Project 2: Human Activity Recognition using smartphone

Contents

1. Introduction

Smartphones have become most useful tool in our daily life for communication with advance technology. It is equipped with variety of sensors from motion detector to optical calibrators. The data collected by these sensors is valuable for better aligning the applications on the phone with user's lifestyle. In this project, we have focused on using data collected from motion sensors (accelerometer and gyroscope) to build a model, which identifies type of activity performed with minimal computation involved. The end goal is to create a model, which can classify the activity performed with high accuracy without sacrificing the limited computational resources available on a single phone

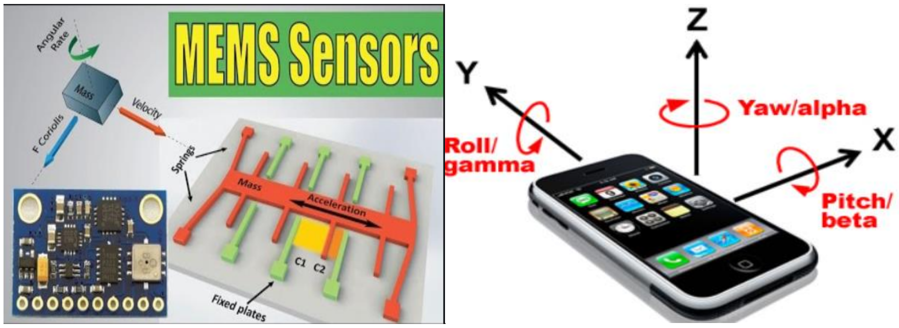


Fig-1 Image of accelerometer and gyroscope

2. Data Set Description and Cleaning

We used the data provided by Human Activity Recognition research project, which built this database from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The complete data & related papers can be accessed at: UCI ML repository page

Data collected for 30 volunteers whose age was between 19-48 years. Each record in the data represents information about features like acceleration along x,y,z axes, velocity along a,y,z axes, 561 attributes derived from these basic measurements, identifier variable for the user & the activity being performed.

There are 6 categories of activities being performed:

1. 'standing'
2. 'sitting'
3. 'laying'
4. 'walk'
5. 'walkdown'
6. 'walkup'

The raw data has separate text files for most of the variable groups & we have used the dataset that saved as text file. In this dataset, a single column ('subject') is used to identify a user and the last column ('activity') was used to identify the activity being performed when the measurements were taken. All other attributes are available in the same column oriented data format. This is important to know, because, the values in the dataset have been normalized.

3. Data Wrangling and Cleaning.

Read Train, test data, and added their features to create Data Frame. Data is already clean and not required any cleaning steps.

4. Exploratory Data Analysis:

Though the observation for each activity are not exactly equal, the data set overall provides a well-balanced distribution of activity observations.

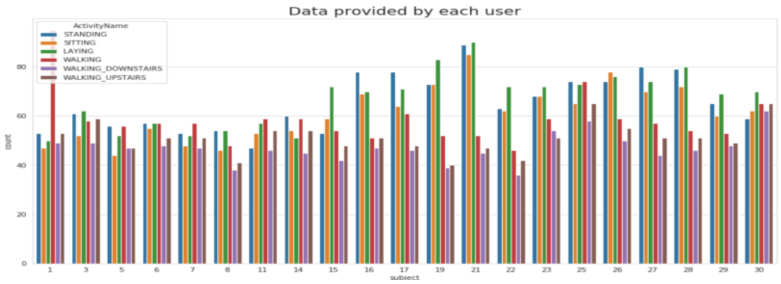


Fig-2 Data provided by each user

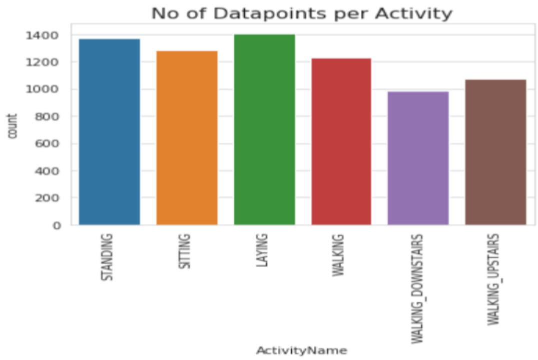


Fig-3. No of data points per activity

Activity is divided in two categories as below. Below figure, provide visualization for Static activity and Dynamic activity.

