Compboost

Modular framework for component-wise boosting

Daniel Schalk

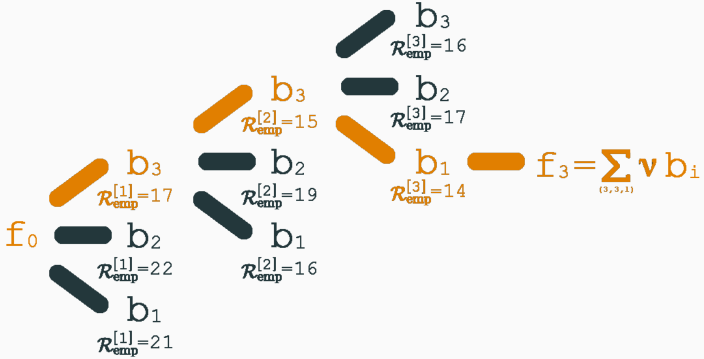
June 27, 2018

LMU Munich
Working Group Computational Statistics

Table of Contents

Why Component-Wise Boosting

Why Component-Wise Boosting?


f_0
b_3
$\mathcal{R}_{\text{emp}}^{[1]}=17$
b_2
$\mathcal{R}_{\text{emp}}^{[1]}=22$
b_1
$\mathcal{R}_{\text{emp}}^{[1]}=21$
b_3
$\mathcal{R}_{\text{emp}}^{[2]}=15$
b_2
$\mathcal{R}_{\text{emp}}^{[2]}=19$
b_1
$\mathcal{R}_{\text{emp}}^{[2]}=16$
b_3
$\mathcal{R}_{\text{emp}}^{[3]}=16$
b_2
$\mathcal{R}_{\text{emp}}^{[3]}=17$
b_1
$\mathcal{R}_{\text{emp}}^{[3]}=14$
$f_3=\sum_{(3,3,1)} \nu b_i$

Why Component-Wise Boosting?

- Inherent (unbiased) feature selection [HHKS11].
- The resulting model is sparse since the important models are selected first.
- The parameters are updated iteratively. Therefore, the parameters are estimated on the fly and can be interpreted due to linearity of the base-learners.
- Efficient model for learning in high-dimensional feature spaces ($p \gg n$).

Available R Packages

Most popular package for model-based boosting is `mboost`:

- Huge functionality: Lots of available base-learner and losses.
- Suited to boost more complex learner to do, for instance, survival analyses.
- Extendible with own base-learner and losses.

So, why another boosting implementation?

- Major parts of the algorithm are implemented in `R` which is not the best choice for expensive algorithms.
- The implementation is hard to debug due to nested scopes and no standard `R` system like `S3` or `S4`.

About Compboost

About Compboost

- With `mboost` as standard, we want to keep the modular principle of defining custom base-learner and losses.
- Completely written in `C++` and exposed by `Rcpp` to obtain high performance and full memory control.
- `R` API is written with `R6` to provide convenient wrapper.
- Major parts of the `compboost` functionality are unit tested against `mboost` to ensure correct use.

Compboost: Functionality

Main components:

- Base-learner and loss classes.
- Logger class for early stopping and logging mechanisms.

Possible extensions:

- Custom `R` or `C++` base-learner.
- Custom `R` or `C++` loss objects.
- Custom logging and stopping rules via custom losses.

Custom classes can be defined without recompiling the whole package, even when using `C++` functions.

Small Usecase

Initializing Model

```r
library(compboost)

data(PimaIndiansDiabetes, package = "mlbench")

cboost = Compboost$new(PimaIndiansDiabetes, "diabetes",
  loss = BinomialLoss$new())

cboost$addBaselearner("age", "spline", PSplineBlearner,
  degree = 3, knots = 10, penalty = 2, differences = 2)
cboost$addBaselearner("mass", "linear", PolynomialBlearner,
  degree = 1, intercept = TRUE)
```

Initializing Model

```
cboost$train(2000, trace = FALSE)
cboost


## Component-Wise Gradient Boosting
##
## Trained on PimaIndiansDiabetes with target diabetes
## Number of base-learners: 2
## Learning rate: 0.05
## Iterations: 2000
## Positive class: neg
## Offset: 0.3118
##
## BinomialLoss Loss:
##
##   Loss function: L(y,x) = log(1 + exp(-2yf(x))
##
##
```

Access Results

```
table(cboost$selected())
```

```
##
##  age_spline mass_linear
##        1511         489
```

```
head(cboost$risk())
```

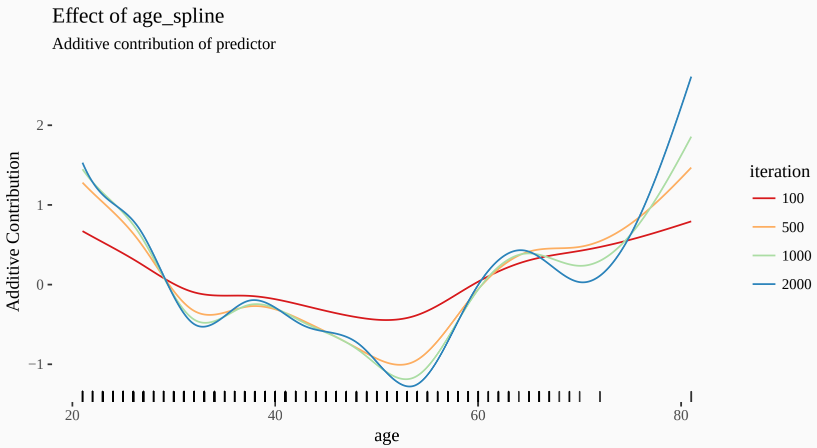
```
## [1] 0.6582 0.6565 0.6549 0.6533 0.6518 0.6503
```

```
str(cboost$coef())
```

```
## List of 3
##  $ age_spline : num [1:14, 1] 4.407 0.913 1.124 -1.044 0.126 ...
##  $ mass_linear: num [1:2, 1] 3.0352 -0.0886
##  $ offset     : num 0.312
```

