Package ‘opdisDownsampling’

August 8, 2021

Type Package
Title Optimal Distribution Preserving Down-Sampling of Bio-Medical Data
Version 0.6.0
Description An optimized method for distribution-preserving class-proportional down-sampling of bio-medical data.
Depends R (>= 3.5.0)
Imports parallel, graphics, methods, stats, caTools, pracma, twosamples, utils, pbmcapply, benchmarkme
LazyData true
Suggests testthat
License GPL-3
URL https://github.com/JornLotsch/opdisDownsampling
Encoding UTF-8
Author Jorn Lotsch [aut,cre] (<https://orcid.org/0000-0002-5818-6958>), Sebastian Malkusch [aut] (<https://orcid.org/0000-0001-6766-140X>), Alfred Ultsch [aut] (<https://orcid.org/0000-0002-7845-3283>)
Maintainer Jorn Lotsch <j.lotsch@em.uni-frankfurt.de>
NeedsCompilation no
Repository CRAN
Date/Publication 2021-08-08 14:10:02 UTC

R topics documented:

FlowcytometricData . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
GMartificialData . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
opdisDownsampling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

Index 4
FlowcytometricData  Example data of hematologic marker expression.

Description
Data set of 6 flow cytometry-based lymphoma makers from 55,843 cells from healthy subjects (class 1) and 55,843 cells from lymphoma patients (class 2).

Usage
data("FlowcytometricData")

Details
Size 111686 x 6, stored in FlowcytometricData$[Var_1,Var_2,Var_3,Var_4,Var_5,Var_6]
Classes 2, stored in FlowcytometricData$Cls

Examples
data(FlowcytometricData)
str(FlowcytometricData)

GMMartificialData  Example data an artificial Gaussian mixture.

Description
Dataset of 30000 instances with 10 variables that are Gaussian mixtures and belong to classes Cls = 1, 2, or 3, with different means and standard deviations and equal weights of 0.5, 0.4, and 0.1, respectively.

Usage
data("GMMartificialData")

Details
Size 30000 x 10, stored in GMMartificialData$[X1,X2,X3,X4,X5,X6,X7,X8,X9,X10]
Classes 3, stored in GMMartificialData$Cls

Examples
data(GMMartificialData)
str(GMMartificialData)
Description

The package provides the necessary functions for optimal distribution-preserving down-sampling of large (bio-medical) data sets.

Usage

```r
opdisDownsampling(Data, Cls, Size, Seed, nTrials = 1000, TestStat = "ad", MaxCores, JobSize = 1000, PCAimportance = FALSE)
```

Arguments

- **Data**: the data as a vector, matrix or data frame.
- **Cls**: the class information, if any, as a vector of similar length as instances in the data.
- **Size**: the percentage of instances per class to be drawn.
- **Seed**: a predefined seed to modify the results.
- **nTrials**: how many samples to choose from should be randomly drawn.
- **TestStat**: statistical criterion for similarity judgment.
- **MaxCores**: maximum number of cpu cores to use for parallel computing.
- **JobSize**: how many samples can be drawn at once.
- **PCAimportance**: PCA based feature selection; only variables important in PCA projection are considered.

Value

Returns a list of data containing the drawn samples and the omitted data.

- **ReducedData**: the selected sample data and class information.
- **RemucedData**: the not-selected sample data and class information.

Author(s)

Jorn Lotsch

References


Examples

```r
## example 1
data(iris)
Iris50percent <- opdisDownsampling(Data = iris[,1:4], Cls = as.integer(iris$Species), Size = 50, MaxCores = 1)
```
Index

* data sampling
  opdisDownsampling, 3
* opdisDownsampling
  opdisDownsampling, 3

FlowcytometricData, 2
GMMartificialData, 2
opdisDownsampling, 3