Parkinson’s Disease Digital Biomarker DREAM Challenge

Junhao Wang, Xinlin Song, Arya Farahi

Michigan Data Science Team, University of Michigan - Ann Arbor

Introduction

Parkinson’s disease (PD) is a degenerative disorder of central nervous system that mainly affects the motor system [1]. Currently, there is no objective test to diagnose PD and the examination by a neurologist remains the most important diagnostic tool [2]. The examination is performed using the assessment of motor symptoms such as shaking, rigidity, slowness of movement and postural instability [3]. However, these motor symptoms begin to occur in very late stage [4]. Smartphones have sensitive sensors (accelerometer, gyroscope and pedometer) that can track the user’s motion more frequently than clinical examinations at much lower cost. Although the movement information is recorded by the sensors, the raw sensor data is hard to interpret and give limited help to PD diagnosis. We are working on a PD prediction application that predicts the probability of a person having PD based the phone sensor data. This application can be a good supplement method for neurologists to monitor the situation of a patient. It can also be helpful to determine whether the medication is effective. (If the medication is effective, the patient will do better than usual. The application predicts less probability of having PD.) Using the phone to record motion is easy to do and can be done during a period with low-cost. The phone sensor data may supplement the clinical examination results and help the neurologists diagnose PD at an earlier stage and know the progression of PD.

Data

The phone sensor training dataset comes from Parkinson’s Disease Digital Biomarker DREAM Challenge. The training data contain time series of mobile device acceleration and rotation data while the participant is walking and at rest. There are 2864 participants (660 PD, 2158 Non-PD/Control, 46 Nan) and ~35,000 walking records (10,000 Non PD, 25000 PD/Control). Each walking record has acceleration, rotation rate and pedometer data when the participants walk outbound, return and at rest. Each walking record has acceleration, rotation rate and pedometer data when the participants walk outbound, return and at rest. The records are of different length with sampling rate of 0.01 second. Figure 1 shows examples for PN and Non-PD/Control acceleration and rotation data.

