Package 'sansa'

August 23, 2022

Title Synthetic Data Generation for Imbalanced Learning in 'R'

Version 0.0.1

Description Machine learning is widely used in information-systems design. Yet, training algorithms on imbalanced datasets may severely affect performance on unseen data. For example, in some cases in healthcare, financial, or internet-security contexts, certain subclasses are difficult to learn because they are underrepresented in training data. This 'R' package offers a flexible and efficient solution based on a new synthetic average neighborhood sampling algorithm ('SANSA'), which, in contrast to other solutions, introduces a novel "placement" parameter that can be tuned to adapt to each datasets unique manifestation of the imbalance. More information about the algorithm's parameters can be found at Nasir et al. (2022) <https://murtaza.cc/SANSA/>.

License GPL (>= 3)

Encoding UTF-8

RoxygenNote 7.1.1

Imports data.table, FNN, ggplot2

NeedsCompilation no

Author Murtaza Nasir [aut, cre] (<https://orcid.org/0000-0002-4481-065X>),
Ali Dag [ctb],
Serhat Simsek [ctb],
Anton Ivanov [ctb],
Asil Oztekin [ths]

Maintainer Murtaza Nasir <mail@murtaza.cc>

Repository CRAN

Date/Publication 2022-08-23 08:40:02 UTC

R topics documented:

sansa *Title*

Description

Title

Usage

```
sansa(x, y, lambda = 0, ksel = 3)
```

Arguments

x	Input predictor as a dataframe
y	Target variable as factor
lambda	Lambda parameter to select distribution of synthetic variables
ksel	K parameter to choose how many neighbors are used in calculations

Value

A list with two elements: x contains predictors with synthetic data, y contains target with synthetic data.

Examples

```
library(sansa)
library(ggplot2)
minority = data.frame(x1 = rnorm(10, 10, 3),
                      x2 = rnorm(10, 25, 10),
                      target = "true")
majority = data.frame(x1 = rnorm(100, 4, 2),
                      x2 = rnorm(100, 30, 10),
                      target = "false")

dataset = rbind(minority, majority)

ggplot(dataset) + geom_point(aes(x1, x2, color = target))
sansaobject = sansa(x = dataset[,1:2], y = dataset$target, lambda = 1, ksel = 3)

balanced <- sansaobject$x
balanced$target = sansaobject$y

ggplot(balanced) + geom_point(aes(x1, x2, color = target))
```

Index