

Classification of Breast Cancer Clinical Stage with Gene Expression Data

Zhu Wang
UT Health San Antonio
wangz1@uthscsa.edu

February 1, 2019

This document presents analysis for the MAQC-II project, human breast cancer data set with boosting algorithms developed in Wang (2018a,b) and implemented in R package `bst`.

Dataset comes from the MicroArray Quality Control (MAQC) II project and includes 278 breast cancer samples with 164 estrogen receptor (ER) positive cases. The data files `GSE20194_series_matrix.txt.gz` and `GSE20194_MDACC_Sample_Info.xls` can be downloaded from <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=rhojvaihkwcsaihq&acc=GSE20194>. After reading the data, some unused variables are removed. From 22283 genes, the dataset is pre-screened to obtain 3000 genes with the largest absolute values of the two-sample t-statistics. The 3000 genes are standardized.

```
# The data files below were downloaded on June 1, 2016
require("gdata")
bc <- t(read.delim("GSE20194_series_matrix.txt.gz", sep = "",
  header = FALSE, skip = 80))
colnames(bc) <- bc[1, ]
bc <- bc[-1, -c(1, 2)]
### The last column is empty with variable name
### !series_matrix_table_end, thus omitted
bc <- bc[, -22284]
mode(bc) <- "numeric" ### convert character to numeric
dat1 <- read.xls("GSE20194_MDACC_Sample_Info.xls", sheet = 1,
  header = TRUE)
y <- dat1$characteristics..ER_status
y <- ifelse(y == "P", 1, -1)
table(y)
## y
## -1  1
## 114 164
res <- rep(NA, dim(bc)[2])
for (i in 1:dim(bc)[2]) res[i] <- abs(t.test(bc[, i] ~ y)$statistic)
### find 3000 largest absolute value of t-statistic
tmp <- order(res, decreasing = TRUE)[1:3000]
```

```

dat <- bc[, tmp]
### standardize variables
dat <- scale(dat)

Set up configuration parameters.

nrun <- 100
per <- c(0, 0.05, 0.1, 0.15)
learntype <- c("tree", "ls")[2]
tuning <- "error"
n.cores <- 4
plot.it <- TRUE
### robust tuning parameters used in bst/rbst function
s <- c(0.9, 1.01, 0.5, -0.2, 0.8, -0.5, -0.2)
nu <- c(0.01, 0.1, 0.01, rep(0.1, 4))
m <- 100 ### boosting iteration number
### whether to truncate the predicted values in each boosting
### iteration?
ctr.trun <- c(TRUE, rep(FALSE, 6))
### used in bst function
bsttype <- c("closs", "gloss", "qloss", "binom", "binom", "hinge",
            "expo")
### and corresponding labels
bsttype1 <- c("ClossBoost", "GlossBoost", "QlossBoost", "LogitBoost",
            "LogitBoost", "HingeBoost", "AdaBoost")
### used in rbst function
rbsttype <- c("closs", "gloss", "qloss", "tbinom", "binomd",
            "thinge", "texpo")
### and corresponding labels
rbsttype1 <- c("ClossBoostQM", "GlossBoostQM", "QlossBoostQM",
            "TLogitBoost", "DlogitBoost", "THingeBoost", "TAdaBoost")

The training data contains randomly selected 50 samples with positive estrogen receptor status and 50 samples with negative estrogen receptor status, and the rest were designated as the test data. The training data is contaminated by randomly switching response variable labels at varying pre-specified proportions per=0, 0.05, 0.1, 0.15. This process is repeated nrun=100 times. The base learner is learntype=ls (with quotes). To select optimal boosting iteration from maximum value of m=100, we run five-fold cross-validation averaging classification errors. In cross-validation, we set the number of cores for parallel computing by n.cores=4. Selected results can be plotted if plot.it=TRUE. Gradient based boosting includes ClossBoost, GlossBoost, QlossBoost, LogitBoost, HingeBoost and AdaBoost. Robust boosting using rbst contains ClossBoostQM, GlossBoostQM, QlossBoostQM, TLogitBoost, DlogitBoost, THingeBoost and TAdaBoost.

summary7 <- function(x) c(summary(x), sd = sd(x))
ptm <- proc.time()
library("bst")

```

```

## Loading required package: gbm
## Loaded gbm 2.1.4

for (k in 1:7) {
  ### k controls which family in bst, and rfamily in rbst
  err.m1 <- err.m2 <- nvar.m1 <- nvar.m2 <- errbest.m1 <- errbest.m2 <- matrix(NA,
    ncol = 4, nrow = nrun)
  mstopbest.m1 <- mstopbest.m2 <- mstopcv.m1 <- mstopcv.m2 <- matrix(NA,
    ncol = 4, nrow = nrun)
  colnames(err.m1) <- colnames(err.m2) <- c("cont-0%", "cont-5%",
    "cont-10%", "cont-15%")
  colnames(mstopcv.m1) <- colnames(mstopcv.m2) <- colnames(err.m1)
  colnames(nvar.m1) <- colnames(nvar.m2) <- colnames(err.m1)
  colnames(errbest.m1) <- colnames(errbest.m2) <- colnames(err.m1)
  colnames(mstopbest.m1) <- colnames(mstopbest.m2) <- colnames(err.m1)
  for (ii in 1:nrun) {
    set.seed(1000 + ii)
    trid <- c(sample(which(y == 1))[1:50], sample(which(y ==
      -1))[1:50])
    dtr <- dat[trid, ]
    dte <- dat[-trid, ]
    ytrold <- y[trid]
    yte <- y[-trid]
    ### number of patients/no. variables in training and test data
    dim(dtr)
    dim(dte)
    ### randomly contaminate data
    ntr <- length(trid)
    set.seed(1000 + ii)
    con <- sample(ntr)
    for (j in 1) {
      ### controls learner type i controls how many percentage of data
      ### contaminated
      for (i in 1:4) {
        ytr <- ytrold
        percon <- per[i]
        ### randomly flip labels of the samples in training set
        ### according to pre-defined contamination level
        if (percon > 0) {
          ji <- con[1:(percon * ntr)]
          ytr[ji] <- -ytrold[ji]
        }
        dat.m1 <- bst(x = dtr, y = ytr, ctrl = bst_control(mstop = m,
          center = FALSE, trace = FALSE, nu = nu[k],
          s = s[k], trun = ctr.trun[k]), family = bsttype[k],
          learner = learner[j])
        err1 <- predict(dat.m1, newdata = dte, newy = yte,
          type = "error")
        err1tr <- predict(dat.m1, newdata = dtr, newy = ytr,
          type = "loss")
      }
    }
  }
}

