemes, shinyWidgets, tidyverse, shinycssloaders
NeedsCompilation no
Date/Publication 2022-06-22 09:40:12 UTC

R topics documented:

appClustering ...................................................... 2
basketball ........................................................... 3
best_ranked_external_metrics .................................... 3
best_ranked_internal_metrics ................................. 4
appClustering

Description
Method that allows us to execute the main algorithm in graphic interface mode instead of through the console.

Usage
appClustering()

Details
The operation of this method is to generate a graphical user interface to be able to execute the clustering algorithm without knowing the parameters. Its operation is very simple, we can change the values and see the behavior quickly.

Value
GUI with the parameters of the algorithm and their representation in tables and graphs.
This data set contains a series of statistics (5 attributes) about 96 basketball players:

```
Description
This data set contains a series of statistics about basketball players:

Usage
data(basketball)

Format
A data frame with 96 observations on 5 variables:

   assists_per_minute Real  average number of assistances per minute
   height Integer     height of the player
   time_played Real    time played by the player
   age Integer        number of years of the player
   points_per_minute Real average number of points per minute

Source
KEEL, <http://www.keel.es/>
```

```
best_ranked_external_metrics
Best rated external metrics.

Description
Method in charge of searching for each algorithm those that have the best external classification. Method that looks for those external attribute that are better classified, making use of the var column. In this way of discard attribute and only work with those that give the best response to the algorithm in question.

Usage
best_ranked_external_metrics(df)

Arguments
  df       Matrix or data frame with the result of running the clustering algorithm.
```
value

Returns a data.frame with the best classified external attribute.

Examples

```r
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='gmm',
    metrics=c("Recall")
)

best_ranked_external_metrics(df = result)
```

---

**Best ranked internal metrics**

**Best rated internal metrics.**

**Description**

Method in charge of searching for each algorithm those that have the best internal classification. Method that looks for those internal attributes that are better classified, making use of the Var column. In this way we discard the attributes and only work with those that give the best response to the algorithm in question.

**Usage**

```r
best_ranked_internal_metrics(df)
```

**Arguments**

- **df**
  - Matrix or data frame with the result of running the clustering algorithm.

**Value**

Returns a data.frame with the best classified internal attributes.

**Examples**

```r
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='gmm',
    metrics=c("Recall")
)

best_ranked_internal_metrics(df = result)
```
Data from an experiment on the affects of machine adjustments on the time to count bolts.

Description

A manufacturer of automotive accessories provides hardware, e.g. nuts, bolts, washers and screws, to fasten the accessory to the car or truck. Hardware is counted and packaged automatically. Specifically, bolts are dumped into a large metal dish. A plate that forms the bottom of the dish rotates counterclockwise. This rotation forces bolts to the outside of the dish and up along a narrow ledge. Due to the vibration of the dish caused by the spinning bottom plate, some bolts fall off the ledge and back into the dish. The ledge spirals up to a point where the bolts are allowed to drop into a pan on a conveyor belt. As a bolt drops, it passes by an electronic eye that counts it. When the electronic counter reaches the preset number of bolts, the rotation is stopped and the conveyor belt is moved forward.

Usage

data(bolts)

Format

A data frame with 40 observations on 8 variables:

RUN
Integer is the order in which the data were collected

SPEED1
Integer a speed setting that controls the speed of rotation of the plate at the bottom of the dish

TOTAL
Integer total number of bolts (TOTAL) to be counted

SPEED2
Integer a second speed setting that is used to change the speed of rotation (usually slowing it down) for the last few bolts

NUMBER2
Integer the number of bolts to be counted at this second speed
**SENSInteger**  the sensitivity of the electronic eye  
**TIMEReal**  The measured response is the time, in seconds  
**T20BOLTRReal**  In order to put times on a equal footing the response to be analyzed is the time to count 20 bolts

