Package 'pRF'

January 11, 2016

Title Permutation Significance for Random Forests

Version 1.2

Type Package

Date 2015-10-02

Author Ankur Chakravarthy

Maintainer Ankur Chakravarthy <ankur.chakravarthy.10@ucl.ac.uk>

Description Estimate False Discovery Rates (FDRs) for importance metrics from random forest runs.

License GPL-3

Depends R(>= 3.1.0)

Imports ggplot2, permute, randomForest, reshape2, dplyr(>= 0.4.1), multtest(>= 2.25.0)

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-01-11 17:12:20

R topics documented:

pRF-package *Permutation based approaches to Random Forest feature selection*

Description

Functions to estimate the statistical significance of the Increase in MSE and Decrease in Gini Coefficient metrics of random forest feature importance. The main functions are pRF and sigplot. See documentation on these functions for information on how to use them.

Details

Package:	pRF
Type:	Package
Version:	1.0
Date:	2015-02-23
License:	GPL-3

Author(s)

Ankur Chakravarthy

Maintainer: ankur.chakravarthy.10@ucl.ac.uk

References

The concept of permuting response variables is loosely based off

Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010 May 15;26(10):1340-7. doi: 10.1093/bioinformatics/btq134. Epub 2010 Apr 12. PubMed PMID: 20385727.

pRF *pRF*

Description

The workhorse function - estimates statistical significance of feature importance by permuting the response variable

Usage

```
pRF(response, predictors, n.perms, alpha = 0.05,
mtry = NULL, type = c("classification", "regression"),
ntree = 500,seed=12345, ...)
```

Arguments

response	a character vector or a factor for classification containing the group memberships for classification, a numeric vector for regression
predictors	A matrix consisting of features (measurements) corresponding to samples. The orientation per se does not matter - the function orients them correctly for Random Forest learning.
n.perms	Number of permutations to estimate significance. If the number of all possible permutations is less than this the latter will be used for estimation.
alpha	The significance level threshold of p.values for estimating false discovery rate using the two-step BH method for correlated test statistics, as implemented in the multtest package's mt.rawp2adjp function.
mtry	see ?randomForest for details - defines how many features are randomly sampled for building trees
type	string, set to "classification" or "regression"
ntree	number of trees in the random forest, see documentation from the randomForest package for details.
seed	set seed to ensure reproducibility from run to run and to standardise runs on actual and permuted data
...	Arguments to pass on to the randomForest function

Value

A standardised list containing

Res.table	A data.frame containing significance, FDR, and the feature name. b= number of permutations yielding a higher importance than observed + 1, m= number of permutations + 1
obs	named numeric vector, contains observed importances
perms	data.frame, contains importance values from permutations
Model	the randomForest model that was fit to the original data

Author(s)

Ankur Chakravarthy

References

The main function is based on the idea presented in

Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010 May 15;26(10):1340-7. doi: 10.1093/bioinformatics/btq134. Epub 2010 Apr 12. PubMed PMID: 20385727.

The permutation p.values in the package are exact, calculated according to

Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010;9:Article39. doi: 10.2202/1544-6115.1585. Epub 2010 Oct 31. PubMed PMID: 21044043.

False discovery rates account for correlations using the Two-Step BH procedure, initially reported in

Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli, 'Adaptive Linear Step-up Procedures That Control the False Discovery Rate', Biometrika, 93 (2006), 491-507.

Examples

```
#Load the iris dataset
data(iris)

#Set up the predictors object

predictors=iris[,c(1:4)]
colnames(predictors)<-colnames(iris[1:4])

#Execute the main pRF function
p.test<-pRF(response=factor(iris$Species),
predictors=iris[,c(1:4)],n.perms=20,mtry=3,
type="classification",alpha=0.05)

#Put together a dataframe that consists of the
#significance stats and observed importance metrics

df<-cbind(p.test$Res.table,p.test$obs)
```

`sigplot` *sigplot*

Description

Plots observed importance and null distribution for a pRF output list.

Usage

```
sigplot(pRF.list,threshold)
```

Arguments

`pRF.list`	output from a pRF run.
`threshold`	False Discovery Rate cutoff for plotting. Default is 0.05

Value

a ggplot2 plot

Author(s)

Ankur Chakravarthy

See Also

pRF

Examples

```
#Load the iris dataset
data(iris)

#Set up the predictors object

predictors=iris[,c(1:4)]
colnames(predictors)<-colnames(iris[1:4])

#Execute the main pRF function
p.test<-pRF(response=factor(iris$Species),
predictors=iris[,c(1:4)],n.perms=20,mtry=3,
type="classification",alpha=0.05)

#Plot

sigplot(pRF.list=p.test,threshold=0.1)
```

Index