Package ‘rfPermute’

October 3, 2016

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  2.1.5
URL  https://github.com/EricArcher/rfPermute
BugReports  https://github.com/EricArcher/rfPermute/issues
Depends  R (>= 3.2.0), randomForest
Imports  abind, ggplot2, graphics, grDevices, gridExtra, parallel, reshape2, stats, swfscMisc (>= 1.1.1)
License  GPL (>= 2)
RoxygenNote  5.0.1.9000
NeedsCompilation  no
Author  Eric Archer [aut, cre]
Maintainer  Eric Archer <eric.archer@noaa.gov>
Repository  CRAN
Date/Publication  2016-10-03 22:44:41

R topics documented:

   classConfiInt ......................................................... 2
   cleanRFdata .......................................................... 3
   confusionMatrix ....................................................... 3
   expdErrRate .......................................................... 4
   impHeatmap ............................................................ 5
### classConfInt

#### Description

Calculate confidence intervals for Random Forest classifications

#### Usage

```r
classConfInt(rf, conf.level = 0.95, threshold = 0.8)
```

#### Arguments

- `rf`: a `randomForest` object
- `conf.level`: confidence level for the `binom.test` confidence interval
- `threshold`: threshold to test observed classification probability against.

#### Value

A matrix with the following columns for each class and overall:

- **pct.correct**: percent correctly classified
- **LCI_##, UCI_##**: the lower and upper central confidence intervals given `conf.level`
- **Pr.gt_##**: the probability that the true classification probability is >= `threshold`

#### Author(s)

Eric Archer <eric.archer@noaa.gov>

#### Examples

```r
data(symb.metab)

rf <- randomForest(type ~ ., symb.metab)
classConfInt(rf)
```
**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` column(s) 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>

---

**confusionMatrix**

*Confusion Matrix*

**Description**

Generate a confusion matrix for Random Forest analyses with error rates translated into percent correctly classified, and columns for confidence intervals and expected classification rates (priors) added.

**Usage**

```
confusionMatrix(rf, conf.level = 0.95, threshold = 0.8)
```

**Arguments**

- `rf` a `randomForest` object.
- `conf.level` confidence level for the `binom.test` confidence interval.
- `threshold` threshold to test observed classification probability against.
Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

classConfInt, exptdErrRate

Examples

data(mtcars)

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

<table>
<thead>
<tr>
<th>exptdErrRate</th>
<th>Expected Error Rate</th>
</tr>
</thead>
</table>

Description

Calculate expected OOB error rates (priors) for randomForest classification model based on random assignment and class sizes.

Usage

exptdErrRate(rf)

Arguments

rf

an object inheriting from link{randomForest}.

Value

a vector of expected error rates (priors) for each class.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

data(mtcars)

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

Importance Heatmap

Description
Plot heatmap of importance scores or ranks from a classification model.

Usage
impHeatmap(rf, n = NULL, ranks = TRUE, plot = TRUE, xlab = NULL, ylab = NULL, scale = TRUE, alpha = 0.05)

Arguments
- `rf`: an object inheriting from `randomForest`.
- `n`: Plot `n` most important predictors.
- `ranks`: plot ranks instead of actual importance scores?
- `plot`: print the plot?
- `xlab`, `ylab`: labels for the x and y axes.
- `scale`: For permutation based measures, should the measures be divided their "standard errors"?
- `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 alpha will be denoted in the heatmap by a black border. If set to NULL, no border is drawn.

Details
`rf` must be a classification model run with `importance = TRUE`.

Value
the `ggplot` object is invisibly returned.

Author(s)
Eric Archer <eric.archer@noaa.gov>

Examples
```r
data(mtcars)
# A randomForest model
rf <- randomForest(factor(am) ~ ., mtcars, importance = TRUE)
importance(rf)
```
impHeatmap(rf, xlab = "Transmission", ylab = "Predictor")

# An rfPermute model with significant predictors identified
rp <- rfPermute(factor(am) ~ ., mtcars, nrep = 100)
impHeatmap(rp, xlab = "Transmission", ylab = "Predictor")

### pctCorrect

**Percent Correctly Classified**

**Description**

Calculate the percent of individuals correctly classified in a specified percent of trees in the forest.

**Usage**

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

**Arguments**

- `rf` a `randomForest` or `rfPermute` 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**

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

Plot the Random Forest importance distributions, with significant p-values as estimated in rfPermute.

Usage

```r
### S3 method for class 'rp.importance'
plot(x, alpha = 0.05, sig.only = FALSE,
     type = NULL, n = NULL, main = NULL, ...)
```

Arguments

- `x`: An object produced by a call to `rp.importance`.
- `alpha`: Critical alpha to identify "significant" predictors.
- `sig.only`: Plot only the significant (<= alpha) predictors?
- `type`: character vector listing which importance measures to plot. Can be class names or names of overall importance measures (e.g., "MeanDecreaseAccuracy") in the `rp.importance` object.
- `n`: Plot `n` most important predictors.
- `main`: Main title for plot.
- `...`: Optional arguments which will be ignored.

Details

The function will generate a panel of plots, one for each importance type.

