Package 'rfPermute'

August 23, 2023

Type Package

Title Estimate Permutation p-Values for Random Forest Importance Metrics

Description Estimate significance of importance metrics for a Random Forest model by permuting the response variable. Produces null distribution of importance metrics for each predictor variable and p-value of observed. Provides summary and visualization functions for 'randomForest' results.

Version 2.5.2

URL https://github.com/EricArcher/rfPermute

BugReports https://github.com/EricArcher/rfPermute/issues

Depends R (>= 4.0.0)

Imports abind (>= 1.4), dplyr (>= 1.0), ggplot2 (>= 3.3), grDevices, gridExtra, magrittr (>= 2.0), methods, parallel, randomForest (>= 4.6), rlang, scales, stats, swfscMisc (>= 1.5), tibble (>= 3.1), tidyr (>= 1.1), utils

Collate rfPermute_package.R dataSets.R internals.R rfPermute.R summary.R importance.R combineRP.R classPriors.R confusionMatrix.R casePredictions.R pctCorrect.R plotInbag.R plotImpPreds.R plotNull.R plotPredictedProbs.R plotProximity.R plotTrace.R plotVotes.R balancedSampsize.R cleanRFdata.R

LazyData TRUE

License GPL (>= 2)

Encoding UTF-8

RoxygenNote 7.2.3

NeedsCompilation no

Author Eric Archer [aut, cre]

Maintainer Eric Archer <eric.archer@noaa.gov>

Repository CRAN

Date/Publication 2023-08-23 17:40:02 UTC

R topics documented:

`balancedSampsize` *Balanced Sample Size*

Description

Create a vector of balanced (equal) sample sizes for use in the `sampsize` argument of `rfPermute` or `randomForest` for a classification model. The values are derived from a percentage of the smallest class sample size.

Usage

```
balancedSampsize(y, pct = 0.5)
```

Arguments

`y`	character, numeric, or factor vector containing classes of response variable. Values will be treated as unique for computing class frequencies.
`pct`	percent of smallest class frequency for `sampsize` vector.

Value

a named vector of sample sizes as long as the number of classes.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
data(mtcars)

# A balanced model with default half of smallest class size
sampsize_0.5 <- balancedSampsize(mtcars$am)
sampsize_0.5

rfPermute(factor(am) ~ ., mtcars, replace = FALSE, sampsize = sampsize_0.5)

# A balanced model with one quarter of smallest class size
sampsize_0.25 <- balancedSampsize(mtcars$am, pct = 0.25)
sampsize_0.25

rfPermute(factor(am) ~ ., mtcars, replace = FALSE, sampsize = sampsize_0.25)
```

---

casePredictions *Case Predictions*

Description

Construct a data frame of case predictions for training data along with vote distributions.

Usage

```
casePredictions(x)
```

Arguments

x	a rfPermte or randomForest model object.

Value

A data frame containing columns of original and predicted cases, whether they were correctly classified, and vote distributions among cases.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars)

cp <- casePredictions(rf)
cp
```

`classPriors` *Class Priors*

Description

Compute the class classification priors and class-specific model binomial p-values using these priors as null hypotheses.

Usage

```
classPriors(x, sampsize)
```

Arguments

`x`	a `rfPermute` or `randomForest` model object.
`sampsize`	the vector of sample sizes used to construct the model. If provided, must have length equal to number of classes. If set to `NULL`, priors will be computed assuming empirical sample sizes.

Author(s)

Eric Archer `<eric.archer@noaa.gov>`

See Also

`balancedSampsize`, `confusionMatrix`

Examples

```
library(randomForest)
data(mtcars)

# random sampling with replacement
rf <- randomForest(factor(am) ~ ., mtcars)
confusionMatrix(rf)
classPriors(rf, NULL)

# balanced design
sampsize <- balancedSampsize(mtcars$am)
```

```
rf <- randomForest(factor(am) ~ ., mtcars, replace = FALSE, sampsize = sampsize)
confusionMatrix(rf)
classPriors(rf, sampsize)
```

---

cleanRFdata *Clean Random Forest Input Data*

Description

Removes cases for a Random Forest classification model with missing data and predictors that are constant.

