Exploring the modeling and data imputation strategies on PPMI data

Nageswara Rao Gurram
University of Rochester
ngurram@ur.rochester.edu

Abstract

As part of this study, I've contributed for analyzing, visualizing, pre-processing, imputing and modeling the PPMI (Parkinson's Progression Markers Initiative) [1] dataset. As a first step, I've merged different files corresponding to UPDRS2 and UPDRS3 scores, patient features and demographics. Next, on this combined dataset some visualizations are plotted to understand the pattern and missingness in data. Due to the evident and lot of missingness, I've explored various imputation strategies from simple interpolation techniques to deep-learning based models [5]. As the data is time-series with interval between visits is irregular, there were some derived variables needed to be calculated such as 'Time from BL', 'Time from First Symptom' and 'Time from Diagnosis' to capture the temporal distance between visits. For modeling time-series in recent times, RNN (Recurrent Neural Networks) models are extensively applied with building blocks as LSTM and GRU cells as they help to capture long range dependencies. In this work also, I've modeled the problem as sequence-to-sequence derivation by giving patient past visits as input and predicting the future visits score. For this kind of problems (especially in NLP settings), encoder-decoder based models showed promising results. In this work, I've used this type of architecture to model PPMI data for predicting the future visits scores based on past visits. The code for the imputation method is at `https://github.com/nageshgurram12/ppmi_imputation` *and for modeling part is at* `https://github.com/nageshgurram12/pd_modeling_ppmi`.

1. Introduction

Parkinson's disease (PD) is a chronic, debilitating neurodegenerative disorder characterized clinically by progressive motor and various non-motor dysfunctions. The aim of this project is to improve the understanding and prediction of PD growth which in turn helps disease management and clinical trial design. For this purpose, I've used PPMI [1] dataset to analyze, model and evaluate the results. The entire study on this dataset can be divided into 3 parts:

- Data Merging
- Data Analysis & Visualization
- Data Imputation
- Data Modeling

2. Data Merging

As the PPMI dataset is split into multiple files for different types of feature sets, to analyze or model its required to merged the data into one file. Although the dataset is huge with multiple chunks of files, I've taken only files corresponding to MDS-UPDRS2, MDS-UPDRS3, PD features, Screening Demographics and Patient Status. This merge process needs to be done based on patient id and visit id columns and empty rows were created if some visits are missing only in some files. Although having more data helps for modeling, I've filtered the dataset based on cohorts and applied modeling techniques only for 'DeNovo PD' patients. It reduced the count of number of samples (patients) to 423.

3. Data Analysis & Visualization

Its always necessary to know about the data before going to model it. So as the next step, I've done some exploratory analysis on the merged data for predictor variables like UPDRS3 and Ambulatory scores. Visualizing the predictor variables always help for detecting any trends and seasonality in it. However, in this dataset there is no pattern observed for these predictor variables. Please refer Fig.1 for visualization of 'Ambulatory Score' and Fig.2 for 'UPDRS3 score' for some random patients.

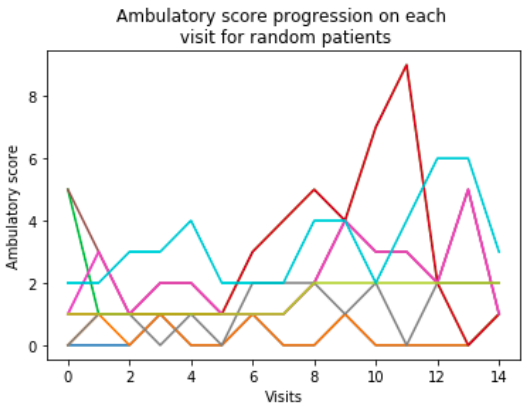

Figure 1. Progression of Ambulatory Score for random patients

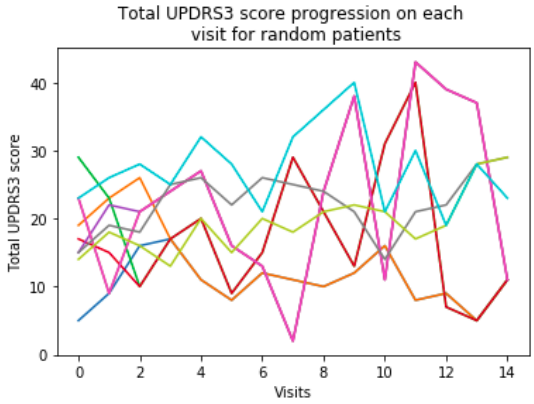

Figure 2. Progression of UPDRS3 Score for random patients

Data visualization also helps to discover any data missingness. So, I've plotted the visit presence graph [Fig.3] (dot if the corresponding visit exists) and visit counts [Fig.4] for some random patients. Its evident from the plots that many patients don't have all visits information and some imputation strategies are needed. In the next section, I've outlined a GAN based model to impute in this kind of data.

