

Package ‘VSURF’

July 18, 2019

Type Package

Title Variable Selection Using Random Forests

Version 1.1.0

Date 2019-07-17

Depends R (>= 3.2.2)

Description Three steps variable selection procedure based on random forests.

Initially developed to handle high dimensional data (for which number of variables largely exceeds number of observations), the package is very versatile and can treat most dimensions of data, for regression and supervised classification problems. First step is dedicated to eliminate irrelevant variables from the dataset. Second step aims to select all variables related to the response for interpretation purpose. Third step refines the selection by eliminating redundancy in the set of variables selected by the second step, for prediction purpose.

Genuer, R. Poggi, J.-M. and Tuleau-Malot, C. (2015)

<<https://journal.r-project.org/archive/2015-2/genuer-poggi-tuleaumalot.pdf>>.

License GPL (>= 2)

LazyData true

URL <https://github.com/robingenuer/VSURF>

BugReports <https://github.com/robingenuer/VSURF/issues>

Imports doParallel, foreach, parallel, Rborist, randomForest, ranger, rpart

Suggests testthat,

RoxygenNote 6.1.1

NeedsCompilation no

Author Robin Genuer [aut, cre],
Jean-Michel Poggi [aut],
Christine Tuleau-Malot [aut]

Maintainer Robin Genuer <Robin.Genuer@u-bordeaux.fr>

Repository CRAN

Date/Publication 2019-07-18 06:36:21 UTC

R topics documented:

plot.VSURF	2
PM10	4
predict.VSURF	5
print.VSURF	7
summary.VSURF	8
toys	9
tune	10
VSURF	11
VSURF_interp	16
VSURF_pred	19
VSURF_thres	21

plot.VSURF

Plot of VSURF results

Description

This function plots 4 graphs illustrating VSURF results.

Usage

```
## S3 method for class 'VSURF'
plot(x, step = "all", var.names = FALSE,
     imp.mean = TRUE, imp.sd = TRUE,
     nvar.imp.mean = length(x$imp.mean.dec),
     nvar.imp.sd = length(x$imp.sd.dec),
     nvar.interp = length(x$vselect.thres),
     nvar.pred = length(x$vselect.pred), ...)

## S3 method for class 'VSURF_thres'
plot(x, var.names = FALSE, imp.mean = TRUE,
     imp.sd = TRUE, nvar.imp.mean = length(x$imp.mean.dec),
     nvar.imp.sd = length(x$imp.sd.dec), ...)

## S3 method for class 'VSURF_interp'
plot(x, var.names = FALSE,
     nvar.interp = length(x$vselect.thres), ...)

## S3 method for class 'VSURF_pred'
plot(x, var.names = FALSE,
     nvar.pred = length(x$vselect.pred), ...)
```

Arguments

<code>x</code>	An object of class <code>VSURF</code> , <code>VSURF_thres</code> , <code>VSURF_interp</code> or <code>VSURF_pred</code> , which is the result of the <code>VSURF</code> function (or resp. <code>VSURF_thres</code> , <code>VSURF_interp</code> or <code>VSURF_pred</code>).
<code>step</code>	A character string indicating which step must be plotted (default is "all"). Available choices are "thres", "interp", "pred".
<code>var.names</code>	If FALSE (default) <code>xticks</code> are the numbering given by the sorting of VI mean, if TRUE they are the variables names.
<code>imp.mean</code>	If TRUE (default) VI mean is plotted, if FALSE it is not.
<code>imp.sd</code>	If TRUE (default) VI standard deviation is plotted, if FALSE it is not.
<code>nvar.imp.mean</code>	The number of variables to be kept for the VI mean plot.
<code>nvar.imp.sd</code>	The number of variables to be kept for the VI standard deviation plot.
<code>nvar.interp</code>	The number of variables to be kept for the "interp" plot.
<code>nvar.pred</code>	The number of variables to be kept for the "pred" plot.
<code>...</code>	Arguments to be passed to <code>par</code> (they will affect all plots) or to others methods of <code>plot</code> .

Details

The 2 graphs of the top row correspond to the "thresholding step" (and only these 2 graphs are plotted by the `plot.VSURF_thres` function). The top left graph plots the mean variable importance in decreasing order (black curve). The red horizontal line represent the value of the threshold. The top right graph plots the standard deviation of variable importance with variables ordered according to their mean variable importance in decreasing order (black curve). The green line represents the predictions given by a CART tree fitted to the black curve (the standard deviations). Finally, the dotted horizontal red line represents the minimum value of the CART predictions, which actually is the value of the threshold.

The bottom left graph corresponds to the "interpretation step" (and only this graph is plotted by the `plot.VSURF_interp` function). It plots the mean OOB error rate of embedded random forests models (from the one with only one variable as predictor, to the one with all variables kept after the "thresholding step"). The vertical red line indicates the retained model.