```

```

### cross-validation to select best boosting iteration
set.seed(1000 + ii)
cvm1 <- cv.bst(x = dtr, y = ytr, K = 5, n.cores = n.cores,
  ctrl = bst_control(mstop = m, center = FALSE,
    trace = FALSE, nu = nu[k], s = s[k], trun = ctr.trun[k]),
  family = bsttype[k], learner = learntype[j],
  main = bsttype[k], type = tuning, plot.it = FALSE)
optmstop <- max(10, which.min(cvm1$cv))
err.m1[ii, i] <- err1[optmstop]
nvar.m1[ii, i] <- nsel(dat.m1, optmstop)[optmstop]
errbest.m1[ii, i] <- min(err1)
mstopbest.m1[ii, i] <- which.min(err1)
mstopcv.m1[ii, i] <- optmstop
dat.m2 <- rbst(x = dtr, y = ytr, ctrl = bst_control(mstop = m,
  iter = 100, nu = nu[k], s = s[k], trun = ctr.trun[k],
  center = FALSE, trace = FALSE), rfamily = rbsttype[k],
  learner = learntype[j])
err2 <- predict(dat.m2, newdata = dte, newy = yte,
  type = "error")
err2tr <- predict(dat.m2, newdata = dtr, newy = ytr,
  type = "loss")
### cross-validation to select best boosting iteration
set.seed(1000 + ii)
cvm2 <- cv.rbst(x = dtr, y = ytr, K = 5, n.cores = n.cores,
  ctrl = bst_control(mstop = m, iter = 100, nu = nu[k],
    s = s[k], trun = ctr.trun[k], center = FALSE,
    trace = FALSE), rfamily = rbsttype[k], learner = learntype[j],
  main = rbsttype[k], type = tuning, plot.it = FALSE)
optmstop <- max(10, which.min(cvm2$cv))
err.m2[ii, i] <- err2[optmstop]
nvar.m2[ii, i] <- nsel(dat.m2, optmstop)[optmstop]
errbest.m2[ii, i] <- min(err2)
mstopbest.m2[ii, i] <- which.min(err2)
mstopcv.m2[ii, i] <- optmstop
}
}
if (ii%nrn == 0) {
  if (bsttype[k] %in% c("closs", "gloss", "qloss"))
    cat(paste("\nbst family ", bsttype1[k], ", s=",
      s[k], ", nu=", nu[k], sep = ""), "\n")
  if (bsttype[k] %in% c("binom", "hinge", "expo"))
    cat(paste("\nbst family ", bsttype1[k], ", nu=",
      nu[k], sep = ""), "\n")
  cat("best misclassification error from bst\n")
  print(round(apply(errbest.m1, 2, summary7), 4))
  cat("CV based misclassification error from bst\n")
  print(round(apply(err.m1, 2, summary7), 4))
  cat("best mstop with best misclassification error from bst\n")
  print(round(apply(mstopbest.m1, 2, summary7), 0))
}

```

```

cat("best mstop with CV from bst\n")
print(round(apply(mstopcv.m1, 2, summary7), 0))
cat("nvar from bst\n")
print(round(apply(nvar.m1, 2, summary7), 1))

cat(paste("\nrbst family ", rbsttype1[k], ", s=",
          s[k], ", nu=", nu[k], sep = ""), "\n")
cat("\nbest misclassification error from rbst\n")
print(round(apply(errbest.m2, 2, summary7), 4))
cat("CV based misclassification error from rbst\n")
print(round(apply(err.m2, 2, summary7), 4))
cat("best mstop with best misclassification error from rbst\n")
print(round(apply(mstopbest.m2, 2, summary7), 0))
cat("best mstop with CV from rbst\n")
print(round(apply(mstopcv.m2, 2, summary7), 0))
cat("nvar from rbst\n")
print(round(apply(nvar.m2, 2, summary7), 1))
res <- list(err.m1 = err.m1, nvar.m1 = nvar.m1, errbest.m1 = errbest.m1,
            mstopbest.m1 = mstopbest.m1, mstopcv.m1 = mstopcv.m1,
            err.m2 = err.m2, nvar.m2 = nvar.m2, errbest.m2 = errbest.m2,
            mstopbest.m2 = mstopbest.m2, mstopcv.m2 = mstopcv.m2,
            s = s[k], nu = nu[k], trun = ctr.trun[k], family = bsttype[k],
            rfamily = rbsttype[k])
if (plot.it) {
  par(mfrow = c(2, 1))
  boxplot(err.m1, main = "Misclassification error",
          subset = "", sub = bsttype1[k])
  boxplot(err.m2, main = "Misclassification error",
          subset = "", sub = rbsttype1[k])
  boxplot(nvar.m1, main = "No. variables", subset = "",
          sub = bsttype1[k])
  boxplot(nvar.m2, main = "No. variables", subset = "",
          sub = rbsttype1[k])
}
check <- FALSE
if (check) {
  par(mfrow = c(3, 1))
  title <- paste("percentage of contamination ",
                percon, sep = "")
  plot(err2tr, main = title, ylab = "Loss value",
        xlab = "Iteration", type = "l", lty = "dashed",
        col = "red")
  points(err1tr, type = "l", lty = "solid", col = "black")
  legend("topright", c(bsttype1[k], rbsttype1[k]),
        lty = c("solid", "dashed"), col = c("black",
        "red"))
  plot(err2, main = title, ylab = "Misclassification error",
        xlab = "Iteration", type = "l", lty = "dashed",
        col = "red")
}

```

```

points(err1, type = "l")
legend("bottomright", c(bsttype1[k], rbsttype1[k]),
      lty = c("solid", "dashed"), col = c("black",
      "red"))
plot(nsel(dat.m2, m), main = title, ylab = "No. variables",
      xlab = "Iteration", lty = "dashed", col = "red",
      type = "l")
points(nsel(dat.m1, m), ylab = "No. variables",
      xlab = "Iteration", lty = "solid", type = "l",
      col = "black")
legend("bottomright", c(bsttype1[k], rbsttype1[k]),
      lty = c("solid", "dashed"), col = c("black",
      "red"))
    }
  }
}
##
## bst family ClossBoost, s=0.9, nu=0.01
## best misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0506 0.0506  0.0449  0.0449
## 1st Qu. 0.0730 0.0730  0.0787  0.0787
## Median  0.0787 0.0843  0.0843  0.1011
## Mean    0.0804 0.0837  0.0971  0.1172
## 3rd Qu. 0.0843 0.0899  0.1081  0.1461
## Max.    0.1292 0.1404  0.2079  0.2528
## sd      0.0135 0.0154  0.0309  0.0480
## CV based misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0618 0.0562  0.0618  0.0618
## 1st Qu. 0.0843 0.0843  0.0899  0.0955
## Median  0.0899 0.0955  0.1011  0.1152
## Mean    0.0909 0.0946  0.1138  0.1338
## 3rd Qu. 0.1011 0.1025  0.1306  0.1573
## Max.    0.1292 0.1798  0.2360  0.2865
## sd      0.0139 0.0190  0.0385  0.0530
## best mstop with best misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.      1      1      1      1
## 1st Qu.   30     29     37     42
## Median    50     48     60     66
## Mean      49     47     56     61
## 3rd Qu.   72     66     80     89
## Max.     100    100    100    100
## sd        30     29     30     31
## best mstop with CV from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.      10     10     10     10

```

```

## 1st Qu.      10      18      33      52
## Median       39      46      50      68
## Mean         40      47      52      65
## 3rd Qu.      60      71      71      86
## Max.         100     100     100     100
## sd           29      30      27      24
## nvar from bst
##           cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     1.0     1.0     1.0
## 1st Qu.       1.0     1.0     1.0     2.0
## Median        1.0     2.0     2.0     4.0
## Mean          2.4     3.1     3.2     3.8
## 3rd Qu.        3.0     4.2     5.0     5.0
## Max.          10.0    12.0    11.0    11.0
## sd            2.1     2.7     2.4     2.4
##
## rbst family ClossBoostQM, s=0.9, nu=0.01
##
## best misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.         0.0506  0.0562  0.0449  0.0506
## 1st Qu.       0.0730  0.0730  0.0730  0.0772
## Median        0.0787  0.0787  0.0843  0.0843
## Mean          0.0792  0.0804  0.0870  0.0958
## 3rd Qu.       0.0843  0.0899  0.0955  0.1067
## Max.          0.1067  0.1180  0.1854  0.2303
## sd            0.0126  0.0125  0.0212  0.0330
## CV based misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.         0.0562  0.0618  0.0562  0.0674
## 1st Qu.       0.0787  0.0843  0.0843  0.0899
## Median        0.0899  0.0899  0.0955  0.1011
## Mean          0.0903  0.0923  0.1024  0.1146
## 3rd Qu.       0.1011  0.1011  0.1067  0.1236
## Max.          0.1180  0.1236  0.2022  0.2640
## sd            0.0138  0.0140  0.0256  0.0401
## best mstop with best misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.           1         1         1         1
## 1st Qu.        14         9         6         9
## Median         28        22        22        19
## Mean           31        26        27        31
## 3rd Qu.        48        38        37        52
## Max.           91       100        99       100
## sd             22        22        25        30
## best mstop with CV from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.           10        10        10        10
## 1st Qu.         10        10        10        10

