Human Activity Recognition Using Smartphones

Introduction

Every month technology companies such as Google, Apple and Samsung deliver new smartphones with new and improved capabilities such as new screens, better internet connectivity and better accelerometers, gyroscopes and GPS. This tools are available to developers and are used on hundreds of applications, the analysis of this data is of vital importance in order to create better experiences to the user; one of the most recent examples of the usage of smartphone data is Google Now [1].

The objective of this analysis was to build a function that predicts the current activity a person is doing, in order to accomplish this objective I used the data of the accelerometer and gyroscope from people carrying a smartphone while doing different activities. Using this data I train a predictor model using the random forests technique, this models proved to be the best one of the different models trained.

Methods

Data collection

The experiments was carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, Laying) wearing a smartphone (Samsung Galaxy S II) on the waist. [2]

The data was processed by the collector in order to make it easier to analyze:

- The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window).
- The time domain signals were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20Hz to remove noise.
- The body linear acceleration and angular velocity were derived in time to obtain Jerk signals. Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm.
- A Fast Fourier Transform (FFT) was applied to some of these signals producing a frequency domain version of those signals. This procedure is very common on analysing signals since the conversion provides easier calculations that in a time domain could require complex calculus [3].

The data consisted of 7352 records with 563 columns that could be potentially used to train the predicted model distributed as follows:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.
- A 561-feature vector with time and frequency domain variables.
- Its activity label.
- An identifier of the subject who carried out the experiment.

The data was downloaded from the *UCI Machine Learning Repository* [4] on monday March 4, 2013.

Exploratory Analysis

The exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data. The data used matches all the requirements of what is considered tidy data, there weren't missing values found on the dataset therefore no imputation methods were needed, also each column correctly displays information about only variable only and finally since most variables are numbers there was no need of transform or clean any of the data.

I found that some of the variables were skewed but given the distribution of the data (very small values close to zero) the transformations done (log (x) and log (x + 1)) produced worse results and therefore were discarded.

Statistical Modeling

The data was divided on 3 sets: Training (using subjects 1 to 20), Validation (using subjects 21 to 26) and Testing (using subjects 27 to 30). The validation dataset was used to test all the models tried and the testing dataset was used at the end to make the final prediction.

Given the nature of the target variables (categorical variable with no order) some statistical models can be discarded at once, for example a linear regression would not be useful at on this case since there is no way of order the target variable; e.g. standing is not greater or lower than walking.

I decide to use the Decision Tree model to create the predictor and validate using the validation Dataset. I used N-Fold cross validation to reduce the risk of overfitting. The final model used was a random forest [5] model which is a model machine learning technique which uses the principles of ensembling small Decision Trees, this model improved the final prediction.

Variable Selection

The number of variables available to make a prediction is really high (561) on this case the decision to which variables use and not use on to train the model is not that simple. In general is

not a good idea to introduce highly correlated variables since in most cases those variables could have the same information. For example if the yearly income and the monthly income of a person are available both variables will be highly correlated to the but the information is duplicated since one can be represented as a linear function of the other, if both variables are used for training the model will be biased.

But is not as simple as just remove the correlated variables since this case is a little bit different; some of the variables will be correlated but that correlation is natural and the different correlation between them is what defines the activity, for example moving an object to the top-right and right will have correlated and not correlated variables and that difference is the key of identifying the different activities.

But even though some correlations are good on this case we also have the case of duplicated information. For example we have the mean, max, min, median absolute value and the Interquartile range, those variables are not linearly independent, that is some of them can be written as a linear function of others. Therefore some variables needs to be removed.

Also on this case we have the variables on two domains: a time domain and frequency domain. On this case the difference is not relevant since we are looking for relation between those variables and the target variable. Therefore I decided to use only the time domain variables, this decision reduce the number of variables to almost half.

Taking into account the previous statements I tried to found the best variables to train the model trying to reduce at most the number of variables selected and getting the best accuracy and RSE. The idea used was use only sets of 3 variables, since most of the variables have 3 components: X, Y and Z and the difference between those variables is of vital importance in order to correctly identify the different activities.

Reproducibility

All analyses performed in this manuscript could be reproduced using the original R markdown file which is available on request, the development was done using RStudio (2013) using R version 2.19. To reproduce the exact results presented in this manuscript, the analysis must be performed on the same data set and using the same packages used on this analysis.

Results

In order to have a benchmark I first train a Random Forest model using all 561 variables and then the 265 time domain variables. The accuracy of the model using all variables was 92.72% and the accuracy of the model using only the time domain variables was 92.65%. The gain of less 0.1% is no worth it to introduce more than 200 variables for this reason the frequency domain variables were rejected.

