Multi Classification Model on Human Activity Prediction using smartphone data sets

sun

Problem

How to use sensor signal data from smart devices to predict human activity:

walking,sitting,lying,walking stairs,running and so on

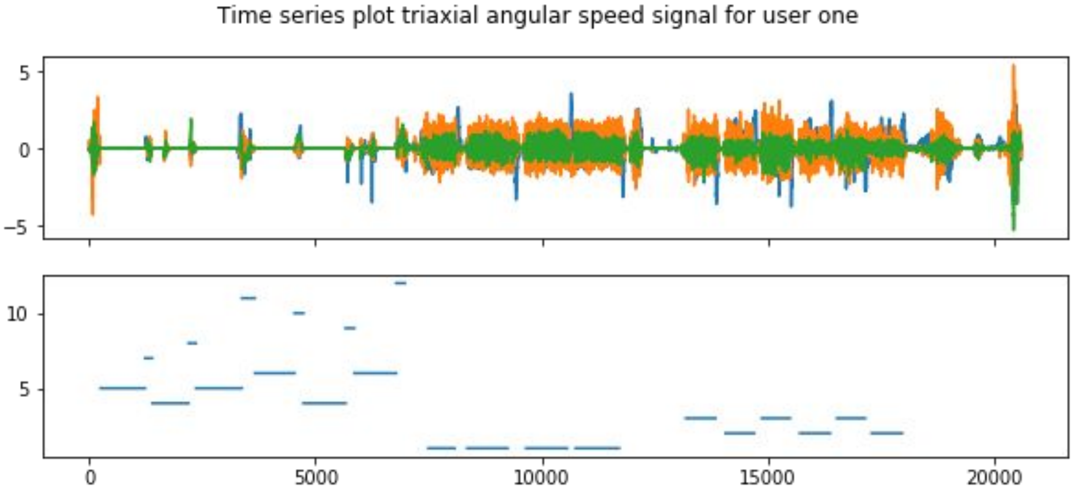
Data

Data: UCI

http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions

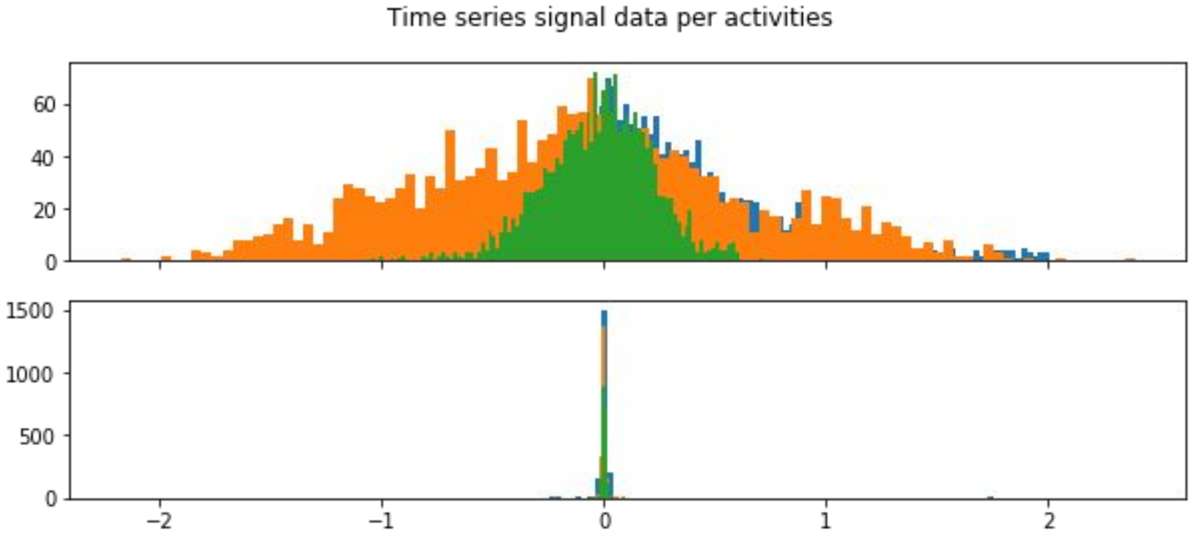
Sensor signals from smartphones of a group of 30 people performed a protocol of activities(such as walking,sitting ,lying and walking stairs) were collected and pre-processed.