```

```

## Median      19      16      19      14
## Mean       28      28      33      31
## 3rd Qu.    41      40      53      47
## Max.       99     100      95     100
## sd        22      24      27      28
## nvar from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1.0     1.0     1.0     1.0
## 1st Qu.   1.0     1.0     1.0     1.0
## Median    2.0     2.0     2.0     2.0
## Mean     3.1     3.3     4.0     4.3
## 3rd Qu.   4.0     4.0     6.0     6.0
## Max.     15.0    16.0    14.0    16.0
## sd       3.1     3.3     3.6     4.1
##
## bst family GlossBoost, s=1.01, nu=0.1
## best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.    0.0449  0.0506  0.0449  0.0562
## 1st Qu. 0.0730  0.0730  0.0787  0.0829
## Median  0.0787  0.0843  0.0843  0.1011
## Mean    0.0812  0.0837  0.0948  0.1126
## 3rd Qu. 0.0899  0.0899  0.1067  0.1348
## Max.    0.1292  0.1236  0.1910  0.2584
## sd      0.0138  0.0142  0.0280  0.0426
## CV based misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.    0.0562  0.0562  0.0562  0.0618
## 1st Qu. 0.0787  0.0843  0.0899  0.0955
## Median  0.0899  0.0927  0.1011  0.1236
## Mean    0.0912  0.0947  0.1121  0.1319
## 3rd Qu. 0.1011  0.1011  0.1292  0.1573
## Max.    0.1798  0.1966  0.2360  0.3258
## sd      0.0166  0.0191  0.0370  0.0496
## best mstop with best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1      1      1      1
## 1st Qu.   19     16     22     23
## Median   45     36     40     48
## Mean    45     40     45     48
## 3rd Qu.  70     59     70     77
## Max.    99    100     99    100
## sd     30     28     31     32
## best mstop with CV from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.     10     10     10     10
## 1st Qu.  10     10     17     28
## Median   20     31     32     44
## Mean    30     39     39     48

```



```

## 3rd Qu.      45      55      54      72
## Max.        95     100     100     98
## sd         24      29      26      28
## nvar from bst
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     1.0     1.0     1.0
## 1st Qu.     1.0     1.0     1.0     2.0
## Median      1.0     2.0     2.0     4.0
## Mean        1.9     2.7     3.0     4.4
## 3rd Qu.     2.0     3.0     4.0     6.2
## Max.        9.0    12.0    10.0    13.0
## sd         1.6     2.3     2.2     3.2
##
## rbst family GlossBoostQM, s=1.01, nu=0.1
##
## best misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506 0.0562 0.0449 0.0562
## 1st Qu.   0.0730 0.0730 0.0730 0.0787
## Median    0.0787 0.0787 0.0843 0.0899
## Mean      0.0811 0.0826 0.0910 0.1037
## 3rd Qu.   0.0899 0.0899 0.1011 0.1250
## Max.      0.1292 0.1124 0.1910 0.2360
## sd        0.0137 0.0132 0.0264 0.0391
## CV based misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618 0.0562 0.0618
## 1st Qu.   0.0829 0.0843 0.0899 0.0899
## Median    0.0899 0.0955 0.1011 0.1096
## Mean      0.0910 0.0940 0.1067 0.1238
## 3rd Qu.   0.1011 0.1011 0.1124 0.1517
## Max.      0.1461 0.1629 0.2135 0.2640
## sd        0.0148 0.0177 0.0315 0.0433
## best mstop with best misclassification error from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1      1      1      1
## 1st Qu.     22     10      8      8
## Median      46     37     29     20
## Mean        46     37     34     32
## 3rd Qu.     73     56     52     55
## Max.       100     97    100    100
## sd          30     29     29     30
## best mstop with CV from rbst
##           cont-0% cont-5% cont-10% cont-15%
## Min.        10     10     10     10
## 1st Qu.     10     10     10     10
## Median      10     19     22     20
## Mean        31     33     35     32
## 3rd Qu.     50     48     54     47

```

```

## Max.      100      98      97      96
## sd        27      27      28      27
## nvar from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1.0      1.0      1.0      1.0
## 1st Qu.   1.0      1.0      1.0      1.0
## Median    1.0      2.0      2.0      2.0
## Mean      2.5      2.6      3.2      3.8
## 3rd Qu.   3.0      3.0      4.2      5.2
## Max.     11.0     11.0     11.0     13.0
## sd        2.4      2.4      2.6      3.4
##
## bst family QlossBoost, s=0.5, nu=0.01
## best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0449  0.0506  0.0449  0.0562
## 1st Qu.  0.0730  0.0730  0.0787  0.0829
## Median   0.0787  0.0843  0.0843  0.1011
## Mean     0.0812  0.0835  0.0948  0.1132
## 3rd Qu.  0.0899  0.0899  0.1067  0.1348
## Max.     0.1292  0.1180  0.1910  0.2584
## sd       0.0139  0.0140  0.0278  0.0434
## CV based misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0562  0.0562  0.0562  0.0562
## 1st Qu.  0.0787  0.0843  0.0899  0.0955
## Median   0.0899  0.0955  0.1011  0.1292
## Mean     0.0908  0.0952  0.1120  0.1340
## 3rd Qu.  0.1011  0.1011  0.1306  0.1545
## Max.     0.1798  0.1966  0.2360  0.3258
## sd       0.0167  0.0192  0.0367  0.0511
## best mstop with best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1       1       1       1
## 1st Qu.   19      17      25      23
## Median    39      38      42      52
## Mean      42      40      47      49
## 3rd Qu.   64      60      72      78
## Max.      98     100     99     100
## sd        29     28     31     33
## best mstop with CV from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      10      10      10      10
## 1st Qu.   10      10      18      28
## Median    24      32      39      45
## Mean      32      38      41      49
## 3rd Qu.   49      56      58      70
## Max.      99      98      97     100
## sd        24      27      27      28

```

```

## nvar from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          1.0    1.0    1.0    1.0
## 1st Qu.        1.0    1.0    1.0    2.0
## Median         1.0    2.0    2.0    3.0
## Mean           2.0    2.6    3.2    4.1
## 3rd Qu.        2.0    3.0    5.0    6.2
## Max.           10.0   11.0   10.0   12.0
## sd             1.7    2.2    2.3    3.0
##
## rbst family QlossBoostQM, s=0.5, nu=0.01
##
## best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506 0.0562  0.0449  0.0562
## 1st Qu.   0.0730 0.0730  0.0772  0.0787
## Median    0.0787 0.0787  0.0843  0.0899
## Mean      0.0811 0.0822  0.0908  0.1026
## 3rd Qu.   0.0899 0.0899  0.1011  0.1250
## Max.      0.1292 0.1124  0.1910  0.2360
## sd        0.0136 0.0131  0.0260  0.0386
## CV based misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0562  0.0618
## 1st Qu.   0.0787 0.0843  0.0899  0.0899
## Median    0.0899 0.0927  0.1011  0.1067
## Mean      0.0909 0.0946  0.1064  0.1225
## 3rd Qu.   0.1011 0.1011  0.1124  0.1461
## Max.      0.1517 0.1573  0.2191  0.2640
## sd        0.0154 0.0170  0.0313  0.0428
## best mstop with best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          1      1      1      1
## 1st Qu.       18     14      6      7
## Median        42     38     26     20
## Mean          43     38     33     33
## 3rd Qu.       69     58     50     54
## Max.         100     96     97     99
## sd           30     28     29     31
## best mstop with CV from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          10     10     10     10
## 1st Qu.        10     10     10     10
## Median         10     24     23     26
## Mean          29     34     33     35
## 3rd Qu.        46     51     46     54
## Max.          93    100     99     96
## sd           24     28     27     28
## nvar from rbst

