

Package ‘CoxBoost’

April 17, 2009

Version 1.1

Title Cox models by likelihood based boosting for a single survival endpoint or competing risks

Author Harald Binder <binderh@fdm.uni-freiburg.de>

Maintainer Harald Binder <binderh@fdm.uni-freiburg.de>

Depends survival

Suggests snowfall

Description This package provides routines for fitting Cox models by likelihood based boosting for a single endpoint or in presence of competing risks

License GPL (>= 2)

Repository CRAN

Date/Publication 2009-02-15 09:25:14

R topics documented:

CoxBoost	2
cv.CoxBoost	5
optimCoxBoostPenalty	7
optimStepSizeFactor	9
predict.CoxBoost	11

Index	14
--------------	-----------

CoxBoost

*Fit a Cox model by likelihood based boosting***Description**

CoxBoost is used to fit a Cox proportional hazards model by componentwise likelihood based boosting. It is especially suited for models with a large number of predictors and allows for mandatory covariates with unpenalized parameter estimates.

Usage

```
CoxBoost (time, status, x, unpen.index=NULL, standardize=TRUE, stepno=100,
          penalty=9*sum(status==1), stepsize.factor=1,
          sf.scheme=c("sigmoid", "linear"), pendistmat=NULL,
          connected.index=NULL, trace=FALSE)
```

Arguments

time	vector of length n specifying the observed times.
status	censoring indicator, i.e., vector of length n with entries 0 for censored observations and 1 for uncensored observations. If this vector contains elements not equal to 0 or 1, these are taken to indicate events from a competing risk and a model for the subdistribution hazard with respect to event 1 is fitted (see e.g. Fine and Gray, 1999; Binder et al. 2009a).
x	$n * p$ matrix of covariates.
unpen.index	vector of length p . unpen with indices of mandatory covariates, where parameter estimation should be performed unpenalized.
standardize	logical value indicating whether covariates should be standardized for estimation. This does not apply for mandatory covariates, i.e., these are not standardized.
penalty	penalty value for the update of an individual element of the parameter vector in each boosting step.
stepsize.factor	determines the step-size modification factor by which the natural step size of boosting steps should be changed after a covariate has been selected in a boosting step. The default (value 1) implies constant penalties, for a value < 1 the penalty for a covariate is increased after it has been selected in a boosting step, and for a value > 1 the penalty it is decreased. If pendistmat is given, penalty updates are only performed for covariates that have at least one connection to another covariate.
sf.scheme	scheme for changing step sizes (via stepsize.factor). "linear" corresponds to the scheme described in Binder and Schumacher (2009b), "sigmoid" employs a sigmoid shape.

<code>pendistmat</code>	connection matrix with entries ranging between 0 and 1, with entry (i, j) indicating the certainty of the connection between covariates i and j . According to this information penalty changes due to <code>stepsize.factor < 1</code> are propagated, i.e., if entry (i, j) is non-zero, the penalty for covariate j is decreased after it has been increased for covariate i , after it has been selected in a boosting step. This matrix either has to have dimension $(p - p.unpen) * (p - p.unpen)$ or the indices of the <code>p.connected</code> connected covariates have to be given in <code>connected.index</code> , in which case the matrix has to have dimension <code>p.connected * p.connected</code> .
<code>connected.index</code>	indices of the <code>p.connected</code> connected covariates, for which <code>pendistmat</code> provides the connection information for distributing changes in penalties. No overlap with <code>unpen.index</code> is allowed. If <code>NULL</code> , and a connection matrix is given, all covariates are assumed to be connected.
<code>stepno</code>	number of boosting steps (m).
<code>trace</code>	logical value indicating whether progress in estimation should be indicated by printing the name of the covariate updated.