Time series plot triaxial angular speed signal per user



Walking(upstair and downstair) involved intense signal, the transition pose or laying pose, signal is weak

Histograms plot of gyro signal per activity



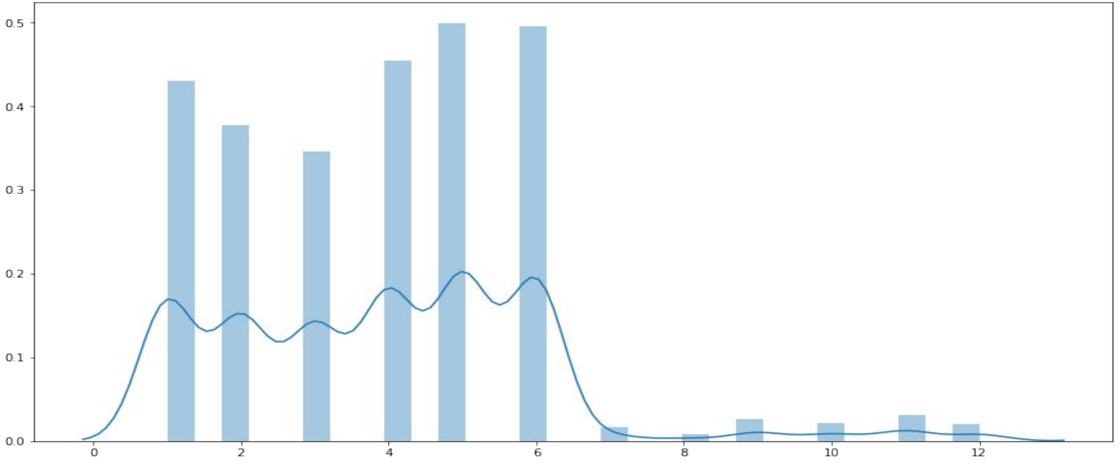
Both walking upstair(top) and laying(bottom) activities have normal distribution ,but walking upstair has a wider standard deviation

Phase one:Data Analysis on Dataset2

- 1.Load Data
- 2.Missing Data
- 3.Unbalance Data
- 4.Outlier
- 5.Correlation
- 6.Null hythophis

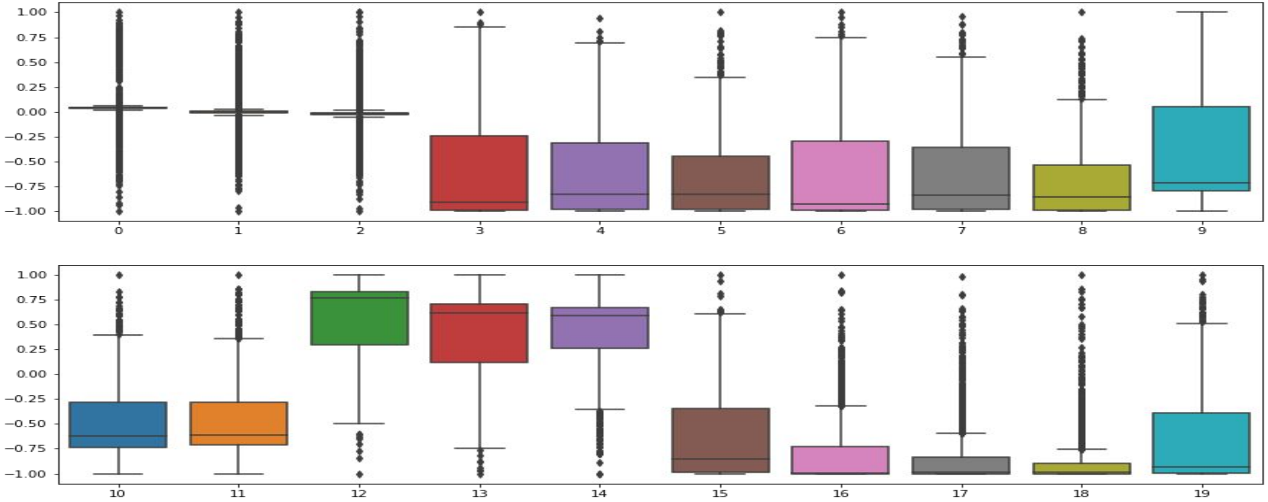
Balance Data

The barplot shows our dataset has 6 major classes and 6 minor classes, it is unbalanced, so we applied SMOTE to resample the data



