Package ‘Clustering’

April 17, 2020

Version 1.4
Date 2020-04-10
Title Execution of Multiple Clustering Algorithm
Author Luis Alfonso Perez Martos [aut, cre]
Maintainer Luis Alfonso Perez Martos <lapm0001@gmail.com>
Depends R (>= 3.5.0)
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.
License GPL (>= 2)
Type Package
Encoding UTF-8
LazyData true
RoxygenNote 7.1.0
Repository CRAN
Imports apclust, cluster, ClusterR, adclust, pvclust, gama, amap, stats, pracma, tools, gmp, utils, xtable, sqldf, data.table, ggplot2, glue
Suggests knitr, rmarkdown, kableExtra, tidyverse
VignetteBuilder knitr
NeedsCompilation no
Date/Publication 2020-04-17 10:00:02 UTC

R topics documented:

- basketball .................................................. 2
- best_ranked_external_metrics ............................. 3
- best_ranked_internal_metrics ............................ 3
- bolts ................................................................ 4
- clustering .................................................... 5
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:
This data set contains a series of statistics about basketball players:

assists_per_minuteReal average number of assistances per minute
heightInteger height of the player
time_playedReal time played by the player
ageInteger number of years of the player
points_per_minuteReal average number of points per minute

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

*Method that calculates the best rated external metrics*

**Description**
Method that calculates the best rated external metrics

**Usage**

```r
closest_ranked_external_metrics(df)
```

**Arguments**

- `df` : data matrix or data frame

**Value**
returns a table with the external metrics that has the best rating

**Examples**

```r
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
best_ranked_external_metrics(df$result)
```

---

**best_ranked_internal_metrics**

*Method that calculates the best rated internal metrics*

**Description**
Method that calculates the best rated internal metrics

**Usage**

```r
closest_ranked_internal_metrics(df)
```

**Arguments**

- `df` : data matrix or data frame

**Value**
returns a table with the internal metrics that has the best rating
Examples

```r
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)

best_ranked_internal_metrics(df$result)
```

bolts

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

```r
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
**clustering**

**SENSInteger**  the sensitivity of the electronic eye

**TIMEReal**  The measured response is the time, in seconds

**T20BOLTReal**  n 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 (T20BOLTReal). Below are the data for 40 combinations of settings. RUNInteger is the order in which the data were collected.

**Source**

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

---

### clustering

**Execute a list of datasets from a route or a dataframe**

**Description**

Execute a list of datasets from a route or a dataframe

**Usage**

```r
clustering(
    path = CONST_NULL,
    df = CONST_NULL,
    packages = CONST_NULL,
    algorithm = CONST_NULL,
    min = CONST_NULL,
    max = CONST_NULL,
    metrics = CONST_NULL,
    variables = CONST_NULL
)
```
evaluate_best_validation_external_by_metrics

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

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

Usage
evaluate_best_validation_external_by_metrics(df)

Arguments
- df: data matrix or data frame
evaluate_best_validation_internal_by_metrics

Value

returns a table with the algorithm and the best performing metric for the datasets

Examples

df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
evaluate_best_validation_external_by_metrics(df$result)

evaluate_best_validation_internal_by_metrics

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

Description

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

Usage

evaluate_best_validation_internal_by_metrics(df)

Arguments

df data matrix or data frame

Value

returns a table with the algorithm and the best performing metric for the datasets

Examples

df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
evaluate_best_validation_internal_by_metrics(df$result)
evaluate_validation_external_by_metrics

Method that calculates which algorithm behaves best for the datasets provided

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

Value
returns a table with the best performing algorithm for the provided datasets

Examples

```r
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
evaluate_validation_external_by_metrics(df$result)
```

evaluate_validation_internal_by_metrics

Method that calculates which algorithm behaves best for the datasets provided

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
**plot_external_validation**

**Value**

returns a table with the best performing algorithm for the provided datasets

**Examples**

```r
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
evaluate_validation_internal_by_metrics(df$result)
```

---

**plot_external_validation**

*Method that graphically compares external evaluation metrics*

---

**Description**

Method that graphically compares external evaluation metrics

**Usage**

```r
plot_external_validation(df, metric)
```

**Arguments**

- `df`: df data matrix or data frame
- `metric`: string with the name of the metric select to evaluate

**Examples**

```r
df <- clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm')
plot_external_validation(df,"precision")
```
plot_internal_validation

*Method that graphically compares internal evaluation metrics*

**Description**
Method that graphically compares internal evaluation metrics

**Usage**

```r
plot_internal_validation(df, metric)
```

**Arguments**
- `df`: df data matrix or data frame
- `metric`: string with the name of the metric select to evaluate

**Examples**

```r
df <- clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm')
plot_internal_validation(df,"dunn")
```

result_external_algorithm_by_metric

*Method that returns a table with the algorithm and the metric indicated as parameters*

**Description**
Method that returns a table with the algorithm and the metric indicated as parameters

**Usage**

```r
result_external_algorithm_by_metric(df, algorithm)
```

**Arguments**
- `df`: data matrix or data frame
- `algorithm`: on which we will calculate the results

**Value**
returns a table with the algorithm and the metric indicated as parameter
Examples

```
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
result_external_algorithm_by_metric(df$result, 'daisy')
```

---

Description

Method that returns a table with the algorithm and the metric indicated as parameters

Usage

```
result_internal_algorithm_by_metric(df, algorithm)
```

Arguments

- **df**: data matrix or data frame
- **algorithm**: on which we will calculate the results

Value

returns a table with the algorithm and the metric indicated as parameter

Examples

```
df = clustering(df = cluster::agriculture, min = 4, max = 5, algorithm='gmm', variables = TRUE)
result_internal_algorithm_by_metric(df$result, 'gmm')
```
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
Company7 company7 details
Company8 company8 details
Company9 company9 details
Company10 company10 details

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

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/>

weather

One of the most known testing data sets in machine learning. This data set 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 set 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 set 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/>
Index

*Topic datasets
  basketball, 2
  bolts, 4
  stock, 12
  stulong, 13
  weather, 13

basketball, 2
best_ranked_external_metrics, 3
best_ranked_internal_metrics, 3
bolts, 4
clustering, 5
evaluate_best_validation_external_by_metrics, 6
evaluate_best_validation_internal_by_metrics, 7
evaluate_validation_external_by_metrics, 8
evaluate_validation_internal_by_metrics, 8
plot_external_validation, 9
plot_internal_validation, 10
result_external_algorithm_by_metric, 10
result_internal_algorithm_by_metric, 11
stock, 12
stulong, 13
weather, 13