Details

In contrast to gradient boosting (implemented e.g. in the `glmboost` routine in the R package `mboost`, using the CoxPH loss function), `CoxBoost` is not based on gradients of loss functions, but adapts the offset-based boosting approach from Tutz and Binder (2007) for estimating Cox proportional hazards models. In each boosting step the previous boosting steps are incorporated as an offset in penalized partial likelihood estimation, which is employed for obtain an update for one single parameter, i.e., one covariate, in every boosting step. This results in sparse fits similar to Lasso-like approaches, with many estimated coefficients being zero. The main model complexity parameter, which has to be selected (e.g. by cross-validation using `cv.CoxBoost`), is the number of boosting steps `stepno`. The penalty parameter `penalty` can be chosen rather coarsely, either by hand or using `optimCoxBoostPenalty`.

The advantage of the offset-based approach compared to gradient boosting is that the penalty structure is very flexible. In the present implementation this is used for allowing for unpenalized mandatory covariates, which receive a very fast coefficient build-up in the course of the boosting steps, while the other (optional) covariates are subjected to penalization. For example in a microarray setting, the (many) microarray features would be taken to be optional covariates, and the (few) potential clinical covariates would be taken to be mandatory, by including their indices in `unpen.index`.

If a group of correlated covariates has influence on the response, e.g. genes from the same pathway, componentwise boosting will often result in a non-zero estimate for only one member of this group. To avoid this, information on the connection between covariates can be provided in `pendistmat`. If then, in addition, a penalty updating scheme with `stepsize.factor < 1` is chosen, connected covariates are more likely to be chosen in future boosting steps, if a directly connected covariate has been chosen in an earlier boosting step (see Binder and Schumacher, 2009b).

Value

`CoxBoost` returns an object of class `CoxBoost`.

`n`, `p` number of observations and number of covariates.

<code>stepno</code>	number of boosting steps.
<code>xnames</code>	vector of length <code>p</code> containing the names of the covariates. This information is extracted from <code>x</code> or names following the scheme <code>V1, V2, ...</code> .
<code>coefficients</code>	$(\text{stepno}+1) * p$ matrix containing the coefficient estimates for the (standardized) optional covariates for boosting steps 0 to <code>stepno</code> .
<code>meanx, sd</code>	vector of mean values and standard deviations used for standardizing the covariates.
<code>unpen.index</code>	indices of the mandatory covariates in the original covariate matrix <code>x</code> .
<code>penalty</code>	$\text{stepno} * (p - p.\text{unpen})$ matrix containing the penalties used for every boosting step and every penalized covariate.
<code>time</code>	observed times given in the <code>CoxBoost</code> call.
<code>status</code>	censoring indicator given in the <code>CoxBoost</code> call.
<code>event.times</code>	vector with event times from the data given in the <code>CoxBoost</code> call.
<code>linear.predictors</code>	$(\text{stepno}+1) * n$ matrix giving the linear predictor for boosting steps 0 to <code>stepno</code> and every observation.
<code>Lambda</code>	matrix with the Breslow estimate for the cumulative baseline hazard for boosting steps 0 to <code>stepno</code> for every event time.
<code>logplik</code>	partial log-likelihood of the fitted model in the final boosting step.

Author(s)

Written by Harald Binder binderh@fdm.uni-freiburg.de.

References

- Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009a). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, accepted.
- Binder, H. and Schumacher, M. (2009b). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*. 10:18.
- Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*. 9:14.
- Tutz, G. and Binder, H. (2007) Boosting ridge regression. *Computational Statistics & Data Analysis*, 51(12):6044-6059.
- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 94:496-509.

See Also

[predict.CoxBoost](#), [cv.CoxBoost](#).

Examples

```

# Generate some survival data with 10 informative covariates
n <- 200; p <- 100
beta <- c(rep(1,10), rep(0, p-10))
x <- matrix(rnorm(n*p), n, p)
real.time <- -(log(runif(n)))/(10*exp(drop(x %*% beta)))
cens.time <- rexp(n, rate=1/10)
status <- ifelse(real.time <= cens.time, 1, 0)
obs.time <- ifelse(real.time <= cens.time, real.time, cens.time)

# Fit a Cox proportional hazards model by CoxBoost

cbfit <- CoxBoost(time=obs.time, status=status, x=x, stepno=100, penalty=100)
summary(cbfit)

# ... with covariates 1 and 2 being mandatory

cbfit.mand <- CoxBoost(time=obs.time, status=status, x=x, unpen.index=c(1, 2),
                      stepno=100, penalty=100)
summary(cbfit.mand)

```

cv.CoxBoost

Determines the optimal number of boosting steps by cross-validation

Description

Performs a K-fold cross-validation for `CoxBoost` in search for the optimal number of boosting steps.