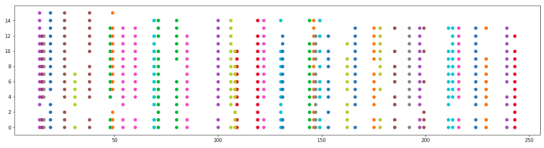

Figure 3. Missing data visualization for some random patients

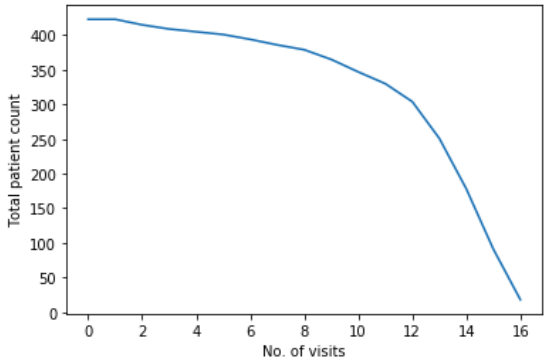

Figure 4. Patient count based on number of visits

4. Data Imputation

The kind of missingness here is different from what we see in other time-series datasets where few points may miss completely random in whole data but here a complete row will be missed(visit information of a patient) if exists. So, the imputation strategies must be smart enough to account for time difference between successive events for the imputation. Initially, I tried different strategies such as imputing missing data with constant (-1) or interpolating with moving average of certain window size. However, these results are not so good compared to the time when I model without any imputation strategies. Refer Fig. 8 for results with different imputation strategies.

This motivated me towards finding better imputation method for our kind of data. In recent times, GAN (Generative Adversarial Networks) [3] based models showed impressive accuracy in data imputation. In this direction, I've explored various architectures and found a closely matching work by Yonghong Luo *et al*. at [5] for multi-variate time series data. In nutshell, in this method they tried to impute the data by learning original distribution through 'Generator' and training it based on the feedback from 'Discriminator'. Both 'Generator' and 'Discriminator' are based on GRU units and trained with corresponding WGAN losses. Once the complete network is trained, the only 'Generator' model is again trained to impute in the left-out data based on reconstruction losses. Both networks are based on RNN network model with a modified GRU unit to account for time lapse between successive visits. For complete architecture details, refer [5] paper. Here, I've listed the architecture and losses in Fig. 5 and 6 for reference. The complete working code of the imputation model can be found at `https://github.com/nageshgurram12/ppmi_imputation`.

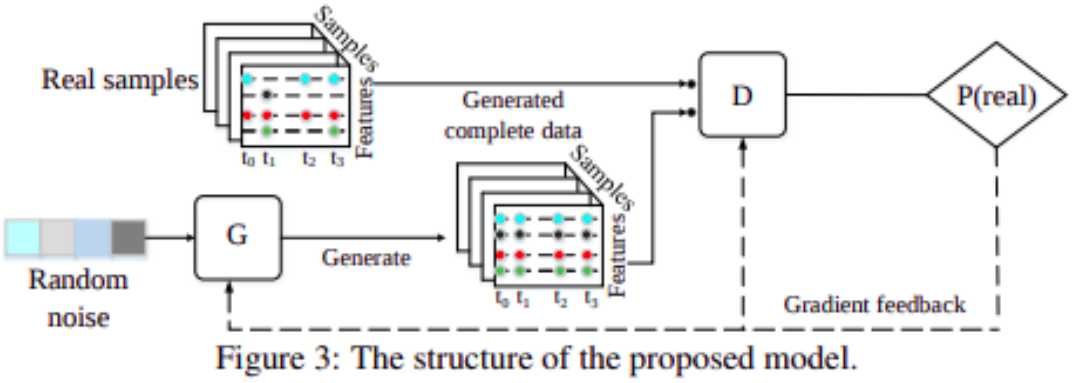

Figure 3: The structure of the proposed model.

Figure 5. Architecture of multi-variate time series data imputation by GAN

$$L_G = \mathbb{E}_{\boldsymbol{z}\sim P_g}\left[-D(G(\boldsymbol{z}))\right],$$
$$L_D = \mathbb{E}_{\boldsymbol{z}\sim P_g}\left[D(G(\boldsymbol{z}))\right] - \mathbb{E}_{\boldsymbol{x}\sim P_r}\left[D(\boldsymbol{x})\right].$$
$$L_r(z) = ||X \odot M - G(z) \odot M||_2$$
$$L_d(z) = -D(G(z))$$
$$L_{imputation}(z) = L_r(z) + \lambda L_d(z)$$
$$x_{imputed} = X \odot M + G(z) \odot (1 - M)$$

Figure 6. Different loss functions for imputation model

The evaluation for imputed data needs to be done by taking out some test data and running trained 'Generator' model on it.

