Qualitative Activity Recognition

Summary

In this document a machine learning algorithim will be trained in order to find the best fit for classifying qualitative activity recognition of wheight lifting exercises. The data is included in Weight Lifting Exercise Dataset [1] and was taken from 9 degrees of freedom IMUs mounted on glove, armband, lumbar belt and dumbbell. Weight lifting exercise was performed correctly and with a set of common mistakes so classifyng correct or four of mistaked performances is our goal.

Loading Data

Data is divided in a training and a test set.

```r
orTraining <- read.csv("data/pml-training.csv")
orTesting <- read.csv("data/pml-testing.csv")
dim(orTraining)
```

```
## [1] 19622   160
```

```r
dim(orTesting)
```

```
## [1]  20 160
```

Up to 160 columns are included in both training and testing data so some previous tidyng and variable preprocessing should be done using large datasets techniques.

First of all, column class wil be verified:

```r
clases <- sapply(orTraining[,1:160], class)
table(clases)
```

```
## clases
##  factor integer numeric
##      37      35      88
```

Data from IMUs are in columns from 8 to 159 so our train and test features are reduced to these columns. Also, many factor columns are found so I will coerce this columns to numeric. NAs will be cleared and low variance features removed.

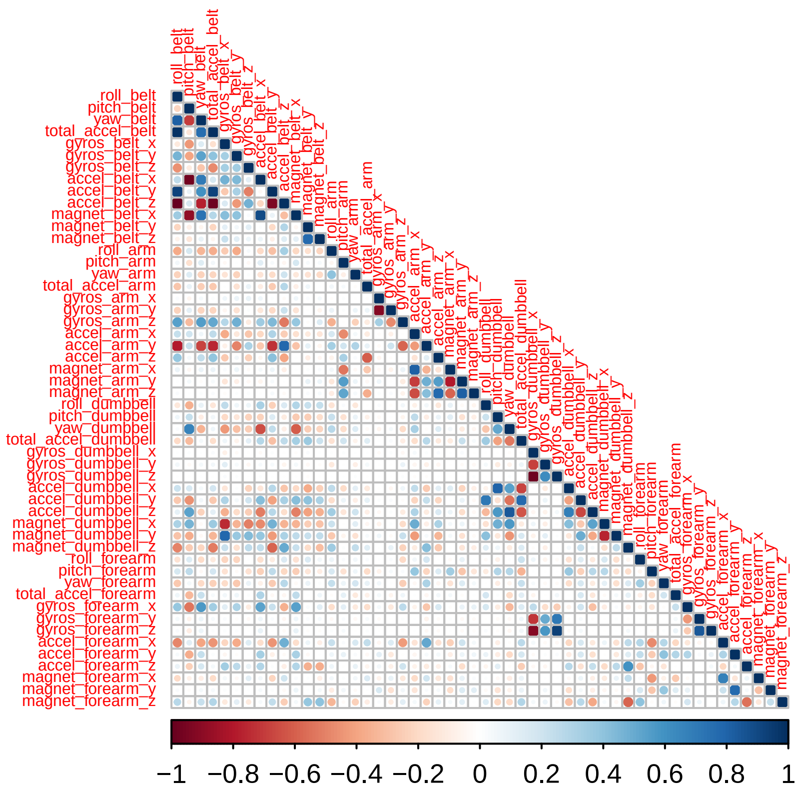
```r
# subset columns for features
trainFeatures <- orTraining[,8:159]
testFeatures <- orTesting[,8:159]
# coerce factors to numeric
for(i in seq(1, dim(trainFeatures)[2])){
        if(is.factor(trainFeatures[,i])){
                trainFeatures[,i] <- as.numeric(as.character(trainFeatures[,i]))
                testFeatures[,i] <- as.numeric(as.character(testFeatures[,i]))
        }
```

```
}
# NAs to zero
trainFeatures[is.na(trainFeatures)] <- 0
testFeatures[is.na(testFeatures)] <- 0
# near zero variance features remotion
nzv <- nearZeroVar(trainFeatures)
trainFeatures <- trainFeatures[,-nzv]
testFeatures <- testFeatures[,-nzv]
```

After removing near zero variance features, only 52 are included in analysis.