**Details**

There are several adjustments on the machine that affect its operation. These include; a speed setting that controls the speed of rotation (SPEED1Integer) of the plate at the bottom of the dish, a total number of bolts (TOTAL) to be counted, a second speed setting (SPEED2Integer) that is used to change the speed of rotation (usually slowing it down) for the last few bolts, the number of bolts to be counted at this second speed (NUMBER2Integer), and the sensitivity of the electronic eye (SENSInteger). The sensitivity setting is to insure that the correct number of bolts are counted. Too few bolts packaged causes customer complaints. Too many bolts packaged increases costs. For each run conducted in this experiment the correct number of bolts was counted. From an engineering standpoint if the correct number of bolts is counted, the sensitivity should not affect the time to count bolts. The measured response is the time (TIMEReal), in seconds, it takes to count the desired number of bolts. In order to put times on a equal footing the response to be analyzed is the time to count 20 bolts (T20BOLTRReal). Below are the data for 40 combinations of settings.  

**Source**

KEEL, <http://www.keel.es/>

---

**clustering**  
*Clustering algorithm.*

**Description**

Discovering the behavior of attributes in a set of clustering packages based on evaluation metrics.

**Usage**

```r
clustering(
  path = NULL,
  df = NULL,
  packages = NULL,
  algorithm = NULL,
  min = 3,
  max = 4,
  metrics = NULL
)
```
Arguments

path
The path of file. NULL It is only allowed to use path or df but not both at the same time. Only files in .dat, .csv or arff format are allowed.

df
data matrix or data frame, or dissimilarity matrix. NULL If you want to use training and test basketball attributes.

packages
character vector with the packets running the algorithm.
NULL The seven packages implemented are: cluster, ClusterR, amap, apcluster, pvclust.
By default runs all packages.

algorithm
character vector with the algorithms implemented within the package. NULL The algorithms implemented are: hclust,apclusterK,agnes,clara,diana,fanny,mona,pam,gmm, kmeans_arma,kmeans_rcpp,mini_kmeans,pvclust.

min
An integer with the minimum number of clusters This data is necessary to indicate the minimum number of clusters when grouping the data. The default value is 3.

max
An integer with the maximum number of clusters. This data is necessary to indicate the maximum number of clusters when grouping the data. The default value is 4.

metrics
Character vector with the metrics implemented to evaluate the distribution of the data in clusters. NULL The night metrics implemented are: Entropy, Variation_information, Precision,Recall,F_measure,Fowlkes_mallows_index,Connectivity,Dunn and Silhouette.

Details
The operation of this algorithm is to evaluate how the attributes of a dataset or a set of datasets behave in different clustering algorithms. To do this, it is necessary to indicate the type of evaluation you want to make on the distribution of the data. To be able to execute the algorithm it is necessary to indicate the number of clusters.

min and max, the algorithms algorithm or packages.

packages that we want to cluster and the metrics metrics.

Value
A matrix with the result of running all the metrics of the algorithms contained in the packages indicated. We also obtain information with the types of metrics, algorithms and packages executed.

- result It is a list with the algorithms, metrics and variables defined in the execution of the algorithm.
- has_internal_metrics Boolean field to indicate if there are internal metrics such as: dunn, silhouette and connectivity.
- has_external_metrics Boolean field to indicate if there are external metrics such as: precision, recall, f-measure, entropy, variation information and fowlkes-mallows.
- algorithms_execute Character vector with the algorithms executed. These algorithms have been mentioned in the definition of the parameters.
- measures_execute Character vector with the measures executed. These measures have been mentioned in the definition of the parameters.

Examples

classification(
    df = cluster::agriculture,
    min = 3,
    max = 3,
    algorithm='clara',
    metrics=c('Precision')
)

convert_toOrdinal

Method to convert columns to ordinal.

Description

Method to convert columns to ordinal.

Usage

convert_toOrdinal(df)

Arguments

df data frame with the results.

Value

convert data frame to Ordinal.
**evaluate_best_validation_external_by_metrics**

*Evaluates algorithms by measures of dissimilarity based on a metric.*

**Description**

Method that calculates which algorithm and which metric behaves best for the datasets provided.