The bottom right graph corresponds to the "prediction step" (and only this graph is plotted by the `plot.VSURF_pred` function). It plots the mean OOB error rate of embedded random forests models (the difference, here, being that variables are added to the model in a step-wise manner). The retained model is the final one.

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236
- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

VSURF, summary.VSURF

Examples

```
## Not run:
data(iris)
iris.vsurf <- VSURF(iris[,1:4], iris[,5])
plot(iris.vsurf)
plot(iris.vsurf, var.names=TRUE)
plot(iris.vsurf, step="thres")

# A more interesting example with toys data (see \link{toys})
# (a few minutes to execute) and intermediate functions
data(toys)
toys.vsurf <- VSURF(toys$x, toys$y)
plot(toys.vsurf)
plot(toys.vsurf, nvar.imp.mean = 50, nvar.imp.sd = 50)
toys.thres <- VSURF_thres(toys$x, toys$y)
plot(toys.thres)
plot(toys.thres, nvar.imp.mean = 70, imp.sd = FALSE)
toys.interp <- VSURF_interp(toys$x, toys$y, vars = toys.thres$vselect.thres)
plot(toys.interp, var.names = TRUE)
toys.pred <- VSURF_pred(toys$x, toys$y, err.interp = toys.interp$err.interp,
                        vselect.interp = toys.interp$vselect.interp)
plot(toys.pred, var.names = TRUE)

## End(Not run)
```

PM10

Real-world data on PM10 pollution in Rouen area, France

Description

These data are TEOM (Tapered Element Oscillating Microbalance) PM10 concentrations from 2004 to 2006 (1096 days) measured by Air Normand, and the associated weather data provided by Meteo France, the French national meteorological service, using six different monitoring sites.

Usage

ail

gcm

gui

hri

jus

rep

Format

Each object is a data frame.

The description of the 18 variables is the following (note that for `gcm` station, only the pollutant SO₂ is available in addition to PM₁₀, and for `ail` station, there is no other pollutant in addition to PM₁₀):

PM10 Daily mean concentration of PM₁₀, in $\mu\text{g}/\text{m}^3$

NO, NO2, SO2 Daily mean concentration of NO, NO₂, SO₂, in $\mu\text{g}/\text{m}^3$

T.min, T.max, T.moy Daily minimum, maximum and mean temperature, in degree Celsius

DV.maxvv, DV.dom Daily maximum speed and dominant wind direction, in degree (for wind direction, 0 degree corresponds to north)

VV.max, VV.moy Daily maximum and mean wind speed, in m/s

PL.som Daily rainfall, in mm

HR.min, HR.max, HR.moy Daily minimum, maximum and mean relative humidity, in %

PA.moy Daily mean air pressure, in hPa

GTrouen, GTlehavre Daily temperature gradient, in degree Celsius

Details

Six different monitoring stations of the Rouen (Haute Normandie, France) area are considered. The urban station `jus`, the traffic station `gui`, the second most polluted in the region, and `gcm` which is located in an industrial area. In Le Havre, are considered the stations `rep` (the most polluted in the region) and `hri` located at the seaside. Lastly, the station `ail` near Dieppe, because it is rural and coastal, and a priori hardly influenced by social and industrial activity. Grouping by categories: `jus` and `hri` are background urban monitoring sites, `gui` and `rep` are urban sites close to traffic, `gcm` is industrial and `ail` is rural.

Source

F.-X. Jollois, J.-M. Poggi, B. Portier, *Three non-linear statistical methods to analyze PM₁₀ pollution in Rouen area*. CSBIGS 3(1): 1-17, 2009

predict.VSURF *Predict method for VSURF object*

Description

This function predicts new data with random forests, using variables selected by VSURF only.

Usage

```
## S3 method for class 'VSURF'
predict(object, newdata, step = c("interp", "pred"), ...)
```

Arguments

object	An object of class VSURF, which is the result of the VSURF function.
newdata	A data frame or matrix containing new data. (Note: If not given, the out-of-bag predictions of the randomForest object is returned.)
step	A character string indicating which variable set must be used to train the randomForest object (default is c("interp", "pred")). Available choices are "thres", "interp", "pred".
...	further parameters passed to randomForest or predict.randomForest functions (depending on their names).

Details

This method applies for a VSURF object. VSURF selects two sets of variables during its two last steps. For each set of variables, a random forest object is created, by running randomForest on training data using this set of variables only. Then the predict.randomForest function is used to predict new data.

Value

If only one step is indicated in step, a vector of predicted values.

