

Package ‘IRSF’

November 13, 2020

Type Package

Title Interaction Random Survival Forest

Version 1.0.3

Date 2020-11-09

Author Jean-Eudes Dazard [aut, cre]

Maintainer Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Description Builds ensemble survival tree models to reveal variable interactions when the response is a time-to-events outcome. Codes contain randomization, interaction modeling, and prediction subroutines to be used in addition to the following R packages: 'survival' for Kaplan-Meier and Cox regression modeling, 'randomForestSRC' for RSF modeling, and optionally 'ggRandomForests' for Random Forest exploration/visualization. The current version contains additional R codes in folder ``/inst/doc'' for the analysis and generation of results shown in the corresponding article (Dazard et al. (2018) <doi:10.1515/sagmb-2017-0038>).

Depends R (>= 3.5.0)

Imports parallel, survival, randomForestSRC, abind

NeedsCompilation no

URL <https://github.com/jedazard/IRSF>

Repository CRAN

Date/Publication 2020-11-13 11:20:02 UTC

License GPL (>= 3)

Archs i386, x64

R topics documented:

IRSF-package	2
cph.int	3
cph.main	7
IRSF.news	11
MACS	12
rsf.int	15
rsf.int.signif	21

rsf.main	22
rsf.main.signif	29

Index	31
--------------	-----------

IRSF-package	<i>Interaction Random Survival Forest</i>
--------------	---

Description

Builds ensemble survival tree models to reveal variable interactions when the response is a time-to-events outcome. Codes contain randomization, interaction modeling, and prediction subroutines to be used in addition to the following R packages: **survival** for Kaplan-Meier and Cox regression modeling, **randomForestSRC** (Ishwaran and Kogalur, 2013, 2007) for RSF modeling, and optionally **ggRandomForests** (Ehrlinger, 2014) for Random Forest exploration/visualization. The current version contains additional R codes in folder "/inst/doc" for the analysis and generation of results shown in the corresponding article (Dazard et al., 2018).

Details

Manuscript Abstract: Unraveling interactions among variables such as genetic, clinical, demographic and environmental factors is essential to understand the development of common and complex diseases. To increase the power to detect such variables interactions associated with clinical time-to-events outcomes, we borrowed established concepts from Random Survival Forest (RSF) models. We introduce a novel RSF-based pairwise interaction estimator and derive a randomization method with bootstrap confidence intervals for inferring interaction significance. Using various linear and non-linear time-to-events survival models in simulation studies, we first show the efficiency of our approach: true pairwise interaction-effects between variables are thus uncovered, while they may not be accompanied with their corresponding main-effects and often not detected by standard semi-parametric Cox regression. Moreover, using a RSF-based cross-validation scheme for generating prediction estimators, we show that informative predictors may thus be inferred. We illustrate the application of our approach in an HIV cohort study recording key host gene polymorphisms and their association with HIV change of tropism or AIDS progression. Altogether, this shows how linear or non-linear pairwise statistical interactions between variables may be uncovered in clinical studies with time-to-event outcomes of interest when the motivation is to discover important variables interactions with a predictive clinical value.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/mac/mac.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from

the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.
- Ehrlinger, J. (2014). "*Contributed R Package ggRandomForests: Visually Exploring Random Forests*" CRAN.

See Also

- Multicenter AIDS Cohort Study (MACS) website at <https://statepi.jhsph.edu/macs/macs.html>
- R package **randomForestSRC**

cph.int

Pairwise interaction effects in a Cox-PH model

Description

Fits a Proportional Hazards Time-To-Event Regression Model saturated with first and second order terms. Computes p-values of significance of regression coefficients of pairwise interaction effects in a Cox-PH model.

Usage

```
cph.int(X,  
        int.term)
```

Arguments

<code>X</code>	data.frame or numeric matrix of input covariates. Dataset <code>X</code> assumes that: - all variables are in columns - the observed times to event and censoring variables are in the first two columns: <code>"stime"</code> : numeric vector of observed times. <code>"status"</code> : numeric vector of observed status (censoring) indicator variable. - each variable has a unique name, excluding the word "noise"
<code>int.term</code>	vector of character string of all possible pairs of covariates names, separated by ":".

Value

list of 2 fields:

<code>raw</code>	Raw p-value of covariates pairwise interaction statistics significance
<code>fdr</code>	FDR-adjusted p-value of covariates pairwise interaction statistics significance

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/macs/macs.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "Random Survival Forests for R". *R News*, 7(2):25-31.

- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- R package **randomForestSRC**

Examples

```
#####
# Loading the library and its dependencies
#####
library("IRSF")

## Not run:
#####
# IRSF package news
#####
IRSF.news()

#####
# MVR package citation
#####
citation("IRSF")

#####
# Loading of the Synthetic and Real datasets
# Use help for descriptions
#####
data("MACS", package="IRSF")
?MACS

head(MACS)

#####
# Synthetic dataset
# Continuous case:
# All variables  $x_j$ ,  $j$  in  $\{1, \dots, p\}$ , are iid
# from a multivariate uniform distribution
# with parameters  $a=1$ ,  $b=5$ , i.e. on  $[1, 5]$ .
#  $\rho = 0.50$ 
# Regression model:  $X_1 + X_2 + X_1X_2$ 
#####
seed <- 1234567
set.seed(seed)
n <- 200
p <- 5
x <- matrix(data=runif(n=n*p, min=1, max=5),
            nrow=n, ncol=p, byrow=FALSE,
            dimnames=list(1:n, paste("X", 1:p, sep="")))

beta <- c(rep(1,2), rep(0,p-2), 1)
covar <- cbind(x, "X1X2"=x[,1]*x[,2])
```

```

eta <- covar %*% beta                # regression function

seed <- 1234567
set.seed(seed)
lambda0 <- 1
lambda <- lambda0 * exp(eta - mean(eta)) # hazards function
tt <- rexp(n=n, rate=lambda)           # true (uncensored) event times
tc <- runif(n=n, min=0, max=3.9)       # true (censored) event times
stime <- pmin(tt, tc)                 # observed event times
status <- 1 * (tt <= tc)               # observed event indicator
X <- data.frame(stime, status, x)

#=====
# Synthetic dataset
# Ranking of pairwise interactions between individual
# or noise variables by bivariate
# Interaction Minimal Depth of a Maximal Subtree (IMDMS)
# Serial mode
#=====
X.int.mdms <- rsf.int(X=X,
                    ntree=1000,
                    method="imdms",
                    splitrule="logrank",
                    importance="random",
                    B=1000,
                    ci=90,
                    parallel=FALSE,
                    conf=NULL,
                    verbose=FALSE,
                    seed=seed)

#=====
# Synthetic dataset
# Proportional Hazards Time-To-Event Regression Model
# saturated with first and second order terms.
#=====
X.int.cph <- cph.int(X=X,
                    int.term=rownames(X.int.mdms))

#=====
# Real dataset
#=====
seed <- 1234567
data("MACS", package="IRSF")

X <- MACS[,c("TTX", "EventX", "Race", "Group3",
            "DEFB.CNV3", "CCR2.SNP", "CCR5.SNP2",
            "CCR5.ORF", "CXCL12.SNP2")]

#=====
# Real dataset
# Ranking of pairwise interactions between individual
# or noise variables by bivariate

```

```

# Interaction Minimal Depth of a Maximal Subtree (IMDMS)
# Entries [i][j] indicate the normalized minimal depth
# of a variable [j] w.r.t. the maximal subtree for variable [i]
# (normalized w.r.t. the size of [i]'s maximal subtree).
#=====
MACS.int.mdms <- rsf.int(X=X,
                        ntree=1000,
                        method="imdms",
                        splitrule="logrank",
                        importance="random",
                        B=1000,
                        ci=80,
                        parallel=TRUE,
                        conf=conf,
                        verbose=TRUE,
                        seed=seed)

#=====
# Real dataset
# Proportional Hazards Time-To-Event Regression Model
# saturated with first and second order terms.
#=====
MACS.int.cph <- cph.int(X=X,
                       int.term=rownames(MACS.int.mdms))

## End(Not run)

```

cph.main

Main effects in a Cox-PH model

Description

Fits a Proportional Hazards Time-To-Event Regression Model saturated with first order terms. Computes p-values of significance of regression coefficients of main effects in a Cox-PH model

Usage

```
cph.main(X,
         main.term)
```

Arguments

X	data.frame or numeric matrix of input covariates. Dataset X assumes that: - all variables are in columns - the observed times to event and censoring variables are in the first two columns: "stime": numeric vector of observed times. "status": numeric vector of observed status (censoring) indicator variable. - each variable has a unique name, excluding the word "noise"
main.term	Vector of character string of each individual covariate name.

Value

List of 2 fields:

raw	Raw p-value covariates importances significance
fdr	FDR-adjusted p-value of covariates importances significance

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/macs/macs.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- R package **randomForestSRC**

Examples

```

#####
# Loading the library and its dependencies
#####
library("IRSF")

## Not run:
#####
# IRSF package news
#####
IRSF.news()

#####
# MVR package citation
#####
citation("IRSF")

#####
# Loading of the Synthetic and Real datasets
# Use help for descriptions
#####
data("MACS", package="IRSF")
?MACS

head(MACS)

#####
# Synthetic dataset
# Continuous case:
# All variables  $x_j$ ,  $j$  in  $\{1, \dots, p\}$ , are iid
# from a multivariate uniform distribution
# with parameters  $a=1$ ,  $b=5$ , i.e. on  $[1, 5]$ .
#  $\rho = 0.50$ 
# Regression model:  $X_1 + X_2 + X_1X_2$ 
#####
seed <- 1234567
set.seed(seed)
n <- 200
p <- 5
x <- matrix(data=runif(n=n*p, min=1, max=5),
            nrow=n, ncol=p, byrow=FALSE,
            dimnames=list(1:n, paste("X", 1:p, sep="")))

beta <- c(rep(1,2), rep(0,p-2), 1)
covar <- cbind(x, "X1X2"=x[,1]*x[,2])
eta <- covar %*% beta # regression function

seed <- 1234567
set.seed(seed)
lambda0 <- 1
lambda <- lambda0 * exp(eta - mean(eta)) # hazards function
tt <- rexp(n=n, rate=lambda) # true (uncensored) event times