**Usage**

```r
evaluate_best_validation_external_by_metrics(df, metric)
```

**Arguments**

- **df**: Data matrix or data frame with the result of running the clustering algorithm.
- **metric**: String with the metric.

**Details**

Method groups the data by algorithm and distance measure, instead of obtaining the best attribute from the data set.

**Value**

A data.frame with the algorithms classified by measures of dissimilarity.

**Examples**

```r
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='kmeans_rcpp',
    metrics=c("F_measure"))

evaluate_best_validation_external_by_metrics(result,'F_measure')
```
evaluate_best_validation_internal_by_metrics

Evaluates algorithms by measures of dissimilarity based on a metric.

Description

Method that calculates which algorithm and which metric behaves best for the datasets provided.

Usage

evaluate_best_validation_internal_by_metrics(df, metric)

Arguments

df Data matrix or data frame with the result of running the clustering algorithm.
metric It’s a string with the metric to evaluate.

Details

This method groups the data by algorithm and distance measure, instead of obtaining the best attribute from the data set.

Value

A data.frame with the algorithms classified by measures of dissimilarity.

Examples

```r
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='gmm',
    metrics=c("Precision","Connectivity")
  )

evaluate_best_validation_internal_by_metrics(result,"Connectivity")
```
evaluate_validation_external_by_metrics

Evaluate external validations by algorithm.

Description

Method that calculates which algorithm behaves best for the datasets provided.

Usage

evaluate_validation_external_by_metrics(df)

Arguments

df data matrix or data frame with the result of running the clustering algorithm.

Details

It groups the results of the execution by algorithms.

Value

A data.frame with all the algorithms that obtain the best results regardless of the dissimilarity measure used.

Examples

```r
result = clustering(
  df = cluster::agriculture,
  min = 4,
  max = 4,
  algorithm='kmeans_arma',
  metrics=c("Precision")
)

evaluate_validation_external_by_metrics(result)
```
evaluate_validation_internal_by_metrics

Evaluate internal validations by algorithm.

Description

Method that calculates which algorithm behaves best for the datasets provided.

Usage

evaluate_validation_internal_by_metrics(df)

Arguments

df data matrix or data frame with the result of running the clustering algorithm.

Details

It groups the results of the execution by algorithms.

Value

A data.frame with all the algorithms that obtain the best results regardless of the dissimilarity measure used.

Examples

```r
result = clustering(
  df = cluster::agriculture,
  min = 4,
  max = 5,
  algorithm='kmeans_rcpp',
  metrics=c("Recall","Silhouette")
)

evaluate_validation_internal_by_metrics(result)

## Not run:
evaluate_validation_internal_by_metrics(result$result)

## End(Not run)
```
export_file_external  

Export result of external metrics in latex.

Description
Method that exports the results of external measurements in latex format to a file.

Usage
export_file_external(df, path = NULL)

Arguments
- df: It's a dataframe that contains as a parameter a table in latex format with the results of the external validations.
- path: It's a string with the path to a directory where a file is to be stored in latex format.

Details
When we work in latex format and we need to create a table to export the results, with this method we can export the results of the clustering algorithm to latex.

Value
A file in Latex format with the results of the external metrics.

Examples
```r
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='gmm',
    metrics=c("Precision")
)
export_file_external(result)
file.remove("external_data.tex")
```
export_file_internal  \textit{Export result of internal metrics in latex.}

\section*{Description}
Method that exports the results of internal measurements in latex format to a file.

\section*{Usage}

\begin{verbatim}
export_file_internal(df, path = NULL)
\end{verbatim}

\section*{Arguments}
\begin{description}
\item[df] It's a dataframe that contains as a parameter a table in latex format with the results of the internal validations.
\item[path] It's a string with the path to a directory where a file is to be stored in latex format.
\end{description}

\section*{Details}
When we work in latex format and we need to create a table to export the results, with this method we can export the results of the clustering algorithm to latex.