If two or more steps are indicated in step, a data frame of predicted values (each column corresponding to a variable set).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

VSURF

Examples

```
## Not run:
data(iris)
iris.learn <- sample(1:nrow(iris), nrow(iris)/2)
iris.vsurf <- VSURF(iris[iris.learn, 1:4], iris[iris.learn, 5], ntree = 100, nfor.thres = 20,
                  nfor.interp = 10, nfor.pred = 10)
iris.predictions <- predict(iris.vsurf, newdata = iris[-iris.learn, 1:4])

# A more interesting example with toys data (see \link{toys})
# (a few minutes to execute)
data(toys)
toys.learn <- 1:(nrow(toys$x) / 2)
toys.vsurf <- VSURF(toys$x[toys.learn, ], toys$y[toys.learn])
toys.predictions <- predict(toys.vsurf, newdata = toys$x[-toys.learn, ])
## End(Not run)
```

`print.VSURF`*Print of VSURF results*

Description

This function display a small description of VSURF results

Usage

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

Arguments

<code>x</code>	An object of class VSURF, which is the result of the VSURF function.
<code>...</code>	Not used.

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

VSURF, plot.VSURF, summary.VSURF

Examples

```
## Not run:
data(iris)
iris.vsurf <- VSURF(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20,
                   nfor.interp = 10, nfor.pred = 10)
iris.vsurf

## End(Not run)
```

summary.VSURF

Summary of VSURF results

Description

This function displays a summary of VSURF results

Usage

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

Arguments

object	An object of class VSURF, which is the result of the VSURF function.
...	Not used.

Details

This function prints the total computation time of VSURF. It also gives the number of selected variables (and the computation time) at each step of VSURF. In addition, it gives the number of cores and the type of cluster if the parallel version of VSURF was used.

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

VSURF, plot.VSURF

Examples

```
## Not run:
data(iris)
iris.vsurf <- VSURF(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20,
                  nfor.interp = 10, nfor.pred = 10)
summary(iris.vsurf)

# A more interesting example with toys data (see \code{\link{toys}})
# (a few minutes to execute)
data(toys)
toys.vsurf <- VSURF(toys$x, toys$y)
summary(toys.vsurf)
## End(Not run)
```

toys

A simulated dataset called toys data

Description

toys is a simple simulated dataset of a binary classification problem, introduced by Weston et.al..

Format

The format is a list of 2 components:

x a dataframe containing input variables: with 100 obs. of 200 variables

y output variable: a factor with 2 levels "-1" and "1"

Details

It is an equiprobable two class problem, Y belongs to $\{-1,1\}$, with six true variables, the others being some noise. The simulation model is defined through the conditional distribution of the X_i for $Y=y$:

- with probability 0.7, $X^j \sim N(yj,1)$ for $j=1,2,3$ and $X^j \sim N(0,1)$ for $j=4,5,6$;
- with probability 0.3, $X^j \sim N(0,1)$ for $j=1,2,3$ and $X^j \sim N(y(j-3),1)$ for $j=4,5,6$;
- the other variables are noise, $X^j \sim N(0,1)$ for $j=7, \dots, p$.

After simulation, the obtained variables are finally standardized.

Source

Weston, J., Elisseeff, A., Schoelkopf, B., Tipping, M. (2003), *Use of the zero norm with linear models and Kernel methods*, J. Machine Learn. Res. 3, 1439-1461

Examples

```
data(toys)
toys.rf <- randomForest::randomForest(toys$x, toys$y)
toys.rf

## Not run:
# VSURF applied for toys data:
# (a few minutes to execute)
data(toys)
toys.vsurf <- VSURF(toys$x, toys$y)
toys.vsurf

## End(Not run)
```

tune

Tuning of the thresholding and interpretation steps of VSURF

Description

This function allows to tune the "thresholding" and "interpretation step" of VSURF, without rerunning all computations.

Usage

```
tune(x, ...)
```

```
## S3 method for class 'VSURF_thres'
tune(x, nmin = 1, ...)
```

```
## S3 method for class 'VSURF_interp'
tune(x, nsd = 1, ...)
```

Arguments

x	An object of class <code>VSURF_thres</code> or <code>VSURF_interp</code> , which is the result of the <code>VSURF_thres</code> or <code>VSURF_interp</code> function.
...	Not used.
nmin	Number of times the "minimum value" is multiplied to set threshold value. See details below.
nsd	Number of times the standard deviation of the minimum value of <code>err.interp</code> is multiplied. See details below.

Details

In `VSURF_thres` function, the actual threshold is performed like this: only variables with a mean VI larger than `nmin * min.thres` are kept. The function `tune.VSURF_thres` allows you to change the value of `nmin` (which multiply the estimated threshold value `min.thres`), without rerunning all computations. To get a softer threshold than default, choose a value of `nmin` less than 1, and to get a harder one, choose a value larger than 1.

In `VSURF_interp` function, the smallest model (and hence its corresponding variables) having a mean OOB error rate less than `err.min + nsd * sd.min` is selected. The function `tune.VSURF_interp` allows to change the value of `nsd` (which multiply the standard deviation of the minimum OOB error rate `sd.min`), without rerunning all computations. To get a larger model than default, choose a value of `nsd` less than 1, and to get a smaller one, choose a value larger than 1.