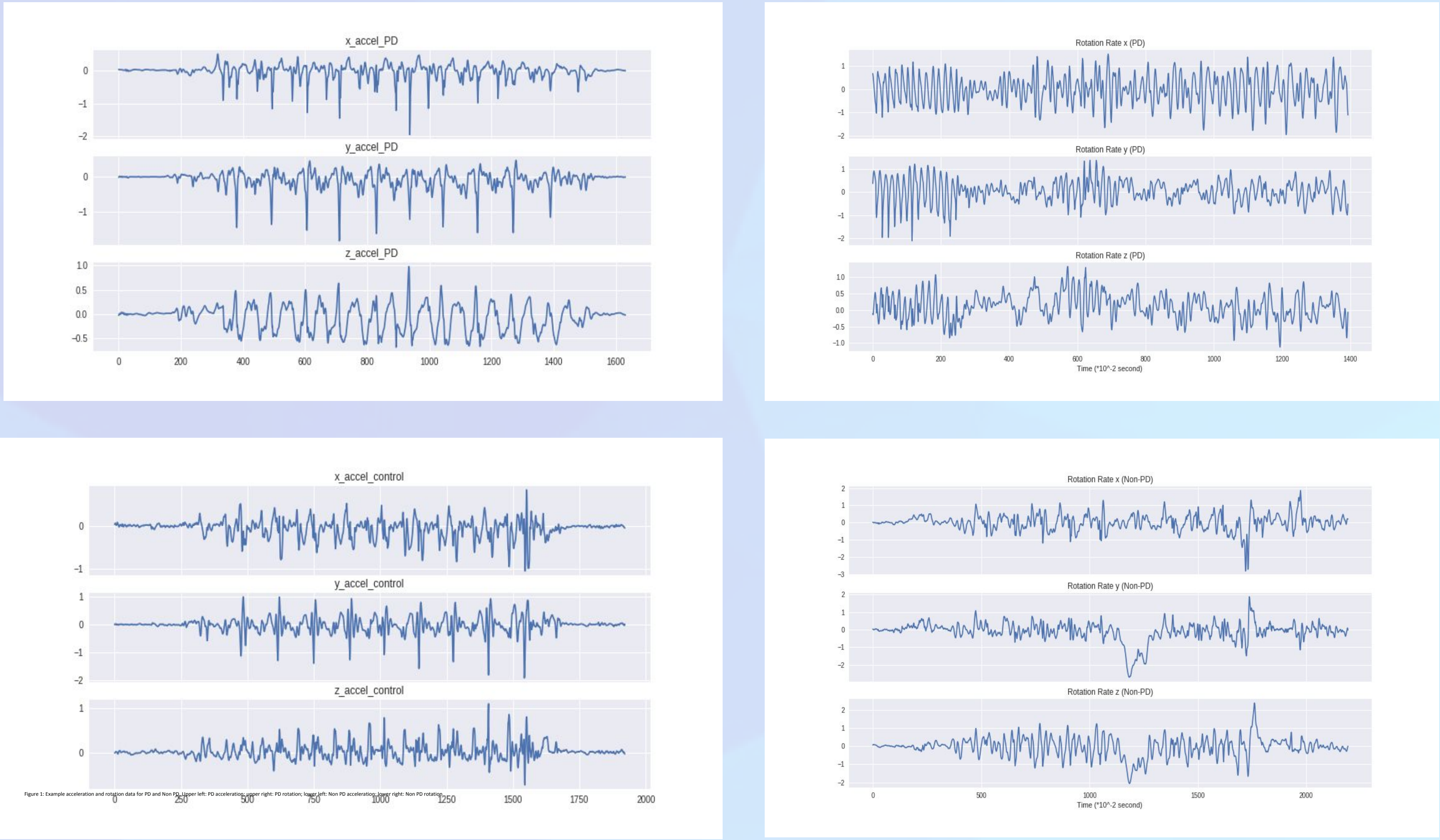


Figure 1: Example acceleration and rotation data for PD and Non PD. Upper left: PD acceleration; upper right: PD rotation; lower left: Non PD acceleration; lower right: Non PD rotation.

