Machine Learning: HAR Project Classifier

Sergio Quadros

2015/07/26

```
library(knitr);library(rmarkdown);library(grid);library(caret);library(xtable)
library(ggplot2);library(png);library(gbm);library(e1071);library(randomForest)
library(dplyr);library(magrittr);library(tidyr)
```

Data

The data for this project come from this source at Human Activity Recognition - **HAR** - group in PUC-Rio, Rio de Janeiro-RJ, Brazil.

```
set.seed(133162)
if ( ! file.exists("pml-training.csv") ) {
        download.file("http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
              ,destfile="pml-training.csv")
        }
if ( ! file.exists("pml-testing.csv") ) {
        download.file("http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
              ,destfile="pml-testing.csv")
        }
trainingOriginal <- tbl_df(read.csv("pml-training.csv", header = TRUE, sep = ",",
                                    na.strings=c("NA","NaN"," ","","#DIV/0!")))
testingOriginal <- tbl_df(read.csv("pml-testing.csv", header = TRUE, sep = ",",
                                   na.strings=c("NA","NaN"," ","","#DIV/0!")))
```

Our data was divided in two sets by Coursera's staff with 160 variables including the outcome one, *classe* with five levels - A, B, C, D and E for activities' types - *sitting, standing, standing up, sitting down and walking* respectively:

- **Traing set** with 19622 observations from link
- **Testing set** with 19622 observations from link

The document was produced with **R version 3.1.2 at a i686-pc-linux-gnu (32-bit) Ubuntu** and for downloading I supressed *'s'* from *'https'* protocol. I didn't find a *readme* file for data and I used this references [1], [2] and [3].

I selected the predictors of interest by nearZeroVar function getting 123 variables, then by name - roll, accelaration, fitch, gyros and classe - and the last choice was predictors' set with absolute correlation index greater than 75%, resulting in 46 predictors.

```
set.seed(133162)
nzv  <-  nearZeroVar (trainingOriginal,  saveMetrics =  FALSE )
training <- select(trainingOriginal,-nzv)
training  %<>%  select(matches("(gy|rol|acc|pi|classe)"))
testing <- select(testingOriginal,-nzv)
testing  %<>%  select(matches("(gy|rol|acc|pi|classe)"))
final <- dim(training)[2]
```

```
matrixCor <- abs(cor(training[,-final],use = "pairwise.complete.obs"))
diag(matrixCor) <- 0
index_cor <- unique(which(matrixCor > 0.75,arr.ind=T)[,1])
training %<>% select(c(index_cor,final))
testing %<>% select(c(index_cor,final))
```

I made the predicting with *Random Forest* algorithm in *caret* package with K-fold equals 10 and repeated five times; the optimistic estimate of accuracy with cross validation into training set is 0.75 $$ 0.07 (errors' range between 19.0% and 26.8%), see below:

```
set.seed(133162)
fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
modFit <- train(classe ~ ., method="rf",data=training,prox=TRUE,
                trControl =fitControl,importance=TRUE)
print(modFit)
```

```
## Random Forest
##
## 19622 samples
##    46 predictors
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 287, 285, 287, 286, 288, 286, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa      Accuracy SD  Kappa SD
##    2    0.7539310  0.6879125  0.07293262   0.09245938
##   24    0.7337049  0.6625789  0.06995949   0.08932506
##   46    0.7152277  0.6392582  0.07448360   0.09522146
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
```

```
print(modFit$finalModel)
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, importance = TRUE,      proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 23.27%
## Confusion matrix:
##    A  B  C  D  E class.error
## A 68  7  4  4  1   0.1904762
## B 14 46  3  1  0   0.2812500
## C  7  5 41  3  0   0.2678571
## D  7  0  4 45  2   0.2241379
## E  5  3  2  2 44   0.2142857
```

```
# big <- varImp(modFit) # 4 most important predictors for outcome
```

We present the 4 most important predictors by using *varImp* function for the 5 types of outcome in a pairs plot:

```
# dimensional problem for estimating 'out' accuracy

# pred <- predict(modFit,testing)
# testing$predRight <- pred==testing$classe
# table(pred,testing$class)
featurePlot(x = training[,c(42,4,22,14)],
            y = training$classe,
            plot = "pairs",
            ## Add a key at the top
            auto.key = list(columns = 3))
```

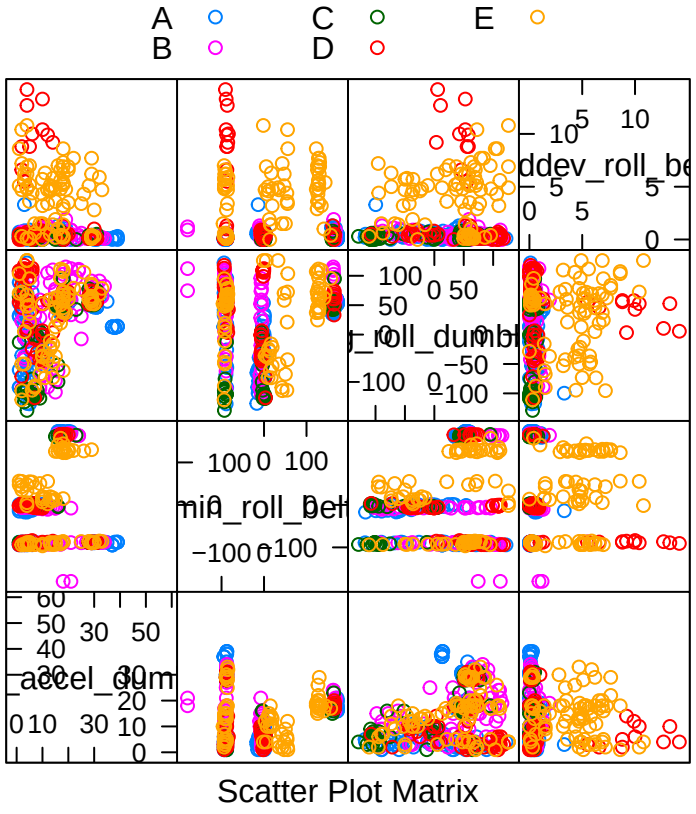

Figure 1:

```
colnames(training[,c(42,4,22,14)])
```

```
## [1] "total_accel_dumbbell" "min_roll_belt"        "avg_roll_dumbbell"
## [4] "stddev_roll_belt"
```

The observed classifier accuracy is 99.4% for authors in [1], [2] and [3], our estimate by cross validation is 75% $$ 7% into training set.

References

[1] Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. DOI: 10.1007/978-3-642-34459-6_6 at http://groupware.les.inf.puc-rio.br/public/papers/2012.Ugulino.WearableComputing.HAR.Classifier.RIBBON.pdf

[2] Ugulino, W.; Ferreira, M.; Velloso, E.; Fuks, H. Virtual Caregiver: Colaboração de Parentes no Acompanhamento de Idosos. Anais do SBSC 2012, IX Simpósio Brasileiro de Sistemas Colaborativos , pp. 43-48. São Paulo, SP: IEEE, 2012. ISBN 978-0-7695-4890-6 at http://groupware.les.inf.puc-rio.br/public/papers/2012.SBSC.Ugulino.VirtualCaregiver.pdf.

[3] Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th Augmented Human (AH) International Conference in cooperation with ACM SIGCHI (Augmented Human'13) . Stuttgart, Germany: ACM SIGCHI, 2013 at http://pubs.acs.org/doi/abs/10.1021/acscentsci.5b00148.