Value

An object with the same structure than the original output (from `VSURF_thres` or `VSURF_interp`).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

`VSURF`, `VSURF_thres`, `VSURF_interp`

Examples

```
## Not run:
data(iris)
iris.thres <- VSURF_thres(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20)
iris.thres.tuned <- tune(iris.thres, nmin = 10)
iris.thres.tuned
iris.interp <- VSURF_interp(iris[,1:4], iris[,5], vars = iris.thres$vselect.thres,
                           nfor.interp = 10)
iris.interp.tuned <- tune(iris.interp, nsd = 10)
iris.interp.tuned

## End(Not run)
```

Description

Three steps variable selection procedure based on random forests for supervised classification and regression problems. First step ("thresholding step") is dedicated to eliminate irrelevant variables from the dataset. Second step ("interpretation step") aims to select all variables related to the response for interpretation purpose. Third step ("prediction step") refines the selection by eliminating redundancy in the set of variables selected by the second step, for prediction purpose.

Usage

```
VSURF(x, ...)

## Default S3 method:
VSURF(x, y, ntree = 2000,
      mtry = max(floor(ncol(x)/3), 1), nfor.thres = 50, nmin = 1,
      nfor.interp = 25, nsd = 1, nfor.pred = 25, nmj = 1,
      RFimplem = "randomForest", parallel = FALSE, ncores = detectCores()
      - 1, clusterType = "PSOCK", verbose = TRUE, ...)

## S3 method for class 'formula'
VSURF(formula, data, ..., na.action = na.fail)
```

Arguments

<code>x</code> , <code>formula</code>	A data frame or a matrix of predictors, the columns represent the variables. Or a formula describing the model to be fitted.
<code>...</code>	others parameters to be passed on to the <code>randomForest</code> function (see <code>?randomForest</code> for further information).
<code>y</code>	A response vector (must be a factor for classification problems and numeric for regression ones).
<code>ntree</code>	Number of trees in each forests grown. Standard parameter of <code>randomForest</code> .
<code>mtry</code>	Number of variables randomly sampled as candidates at each split. Standard parameter of <code>randomForest</code> .
<code>nfor.thres</code>	Number of forests grown for "thresholding step" (first of the three steps).
<code>nmin</code>	Number of times the "minimum value" is multiplied to set threshold value. See details below.
<code>nfor.interp</code>	Number of forests grown for "interpretation step" (second of the three steps).
<code>nsd</code>	Number of times the standard deviation of the minimum value of <code>err.interp</code> is multiplied. See details below.
<code>nfor.pred</code>	Number of forests grown for "prediction step" (last of the three steps).
<code>nmj</code>	Number of times the mean jump is multiplied. See details below.

<code>RFimplem</code>	Choice of the random forests implementation to use : "randomForest" (default), "ranger" or "Rborist" (not that if "Rborist" is chosen, "randomForest" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>parallel</code>	A logical indicating if you want VSURF to run in parallel on multiple cores (default to FALSE). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>ncores</code>	Number of cores to use. Default is set to the number of cores detected by R minus 1.
<code>clusterType</code>	Type of the multiple cores cluster used to run VSURF in parallel. Must be chosen among "PSOCK" (default: SOCKET cluster available locally on all OS), "FORK" (local too, only available for Linux and Mac OS), "MPI" (can be used on a remote cluster, which needs <code>snow</code> and <code>Rmpi</code> packages installed), "ranger" and "Rborist" for internal parallelizations of those packages (not that if "Rborist" is chosen, "SOCKET" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 2 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , in this order.
<code>verbose</code>	A logical indicating if information about method's progress (included progress bars for each step) must be printed (default to TRUE). Adds a small extra overload.
<code>data</code>	a data frame containing the variables in the model.
<code>na.action</code>	A function to specify the action to be taken if NAs are found. (NOTE: If given, this argument must be named, and as <code>randomForest</code> it is only used with the formula-type call.)

Details

- First step ("thresholding step"): first, `nfor.thres` random forests are computed using the function `randomForest` with arguments `importance=TRUE`, and our choice of default values for `ntree` and `mtry` (which are higher than default in `randomForest` to get a more stable variable importance measure). Then variables are sorted according to their mean variable importance (VI), in decreasing order. This order is kept all along the procedure. Next, a threshold is computed: `min.thres`, the minimum predicted value of a pruned CART tree fitted to the curve of the standard deviations of VI. Finally, the actual "thresholding step" is performed: only variables with a mean VI larger than `nmin * min.thres` are kept.
- Second step ("interpretation step"): the variables selected by the first step are considered. `nfor.interp` embedded random forests models are grown, starting with the random forest build with only the most important variable and ending with all variables selected in the first step. Then, `err.min` the minimum mean out-of-bag (OOB) error of these models and its associated standard deviation `sd.min` are computed. Finally, the smallest model (and hence its corresponding variables) having a mean OOB error less than `err.min + nsd * sd.min` is selected.

Note that for this step (and the next one), the `mtry` parameter of `randomForest` is set to its default value (see `randomForest`) if `nvm`, the number of variables in the model, is not

greater than the number of observations, while it is set to $nvm/3$ otherwise. This is to ensure quality of OOB error estimations along embedded RF models.