Usage

```
cleanRFdata(x, y, data, max.levels = 30)
```

Arguments

x	columns used as predictor variables as character or numeric vector.
y	column used as response variable as character or numeric.
data	data.frame containing x and y columns.
max.levels	maximum number of levels in response variable y.

Value

a data.frame containing cleaned data.

Author(s)

Eric Archer <eric.archer@noaa.gov>

---

combineRP *Combine rfPermute objects*

Description

Combines two or more ensembles of rfPermute objects into one, combining randomForest results, null distributions, and re-calculating p-values.

Usage

```
combineRP(...)
```

Arguments

...	two or more objects of class `rfPermute`, to be combined into one.

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

combine

Examples

```
data(iris)
rp1 <- rfPermute(
  Species ~ ., iris, ntree = 50, norm.votes = FALSE, nrep = 100, num.cores = 1
)
rp2 <- rfPermute(
  Species ~ ., iris, ntree = 50, norm.votes = FALSE, nrep = 100, num.cores = 1
)
rp3 <- rfPermute(
  Species ~ ., iris, ntree = 50, norm.votes = FALSE, nrep = 100, num.cores = 1
)
rp.all <- combineRP(rp1, rp2, rp3)
rp.all

plotNull(rp.all)
```

---

confusionMatrix *Confusion Matrix*

Description

Generate a confusion matrix for Random Forest classification models with error rates translated into percent correctly classified, and columns for confidence intervals added.

Usage

```
confusionMatrix(x, conf.level = 0.95, threshold = NULL)

plotConfMat(x, title = NULL, plot = TRUE)
```

Arguments

x	a rfPermute or randomForest model object.
conf.level	confidence level for the binom.test confidence interval
threshold	threshold to test observed classification probability against. Should be a number between 0 and 1. If not NULL, the output matrix will have extra columns giving the one-tailed probability that the true correct classification is >= threshold.
title	a title for the plot.
plot	display the plot?

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

classPriors

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars)
confusionMatrix(rf)

confusionMatrix(rf, conf.level = 0.75)

confusionMatrix(rf, threshold = 0.7)
confusionMatrix(rf, threshold = 0.8)
confusionMatrix(rf, threshold = 0.95)
```

---

importance *Extract rfPermute Importance Scores and p-values.*

Description

The importance function extracts a matrix of the observed importance scores and p-values from the object produced by a call to rfPermute. plotImportance produces a visualization of importance scores as either a barchart or heatmap.

Usage

```
## S3 method for class 'rfPermute'
importance(x, scale = TRUE, sort.by = NULL, decreasing = TRUE, ...)

plotImportance(
  x,
  plot.type = c("bar", "heatmap"),
  imp.type = NULL,
  scale = TRUE,
  sig.only = FALSE,
  alpha = 0.05,
  n = NULL,
  ranks = TRUE,
  xlab = NULL,
  ylab = NULL,
  main = NULL,
  size = 3,
  plot = TRUE
)
```

Arguments

x	for importance, an object produced by a call to rfPermute. For plotImportance, either a rfPermute or randomForest model object. If the latter, it must have been run with importance = TRUE.
scale	for permutation based measures, should the measures be divided their "standard errors"?
sort.by	character vector giving the importance metric(s) or p-values to sort by. If NULL, defaults to "MeanDecreaseAccuracy" for classification models and "%IncMSE" for regression models.
decreasing	logical. Should the sort order be increasing or decreasing?
...	arguments to be passed to and from other methods.
plot.type	plot importances as a bar chart or heatmap?
imp.type	character vector listing which importance measures to plot. Can be class names (for classification models) or names of overall importance measures (e.g., "MeanDecreaseAccuracy").
sig.only	Plot only the significant (<= alpha) predictors?
alpha	a number specifying the critical alpha for identifying predictors with importance scores significantly different from random. This parameter is only relevant if rf is a rfPermute object with p-values. Importance measures with p-values less than or equal to alpha will be denoted in barcharts in red and in the heatmap by a white diamond. If set to NULL, significance is not denoted.
n	plot n most important predictors.
ranks	plot ranks instead of actual importance scores?
xlab, ylab	labels for the x and y axes.

main	main title for plot.
size	a value specifying the size of the significance diamond in the heatmap if the p-value <= alpha.
plot	display the plot?

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
data(mtcars)

# A classification model classifying cars to manual or automatic transmission
am.rp <- rfPermute(factor(am) ~ ., mtcars, ntree = 100, nrep = 50)

imp.scaled <- importance(am.rp, scale = TRUE)
imp.scaled

# plot scaled importance scores
plotImportance(am.rp, scale = TRUE)

# plot unscaled and only significant scores
plotImportance(am.rp, scale = FALSE, sig.only = TRUE)
```

---

pctCorrect *Percent Correctly Classified*

Description

For classification models, calculate the percent of individuals correctly classified in a specified percent of trees in the forest.

