Prediction Model by Activity in Samsung Smartphones

1 Introduction

The purpose of this analysis is to find the model to predict the activities that a person does, using 562 measurements taken from a Samsung Smartphone by activity (Jorge L. Reyes-Ortiz, 2012).

Predictive models used to be Support Vector Machine(SVM) with its variants, Random Forest and Decision Trees to identify the model that best predicts the evaluation data from the training data.

2 Methods

This analysis used the following methods or methodologies:

2.1 Preparation of the training data set and evaluating

Through Sansumg data set comprising of 7352 records and 563 variables and was obtained on March 4 of 212 at 14h00.

Over 563 variables were found the following columns with names repeated 3 times:

times:fBodyAcc-bandsEnergy()-1,16
fBodyAcc-bandsEnergy()-1,8
fBodyAcc-bandsEnergy()-17,32
fBodyAcc-bandsEnergy()-25,48
fBodyAcc-bandsEnergy()-33,48
fBodyAcc-bandsEnergy()-49,56
fBodyAcc-bandsEnergy()-57,64
fBodyAccJerk-bandsEnergy()-9,16

fBodyAcc-bandsEnergy()-1,24
fBodyAcc-bandsEnergy()-17,24
fBodyAcc-bandsEnergy()-25,32
fBodyAcc-bandsEnergy()-33,40
fBodyAcc-bandsEnergy()-41,48
fBodyAcc-bandsEnergy()-49,64
fBodyAcc-bandsEnergy()-9,16
fBodyAccJerk-bandsEnergy()-1,16

fBodyAccJerk-bandsEnergy()-1,24
fBodyAccJerk-bandsEnergy()-1,8
fBodyAccJerk-bandsEnergy()-17,24
fBodyAccJerk-bandsEnergy()-17,32
fBodyAccJerk-bandsEnergy()-25,32
fBodyAccJerk-bandsEnergy()-25,48
fBodyAccJerk-bandsEnergy()-33,40
fBodyAccJerk-bandsEnergy()-33,48
fBodyAccJerk-bandsEnergy()-41,48
fBodyAccJerk-bandsEnergy()-49,56
fBodyAccJerk-bandsEnergy()-49,64
fBodyAccJerk-bandsEnergy()-57,64
fBodyGyro-bandsEnergy()-1,16
fBodyGyro-bandsEnergy()-1,24
fBodyGyro-bandsEnergy()-1,8
fBodyGyro-bandsEnergy()-17,24
fBodyGyro-bandsEnergy()-17,32
fBodyGyro-bandsEnergy()-25,32
fBodyGyro-bandsEnergy()-25,48
fBodyGyro-bandsEnergy()-33,40
fBodyGyro-bandsEnergy()-33,48
fBodyGyro-bandsEnergy()-41,48
fBodyGyro-bandsEnergy()-49,56
fBodyGyro-bandsEnergy()-49,64
fBodyGyro-bandsEnergy()-57,64
fBodyGyro-bandsEnergy()-9,16

For the names of the columns are repeated at the end use of the variable character ".1", ".2" and ".3" correspondingly.

2.2 Missing data

There were no missing data in the data set of Samsung smartphones for training or test data sets.

2.3 Preparation of training data and test data sets.

In this analysis, the training data set had included subjects 1, 3, 5 and 6, with 1315 rows and 563 columns. While that for the test data set had included subjects 27, 28, 29 and 30 with 1485 rows and 563 columns.

2.4 Outliers

Methods were used to check the behavior of a variable with respect to the mean, variance and outliers. For the behavior of two variables, were revised behavioral variables with respect to the variable activity.

2.5 Prediction Models

This analysis includes an exploration of the prediction model with the following models: SVM (linear, polynomial, radial) (machines, 2012), Random Forest (R, 2012) and Decison Tree (Zhao, 2012) to obtain the error rate in the training dataset and evaluation.

2.6 Model Comparison

The obtained pattern matching comparison between actual and predicted activities in each of the records of the training and evaluated data sets. The indicator used to measure error rate was Mean Square Error (MSE).

2.7 Reproductible

This analysis is reproductible, beacuse used the next packages in R: randomForest (Forest), e1071 (e1071, 2012) and Decision Tree (Tree, 2012). Floowing the instuction in lectures 6 and 7.