```

```

##          cont-0% cont-5% cont-10% cont-15%
## Min.          1.0    1.0    1.0    1.0
## 1st Qu.        1.0    1.0    1.0    1.0
## Median         1.0    2.0    2.0    3.0
## Mean           2.4    2.8    3.1    4.1
## 3rd Qu.        3.0    3.0    4.0    7.0
## Max.           11.0   12.0   12.0   15.0
## sd             2.2    2.6    2.6    3.6
##
## bst family LogitBoost, nu=0.1
## best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        0.0449 0.0562 0.0449 0.0506
## 1st Qu.      0.0730 0.0730 0.0787 0.1053
## Median       0.0843 0.0843 0.1124 0.1433
## Mean         0.0824 0.0896 0.1146 0.1487
## 3rd Qu.      0.0899 0.1067 0.1419 0.1798
## Max.         0.1461 0.1517 0.2303 0.3258
## sd           0.0152 0.0208 0.0419 0.0606
## CV based misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.         0.0506 0.0562 0.0618 0.0562
## 1st Qu.      0.0843 0.0885 0.0997 0.1222
## Median       0.0899 0.1039 0.1348 0.1657
## Mean         0.0907 0.1027 0.1341 0.1735
## 3rd Qu.      0.1011 0.1180 0.1587 0.2107
## Max.         0.1573 0.1573 0.2697 0.3876
## sd           0.0145 0.0218 0.0444 0.0703
## best mstop with best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          1     4     1     5
## 1st Qu.        3    14    26    33
## Median        44    37    48    69
## Mean          42    42    50    61
## 3rd Qu.       72    60    80    92
## Max.          100   99   100  100
## sd            34    28    30    31
## best mstop with CV from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          10    15    16    10
## 1st Qu.       28    33    36    39
## Median        55    52    58    60
## Mean          53    53    58    60
## 3rd Qu.       74    70    76    84
## Max.          100   99    98   100
## sd            26    23    24    26
## nvar from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.          1.0    1.0    1.0    1.0

```

```

## 1st Qu.      1.0      2.0      3.0      4.0
## Median      2.0      3.0      5.0      7.0
## Mean        2.5      3.6      5.3      6.9
## 3rd Qu.     3.0      5.0      8.0      9.2
## Max.        9.0     10.0     13.0     15.0
## sd          1.8      2.2      2.9      3.6
##
## rbst family TLogitBoost, s=-0.2, nu=0.1
##
## best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0506  0.0506  0.0449
## 1st Qu.   0.0730 0.0674  0.0787  0.0787
## Median    0.0843 0.0843  0.0899  0.1011
## Mean      0.0838 0.0825  0.1022  0.1156
## 3rd Qu.   0.0955 0.0955  0.1124  0.1404
## Max.      0.1180 0.1292  0.2360  0.2865
## sd        0.0136 0.0164  0.0395  0.0539
## CV based misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0562  0.0562  0.0562
## 1st Qu.   0.0787 0.0787  0.0843  0.0885
## Median    0.0899 0.0899  0.1011  0.1096
## Mean      0.0884 0.0920  0.1125  0.1326
## 3rd Qu.   0.0955 0.1067  0.1236  0.1699
## Max.      0.1180 0.1348  0.3146  0.3933
## sd        0.0131 0.0182  0.0453  0.0646
## best mstop with best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           1      4      1      4
## 1st Qu.        1     12     20     26
## Median         11     31     50     50
## Mean           29     37     49     53
## 3rd Qu.        58     62     83     81
## Max.           99     99    100    100
## sd             32     29     34     30
## best mstop with CV from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           10     13     10     11
## 1st Qu.        29     33     32     35
## Median         42     50     53     57
## Mean           48     53     53     57
## 3rd Qu.        68     74     75     81
## Max.           99    100     99    100
## sd             25     24     24     26
## nvar from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           1.0     1.0     1.0     1.0
## 1st Qu.         1.0     1.0     1.0     1.0

```

```

## Median      2.0      2.0      2.0      2.0
## Mean        1.8      1.8      2.1      2.4
## 3rd Qu.     2.0      2.0      3.0      3.0
## Max.        7.0      5.0      7.0      8.0
## sd          1.1      1.0      1.3      1.5
##
## bst family LogitBoost, nu=0.1
## best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449 0.0562 0.0449 0.0506
## 1st Qu.    0.0730 0.0730 0.0787 0.1053
## Median    0.0843 0.0843 0.1124 0.1433
## Mean      0.0824 0.0896 0.1146 0.1487
## 3rd Qu.   0.0899 0.1067 0.1419 0.1798
## Max.      0.1461 0.1517 0.2303 0.3258
## sd        0.0152 0.0208 0.0419 0.0606
## CV based misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506 0.0562 0.0618 0.0562
## 1st Qu.   0.0843 0.0885 0.0997 0.1222
## Median    0.0899 0.1039 0.1348 0.1657
## Mean      0.0907 0.1027 0.1341 0.1735
## 3rd Qu.   0.1011 0.1180 0.1587 0.2107
## Max.      0.1573 0.1573 0.2697 0.3876
## sd        0.0145 0.0218 0.0444 0.0703
## best mstop with best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1      4      1      5
## 1st Qu.      3     14     26     33
## Median      44     37     48     69
## Mean        42     42     50     61
## 3rd Qu.     72     60     80     92
## Max.       100     99    100    100
## sd          34     28     30     31
## best mstop with CV from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        10     15     16     10
## 1st Qu.     28     33     36     39
## Median     55     52     58     60
## Mean       53     53     58     60
## 3rd Qu.    74     70     76     84
## Max.     100     99     98    100
## sd         26     23     24     26
## nvar from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     1.0     1.0     1.0
## 1st Qu.     1.0     2.0     3.0     4.0
## Median     2.0     3.0     5.0     7.0
## Mean       2.5     3.6     5.3     6.9

```

```

## 3rd Qu.      3.0      5.0      8.0      9.2
## Max.        9.0     10.0     13.0     15.0
## sd          1.8      2.2      2.9      3.6
##
## rbst family DlogitBoost, s=0.8, nu=0.1
##
## best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0506  0.0506  0.0562
## 1st Qu.   0.0787 0.0730  0.0843  0.1011
## Median    0.0899 0.0899  0.1124  0.1461
## Mean      0.0869 0.0907  0.1197  0.1556
## 3rd Qu.   0.0955 0.1067  0.1461  0.2079
## Max.      0.1685 0.1685  0.2978  0.3652
## sd        0.0162 0.0240  0.0503  0.0730
## CV based misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0562  0.0562  0.0618
## 1st Qu.   0.0787 0.0787  0.0899  0.1067
## Median    0.0899 0.0955  0.1124  0.1601
## Mean      0.0898 0.0997  0.1316  0.1708
## 3rd Qu.   0.1011 0.1180  0.1573  0.2317
## Max.      0.1798 0.1966  0.3202  0.3708
## sd        0.0163 0.0265  0.0562  0.0787
## best mstop with best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1         6         1         7
## 1st Qu.      1        19        34        41
## Median       1        52        62        70
## Mean        19        51        59        66
## 3rd Qu.     31        78        89        92
## Max.       100       100       100       100
## sd          31        31        32        28
## best mstop with CV from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        10        18        15        11
## 1st Qu.     37        47        52        52
## Median      56        64        68        74
## Mean       56        63        68        70
## 3rd Qu.    74        78        84        90
## Max.      99       100       100       100
## sd         24        21        21        24
## nvar from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.     1.0      1.0      1.0      2.0
## Median     1.0      1.0      2.0      3.0
## Mean       1.4      1.7      2.5      2.9
## 3rd Qu.    2.0      2.0      3.0      4.0

