Prediction Assignment Writeup

Atoosa Madadkar

11/26/2021

Introduction

Human Activity Recognition - HAR - has emerged as a key research area in the last years and is gaining increasing attention by the pervasive computing research community, especially for the development of context-aware systems.

This human activity recognition research has traditionally focused on discriminating between different activities, i.e. to predict "which" activity was performed at a specific point in time.

Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions which are explained in the table below.

Class Category	Descriptions
Class A	exactly according to the specification
Class B	throwing the elbows to the front
Class C	lifting the dumbbell only halfway
Class D	lowering the dumbbell only halfway
Class E	and throwing the hips to the front

Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes. Participants were supervised by an experienced weight lifter to make sure the execution complied to the manner they were supposed to simulate. The exercises were performed by six male participants aged between 20-28 years, with little weight lifting experience. The puporse of this data processing is to introduce a classification algorithm which can accurately predict the exercise type.

Data Processing and Data Cleaning

The data consists of 160 variables, with "classe" as the exercise type and 19622 observations from six participants. 71 variables consisted of 98% of missing data, the index and time related feature were removed. Furthermore, 59 features which had near zero variablity, were put aside. With the 55 remained variables, the variables with the suffixes _x, _y, _z were excluded and the ones with total measurements remained the final model. Ultimately, 17 variables were considered for further analysis.

```
#Find columns with missing values and remove them
col_names <- names(pml[,!sapply(pml, function(x) sum(is.na(x)) > 0)])
keep_cols <- names(pml) %in% col_names;
pml <- pml[,keep_cols] #67 variables are thrown out

#Exclude time and index variables
pml <- pml %>% select(-c("X", "num_window", "cvtd_timestamp", "raw_timestamp_part_1", "raw_timestamp_pa

#Find near zero values and remove them
nzv <- nearZeroVar(pml, saveMetrics = TRUE)
```

```
pml <- pml[,nzv$nzv == FALSE] #59 variables are gone

finalCols <- !grepl("_x|_y|_z", names(pml))
pml <- pml[finalCols]
```

Cross-Validation

The properly cleaned data was partitioned into training and testing datasets. The training dataset was set to include 70% of the whole data. The testing dataset will be used in the following section, after the finalized model is introduced. The characteristic of the training dataset is shown in the following table.

```
set.seed(2223)
inTrain <- createDataPartition(pml$classe, p=.7, list=FALSE)
training <- pml[inTrain,]
testing <- pml[-inTrain,]

skim(training)[,1:11]
```

Table 2: Data summary

Name	training
Number of rows	13737
Number of columns	18
Column type frequency:	
character	2
numeric	16
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
user_name	0	1	5	8	0	6	0
classe	0	1	1	1	0	5	0

