PH290: Kaggle Predicting Parkinson's Disease Progression with Smartphone Data

Jin Rou New
SID: 25944841

1 Summary

A number of classifiers, including LDA, QDA, SVM and RF were applied to hourly accelerometer data passively collected from 16 smartphone users, 9 of which have Parkinson's Disease (PD). The first three classifiers performed poorly, predicting the test/left-out user to have PD at every try. However, the RF had superior performance with 100% prediction accuracy determined by leave-one-out cross validation. The RF hence shows promising results for predicting if a user has PD based on his/her accelerometer data.

2 Introduction

The key aim of the Kaggle competition is to use the data collected from 16 smartphone users to distinguish Parkinson's Disease (PD) patients from healthy controls. Of the 16 users, 9 are PD patients at varying stages of the disease. The smarthphone data includes audio, accelorometry, compass, ambient light, proximity, battery level and GPS data.

3 Data processing and feature engineering

The data is organized into about 7000 compressed folders each containing a set of csv and log files. I wrote a bash shell script to programmatically uncompress and concatenate csv files of each type of smartphone data for each user and discarded the log files. Ultimately, only accelerometer and GPS data were used in my analysis.

For both accelerometer and GPS data, I aggregated them by second (as multiple samples were usually available in one second) by taking the average for every variable. Next, for the accelerometer data, I aggregated them by hour and discarded hourly windows with less than 5 seconds of data. Finally, as pointed out by Wang (2012), since the x, y and z axes were measured with respect to the phone's orientation, which is unclear based on the accelerometer alone, I took the root mean square of the 3 channels for all variables.

The resulting data set has 4615 observations and 8 features aside from the class (PD vs Control) for the 16 users: average hourly 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 (using root mean square). The relationship between each pair of variables is visualized in Figure 1. We see some separation between PD patients and healthy controls.

4 Proposed method

I applied a range of classification methods to the processed data set to determine the one that gave the best results, including Linear Discriminant Analysis, Quadratic Discriminant Analysis, Support Vector Machine and Random Forest (RF).

To assess the performance of each method, I used leave-one-out cross validation, as was also done by Brunato et al. (2012). Each classifier was trained with all data points except those of one user, and then used to

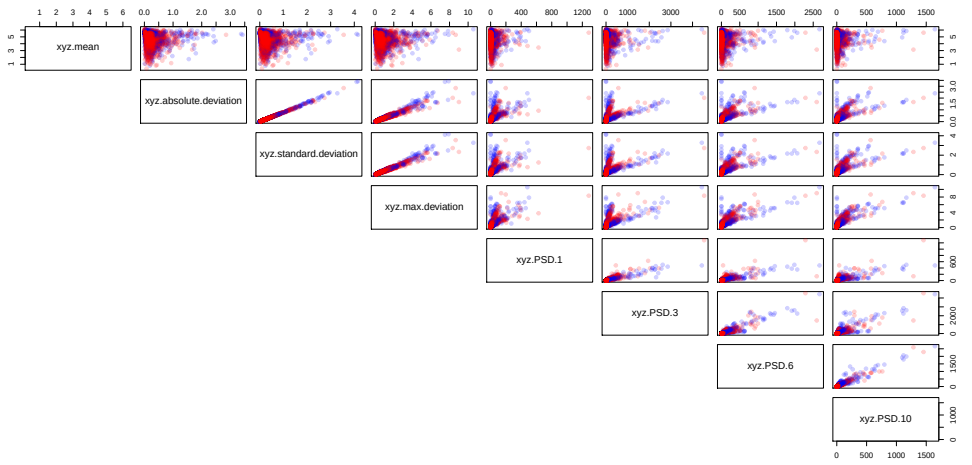

Figure 1: **Pairs plot of all 8 features for PD patients (in red) and healthy controls (in blue).**

predict the class of all data points of that left-out user. The proportion of of data points of that user that are predicted to have the "PD" class gives the probability that the user has PD. If that probability is more than half, that user is predicted to have PD.