```

```

tc <- runif(n=n, min=0, max=3.9)           # true (censored) event times
stime <- pmin(tt, tc)                     # observed event times
status <- 1 * (tt <= tc)                  # observed event indicator
X <- data.frame(stime, status, x)

#####
# Synthetic dataset
# Ranking of individual and noise variables by univariate
# Minimal Depth of a Maximal Subtree (MDMS)
# Serial mode
#####
X.main.mdms <- rsf.main(X=X,
                       ntree=1000,
                       method="mdms",
                       splitrule="logrank",
                       importance="random",
                       B=1000,
                       ci=90,
                       parallel=FALSE,
                       conf=NULL,
                       verbose=TRUE,
                       seed=seed)

#####
# Synthetic dataset
# Proportional Hazards Time-To-Event Regression Model
# saturated with first order terms only.
#####
X.main.cph <- cph.main(X=X,
                      main.term=rownames(X.main.mdms))

#####
# Real dataset
#####
seed <- 1234567
data("MACS", package="IRSF")

X <- MACS[,c("TTX", "EventX", "Race", "Group3",
            "DEFB.CNV3", "CCR2.SNP", "CCR5.SNP2",
            "CCR5.ORF", "CXCL12.SNP2")]

#####
# Real dataset
# Ranking of individual and noise variables by univariate
# Minimal Depth of a Maximal Subtree (MDMS)
# Parallel mode
#####
MACS.main.mdms <- rsf.main(X=X,
                          ntree=1000,
                          method="mdms",
                          splitrule="logrank",
                          importance="random",
                          B=1000,

```

```

                                ci=80,
                                parallel=TRUE,
                                conf=conf,
                                verbose=TRUE,
                                seed=seed)

#####
# Real dataset
# Proportional Hazards Time-To-Event Regression Model
# saturated with first order terms only.
#####
MACS.main.cph <- cph.main(X=X,
                          main.term=rownames(MACS.main.mdms))

## End(Not run)

```

IRSF.news

*Display the **IRSF** Package News*

Description

Function to display the log file NEWS of updates of the **IRSF** package.

Usage

```
IRSF.news(...)
```

Arguments

... Further arguments passed to or from other methods.

Value

None.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/mac/macs.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and

the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Note

End-user function.

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

MACS

Real Dataset: MACS Cohort Study

Description

Publicly available dataset from the Multicenter AIDS Cohort Study (MACS) available at (<http://www.statepi.jhsph.edu/macsv/>). The dataset provides longitudinal account of viral tropism in relation to the HIV full spectrum of rates of HIV-1 disease progression (Shepherd, et al. 2008). To our knowledge, this cohort provides a unique dataset with well characterized clinical information for analyzing associations between host genetic variation and viral tropism as well as disease progression. Here, we determined whether copy number variation in beta-defensin and its interactions with certain polymorphisms in chemokine receptors and ligand genes are associated, either alone or jointly, with clinical events in HIV-seropositive patients, such as time to HIV change of tropism or time to AIDS diagnosis (See Dazard et al. (2017) for additional descriptions of the dataset and materials).

Usage

```
data("MACS", package="IRSF")
```

Format

The dataset consists of a numeric data frame containing $n = 50$ complete observations (samples) by rows and $p = 7$ covariates by columns, not including the censoring indicator and (censored) time-to-event variables.

The variables included in the MACS cohort study were 5 genetic variants (DEFB4/103A CNV [1-5], CCR2 SNP [190G>A], CCR5 [SNP -2459G>A, ORF], CXCL12 SNP [801G>A]) and 2 non-genetic variables, taken as two additional covariates. All input variables were categorical with no more than three levels (experimental groups) each. We used genetic variables with original and aggregated categories as follows: DEFB CNV [CNV = 2 or CNV > 2]; CCR2 SNP [GG or GA], CCR5 SNP [GG or GA]; CCR5 ORF [WT or D32], CXCL12 SNP [GG or GA]. The first covariate was the two-level disease progression Group variable [Fast, Slow], and the second was the three-level Race/Ethnicity variable [White, Hispanic, Black]. For each observation $i \in \{1, \dots, n\}$, we denote the j -th variable by the n -dimensional vector $\mathbf{x}_j = (x_{1,j}, \dots, x_{n,j})^T$, where $j \in \{1, \dots, p\}$. Here, p denotes the number of variables. Hereafter, we denoted the $p = 7$ included variables as follows:

- \mathbf{x}_1 =DEFB CNV
- \mathbf{x}_2 =CCR2 SNP
- \mathbf{x}_3 =CCR5 SNP
- \mathbf{x}_4 =CCR5 ORF
- \mathbf{x}_5 =CXCL12 SNP
- \mathbf{x}_6 =Group
- \mathbf{x}_7 =Race

The time-to-event outcomes included in the MACS cohort study, generically denoted E , were the time-to-X4-Emergence (denoted XE) and the time-to-AIDS-Diagnosis (denoted AD), whether each was observed or not during each patient's follow-up time. The corresponding event-free (EF) ("survival") probability function $S(t)$ of time-to-event $E := XE$ (X4-Emergence) or $E := AD$ (AIDS-Diagnosis), were called X4-Emergence-Free ($E := XEF$) or AIDS-Diagnosis-Free ($E := ADF$) probability.