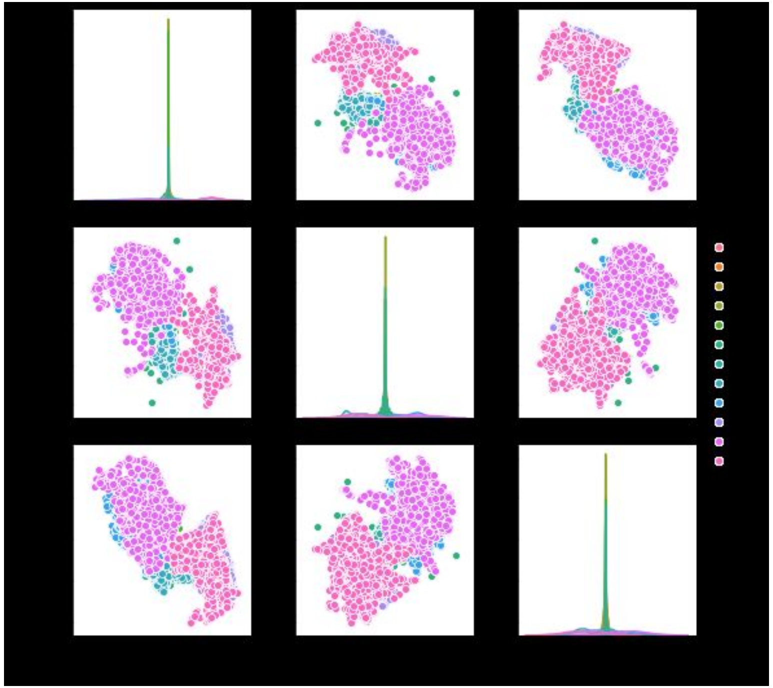
Outliers

According to the boxplot , we can see that there are a lot of features having outliers.For example feature 8 has a lot of larger outliers,we refill the outlier with 75% value ,feature 13 has smaller outliers ,we will refill those with 25% value



Correlation

The results of plot pairs of the first 3 numeric variables shows some variables are correlated with each other , for example feature 0 and feature 1



Null hythophis

Hythophis:There is no significent difference between below two sets:Set a) feature 2 value larger than 0.5, y belong to class 4 Set b) feature 2 value less than 0.5, y not belong to class 4

Result:The pvalue is 7.363006721969925e-25 < 0.05, so the null hythophis is rejected, there is a big chance if feature 2>0.5 ,it will belong to class 4.

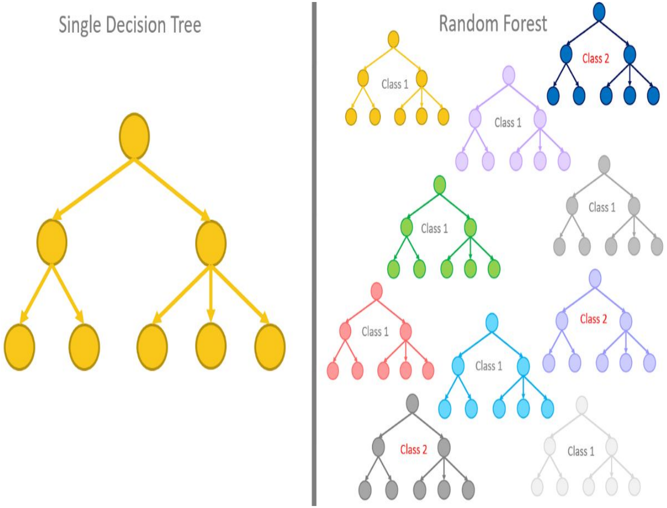
DataSet2 Analysis Summary

- No missing data
- All numeric fields
- Unbalance dataset
- A lot of Outliers
- Pairbox shows some features are correlated
- When feature 2>0.5, there is a big chance ,the final class will be 4.