The first three classifiers performed poorly, predicting the test/left-out user to have PD in all 16 cases. However, the RF had superior performance with 100% prediction accuracy, i.e. the test/left-out user was predicted correctly to be in either group in all 16 cases. Finally, I applied RF to the full data set and determined that the average hourly mean acceleration and PSD in the 6Hz band seem to be the most important variables.

5 Method benchmarking

Brunato et al. (2012) and Wang (2012) first applied k-means clustering to the compass and GPS data to identify windows in which the users were in motion. Accelerometer data from these windows were then discarded. A SVM was then applied to these data averaged in each window. With leave-one-out cross validation, they achieved 100% prediction accuracy.

Wang averaged acceleration data over hourly windows and combined the x, y and z channels into one by taking their root mean square. He also used a SVM on the data with 5 features: the mean acceleration value and the PSD across 4 separate bands. The k-fold cross validation error for his classifiers were about 15% and 9% respectively.

Teo and Nachev (2012), on the other hand, opted for a simple hypothesis-driven univariate analysis of the data. They calculated the intersessional absolute deviation (ISAD) of the accelerometer mean for randomly selected 1-hour recording sessions and found that all users could be categorised into three groups based on their ISAD level: patients taking the medication levodopa, healthy controls and patients not on that medication.

With its 100% prediction accuracy , my proposed method performed as well as Brunato et al. (2012)'s method, despite its comparative simplicity and its use of only the accelerometer data. A fair comparison of

my method and the other two winning methods could not be made as different metrics were used, however, it does appear to be superior given the high prediction accuracy.

6 Additional analyses

I was interested in determining how much could be gleaned from GPS data of smartphone users. After reducing the GPS data to per-second data, I plotted the data by data for all users. An example series is shown in Figure 2 for the user Crocus. It was possible to determine the user's home, daily movement patterns and travel patterns by simply examining these plots over time. More sophisticated analyses could possibly determine the time spent at each location, guess at the mode of transport (e.g. car vs bicyle vs public bus/subway) taken by the user by calculating the speed of travel and comparing the routes taken with bus/subway maps. I tried to determine if the user is a PD patient or a healthy control based on the GPS data (the hypothesis being that someone with PD is likely to move around less). While my guess that Crocus is a healthy control is right, given that Crocus travels long distances frequently, this could not be generalized as there were many confounding variables such as age and occupation. Moreover, these raise important ethical questions and privacy concerns for the use and analysis of smartphone data as too much information aside from the study aims is revealed through the GPS data.

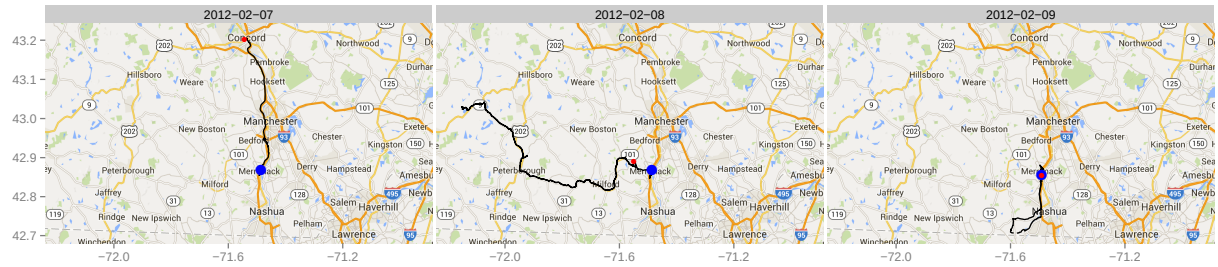

Figure 2: **Plot of GPS coordinates for user Crocus over three days. The blue and red spots show the starting and ending points respectively of each day.**

7 Discussion

While initial results indeed show a lot of potential, more training and testing of my proposed solution on data from a larger study group of smartphone users would be necessary to assess if the results can indeed be generalized to a larger population.

8 References

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. Teo, J.T.H. and Nachev, P. (2012). Remote monitoring of Levodopa response in Parkinson's disease.

3. Wang, M. (2012). Identifying Parkinson's Disease from Passively Collected Data.