```

```

## Max.      5.0      5.0      8.0      8.0
## sd        0.8      0.9      1.5      1.4
##
## bst family HingeBoost, nu=0.1
## best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0449 0.0506 0.0449 0.0562
## 1st Qu.  0.0730 0.0730 0.0772 0.0787
## Median   0.0787 0.0843 0.0843 0.1011
## Mean     0.0788 0.0839 0.0954 0.1130
## 3rd Qu.  0.0843 0.0899 0.1081 0.1362
## Max.     0.1292 0.1348 0.2079 0.2528
## sd       0.0137 0.0177 0.0296 0.0414
## CV based misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0562 0.0618 0.0562 0.0674
## 1st Qu.  0.0843 0.0843 0.0899 0.1011
## Median   0.0955 0.0955 0.1011 0.1264
## Mean     0.0931 0.0996 0.1165 0.1379
## 3rd Qu.  0.1011 0.1067 0.1348 0.1685
## Max.     0.1629 0.1910 0.3427 0.2809
## sd       0.0158 0.0230 0.0414 0.0492
## best mstop with best misclassification error from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1       1       1       1
## 1st Qu.   18      18      15      21
## Median    24      23      25      44
## Mean      27      28      36      48
## 3rd Qu.   32      30      56      76
## Max.      89     100     100     99
## sd        17      23      30      32
## best mstop with CV from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      10      10      10      10
## 1st Qu.   10      16      19      27
## Median    23      26      30      42
## Mean      28      35      37      50
## 3rd Qu.   35      54      51      74
## Max.      99      95     100     100
## sd        20      25      23      28
## nvar from bst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      1.0     1.0     1.0     1.0
## 1st Qu.   1.0     1.0     2.0     3.0
## Median    3.0     4.0     5.0    11.0
## Mean      6.6     8.9     8.8    12.4
## 3rd Qu.   9.0    15.0    14.2    20.0
## Max.     36.0    30.0    32.0    36.0
## sd        7.8     9.1     8.4     9.8

```



```

##
## rbst family THingeBoost, s=-0.5, nu=0.1
##
## best misclassification error from rbst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0506 0.0506  0.0449  0.0506
## 1st Qu.  0.0730 0.0730  0.0772  0.0787
## Median   0.0787 0.0815  0.0843  0.0871
## Mean     0.0790 0.0817  0.0906  0.1007
## 3rd Qu.  0.0843 0.0899  0.1011  0.1236
## Max.     0.1348 0.1348  0.2022  0.1910
## sd       0.0135 0.0147  0.0253  0.0336
## CV based misclassification error from rbst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0562 0.0618  0.0618  0.0618
## 1st Qu.  0.0843 0.0843  0.0899  0.0941
## Median   0.0955 0.0955  0.1011  0.1067
## Mean     0.0937 0.0957  0.1076  0.1197
## 3rd Qu.  0.1011 0.1011  0.1180  0.1362
## Max.     0.1461 0.1517  0.2022  0.2640
## sd       0.0149 0.0173  0.0295  0.0411
## best mstop with best misclassification error from rbst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         1      1      1      1
## 1st Qu.       17     17     16     20
## Median        23     23     24     29
## Mean          25     28     35     38
## 3rd Qu.       29     30     50     59
## Max.          93     99    100     99
## sd            18     22     30     27
## best mstop with CV from rbst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         10     10     10     10
## 1st Qu.       10     18     19     22
## Median        23     28     34     44
## Mean          33     38     41     48
## 3rd Qu.       45     56     59     68
## Max.          99    100     96     99
## sd            25     27     26     28
## nvar from rbst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         1.0    1.0    1.0    1.0
## 1st Qu.       1.0    1.8    2.0    2.0
## Median        4.0    4.0    7.5   11.0
## Mean          8.4    9.4    9.6   11.2
## 3rd Qu.      13.2   16.0   14.0   18.0
## Max.         33.0   34.0   30.0   33.0
## sd           9.0    9.5    8.7    9.0
##

```

```

## bst family AdaBoost, nu=0.1
## best misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0449 0.0562  0.0449  0.0506
## 1st Qu. 0.0716 0.0730  0.0787  0.1011
## Median  0.0787 0.0843  0.1039  0.1264
## Mean    0.0797 0.0870  0.1073  0.1310
## 3rd Qu. 0.0857 0.0955  0.1292  0.1573
## Max.    0.1461 0.1461  0.2079  0.3034
## sd      0.0147 0.0198  0.0339  0.0443
## CV based misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.    0.0618 0.0562  0.0618  0.0730
## 1st Qu. 0.0787 0.0899  0.1053  0.1222
## Median  0.0955 0.1067  0.1348  0.1517
## Mean    0.0923 0.1072  0.1333  0.1585
## 3rd Qu. 0.1011 0.1236  0.1573  0.1910
## Max.    0.1573 0.1573  0.2360  0.3146
## sd      0.0156 0.0237  0.0407  0.0473
## best mstop with best misclassification error from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         1      1      1      2
## 1st Qu.       9      4      7      9
## Median      21     12     14     23
## Mean        25     17     25     39
## 3rd Qu.     35     26     32     72
## Max.        93     98    100    100
## sd          21     17     26     33
## best mstop with CV from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         10     10     10     10
## 1st Qu.      13     12     14     15
## Median      21     18     24     34
## Mean        28     33     33     44
## 3rd Qu.     35     49     46     70
## Max.        92     99    100     99
## sd          20     28     26     30
## nvar from bst
##      cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     1.0     1.0     1.0
## 1st Qu.       1.0     2.0     3.0     4.8
## Median       3.0     4.0     7.0    11.0
## Mean        4.1     6.6     8.6    11.7
## 3rd Qu.      6.0    10.2    12.0    17.0
## Max.       17.0    21.0    25.0    27.0
## sd          3.8     5.6     6.0     7.0
##
## rbst family TAdaBoost, s=-0.2, nu=0.1
##

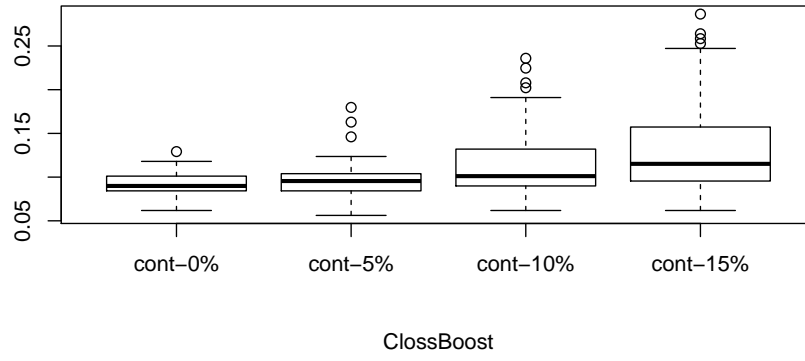
```

```

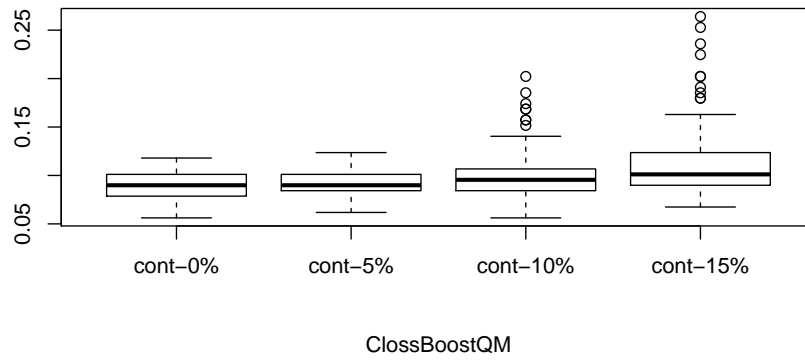
## best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0562  0.0562
## 1st Qu.   0.0787 0.0787  0.0899  0.0955
## Median    0.0843 0.0899  0.1039  0.1236
## Mean      0.0856 0.0937  0.1059  0.1258
## 3rd Qu.   0.0955 0.1011  0.1180  0.1517
## Max.      0.1348 0.1573  0.1910  0.2360
## sd        0.0150 0.0195  0.0236  0.0380
## CV based misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0730  0.0674
## 1st Qu.   0.0843 0.0955  0.1067  0.1236
## Median    0.0955 0.1067  0.1236  0.1461
## Mean      0.0978 0.1097  0.1289  0.1511
## 3rd Qu.   0.1067 0.1194  0.1419  0.1798
## Max.      0.1742 0.2472  0.2584  0.2921
## sd        0.0198 0.0269  0.0344  0.0425
## best mstop with best misclassification error from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           1      2      1      1
## 1st Qu.        14     15     14     12
## Median         36     37     40     39
## Mean           42     39     44     42
## 3rd Qu.        70     60     70     69
## Max.           97     98    100     94
## sd             31     28     31     30
## best mstop with CV from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           10     10     10     10
## 1st Qu.        14     15     19     13
## Median         28     36     38     30
## Mean           38     43     42     38
## 3rd Qu.        62     70     61     64
## Max.          100    100    100     98
## sd             28     28     26     28
## nvar from rbst
##          cont-0% cont-5% cont-10% cont-15%
## Min.           1.0    1.0    1.0    1.0
## 1st Qu.         2.0    3.0    4.0    3.0
## Median          3.0    6.0    8.0    8.0
## Mean            4.5    7.6    8.8    8.7
## 3rd Qu.         7.0   11.0   12.0   12.2
## Max.           18.0   22.0   26.0   28.0
## sd              3.7    5.4    5.7    6.2
print(proc.time() - ptm)
##          user      system elapsed
## 49684.357  3190.518 53841.009

