# Predicting Qualities of Physical Activities

## Introduction

Most people agree that doing more physical activities can help to improve people’s health and life. It is very easy to know how much a particular activity one has done by count time of the activity, but it is hard to quantitatively measure how well one has done the activity. Recent developments in mobile and health devices have enabled possibilities to collect a large amount of data about how people do an activity. By analyze this data, quantitative measurements of the qualities of physical activities are no longer a difficult problem.

## Description of data

Researchers from Human Activity Recognition (HAR) research (<http://groupware.les.inf.puc-rio.br/har>) ask 6 male participants (aged between 20 and 28) to perform lifting a 1.25kg dumbbell in 5 different manners, which are described as A, B, C, D and E. The manner A is the only one exactly following the specification. The other four manners corresponding to four different common mistakes respectively. The five manners are stored in the last column of the data named “classe”. The data also include more than 150 variables, most of which are measured from 4 sensors attached on various positions on the participants’ body.

## Methodology

The first 7 columns in the data represent the indices, usernames and information about time stamps and time windows for measurements. I believe this should not be related to the types of manner, since they should be the uniformly designed during the experiments for all the manners. Therefore, it is better to delete the first 7 columns.

The data include a large number of missing values and invalid numbers such as “#DIV/0!” and “NA”. Most of them do not provide any important information regarding the type of manner. For example,

Table 1 Fraction of NAs in the variable "var\_yaw\_forearm"

|  |  |  |
| --- | --- | --- |
|  | Missing values and invalid numbers | Valid numbers |
| Count | 19216 (98%) | 406 (2%) |

Therefore, all the columns that contain missing values and invalid numbers are also dropped. These processes cut the number of variables down to 52 variables.

The remaining date is divided into two groups: training (75%) and validation (25%). The training data is used to train a model. The validation date is used to validate data and evaluate the accuracy of the data.

Some of the variables are obviously correlated. For example, “accel\_arm\_x” and “magnet\_arm\_x” are measurements of acceleration and magnetometer reading of an arm on the x direction. In the lifting dumbbell activity, this two variables are is strongly correlated because in the activity, the arm will move over a relatively fixed path with relatively fixed speed (see Fig. 1)

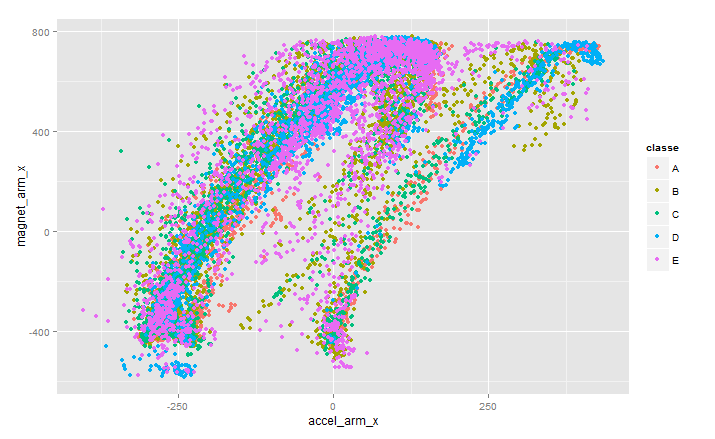


Figure 1: The plot show a special pattern between “accel\_arm\_x" and "magnet\_arm\_x", which indicates they have a close relation. The correlation between the two variables is 0.8133068.

Considering this situation, it is necessary to preprocess the data using principal components analysis (PCA) to reduce number of variables. After PCA process (with 0.95 thresh), there are only 25 variables left.

At this stage, there are two options to train the data: 1) train the original training set which has 53 predictors and will be more accurate; 2) train the new training data set with 25 principle components, which is faster but will lose some accuracy. The random forest (RF) method can be used to train the two data sets. The trained models is applied to the validation data set to find out which option is better and then the better model will be used to find out the final results.

## Results

### Training with original 53 predictors

The original training data set which has 53 predictors is used to train a random forest model. The model is applied to the validation data set and confusion matrix is calculated. The accuracy of the model in the validation set is 0.996 with 95% confidence interval is (0.994, 0.998).



