Package 'EFS'

August 2, 2016

Title Tool for Ensemble Feature Selection

Description Provides a function to check the importance of a feature based on a dependent classification variable. An ensemble of correlation and importance measure tests are used to determine the normed importance value of all features. Combining these methods in one function (building the sum of the importance values) leads to a better tool for selecting most important features. This selection can also be viewed in a barplot using the barplot_fs() function and proved using an also provided function for a logistic regression model, namely logreg_test().

Type Package

Version 1.0.0

Date 2016-04-13

License GPL (>= 2)

Encoding UTF-8

Author Nikita Genze, Ursula Neumann

Maintainer Ursula Neumann <u.neumann@wz-straubing.de>

LazyLoad yes

Imports party, pROC, randomForest, ROCR, grDevices, graphics, stats

Repository CRAN

RoxygenNote 5.0.1

NeedsCompilation no

Date/Publication 2016-08-02 14:45:33

R topics documented:

barplot_fs *Visualization of* ensemble_fs *in barplot*

Description

Generates a barplot from the output of ensemble_fs and produces a pdf-file. This file will be located in the working directory. A barplot will only be provided, when the number of features does not exceed 100.
x-axis: sum of all normed importance values of each feature ranging from 0 to 1
y-axis: names of features

Usage

```
barplot_fs(name, efs_table)
```

Arguments

name	a character string giving the name of the file. If it is NULL, then no external file is created (effectively, no drawing occurs), but the device may still be queried.
efs_table	table object of class matrix (retrieved from ensemble_fs)

Author(s)

Ursula Neumann

See Also

barplot, pdf

Examples

```
##loading dataset in Environment
data(efsdata)
##Generate a ranking based on inportance (with default
##NA_threshold = 0.7,cor_threshold = 0.2)
efs<-ensemble_fs(efsdata,5,runs=2)
##Create a ROC Curve based on the output from
##efs <- ensemble_fs
barplot_fs("test",efs)
```

efsdata *Meteorological data for feature selection analysis*

Description

A dataset with meteorological data from a weather station in Frankfurt (Oder), Germany from february 2016

Usage

```
data(efsdata)
```

Format

a data frame with 29 entries and following 7 variables

`date` index variable from 1 to 29

`Tmin` temperature minimum of the day

`Tmax` temperature maximum of the day

`SunAvg` sunshine duration of the day

`RainBool` classification variable: if it has not rained: 0, if it has rained: 1

`RelHumAvg` average relative humidity of the day

`WindForceAvg` average wind force of the day

References

modified data from http://wetterstationen.meteomedia.de/

ensemble_fs *Ensemble Feature Selection*

Description

Uses an ensemble of feature selection methods to create a normalized quantitative ranking of all relevant features. Irrelevant features (e.g. too many NA or variance = 1) will be deleted. See Details for a list of tests used in this function.

Usage

```
ensemble_fs(data, classnumber, NA_threshold = 0.2, cor_threshold = 0.7,
  runs = 100, selection = c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE,
  FALSE))
```

Arguments

`data`	the name of the dataset, which should already be loaded in the environment.
`classnumber`	nominal, dichotomous classification. variable, number of column in dataset, which should be the dependent variable for classification.
`NA_threshold`	(optional) decimal number in range of [0,1]. Threshold for deletion of features with a greater proportion of NAs than NA_threshold. than `NA_threshold`
`cor_threshold`	(optional) used only for Spearman and Pearson correlation. The correlation within features is tested. If the correlation of 2 features is greater than. `cor_threshold` the dependent feature is deleted
`runs`	(optional) amount of runs for randomForest and cforest to gain higher robustness.
`selection`	(optional) vector of length eight with TRUE or FALSE values. Selection of feature selection methods to be conducted.

Details

Following methods are provided in the `ensemble_fs`:

- Median: p-values from Wilcoxon signed-rank test (wilcox.test)
- Spearman: Spearman's rank correlation test arccording to Yu et al. (2004) (cor)
- Pearson: Pearson's product moment correlation test arccording to Yu et al. (2004) (cor)
- LogReg: beta-Values of logistic regression (glm)
- Accuracy//Error-rate randomForest: Error-rate-based variable importance measure embedded in randomForest according to Breiman (2001) (randomForest)
- Gini randomForest: Gini-index-based variable importance measure embedded in randomForest according to Breiman (2001) (randomForest)
- Error-rate cforest: Error-rate-based variable importance measure embedded in cforest according Strobl et al. (2009) (cforest)
- AUC cforest: AUC-based variable importance measure embedded in cforest according to Janitza et al. (2013) (cforest)

By the argument `selection` the user decides which feature selection methods are used in `ensemble_fs`. Default value is `selection = c(TRUE, TRUE, TRUE,TRUE, TRUE, TRUE, FALSE, FALSE)`, i.e., the function does not use either of the cforest variable importance measures. The maximum score for features depends on the input of `selection`. The scores are always divided through the amount of selected feature selection, respectively the amount of TRUEs.

Value

table of normalized importance values of class matrix (used methods as rows and features of the imported file as columns).

Author(s)

Ursula Neumann

References

- Yu, L. and Liu H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 2004, 5:1205-1224.

- Breiman, L.: Random Forests, Machine Learning. 2001, 45(1): 5-32.

- Strobl, C., Malley, J. anpercentaged Tutz, G.: An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random forests. Psychological Methods. 2009, 14(4), 323–348.

- Janitza, S., Strobl, C. and Boulesteix AL.: An AUC-based Permutation Variable Importance Measure for Random Forests. BMC Bioinformatics.2013, 14, 119.

See Also

wilcox.test, randomForest, cforest, cor, glm

Examples

```
##loading dataset in Environment
data(efsdata)
##Generate a ranking based on importance (with default NA_threshold = 0.2,
##cor_threshold = 0.7, selection = c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE))
efs<-ensemble_fs(efsdata,5,runs=2)
```

---

logreg_test *Evaluation of* ensemble_fs *via logistic regression*

Description

Evaluates the accuracy of the ouput of ensemble_fs using logistic regression analysis. It selects features which have an importance value above average and generates a logistic regression model. The function compares that model with a logistic regression model generated from all features (without ensemble_fs) and shows the results in a ROC-curve. For further evaluation the function also shows the p-value as result from roc.test in the pdf-file.

Usage

```
logreg_test(data, efs_table, file_name, classnumber, NA_threshold)
```

Arguments

`data`	the name of the dataset, which should already be loaded in the environment.
`efs_table`	table object of class matrix (retrieved from `ensemble_fs`)
`file_name`	name of the file, wil be saved as `file_name` + "-ROC.pdf"
`classnumber`	nominal, dichotomous classification. variable, number of column in dataset, which should be the dependent variable for classification.
`NA_threshold`	(optional) decimal number in range of [0,1]. Threshold for deletion of features with a greater proportion of NAs than NA_threshold. than `NA_threshold`

Value

ROC-curve and p-value of `roc.test` in a pdf-File

Author(s)

Ursula Neumann

See Also

glm, roc

Examples

```
##loading dataset in Environment
data(efsdata)
##Generate a ranking based on importance (with default
##NA_threshold = 0.7,cor_threshold = 0.2)
efs<-ensemble_fs(efsdata,5,runs=2)
##Create a ROC Curve based on the output from ensemble_fs
logreg_test(efsdata,efs,"test",5)
```

Index