The dataset comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/macs/mac.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants

from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Source

See real data application in Dazard et al., 2017.

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- Multicenter AIDS Cohort Study (MACS) website at <https://statepi.jhsph.edu/macs/macs.html>

Examples

```
#####
# Loading the library and its dependencies
#####
library("IRSF")

#####
# Help on MACS dataset
#####
data("MACS", package="IRSF")
?MACS
```

rsf.int

*Bivariate Interaction Minimal Depth of a Maximal Subtree (IMDMS)***Description**

Ranking of pairwise interactions between individual or noise variables by bivariate interaction Minimal Depth of a Maximal Subtree (IMDMS)

Usage

```
rsf.int(X,
        ntree = 1000,
        method = "imdms",
        splitrule = "logrank",
        importance = "random",
        B = 1000,
        ci = 90,
        parallel = FALSE,
        conf = NULL,
        verbose = TRUE,
        seed = NULL)
```

Arguments

X	data.frame or numeric matrix of input covariates. Dataset X assumes that: - all variables are in columns - the observed times to event and censoring variables are in the first two columns: "stime": numeric vector of observed times. "status": numeric vector of observed status (censoring) indicator variable. - each variable has a unique name, excluding the word "noise"
ntree	Number of trees in the forest. Defaults to 1000.
method	Method for ranking of interactions between pairs of individual and noise variables. character string "imdms" (default) that stands for interaction minimal depth of a maximal subtree (IMDMS).
splitrule	Splitting rule used to grow trees. For time-to-event analysis, use "logrank" (default), which implements log-rank splitting (Segal, 1988; Leblanc and Crowley, 1993).
importance	Method for computing variable importance. Defaults to Character string "random". See details below.
B	Positive integer of the number of replications of the cross-validation procedure. Defaults to 1000.
ci	Confidence Interval for inferences of individual and noise variables. numeric scalar between 50 and 100. Defaults to 90.
parallel	logical. Is parallel computing to be performed? Defaults to FALSE.

conf	<p>list of 5 fields containing the parameters values needed for creating the parallel backend (cluster configuration). See details below for usage. Optional, defaults to NULL, but all fields are required if used:</p> <ul style="list-style-type: none"> • type : character vector specifying the cluster type ("SOCKET", "MPI"). • spec : A specification (character vector or integer scalar) appropriate to the type of cluster. • homogeneous : logical scalar to be set to FALSE for inhomogeneous clusters. • verbose : logical scalar to be set to FALSE for quiet mode. • outfile : character vector of an output log file name to direct the stdout and stderr connection output from the workernodes. "" indicates no redirection.
verbose	logical scalar. Is the output to be verbose? Optional, defaults to TRUE.
seed	Positive integer scalar of the user seed to reproduce the results. Defaults to NULL.

Details

The option `importance` allows several ways to calculate Variable Importance (VIMP). The default `"permute"` returns Breiman-Cutler permutation VIMP as described in Breiman (2001). For each tree, the prediction error on the out-of-bag (OOB) data is recorded. Then for a given variable x , OOB cases are randomly permuted in x and the prediction error is recorded. The VIMP for x is defined as the difference between the perturbed and unperturbed error rate, averaged over all trees. If `"random"` is used, then x is not permuted, but rather an OOB case is assigned a daughter node randomly whenever a split on x is encountered in the in-bag tree. If `"anti"` is used, then x is assigned to the opposite node whenever a split on x is encountered in the in-bag tree.

The function `rsf.int` relies on the R package **parallel** to create a parallel backend within an R session, enabling access to a cluster of compute cores and/or nodes on a local and/or remote machine(s) and scaling-up with the number of CPU cores available and efficient parallel execution. To run a procedure in parallel (with parallel RNG), argument `parallel` is to be set to TRUE and argument `conf` is to be specified (i.e. non NULL). Argument `conf` uses the options described in function `makeCluster` of the R packages **parallel** and **snow**. **IRSF** supports two types of communication mechanisms between master and worker processes: 'Socket' or 'Message-Passing Interface' ('MPI'). In **IRSF**, parallel 'Socket' clusters use sockets communication mechanisms only (no forking) and are therefore available on all platforms, including Windows, while parallel 'MPI' clusters use high-speed interconnects mechanism in networks of computers (with distributed memory) and are therefore available only in these architectures. A parallel 'MPI' cluster also requires R package **Rmpi** to be installed. Value `type` is used to setup a cluster of type 'Socket' ("SOCKET") or 'MPI' ("MPI"), respectively. Depending on this type, values of `spec` are to be used alternatively:

- For 'Socket' clusters (`conf$type="SOCKET"`), `spec` should be a character vector naming the hosts on which to run the job; it can default to a unique local machine, in which case, one may use the unique host name "localhost". Each host name can potentially be repeated to the number of CPU cores available on the local machine. It can also be an integer scalar specifying the number of processes to spawn on the local machine; or a list of machine specifications if you have `ssh` installed (a character value named `host` specifying the name or address of the host to use).

- For 'MPI' clusters (`conf$type="MPI"`), `spec` should be an integer scalar specifying the total number of processes to be spawned across the network of available nodes, counting the workernodes and masternode.

