Explanable Projection: Application to PREDICT-HD

PROPOSAL

Sungmin Hong
sungmin.hong@nyu.edu (sh4174)

What is the problem you want to solve and who has this problem?

Huntington's disease(HD) is a fatal neurodegenerative disease which is caused by a known genetic pathology in Huntington's protein. Clinical symptoms of HD include motor, cognitive and psychiatric abilities of patients. HD diagnosis is determined based on unequivocal movement disorder. Despite of an utmost demand of identifying features which can be used as a biomarker of the HD progression prior to disease manifestation, significant features have not yet been discovered. Most of previous HD studies relied on brain sub-region volumes, especially in the striatum, to show the correlation of the degeneration of brain and HD progression. Although the degeneration is most significantly observed in the striatum, the other parts of brain also degenerate while HD progresses. Also, the brain sub-regions are not independent from each other since they are all connected and adjacent. HD does not only affect the particular sub-regions of brains as already discovered, but it might also affect the other parts of brain. Thus, the relationship of changes in brain sub-regions and HD progression is complicated since there are numerous sub-regions which results in high dimensional features.

CAP score, which is the product of age and a gene status, is suggested as the estimated HD progression score in clinical studies. Prodromal HD patients are classified into different risk groups (controls, low, medium, and high groups) based on their CAP scores. Although a few recent studies have been published to support the relationship between CAP scores and prodromal HD progression, solid proofs have not been suggested. Thus, the clinical risk group analysis based on CAP score needs more supports from data to prove the correlation to prodromal HD progression.

To summarize, there are two main problems in HD study which we want to tackle in this project as follows,

1. ***High dimensional brain sub-region features and their correlations with respect to HD progression***
2. ***Validation of clinical risk groups***

What questions do you want to be able to answer with your visualization?

In this project, we will develop a visualization framework which explains following problems.

1. What brain sub-region volumes measured from medical imaging discriminately affect unsupervised clustering results
2. How brain sub-region volumes are correlated
3. What brain sub-region features can be used to generate the best clustering results.
4. How much data-driven clustering results are related with clinical risk groups

What is your data about? Where does it come from? What attributes are you going to use? What is their meaning? What are their attribute types (data abstraction)? Do you plan to generate derived attributes? If yes, which and why?

PREDICT-HD is a NIH funded project which aims to document the natural history of prodromal HD progression in the largest cohort ever studied and to develop neurobiological predictors of HD progression. There are 321 participants with 857 MRI T1 scans (64% female, 36% male). The number of scans of a control group (Non-HD) is 243, and the number of risk group scans is 614. Most of data attributes are quantitative, including age, brain sub-region volumes, CAP score, and motor score. There are also a few categorical attributes, such as, gender, diagnosis (whether a patient is diagnosed as HD or not eventually. CAP group is an ordinal attribute, which is derived from CAP score.

Attribute Names	Attribute Types	Explanation
Age	Quantitative	Age of a patient
Brain Sub-region Volumes	Quantitative	Volumes of brain sub-regions of a patient. Total 137 brain sub-regions are segmented and manually cleaned
Gender	Categorical	Gender of a patient
CAP Score	Quantitative	Clinically estimated disease burden score by genetic impairment of age
CAP Group	Ordinal (Derived from CAP score)	Clustered risk groups by thresholding CAP scores into 4 groups (Control, low, medium and high groups)
Motor Score	Quantitative	Definitive measure to diagnose HD by observing patient's movement

Other than given data from PREDICT-HD database, we will derive the correlation between volumetric features by conventional correlation measures, such as, Spearman correlation or Pearson correlation to show how features are related and affect each other. To see what features play significant roles for clustering, we will derive feature importance on data-driven clusters.
The data-driven clusters will be also a derived attribute (categorical) which will be compared with CAP group, which is a clinical clustering result, to see how much the data-driven clusters coincide with clinical clusters. Since there is no way to order the data-driven clusters in a desirable way like given CAP groups, the data-driven clusters are categorical, not ordinal.

Attribute Names	Attribute Types	Explanation
Correlation between Volumetric Features	Quantitative	The correlation between brain sub-region volumetric features.
Feature Importance	Quantitative	Feature importance of brain sub-region volumetric features for a given cluster
Data-driven Cluster	Categorical	Data-driven clustering results by using all or selected features

What have others done to solve this or related problems?

There are no directly related works published in my limited knowledge. Paulsen et al. showed that the statistical significance of the correlation of volume changes in basal ganglia regions and HD progression [1]. Like Paulsen et al., the most of HD studies have focused on revealing the relationship between HD progression and an individual feature or a few selected features rather than using dimensionality reduction techniques, or change feature selections. Stahnke et al. suggested an interactive framework which aims to interpret dimensionality reduced data [2]. The suggested framework allows users to select data and look into the features with respect to selected data more closely to understand why the data are clustered together, or why they have small or large distance in the given projection space. Although the feature correlation and the axes of the projection space were not well defined, Stahnke's work shows comprehensive ideas how high dimensional data are projected on 2D space.

What solution do you propose? How does the solution help you answer the questions stated above?