2.8 Confunders

This analysis did a covariance to extract correlation between las variables. Where obtained 8 possibles confunders, between tBodyAccJerkMag.entropy..(235) and fBodyAcc.entropy...X (288), fBodyAccJerk.entropy...X (367) and fBodyAccJerk.entropy...Y (368)

Analysis

After obtaining the results by model and type of MSE, we obtained the following results detailed in Table 1 and their presentation in Figure 1.

Model	**MSE Train**	**MSE Test**
Random Forest	0	0.0713805
SVM Linear	0	0.0875421

SVM Radial	0.01	0.0989899
SVM Polynomial	0.0205323	0.1353535
Decision Tree	0.0425856	0.1649832

Table 1. Comparison of MSE in the training and test data sets, considering the comparisons between actual and predicted values.

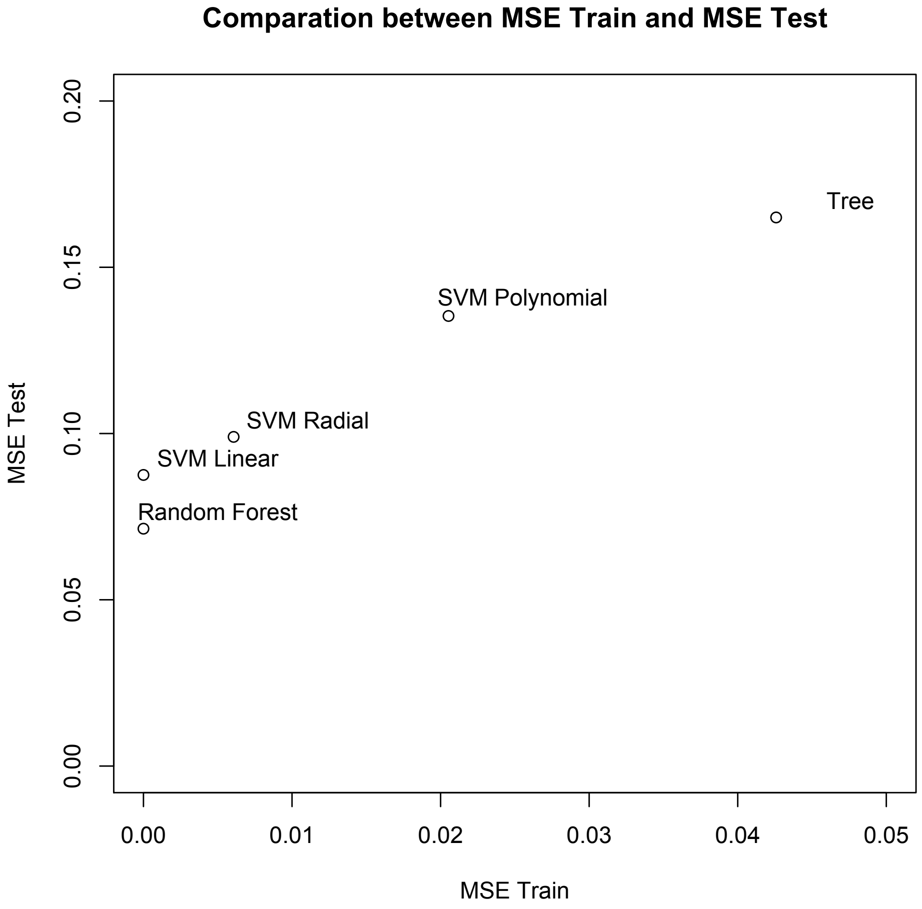
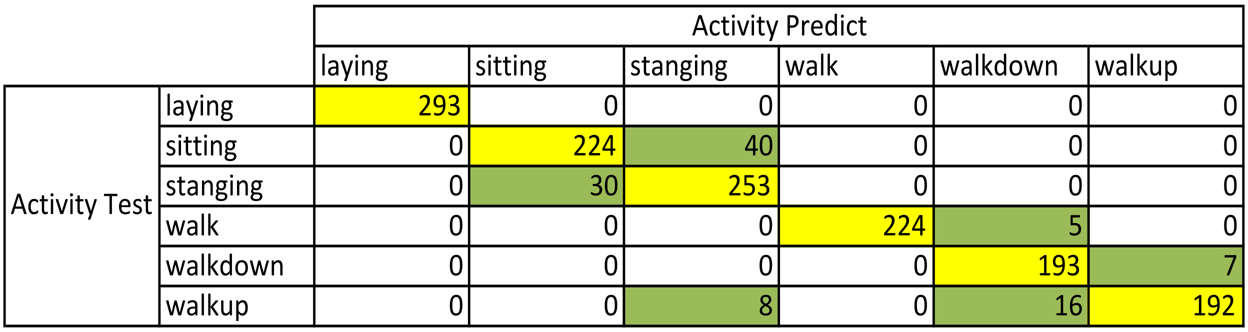