Usage

```
pctCorrect(x, pct = c(seq(0.8, 0.95, 0.05), 0.99))
```

Arguments

x	a rfPermte or randomForest model object.
pct	vector of minimum percent of trees voting for each class. Can be 0:1 or 0:100.

Value

a matrix giving the percent of individuals correctly classified in each class and overall for each threshold value specified in pct.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars, importance = TRUE)
pctCorrect(rf)
```

`plotImpPreds` *Plot Important Predictor Distribution*

Description

For classification models, plot distribution of predictor variables on classes sorted by order of importance in model.

Usage

```
plotImpPreds(
  x,
  df,
  class.col,
  imp.type = NULL,
  max.vars = 16,
  scale = TRUE,
  size = 1,
  point.alpha = 0.2,
  violin.alpha = 0.5,
  plot = TRUE
)
```

Arguments

`x`	a `rfPermute` or `randomForest` model object.
`df`	data.frame with predictors in `rf` model.
`class.col`	response column name in `df`.
`imp.type`	character string representing importance type to use for sorting predictors.
`max.vars`	number of variables to plot (from most important to least).
`scale`	For permutation based importance measures, should they be divided their "standard errors"?
`size, point.alpha, violin.alpha`	controls size of points and alpha values (transparency) for points and violin plots.
`plot`	display the plot?

Value

the ggplot2 object is invisibly returned.

Note

If the model in x is from randomForest and was run with importance = TRUE, then 'MeanDecreaseAccuracy' is used as the default importance measure for sorting. Otherwise, 'MeanDecreaseGini' is used.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

df <- mtcars
df$am <- factor(df$am)

rf <- randomForest(am ~ ., df, importance = TRUE)
plotImpPreds(rf, df, "am")
```

---

plotInbag *Plot Inbag distribution*

Description

Plot distribution of the fraction of trees that samples were inbag in the Random Forest model.

Usage

```
plotInbag(x, bins = 10, replace = TRUE, sampsize = NULL, plot = TRUE)
```

Arguments

x	a rfPermute or randomForest model object..
bins	number of bins in histogram.
replace	was sampling done with or without replacement?
sampsize	sizes of samples drawn. Either a single value or vector of sample sizes as long as the number of classes.
plot	display the plot?

Value

the ggplot2 object is invisibly returned.

Note

Red vertical lines on the plot denote the expected inbag rate(s). These rates are based on the values of `replace` and `sampsize` supplied. If not specified, they are set to the `randomForest` defaults. If this is not the same as the arguments used to run the model, there will be a mismatch in the location of these indicator lines and the inbag frequency distribution.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

sampsize = c(5, 5)

rf <- randomForest(factor(am) ~ ., data = mtcars, ntree = 10)
plotInbag(rf)

rf <- randomForest(factor(am) ~ ., data = mtcars, ntree = 1000)
plotInbag(rf)

rf <- randomForest(factor(am) ~ ., data = mtcars, ntree = 10000)
plotInbag(rf)
```

---

plotNull *Plot Random Forest Importance Null Distributions*

Description

Plot the Random Forest null distributions importance metrics, observed values, and p-values for each predictor variable from the object produced by a call to `rfPermute`.

Usage

```
plotNull(
  x,
  preds = NULL,
  imp.type = NULL,
  scale = TRUE,
  plot.type = c("density", "hist"),
  plot = TRUE
)
```

Arguments

x	An object produced by a call to rfPermute.
preds	a character vector of predictors to plot. If NULL, then all predictors are plotted.
imp.type	A character vector giving the importance metric(s) to plot.
scale	Plot importance measures scaled (divided by) standard errors?
plot.type	type of plot to produce: "density" for smoothed density plot, or "hist" for histogram.
plot	display the plot?

Details

The function will generate an plot for each predictor, with facetted importance metrics. The vertical red line shows the observed importance score and the _p_-value is given in the facet label.

Value

A named list of the ggplot figures produced is invisibly returned.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
# A regression model using the ozone example
data(airquality)
ozone.rp <- rfPermute(
  Ozone ~ ., data = airquality, ntree = 100,
  na.action = na.omit, nrep = 50, num.cores = 1
)

# Plot the null distributions and observed values.
plotNull(ozone.rp)
```