```

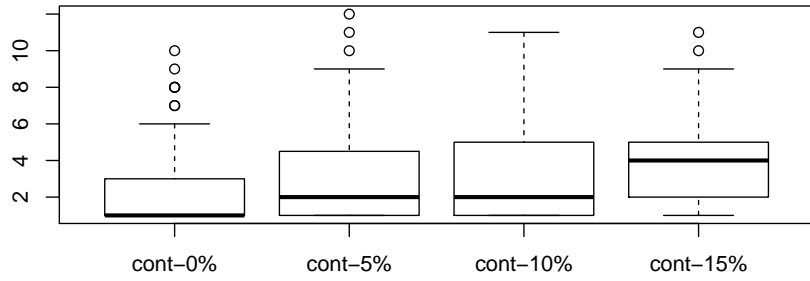
Misclassification error



Misclassification error

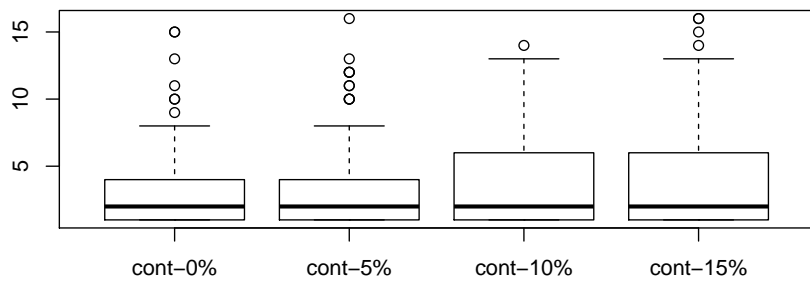


No. variables



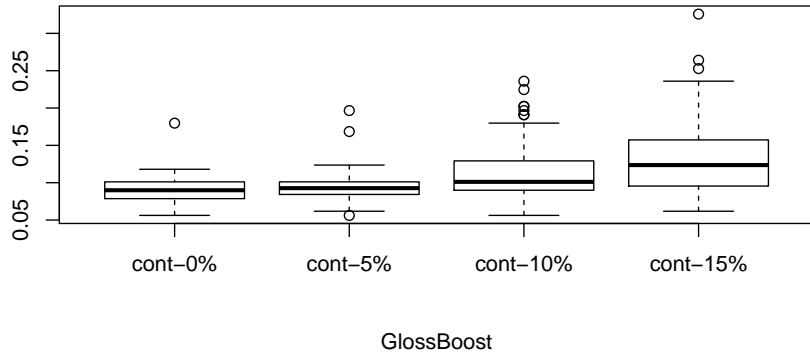
ClossBoost

No. variables

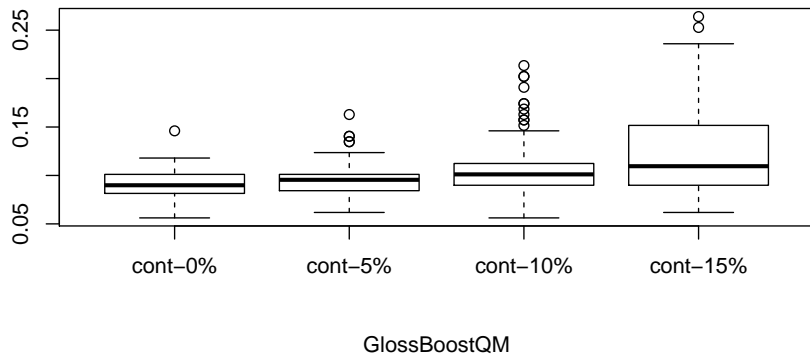


ClossBoostQM

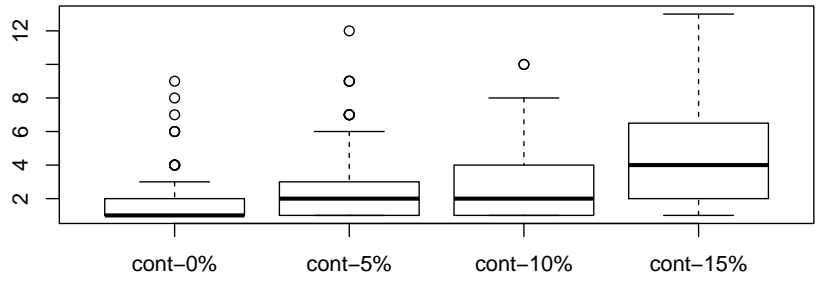
Misclassification error



Misclassification error

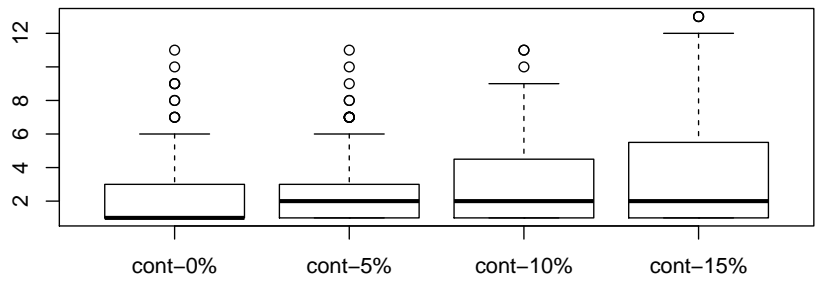


No. variables



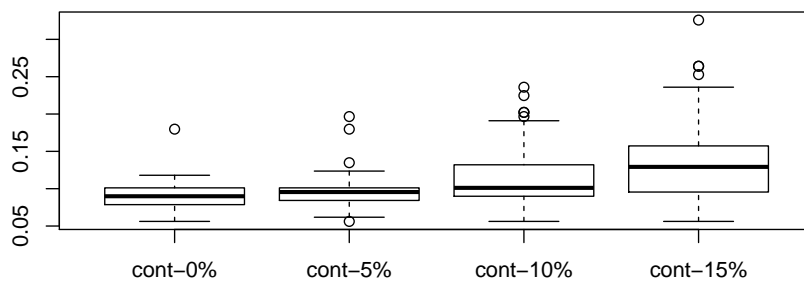
GlossBoost

No. variables



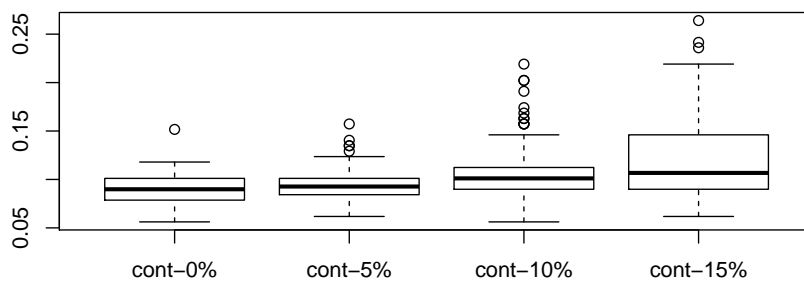
GlossBoostQM

Misclassification error



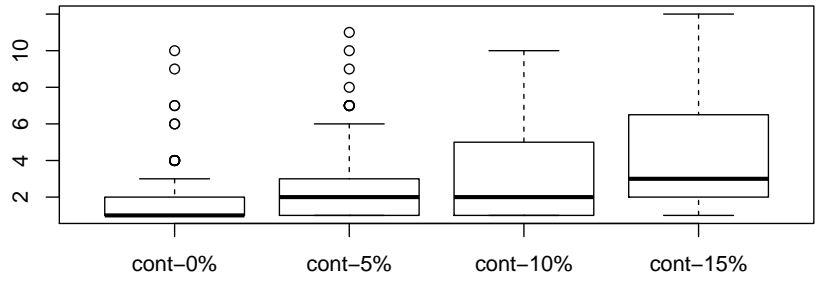
QlossBoost

Misclassification error



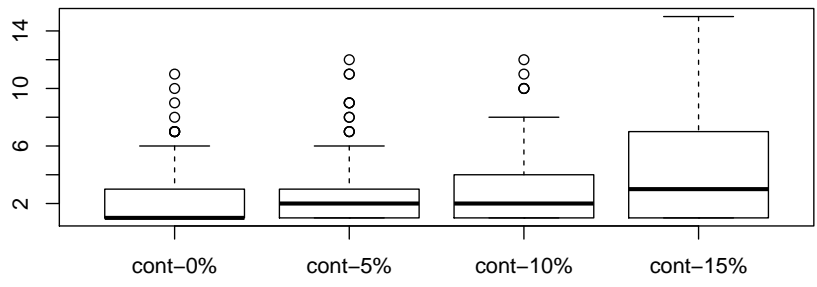
QlossBoostQM

No. variables



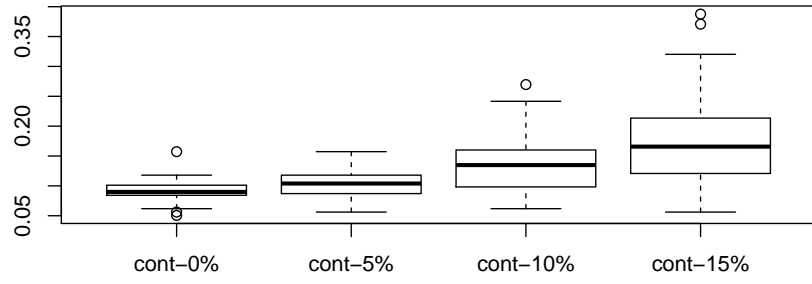
QlossBoost

No. variables



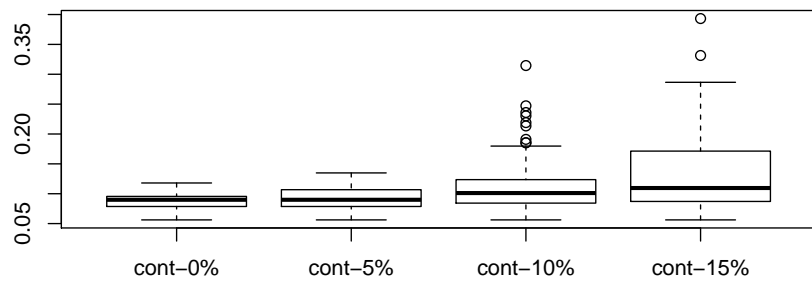
QlossBoostQM

Misclassification error



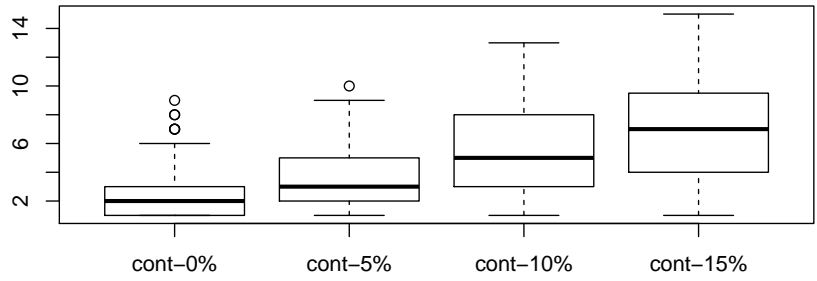
LogitBoost

Misclassification error



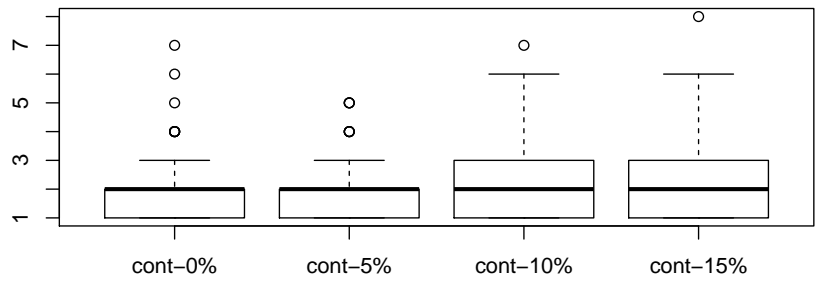
TLogitBoost

No. variables



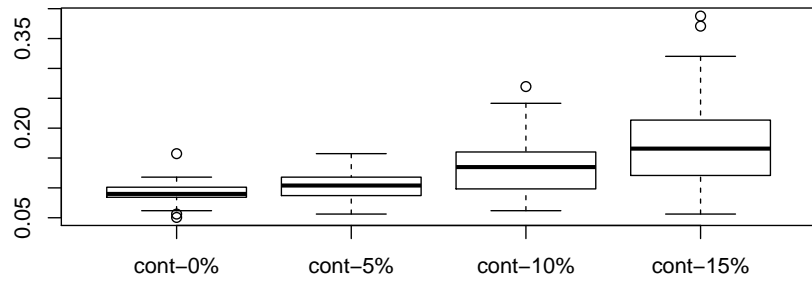