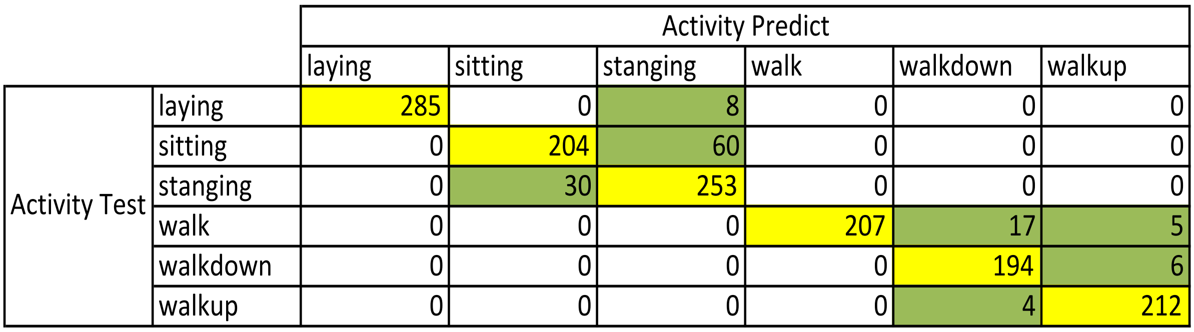
Figure 1. The results of the measurement errors for the training data and evaluation for each model

Random Forest and SVM are Linear models get a 100% training, there are activities that generate differences in prediction. As shown in Table 2 and Table 3.



		Activity Predict					
		laying	sitting	stanging	walk	walkdown	walkup
Activity Test	laying	293	0	0	0	0	0
	sitting	0	224	40	0	0	0
	stanging	0	30	253	0	0	0
	walk	0	0	0	224	5	0
	walkdown	0	0	0	0	193	7
	walkup	0	0	8	0	16	192

Table 2. Count the number of hits of activities between assessment data and the prediction for the Random Forest method. The results are in yellow are correctly predicted values and green values are incorrect predictions.



		Activity Predict					
		laying	sitting	stanging	walk	walkdown	walkup
Activity Test	laying	285	0	8	0	0	0
	sitting	0	204	60	0	0	0
	stanging	0	30	253	0	0	0
	walk	0	0	0	207	17	5
	walkdown	0	0	0	0	194	6
	walkup	0	0	0	0	4	212

Table 3. Count the number of hits of activities between assessment data and the prediction for the SVM Linear. The results are in yellow are correctly predicted values and green values are incorrect predictions.

Conclusions

The prediction model for Random Forest is more accurate at predicting an activity, especially in laying, sitting and walk. While for waldown activities and walkup, Linear SVM model, improves it.

The most relevant variables prediction model with Random Forest are presented in Figure 2, being explained by the 10 first variables that are most GINI.