---

plotPredictedProbs *Plot Predicted Probabilities*

Description

Plot histogram of assignment probabilities to predicted class. This is used for determining if the model differentiates between correctly and incorrectly classified samples in terms of how strongly they are classified.

Usage

```
plotPredictedProbs(x, bins = 30, plot = TRUE)
```

Arguments

x	a rfPermute or randomForest model object.
bins	number of bins in histogram. Defaults to number of samples / 5.
plot	display the plot?

Value

the ggplot2 object is invisibly returned.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars)
plotPredictedProbs(rf, bins = 20)
```

---

plotProximity *Plot Random Forest Proximity Scores*

Description

Create a plot of Random Forest proximity scores using multi-dimensional scaling.

Usage

```
plotProximity(
  x,
  dim.x = 1,
  dim.y = 2,
  class.cols = NULL,
  legend.type = c("legend", "label", "none"),
  legend.loc = c("top", "bottom", "left", "right"),
  point.size = 2,
  circle.size = 8,
  circle.border = 1,
  group.type = c("ellipse", "hull", "contour", "none"),
  group.alpha = 0.3,
  ellipse.level = 0.95,
  n.contour.grid = 100,
  label.size = 4,
  label.alpha = 0.7,
  plot = TRUE
)
```

Arguments

x	a rfPermute or randomForest model object.
dim.x, dim.y	numeric values giving x and y dimensions to plot from multidimensional scaling of proximity scores.
class.cols	vector of colors to use for each class.
legend.type	type of legend to use to label classes.
legend.loc	character keyword specifying location of legend. Can be "bottom", "top", "left", "right".
point.size	size of central points. Set to NULL for no points.
circle.size	size of circles around points indicating classification. Set to NULL for no circles.
circle.border	width of circle border.
group.type	type of grouping to display. Ignored for regression models.
group.alpha	value giving alpha transparency level for group shading. Setting to 0 produces no shading.
ellipse.level	the confidence level at which to draw the ellipse.
n.contour.grid	number of grid points for contour lines.
label.size	size of label if legend.type = `label`.
label.alpha	transparency of label background.
plot	logical determining whether or not to show plot.

Details

Produces a scatter plot of proximity scores for dim.x and dim.y dimensions from a multidimensional scale (MDS) conversion of proximity scores from a randomForest object. For classification models, points are colored according to original (inner) and predicted (outer) class.

Value

a list with:

prox.mds	the MDS scores of the selected dimensions
g	ggplot object

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(symb.metab)

rf <- randomForest(type ~ ., symb.metab, proximity = TRUE)
```

```
# With confidence ellipses
plotProximity(rf)

# With convex hulls
plotProximity(rf, group.type = "hull")

# With contours
plotProximity(rf, group.type = "contour")

# Remove the points and just show ellipses
plotProximity(rf, point.size = NULL, circle.size = NULL, group.alpha = 0.5)

# Labels instead of a legend
plotProximity(rf, legend.type = "label", point.size = NULL, circle.size = NULL, group.alpha = 0.5)
```

---

plotTrace *Plot Trace*

Description

Plot trace of cumulative OOB (classification) or MSE (regression) error rate by number of trees.

Usage

```
plotTrace(x, pct.correct = TRUE, plot = TRUE)
```

Arguments

x	a rfPermute or randomForest model object.
pct.correct	display y-axis as percent correctly classified (TRUE) or OOB error rate (FALSE).
plot	display the plot?