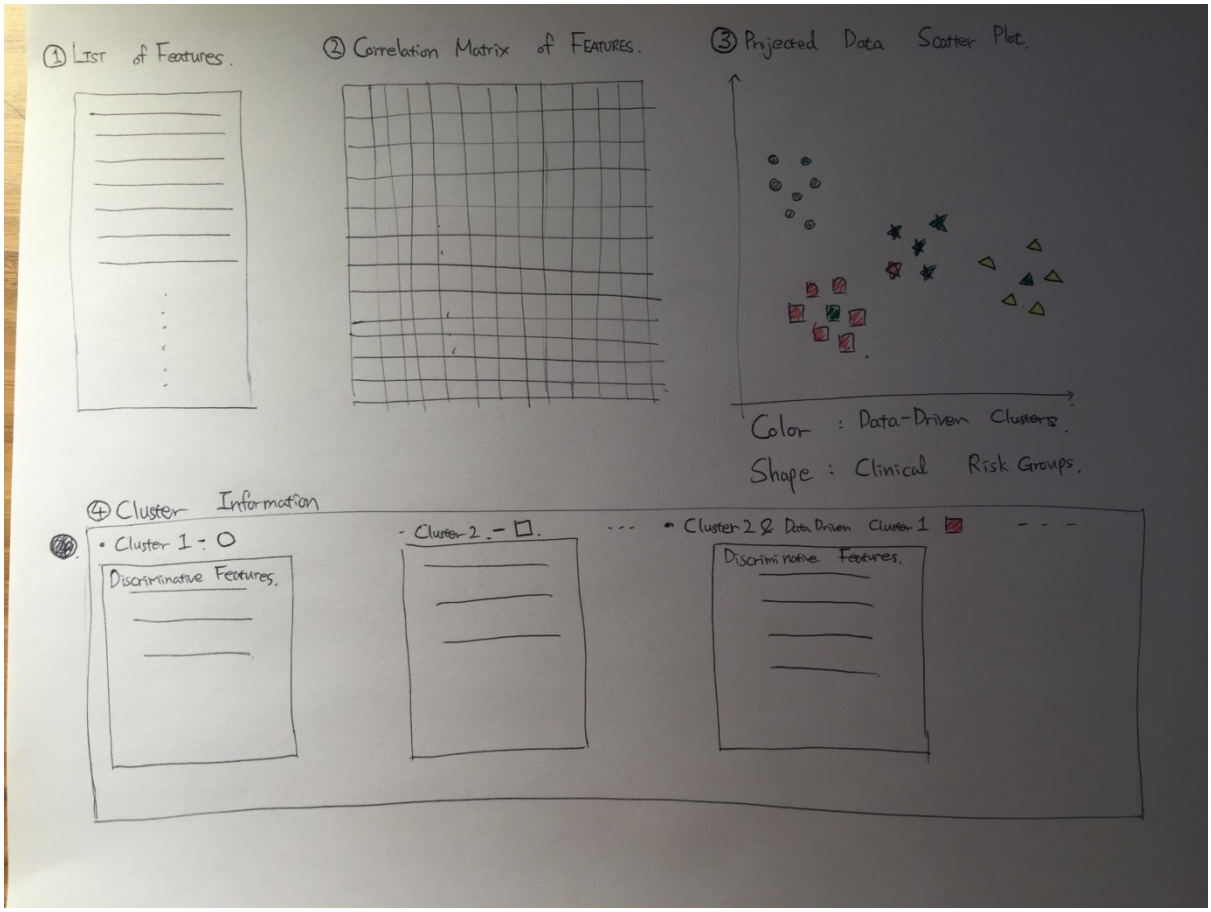

Figure 1. The mock-up of the suggested framework.

Figure 1. shows the mock-up of the suggested framework. The framework consists of four parts. In the first section, we will list the sub-region volume features in the predefined order. The list can be sorted by individual feature importance or the volume values. To show the correlation between features, we will plot the correlation matrix of volumetric features by ordinal color. Since, the number of volumetric features is 137, it is expected to be visualized in a single matrix with an adequate scale. The third scatter plot will display the projected data

on 2D space and data-driven clusters and clinical risk groups. The data driven clusters will be represented by categorical colors since the number of clusters can vary. The clinical risk groups will be displayed by shapes since it only has four different groups. The colored shapes will display how the data-driven clusters and the clinically decided clusters are coincided, or, at least, related with given data points. In the fourth part, we will show the cluster information from either data-driven clusters, clinical clusters, or both. The most discriminative features will be displayed for each cluster to show what features devote most to determine the cluster. We can list data-driven clusters by colors, clinical clusters by shapes, or even clusters defined by both data and clinical disease burden to show what features decides the clusters. In this way, we can show what features can be important to form data-driven clusters and also clinical clusters and how the features are related by visualization. Also, by selecting a few features in the first list or the second correlation matrix, we will do the clustering again only with the selected features to see the effect of selected features.

How do you plan to verify whether you have met your goals with this project?

We will make use of CAP score, CAP groups to verify how well data-driven clusters and clinical clusters are coincided. Clusters generated by selected features will be examined by a same manner with CAP score and CAP groups. The motor score will be used to verify both data-driven clusters and clinical clusters. If the data in a single cluster has similar motor scores, it is likely to believe that the cluster has similar HD progress. In the fourth cluster information section, we can also analyze what features are important and match them with previous literatures. For example, it is suggested that basal ganglia area has significant correlation to HD progression in the previous studies. We can compare the knowledge from medical society with our observations to see any insights we can find from our framework.

How is your project team going to work on the project? Who is going to do what?

Since this is a solitary project, there will be no need to divide the work.

References

1. Paulsen JS, et al., “Clinical and Biomarker Changes in Premanifest Huntington Disease Show Trial Feasibility: A Decade of the PREDICT-HD Study,” Front Aging NeuroSci, 6:78, Apr 2014.

2. Stahnke J, et al., “Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions,” IEEE TVCG, 22:1, Aug 2015.