The actual creation of the cluster, its initialization, and closing are all done internally. For more details, see the reference manual of R package **snow** and examples below.

When random number generation is needed, the creation of separate streams of parallel RNG per node is done internally by distributing the stream states to the nodes. For more details, see the vignette of R package **parallel**. The use of a seed allows to reproduce the results within the same type of session: the same seed will reproduce the same results within a non-parallel session or within a parallel session, but it will not necessarily give the exact same results (up to sampling variability) between a non-parallelized and parallelized session due to the difference of management of the seed between the two (see parallel RNG and value of returned seed below).

Value

data.frame containing the following columns:

- "obs.mean" observed mean of covariates pairwise interaction statistics
- "obs.se" observed standard error of covariates pairwise interaction statistics
- "obs.LBCI" observed Lower Bound Confidence Interval of covariates pairwise interaction statistics
- "obs.UBCI" observed Upper Bound Confidence Interval of covariates pairwise interaction statistics
- "noise.mean" observed mean of noise covariates statistics
- "noise.se" observed standard error of noise covariates pairwise interaction statistics
- "noise.LBCI" observed Lower Bound Confidence Interval of noise covariates pairwise interaction statistics
- "noise.UBCI" observed Upper Bound Confidence Interval of noise covariates pairwise interaction statistics
- "signif.ISE" calls of covariates pairwise interaction statistics significance using the ISE rule
- "signif.CI" calls of covariates pairwise interaction statistics significance using the CI rule at `ci%` confidence level

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/macs/macs.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and

the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- R package **randomForestSRC**

Examples

```
#####
# Loading the library and its dependencies
#####
library("IRSF")

## Not run:
#####
# IRSF package news
#####
IRSF.news()

#####
# MVR package citation
#####
citation("IRSF")

#####
# Loading of the Synthetic and Real datasets
# Use help for descriptions
#####
data("MACS", package="IRSF")
```

```

?MACS

head(MACS)

#=====
# Synthetic dataset
# Continuous case:
# All variables  $x_j$ ,  $j$  in  $\{1, \dots, p\}$ , are iid
# from a multivariate uniform distribution
# with parameters  $a=1$ ,  $b=5$ , i.e. on  $[1, 5]$ .
#  $\rho = 0.50$ 
# Regression model:  $X_1 + X_2 + X_1X_2$ 
#=====
seed <- 1234567
set.seed(seed)
n <- 200
p <- 5
x <- matrix(data=runif(n=n*p, min=1, max=5),
            nrow=n, ncol=p, byrow=FALSE,
            dimnames=list(1:n, paste("X", 1:p, sep="")))

beta <- c(rep(1,2), rep(0,p-2), 1)
covar <- cbind(x, "X1X2"=x[,1]*x[,2])
eta <- covar %*% beta # regression function

seed <- 1234567
set.seed(seed)
lambda0 <- 1
lambda <- lambda0 * exp(eta - mean(eta)) # hazards function
tt <- rexp(n=n, rate=lambda) # true (uncensored) event times
tc <- runif(n=n, min=0, max=3.9) # true (censored) event times
stime <- pmin(tt, tc) # observed event times
status <- 1 * (tt <= tc) # observed event indicator
X <- data.frame(stime, status, x)

#=====
# Synthetic dataset
# Ranking of pairwise interactions between individual
# or noise variables by bivariate
# Interaction Minimal Depth of a Maximal Subtree (IMDMS)
# Serial mode
#=====
X.int.mdms <- rsf.int(X=X,
                    ntree=1000,
                    method="imdms",
                    splitrule="logrank",
                    importance="random",
                    B=1000,
                    ci=90,
                    parallel=FALSE,
                    conf=NULL,
                    verbose=FALSE,
                    seed=seed)

```

```

#=====
# Examples of parallel backend parametrization
#=====
if (require("parallel")) {
  cat("'parallel' is attached correctly \n")
} else {
  stop("'parallel' must be attached first \n")
}
#=====
# Ex. #1 - Multicore PC
# Running WINDOWS
# SOCKET communication cluster
# Shared memory parallelization
#=====
cpus <- parallel::detectCores(logical = TRUE)
conf <- list("spec" = rep("localhost", cpus),
            "type" = "SOCKET",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")
#=====
# Ex. #2 - Master node + 3 Worker nodes cluster
# All nodes equipped with identical setups of multicores
# (8 core CPUs per machine for a total of 32)
# SOCKET communication cluster
# Distributed memory parallelization
#=====
masterhost <- Sys.getenv("HOSTNAME")
slavehosts <- c("compute-0-0", "compute-0-1", "compute-0-2")
nodes <- length(slavehosts) + 1
cpus <- 8
conf <- list("spec" = c(rep(masterhost, cpus),
                        rep(slavehosts, cpus)),
            "type" = "SOCKET",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")
#=====
# Ex. #3 - Enterprise Multinode Cluster w/ multicore/node
# Running LINUX with SLURM scheduler
# MPI communication cluster
# Distributed memory parallelization
# Below, variable 'cpus' is the total number of requested
# taks (threads/CPUs), which is specified from within a
# SLURM script.
#=====
if (require("Rmpi")) {
  print("Rmpi is loaded correctly \n")
} else {
  stop("Rmpi must be installed first to use MPI\n")
}
cpus <- as.numeric(Sys.getenv("SLURM_NTASKS"))

