DREAM-PD: Parkinson’s Disease Healthcare Prediction Challenge

Predicting Tremor, Dyskinesia and Medication levels on Parkinson’s disease patients

Kanishk Verma
Student ID: 19210344
kanishk.verma2@mail.dcu.ie

Abstract: In this paper, I will be exploring the techniques for processing digital signals from smartwatches and smartphones of Parkinson’s disease patients. Extracting univariate and multivariate features from these signal data. Also using conventional techniques to segment data based on the Human Activity. Thereby enabling me to make healthcare predictions based on Tremor, Dyskinesia, and Medication levels of the patients using ensemble learning techniques like Random Forest and Voting Classifier.

Keywords: Signal Processing, Human Activity Recognition, Machine Learning, Random Forest Classifier, Voting Classifier

Introduction

BEAT-PD DREAM Challenge: Parkinson’s disease (PD) is a neurodegenerative disease that along with the motor system also shows other symptoms. [1] The Biomarker and Endpoint Assessment to track Parkinson’s Disease (BEAT-PD) challenge, is a challenge to benchmark methods for the processing of sensor data. [1] The challenge comprises 3 sub-challenges to predict on/off medication status, dyskinesia, and tremor severity.

About the Data

The training data consists of sensor files and symptom labels for two different datasets: CIS-PD and REAL-PD.

Data characteristics for both CIS and REAL-PD

Sample Size = 25

Age = 64 ± 8.9

Diagnosis of PD (years) = 8.3 ± 4.0

Sensor Data Sources

CIS-PD	REAL-PD
Smartwatch Accelerometer Data (50 hertz frequency)	Smartwatch Accelerometer (50 hertz frequency), Smartphone Accelerometer and Gyro-sensor Data (100 hertz frequency)

Methodology

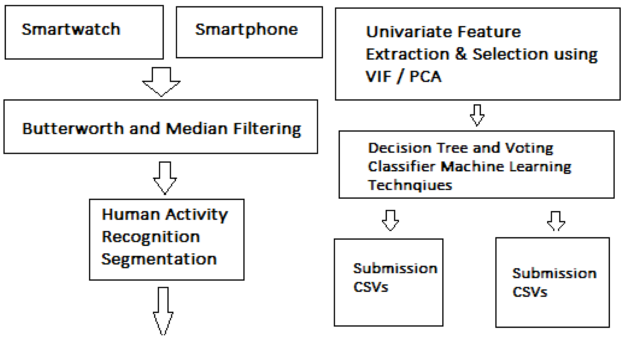
Smartwatch
Smartphone
Butterworth and Median Filtering
Human Activity Recognition Segmentation
Univariate Feature Extraction & Selection using VIF / PCA
Decision Tree and Voting Classifier Machine Learning Technqiues
Submission CSVs
Submission CSVs

Fig 1. Methodology

As seen above, the methodology is inclusive of multiple modules.

1. Filtering / Pre-processing

Since, the data available was raw sensor data. And probably consisted of noise elements. The approach I took for filtering and pre-processing the signal data was based on the techniques deployed by UCI HAR Dataset.

The sensor signals were pre-processed by applying noise filters, sampled in fixed width sliding windows of 3sec and 50% overlap. The accelerometer signal accounts for both gravitational and body acceleration. [2] I passed the raw dataset first on Median Filter, and then on two Low-pass Butterworth Filters with cut-off frequencies of 20hertz and 0.3hertz respectively. Since gravitational force is assumed to have low frequency components gravity was segregated using cut-off frequencies of 0.3hertz.

2. Human Activity Recognition

In the paper [3] to recognise the Human Activity, I have adopted the similar strategy of a conventional cut-point approach wherein the acceleration magnitude and angle-z metric are used to assign each of the 10-second epochs to one of the following three categories: Sustained Inactivity (denoted by Inactivity label in the modified dataset), Light Physical activity (denoted by Light label in modified dataset) and Moderate or Vigorous physical activity (denoted by Moderate label in modified dataset).[4]. Wherein, based on the Euclidean Norm Minus One

(ENMO) Magnitude of acceleration to separate sedentary behaviours from raw acceleration data collected from wrist/hip worn tri-axial accelerometers. [5] The ENMO magnitude of acceleration varies from 20mg to over 120mg. According to the paper [3], the movement across Angle-Z metric along with the ENMO magnitude of acceleration with specific bouts of 10 minutes help classify the pre-processed data as per three labels.

3. Univariate Feature Extraction

Once the data was segmented and labelled as per the activity, I was able to split all the measurement files for each user based on the activity label. For instance, all LIGHT labelled measurements were segmented into a new data-frame. By the end, each user's measurements were divided into three separate measurement frames.