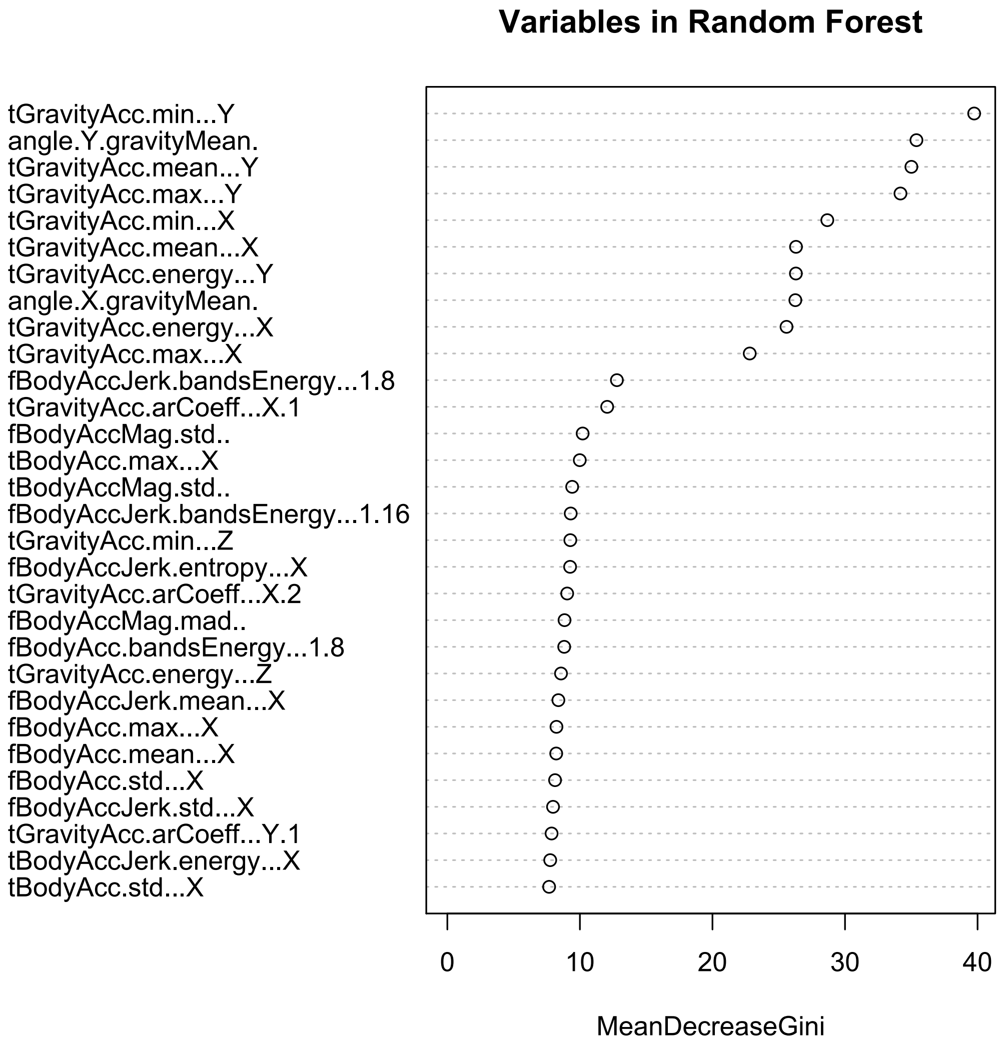


Figure 3. Variables in Random Forest with GINI for prediction model.

References

R, R. F. (08 de 02 de 2012). *RANDOM FOREST IN R.* From COMPUTATIONAL PREDICTION: http://mkseo.pe.kr/stats/?p=220

Drew Conway, J. M. (2012). *Machine Learning for Hacking.* O´Reilly.

machines, S. v. (2012). *Support vector machines.* From Stats 202 Data Mining: http://www.stanford.edu/class/stats202/svms.html

Jorge L. Reyes-Ortiz, D. A. (10 de 12 de 2012). Human Activity Recognition Using Smartphones Data Set . *Human Activity Recognition Using Smartphones Data Set* . Genova.

Zhao, Y. (2012). *R and Data Mining, Example and Case Studies.* Elsevier.

e1071. (12 de 09 de 2012). *e1071.* From CRAN: http://cran.r-project.org/web/packages/e1071/e1071.pdf

Forest, R. (n.d.). *Random Forest.* From CRAN: http://cran.r-project.org/web/packages/randomForest/index.html

Tree, C. a. (03 de 11 de 2012). *Clasification and Regression Tree.* From CRAN: http://cran.r-project.org/web/packages/tree/index.html