```

```

conf <- list("spec" = cpus,
            "type" = "MPI",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")

#####
# Real dataset
#####
seed <- 1234567
data("MACS", package="IRSF")

X <- MACS[,c("TTX", "EventX", "Race", "Group3",
            "DEFB.CNV3", "CCR2.SNP", "CCR5.SNP2",
            "CCR5.ORF", "CXCL12.SNP2")]

#####
# Real dataset
# Ranking of pairwise interactions between individual
# or noise variables by bivariate
# Interaction Minimal Depth of a Maximal Subtree (IMDMS)
# Entries [i][j] indicate the normalized minimal depth
# of a variable [j] w.r.t. the maximal subtree for variable [i]
# (normalized w.r.t. the size of [i]'s maximal subtree).
#####
MACS.int.mdms <- rsf.int(X=X,
                       ntree=1000,
                       method="imdms",
                       splitrule="logrank",
                       importance="random",
                       B=1000,
                       ci=80,
                       parallel=TRUE,
                       conf=conf,
                       verbose=TRUE,
                       seed=seed)

## End(Not run)

```

rsf.int.signif

Subroutine of rsf.int function

Description

Subroutine for ranking of pairwise interactions between individual or noise variables by bivariate interaction Minimal Depth of a Maximal Subtree (IMDMS)

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/macs/macs.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- R package **randomForestSRC**

rsf.main

Univariate Minimal Depth of a Maximal Subtree (MDMS)

Description

Ranking of individual and noise variables main effects by univariate Minimal Depth of a Maximal Subtree (MDMS)

Usage

```
rsf.main(X,
         ntree = 1000,
         method = "mdms",
         splitrule = "logrank",
         importance = "random",
         B = 1000,
         ci = 90,
         parallel = FALSE,
         conf = NULL,
         verbose = TRUE,
         seed = NULL)
```

Arguments

X	data.frame or numeric matrix of input covariates. Dataset X assumes that: - all variables are in columns - the observed times to event and censoring variables are in the first two columns: "stime": numeric vector of observed times. "status": numeric vector of observed status (censoring) indicator variable. - each variable has a unique name, excluding the word "noise"
ntree	Number of trees in the forest. Defaults to 1000.
method	Method for ranking of individual and noise variables. character string "mdms" (default) that stands for Univariate Minimal Depth of a Maximal Subtree (MDMS).
splitrule	Splitting rule used to grow trees. For time-to-event analysis, use "logrank" (default), which implements log-rank splitting (Segal, 1988; Leblanc and Crowley, 1993).
importance	Method for computing variable importance. Defaults to Character string "random". See details below.
B	Positive integer of the number of replications of the cross-validation procedure. Defaults to 1000.
ci	Confidence Interval for inferences of individual and noise variables. numeric scalar between 50 and 100. Defaults to 90.
parallel	logical. Is parallel computing to be performed? Defaults to FALSE.
conf	list of 5 fields containing the parameters values needed for creating the parallel backend (cluster configuration). See details below for usage. Optional, defaults to NULL, but all fields are required if used: <ul style="list-style-type: none"> • type : character vector specifying the cluster type ("SOCKET", "MPI"). • spec : A specification (character vector or integer scalar) appropriate to the type of cluster. • homogeneous : logical scalar to be set to FALSE for inhomogeneous clusters. • verbose : logical scalar to be set to FALSE for quiet mode. • outfile : character vector of an output log file name to direct the stdout and stderr connection output from the workernodes. "" indicates no redirection.

verbose	logical scalar. Is the output to be verbose? Optional, defaults to TRUE.
seed	Positive integer scalar of the user seed to reproduce the results. Defaults to NULL.

Details

The option `importance` allows several ways to calculate Variable Importance (VIMP). The default `"permute"` returns Breiman-Cutler permutation VIMP as described in Breiman (2001). For each tree, the prediction error on the out-of-bag (OOB) data is recorded. Then for a given variable `x`, OOB cases are randomly permuted in `x` and the prediction error is recorded. The VIMP for `x` is defined as the difference between the perturbed and unperturbed error rate, averaged over all trees. If `"random"` is used, then `x` is not permuted, but rather an OOB case is assigned a daughter node randomly whenever a split on `x` is encountered in the in-bag tree. If `"anti"` is used, then `x` is assigned to the opposite node whenever a split on `x` is encountered in the in-bag tree.

