Package 'imbalanceDatRel'

April 28, 2023

Type Package

Title Relocated Data Oversampling for Imbalanced Data Classification

Version 0.1.5

Description Relocates oversampled data from a specific oversampling method to cover area determined by pure and proper class cover catch digraphs (PCCCD). It prevents any data to be generated in class overlapping area.

Depends R (>= 4.2), rcccd

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.1

Imports RANN, Rfast, SMOTEWB

NeedsCompilation no

Author Fatih Saglam [aut, cre] (<https://orcid.org/0000-0002-2084-2008>)

Maintainer Fatih Saglam <saglamf89@gmail.com>

Repository CRAN

Date/Publication 2023-04-28 18:10:05 UTC

R topics documented:

DatRel *Data Relocation for Resampled Data using Pure and Proper Class Cover Catch Digraph*

Description

DatRel relocates resampled data using Pure and Proper Class Cover Catch Digraph

Usage

```
DatRel(x, y, x_syn, proportion = 1, p_of = 0, class_pos = NULL)
```

Arguments

x	feature matrix or dataframe.
y	class factor variable.
x_syn	synthetic data generated by an oversampling method.
proportion	proportion of covered samples. A real number between $(0, 1]$. 1 by default. Algorithm stops when desired percent of coverage achieved in each class. Smaller numbers results in less dominant samples.
p_of	proportion to increase cover radius. A real number between $(0, \infty)$. Default is 0. Higher values tolerate other classes more.
class_pos	Class name of synthetic data. Default is NULL. If NULL, positive class is minority class.

Details

Calculates cover areas using pure and proper class cover catch digraphs (PCCCD) for original dataset. Any sample outside of cover area is relocated towards a specific dominant point. Determination of dominant point to move towards is based on distance based on radii of PCCCD balls. p_of is to increase obtained radii to be more tolerant to noise. prooportion argument is cover percentage for PCCCD to stop when desired percentage is covered for each class. PCCCD models are determined using rcccd package. class_pos argument is used to specify oversampled class.

Value

an list object which includes:

x_new	Oversampled and relocated feature matrix
y_new	Oversampled class variable
x_syn	Generated and relocated sample matrix
i_dominant	Indexes of dominant samples
x_pos_dominant	Dominant samples for positive class
radii_pos_dominant	Positive class cover percentage

Author(s)

Fatih Saglam, saglamf89@gmail.com

Examples

```
library(SMOTEWB)
library(rcccd)

set.seed(10)
# adding data
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(60, 6, 1), ncol = 2, nrow = 30))
y <- as.factor(c(rep("negative", 1000), rep("positive", 30)))

# adding noise
x[1001,] <- c(3,3)
x[1002,] <- c(2,2)
x[1003,] <- c(4,4)

# resampling
m_SMOTE <- SMOTE(x = x, y = y, k = 3)

# relocation of resampled data
m_DatRel <- DatRel(x = x, y = y, x_syn = m_SMOTE$x_syn)

# resampled data
plot(x, col = y, main = "SMOTE")
points(m_SMOTE$x_syn, col = "green")

# resampled data after relocation
plot(x, col = y, main = "SMOTE + DatRel")
points(m_DatRel$x_syn, col = "green")
```

`f_dominate` *Determining cover balls*

Description

Determining cover balls

Usage

```
f_dominate(x_main, x_other, proportion = 1, p_of = 0)
```

Arguments

x_main	Target class samples.
x_other	Non-target class samples.
proportion	proportion of covered samples. A real number between $(0, 1]$. 1 by default. Smaller numbers results in less dominant samples.
p_of	roportion to increase cover radius. A real number between $(0, \infty)$. Default is 0. Higher values tolerate other classes more.

Details

To be used in DatRel.

Value

a list object with following:

i_dominant	dominant sample indexes
dist_main2other	distance matrix of target class samples to non-target class samples
dist_main2main	distance matrix of target class samples to target class samples

Author(s)

Fatih Saglam, saglamf89@gmail.com

---

f_relocate *Relocation function*

Description

Relocation function

Usage

```
f_relocate(x_pos_dominant, x_syn, radii_pos_dominant, p_of = 0)
```

Arguments

x_pos_dominant	positive class dominant sample matrix or dataframe
x_syn	synthetically generated positive class sample matrix or dataframe
radii_pos_dominant	positive class dominant sample radii
p_of	proportion to increase cover radius. A real number between $(0, \infty)$. Default is 0. Higher values tolerate other classes more.

Value

relocated data matrix

Author(s)

Fatih Saglam, saglamf89@gmail.com

`oversampleDatRel` *Oversampling and Data Relocation for Resampled Data*

Description

`oversampleDatRel` first oversamples using selected method then relocates resampled data using Pure and Proper Class Cover Catch Digraph.

Usage

```
oversampleDatRel(
  x,
  y,
  method = "SMOTE",
  proportion = 1,
  p_of = 0,
  class_pos = NULL,
  ...
)
```

Arguments

`x`	feature matrix or dataframe.
`y`	class factor variable.
`method`	oversampling method. Default is "SMOTE". Available methods are: "ADASYN": Adaptive Synthetic Sampling "ROS": Random Oversampling "ROSE": Randomly Over Sampling Examples "RSLSMOTE": Relocating safe-level SMOTE with minority outcast handling "RUS": Random Undersampling "SLSSMOTE": Safe-level Synthetic Minority Oversampling Technique "SMOTE": Synthetic Minority Oversampling Technique "SMOTEWB": SMOTE with boosting
`proportion`	proportion of covered samples. A real number between $(0, 1]$. 1 by default. Smaller numbers results in less dominant samples.
`p_of`	proportion to increase cover radius. A real number between $(0, \infty)$. Default is 0. Higher values tolerate other classes more.
`class_pos`	Class name of synthetic data. Default is NULL. If NULL, positive class is minority class.
`...`	arguments to be used in specified method.

Details

Oversampling using DatRel. Available oversampling methods are from SMOTEWB package. "ROSE" generates samples from all classes. DatRel relocates all class samples.

Value

an list which includes:

x_new	dominant sample indexes.
y_new	dominant samples from feature matrix, x
x_syn	Radiuses of the circle for dominant samples
i_dominant	class names
x_pos_dominant	number of classes
radii_pos_dominant	proportions each class covered

Author(s)

Fatih Saglam, saglamf89@gmail.com

Examples

```
library(SMOTEWB)
library(rcccd)

set.seed(10)
# adding data
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(60, 6, 1), ncol = 2, nrow = 30))
y <- as.factor(c(rep("negative", 1000), rep("positive", 30)))

# adding noise
x[1001,] <- c(3,3)
x[1002,] <- c(2,2)
x[1003,] <- c(4,4)

# resampling
m_SMOTE <- SMOTE(x = x, y = y, k = 3)

# resampled data
plot(x, col = y, main = "SMOTE")
points(m_SMOTE$x_syn, col = "green")

m_DatRel <- oversampleDatRel(x = x, y = y, method = "SMOTE")

# resampled data after relocation
plot(x, col = y, main = "SMOTE + DatRel")
points(m_DatRel$x_syn, col = "green")
```

Index