Variable type: numeric

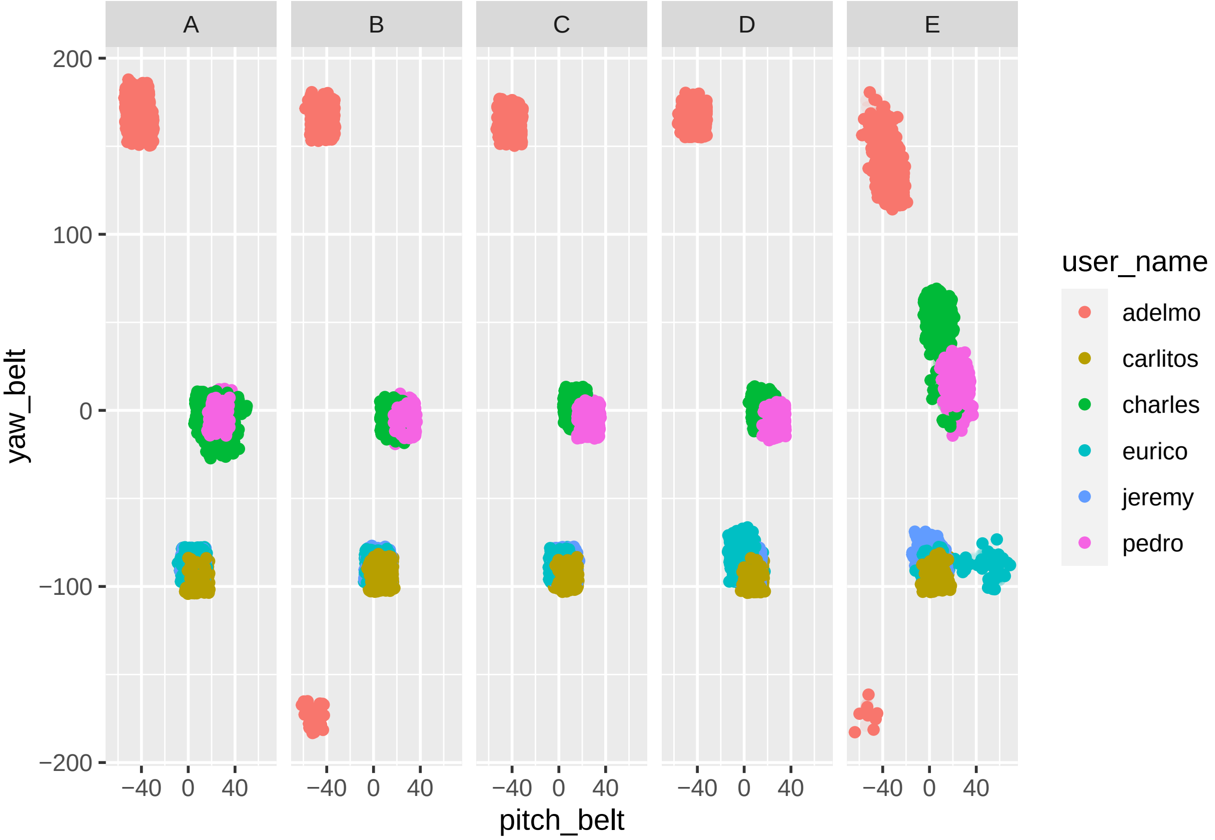
skim_variable	n_missing	complete_rate	mean	sd
roll_belt	0	1	64.39	62.78
pitch_belt	0	1	0.29	22.35
yaw_belt	0	1	-11.13	95.23
total_accel_belt	0	1	11.32	7.75
roll_arm	0	1	17.64	72.52
pitch_arm	0	1	-4.35	30.64
yaw_arm	0	1	-1.02	71.41
total_accel_arm	0	1	25.55	10.50
roll_dumbbell	0	1	23.83	69.66
pitch_dumbbell	0	1	-10.94	37.02

skim_variable	n_missing	complete_rate	mean	sd
yaw_dumbbell	0	1	1.65	82.76
total_accel_dumbbell	0	1	13.76	10.26
roll_forearm	0	1	33.13	108.19
pitch_forearm	0	1	10.66	28.04
yaw_forearm	0	1	19.69	103.17
total_accel_forearm	0	1	34.82	10.06

