Project2:Predicting Parksinson's Disease Progression with Smartphone

Hengcheng Zhu, Yuhui Zhang

1 Summary

We demonstrate several different ways to apply LDA, QDA, KNN, RF classifiers to the MJFF data that were passively collected from 16 smartphone users, 9 of which have Parkinson's Disease(PD). We mainly focus on the acceleromater data after we filter some outliers by doing clustering analysis on the other two dataset:GPS data and Compass data.Our final result shows that the QDA is relatively better than the other methods by achieving $13/16 = 81.25\%$ accuracy.We think the spot of our report is the balance between model complexity and model accuracy.

2 Data Overview and preprocessing

The data is organized into about 7000 compressed folders each containing a set of csv and log files.We firstly use the R.script (provided by Kaggle and GSI) to unzip and combine these csv files,but takes nearly more than 10 hours just to extract and combine the data.Furthermore, the output combined data doesn't include subject's name,which is crucial for our analysis.Instead, we write a bash shell script to programmatically uncompress and concatenate csv files of each type of smartphone data for each user and discarded the used extracted files to save disc space.To our surprise, it took only 20 minutes to perform all these work.We now know how to deal with the large dataset at the very begining step.

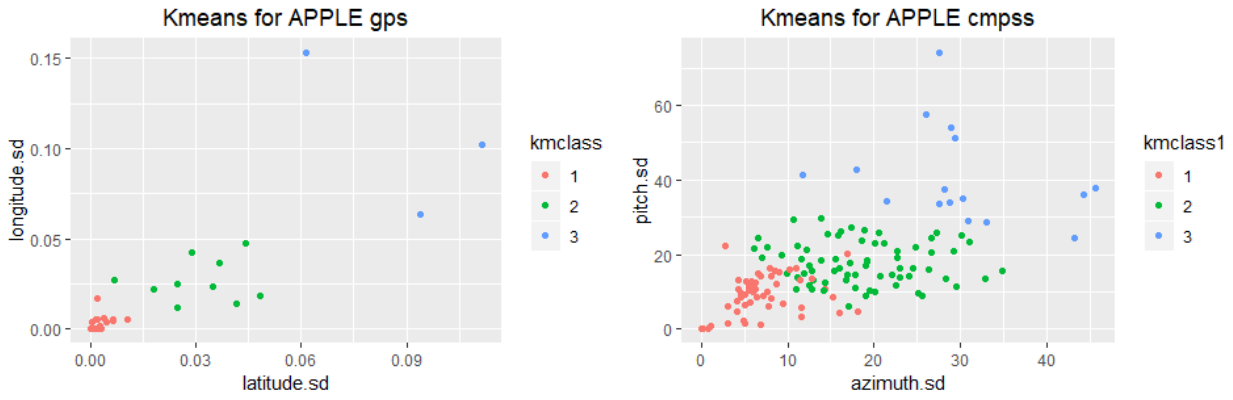
The acceleration data includes average second mean acceleration, acceleration standard deviation, acceleration absolute deviation, acceleration maximum deviation, power spectral density (PSD) for 1Hz, 3Hz, 6Hz and 10Hz bands combined across all x, y and z axes.The compass data comprises the pitch, roll, azimuth information of every second to describe the inclination of the device. The GPS data record the mean and deviation of the subject's latitude, longitude, altitude for every second.

According to the movement-related symptoms of Parkinson's (shaking, rigidity, slowness of movement), the pattern of the acceleation data is expected to differ between control group and patients.So we foucus on the acceleration data to see the different patterns for two groups.We believe that the collected data is bound to to be affected by external noise: for example, walking or sitting on a moving vehicle can influence the tremors recorded by the accelerometers. what's more , the movement of the phone itself can also add great noise to the acceleation data.Therefore, the accelerometer data should be filtered by considering compass data and gps data.

3 Clustering analysis on Cmpss and GPS data

As mentioned before, Compass and GPS readings are used to isolate time intervals in which the phone wasn't being moved or rotated in an excessive manner.The importance of isolating such moments is underlined by the fact that tremors in Parkinson patients tend to be stronger when the subject is at rest.

For GPS data, we firstly aggregate them by minute by taking the average of every variable.Then we calculate the standard deviation for latitude, longitude and altitude respectively to see whether the phone position changed too frequent or not.To overcome vaiability, I scale them three to have the same measure.Next we apply Kmeans clustering on these data to see is there any apparent pattern in the data in the EDA step.



We do the similar processing step for Cmpss data: firstly aggregate them by minute by taking the average of every variable and then calculate the standard deviation for latitude, longitude and altitude respectively.As an example shown above, we found that kmeans clustering is an effective unsupervised leaning alghorim to identify some outliers in the dataset.The left plot indicates that in these time intervals, the APPLE subject were moved greatly(like walking, up-down stairs), so we choose to filter out the corresponding time-series data in the acceleration dataset by matching the same time interval(year-month0day-hour).It also applys to the cmpss data.The points belong to class 3 are highly possibily influenced by frequent movement of the cell phone in this hour interval.Each of the 16 subjects were applied to this process to filter out their 'unreliable' data in acceleration dataset.