Using only the time domain variables then I began testing groups of 3 variables (X,Y,Z) trying to improve the accuracy while introducing fewer variables. The following table contains the accuracy using different combinations of variables

ID	Introduced Variables	Accuracy	Total number of variables
1	BodyAcc (XYZ) - mean BodyAcc (XYZ) -std	64.37%	6
2	BodyGyro (XYZ) - mean BodyGyro (XYZ) - std	81.68%	12
3	Row 1 + Row 2 BodyAcc (XYZ) - iqr	79.93%	15
4	Row 1 + Row 2 GravityAcc (XYZ) - mean GravityAcc (XYZ) - std	84.71%	18
5	Row 4 BodyJerk (XYZ) - mean BodyJerk (XYZ) - std	90.37%	24

From the following table I was able to conclude that with only 24 variables of the 561 I was able to get a 90.37% accuracy which is less than 2% than the case of using the 265 time domain variables. Also on the third row I was able to note that the introduction of some variables reduces the accuracy of the model.

The previous conclusions tell me than the variables selection is of vital importance on this case, at the end I decide the use the 24 variables on row 5.

Model comparison

In order to improve the model I used a random forest classifier [5] which greatly improves the accuracy of the model by using ensemble methods. In average the accuracy is improved by 3-4%. Using the selected 24 variables and a single decision tree the accuracy found was 87.47%.

Also using the selected 24 variables the Mean Squared Error of the Decision Tree was 0.2224 and the random forest got and MSE 0.1427 improving again the result of a single Decision Tree.

Confusion Matrix

The heat map of the final confusion matrix can be seen in Figure 1. Is possible to see the

misclassification rate is very low as the diagonal has very high red colors and the rest has a deep blue fill.

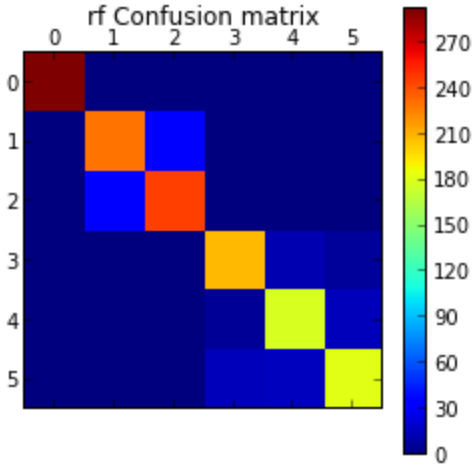

Figure 1. Confusion matrix heat map for the Random Forest classifier. The encoding of the variable is the following: 0: Standing, 1: Walkdown, 2: Sitting, 3: Walkup, 4: Laying, 5: Walk.

The misclassification for a person standing was 0% while the other activities do have some error. The lower accuracy obtained was for activity 4, Laying and really close also the walking activity had the lower accuracy, 85% and 89% respectively.

Conclusion

The accuracy and MSE obtained for the final Random Forest model tell us that the model has a very high accuracy and a low MSE, in general the classifier does a very good job predicting the current activity a person could be doing.

Using only 24 variables of the total 561 variables that could be used I achieved an accuracy on the testing dataset of +90%. The selection of the selected variables was done both manually and looking at the accuracy and MSE improvements. If adding a set of variables did not generate any relevant increase in accuracy the variables were rejected. Also I used set of 3 variables (X, Y and Z axis) since the difference between those variables is the key to identify the current state of a person.

Given the amount of variables there is a possibility of found confounders in the variables. The manual selection tried to reduce the number of possible confounders but given the high amount of combinations possible that some confounders are still present and not included on the model.

The manual variable selection also tried to reduce the possibility of overfitting I can say that using the selected 24 variables the fitting of the data is much less noisy than using the complete 561 variables. Also I trimmed both the decision tree and random forest generated trees to a maximum depth of 10, this is a common practice in order to overfitting.

Even though the trained model has very high accuracy and low error rates is possible to improve the prediction. Using more data (records) will improve the predicted values. While there is a high amount of records (+7000) all records are from only 30 people and some bias can be found on the data. Finally I believe the trained model can be highly improved by an algorithm which automatically selects the best variables.

References

[1] Google Now. Available online [http://www.google.com/landing/now/]
[2] Information from the 'README.txt' file included with the downloaded data.
[3] Fast Fourier Transform. Available online [http://en.wikipedia.org/wiki/Fast_Fourier_transform]
[4] Human Activity recognition using smartphones. UCI Machine learning. Available online [http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones]
[5] 'randomForest: Breiman and Cutler's random forests for classification and regression' Available Online ['http://cran.r-project.org/web/packages/randomForest/index.html]