### Training with 25 principal components

The same process can be applied to the data set with 25 principal components. The model is the used on the validation data set and confusion matrix is obtained. The accuracy of the model in the validation set is 0.98 with 95% confidence interval is (0.975, 0.983).



### Properties of the random forest model

From the accuracies in validation data set of the two models, the model trained with original data set (53 predictors) is better. Although it spend more time for training process, the excellent accuracy and easier to implement relative to the principle components make it a better model.

The random forest model contains 500 trees. The following plot shows the training error as a function of number of trees. The errors for 5 manners are almost constant after 200 trees, so about 200 trees are enough to get the similar accuracy.

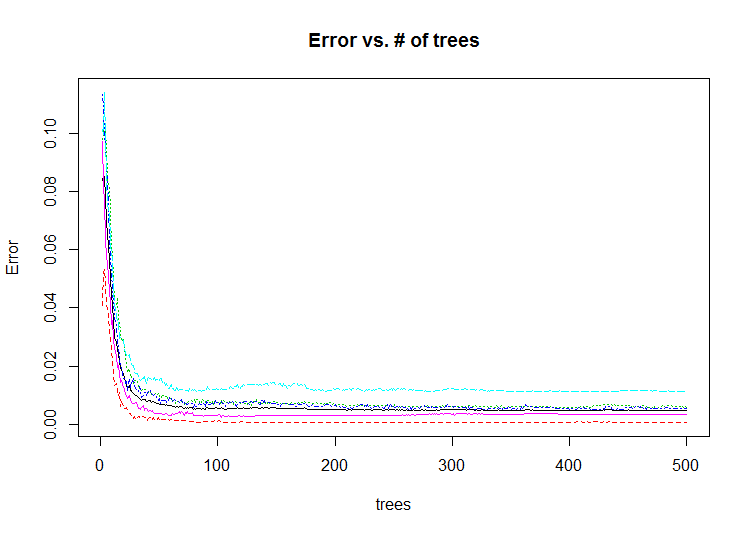


Figure 2: Error of the model as a function of number of trees

The variable importance plot is shown in the following figure, where displays decrease of accuracy and Gini coefficient of the model (only the highest 20 variables are shown). The mean decrease of accuracy indicates that the most relevant predictors in the model are "yaw\_belt" and "roll\_belt" and these two variables also have the highest Gini coefficients. Therefore, they are the most important variables that can be used to determine the quality of one's physical activity. It is interesting that the quality of a lifting-dumbbell activity which uses arms only is most relevant to two measurements on the belt. This is a point that HAR researchers should study in future.

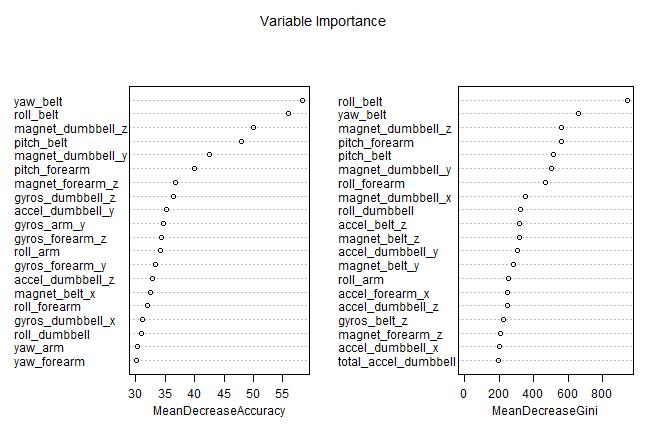


Figure 3: The mean decrease in accuracy and Gini coefficient (sorted decreasingly from top to bottom) of 25 predictors in the random forest model.

### Results of the project

The aim of the project is to predict the type of manner of 20 testing data sets from a physical activity. The results are obtain by applying the trained random forest model to the 20 testing data sets and making predictions for them.

The final model is applied to the testing data set and provide following results:



## Summary

The HAR data is used to distinguish the 5 different types of manner in lifting-dumbbell activity. The 53 sensor measurements are picked as predictors to train a random forest model. The final model is testified with a validation data set which achieves 99.6% accuracy, so the model is proper for predicting the type of manner in the activity. The model shows that the most important measurements of the quality of the lifting-dumbbell activity are "yaw\_belt" and "roll\_belt", which both set on the belt.