4 Feature selection and classification on Acceleration data

For the acceleration data, we choose to focus on 5 variables:the mean acceleration, power spectral density (PSD) for 1Hz, 3Hz, 6Hz and 10Hz bands combined across all x, y and z axes (using root mean square).Similar as before, we aggregrate them within one hour interval and filtering out those were considered unreliable on the previous step.Finally, we got 3657 observations and each observation with 5 predictors without counting the corresponding subject's name and PK's status.

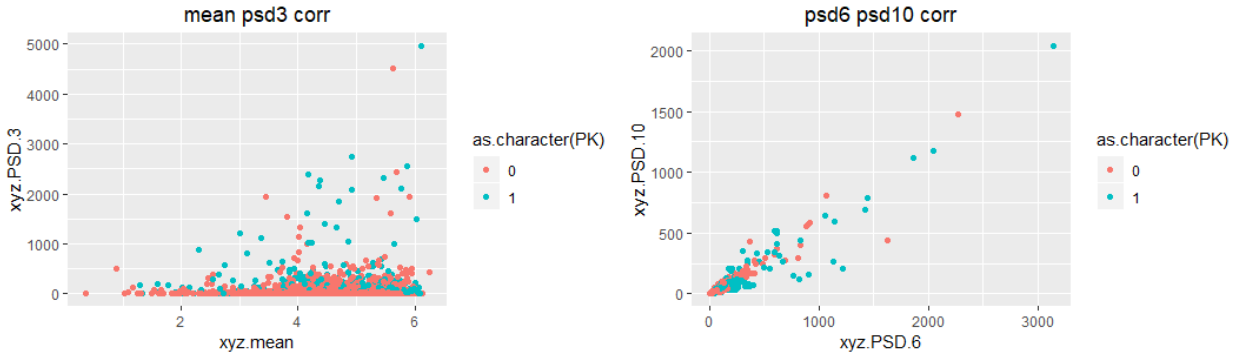
Then we apply different classfication methods to the final dataset, including Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbour and Random Forest (RF).To assess the performance of each method, I used leave-one-out cross validation.That means instead of randomly split the whole 3647 observations into training and testing set, we trained with all data points except those of one user, and then use the trained model to predict the class of the leaved-out subject.If the proportion of the correct predicted points exceeds 0.5, then we choose to count this outcome as true.

Except from RF, the other methods performed badly by only getting 56.25% or 62.5% accuracy.Rf got 75.6% on the whole 3657 data and 68.75% accurcay on the cross-validation setup.We also found that all the classification methods prefer to predict subject as PD subject.So to improve our accuracy, next we choose to invetigate is there any distribution difference of the variables.

5 Further improvement

The following are three examples to show the different distribution of two kind of subjects.We found that the control group tend to be more scattered than PD group, which is similar to the conclusion of Brunito

et.al.So we then want to utilize this 'differentrial expression' property to improve our classfication accuracy.



Inspired by the idea of Brunito et.al, we did a new feature extraction by space quantization.However, this time we split each variable distribution into Q equal intervals and then calculate the proporation of points falling into each one interval, thus can quantify the distribution of each variable and then can compare the difference between two groups.In our setting, we pick the Q from 4 to 8, and thus the desgin matrix is a 16 by 5Q matrix.And it is clear that the smallest is 20, which is larger than 16.Due to the smaller size of the datset, we have to deal with this n > p problem because some classification methods only proper to $n < p$ problem.To make our problem easier, We choose to pick the idea from high-throughput genoimcs setting that use two sample-test between control group and PK group to choose the differential expression variables. We finally choose Q = 4 and threhold α being 0.3 to pick 8 significant variables.The best result of this new method is conducted by QDA method to achieve 81.5% accuracy.

6 Disussion

We have to admit that the big size of the data but smaller size of the subjects pose us a really hard problem.Although we found that using space quantiztion of variables is useful to try this problem, our method is too rough to identify different distribution patterns between 2 groups.More carfully space quantization method are easy to come up with and the accuracy is expected to be better than us. The method Brunito et.al they pick was quite complex but achieved 100% accuracy with the help of SVM.

7 Reference

1. Brunato, M., Battiti, R., Pruitt, D. and Sartori, E. (2012). Supervised and unsupervised machine learning for the detection, monitoring and management of Parkinson's disease from passive mobile phone data.
2. Wang, M. (2012). Identifying Parkinson's Disease from Passively Collected Data.
3. David Ireland et al.(2013) A Review of the Kaggle Data And a Proposed Response to the Kaggle Challenge.
4. Pramod Anantharm et al.(2013) Predicting Parkinson's Disease Progression with Smartphone Data