Package 'balancedrandomforest'

December 11, 2023

Type Package

Title Refined random forest algorithms for imblanced binary classification

Version 0.1.0

Author Lin Gui and Siyu Kong

Maintainer Lin Gui <lg625@cornell.edu> Siyu Kong <sk6666@cornell.edu>

Description Conduct classfication using different variants of random forest and make comparison. Run the `demo.R` and `demo_clustered.R` to see the examples.

Encoding UTF-8

License TBA

LazyData true

Imports MASS,
randomForest,
ggplot2,
gridExtra

RoxygenNote 7.2.3

R topics documented:

`fit_random_forest` *Train the data with random forest-based algorithms*

Description

Train the data with random forest-based algorithms

Usage

```
fit_random_forest(train_data, test_data, method)
```

Arguments

`train_data`	Training data in the 2d plane.
`test_data`	Testing data.
`method`	Five algorithms are included. "rf": random forest, "wrf": weighted random forest, "brf": balanced random forest, "brf1": balanced random forest with refined weights using method 1, "brf2": balanced random forest with refined weights using method 2.

Value

Return a list including ‘model‘, the model after trianing, and ‘measure‘, the six measures on the testing data.

`fit_random_forest_clustered`
Train the clustered data with random forest-based algorithms

Description

Train the clustered data with random forest-based algorithms

Usage

```
fit_random_forest_clustered(train_data, test_data, method)
```

Arguments

`train_data`	Training data in the 2d plane.
`test_data`	Testing data.
`method`	Five algorithms are included. "rf": random forest, "wrf": weighted random forest, "brf": balanced random forest, "brf1": balanced random forest with refined weights using method 1, "brf2": balanced random forest with refined weights using method 2.

Value

Return a list including ‘model‘, the model after trianing, and ‘measure‘, the six measures on the testing data.

generate_clustered_data

Generate data with two groups, for which the major group comes from a Gaussian distribution and the minor group comes from a mixture of two Gaussian distributions.

Description

Generate data with two groups, for which the major group comes from a Gaussian distribution and the minor group comes from a mixture of two Gaussian distributions.

Usage

```
generate_clustered_data(parameters, n_train_zero, n_train_one, n_train_two)
```

Arguments

parameters	A list containing the mean of the major distribution, the covariance of the major distribution, the mean of the minor distribution and the covariance of the minor distribution
n_train_zero	The sample size of the major group.
n_train_one	The sample size of the minor group from the first distribution of the mixture of Gaussian.
n_train_two	The sample size of the minor group from the second distribution of the mixture of Gaussian.

Value

This function returns a list including ‘train_data‘ the training data and ‘test_data‘ the testing data

generate_data

Generate data with two groups, each of which comes from a Gaussian distribution respectively.

Description

Generate data with two groups, each of which comes from a Gaussian distribution respectively.

Usage

```
generate_data(parameters, n_train_zero, n_train_one)
```

Arguments

parameters	A list containing the mean of the major distribution, the covariance of the major distribution, the mean of the minor distribution and the covariance of the minor distribution
n_train_zero	The sample size of the major group.
n_train_one	The sample size of the minor group.

Value

This function returns a list including ‘train_data‘ the training data and ‘test_data‘ the testing data

---

`get_mean_covariance` *Get prespecified mean and covariances for the Gaussian distribution from which the data come.*

Description

Get prespecified mean and covariances for the Gaussian distribution from which the data come.

Usage

```
get_mean_covariance(choice, mean_dist)
```

Arguments

`choice`	Taking 1, 2, 3, or 4, this function offers four choices to allow different settings of data distribution.
`mean_dist`	This argument specifies the distance between the mean of two Gaussian distribution.

Value

This function returns a list including ‘mean_zero‘, ‘mean_one‘, ‘cov_zero‘, ‘cov_one‘

`get_mean_covariance_clustered` *Get prespecified mean and covariances for the Gaussian distribution from which the data come.*

Description

Get prespecified mean and covariances for the Gaussian distribution from which the data come.

Usage

```
get_mean_covariance_clustered(choice, mean_dist)
```

Arguments

`choice`	Taking 1, 2, 3, or 4, this function offers four choices to allow different settings of data distribution.
`mean_dist`	This argument specifies the distance between the mean of two Gaussian distribution.

Value

This function returns a list including ‘mean_zero‘, ‘mean_one‘, ‘mean_two‘, ‘cov_zero‘, ‘cov_one‘, ‘cov_two‘. Zero corresponds to the major distribution, one and two correspond to two distributions in the mixture of Gaussian.

get_print_results *Compute six measures after training*

Description

Compute six measures after training

Usage

```
get_print_results(rf_model)
```

Arguments

rf_model	The trained model.

Value

Return six measures: the classiciation accuracy on the major group, the classiciation accuracy on the minor group, precision, F-measure, G-mean, and weighted accuracy.

modify_rf_with_prior *Calibrate the random forest model with prior information and test it on the testing data.*

Description

Calibrate the random forest model with prior information and test it on the testing data.

Usage

```
modify_rf_with_prior(rf_model, class_zero_ratio, test_data)
```

Arguments

rf_model	The model after training with the ordinary random forest.
class_zero_ratio	The value of $P(Y=0)$, which represents the ratio of data with label 0 among all the data.
test_data	Testing data.

Value

This function returns six measures of the calibrated model with prior information on the testing data.

Index