Introduction

Improving the quality of the diagnostics is essential in modern life. Although neuroimaging tools become available to use in everyday life in clinic, the potential of these methods is still not fully unveiled.

In my Coursera Capstone Project I propose to test a following hypothesis on openly available data: not only the psychological and demographic data (i.e. age, education, SES and MMSE) could help us to diagnose neurodegenerative disorders, but it is also possible suggest a diagnostic tool based on the application of machine learning in neuroscience. I propose to set at openly available dataset with data derived from MRI scans and psychological, demographic characteristics conduct an analysis with Support Vector Machine and Logistic Regression to discover, if the combination of brain and psychological data allow better classification between patients and healthy controls, than psychological-demographical data only.

Such tool could potentially be of interest to private clinics, hospitals and medical centres who specialize mostly on preventive care and diagnostics.

Data description

Data used in this project were found at Kaggle as a open database - available under Creative Commons CCO: Public Domain license.

https://www.kaggle.com/jboysen/mri-and-alzheimers

Longitudinal MRI Data in Nondemented and Demented Older Adults: This set consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer’s disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

For the analysis, the following data used: Age, Gender, Education, Handness, MMSE, CDR, eTIV, nWBV.eTIV, nWBV comprise brain data, while Age, Gender, Education, Handness, MMSE, CDR comprise psychological examination and demographic data.

Methodology

We used these sets of variables to predict the Group of the current data point (assigned to the patient) – diagnosed with dementia or not.

The data are evaluated with Support Vector Machine and Logistic Regression to obtain a better predictive model by comparison. Application of Logistic Regression is possible because the dependent variable(target) is categorical. It also allows us to estimate a probability of predicted value to be actual value, which provides an advantage in comparison with Support Vector Machine.

Results

1) Support Vector Machine.

We analyzed the complex set of variables (brain+test data) to see, if the complex model is working well in accordance to our prediction. However, the precision obtained for a class "nondemented" was 0.58, while for class "demented" the model failed (precision = 0.00, see ipynb for the illustrations). F1-score of the model was 0.58, so we assume low accuracy and precision for this particular model.

Thus, we switched to second model to test three combinations of independent variables "brain+test" (all variables), "brain" (only MRI–derived data) and "test" (psychological and demographic charasteristics).

2) Logistic Regression.

2.1. The precision of the model on the parameters' set "brain+test" was the highest: 0.96 for the class "demented" and 0.90 for the class "nondemented". The accuracy estimated by f1-score was equal to 0.92.

2.2. When the parameters' set "brain data" only was taken into account, the precision and accuracy of the model dropped significantly. Data points were almost equally distributed for four cells of confusion matrix (see ipynb for the illustrations)

2.3. However, when the parameters' set "test data" only was taken into account, the precision and accuracy became comparable to the results of the model with the complex parameters'set – "brain+test" (see ipynb for the illustrations).

Conclusion

So, if we compare confusion matrices for "brain&tests", "brain only" and "tests only" models, we will conclude that the "brain&tests" and "tests only" provides comparable results, while "brain only" model is not sufficient. Thus, based on this dataset we could assume that the brain data without the test data do not provide sufficient evidence to classify these two groups of observations from the current dataset. Therefore, our hypothesis was only partially confirmed, and the diagnostic tool for neurodegenerative disorders may be comprised from the psychological–demographical data, as well as brain data, with the prevalent weight of "test data" (psychological–demographical) component.