Usage

```

cv.CoxBoost(time, status, x, maxstepno=100, K=10, type=c("verweij", "naive"),
            parallel=FALSE, upload.x=TRUE, folds=NULL, trace=FALSE, ...)

```

Arguments

<code>time</code>	vector of length <code>n</code> specifying the observed times.
<code>status</code>	censoring indicator, i.e., vector of length <code>n</code> with entries 0 for censored observations and 1 for uncensored observations. If this vector contains elements not equal to 0 or 1, these are taken to indicate events from a competing risk and a model for the subdistribution hazard with respect to event 1 is fitted (see e.g. Fine and Gray, 1999).
<code>x</code>	<code>n * p</code> matrix of covariates.
<code>maxstepno</code>	maximum number of boosting steps to evaluate, i.e, the returned “optimal” number of boosting steps will be in the range <code>[0, maxstepno]</code> .

<code>K</code>	number of folds to be used for cross-validation. If <code>K</code> is larger or equal to the number of non-zero elements in <code>status</code> , leave-one-out cross-validation is performed.
<code>type</code>	way of calculating the partial likelihood contribution of the observation in the hold-out folds: <code>"verweij"</code> uses the more appropriate method described in Verweij and van Houwelingen (1996), <code>"naive"</code> uses the approach where the observations that are not in the hold-out folds are ignored (often found in other R packages).
<code>parallel</code>	logical value indicating whether computations in the cross-validation folds should be performed in parallel on a compute cluster. Parallelization is performed via the package <code>snowfall</code> and the initialization function of of this package, <code>sfInit</code> , should be called before calling <code>cv.CoxBoost</code> .
<code>upload.x</code>	logical value indicating whether <code>x</code> should/has to be uploaded to the compute cluster for parallel computation. Uploading this only once (using <code>sfExport(x)</code> from library <code>snowfall</code>) can save much time for large data sets.
<code>folds</code>	if not <code>NULL</code> , this has to be a list of length <code>K</code> , each element being a vector of indices of fold elements. Useful for employing the same folds for repeated runs.
<code>trace</code>	logical value indicating whether progress in estimation should be indicated by printing the number of the cross-validation fold and the index of the covariate updated.
<code>...</code>	miscellaneous parameters for the calls to <code>CoxBoost</code>

Value

List with the following components:

<code>mean.logplik</code>	vector of length <code>maxstepno+1</code> with the mean partial log-likelihood for boosting steps 0 to <code>maxstepno</code>
<code>se.logplik</code>	vector with standard error estimates for the mean partial log-likelihood criterion for each boosting step.
<code>optimal.step</code>	optimal boosting step number, i.e., with minimum mean partial log-likelihood.
<code>folds</code>	list of length <code>K</code> , where the elements are vectors of the indices of observations in the respective folds.

Author(s)

Harald Binder <binderh@fdm.uni-freiburg.de>

References

Verweij, P. J. M. and van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305-2314.

See Also

[CoxBoost](#), [optimCoxBoostPenalty](#)

