Package 'SMOTEWB'

January 22, 2023

Type Package

Title Imbalanced Resampling using SMOTE with Boosting (SMOTEWB)

Version 0.1.3

Description Provides the SMOTE with Boosting (SMOTEWB) algorithm. See F. Sağlam, M. A. Cengiz (2022) <doi:10.1016/j.eswa.2022.117023>.

Depends R (>= 4.2)

Imports stats, FNN, rpart

Suggests

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.1

NeedsCompilation no

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

Maintainer Fatih Saglam <saglamf89@gmail.com>

Repository CRAN

Date/Publication 2023-01-21 23:20:17 UTC

R topics documented:

ROS *Random Oversampling (ROS)*

Description

Resampling with ROS.

Usage

```
ROS(x, y)
```

Arguments

x	feature matrix.
y	a factor class variable with two classes.

Details

Random Oversampling (ROS) is a method of copying and pasting of positive samples until balance is achieved.

Value

a list with resampled dataset.

x_new	Resampled feature matrix.
y_new	Resampled target variable.

Author(s)

Fatih Saglam, saglamf89@gmail.com

Examples

```
set.seed(1)
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(100, 5, 1), ncol = 2, nrow = 50))
y <- as.factor(c(rep("negative", 1000), rep("positive", 50)))

plot(x, col = y)

# resampling
m <- ROS(x = x, y = y)

plot(m$x_new, col = m$y_new)
```

RUS *Random Undersampling (RUS)*

Description

Resampling with RUS.

Usage

```
RUS(x, y)
```

Arguments

x	feature matrix.
y	a factor class variable with two classes.

Details

Random Undersampling (RUS) is a method of removing negative samples until balance is achieved.

Value

a list with resampled dataset.

x_new	Resampled feature matrix.
y_new	Resampled target variable.

Author(s)

Fatih Saglam, saglamf89@gmail.com

Examples

```
set.seed(1)
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(100, 5, 1), ncol = 2, nrow = 50))
y <- as.factor(c(rep("negative", 1000), rep("positive", 50)))

plot(x, col = y)

# resampling
m <- RUS(x = x, y = y)

plot(m$x_new, col = m$y_new)
```

SMOTE *Synthetic Minority Oversampling Technique (SMOTE)*

Description

Resampling with SMOTE.

Usage

```
SMOTE(x, y, k = 5)
```

Arguments

x	feature matrix.
y	a factor class variable with two classes.
k	number of neighbors.

Details

SMOTE (Chawla et al., 2002) is an oversampling method which creates links between positive samples and nearest neighbors and generates synthetic samples along that link.

It is well known that SMOTE is sensitive to noisy data. It may create more noise.

Value

a list with resampled dataset.

x_new	Resampled feature matrix.
y_new	Resampled target variable.
C	Number of synthetic samples for each positive class samples.

Author(s)

Fatih Saglam, saglamf89@gmail.com

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Examples

```
set.seed(1)
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(100, 5, 1), ncol = 2, nrow = 50))
y <- as.factor(c(rep("negative", 1000), rep("positive", 50)))

plot(x, col = y)

# resampling
m <- SMOTE(x = x, y = y, k = 7)

plot(m$x_new, col = m$y_new)
```

SMOTEWB *SMOTE with boosting (SMOTEWB)*

Description

Resampling with SMOTE with boosting.

Usage

```
SMOTEWB(x, y, n_weak_classifier = 100, class_weights = NULL, k_max = NULL, ...)
```

Arguments

x	feature matrix.
y	a factor class variable with two classes.
n_weak_classifier	number of weak classifiers for boosting.
class_weights	numeric vector of length two. First number is for positive class, and second is for negative. Higher the relative weight, lesser noises for that class. By default, $2 \times n_{neg}/n$ for positive and $2 \times n_{pos}/n$ for negative class.
k_max	to increase maximum number of neighbors. It is `ceiling(n_neg/n_pos)` by default.
...	additional inputs for ada::ada().

Details

SMOTEWB (Saglam & Cengiz, 2022) is a SMOTE-based oversampling method which can handle noisy data and adaptively decides the appropriate number of neighbors to link during resampling with SMOTE.

This function first scales features into [0-1] range then applies resampling. Descaling is based on minimum and maximum values of original dataset features. This is to detect better nearest neighbours.

Value

a list with resampled dataset.

x_new	Resampled feature matrix.
y_new	Resampled target variable.
w	Boosting weights for original dataset.
k	Number of nearest neighbors for positive class samples.
C	Number of synthetic samples for each positive class samples.

Author(s)

Fatih Saglam, saglamf89@gmail.com

References

Sağlam, F., & Cengiz, M. A. (2022). A novel SMOTE-based resampling technique trough noise detection and the boosting procedure. Expert Systems with Applications, 200, 117023.

Examples

```
set.seed(1)
x <- rbind(matrix(rnorm(2000, 3, 1), ncol = 2, nrow = 1000),
           matrix(rnorm(100, 5, 1), ncol = 2, nrow = 50))
y <- as.factor(c(rep("negative", 1000), rep("positive", 50)))

plot(x, col = y)

# resampling
m <- SMOTEWB(x = x, y = y, n_weak_classifier = 150)

plot(m$x_new, col = m$y_new)
```

Index