Package 'icRSF'

February 27, 2018

Type Package

Title A Modified Random Survival Forest Algorithm

Version 1.2

Date 2018-2-26

Author Hui Xu and Raji Balasubramanian

Maintainer Hui Xu <huix@schoolph.umass.edu>

Description Implements a modification to the Random Survival Forests algorithm for obtaining variable importance in high dimensional datasets. The proposed algorithm is appropriate for settings in which a silent event is observed through sequentially administered, error-prone self-reports or laboratory based diagnostic tests. The modified algorithm incorporates a formal likelihood framework that accommodates sequentially administered, error-prone self-reports or laboratory based diagnostic tests. The original Random Survival Forests algorithm is modified by the introduction of a new splitting criterion based on a likelihood ratio test statistic.

License GPL (>= 2)

Imports Rcpp (>= 0.11.3), icensmis, parallel, stats

LinkingTo Rcpp

RoxygenNote 6.0.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-02-27 05:25:02 UTC

R topics documented:

icrsf *Permutation-based variable importance metric for high dimensional datasets appropriate for time to event outcomes, in the presence of imperfect self-reports or laboratory-based diagnostic tests.*

Description

Let N and P denote the number of subjects and number of variables in the dataset, respectively. Let N** denote the total number of visits, summed over all subjects in the study [i.e. N** denotes the number of diagnostic test results available for all subjects in the study]. This algorithm builds a user-defined number of survival trees, using bootstrapped datasets. Using the out of bag (OOB) data in each tree, a permutation-based measure of variable importance for each of the P variables is obtained.

Usage

```
icrsf(data, subject, testtimes, result, sensitivity, specificity, Xmat,
  root.size, ntree, ns, node, pval = 1)
```

Arguments

data	name of the data frame that includes the variables subject, testtimes, result
subject	vector of subject IDs of length N**x1.
testtimes	vector of visit or test times of length N**x1.
result	vector of binary diagnostic test results (0 = negative for event of interest; 1 = positive for event of interest) of length N**x1.
sensitivity	the sensitivity of the diagnostic test.
specificity	the specificity of the diagnostic test.
Xmat	a N x P matrix of covariates.
root.size	minimum number of subjects in a terminal node.
ntree	number of survival trees.
ns	number of covariate selected at each node to split the tree.
node	For parallel computation, specify the number of nodes.
pval	P-value threshold of the Likelihood Ratio Test.

Value

a vector of the ensembled variable importance for modified random survival forest (icRSF).

Examples

```
library(parallel)
data(Xmat)
data(pheno)
vimp <- icrsf(data=pheno, subject=ID, testtimes=time, result=result, sensitivity=1,
          specificity=1, Xmat=Xmat, root.size=30, ntree=1, ns=sqrt(ncol(Xmat)), node=1, pval=1)
```

---

pheno *A longitudinal data with diagnostic results for pre-determined time*

Description

A longitudinal data (629 x 3) with diagnostic results for 4 pre-determined times.

Usage

```
data(pheno)
```

Format

A data frame with 629 observations on 3 variables.

- ID: subject IDs
- time: Pre-determined diagnostic times. (1-4)
- result: diagnostic results (0=No, 1=Yes)

---

simout *Simulate error-prone test results for a user-defined vector of test times for each of the N subjects, for a user input NxP design matrix (Xmat).*

Description

This function simulates test results, subject to user-defined values of sensitivity, specificity, test times and matrix of covariates (Xmat). The first user-defined number of columns of the Xmat matrix are assumed to be true biomarkers, influencing the hazard function of the time to event of interest. In the reference group, event times are simulated assuming an exponential distribution, corresponding to user-defined parameter for the cumulative incidence in the study period of 8 years [1-noevent]. Assuming the PH model and user-defined vector of regression coefficients [betas], the time to event for individuals in each covariate stratum is simulated. Assuming that all subjects are tested at the same test times [testtimes], and user-defined values of sensitivity and specificity of the diagnostic test or self-report, test results are simulated at each test time, for each subject. When the parameter 'design' is set to its default value ['NMISS'], we assume that there are no missing test results. When the parameter 'design' is set to 'NTFP', no further test results are simulated following the first positive test result, for each subject.

Usage

```
simout(Xmat, testtimes, sensitivity, specificity, noevent, betas,
  design = "NMISS")
```

Arguments

Xmat	a 300x1000 covariate matrix that we simulated independently from binomial distribution with p=0.4.
testtimes	a vector of times at which self-reported outcomes are collected for all subjects.
sensitivity	the sensitivity of the self-report.
specificity	the specificity of the self-report.
noevent	denotes the probability of remaining event free by study end (or the complement of cumulative incidence)
betas	denotes the vector of regression coefficients associated with the set of biomarkers in the Cox PH model
design	denotes whether tests are missing after first positive result. 'NMISS' denotes no missing test after first positive and 'NTFP' denotes all tests are missing after first positive. (The default is 'NMISS').

Value

data frame: simulated longitudinal form of observed test results [1 row per subject per test time]. The dimensions of this dataframe are N** x 3, where first column is the subject ID, second column is the test times and the third column is the binary test result [1 = positive, indicating event of interest has occurred; 0=negative].

Examples

```
data(Xmat)
sim <- simout(Xmat, testtimes=1:4, sensitivity=1, specificity=1, noevent=0.7,
              betas=c(rep(0.81, 5), rep(0, ncol(Xmat)-5)), design="NTFP")
```

---

treebuilder *Permutation-based variable importance metric for high dimensional datasets appropriate for time to event outcomes, in the presence of imperfect self-reports or laboratory-based diagnostic tests.*

Description

Let N and P denote the number of subjects and number of variables in the dataset, respectively. Let N** denote the total number of visits, summed over all subjects in the study [i.e. N** denotes the number of diagnostic test results available for all subjects in the study]. This algorithm builds a user-defined number of survival trees, using bootstrapped datasets. Using the out of bag (OOB) data in each tree, a permutation-based measure of variable importance for each of the P variables is obtained.

Usage

```
treebuilder(data, subject, testtimes, result, sensitivity, specificity, Xmat,
  root.size, ns, pval = 1)
```

Arguments

data	name of the data frame that includes the variables subject, testtimes, result
subject	vector of subject IDs of length N**x1.
testtimes	vector of visit or test times of length N**x1.
result	vector of binary diagnostic test results (0 = negative for event of interest; 1 = positive for event of interest) of length N**x1.
sensitivity	the sensitivity of the diagnostic test.
specificity	the specificity of the diagnostic test.
Xmat	a N x P matrix of covariates.
root.size	the minimum number of subjects in a terminal node.
ns	number of covariate selected at each node to split the tree.
pval	P-value threshold of the Likelihood Ratio Test.

Value

a vector of the ensembled variable importance for modified random survival forest (icRSF).

Examples

```
data(Xmat)
data(pheno)
tree <- treebuilder(data=pheno, subject=ID, testtimes=time, result=result, sensitivity=1,
                    specificity=1, Xmat=Xmat, root.size=30, ns=sqrt(ncol(Xmat)), pval=1)
```

Xmat *A covariate matrix*

Description

a 300 x 1000 covariate matrix that we simulated independently from binomial distribution with p=0.4.

Usage

```
data(Xmat)
```

Format

A matrix with 300 observations on 1000 variables.

Index