5. Data Modeling

In this section, I've briefed the modeling work done on the PPMI dataset. As the data is multi-variate time series with complex distribution, I've applied RNN based model to predict the future scores. In recent years, for modeling sequence-to-sequence and long temporal sequences LSTM or GRU based networks showed promising results. Taking inspiration from Dzmitry Bahdanau *et al*. work in [2], I've modeled PPMI time-series as sequence-to-sequence prediction with Encoder-Decoder architecture. The Encoder takes the past visits and learns a overall representation of patient's disease growth and then supplying this to 'Decoder' network yields prediction scores for future. I've used 'Total UPDRS3 Score' and 'Ambulatory Capacity Score' as predictor variables for Decoder and complete feature set of single visit as input to Encoder cell. Both of them are based on Recurrent Neural Network architecture with GRU cell at each time step(visit). As the patient visit sequence length is not uniform, padding or imputation has to be done before giving as input. When Encoder is unrolled, it will have 17 total GRU units as the maximum number of visits for any patient is 17. The prediction sequence length for Decoder can be configurable but for predicting scores far in future is quite difficult, so its set to 3. For architecture reference, please look at Fig. 7.

The complete working code of the model can be found at `https://github.com/nageshgurram12/pd_modeling_ppmi`

For initial modeling tasks, I've considered only UPDRS2, UPDRS3 features and 'PD' cohort to predict the total scores. I've written the code to take different configurable parameters like cohorts of the patients, predictor variable, prediction sequence length etc.. In the initial training tasks, I've trained the model with only simple imputation strategies like padding missing visits with '-1' or taking moving average. For complete set of initial results on test data, please look at Fig. 8.

Data	Cohorts	Predictions	Imputation Strategy	Prediction Seq Length	Test error
UPDRS2, UPDRS3	PD	Total UPDRS3 score	Padding with -1	3	123.64
UPDRS2, UPDRS3	PD	Total Ambulatory Score	Padding with -1	3	8.43
UPDRS3	PD	Total UPDRS3 score	Padding with -1	3	66.75
UPDRS3	PD	Total UPDRS3 score	Moving average	3	576.68

Figure 8. Initial set of results with Encoder-Decoder model

6. Future Work

There is still lot of aggregating and novel work to be done for getting better prediction results. The to-do work can be iterated step-by-step like below:

- Evaluating GAN model imputation results on test data

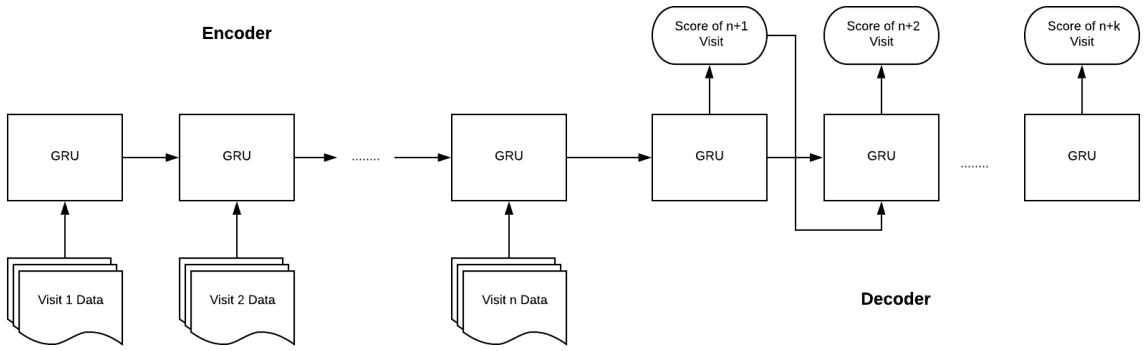

Figure 7. Encoder-Decoder model to predict future scores

- Using the imputed data to Encoder-Decoder model
- Tune the model with Attention mechanism. Yao Qin *et al*. proposed a 'Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction' [4] on financial data. However, this work also closely matches with the PPMI data and can be used to model for getting better results.

References

[1] Ppmi data - www.ppmi-info.org/data.

[2] Y. B. Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate.

[3] M. M. B. X. D. W.-F. S. O. A. C. Y. B. Ian J. Goodfellow, Jean Pouget-Abadie. Generative adversarial networks.

[4] H. C. W. C. G. J. G. C. Yao Qin, Dongjin Song. A dual-stage attention-based recurrent neural network for time series prediction.

[5] Y. Z. J. X. X. Y. Yonghong Luo, Xiangrui Cai. Multivariate time series imputation with generative adversarial networks, 2018.