# Package 'FairMclus'

November 19, 2021

**Type** Package

**Title** Clustering for Data with Sensitive Attribute

**Version** 2.2.1

**Author** Carlos Santos-Mangudo [aut, cre]

**Maintainer** Carlos Santos-Mangudo `<carlossantos.csm@gmail.com>`

**Description**
Clustering for categorical and mixed-type of data, to preventing classification biases due to race, gender or others sensitive attributes.
This algorithm is an extension of the methodology proposed by ``Santos & Heras (2020) <doi:10.28945/4643>".

**License** GPL (>= 2)

**Encoding** UTF-8

**Imports** dplyr, irr, rlist, tidyr, parallel, magrittr, cluster, base, data.table, foreach, doParallel

**RoxygenNote** 7.1.2

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2021-11-19 07:30:14 UTC

## R topics documented:

## FairMclus — *FairMclus Clustering for Data with Sensitive Attribute*

### Description

Clustering for categorical and mixed-type of data, to preventing classification biases due to race, gender or others sensitive attributes. This algorithm is an extension of the methodology proposed by "Santos & Heras (2020) <doi:10.28945/4643>".

### Usage

```
FairMclus(f, typedata, protected, ncores, kclus, numpos)
```

### Arguments

`f` A matrix or data frame, categorical or mixed-type of data. Objects have to be in rows, variables in columns.

`typedata` Type of data included in dataset. For categorical data should be "C" and for mixed data should be "M".

`protected` Name of Protected attribute column included in the data (only one).

`ncores` Number of logical cores of computer to execute parallel process (if = 0, then will take it 2 cores by default).

`kclus` Either the number of clusters, say k,

`numpos` A vector specifying the numerical positions included in dataset. If dataset is a categorical dataset then should be c(0).

### Details

The data given by "f" parameter is clustered by the FairMclus method (Santos & Heras, 2020), which aims to partition the objects into k groups such that the distance from objects to the assigned cluster is minimized, maintaining the ratio of the protected attribute from the original data

All executions return the specified values, however and depending on how large the number of rows in the dataset is, it will take a little longer to execute.

If the typedata value is different from "m" or "c", the algorithm will give an error, and if no protected attribute is included, the algorithm will give an error as well.

FairMclus is a clustering algorithm for finding homogeneous and fair clusters in data files, with categorical only or also with mixed-type attributes, with a better grouping effect that preventing classification biases due to race, gender, social status, others.

Stability, classification efficiency and fairness are the major benefits of FairMclus.

## Value

An object of output "FairMclus" contain a list with following components:

$cluster - A vector of integers (from 1 to k) indicating the cluster to which each point is allocated.

$fairdis - Total Fairness distributed in the data

$fairatio - Fairness ratio obtained by FairMclus

$fairclus - A matrix with percentage of objects in each cluster and per each value of the protected attribute.

$clusize - A table with number of objects in each cluster.

$fairsize - A matrix with number of objects in each cluster and per each value of the protected attribute.

## Author(s)

Carlos Santos-Mangudo, carlossantos.csm@gmail.com

## References

Santos M., C. & J. Heras, A. (2020). A Multicluster Approach to Selecting Initial Sets for Clustering of Categorical Data. Interdisciplinary Journal of Information, Knowledge, and Management, 15, 227-246, https://doi.org/10.28945/4643

## Examples

```
### a toy-example
#
# Some required libraries to be used:
library(dplyr); library(utils); library(data.table); library(tidyr)
library(cluster); library(rlist); library(magrittr); library(irr)
library(stats); library(parallel); library(foreach); library(doParallel);
#
## generate data set with 4 columns and 20 rows:
a <- c(1:20)                                  # name of element
b <- c(1:5)                                   # categorical attribute
c <- c(1:2)                                   # protected attribute
d <- rbind(matrix(rnorm(20, mean=10, sd = 1), ncol = 1))   # numerical value
e <- c(1:4)                                                # categorical value
#
dataM <- cbind(a,b,c,d)
dataC <- cbind(a,b,c,e)
colnames(dataM) <- colnames(dataC) <- c("V0", "V1", "V2", "V3")
#
## run algorithm on mixed-type of data: FairMclus(dataM, "m", "V2", 0, 2, c(3))

## run algorithm on categorical data: FairMclus(dataC, "c", "V2", 0, 2, c(0))
```

# Index