Value

the ggplot2 object is invisibly returned.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars)
plotTrace(rf)
```

plotVotes *Plot Vote Distribution*

Description

For classification models, plot distribution of votes for each sample in each class.

Usage

```
plotVotes(x, type = NULL, freq.sep.line = TRUE, plot = TRUE)
```

Arguments

x	a rfPermute or randomForest model object.
type	either area for stacked continuous area plot or bar for discrete stacked bar chart. The latter is prefered for small numbers of cases. If not specified, a bar chart will be used if all classes have <= 30 cases.
freq.sep.line	put frequency of original group on second line in facet label? If FALSE, labels are single line. If NULL frequencies will not be included in labels.
plot	display the plot?

Value

the ggplot2 object is invisibly returned.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
library(randomForest)
data(mtcars)

rf <- randomForest(factor(am) ~ ., mtcars)
plotVotes(rf)
```

`rfPermute` *Estimate Permutation p-values for Random Forest Importance Metrics*

Description

Estimate significance of importance metrics for a Random Forest model by permuting the response variable. Produces null distribution of importance metrics for each predictor variable and p-value of observed.

Usage

```
rfPermute(x, ...)

## Default S3 method:
rfPermute(x, y = NULL, ..., num.rep = 100, num.cores = 1)

## S3 method for class 'formula'
rfPermute(
  formula,
  data = NULL,
  ...,
  subset,
  na.action = na.fail,
  num.rep = 100,
  num.cores = 1
)

as.randomForest(x)

## S3 method for class 'rfPermute'
print(x, ...)

## S3 method for class 'rfPermute'
predict(object, ...)
```

Arguments

`x, y, formula, data, subset, na.action, ...`	See `randomForest` for definitions. In `as.randomForest` this is either a `randomForest` or `rfPermute` object to be converted to a `randomForest` object.
`num.rep`	Number of permutation replicates to run to construct null distribution and calculate p-values (default = 100).
`num.cores`	Number of CPUs to distribute permutation results over. Defaults to `NULL` which uses one fewer than the number of cores reported by `detectCores`.
`object`	an `rfPermute` model to be used for prediction. See `predict.randomForest`

Details

All other parameters are as defined in randomForest.formula. A Random Forest model is first created as normal to calculate the observed values of variable importance. The response variable is then permuted num.rep times, with a new Random Forest model built for each permutation step.

Value

An rfPermute object.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

```
# A regression model predicting ozone levels
data(airquality)
ozone.rp <- rfPermute(Ozone ~ ., data = airquality, na.action = na.omit, ntree = 100, num.rep = 50)
ozone.rp

# Plot the scaled importance distributions
# Significant (p <= 0.05) predictors are in red
plotImportance(ozone.rp, scale = TRUE)

# Plot the importance null distributions and observed values for two of the predictors
plotNull(ozone.rp, preds = c("Solar.R", "Month"))


# A classification model classifying cars to manual or automatic transmission
data(mtcars)

am.rp <- rfPermute(factor(am) ~ ., mtcars, ntree = 100, num.rep = 50)
summary(am.rp)


plotImportance(am.rp, scale = TRUE, sig.only = TRUE)
```

---

rfPermute_package rfPermute *package*

Description

Random Forest Predictor Importance Significance and Model Diagnostics.

Usage

```
rfPermuteTutorial()
```

summary *Diagnostics of* rfPermute *or* randomForest *models.*

Description

Combine plots of error traces and inbag rates.

Usage

```
## S3 method for class 'randomForest'
summary(object, ...)

## S3 method for class 'rfPermute'
summary(object, ...)
```

Arguments

object	a rfPermute or randomForest model object to summarize.
...	arguments passed to plotInbag.

Value

A combination of plots from plotTrace and plotInbag as well as summary confusion matrices (classification) or error rates (regression) from the model.

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

plotTrace, plotInbag

Examples

```
# A regression model using the ozone example
data(airquality)
ozone.rp <- rfPermute(
  Ozone ~ ., data = airquality, na.action = na.omit,
  ntree = 100, nrep = 50, num.cores = 1
)

summary(ozone.rp)
```

`symb.metab` *Symbiodinium type metabolite profiles*

Description

A data.frame of 155 metabolite relative concentrations for 64 samples of four Symbiodinium clade types.

Usage

```
data(symb.metab)
```

Format

data.frame

References

Klueter, A.; Crandall, J.B.; Archer, F.I.; Teece, M.A.; Coffroth, M.A. Taxonomic and Environmental Variation of Metabolite Profiles in Marine Dinoflagellates of the Genus Symbiodinium. Metabolites 2015, 5, 74-99.

Index