Big Data for Computational Medicine Project Plan (Part-I)

Input: Raw feature tables of PPMI dataset.
Output: PD or healthy control prediction; H&Y stage prediction; MoCA scale prediction; Patient Subtyping; Ranked list of dominant features for each prediction task.

1. Data Manipulation

Record Concatenation. Concatenate the patient records according to their occurring time.

Data Imputation. To deal with missing values, imputation procedure was conducted. We used the last occurrence carry forward strategy for most of the missing values. Concretely,

- The first ever observed record of the patient was used if the first record of a patient was missing;
- The mean observed value of the certain feature across the entire population was used to impute if all entries of the feature of a patient were missing;
- For features with integer values, we rounded up such mean values;
- For categorical features, we transformed and normalized them into a set of one-hot input vectors.

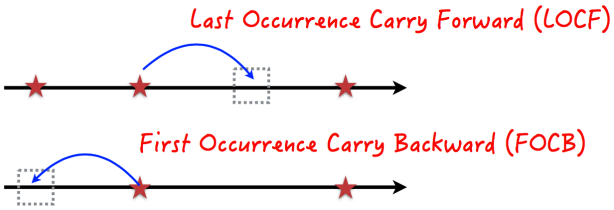

Figure 1. Imputation strategies

Besides, multiple imputation can also be used instead of the LOCF strategy. Specifically, the imputed values are drawn *m* times from a distribution rather than just once. One advantage that multiple imputation is that, multiple imputation can be used in cases where the data is missing completely, missing at random, and even when the data is missing not at random. Then we obtained m datasets, and utilize the pooling strategy to consolidate *m* results into one result. One multiple imputation method we can try here is Multiple Imputation by Chained Equations (MICE).

Feature Combination. Since PPMI is multi-source study that has various Parkinson's disease progression markers including demographics, clinical, imaging and biospecimen for more than six years, the features dimensions should contain above meaningful disease variables. Figure 2 shows

a simple illustration of incorporating the multiple data resources together. For each patient, the clinical records were extracted from patient data after imputation according to timestamps. As some of the records are continuous while others are categorical, we transformed the categorical features into one-hot dimensions. For example, if one variable contains three values: 1, 2, and 3, we encoded the values as 001, 010, and 100 respectively. For corresponding clinical variables at each timestamp, multiple records represented in fixed-length vectors can be obtained.

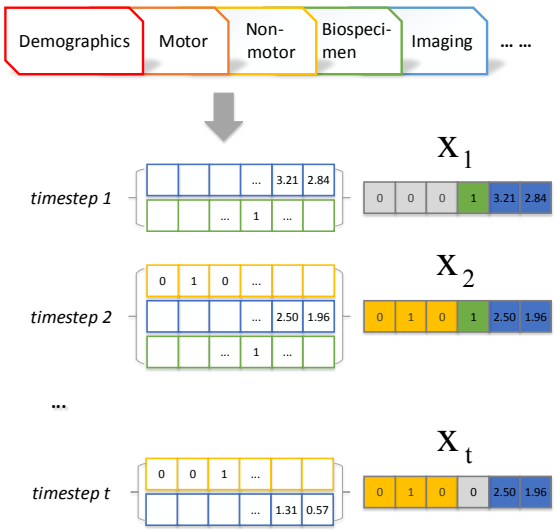

Figure 2. A simple illustration of the preprocessing for multiple data resources

2. Prediction Tasks

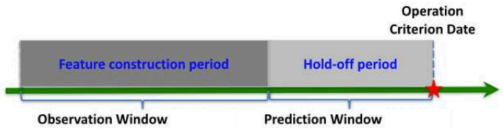

Figure 3. The experimental setting of the prediction tasks

Once the data imputation is finished, the sequential input vectors can be further aggregated by merging vectors of *t* timesteps. In Particular, the entire follow-up period can be divided into observation window and prediction window. The available patient features in observation window are utilized to train prediction models, while the prediction window is set for disease severity prediction. For the 6-year follow-up period, we firstly set the length of observation window as 4 years and set the length of observation window as 2 years. Thus, an input representation vector can be obtained for each patient with feature aggregation of observation window. Following the work [1], the disease status including Health Control (HC) and Parkinson's Disease (PD) can be

conducted. Also, With the aggregated patient representations, two prediction tasks including MDS-UPDRS score prediction and MoCA cognitive scores predictions could be conducted. The MDS-UPDRS and MoCA scores are capable to reflect the function disabilities for patients from motor and cognitive aspects, respectively. Therefore, they are important clinical phenotypes in Parkinson's Disease.

- MDS-UPDRS score or H&Y stage prediction: MDS-UPDRS (Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale) is a sort ß clinical scores that plays a critical role in predicting diagnosis of PD. The MDS-UPDRS has four parts, namely, I: Non-motor Experiences of Daily Living; II: Motor Experiences of Daily Living; III: Motor Examination; IV: Motor Complications. Previous clinical studies have shown that the motor assessment results apparently associate with other variables. Thus, we can use these variables to predict MDS-UPDRS. Alternatively, we can choose to predict H&Y stages (0-5) that is a clinical validation of the level of motor impairment.

- MoCA score prediction: cognitive performance of patients can be assessed by Montreal Cognitive Assessment (MoCA) scores. Similar to MDS-UPDRS, we can predict the MoCA scores by demographics, clinical data, APOE status, and biomarkers including CSF and dopamine transporter imaging results (DAT). In previous work[2], logistic regression model is exploited in model for prediction of cognitive impairment in patients with newly diagnosed Parkinson's Disease. The results of ten-fold cross validation confirmed the final model.