Phase two: Machine learning Analysis

1. **Model Selection:**
   a. **Rodam Forest**
   b. **Logisist Regression**
2. **Train the Model:**
3. **Model Optimization:**
   a. **GridSearchCV()**
4. **Model Evaluation *:***
   a. ***Accuracy_score***
   b. ***Confusion Matrix***
   c. **Roc_auc_score**
   d. **Classification report**

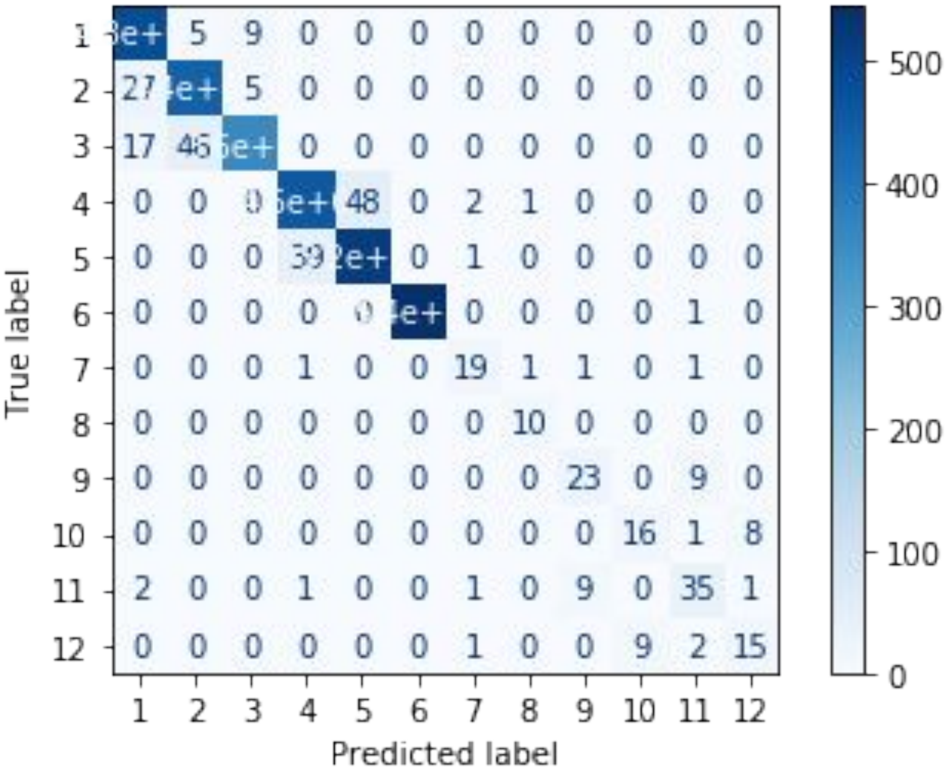
Random Forest



Random Forest is an ensemble of randomized decision trees model: simple,efficient,versatile and also helps prevent overfitting

- Random Forests are less influenced by outliers ,can handle noisy data
- no assumptions about the underlying distribution of data, and can implicitly handle collinearity in features
- Random Forests can be used for feature selection
- "robust": can work with practically any kind of data, when mixing categorical and numerical features, or mixing completely different ranges of values, no need to scaling

