Initial Plan – Identifying diagnostic features of Parkinson's disease using classical machine learning.

Author: Neofytos Neokleous (C16150333)

Supervisor: Matthias Treder

Module Number: CM3203
Module Name: One Semester Individual Project (40 Credits)

Ethics:

This project has undergone ethics approval by the Health Research Ethics Committee of the University of Stellenbosch. The University of Cardiff Ethics Committee has also been informed about this project; the application will be considered at the upcoming committee meeting. The Shared Roots data includes metrics from the different patients' brains as well as general information about them such as patient demographics and other general health metrics.

Project Description

In today's society more and more people suffer from diseases that are related to an abnormal function of the brain. Neurons (one special type of cells) are the basic working unit of our brain and neurons' function is to transfer information in the form of signals. Yet, we are still not able to know exactly how the human brain functions as causes and treatment of some brain diseases are currently being researched.

Through structural MRI (Magnetic Resonance Imaging) of the brain, scientists are able to obtain more information about the brain's function and structure. The way that MRIs work is by creating a strong magnetic field in the area of interest. These leads to the cells of that area to be aligned as each cell's nucleus has a positive charge (protons in the nucleus). Radio waves are then sent to the same area; these are received by a detector on the other side causing the protons to get out of alignment. Subsequently, a detailed map of where each proton is located can be constructed and based on that we can create a 2D or even 3D images. In my project extracted preprocessed tabular data from the MRI brain images that has been partitioned into regions-of-interest (ROIs) will be used in order to gain valuable insights about the 'diagnostic features' of Parkinson's disease patients. ROIs are subsets of a dataset that help to summarize characteristics of the brain such as cortical thickness or grey matter volume.

The clinical population in this project consists of patients who have Parkinson's disease (PD) originating from a number of causes such as medication side effects, genetic predisposition and environmental factors. Parkinson's disease is one of the diseases which are related to the loss of nerve cells in a particular part of the human brain (substantia nigra). The malfunction of those cells have as a result the reduced production of dopamine; a chemical related to the contraction and relaxation of muscles. Statistics show that about 1 in 500 people is affected by the disease and most of them are over the age of 50. Unfortunately, there is not a cure that can treat successfully Parkinson's disease but some helpful measures can be provided in order to reduce the effect of the disease such as medication or even brain surgery.

It has been observed in many cases the concurrence of PDs with other medical conditions such as cardiovascular disease (CVD) and metabolic syndrome (MetS). Up to now, it is known that the development of MetS is an important predicator of cardiovascular disease (CVD). The purpose of this project is to build a predictive model using classical machine learning in order to identify the diagnostic features that make patients to suffer from PD. As the molecular pathophysiology relating those conditions led to the disease is poorly understood, my model would try to spot the features that make people suffer from PD and hopefully help to prevent the development of CVDs.

My research will be based upon a clinical population consisting of 600 patients who suffer from one type of neuropsychiatric disorders (NPDs). The neuropsychiatric diseases that can be found in the clinical population are Post - traumatic Stress disorder (PTSD), Parkinson's disease (PD) and schizophrenia. For each patient, a variety of data points is held including genomics, transcriptomic, epigenetics and complementary phenotypic and multimodal neuroimaging data. Within the sample, we have identified 200 patients with PTSD, 200 patients with a diagnosis of Parkinson's disease (PD) and 200 patients with a diagnosis of schizophrenia. In addition, 300 of those met the criteria for MetS and 300 of those don't met the criteria for MetS.

My goal is to compare and identify the best classifier method that helps to identify patients with PD based on the preprocessed ROI data. Classifiers are methods which can predict the class of an entry based on its data points in a way that a given function (f) with input variables (x) can give some discrete output variables (y). Classifiers fall into the category of supervised machine learning meaning that the target is provided with the data and in that way the algorithm can learn to classify the entries based on experience. After deriving the different classifiers, the evaluation process needs to be carried out in order to examine the effectiveness of each one. Evaluation methods such as cross-validation and ROC (Receiving Operating Classifiers) curve are the most suitable for the task. Cross-validation involve the division of the dataset into smaller subsets where one would be the test set and the others would be the training set. Using the training set the model will be trained using the classifier algorithm and then its performance will be evaluated using the test set. This will occur in many different iterations by switching the test and train sets and then get the average results out of all the iterations. ROC curve validation is used to visualize the comparison between the different classifier models in a way that the difference in their performance can be observed. The graph is used to show the trade-off between the true positive and false positive rate of each of the classifiers used. The area under the curves defines the accuracy of the different models and a model with a perfect accuracy has a value of 1.0. In addition to the above, the project will include the comparison of different feature selection methods that can help to increase the performance of the different classifiers. Feature selection methods involves the selection of the features of the dataset that contribute the maximum to the performance of the model. For example, feature selection methods will help to identify an area 'x' of the brain when the presence of grey matter volume in 'x' is strongly correlated with the occurrence of PD.

Project Aims and Objectives

The aim of my project is to use different classical machine learning classifier methods and identify the one that performs better on the given dataset. In the project 5 different classifiers are going to be developed and evaluated. Following that, a comparison of feature selection methods such as feature importance in random forests and activation patterns for linear methods will be used in order to investigate how consisted different methods are at identifying diagnostic features. Feature selection techniques will enable the developed models to gain significant increase in performance. Both the classifier methods and feature selection methods will be developed using predefined python libraries such as Tensorflow which is the low level mathematical library and Keras which is a high level neural network API that can run on top of Tensorflow.