Examples

```
## Not run:
# Generate some survival data with 10 informative covariates
n <- 200; p <- 100
beta <- c(rep(1,10), rep(0,p-10))
x <- matrix(rnorm(n*p), n, p)
real.time <- -(log(runif(n)))/(10*exp(drop(x %*% beta)))
cens.time <- rexp(n, rate=1/10)
status <- ifelse(real.time <= cens.time, 1, 0)
obs.time <- ifelse(real.time <= cens.time, real.time, cens.time)

# 10-fold cross-validation

cv.res <- cv.CoxBoost(time=obs.time, status=status, x=x, maxstepno=500,
                      K=10, type="verweij", penalty=100)

# examine mean partial log-likelihood in the course of the boosting steps
plot(cv.res$mean.logplik)

# Fit with optimal number of boosting steps

cbfit <- CoxBoost(time=obs.time, status=status, x=x, stepno=cv.res$optimal.step,
                  penalty=100)
summary(cbfit)

## End(Not run)
```

optimCoxBoostPenalty

Coarse line search for adequate penalty parameter

Description

This routine helps in finding a penalty value that leads to an “optimal” number of boosting steps for CoxBoost, determined by cross-validation, that is not too small/in a specified range.

Usage

```
optimCoxBoostPenalty(time, status, x, minstepno=50, maxstepno=200,
                     start.penalty=9*sum(status==1), iter.max=10,
                     upper.margin=0.05, parallel=FALSE,
                     trace=FALSE, ...)
```

Arguments

time vector of length n specifying the observed times.

<code>status</code>	censoring indicator, i.e., vector of length n with entries 0 for censored observations and 1 for uncensored observations. If this vector contains elements not equal to 0 or 1, these are taken to indicate events from a competing risk and a model for the subdistribution hazard with respect to event 1 is fitted (see e.g. Fine and Gray, 1999).
<code>x</code>	$n * p$ matrix of covariates.
<code>minstepno, maxstepno</code>	range of boosting steps in which the “optimal” number of boosting steps is wanted to be.
<code>start.penalty</code>	start value for the search for the appropriate penalty.
<code>iter.max</code>	maximum number of search iterations.
<code>upper.margin</code>	specifies the fraction of <code>maxstepno</code> which is used as an upper margin in which a cross-validation minimum is not taken to be one. This is necessary because of random fluctuations of cross-validated partial log-likelihood.
<code>parallel</code>	logical value indicating whether computations in the cross-validation folds should be performed in parallel on a compute cluster. Parallelization is performed via the package <code>snowfall</code> and the initialization function of of this package, <code>sfInit</code> , should be called before calling <code>cv.CoxBoost</code> .
<code>trace</code>	logical value indicating whether information on progress should be printed.
<code>...</code>	miscellaneous parameters for <code>cv.CoxBoost</code> .

Details

The penalty parameter for `CoxBoost` has to be chosen only very coarsely. In Tutz and Binder (2006) it is suggested for likelihood based boosting just to make sure, that the optimal number of boosting steps, according to some criterion such as cross-validation, is larger or equal to 50. With a smaller number of steps, boosting may become too “greedy” and show sub-optimal performance. This procedure uses a very coarse line search and so one should specify a rather large range of boosting steps.

Value

List with element `penalty` containing the “optimal” penalty and `cv.res` containing the corresponding result of `cv.CoxBoost`.

Author(s)

Written by Harald Binder (binderh@fdm.uni-freiburg.de).

References

Tutz, G. and Binder, H. (2006) Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, 62:961-971.

See Also

[CoxBoost](#), [cv.CoxBoost](#)

Examples

```
## Not run:
#   Generate some survival data with 10 informative covariates
n <- 200; p <- 100
beta <- c(rep(1,10), rep(0,p-10))
x <- matrix(rnorm(n*p), n, p)
real.time <- -(log(runif(n)))/(10*exp(drop(x %*% beta)))
cens.time <- rexp(n, rate=1/10)
status <- ifelse(real.time <= cens.time, 1, 0)
obs.time <- ifelse(real.time <= cens.time, real.time, cens.time)

#   determine penalty parameter

optim.res <- optimCoxBoostPenalty(time=obs.time, status=status, x=x,
                                trace=TRUE, start.penalty=500)

#   Fit with obtained penalty parameter and optimal number of boosting
#   steps obtained by cross-validation

cbfit <- CoxBoost(time=obs.time, status=status, x=x,
                 stepno=optim.res$cv.res$optimal.step,
                 penalty=optim.res$penalty)

summary(cbfit)

## End(Not run)
```

optimStepSizeFactor

Coarse line search for optimum step-size modification factor

Description

This routine helps in finding an optimum step-size modification factor for `CoxBoost`, i.e., that results in an optimum in terms of cross-validated partial log-likelihood.