- Third step ("prediction step"): the starting point is the same than in the second step. However, now the variables are added to the model in a stepwise manner. `mean.jump`, the mean jump value is calculated using variables that have been left out by the second step, and is set as the mean absolute difference between mean OOB errors of one model and its first following model. Hence a variable is included in the model if the mean OOB error decrease is larger than $nmj * mean.jump$.

As for interpretation step, the `mtry` parameter of `randomForest` is set to its default value if nvm , the number of variables in the model, is not greater than the number of observations, while it is set to $nvm/3$ otherwise.

VSURF is able to run using multiple cores in parallel (see `parallel`, `clusterType` and `ncores` arguments).

Value

An object of class `VSURF`, which is a list with the following components:

<code>vartable.thres</code>	A vector of indexes of variables selected after "thresholding step", sorted according to their mean VI, in decreasing order.
<code>vartable.interp</code>	A vector of indexes of variables selected after "interpretation step".
<code>vartable.pred</code>	A vector of indexes of variables selected after "prediction step".
<code>num.vartable</code>	A vector of the 3 numbers of variables selected resp. by "thresholding step", "interpretation step" and "prediction step".
<code>imp.vartable.thres</code>	A vector of importance of the <code>vartable.thres</code> variables.
<code>min.thres</code>	The minimum predicted value of a pruned CART tree fitted to the curve of the standard deviations of VI.
<code>imp.mean.dec</code>	A vector of the variables importance means (over <code>nfor.thres</code> runs), in decreasing order.
<code>imp.mean.dec.ind</code>	The ordering index vector associated to the sorting of variables importance means.
<code>imp.sd.dec</code>	A vector of standard deviations of all variables importance. The order is given by <code>imp.mean.dec.ind</code> .
<code>mean.perf</code>	Mean OOB error rate, obtained by a random forests build on all variables.
<code>pred.pruned.tree</code>	Predictions of the CART tree fitted to the curve of the standard deviations of VI.
<code>err.interp</code>	A vector of the mean OOB error rates of the embedded random forests models build during the "interpretation step".

<code>sd.min</code>	The standard deviation of OOB error rates associated to the random forests model attaining the minimum mean OOB error rate during the "interpretation step".
<code>err.pred</code>	A vector of the mean OOB error rates of the random forests models build during the "prediction step".
<code>mean.jump</code>	The mean jump value computed during the "prediction step".
<code>nmin, nsd, nmj</code>	Corresponding parameters values.
<code>overall.time</code>	Overall computation time.
<code>comput.times</code>	A list of the 3 computation times respectively associated with the 3 steps: "thresholding", "interpretation" and "prediction".
<code>RFimplem</code>	The RF implementation used to run VSURF, among "randomForest" (default), "ranger" and "Rborist" or a vector of length 3 with those.
<code>ncores</code>	The number of cores used to run VSURF in parallel (NULL if VSURF did not run in parallel).
<code>clusterType</code>	The type of the cluster used to run VSURF in parallel (NULL if VSURF did not run in parallel).
<code>call</code>	The original call to VSURF.
<code>terms</code>	Terms associated to the formula (only if formula-type call was used).
<code>na.action</code>	Method used to deal with missing values (only if formula-type call was used).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236
- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

`plot.VSURF`, `summary.VSURF`, `VSURF_thres`, `VSURF_interp`, `VSURF_pred`, `tune`

Examples

```
data(iris)
iris.vsurf <- VSURF(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20,
                  nfor.interp = 10, nfor.pred = 10)

iris.vsurf

## Not run:
# A more interesting example with toys data (see \code{\link{toys}})
# (a few minutes to execute)
```

```

data(toys)
toys.vsurf <- VSURF(toys$x, toys$y)
toys.vsurf

# VSURF run on 2 cores in parallel (using a SOCKET cluster):
data(toys)
toys.vsurf.parallel <- VSURF(toys$x, toys$y, parallel = TRUE, ncores = 2)

## End(Not run)

```

VSURF_interp

Interpretation step of VSURF

Description

Interpretation step aims to select all variables related to the response for interpretation purpose. This is the second step of the VSURF function. It is designed to be executed after the thresholding step VSURF_thres.

Usage

```

VSURF_interp(x, ...)

## Default S3 method:
VSURF_interp(x, y, ntree = 2000, vars,
  nfor.interp = 25, nsd = 1, RFimplem = "randomForest",
  parallel = FALSE, ncores = detectCores() - 1,
  clusterType = "PSOCK", verbose = TRUE, ...)

## S3 method for class 'formula'
VSURF_interp(formula, data, ..., na.action = na.fail)

```

Arguments

<code>x</code> , <code>formula</code>	A data frame or a matrix of predictors, the columns represent the variables. Or a formula describing the model to be fitted.
<code>...</code>	others parameters to be passed on to the <code>randomForest</code> function (see <code>?randomForest</code> for further information).
<code>y</code>	A response vector (must be a factor for classification problems and numeric for regression ones).
<code>ntree</code>	Number of trees in each forests grown. Standard parameter of <code>randomForest</code> .
<code>vars</code>	A vector of variable indices. Typically, indices of variables selected by thresholding step (see value <code>varselect.thres</code> of <code>VSURF_thres</code> function).
<code>nfor.interp</code>	Number of forests grown.