\section*{Value}
A file in Latex format with the results of the internal metrics.

\section*{Examples}

\begin{verbatim}
result = clustering(
    df = cluster::agriculture,
    min = 4,
    max = 5,
    algorithm='gmm',
    metrics=c("Recall","Dunn")
)

export_file_internal(result)
file.remove("internal_data.tex")
\end{verbatim}
**plot_clustering**  

**Graphic representation of the evaluation measures.**

---

**Description**

Graphical representation of the evaluation measures grouped by cluster.

**Usage**

```r
plot_clustering(df, metric)
```

**Arguments**

- `df`: data matrix or data frame with the result of running the clustering algorithm.
- `metric`: it’s a string with the name of the metric select to evaluate.

**Details**

In certain cases the review or filtering of the data is necessary to select the data, that is why thanks to the graphic representations this task is much easier. Therefore with this method we will be able to filter the data by metrics and see the data in a graphical way.

**Value**

Generate an image with the distribution of the clusters by metrics.

**Examples**

```r
result = clustering(
    df = cluster::agriculture, 
    min = 4, 
    max = 5, 
    algorithm='gmm', 
    metrics=c("Precision")
)

plot_clustering(result,c("Precision"))
```
result_external_algorithm_by_metric

*External results by algorithm.*

**Description**

It is used for obtaining the results of an algorithm indicated as a parameter grouped by number of clusters.

**Usage**

```
result_external_algorithm_by_metric(df, metric)
```

**Arguments**

- `df`: data matrix or data frame with the result of running the clustering algorithm.
- `metric`: It’s a string with the metric to evaluate.

**Value**

A data.frame with the results of the algorithm indicated as parameter.

**Examples**

```r
result = clustering(
  df = cluster::agriculture,
  min = 4,
  max = 5,
  algorithm='gmm',
  metrics=c("Precision")
)

result_external_algorithm_by_metric(result,'Precision')
```

result_internal_algorithm_by_metric

*Internal results by algorithm*

**Description**

It is used for obtaining the results of an algorithm indicated as a parameter grouped by number of clusters.
sort.clustering

Usage
result_internal_algorithm_by_metric(df, metric)

Arguments
df data matrix or data frame with the result of running the clustering algorithm.
metric It’s a string with the metric we want to evaluate your results.

Value
A data.frame with the results of the algorithm indicated as parameter.

Examples

result = clustering(
  df = cluster::agriculture,
  min = 4,
  max = 5,
  algorithm='gmm',
  metrics=c("Recall","Silhouette")
)

result_internal_algorithm_by_metric(result,'Silhouette')

sort.clustering Returns the clustering result sorted by a set of metrics.

Description
This function receives a clustering object and sorts the columns by parameter. By default it performs
sorting by the algorithm field.

Usage
## S3 method for class 'clustering'
sort(x, decreasing = TRUE, ...)

Arguments
x It’s an clustering object.
decreasing A logical indicating if the sort should be increasing or decreasing. By default,
            decreasing.
... Additional parameters as "by", a String with the name of the evaluation measure
to order by. Valid values are: Algorithm, Distance,Clusters, Data, Var,
            Time, Entropy, Variation_information, Precision, Recall, F_measure,
            Fowlkes_mallows_index, Connectivity, Dunn, Silhouette and TimeAtt.
Details

The additional argument in "..." is the 'by' argument, which is a array with the name of the evaluation measure to order by. Valid value are: Algorithm, Distance, Clusters, Data, Var, Time, Entropy, Variation_information, Precision, Recall, F_measure, Fowlkes_mallows_index, Connectivity, Dunn, Silhouette, TimeAtt.

Value

Another clustering object with the evaluation measures sorted

Examples

library(Clustering)

result <- clustering(df = cluster::agriculture, min = 4, max = 4, algorithm='gmm', metrics='Recall')

sort(result, FALSE, 'Recall')

stock

The data provided are daily stock prices from January 1988 through October 1991, for ten aerospace companies.