Usage

```
optimStepSizeFactor(time, status, x,
                   direction=c("down", "up", "both"), start.stepsize=0.1,
                   iter.max=10, constant.cv.res=NULL,
                   parallel=FALSE, trace=FALSE, ...)
```

Arguments

`time` vector of length `n` specifying the observed times.

<code>status</code>	censoring indicator, i.e., vector of length <code>n</code> with entries 0 for censored observations and 1 for uncensored observations. If this vector contains elements not equal to 0 or 1, these are taken to indicate events from a competing risk and a model for the subdistribution hazard with respect to event 1 is fitted (see e.g. Fine and Gray, 1999).
<code>x</code>	<code>n * p</code> matrix of covariates.
<code>direction</code>	direction of line search for an optimal step-size modification factor (starting from value 1).
<code>start.stepsize</code>	step size used for the line search. A final step is performed using half this size.
<code>iter.max</code>	maximum number of search iterations.
<code>constant.cv.res</code>	result of <code>cv.CoxBoost</code> for <code>stepsize.factor=1</code> , that can be provided for saving computing time, if it already is available.
<code>parallel</code>	logical value indicating whether computations in the cross-validation folds should be performed in parallel on a compute cluster. Parallelization is performed via the package <code>snowfall</code> and the initialization function of of this package, <code>sfInit</code> , should be called before calling <code>cv.CoxBoost</code> .
<code>trace</code>	logical value indicating whether information on progress should be printed.
<code>...</code>	miscellaneous parameters for <code>cv.CoxBoost</code> .

Details

A coarse line search is performed for finding the best parameter `stepsize.factor` for `CoxBoost`. If an `pendistmat` argument is provided (which is passed on to `CoxBoost`), a search for factors smaller than 1 is sensible (corresponding to `direction="down"`). If no connection information is provided, it is reasonable to employ `direction="both"`, for avoiding restrictions without subject matter knowledge.

Value

List with the following components:

<code>factor.list</code>	array with the evaluated step-size modification factors.
<code>critmat</code>	matrix with the mean partial log-likelihood for each step-size modification factor in the course of the boosting steps.
<code>optimal.factor.index</code>	index of the optimal step-size modification factor.
<code>optimal.factor</code>	optimal step-size modification factor.
<code>optimal.step</code>	optimal boosting step number, i.e., with minimum mean partial log-likelihood, for step-size modification factor <code>optimal.factor</code> .

Author(s)

Written by Harald Binder (binderh@fdm.uni-freiburg.de).

References

Binder, H. and Schumacher, M. (2008b). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. Manuscript.

See Also

[CoxBoost](#), [cv.CoxBoost](#)

Examples

```
## Not run:
# Generate some survival data with 10 informative covariates
n <- 200; p <- 100
beta <- c(rep(1,10), rep(0,p-10))
x <- matrix(rnorm(n*p), n,p)
real.time <- -(log(runif(n)))/(10*exp(drop(x %*% beta)))
cens.time <- rexp(n, rate=1/10)
status <- ifelse(real.time <= cens.time, 1, 0)
obs.time <- ifelse(real.time <= cens.time, real.time, cens.time)

# Determine step-size modification factor. As there is no connection matrix,
# perform search into both directions

optim.res <- optimStepSizeFactor(direction="both",
                                time=obs.time, status=status, x=x,
                                trace=TRUE)

# Fit with obtained step-size modification parameter and optimal number of boosting
# steps obtained by cross-validation

cbfit <- CoxBoost(time=obs.time, status=status, x=x,
                 stepno=optim.res$optimal.step,
                 stepsize.factor=optim.res$optimal.factor)

summary(cbfit)

## End(Not run)
```

predict.CoxBoost *Predict method for CoxBoost fits*

Description

Obtains predictions at specified boosting steps from a CoxBoost object fitted by [CoxBoost](#).