- Static Activity (Motion information is not useful)
  - Sit
  - Stand
  - Lie down
- Dynamic Activity (Motion information is useful)
  - Walking,
  - Walking Upstairs,
  - Walking Downstairs

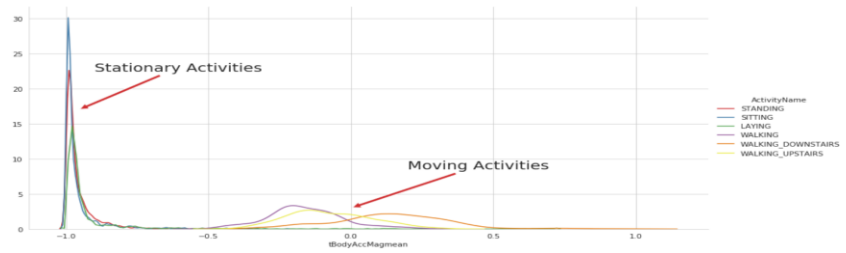


Fig-4 Stationary and moving activities graph

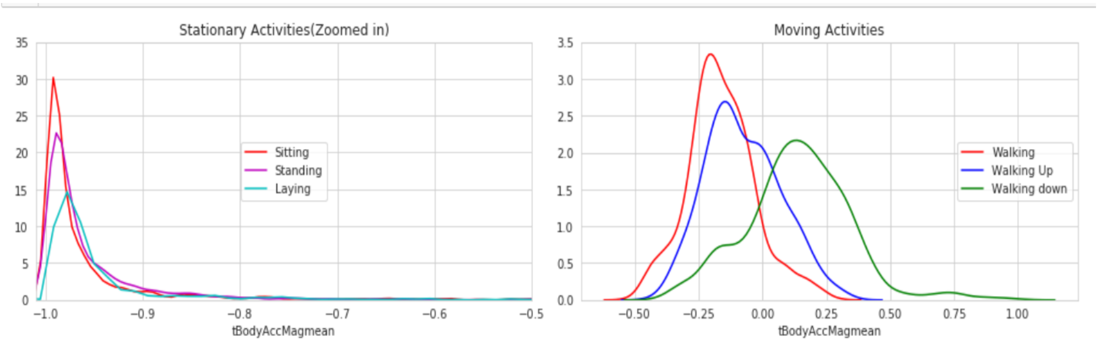


Fig-5 Stationary and moving activities graph

The next set of analysis is carried out to understand the variation in data for each activity. The feature of 'Mean Body Acceleration' along the X, Y, Z axis was documented as a scatter graph. The graph in Figure 6 and table indicates that the mean value of the body acceleration is more variable for the activities of walking, walking upstairs and walking downstairs than the passive activities of sitting, standing and laying.

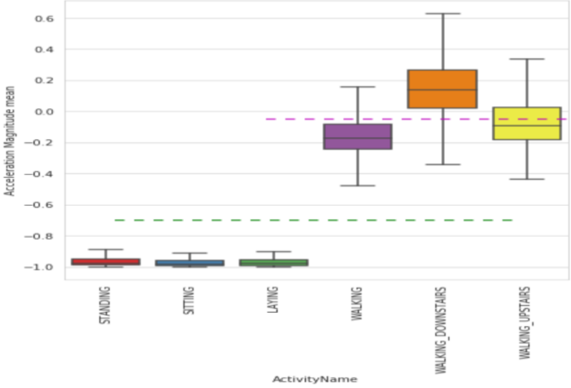


Fig-6

Activities	Acceleration magnitude mean
Standing	-0.1< tAccMean < -0.8
Sitting	-0.1< tAccMean < -0.8
Laying	-0.1< tAccMean < -0.8
Walking	-0.5< tAccMean < 0.2
Walking Downstairs	-0.4< tAccMean < -0.6
Walking Upstairs	-0.4< tAccMean < -0.4

The feature of ‘Angel Gravity mean’ along the X, Y, Z axis was documented as a box plot graph. We can classify all data point belongs to Laying activity with just observation on angle x gravity mean >0

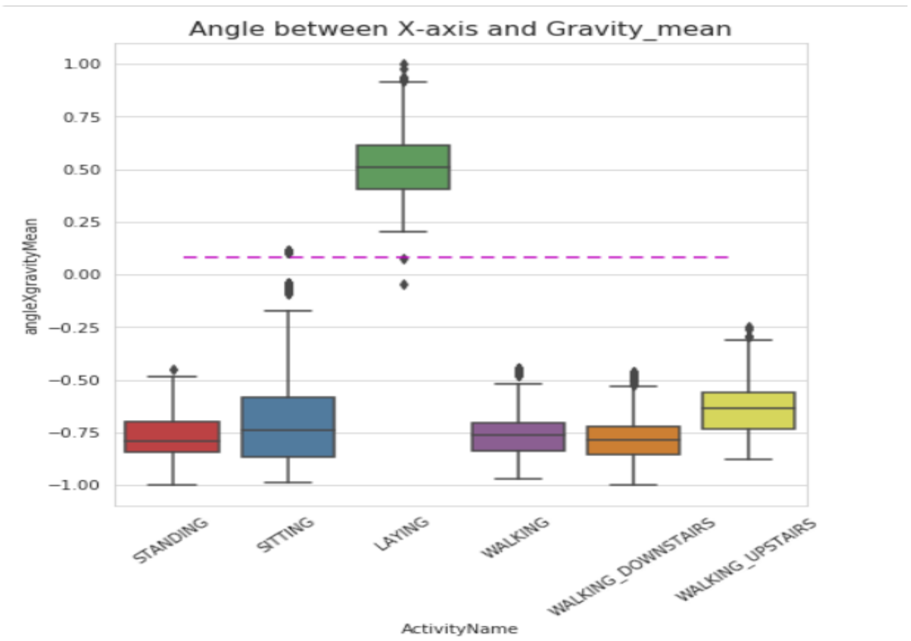


Fig-7

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. It is extensively applied in image processing, NLP, genomic data and speech processing. We can see with different perplexities, all the different activity names are clustered Pretty well, except standing and siting.