LogitBoost

No. variables



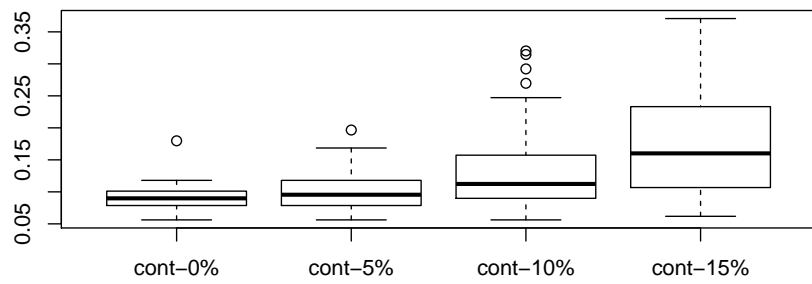
TLogitBoost

Misclassification error



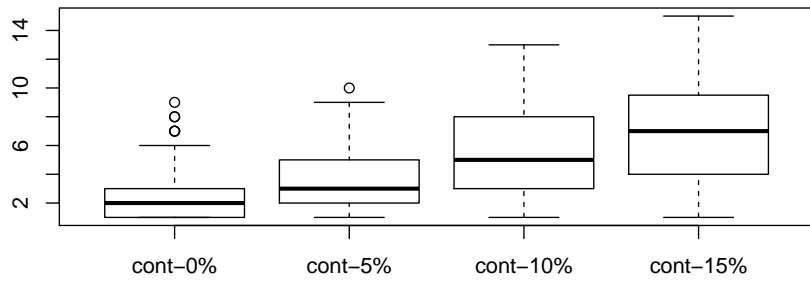
LogitBoost

Misclassification error



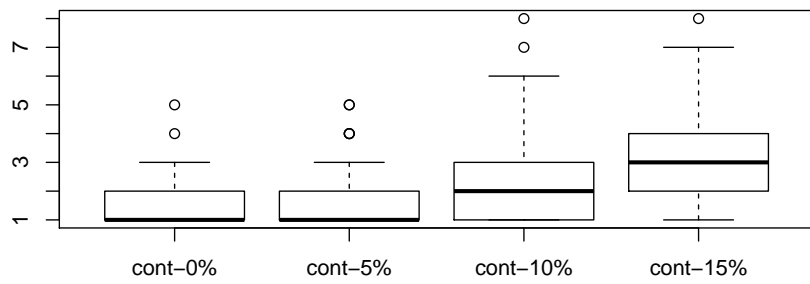
DlogitBoost

No. variables



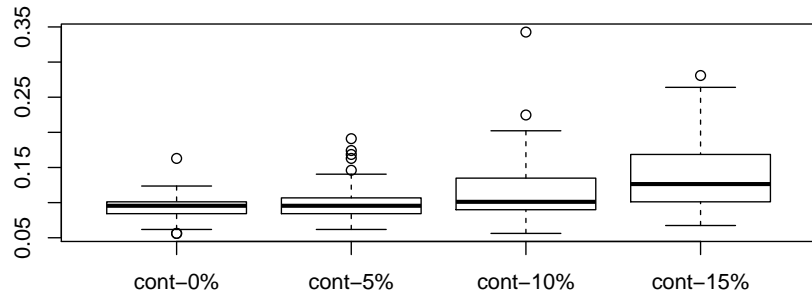
LogitBoost

No. variables



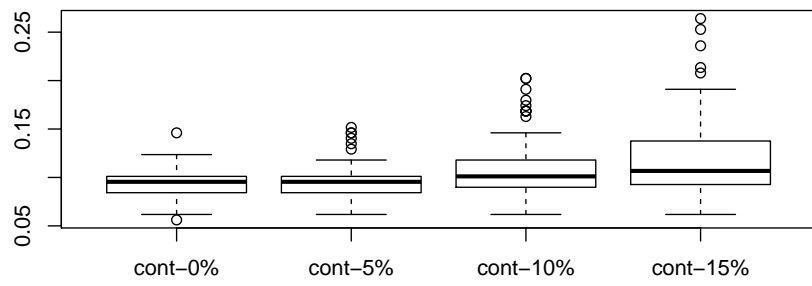
DlogitBoost

Misclassification error



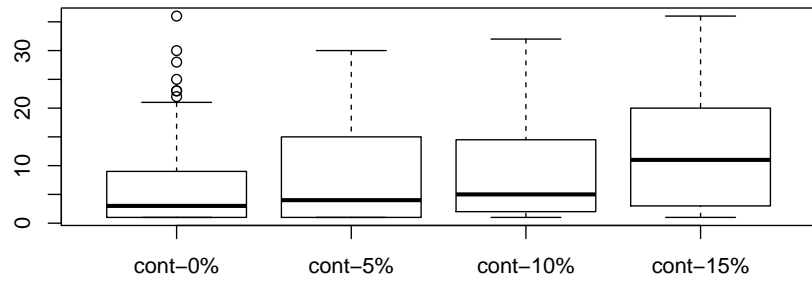
HingeBoost

Misclassification error



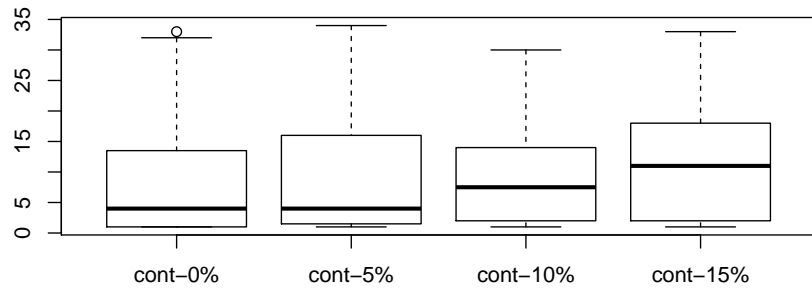
THingeBoost

No. variables



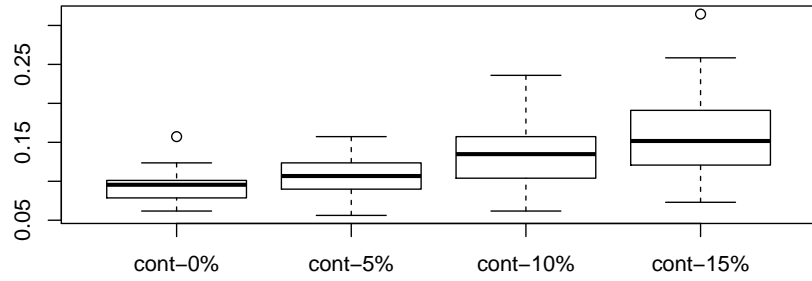
HingeBoost

No. variables



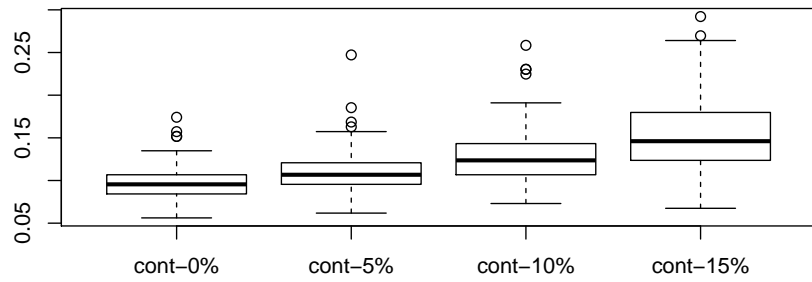
THingeBoost

Misclassification error



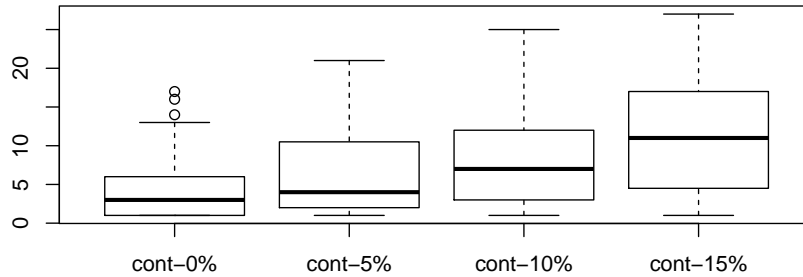
AdaBoost

Misclassification error



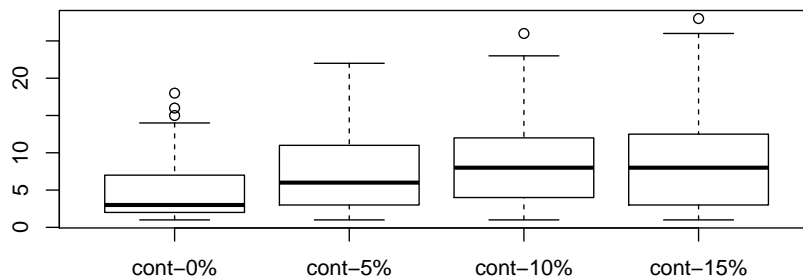
TAdaBoost

No. variables



AdaBoost

No. variables



TAdaBoost

```
sessionInfo()
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.1 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