Description

The data provided are daily stock prices from January 1988 through October 1991, for ten aerospace companies.

Usage

data(stock)

Format

A data frame with 950 observations on 10 variables:

The data provided are daily stock prices from January 1988 through October 1991, for ten aerospace companies.

Company1  company1 details
Company2  company2 details
Company3  company3 details
Company4  company4 details
Company5  company5 details
Company6  company6 details
The study was performed at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital. The data were transferred to electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences.

Description

The study was performed at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital. The data were transferred to electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences.

Usage

data(stulong)

Format

A data frame with 1417 observations on 5 variables.

The study was performed at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital. The data were transferred to electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences.

a1 Height
a2 Weight
a3 Blood pressure I systolic (mm Hg)
a4 Blood pressure I diastolic (mm Hg)
a5 Percentage Cholesterol in mg

Source

KEEL, <http://www.keel.es/>
transform_dataset

Method for filtering external columns of a dataset.

Description

Method for filtering external columns of a dataset.

Usage

transform_dataset(df)

Arguments

df  Data frame with clustering results.

Value

data frame filtered with the columns of the external measurements.

Exists internal measure

transform_dataset_internal

Method for filtering internal columns of a dataset.

Description

Method for filtering internal columns of a dataset.

Usage

transform_dataset_internal(df)

Arguments

df  data frame with clustering results.

Value

data frame filtered with the columns of the internal measurements.

Exists internal measure
weather

One of the most known testing data sets in machine learning. This data sets describes several situations where the weather is suitable or not to play sports, depending on the current outlook, temperature, humidity and wind.

Description

One of the most known testing data sets in machine learning. This data sets describes several situations where the weather is suitable or not to play sports, depending on the current outlook, temperature, humidity and wind.

Usage

data(weather)

Format

A data frame with 14 observations on 5 variables:

One of the most known testing data sets in machine learning. This data sets describes several situations where the weather is suitable or not to play sports, depending on the current outlook, temperature, humidity and wind.

Outlook sunny, overcast, rainy
Temperature hot, mild, cool
Humidity high, normal
Windy true, false
Play yes, no

Source

KEEL, <http://www.keel.es/>

[.clustering Filter metrics in a clustering object returning a new clustering object.

Description

Generates a new filtered clustering object.

Usage

## S3 method for class 'clustering'
clustering[condition = TRUE]
Arguments

clustering  The clustering object to filter.
condition  Expression to filter the clustering object.

Details

This function allows you to filter the data set for a given evaluation metric. The evaluation metrics available are: Algorithm, Distance, Clusters, Data, Var, Time, Entropy, Variation_information, Precision, Recall, F_measure, Fowlkes_mallows_index, Connectivity, Dunn, Silhouette and TimeAtt.

Value

A clustering object filtered from the input parameters.

Examples

library(Clustering)

result <- clustering(df = Clustering::basketball, algorithm = 'clara',
min=3, max=4, metrics = c('Precision','Recall'))

result[Precision > 0.14 & Recall > 0.11]
## Index

* datasets
  - basketball, 3
  - bolts, 5
  - stock, 18
  - stulong, 19
  - weather, 21
  - [.clustering, 21

appClustering, 2

basketball, 3
best_ranked_external_metrics, 3
best_ranked_internal_metrics, 4
bolts, 5

clustering, 6
custom_toOrdinal, 8

evaluate_best_validation_external_by_metrics, 9
evaluate_best_validation_internal_by_metrics, 10
evaluate_validation_external_by_metrics, 11
evaluate_validation_internal_by_metrics, 12
export_file_external, 13
export_file_internal, 14

plot_clustering, 15

result_external_algorithm_by_metric, 16
result_internal_algorithm_by_metric, 16

sort_clustering, 17
stock, 18
stulong, 19

transform_dataset, 20

transform_dataset_internal, 20

weather, 21

23