<code>nsd</code>	Number of times the standard deviation of the minimum value of <code>err.interp</code> is multiplied. See details below.
<code>RFimplem</code>	Choice of the random forests implementation to use : "randomForest" (default), "ranger" or "Rborist" (not that if "Rborist" is chosen, "randomForest" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>parallel</code>	A logical indicating if you want VSURF to run in parallel on multiple cores (default to FALSE). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>ncores</code>	Number of cores to use. Default is set to the number of cores detected by R minus 1.
<code>clusterType</code>	Type of the multiple cores cluster used to run VSURF in parallel. Must be chosen among "PSOCK" (default: SOCKET cluster available locally on all OS), "FORK" (local too, only available for Linux and Mac OS), "MPI" (can be used on a remote cluster, which needs <code>snow</code> and <code>Rmpi</code> packages installed), "ranger" and "Rborist" for internal parallelizations of those packages (not that if "Rborist" is chosen, "SOCKET" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 2 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , in this order.
<code>verbose</code>	A logical indicating if information about method's progress (included progress bars for each step) must be printed (default to TRUE). Adds a small extra overload.
<code>data</code>	a data frame containing the variables in the model.
<code>na.action</code>	A function to specify the action to be taken if NAs are found. (NOTE: If given, this argument must be named, and as <code>randomForest</code> it is only used with the formula-type call.)

Details

`nfor.interp` embedded random forests models are grown, starting with the random forest build with only the most important variable and ending with all variables. Then, `err.min` the minimum mean out-of-bag (OOB) error rate of these models and its associated standard deviation `sd.min` are computed. Finally, the smallest model (and hence its corresponding variables) having a mean OOB error less than `err.min + nsd * sd.min` is selected.

Note that, the `mtry` parameter of `randomForest` is set to its default value (see `randomForest`) if `nvm`, the number of variables in the model, is not greater than the number of observations, while it is set to `nvm/3` otherwise. This is to ensure quality of OOB error estimations along embedded RF models.

Value

An object of class `VSURF_interp`, which is a list with the following components:

`vareselect.interp`
A vector of indices of selected variables.

<code>err.interp</code>	A vector of the mean OOB error rates of the embedded random forests models.
<code>sd.min</code>	The standard deviation of OOB error rates associated to the random forests model attaining the minimum mean OOB error rate.
<code>num.vareselect.interp</code>	The number of selected variables.
<code>vareselect.thres</code>	A vector of indexes of variables selected after "thresholding step", sorted according to their mean VI, in decreasing order.
<code>nsd</code>	Value of the parameter in the call.
<code>comput.time</code>	Computation time.
<code>RFimplem</code>	The RF implementation used to run <code>VSURF_interp</code> .
<code>ncores</code>	The number of cores used to run <code>VSURF_interp</code> in parallel (NULL if <code>VSURF_interp</code> did not run in parallel).
<code>clusterType</code>	The type of the cluster used to run <code>VSURF_interp</code> in parallel (NULL if <code>VSURF_interp</code> did not run in parallel).
<code>call</code>	The original call to <code>VSURF</code> .
<code>terms</code>	Terms associated to the formula (only if formula-type call was used).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236
- Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

`VSURF`, `tune`

Examples

```
data(iris)
iris.thres <- VSURF_thres(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20)
iris.interp <- VSURF_interp(iris[,1:4], iris[,5],
  vars = iris.thres$vareselect.thres, nfor.interp = 10)
iris.interp

## Not run:
# A more interesting example with toys data (see \code{\link{toys}})
# (a few minutes to execute)
data(toys)
toys.thres <- VSURF_thres(toys$x, toys$y)
```

```

toys.interp <- VSURF_interp(toys$x, toys$y,
  vars = toys.thres$vareselect.thres)
toys.interp
## End(Not run)

```

VSURF_pred

Prediction step of VSURF

Description

Prediction step refines the selection of interpretation step `VSURF_interp` by eliminating redundancy in the set of variables selected, for prediction purpose. This is the third step of the `VSURF` function.