Plot Results

```
cboost$plot("age_spline", iters = c(100, 500, 1000, 2000))
```



Custom Poisson Loss: Definition

```r
lossPoisson = function (truth, response) {
  return (-log(exp(response)^truth * exp(-exp(response)) / gamma(truth + 1)))
}
gradPoisson = function (truth, response) {
  return (exp(response) - truth)
}
constInitPoisson = function (truth) {
  return (log(mean.default(truth)))
}
# Define custom loss:
my.poisson.loss = CustomLoss$new(lossPoisson, gradPoisson, constInitPoisson)
```

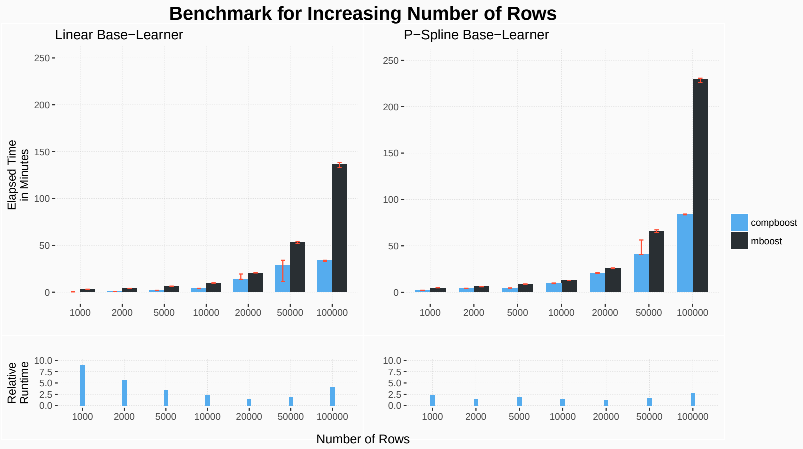
Custom Poisson Loss: Train Model

```r
data(VonBort, package = "vcd")

cboost = Compboost$new(VonBort, "deaths", loss = my.poisson.loss)
cboost$addBaselearner("year", "spline", PSplineBlearner,
  degree = 3, knots = 10, penalty = 2, differences = 2)
cboost$train(100, trace = FALSE)

mod = mboost(deaths ~ bbs(year, differences = 2, lambda = 2,
  degree = 3, knots = 10), data = VonBort, family = Poisson(),
  control = boost_control(mstop = 100, nu = 0.05))

head(cbind(
  compboost = as.numeric(cboost$coef()[[1]]),
  mboost = as.numeric(coef(mod)[[1]])))
```

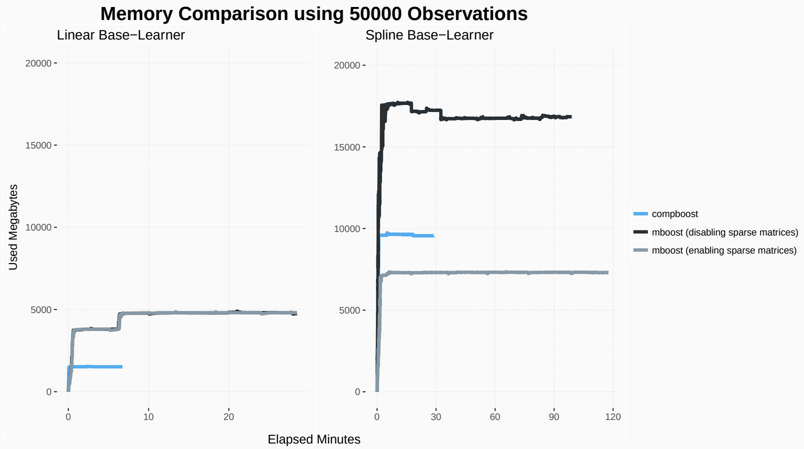
```
##      compboost  mboost
## [1,]   -1.4093 -1.4093
## [2,]   -0.8975 -0.8975
## [3,]   -0.4040 -0.4040
## [4,]    0.0947  0.0947
## [5,]    0.3384  0.3384
## [6,]    0.1531  0.1531
```

Small Benchmark

Runtime Comparison Against Mboost



Memory Comparison Against Mboost



Next Steps

Next Steps

- Implementing more base-learner and loss functions.
- Support for multiclass classification.
- Parallel computations.
- Using sparse matrices to store, e.g., spline data matrices.

Where to Find

- Developed on *GitHub*:

  www.github.com/schalkdaniel/compboost

- Additional resources on the project page:

  www.compboost.org

- Bug reports via the issue tracker.
- Contributions are welcome!

Questions?

References

Dirk Eddelbuettel.
Seamless R and C++ Integration with Rcpp.
Springer, New York, 2013.
ISBN 978-1-4614-6867-7.

Dirk Eddelbuettel and Romain François.
Exposing C++ functions and classes with rcpp modules.
Vignette included in R package Rcpp, URL http://CRAN. R-Project. org/package= Rcpp, 2017.

Torsten Hothorn, Peter Buehlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner.
mboost: Model-Based Boosting, 2017.
R package version 2.8-1.

Benjamin Hofner, Torsten Hothorn, Thomas Kneib, and Matthias Schmid.
A framework for unbiased model selection based on boosting.
Journal of Computational and Graphical Statistics, 20(4):956–971, 2011.