Usage

```
## S3 method for class 'CoxBoost':
predict(object, newdata=NULL, newtime=NULL, newstatus=NULL,
        at.step=NULL, times=NULL, type=c("lp", "logplik", "risk", "CIF"), ...)
```

Arguments

<code>object</code>	fitted CoxBoost object from a <code>CoxBoost</code> call.
<code>newdata</code>	$n.new * p$ matrix with new covariate values. If just prediction for the training data is wanted, it can be omitted.
<code>newtime, newstatus</code>	vectors with observed time and censoring indicator (0 for censoring, 1 for no censoring, and any other values for competing events in a competing risks setting) for new observations, where prediction is wanted. Only required if predicted partial log-likelihood is wanted, i.e., if <code>type="logplik"</code> . This can also be omitted when prediction is only wanted for the training data, i.e., <code>newdata=NULL</code> .
<code>at.step</code>	scalar or vector of boosting step(s) at which prediction is wanted. If <code>type="risk"</code> is used, only one step is admissible. If no step is given, the final boosting step is used.
<code>times</code>	vector with T time points where prediction is wanted. Only needed for <code>type="risk"</code>
<code>type</code>	type of prediction to be returned: "lp" gives the linear predictor, "logplik" the partial log-likelihood, "risk" the predicted probability of not yet having had the event at the time points given in <code>times</code> , and "CIF" the predicted cumulative incidence function, i.e., the predicted probability of having had the event of interest.
<code>...</code>	miscellaneous arguments, none of which is used at the moment.

Value

For `type="lp"` and `type="logplik"` a vector of length $n.new$ (`at.step` being a scalar) or a $n.new * \text{length}(\text{at.step})$ matrix (`at.step` being a vector) with predictions is returned. For `type="risk"` or `type="CIF"` a $n.new * T$ matrix with predicted probabilities at the specific time points is returned.

Author(s)

Harald Binder (binderh@fdm.uni-freiburg.de)

Examples

```
# Generate some survival data with 10 informative covariates
n <- 200; p <- 100
beta <- c(rep(1,10), rep(0,p-10))
x <- matrix(rnorm(n*p), n,p)
real.time <- -(log(runif(n)))/(10*exp(drop(x %*% beta)))
cens.time <- rexp(n,rate=1/10)
status <- ifelse(real.time <= cens.time,1,0)
obs.time <- ifelse(real.time <= cens.time,real.time,cens.time)

# define training and test set

train.index <- 1:100
test.index <- 101:200
```

```
# Fit CoxBoost to the training data

cbfit <- CoxBoost(time=obs.time[train.index], status=status[train.index],
                 x=x[train.index,], stepno=300, penalty=100)

# mean partial log-likelihood for test set in every boosting step

step.logplik <- predict(cbfit, newdata=x[test.index,],
                      newtime=obs.time[test.index],
                      newstatus=status[test.index],
                      at.step=0:300, type="logplik")

plot(step.logplik)

# names of covariates with non-zero coefficients at boosting step
# with maximal test set partial log-likelihood

print(cbfit$xnames[cbfit$coefficients[which.max(step.logplik),] != 0])
```

Index

*Topic **models**

- CoxBoost, 1
- cv.CoxBoost, 5
- optimCoxBoostPenalty, 7
- optimStepSizeFactor, 9
- predict.CoxBoost, 11

*Topic **regression**

- CoxBoost, 1
- cv.CoxBoost, 5
- optimCoxBoostPenalty, 7
- optimStepSizeFactor, 9
- predict.CoxBoost, 11

*Topic **smooth**

- optimCoxBoostPenalty, 7
- optimStepSizeFactor, 9

*Topic **survial**

- CoxBoost, 1
- cv.CoxBoost, 5
- predict.CoxBoost, 11

CoxBoost, 1, 5, 6, 8–11

cv.CoxBoost, 3, 4, 5, 7–10

optimCoxBoostPenalty, 3, 6, 7

optimStepSizeFactor, 9

predict.CoxBoost, 4, 11