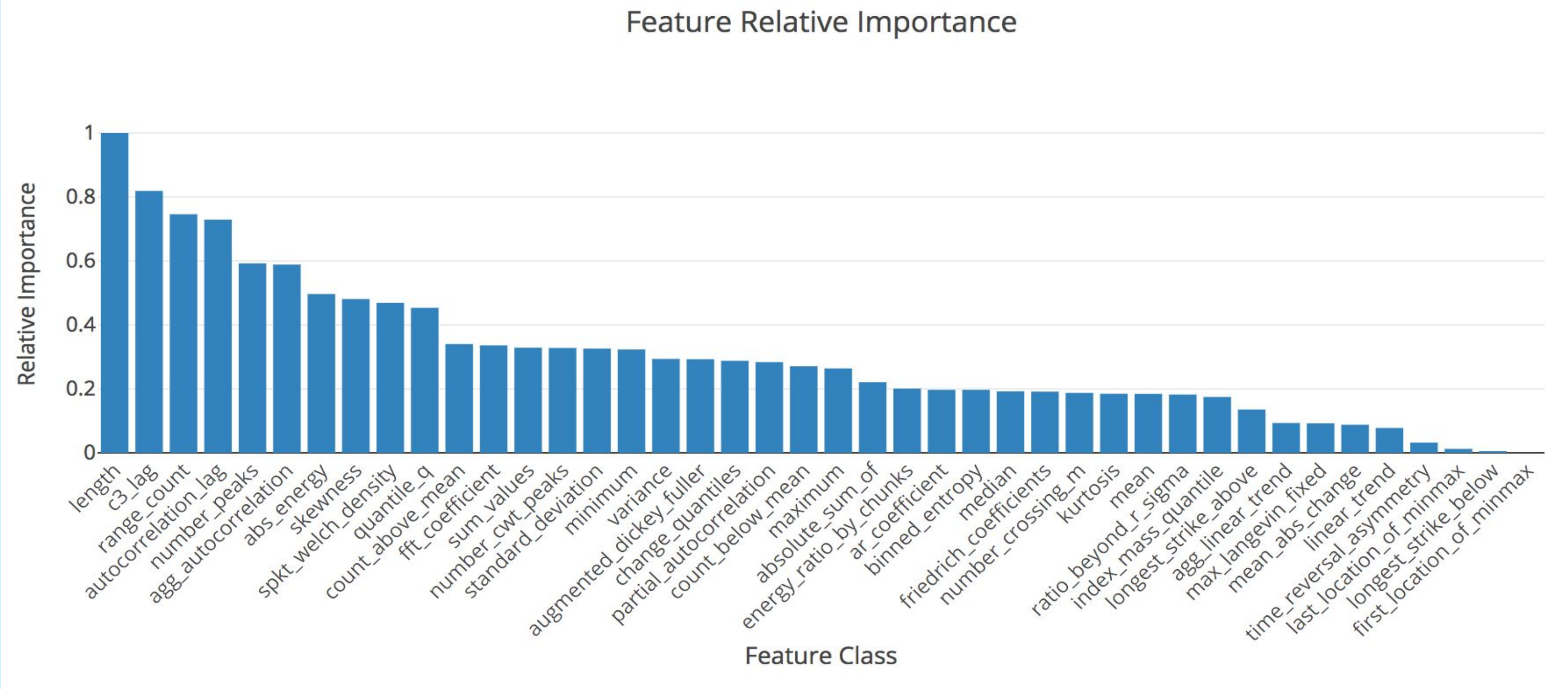
Methodology

First we preprocess acceleration data with Quaternion spatial transformation to convert phone reference frame to earth reference frame and drop all NaN data entries. Then three directions of research are carried out: (1) using global time series statistics from acceleration and rotation data and tree-based ensemble model to classify record into PD or Non-PD. (2) applying neural network classifier to learn from fixed length segments of acceleration and rotation data and classify record into PD or Non-PD. (3) identifying high frequent pattern in acceleration and rotation data and calculate morphological distance profile of each record using dynamic time warping (DTW) and classify record based on the distance profile.

Most emphasis is put on research direction (1) for three reasons: simplicity of feature engineering, independence of time series length, and interpretability. Research direction (2) and (3) were briefly experimented with limited success. With research direction (1) method, an exhaustive list of time series statistics are calculated for each record of 18 one-dimensional arrays: acceleration_x, acceleration_y, acceleration_z, rotation_x, rotation_y and rotation_z for walking_out, walking_return and rest. These statistics include approximate entropy [7], friedrich coefficient, max langevin at fixed point [8], continuous wavelet transform coefficient, Fourier transform coefficient, partial autocorrelation, etc. Non-parametric statistics are calculated directly and parametric statistics are calculated using a grid of commonly used hyperparameters. For each record of 18 one-dimensional arrays, 14,076 statistical features are calculated.

Then an importance order of these features are created by fitting an extra tree classifier or random forest classifier on the data (~20,000 records * 14,076 features). Then all data are grouped by person and each statistical feature is taken median across person. The number of records is reduced to the number of people (~3000). Then using top k features from the importance list, a logistic tree model is trained on 80% of the data (~3000*80%) and then tested on 20% of the data (~3000*20%). The training and testing are done multiple times, each with a random stratified split and average AUROC is calculated.

Most of the statistics calculations are computed using Python, with the help of Pandas, Jupyter, Numpy, Tsfresh [9]. Visualizations are created with Plotly. Model building and validation modules are created with Sklearn and Xgboost. Neural network classifier is built with Tensorflow. High frequent patterns in time series are created using Matrix Profile algorithm [10].