The function `rsf.main` relies on the R package **parallel** to create a parallel backend within an R session, enabling access to a cluster of compute cores and/or nodes on a local and/or remote machine(s) and scaling-up with the number of CPU cores available and efficient parallel execution. To run a procedure in parallel (with parallel RNG), argument `parallel` is to be set to TRUE and argument `conf` is to be specified (i.e. non NULL). Argument `conf` uses the options described in function `makeCluster` of the R packages **parallel** and **snow**. **IRSF** supports two types of communication mechanisms between master and worker processes: `'Socket'` or `'Message-Passing Interface'` (`'MPI'`). In **IRSF**, parallel `'Socket'` clusters use sockets communication mechanisms only (no forking) and are therefore available on all platforms, including Windows, while parallel `'MPI'` clusters use high-speed interconnects mechanism in networks of computers (with distributed memory) and are therefore available only in these architectures. A parallel `'MPI'` cluster also requires R package **Rmpi** to be installed. Value type is used to setup a cluster of type `'Socket'` (`"SOCKET"`) or `'MPI'` (`"MPI"`), respectively. Depending on this type, values of `spec` are to be used alternatively:

- For `'Socket'` clusters (`conf$type="SOCKET"`), `spec` should be a character vector naming the hosts on which to run the job; it can default to a unique local machine, in which case, one may use the unique host name `"localhost"`. Each host name can potentially be repeated to the number of CPU cores available on the local machine. It can also be an integer scalar specifying the number of processes to spawn on the local machine; or a list of machine specifications if you have `ssh` installed (a character value named `host` specifying the name or address of the host to use).
- For `'MPI'` clusters (`conf$type="MPI"`), `spec` should be an integer scalar specifying the total number of processes to be spawned across the network of available nodes, counting the `workernodes` and `masternode`.

The actual creation of the cluster, its initialization, and closing are all done internally. For more details, see the reference manual of R package **snow** and examples below.

When random number generation is needed, the creation of separate streams of parallel RNG per node is done internally by distributing the stream states to the nodes. For more details, see the vignette of R package **parallel**. The use of a seed allows to reproduce the results within the same type of session: the same seed will reproduce the same results within a non-parallel session or within a parallel session, but it will not necessarily give the exact same results (up to sampling variability) between a non-parallelized and parallelized session due to the difference of management of the seed between the two (see parallel RNG and value of returned seed below).

Value

data.frame containing the following columns:

- "obs.mean" observed mean of covariates importances
- "obs.se" observed standard error of covariates importances
- "obs.LBCI" observed Lower Bound Confidence Interval of covariates importances
- "obs.UBCI" observed Upper Bound Confidence Interval of covariates importances
- "noise.mean" observed mean of noise covariates importances
- "noise.se" observed standard error of noise covariates ranks
- "noise.LBCI" observed Lower Bound Confidence Interval of noise covariates importances
- "noise.UBCI" observed Upper Bound Confidence Interval of noise covariates importances
- "signif.1SE" calls of covariates importances significance using the 1SE rule
- "signif.CI" calls of covariates importances significance using the CI rule at ci% confidence level

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/mac/mac.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.

- Ishwaran, H. and Kogalur, U.B. (2007). "Random Survival Forests for R". R News, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "Contributed R Package **randomForestSRC**: Random Forests for Survival, Regression and Classification (RF-SRC)" CRAN.

See Also

- R package **randomForestSRC**

Examples

```
#####
# Loading the library and its dependencies
#####
library("IRSF")

## Not run:
#####
# IRSF package news
#####
IRSF.news()

#####
# MVR package citation
#####
citation("IRSF")

#####
# Loading of the Synthetic and Real datasets
# Use help for descriptions
#####
data("MACS", package="IRSF")
?MACS

head(MACS)

#####
# Synthetic dataset
# Continuous case:
# All variables  $x_j$ ,  $j$  in  $\{1, \dots, p\}$ , are iid
# from a multivariate uniform distribution
# with parameters  $a=1$ ,  $b=5$ , i.e. on  $[1, 5]$ .
#  $\rho = 0.50$ 
# Regression model:  $X_1 + X_2 + X_1X_2$ 
#####
seed <- 1234567
set.seed(seed)
n <- 200
p <- 5
x <- matrix(data=runif(n=n*p, min=1, max=5),
            nrow=n, ncol=p, byrow=FALSE,
            dimnames=list(1:n, paste("X", 1:p, sep="")))
```

```

beta <- c(rep(1,2), rep(0,p-2), 1)
covar <- cbind(x, "X1X2"=x[,1]*x[,2])
eta <- covar %*% beta # regression function

seed <- 1234567
set.seed(seed)
lambda0 <- 1
lambda <- lambda0 * exp(eta - mean(eta)) # hazards function
tt <- rexp(n=n, rate=lambda) # true (uncensored) event times
tc <- runif(n=n, min=0, max=3.9) # true (censored) event times
stime <- pmin(tt, tc) # observed event times
status <- 1 * (tt <= tc) # observed event indicator
X <- data.frame(stime, status, x)

#=====
# Synthetic dataset
# Ranking of individual and noise variables by univariate
# Minimal Depth of a Maximal Subtree (MDMS)
# Serial mode
#=====
X.main.mdms <- rsf.main(X=X,
                      ntree=1000,
                      method="mdms",
                      splitrule="logrank",
                      importance="random",
                      B=1000,
                      ci=90,
                      parallel=FALSE,
                      conf=NULL,
                      verbose=TRUE,
                      seed=seed)

#=====
# Examples of parallel backend parametrization
#=====
if (require("parallel")) {
  cat("'parallel' is attached correctly \n")
} else {
  stop("'parallel' must be attached first \n")
}
#=====
# Ex. #1 - Multicore PC
# Running WINDOWS
# SOCKET communication cluster
# Shared memory parallelization
#=====
cpus <- parallel::detectCores(logical = TRUE)
conf <- list("spec" = rep("localhost", cpus),
            "type" = "SOCKET",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")