Usage

```

VSURF_pred(x, ...)

## Default S3 method:
VSURF_pred(x, y, ntree = 2000, err.interp,
  vareselect.interp, nfor.pred = 25, nmj = 1,
  RFimplem = "randomForest", parallel = FALSE, ncores = detectCores()
  - 1, verbose = TRUE, ...)

## S3 method for class 'formula'
VSURF_pred(formula, data, ..., na.action = na.fail)

```

Arguments

<code>x</code> , <code>formula</code>	A data frame or a matrix of predictors, the columns represent the variables. Or a formula describing the model to be fitted.
<code>...</code>	others parameters to be passed on to the <code>randomForest</code> function (see <code>?randomForest</code> for further information).
<code>y</code>	A response vector (must be a factor for classification problems and numeric for regression ones).
<code>ntree</code>	Number of trees in each forests grown. Standard parameter of <code>randomForest</code> .
<code>err.interp</code>	A vector of the mean OOB error rates of the embedded random forests models build during interpretation step (value <code>err.interp</code> of function <code>VSURF_interp</code>).
<code>vareselect.interp</code>	A vector of indices of variables selected after interpretation step.
<code>nfor.pred</code>	Number of forests grown.
<code>nmj</code>	Number of times the mean jump is multiplied. See details below.

<code>RFimplem</code>	Choice of the random forests implementation to use : "randomForest" (default), "ranger" or "Rborist" (not that if "Rborist" is chosen, "randomForest" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>parallel</code>	A logical indicating if you want VSURF to run in parallel on multiple cores (default to FALSE). If a vector of length 3 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , <code>VSURF_pred</code> , in this order.
<code>ncores</code>	Number of cores to use. Default is set to the number of cores detected by R minus 1.
<code>verbose</code>	A logical indicating if information about method's progress (included progress bars for each step) must be printed (default to TRUE). Adds a small extra overload.
<code>data</code>	a data frame containing the variables in the model.
<code>na.action</code>	A function to specify the action to be taken if NAs are found. (NOTE: If given, this argument must be named, and as <code>randomForest</code> it is only used with the formula-type call.)

Details

`nfor.pred` embedded random forests models are grown, starting with the random forest build with only the most important variable. Variables are added to the model in a stepwise manner. The mean jump value `mean.jump` is calculated using variables that have been left out by interpretation step, and is set as the mean absolute difference between mean OOB errors of one model and its first following model. Hence a variable is included in the model if the mean OOB error decrease is larger than `nmj * mean.jump`.

Note that, the `mtry` parameter of `randomForest` is set to its default value (see `randomForest`) if `nvm`, the number of variables in the model, is not greater than the number of observations, while it is set to `nvm/3` otherwise. This is to ensure quality of OOB error estimations along embedded RF models.

Value

An object of class `VSURF_pred`, which is a list with the following components:

<code>vartselect.pred</code>	A vector of indices of variables selected after "prediction step".
<code>err.pred</code>	A vector of the mean OOB error rates of the random forests models build during the "prediction step".
<code>mean.jump</code>	The mean jump value computed during the "prediction step".
<code>num.vartselect.pred</code>	The number of selected variables.
<code>nmj</code>	Value of the parameter in the call.
<code>comput.time</code>	Computation time.
<code>RFimplem</code>	The RF implementation used to run <code>VSURF_pred</code> .

call The original call to VSURF.
 terms Terms associated to the formula (only if formula-type call was used).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

VSURF

Examples

```
data(iris)
iris.thres <- VSURF_thres(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20)
iris.interp <- VSURF_interp(iris[,1:4], iris[,5],
  vars = iris.thres$vselect.thres, nfor.interp = 10)
iris.pred <- VSURF_pred(iris[,1:4], iris[,5],
  err.interp = iris.interp$err.interp,
  vselect.interp = iris.interp$vselect.interp, nfor.pred = 10)
iris.pred

## Not run:
# A more interesting example with toys data (see \code{\link{toys}})
# (a few minutes to execute)
data(toys)
toys.thres <- VSURF_thres(toys$x, toys$y)
toys.interp <- VSURF_interp(toys$x, toys$y,
  vars = toys.thres$vselect.thres)
toys.pred <- VSURF_pred(toys$x, toys$y, err.interp = toys.interp$err.interp,
  vselect.interp = toys.interp$vselect.interp)
toys.pred
## End(Not run)
```

Description

Thresholding step is dedicated to roughly eliminate irrelevant variables a the dataset. This is the first step of the VSURF function. For refined variable selection, see VSURF other steps: VSURF_interp and VSURF_pred.

Usage

```
VSURF_thres(x, ...)

## Default S3 method:
VSURF_thres(x, y, ntree = 2000,
            mtry = max(floor(ncol(x)/3), 1), nfor.thres = 50, nmin = 1,
            RFimplem = "randomForest", parallel = FALSE, clusterType = "PSOCK",
            ncores = parallel::detectCores() - 1, verbose = TRUE, ...)

## S3 method for class 'formula'
VSURF_thres(formula, data, ..., na.action = na.fail)
```