Set aims of the study:

- Primary aim for the thesis will be the comprehensive comparison between different classifiers with an appropriate visualization of the results
- Secondary aim of the thesis will be to compare feature selection methods such as importance in random forests and activation patterns for linear methods. The reason for the comparison is to identify the consistency of different methods at identifying the diagnostic features.

As a first step, I will start reading about the use of classical machine learning in similar projects. Through research, I will also identify the different types of classifiers that can be applied to the project such as LDA (Linear Discriminant Analysis), Logistic Regression, SVM (Support Vector Machine) and Random Forest. Furthermore, a research will be conducted in order to identify the feature selection methods that can be used in the project.

Set objectives of the study:

- Conduct research on python libraries that help in the analysis of the dataset (Tensorflow, Keras, SciKit, Pandas, MatPLotLib)
- Conduct research on classical machine learning classifiers
- Identify the best suited classifiers that can apply to the project and apply them.
- Evaluate the classifier models that have been developed using cross-validation and ROC curve.
- Conduct research on feature selection methods and identify which methods will be more suitable to use in the project.
- Select the feature selection methods and evaluate their effectiveness on the models already developed.

Work Plan

Deliverables

Before the deadline of the project (7th May), all the work that has been done through the semester should be finished and well prepared. That work involves the final report, the code developed to analyze the data, any supportive material that have been used and finally any figures such as graphs.

First of all, reasonable time has to be allowed in order to get familiar with machine learning techniques and practice the algorithms described above. That would take approximately some days before the level of understanding is sufficient in order be able to produce solutions using the particular topic. Subsequently a review of the data will be conducted in order to get myself familiar with it. Then different classifiers will be developed and train them using the dataset. That will involve applying different machine learning techniques in order to construct the best applicable models.

The work plan will begin at the first week of the semester and finish at the submission date of the final report, a total of 14 weeks and 3 days (including Easter recess). During that period, the aims as well as the final report will have to be completed. In addition, there is a plan for weekly group meetings with my supervisor and three other students. During those meetings, the supervisor will track our progress in the project and the opportunity will be given to ask any project-relevant questions. Additionally, they are going to be individual meetings with the supervisor once every 3 weeks. To conclude, all the code will be developed using version control system (GitHub) in order to allow the comparison of same files from different versions, as well as enabling myself to go back to develop a previous version of the code if needed.

Milestones:

Week	Plan for the week
1 (27 Jan – 2 Feb)	• Frist meeting with supervisor to discuss the initial plan and ask any questions • Write the initial plan • Hand in a draft of the initial plan in order to get feedback before final submission of the initial plan • Correct any suggestions made from the feedback for the initial plan
2 (3 Feb -9 Feb)	• Submit the initial plan • Research on the web for classical Machine Learning and the different classification algorithms and identify the once that can be applied to the project. (conclude to 5 different classifiers to use in the project) • Weekly meeting
3 (10 Feb – 16 Feb)	• Program and train the first 2 models (using 2 of the 5 selected classifiers) and evaluate their effectiveness using different evaluation methods (Cross-Validation and ROC curve) • Weekly Meeting

4 (17 Feb – 23 Feb)	• Program and train the next 3 models (using the remaining 3 of the 5 selected classifiers) and evaluate their effectiveness using different evaluation methods (Cross-Validation and ROC curve) • Weekly Meeting
5 (24 Feb – 1 Mar)	• Finalize the work made in weeks 3 and 4. Complete any task that may have not finished during the previous 2 weeks. • Weekly meeting
6 (2 Mar – 8 Mar)	• Evaluate and compare the 5 classifiers that have been used in the project. Identify which one works better for the given dataset. • Conduct research on feature selection methods and identify the once that can be used in the project. Conclude to 3 methods that can be applied in the study. • Weekly meeting
7 (9 Mar -15 Mar)	• Apply (if applicable) the first feature selection method on all of the 5 models that have been developed and observe how their performance is affected. • Weekly meeting
8 (16 Mar - 22 Mar)	• Apply (if applicable) the second and third feature selection methods on all of the 5 models that have been developed and observe how their performance is affected. • Weekly meeting
9 (23 Mar - 29 Mar)	• Evaluate which one of the 3 selection methods works better overall on all of the 5 classifiers. Emphasize on the classifier that has the best performance and find the selection method that boost its performance to the maximum. • Weekly meeting
10 (30 Mar – 5 Apr) Easter Recess	• Evaluate code and debug to ensure correctness • Finalize work that has been done in weeks 6-9. Complete any task that may have not finished during those weeks.
11 (6 Apr – 12 Apr) Easter Recess	• Produce graphs that show how the performance of different classifier algorithms vary • Produce graphs that show how the performance of different feature selection methods vary • Clean up the code • Make comments on the code so it can be more readable • Gather all the references • Develop a skeleton on how to construct the final report.
12 (13 Apr – 19 Apr) Easter Recess	• Find all the figures that will be included in the report • Start writing the final report
13 (20 Apr – 26 Apr)	• Write the final report • Weekly meeting
14 (27 Apr – 7 May)	• Final meeting with supervisor • Finalize and submit the report