Samsung Human Activity recognition report

Siyu Nan

Feb 2017

Introduction

Samsung smart phones with an embedded accelerometer and gyroscope can capture 3-axial angular velocity at a constant rate of 50Hz. In order to predict activities of users, experiment is conducted with 30 individuals participated. In the experiment, 561 different features are recorded from the raw accelerometer and gyroscope signals, which contain 6 different activities, walking, walking upstairs, walking downstairs, standing, laying and sitting. Using all 561 features, it is proved that the accuracy rate of 98% can be achieved. However, this is a very expensive way to make the prediction. In this project, minimize number of features as predictors but still achieve a decent rate of accuracy (more than 80%) is the objective. Existing algorithm is used for classification and optimization purposes.

Data collection and preparation

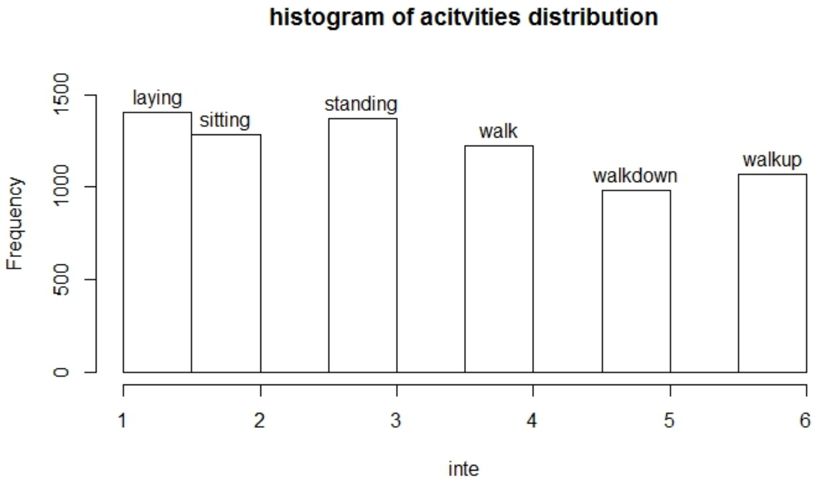
Data collection

Two datasets are given by Instructor Kirk Ocke. The original one includes data collected from 30 individuals from 19 to 48 performing six activities with Sumsung on them. The second datasets is used in our project which contains the exact same information but splitting into training and testing sets. The .rda file, training set is a dataframe file with 21 individuals' data. In this file, 7352 observations and 563 columns, which consisted of 561 features, subject number and activity performed. Same format and information is contained in testing set, which 9 individuals' 2202 records. Validation set is not used here since 10 fold cross validation is used in our model.

Exploratory Analysis

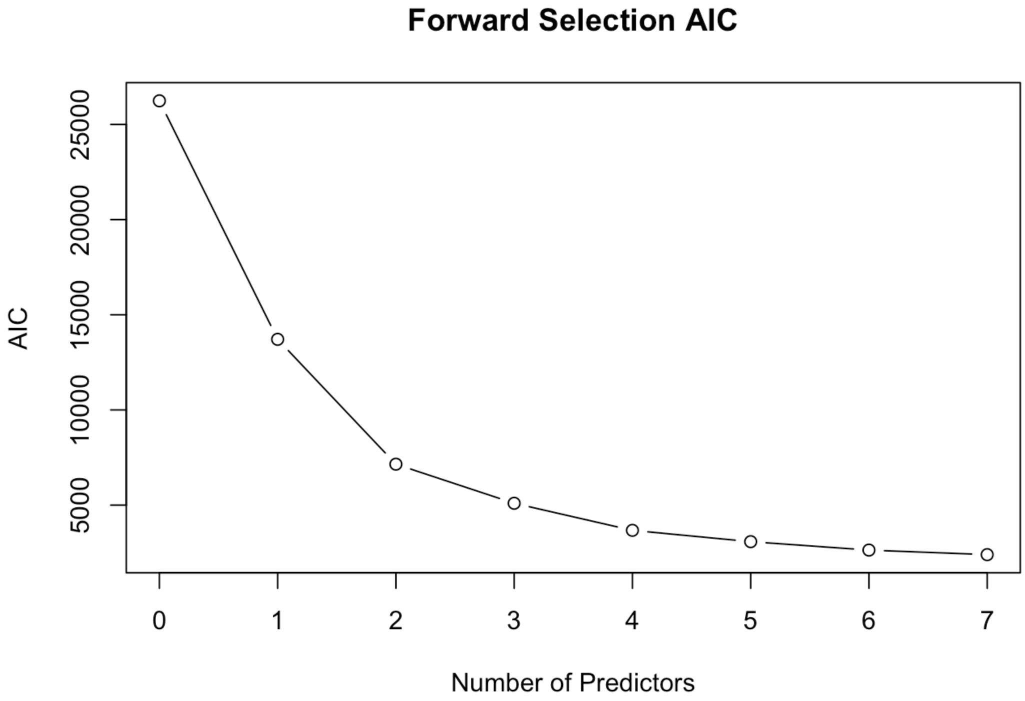
The completeness and distribution of the data is checked. Results show that there are no missing values and different activities, as shown below, are evenly distributed. However, when checking the attribute's name, error appears which indicates there're replicated name. Recorded data in those attributes with identical name are very close. In order to decide whether to keep both, hypothesis tests are performed, which give a very low p value. That is to say, though with the same name, these features record different information. New names are given to deal with this issue thereafter.

