Parkinsons Disease Digital Biomarker DREAM Challenge write-up for L-Dopa Submissions

Bálint Ármin Pataki
patbaa@gmail.com

Physics of Complex Systems Department,
Eötvös Loránd University,
Budapest, Hungary

November 3, 2017

Abstract

This paper contains the documentation of the idea and methods behind the code can be found at `https://github.com/patbaa/synapseParkinson` written for DREAM Challenge [1].

1 Feature engineering

My method had 3 main parts:

- creating features by hand
- creating many features with the python package tsfresh [2]
- selecting the best features

I followed this three steps for all the 3 subchallenges with smaller modifications.

1.1 Creating features by hand

1.1.1 Empirical features

Firstly I plotted the percentage of the positive samples $\frac{\#positive}{\#positive+\#negative}$ for different metadata conditions, see Figure 1. It seems that if the task is ftnr2 and the device is GENEActiv we have a pretty high hitrate. Visually analyzing all the possible metadata configurations I created 1-2 empirical features for each subchallenge.

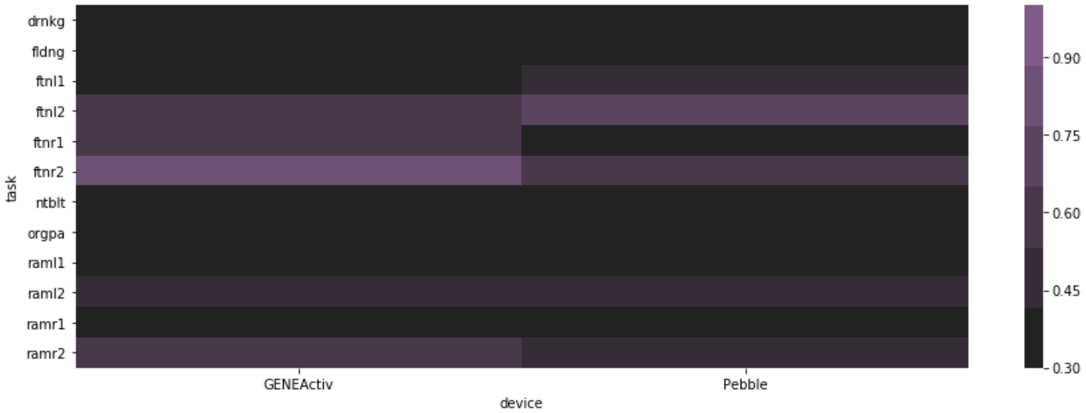

Figure 1: hitrate by device and task

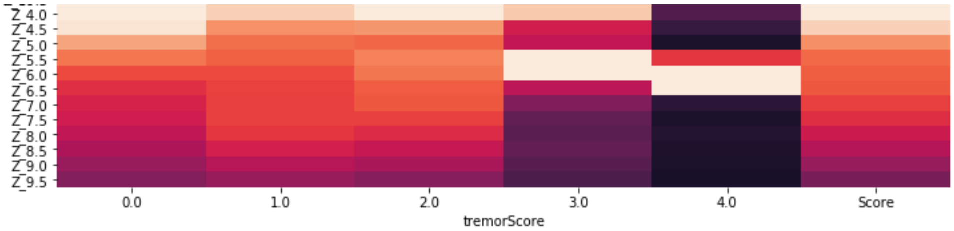

Figure 2: different power spectrum distribution averaged over different tremor score categories

Secondly I binaryzed the metadata except for the session and visit. So instead of having 'a', 'b' and 'c' for feature1, I had feature1_a, feature1_b and feature1_c. And feature1_c is 1 if feature1 is 'c' else it is 0.

1.1.2 Reasonable features

Then I created features that seemed to contain useful information, such as:

1. power spectrum of the acceleration data
2. power spectrum of the autocorrelation function
3. range and standard deviation of the acceleration data + correlation of the acceleration data

The idea behind them is:

1. The hand shaking could have a unique frequency profile that could be captured by the power spectrum.
2. This is a bit smoothed version of the previous feature.
3. Slowness and rapid shaking has pretty different range of acceleration and uncontrolled and controlled movement could have different correlation between its coordinates.

For the power spectrum related features I used binned frequencies (see Figure 2) and the sum of them for each coordinate.
For the range features and standard dev. features I used 80% and 20% percentiles to omit the outliers.

1.2 Tsfresh features

With the python packge tsfresh [2] I extracted 1000s of features. This package offers tons of features for time series.

1.3 Feature selection

I applied standard scaling for the whole dataset.

After having numerous features it is sure that many of them are not useful but some models doesn't tolerate the presence of many unrelated features so it could be fruitful to get rid of them. I used random forest model and selected the top features from it's feature ranking. For different subchallanges I found different number of features to be the best (around 20-200).

2 Conclusion

The metadata features seemed to be quiet powerful.

Untounched ideas:

- Proper ROI (region of interest) cut. Many tracks has junk at the beginning or at the end, see Figure 3
- Deep learning on STFT(short-time Fourier transform)/CWT(continuous wavelet transform) images.

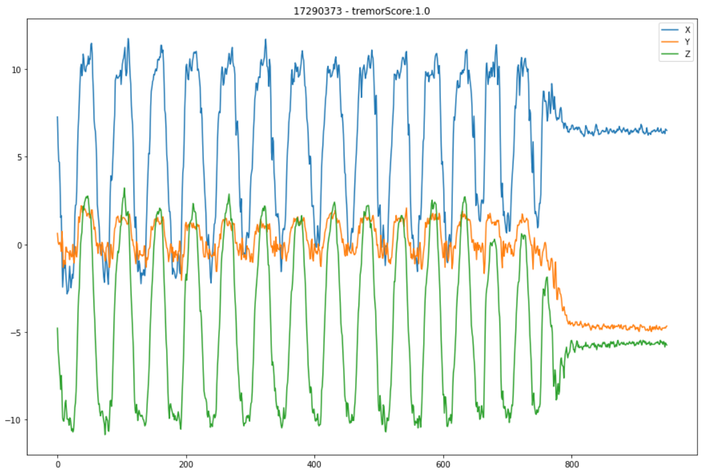

Figure 3: junk at the end of the track

3 Authors Statement

All work was done by Bálint Ármin Pataki.

References

[1] https://www.synapse.org/#!Synapse:syn8717496.

[2] https://github.com/blue-yonder/tsfresh.