Package ‘RAEN’

February 21, 2021

Title Random Approximate Elastic Net (RAEN) Variable Selection Method
Version 0.2
Encoding UTF-8
Description The Proportional Subdistribution Hazard (PSH) model has been popular for estimating the effects of the covariates on the cause of interest in Competing Risks analysis. The fast accumulation of large scale datasets has posed a challenge to classical statistical methods. Current penalized variable selection methods show unsatisfactory performance in ultra-high dimensional data. We propose a novel method, the Random Approximate Elastic Net (RAEN), with a robust and generalized solution to the variable selection problem for the PSH model. Our method shows improved sensitivity for variable selection compared with current methods.

Author Han Sun and Xiaofeng Wang
Maintainer Han Sun <han.sunny@gmail.com>

URL https://github.com/saintland/RAEN
Imports boot, foreach, doParallel, glmnet, fastcmprsk
Depends R(>= 3.5.0), lars
Suggests testthat, knitr, rmarkdown
License GPL (>= 2)
RoxygenNote 7.1.1
VignetteBuilder knitr
NeedsCompilation no
Repository CRAN
Date/Publication 2021-02-21 06:00:16 UTC

R topics documented:

decorr ................................................................. 2
grpselect ............................................................. 2
lossTrans ............................................................ 3
r2select ............................................................... 3
### deCorr

**Description**
Divide the highly correlated variables into exclusive groups

**Usage**
```
deCorr(x, rho = 0.7, ngrp = floor(15 * ncol(x)/nrow(x)))
```

**Arguments**
- `x`: the predictor matrix
- `rho`: the preset correlation threshold. Variables with correlation higher than rho will be separate into exclusive groups. Default is set to 0.7
- `ngrp`: the number of blocks to separate variables

**Value**
a dataframe of variable names 'varname' and the variable subgroup membership 'grp'

### grpselect

**Description**
This is the split step, where variable in subgroups are selected

**Usage**
```
grpselect(fgrp, x, y, B = 50, parallel = TRUE)
```

**Arguments**
- `fgrp`: the variable group object from 'deCorr'
- `x`: the predictor matrix
- `y`: a dataframe of time to event and event status. The primary outcome status is coded 1, the secondary outcome as 2, etc. The censored is coded as 0.
- `B`: the number of bootstraps
- `parallel`: whether to use multiple cores for parallel computing. Default is TRUE.
### lossTrans

**Value**

- a list of
  - fselect: Names of the selected variables.
  - prob: the generalized ridge variable importance.
  - weight: the inverse of the ridge variable importance.

---

### lossTrans

*Linear Approximation of the object function*

#### Description

Linear Approximation of the object function

#### Usage

```r
mod_lsa(obj, n)
```

#### Arguments

- `obj` the regression object from R output
- `n` the sample size

---

### r2select

*Variable Selection with the candidate pool*

#### Description

Perform variable selection with pooled candidates

#### Usage

```r
r2select(x.tr, y.tr, B, weight, prob, parallel = TRUE, m = 8)
```

#### Arguments

- `x.tr` the predictor matrix
- `y.tr` the time and status object for survival
- `B` times of bootstrap
- `weight` variable weight
- `prob` variable selection probability
- `parallel` Logical TRUE or FALSE. Whether to use multithread computing, which can save considerable amount of time for high dimensional data. Default is TRUE.
- `m` the number of variables to be randomly included in the model in this step. Default is 8.
RAEN

Random Ensemble Variable Selection for High Dimensional Data

Value

the estimates of variables with B bootstraps, which is a dataframe with B rows and \( \text{ncol}(x) \) columns.

Description

Perform variable selection for high dimensional data

Usage

RAEN(
  x,
  y,
  B,
  ngrp = floor(15 * ncol(x)/nrow(x)),
  parallel = TRUE,
  family = "competing",
  ncore = 2
)

## S3 method for class 'RAEN'
predict(object, newdata, ...)

Arguments

x         the predictor matrix
y         the time and status object for survival
B         times of bootstrap
ngrp      the number of blocks to separate variables into. Default is 15*p/N, where p is
          the number of predictors and N is the sample size.
parallel   Logical TRUE or FALSE. Whether to use multithread computing, which can
          save considerable amount of time for high dimensional data. Default is TRUE.
family     what family of data types. Default is ’competing’. Quantile regression for com-
          peting risks will be available through the developmental version on github
ncore      Number of cores used for parallel computing, if parallel=TRUE
object     the RAEN object containing the variable selection results
newdata    the predictor object containing the variable selection results
...        other parameters to pass
Value
a dataframe with the variable names and the regression coefficients
the linear predictor of the outcome risk

Examples

```r
library(RAEN)
data(toydata)
x=toydata[,c(1:2)]
y=toydata[,1:2]
fgrp<-deCorr(x, ngrp=20)
```

Description
We provide a novel solution to the variable selection problem in the ultra-high dimensional setting with a robust and generalized method.

Author(s)
Han Sun and Xiaofeng Wang

Description
toydata
Toy data for demonstration

Description
This simulated datasets contains 1000 predictors, of which X1-X20, X41-X60 are true predictors. The first two columns are time to event competing risks and status.

Format
A dataframe of 200 rows and 1002 columns.
Index

* datasets
  toydata, 5

deCorr, 2

grpselect, 2

lossTrans, 3

mod_lsa (lossTrans), 3

predict.RAEN (RAEN), 4

r2select, 3

RAEN, 4

RAEN-Package, 5

toydata, 5