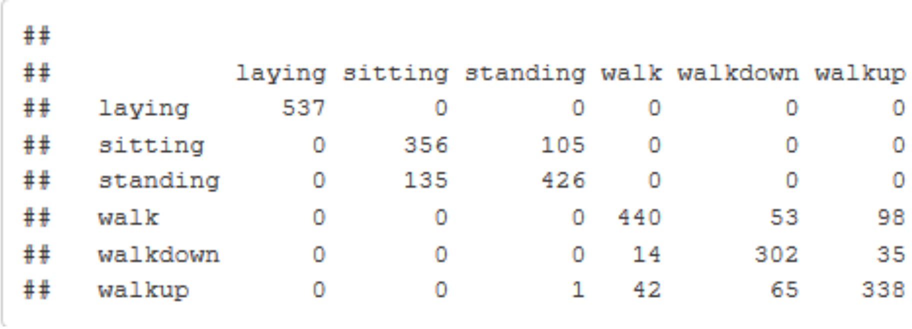
Found similar data in features with identical name inspires us to do a correlation check. The results turn out to be there are features having 0.9 or higher correlation. One of them is randomly picked and dropped in order to avoid over-fitting issue in this project [3].



Modeling and Analysis

Modeling in our project is random forest [2] with stepwise multinomial selection. Since this is a high dimensional dataset, we used random forest to pick out important features, after which stepwise recursive feature elimination is implemented. Stepwise recursive feature elimination works in the way that is based on the idea to select features by recursively considering smaller and smaller sets of features. It repeatedly constructs a model and chooses either the best or worst performing feature. It then sets this feature aside and repeats the process with the rest of the features. As a greedy algorithm, this process is repeated until all features in the dataset are exhausted. Features are ranked according to when they were eliminated [1]. After getting top 50 features in random forest, feature elimination narrow down the scope to 4 features. The reason we chose 4 features are shown in graph below. AIC score is an important measurement for how good a model is. Lower the score is, better the model is. In the graph, it is obvious that the score drops a lot from one feature to four features as predictor. The drop slows down afterward. Therefore, 4 features are good in our model. After stepwise regression with AIC criterion filtered through we tried the four features left to train our model, which achieve about 81.4% accuracy rate in the testing set, as it can be seen in the confusion matrix below.





```
##
##            laying sitting standing walk walkdown walkup
##   laying      537       0        0    0        0      0
##   sitting       0     356      105    0        0      0
##   standing      0     135      426    0        0      0
##   walk          0       0        0  440       53     98
##   walkdown      0       0        0   14      302     35
##   walkup        0       0        1   42       65    338
```

Conclusion

With the Samsung data consisted of 7352 observations and 561 features collected from 30 subjects, our analysis shows that we are able to predict subject's activity with 81% accuracy using 4 features. Although the accuracy rate drops from 98% with all features to 81% with four features, it costs much less. There are some potential problems in this project. Firstly, there are only 30 individual participated in the experiment. Everyone may have their own pattern of moving. Therefore splitting training and testing sets by 21/9 individuals may lose a lot of information. Secondly, the last four features are there because other features, individually, perform worse than them. With the nature of greedy algorithm, we can't say these four features are the best combination. If these problems can be solved, the results would definitely get improved.

Reference

[1]An introduction to feature selections
http://machinelearningmastery.com/an-introduction-to-feature-selection/

[2]Random Forest
https://en.wikipedia.org/wiki/Random_forest

[3]Preprocessing about high correlation variables.
http://topepo.github.io/caret/pre-processing.html#identifying-correlated-predictors