## [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8    LC_NAME=C
## [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats    graphics  grDevices  utils      datasets
## [6] methods  base
```

```
##
## other attached packages:
## [1] bst_0.3-15    gbm_2.1.4      gdata_2.18.0  knitr_1.21
##
## loaded via a namespace (and not attached):
## [1] codetools_0.2-16  lattice_0.20-38  gtools_3.8.1
## [4] foreach_1.4.4    grid_3.5.2       gtable_0.2.0
## [7] formatR_1.5      magrittr_1.5     evaluate_0.12
## [10] stringi_1.2.4    doParallel_1.0.14  rpart_4.1-13
## [13] Matrix_1.2-15    splines_3.5.2     iterators_1.0.10
## [16] tools_3.5.2      stringr_1.3.1     parallel_3.5.2
## [19] xfun_0.4         survival_2.43-3   compiler_3.5.2
## [22] gridExtra_2.3
```

References

- Zhu Wang. Robust boosting with truncated loss functions. *Electronic Journal of Statistics*, 12(1):599–650, 2018a. doi: 10.1214/18-EJS1404.
- Zhu Wang. Quadratic majorization for nonconvex loss with applications to the boosting algorithm. *Journal of Computational and Graphical Statistics*, 2018b. doi: 10.1080/10618600.2018.1424635.