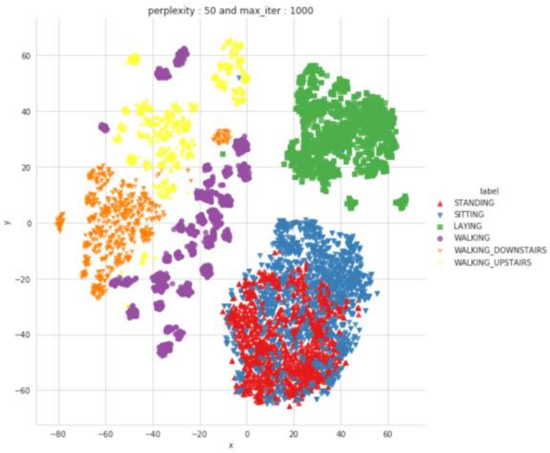
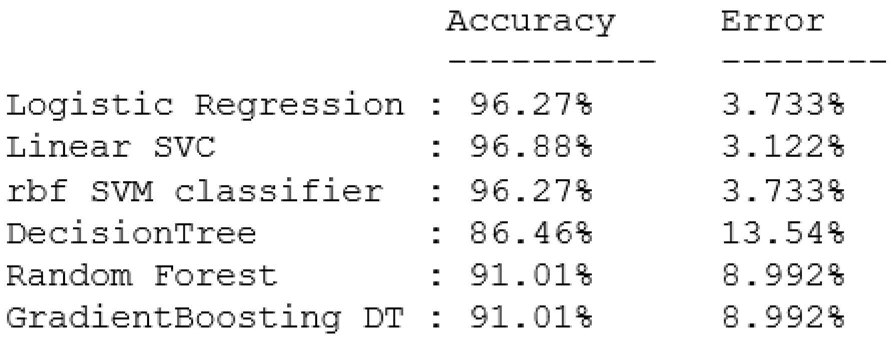


Fig-7

5. Results and In-depth analysis using machine learning

Below six machine learning classification models have been performed on data and selected features for prediction

1. Logistic Regression with Grid search
2. Linear SVC with gridSearch
3. Kernel SVM with gridsearch
4. Decision Trees with GridSearchCV
5. Random Forest Classifier with GridSearch
6. Gradient Boosted Decision Trees With GridSearch



```
                      Accuracy     Error
                     ----------   --------
Logistic Regression : 96.27%      3.733%
Linear SVC          : 96.88%      3.122%
rbf SVM classifier  : 96.27%      3.733%
DecisionTree        : 86.46%      13.54%
Random Forest       : 91.01%      8.992%
GradientBoosting DT : 91.01%      8.992%
```

Models Accuracy and Error

We can choose Logistic regression or Linear SVC or rbf SVM as they are giving the highest accuracies but we can see that there is still a little confusion between the Activity Names - 'SITTING' and 'STANDING', as we saw in the T-SNE visualization as well. We can send this information to the experts and let them know about this discrepancy and ask them if they can work on it and try to improve it. If they cannot, it is still a good model with 96% accuracy. However, we can try to work with the raw, non-expertized data with Deep Learning models and see if there is any improvement. If we can get the same 96% accuracy on the model without the expertise, it would be a huge accomplishment

LSTM (Long Short-Term Memory layer) : LSTM is an artificial recurrent neural network(RNN) architecture used in deep learning. LSTM networks are well-suited to classification, processing and making decision

```
 Score[0.41655886096877154, 0.8954869508743286]
```

With a simple 2-layer architecture, we got approx. 90% accuracy and a loss of 0.49. We can further improve the performance with Hyper parameter tuning

6. Conclusion

Model Linear SVC or rbf SVM are providing higher accurate prediction. Better feature selection methods and improvement in tuning the parameters can assist further to improve accuracy and decrease computational cost. We have successfully shown that with the use of the LSTM network model, built to train on the sequences of raw inertial signals, features can be learned automatically by the network, and a significant accuracy is achieved. The accuracy achieved by a recurrent neural network on raw signal data is at par with other classification models, which are built on handcrafted features. Adding further layers to the network or increasing the complexity would further boost the recognition accuracy of the deep learning algorithm.

7. References

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
http://scikit-learn.org/stable/modules/ensemble.html#forest
https://en.wikipedia.org/wiki/Random_forest
http://scikit-learn.org/stable/modules/svm.html
https://en.wikipedia.org/wiki/Support_vector_machine
https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr
http://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html
https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
http://scikit-learn.org/stable/index.html