Confusion Matrix

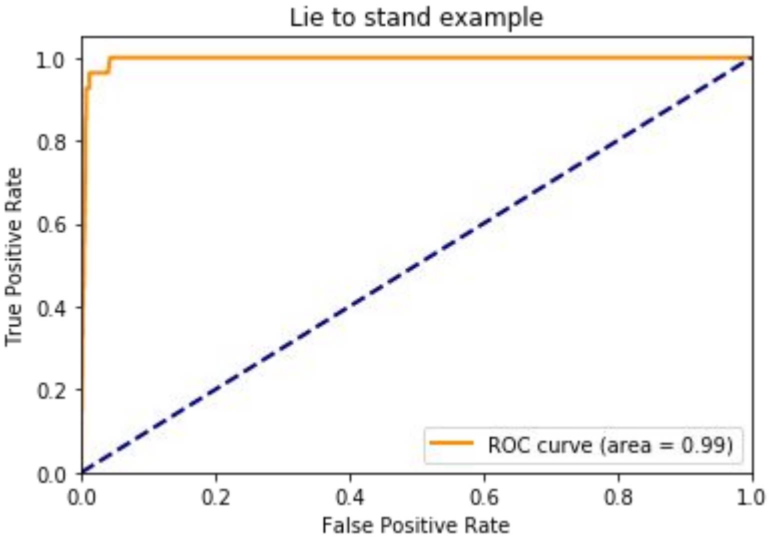


	1	2	3	4	5	6	7	8	9	10	11	12
1	3e+	5	9	0	0	0	0	0	0	0	0	0
2	27	4e+	5	0	0	0	0	0	0	0	0	0
3	17	46	5e+	0	0	0	0	0	0	0	0	0
4	0	0	0	5e+(	48	0	2	1	0	0	0	0
5	0	0	0	39	2e+	0	1	0	0	0	0	0
6	0	0	0	0	0	4e+	0	0	0	0	1	0
7	0	0	0	1	0	0	19	1	1	0	1	0
8	0	0	0	0	0	0	0	10	0	0	0	0
9	0	0	0	0	0	0	0	0	23	0	9	0
10	0	0	0	0	0	0	0	0	0	16	1	8
11	2	0	0	1	0	0	1	0	9	0	35	1
12	0	0	0	0	0	0	1	0	0	9	2	15

The classifier has very good predictions on the 6 Major class(walking,walking_upstairs,walking_downstairs,sitting, laying)

Less performance on the transition poses. both false positive and false negative case are exist, especially for class 10 and 12, it hard to seperate 'lie to sit' from 'lie to stand' which make sense

Roc_auc_score



Roc_auc_score: Compute Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. By 'ovr' mode , we got score value `0.995625`

ROC curves: Below is the ROC for last class ‘lie to stand’ transition pose, we can see it still has good AUC score.

Classification report

```
              precision    recall  f1-score   support

           1       0.90      0.98      0.94       496
           2       0.89      0.93      0.91       471
           3       0.96      0.84      0.89       420
           4       0.96      0.88      0.92       508
           5       0.90      0.97      0.94       556
           6       1.00      1.00      1.00       545
           7       0.74      0.74      0.74        23
           8       0.82      0.90      0.86        10
           9       0.64      0.72      0.68        32
          10       0.65      0.68      0.67        25
          11       0.70      0.57      0.63        49
          12       0.67      0.52      0.58        27

    accuracy                           0.92      3162
   macro avg       0.82      0.81      0.81      3162
weighted avg       0.92      0.92      0.92      3162
```

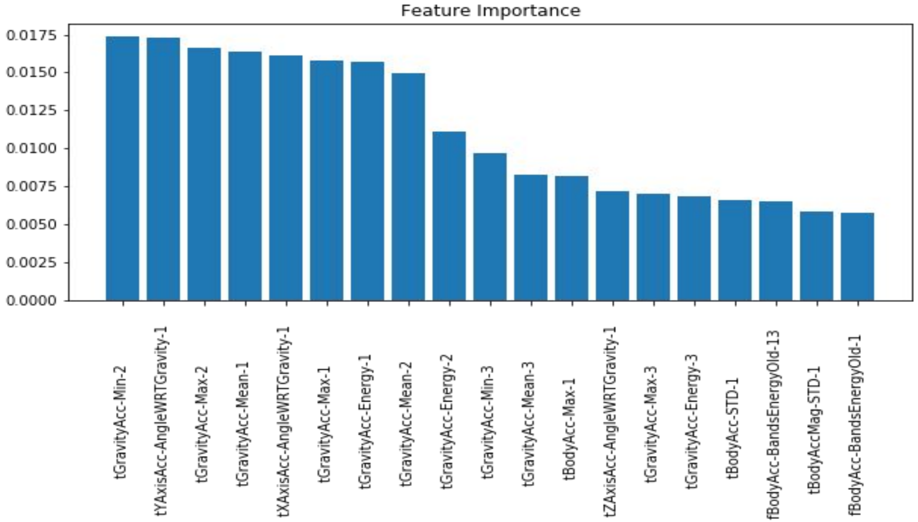
Class 6 laying has highest f1 score,the classfiler is good to predict laying pose.

Standing pose has good recall than precision, in opposite, walking downstairs has good precision than recall, for transition poses both are not good enough

Feature selection

PCA:Since our data has 560 features, PCA analysis shows that if we can reduce the dimension to 14 and we still can keep 80% varierious

Feature_importantce: below is top 20 important features, it sees tgr features matters



Final thoughts

- Both LR and RF are good classifiers for our data, Logistic Regression seems better, that might because the dataset be normal or binomial distribution and features are mostly independent. We may tune LR as our final model.
- For dataset2,Confusion Matrix and Classification report are good evaluation metrics, roc curve is not good to seperate class
- We also can build deep learning models on Dataset1(time series data), it will be interesting to compare those results.