```

```

#####
# Ex. #2 - Master node + 3 Worker nodes cluster
# All nodes equipped with identical setups of multicores
# (8 core CPUs per machine for a total of 32)
# SOCKET communication cluster
# Distributed memory parallelization
#####
masterhost <- Sys.getenv("HOSTNAME")
slavehosts <- c("compute-0-0", "compute-0-1", "compute-0-2")
nodes <- length(slavehosts) + 1
cpus <- 8
conf <- list("spec" = c(rep(masterhost, cpus),
                        rep(slavehosts, cpus)),
            "type" = "SOCKET",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")

#####
# Ex. #3 - Enterprise Multinode Cluster w/ multicore/node
# Running LINUX with SLURM scheduler
# MPI communication cluster
# Distributed memory parallelization
# Below, variable 'cpus' is the total number of requested
# taks (threads/CPUs), which is specified from within a
# SLURM script.
#####
if (require("Rmpi")) {
  print("Rmpi is loaded correctly \n")
} else {
  stop("Rmpi must be installed first to use MPI\n")
}
cpus <- as.numeric(Sys.getenv("SLURM_NTASKS"))
conf <- list("spec" = cpus,
            "type" = "MPI",
            "homo" = TRUE,
            "verbose" = TRUE,
            "outfile" = "")

#####
# Real dataset
#####
seed <- 1234567
data("MACS", package="IRSF")

X <- MACS[,c("TTX", "EventX", "Race", "Group3",
            "DEFB.CNV3", "CCR2.SNP", "CCR5.SNP2",
            "CCR5.ORF", "CXCL12.SNP2")]

#####
# Real dataset
# Ranking of individual and noise variables by univariate
# Minimal Depth of a Maximal Subtree (MDMS)
# Parallel mode

```

```

#####
MACS.main.mdms <- rsf.main(X=X,
                          ntree=1000,
                          method="mdms",
                          splitrule="logrank",
                          importance="random",
                          B=1000,
                          ci=80,
                          parallel=TRUE,
                          conf=conf,
                          verbose=TRUE,
                          seed=seed)

## End(Not run)

```

rsf.main.signif *Subroutine of rsf.main function*

Description

Subroutine for ranking of individual and noise variables main effects by univariate Minimal Depth of a Maximal Subtree (MDMS)

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. We are thankful to Ms. Janet Schollenberger, Senior Project Coordinator, CAMACS, as well as Dr. Jeremy J. Martinson, Sudhir Penugonda, Shehnaz K. Hussain, Jay H. Bream, and Priya Duggal, for providing us the data related to the samples analyzed in the present study. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) at (<https://www.statepi.jhsph.edu/mac/mac.html>) with centers at Baltimore, Chicago, Los Angeles, Pittsburgh, and the Data Coordinating Center: The Johns Hopkins University Bloomberg School of Public Health. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by Johns Hopkins University CTSA. This study was supported by two grants from the National Institute of Health: NIDCR P01DE019759 (Aaron Weinberg, Peter Zimmerman, Richard J. Jurevic, Mark Chance) and NCI R01CA163739 (Hemant Ishwaran). The work was also partly supported by the National Science Foundation grant DMS 1148991 (Hemant Ishwaran) and the Center for AIDS Research grant P30AI036219 (Mark Chance).

Author(s)

Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Maintainer: Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

References

- Dazard J-E., Ishwaran H., Mehlotra R.K., Weinberg A. and Zimmerman P.A. (2018). "*Ensemble Survival Tree Models to Reveal Pairwise Interactions of Variables with Time-to-Events Outcomes in Low-Dimensional Setting*" *Statistical Applications in Genetics and Molecular Biology*, 17(1):20170038.
- Ishwaran, H. and Kogalur, U.B. (2007). "*Random Survival Forests for R*". *R News*, 7(2):25-31.
- Ishwaran, H. and Kogalur, U.B. (2013). "*Contributed R Package randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*" CRAN.

See Also

- R package **randomForestSRC**

Index

* Epistasis

IRSF-package, 2
MACS, 12

* Genetic Variations Interactions

IRSF-package, 2
MACS, 12

* Interaction Detection and Modeling

cph.int, 3
cph.main, 7
IRSF-package, 2
MACS, 12
rsf.int, 15
rsf.int.signif, 21
rsf.main, 22
rsf.main.signif, 29

* Random Survival Forest

cph.int, 3
cph.main, 7
IRSF-package, 2
MACS, 12
rsf.int, 15
rsf.int.signif, 21
rsf.main, 22
rsf.main.signif, 29

* Real Dataset

MACS, 12

* Time-to-Event Analysis

cph.int, 3
cph.main, 7
IRSF-package, 2
MACS, 12
rsf.int, 15
rsf.int.signif, 21
rsf.main, 22
rsf.main.signif, 29

* documentation

IRSF.news, 11

cph.int, 3
cph.main, 7

IRSF (IRSF-package), 2

IRSF-package, 2

IRSF.news, 11

MACS, 12

rsf.int, 15

rsf.int.signif, 21

rsf.main, 22

rsf.main.signif, 29