**Predicting Parkinson’s disease using smartphone data**

Shuyao Wang

**Abstract**  GitHub: https://github.com/Ginawsy/Parkinson.git

In this project, I compared the methods of SVM and Random Forest utilizing accelerometry, compass, and GPS data of predicting Parkinson’s disease. The results indicated that RF shows a better performance with 100% accuracy than SVM with just 87.5% accuracy. Altitude, longitude, and latitude are the three most significant features conducted by RF and Parkinson's group people show higher altitude values compared to the control group. However, the model which just includes the three most important features shows smaller 62.5% accuracy which may imply correlations between variables that needs to bee further explored.

**Introduction**

Parkinson's disease exhibits several symptoms and features that can be objectively measured and monitored with the aid of common technological devices, equipped with basic sensors that grant a glimpse into the patient's life, such as mobile phones we carry daily. The main objective of this contest is to differentiate individuals who have Parkinson's Disease (PD) from those who are healthy, using data obtained from 16 smartphone users to quantify Parkinson's disease symptoms, thereby facilitating the measurement of the progression of the ailment. Based on the finding, I am also further explore the most important features and comparing the original model with the model that just including these significant features to see if it could help increase the accuracy of the prediction.

**Data Processing**

In the original dataset, there are 16 users, 9 of which had PD at different phases of the illness. The researchers collected various types of smartphone data including audio, accelerometry, compass, ambient light, proximity, battery level, and GPS data. However, the audio and battery data were not included in further analysis of this project due to limitations in determining their relevance. Specifically, in my opinion, the researchers could not determine if the sound was made by the participants or external sources, and if the battery data was influenced by the phone's power source or the time spent on the screen. Additionally, the data of the proximity was also excluded after processing csv files into data frame in R since proximity data only included values of 0 and 1, which may not accurately reflect the actual distance from the phone. For the remaining data (i.e. accelerometry, compass, GPS and the ambient light), the values were turned into numeric format and aggregated by hour. I further discarded the results with fewer than 5 accelerometry data points per hour for accelerometry data and then followed Wang's (2012) approach of averaging the acceleration data over hourly windows, combining the x, y, and z channels into one by taking their root mean square. However, for the data of ambient light, I found that it did not include data for all 16 users by filling the data frames that corresponding to specific users. Thus, I decided not to include it into the model building.

The resulting data has 2656 observations with 25 variables aside from the class (PD vs Control) for the 16 users including average hourly mean, standard deviation, absolute deviation, max deviation for acceleration, roll, pitch, and azimuth; power spectral density (PSD) for 1Hz, 3Hz, 6Hz and 10Hz bands combined across all x, y and z axes (using root mean square); and latitude, longitude, and altitude.

**Data Analysis**

Reflecting on the literature review, some previous teams employed various innovative techniques in their approach to this contest. SVM emerged as a popular choice since SVM is a great choice for supervised learning classification with medium size dataset which can deal with both linear and noon-linear problems. Random Forest is also a popular choice known by its capability to reduce overfitting of the model and is also simple and easy to interpret. For this project, I would further compare the methods of SVM and Random Forest to see which methods with the help of cross-validation have higher prediction accuracy. The performance of the machine learning algorithms was assessed using leave-one-out cross-validation, a common technique used in small datasets. The approach used was similar to that of Brunato et al. (2012), whereby each classifier was trained using all data points except for those of one user, which were then used to predict the class of all data points of that left-out user. The proportion of data points of that user that are predicted to have the “PD” class gives the probability that the user has PD. If the probability is more than half, then the user is predicted to have PD. This approach allowed for a fair and robust comparison between the SVM and Random Forest models to determine which method had the higher prediction accuracy for this specific dataset.

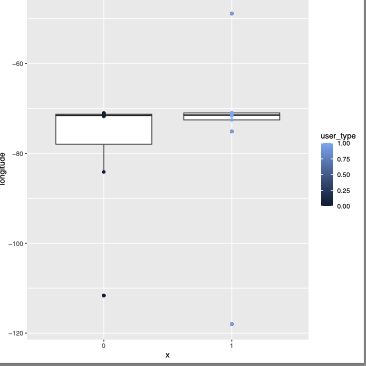
My result shows that SVM leads to an 87.5% accuracy while RF leads to 100% accuracy, predicting the test/left-out user to have PD in all 16 cases. The feature importance conducted by the RF shows that altitude, longitude, and latitude seem to be the most important factors. The max deviation of pitch and mean roll also indicated large significant which can show the importance of location contributed to the prediction of Parkinson’s diseases.

I further visualized the of altitude, longitude, and latitude of people with Parkinson and control individual. From the boxplot, we can see that there is no significant difference found between Parkinson and control group for latitude and longitude data, but it did indicate that majority people in the Parkinson’s group have lager value of altitude. Then I further duplicated the experiment by revising the model to just include altitude, longitude, and latitude. However, the accuracy decreased to 62.5% by running SVM model which might imply some correlation between different types of smartphone data.

A picture containing table

Description automatically generatedChart, box and whisker chart

Description automatically generated



**Conclusion**

My method that combined accelerometry, compass, and GPS data with Random Forest model shows 100% strong accuracy of predicting Parkinson's disease which shows better performance than SVM model. Altitude, longitude, and latitude seem to be the most important features conducted by the RF. And People in the Parkinson's group shows higher altitude values compared to the control group. But the model which also includes the GPS data does not show any improvement from the combined model. More analyses needed to be conducted to examine whether there exist some correlations within variables later.

**Reference**

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.

New, J. (2015). Kaggle Predicting Parkinson’s Disease Progression with Smartphone Data

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