Post the split, I performed a few univariate computational analyses to calculate new features such by identifying Peaks. Applied a band power calculating function to gather frequencies and spectral density by using Welch and Simpson's rule. Also, other univariate features were calculated for each of the tri-axial data point.

At the end of the univariate function, I calculated the average of each measurement file for a user subject. So, for instance if a user had 82 measurement files, wherein each file consisted of 50,000+ data points. An average of each of the measurement file was calculated and merged into one thereby transforming the dataset as completely new.

4. Feature Selection

In order to remove the collinear variables. I found this [6] link pretty useful. It makes use of Variance Inflation Factor to calculate the VIF multicollinearity score. Wherein, a high score denoted strong and high collinearity between independent variables.

I set a threshold of 10.0 to select only relevant features that have low co-linearity.

5. Machine Learning Algorithms

For the purpose of this assignment, I used two

machine learning algorithms random forest classifier and another Ensemble learning technique like the Voting Classifier to predict the Medication status, Tremor and Dyskinesia levels for each of the Test measurement of the subject patients.

The random forest builds a classification ensemble learning model that constructs multitude of decision trees and predicts the output based on mode of each class of classification. [7]

Voting classifier is an ensemble technique that trains on numerous models and for the purpose of this assignment I added the estimators as the Random Forest Classifier, Extra Trees classifier and Bagging Classifier. Based on these three classifying algorithms I designed a Soft Voting classifier to predict the classification class based on their highest probability. [8]

Based on these two multi-class classification problems the prediction ROC-AUC score was 0.549 and 0.449 for Random forest and Voting classifier respectively.

What were the problems?

1. Imbalance in dataset: This scenario was encountered when number of Dyskinesia, Tremor and medication on-off labels belonging to one class was significantly lower than other classes.

2. Computationally heavy univariate calculations: Initially I was using TSFRESH to automatically calculate many time series characteristics. Since I was using Google Colab and Jupyter notebook for the code development and with lack of a strong and heavy GPU available, the IDEs kept crashing.

3. Lack of GAIT Features: My human activity recogniser model was crude, in the sense that it used a conventional cut-point approach to identify activity labels. I did not consider human gait features i.e. locomotion achieved through movement of limbs. [9]. My features were based on univariate calculations and not identifying gait features which would have made the model richer.

Conclusion

To conclude, this was a learning experience in terms to understand how a real challenge project works and how in-depth knowledge of gait features is important to make healthcare predictions. My model though crude were rich in the sense that each data point from the tri-axial accelerometer and gyro-sensor data was calculated. I had not experienced data imbalance for multi-classification. Random forest proved efficient for making multi-class classification prediction.

Future Work

Currently I am developing a Convolutional neural network model to use processed signal images. I generated images from the processed tri-axial accelerometer and gyro-sensor data plots using Seaborn and Matplotlib libraries. I am using Transfer Learning. Because it is a technique that while resolving a problem stores knowledge gained for solving other problem. I aim to produce result of this model before the end of the 4th round of submission.

[1] "Synapse | Sage Bionetworks." https://www.synapse.org/#!Synapse:syn20825169/wiki/596118 (accessed Apr. 29, 2020).

[2] "UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set." https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones# (accessed Apr. 29, 2020).

[3] D. van Kuppevelt, J. Heywood, M. Hamer, S. Sabia, E. Fitzsimons, and V. van Hees, "Segmenting accelerometer data from daily life with unsupervised machine learning," *PLoS ONE*, vol. 14, no. 1, p. e0208692, Jan. 2019, doi: 10.1371/journal.pone.0208692.

[4] V. T. van Hees *et al.*, "A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer," *PLOS ONE*, vol. 10, no. 11, p. e0142533, Nov. 2015, doi: 10.1371/journal.pone.0142533.

[5] K. Bakrania *et al.*, "Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches," *PLoS One*, vol. 11, no. 10, Oct. 2016, doi: 10.1371/journal.pone.0164045.

[6] "multicollinearity - How to systematically remove collinear variables in Python?," *Cross Validated*. https://stats.stackexchange.com/questions/155028/how-to-systematically-remove-collinear-variables-in-python (accessed Apr. 29, 2020).

[7] "Random forest," *Wikipedia*. Apr. 17, 2020, Accessed: Apr. 29, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=951450315.

[8] "ML | Voting Classifier using Sklearn," *GeeksforGeeks*, Nov. 23, 2019. https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/ (accessed Apr. 29, 2020).

[9] "Gait (human)," *Wikipedia*. Apr. 14, 2020, Accessed: Apr. 29, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Gait_(human)&oldid=950885465.