Objective function:

- Logistic regression:

$$\sum_{t}\sum_{j=1}^{m_b}\log\left(1+\exp\left(-y_{t,j}^{b}(w_{b,j}^{T}x_t)\right)\right)+\lambda||W_b||_{2,1}$$

- Linear regression:

$$\frac{1}{2}\sum_{t}\left\|y_t^g-W_g x_t\right\|_2^2+\lambda||W_g||_{2,1}$$

W_b and W_g are parameters needed to be learned from the regression models. The reason we resort to two different objective functions is that the target values could be binary (logistic regression) or continuous (linear regression). t indicates the index of patients here.

Machine learning based on logistic regression with penalization by a least absolute shrinkage and selection operator: Regression with least absolute shrinkage and selection operator. Figure 2 shows an example of factor selections by the learned parameters. The assumptions behind l-21 norm is that not all the variables have contribution on the prediction task.

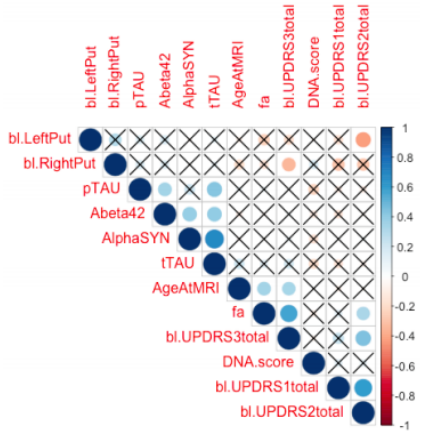

Figure 4. Example of factor selections (from PPMI DATA CHALLENGE 2016, D. Tosun et al.)

3. Clustering Tasks

Basic task. Most of the previous PD studies group patients into subtypes using k-means cluster analysis[3]. For the PD cohort, the task can further be divided into a 2-class solution as well as a 3-class solution. In order to cluster patients into subgroups, the representation for each patient is obtained by aggregating all the sequential vectors at timestamps. Note that patients may have different number of records, normalization before k-means cluster analysis is necessary. Hence, t-Stochastic Neighbor Embedding (t-SNE) can be conducted to visualize the clustering results.

Advanced task. Instead of merging sequential vectors into one representation vector, the imputation results are used directly in patient clustering task. The following figure shows the designed clustering procedure.

In particular, there are two steps in order to cluster patients using the sequential representation: Dynamic Time Warping (DTW) and t-Stochastic Neighbor Embedding (t-SNE). DTW is an approximate pattern detection algorithm that measures similarity using a dynamic programming approach to minimize a predefined distance measure. It can optimally align two sequences and learn pairwise similarity of every two patients. On the other hand, t-SNE is widely used as a visualization method in various machine learning tasks. The core idea is to reduce the dimensions and approximate the input similarity matrix by minimizing the Kullback-Leibler divergence.

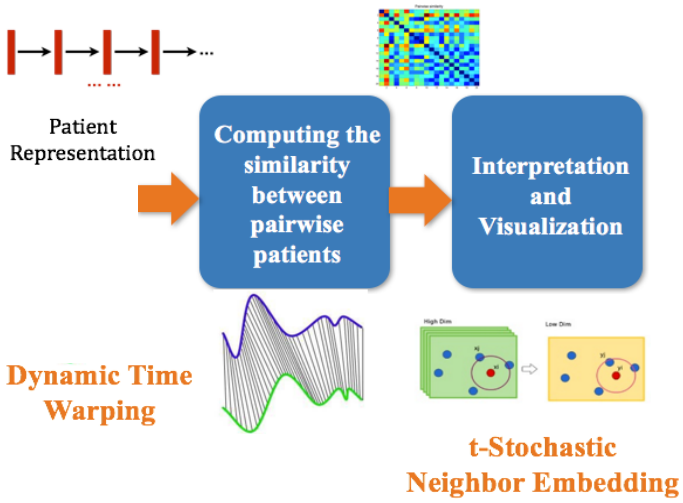

Figure 5. Clustering task procedure

4. Statistical Analysis

Statistical Study can be conducted on both prediction and clustering results.

For clustering task.

- Statistical analysis on each interpretable PD variable can be used to characterize the clusters as meaningful subtypes. Namely, obtained clusters are compared using Chi-square test for the categorical variables, one-way for the normal continuous variables, and Kruskal-Wallis test for the non-normal continuous variables, and Fisher's exact test for the high sparsity variables. For the tests with significant p-value, Tukey post hoc analysis were performed on each of two subtypes to find specific difference.
- To demonstrate that the clusters are progressive subtypes. The changes of mean values of some prevalent variables be draw to show slow-progressing subtypes as well as fast-progressing subtypes.
- According to the significance of p-value, we can not only find the dominant disease variables of clusters, but also get a ranking list of the variables that indicates the discriminative degrees of the variables. Eventually, the top characteristics are summarized.

For prediction task.

- Using change in MoCA scores over years, for example, 2 years, MoCA scores at 2 years' follow-up, and a diagnosis of cognitive impairment (combined mild cognitive impairment or dementia) at 2 years as outcome measures, the predictive values of baseline clinical variables and separate or combined additions of APOE status, DAT imaging, and CSF biomarkers can be assessed. We did univariate and multivariate linear analyses with MoCA

change scores between baseline and 2 years, and with MoCA scores at 2 years as dependent variables, using backwards linear regression analysis.

Reference

1. Dinov, I.D., et al., *Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations.* PloS one, 2016. **11**(8): p. e0157077.
2. Schrag, A., et al., *Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study.* The Lancet Neurology, 2017. **16**(1): p. 66-75.
3. Post, B., et al., *Clinical heterogeneity in newly diagnosed Parkinson's disease.* Journal of neurology, 2008. **255**(5): p. 716-722.