Let's have a look to correlation matrix for these features.

```
corMatrix <- cor(trainFeatures)
corrplot(corMatrix, method="circle", tl.cex = .5, type = "lower")
```



There are some high correlated features so these wil be removed with default cutoff of .9.

```
corFeatures <- findCorrelation(corMatrix)
testFeatures <- testFeatures[, -corFeatures]
trainFeatures <- trainFeatures[, -corFeatures]
```

Finally, only 45 features are included in analysis. Classe colum is binded to train set.

```r
train <- cbind(classe = orTraining$classe, trainFeatures)
```

Prediction model

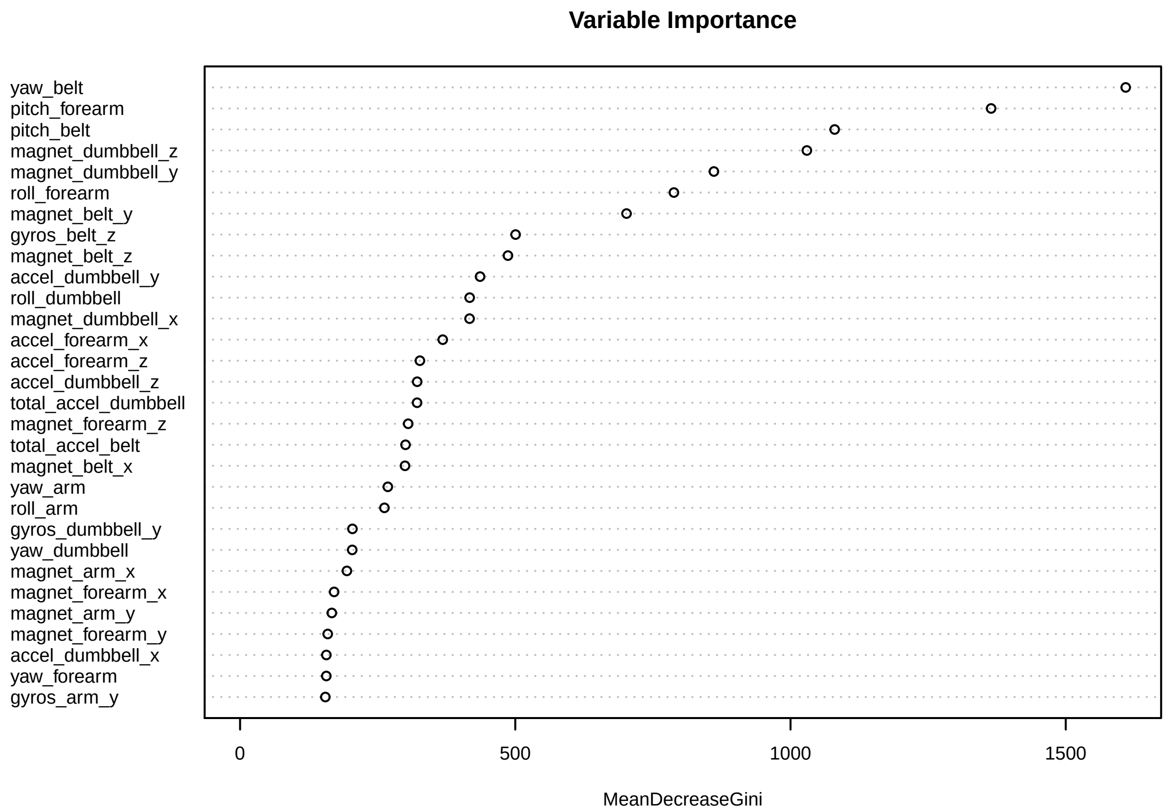
As seen in E Velloso, A Bulling, H Gellersen, W Ugulino, H Fuks [1], because of the characteristic noise in the sensor data Random Forest approach is used. A cross validation, with five folds is defined via trainControl.

```r
set.seed(2302)
trCtrl <- trainControl(method = "cv", number = 5)
rfModel <- train(classe ~., data=train, method = "rf", trControl = trCtrl)
rfModel$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 23
##
##         OOB estimate of  error rate: 0.41%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 5575    3    1    0    1 0.0008960573
## B   16 3775    6    0    0 0.0057940479
## C    0   12 3403    7    0 0.0055523086
## D    0    0   23 3190    3 0.0080845771
## E    0    0    3    6 3598 0.0024951483
```

The mean decrease in feature impurity plot:

```r
varImpPlot(rfModel$finalModel, main="Variable Importance", cex = .6)
```



Test set result

Random forest fitted is applied on test set to predict values asked for Coursera Practical Machine Learning Course Project.

```
result <- predict(rfModel$finalModel, testFeatures)
result
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  B  A  B  A  A  E  D  B  A  A  B  C  B  A  E  E  A  B  B  B
## Levels: A B C D E
```

References

[1] Qualitative activity recognition of weight lifting exercises E Velloso, A Bulling, H Gellersen, W Ugulino, H Fuks Proceedings of the 4th Augmented Human International Conference, 116-123