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

rfPermute, rp.importance

Examples

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

# Plot the unscaled importance distributions and highlight significant predictors
plot(rp.importance(ozone.rfP, scale = FALSE))
```
# ... and the scaled measures
plot(rp.importance(ozone.rfP, scale = TRUE))

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"))

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**: Either a numeric or 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.

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.rfP <- rfPermute(Ozone ~ ., data = airquality, ntree = 100, na.action = na.omit, nrep = 50)

# Plot the null distributions and observed values.
plotNull(ozone.rfP)
**plotVotes**

**Vote Distribution**

**Description**

Plot distribution of votes for each sample in each class.

**Usage**

```
plotVotes(rf, type = NULL, plot = TRUE)
```

**Arguments**

- `rf`: an object inheriting from `randomForest`.
- `type`: either `area` for stacked continuous area plot or `bar` for discrete stacked bar chart. The latter is preferred for small numbers of cases. If not specified, a bar chart will be used if all classes have <= 30 cases.
- `plot`: display the plot?

**Value**

The `ggplot` object is invisibly returned.

**Author(s)**

Eric Archer <eric.archer@noaa.gov>

**Examples**

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

**proximityPlot**

**Plot Random Forest Proximity Scores**

**Description**

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

**Usage**

```
proximityPlot(rf, dim.x = 1, dim.y = 2, legend.loc = c("top", "bottom", "left", "right"), point.size = 2, circle.size = 8, circle.border = 1, hull.alpha = 0.3, plot = TRUE)
```
Arguments

rf
A randomForest object.
dim.x, dim.y
Numeric values giving x and y dimensions to plot from multidimensional scaling of proximity scores.
legend.loc
Character keyword specifying location of legend. Can be "bottom", "top", "left", "right".
point.size
Size of central points.
circle.size
Size of circles around correctly classified points as argument to 'cex'. Set to NULL for no circles.
circle.border
Width of circle border.
hull.alpha
value giving alpha transparency level for convex hull shading. Setting to NULL produces no shading. Ignored for regression models.
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, a convex hull is drawn around the a-priori classes with points colored according to original (inner) and predicted (outer) class.

Value

a list with prox.cmd: the MDS scores of the selected dimensions, and g the ggplot object.

Author(s)

Eric Archer <eric.archer@noaa.gov>

Examples

data(symb.metab)

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

Usage

rfPermute(x, ...)

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

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

Arguments

x, y, formula, data, subset, na.action, ...  
See randomForest for definitions.
nrep   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.

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 nrep times, with a new Random Forest model built for each permutation step.

Value

An rfPermute object which contains all of the components of a randomForest object plus:

null.dist   A list containing two three-dimensional arrays of null distributions for unscaled and scaled importance measures.
pval   A three dimensional array containing permutation p-values for unscaled and scaled importance measures.

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

plotNull for plotting null distributions from the rfPermute objects.
 rp.importance for extracting importance measures.
 rp.combine for combining multiple rfPermute objects.
 proximityPlot for plotting case proximities.
 impHeatmap for plotting a heatmap of importance scores.
 randomForest

Examples

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

# Plot the null distributions and observed values.
layout(matrix(1:6, nrow = 2))
plotNull(ozone.rfP)
layout(matrix(1))

# Plot the unscaled importance distributions and highlight significant predictors
plot(rp.importance(ozone.rfP, scale = FALSE))

# ... and the scaled measures
plot(rp.importance(ozone.rfP, scale = TRUE))

---

**rp.combine**

**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

rp.combine(...)

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)
rp2 <- rfPermute(Species ~ ., iris, ntree = 50, norm.votes = FALSE, nrep = 100)
rp3 <- rfPermute(Species ~ ., iris, ntree = 50, norm.votes = FALSE, nrep = 100)
rp.all <- rp.combine(rp1, rp2, rp3)
rp.importance

Extract rfPermute Importance Scores and p-values.

Description

Extract a matrix of the observed importance scores and p-values from the object produced by a call to rfPermute.

Usage

rp.importance(x, scale = TRUE, sort.by = NULL, decreasing = TRUE)

Arguments

x An object produced by a call to rfPermute.
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?

Details

p-values can be given to the sort.by argument by adding '.pval' to the column name of the desired column from the importance element of the rfPermute object.

Author(s)

Eric Archer <eric.archer@noaa.gov>

See Also

rfPermute, plot rp.importance
**Examples**

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

imp.unscaled <- rp.importance(ozone.rfp, scale = TRUE)
imp.unscaled

imp.scaled <- rp.importance(ozone.rfp, scale = TRUE)
imp.scaled
```

---

**symb.metab**

<table>
<thead>
<tr>
<th>Symbiodinium type metabolite profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>symb.metab</td>
</tr>
</tbody>
</table>

**Description**

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

**Usage**

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

**Format**

data.frame

**References**

Index

*Topic classif
    rfPermute, 10
*Topic datasets
    symb.metab, 14
*Topic regression
    rfPermute, 10
*Topic tree
    rfPermute, 10
binom.test, 2, 3
classConfInt, 2, 4
cleanRFdata, 3
combine, 12
confusionMatrix, 3
exptdErrRate, 4, 4
ggplot, 10
impHeatmap, 5, 11
pctCorrect, 6
plot.rp.importance, 7, 13
plotNull, 8, 11
plotVotes, 9
proximityPlot, 9, 11
randomForest, 2, 3, 5, 6, 9, 11
rfPermute, 5, 7, 8, 10, 13
rp.combine, 11, 12
rp.importance, 7, 11, 13
symb.metab, 14