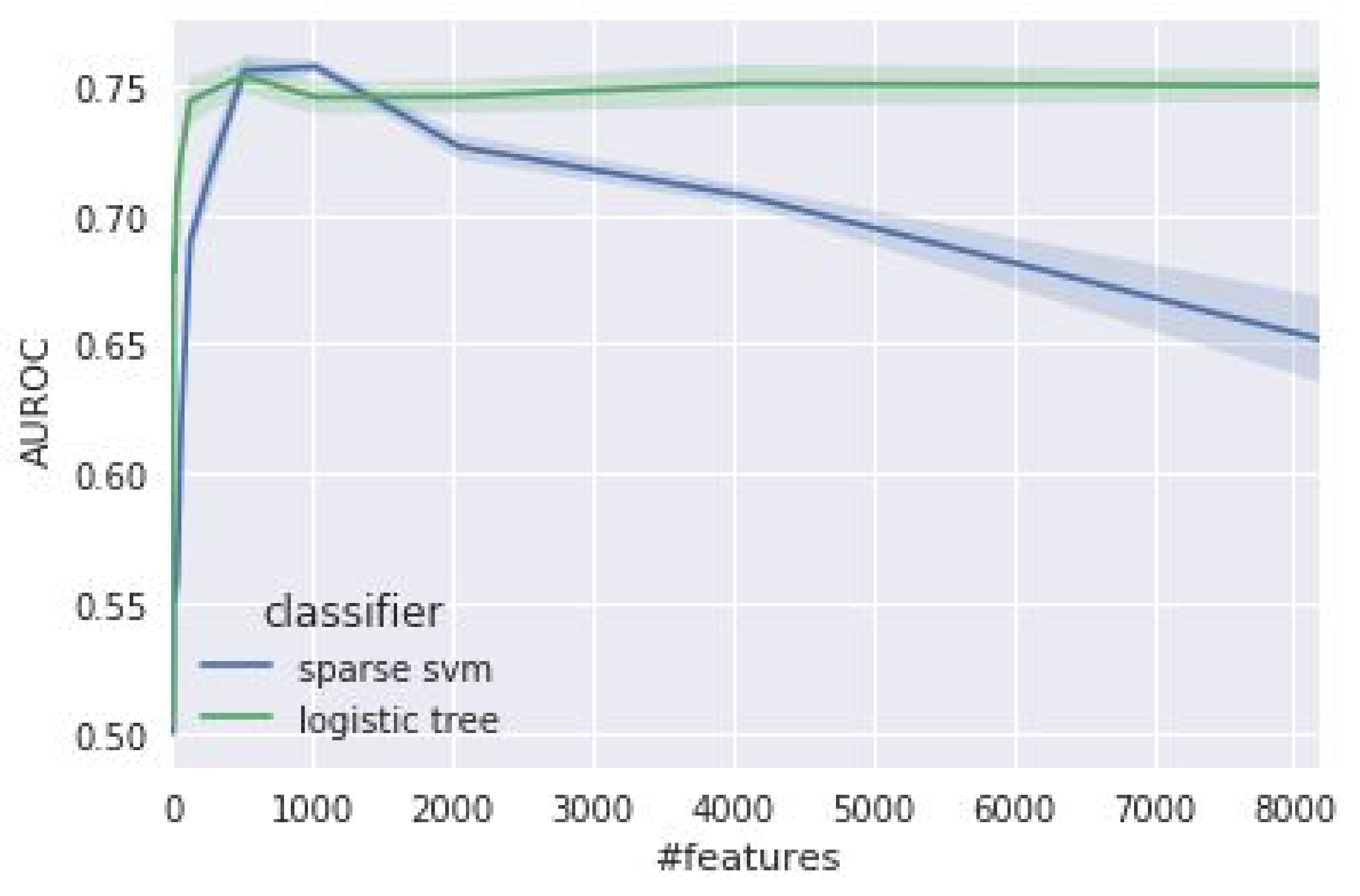


Next Step

Continue to explore direction (1) in two aspects: understand why certain statistical features are ranked more important than others, and compute multivariate statistics instead of single-array statistics. Also research more about direction (2) and (3).

Results

Since the positive and negative labels in the data set are not even, we use AUROC to evaluate the model performance. Logistic tree model consistently achieve AUROC of 0.75 after convergence.Comparatively sparse SVM drops performance once the highest AUROC is reached. The error bars are estimated using repeated train test split and evaluation 5 times.



Reference

[1] “Parkinson’s Disease Information Page”. NINDS. June 30, 2016. Retrieved July 18, 2016

[2] “Parkinson’s Disease Diagnosis Page”. National Parkinson Foundation. Retrieved Aug. 20, 2017

[3] Saba Emrani, Anya McGuirk and Wei Xiao. 2017. Prognosis and Diagnosis of Parkinson’s Disease Using Multi-Task Learning. KDD’ 17, 1457-1466.

[4] Diane B Miller and James P O’Callaghan. 2015. Biomarkers of Parkinson’s disease present and future. Metabolism 64, 3 (2015), S40–S46.

[7] Yentes et al. (2012) - *The Appropriate Use of Approximate Entropy and Sample Entropy with Short Data Sets*

[8] Friedrich et al. (2000): Physics Letters A 271, p. 217-222 *Extracting model equations from experimental data*

[9] Maximilian Christ, Andreas W. Kempa-Liehr, Michael Feindt (2017) *Distributed and parallel time series feature extraction for industrial big data applications*, arXiv:1610.07717 [cs.LG]

[10] Chin-Chia Michael Yeh, Helga Van Herle, Eamonn Keogh (2016). Matrix Profile III: The Matrix Profile allows Visualization of Salient Subsequences in Massive Time Series. IEEE ICDM 2016.

Acknowledgements

Authors recognize the support of Michigan Institute for Data Science.