Exploratory Data Analysis

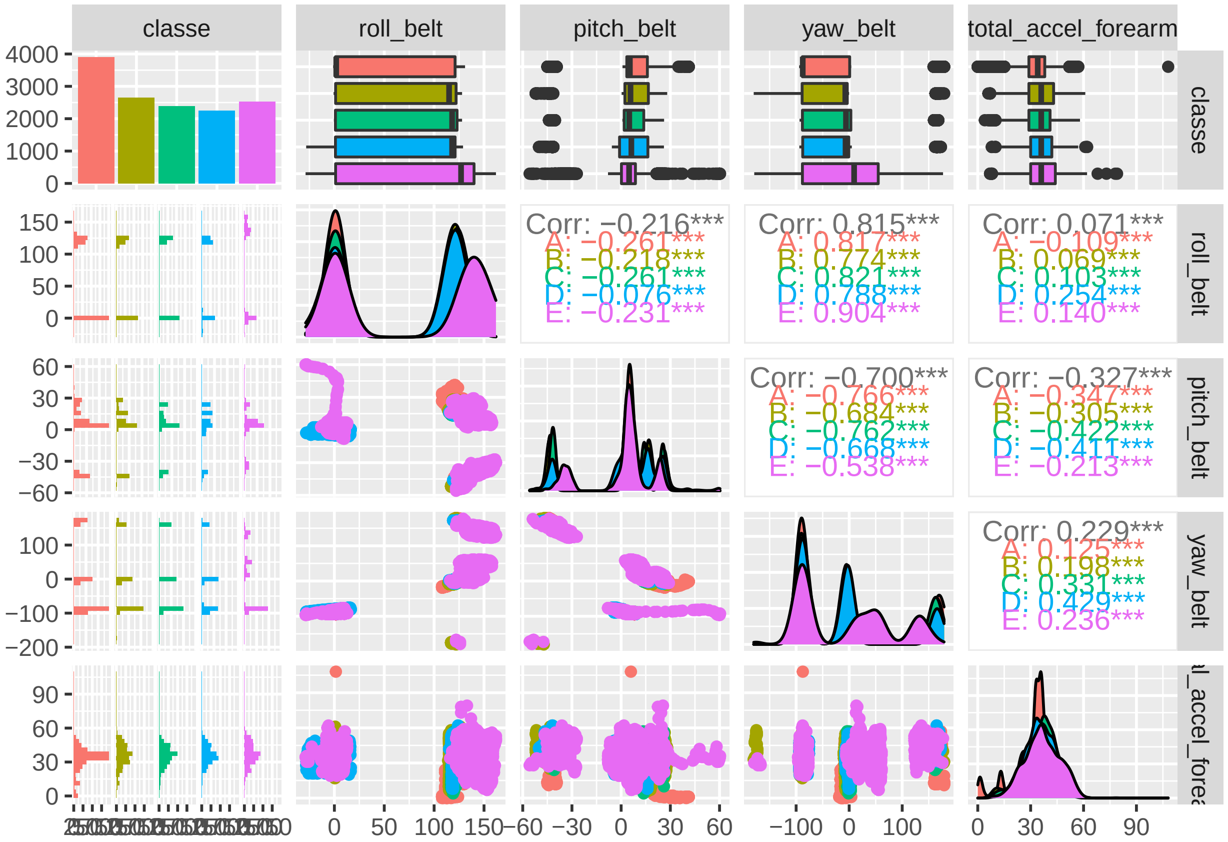
Various relationships were explored, out of which the **roll_belt**, **pitch_bell**, and **yaw_belt** showed the most variability with exercise type and the individual. The graph below shows this kind of variability in **pitch_belt** and **yaw_belt** in classe and individual's categories.

```r
ggplot(training, aes(pitch_belt, yaw_belt, color = user_name)) +
  geom_point(alpha=.1) + geom_jitter(width=10, height=10) +
  facet_grid(~ classe)
```



```r
pairsvar <- select(training, c("classe", "roll_belt",
                               "pitch_belt", "yaw_belt", "total_accel_forearm"))
ggpairs(pairsvar, aes(colour=classe))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Machine Learning Algorithm

The finalized training dataset was used in various machine learning algorithms including Decision Trees, Boosting, and K-nearest neighbor. None of them had a better performance than Random Forest in terms of accuracy, sensitivity and specificity. For the cross validation, the fitted model was implemented in the test outcomes, and the result was impressive.

```
set.seed(5742)
rfFit <- randomForest(as.factor(classe) ~ ., data=training)
rf_prediction <- predict(rfFit, newdata=testing)
confusionMatrix(rf_prediction, as.factor(testing$classe))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1671   10    0    1    0
##          B    0 1109    2    0    0
##          C    1   19 1018    8    7
##          D    2    0    6  955    3
##          E    0    1    0    0 1072
##
## Overall Statistics
##
##                Accuracy : 0.9898
##                  95% CI : (0.9869, 0.9922)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
```

```
##
##                  Kappa : 0.9871
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9982   0.9737   0.9922   0.9907   0.9908
## Specificity            0.9974   0.9996   0.9928   0.9978   0.9998
## Pos Pred Value         0.9935   0.9982   0.9668   0.9886   0.9991
## Neg Pred Value         0.9993   0.9937   0.9983   0.9982   0.9979
## Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate         0.2839   0.1884   0.1730   0.1623   0.1822
## Detection Prevalence   0.2858   0.1888   0.1789   0.1641   0.1823
## Balanced Accuracy      0.9978   0.9866   0.9925   0.9942   0.9953
```

There were 500 trees built in the model and the model tried 4 different variables at each split. Based on the Mean Decrease Gini of our 17 variables, **roll_belt**, **yaw_belt**, and **pitch_belt** are the most important variables in the model since they contribute the most to the homogeneity of the nodes and leaves. The average out of bound (OOB) error rate is 0.012 in this model, calculated by mean(rfFit$err.rate[,1]).

Reference

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013.