Arguments

<code>x</code> , <code>formula</code>	A data frame or a matrix of predictors, the columns represent the variables. Or a formula describing the model to be fitted.
<code>...</code>	others parameters to be passed on to the <code>randomForest</code> function (see <code>?randomForest</code> for further information).
<code>y</code>	A response vector (must be a factor for classification problems and numeric for regression ones).
<code>ntree</code>	Number of trees in each forests grown. Standard parameter of <code>randomForest</code> .
<code>mtry</code>	Number of variables randomly sampled as candidates at each split. Standard parameter of <code>randomForest</code> .
<code>nfor.thres</code>	Number of forests grown.
<code>nmin</code>	Number of times the "minimum value" is multiplied to set threshold value. See details below.
<code>RFimplem</code>	Choice of the random forests implementation to use : "randomForest" (default), "ranger" or "Rborist" (not that if "Rborist" is chosen, "randoForest" will still be used for the first step VSURF_thres). If a vector of length 3 is given, each coordinate is passed to each intermediate function: VSURF_thres, VSURF_interp, VSURF_pred, in this order.
<code>parallel</code>	A logical indicating if you want VSURF to run in parallel on multiple cores (default to FALSE). If a vector of length 3 is given, each coordinate is passed to each intermediate function: VSURF_thres, VSURF_interp, VSURF_pred, in this order.
<code>clusterType</code>	Type of the multiple cores cluster used to run VSURF in parallel. Must be chosen among "PSOCK" (default: SOCKET cluster available locally on all OS), "FORK" (local too, only available for Linux and Mac OS), "MPI" (can be used on a remote cluster, which needs <code>snow</code> and <code>Rmpi</code> packages installed), "ranger" and "Rborist" for internal parallelizations of those packages (not that if "Rborist"

	is chosen, "SOCKET" will still be used for the first step <code>VSURF_thres</code>). If a vector of length 2 is given, each coordinate is passed to each intermediate function: <code>VSURF_thres</code> , <code>VSURF_interp</code> , in this order.
<code>ncores</code>	Number of cores to use. Default is set to the number of cores detected by R minus 1.
<code>verbose</code>	A logical indicating if information about method's progress (included progress bars for each step) must be printed (default to TRUE). Adds a small extra overload.
<code>data</code>	a data frame containing the variables in the model.
<code>na.action</code>	A function to specify the action to be taken if NAs are found. (NOTE: If given, this argument must be named, and as <code>randomForest</code> it is only used with the formula-type call.)

Details

First, `nfor.thres` random forests are computed using the function `randomForest` with arguments `importance=TRUE`, and our choice of default values for `ntree` and `mtry` (which are higher than default in `randomForest` to get a more stable variable importance measure). Then variables are sorted according to their mean variable importance (VI), in decreasing order. This order is kept all along the procedure. Next, a threshold is computed: `min.thres`, the minimum predicted value of a pruned CART tree fitted to the curve of the standard deviations of VI. Finally, the actual thresholding is performed: only variables with a mean VI larger than `nmin * min.thres` are kept.

Value

An object of class `VSURF_thres`, which is a list with the following components:

<code>vareselect.thres</code>	A vector of indices of selected variables, sorted according to their mean VI, in decreasing order.
<code>imp.vareselect.thres</code>	A vector of importance of the <code>vareselect.thres</code> variables.
<code>min.thres</code>	The minimum predicted value of a pruned CART tree fitted to the curve of the standard deviations of VI.
<code>num.vareselect.thres</code>	The number of selected variables.
<code>imp.mean.dec</code>	A vector of the variables importance means (over <code>nfor.thres</code> runs), in decreasing order.
<code>imp.mean.dec.ind</code>	The ordering index vector associated to the sorting of variables importance means.
<code>imp.sd.dec</code>	A vector of standard deviations of all variables importance. The order is given by <code>imp.mean.dec.ind</code> .
<code>mean.perf</code>	The mean OOB error rate, obtained by a random forests build with all variables.

<code>pred.pruned.tree</code>	The predictions of the CART tree fitted to the curve of the standard deviations of VI.
<code>nmin</code>	Value of the parameter in the call.
<code>comput.time</code>	Computation time.
<code>RFimplem</code>	The RF implementation used to run <code>VSURF_thres</code> .
<code>ncores</code>	The number of cores used to run <code>VSURF_thres</code> in parallel (NULL if <code>VSURF_thres</code> did not run in parallel).
<code>clusterType</code>	The type of the cluster used to run <code>VSURF_thres</code> in parallel (NULL if <code>VSURF_thres</code> did not run in parallel).
<code>call</code>	The original call to <code>VSURF</code> .
<code>terms</code>	Terms associated to the formula (only if formula-type call was used).

Author(s)

Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

References

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters 31(14), 2225-2236

Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015), *VSURF: An R Package for Variable Selection Using Random Forests*, The R Journal 7(2):19-33

See Also

`VSURF`, `tune`

Examples

```
data(iris)
iris.thres <- VSURF_thres(iris[,1:4], iris[,5], ntree = 100, nfor.thres = 20)
iris.thres

## Not run:
# A more interesting example with toys data (see \code{\link{toys}})
# (a few minutes to execute)
data(toys)
toys.thres <- VSURF_thres(toys$x, toys$y)
toys.thres
## End(Not run)
```