Machine Learning for Autism Spectrum Disorder Diagnosis

{ Giulia Lorini, Dario Ferrero and Gaganjot Shan }

Introduction

Autism Spectrum Disorder (ASD) is a range of neurodevelopmental disorders characterized by impaired social skills, repetitive behaviors, sensory issues, and language delay. More than 1% of the population falls into this spectrum, with a high imbalance between the sexes, males being 4 to 5 times more likely to be affected than females.

Currently, ASD diagnosis involves long processes and multiple specialists evaluations, using behavioural assessment instruments. Application of Machine Learning methods could significantly speed up the diagnostic process.

Objectives

Following exploration of ASD Diagnosis current state, we decided to focus on female patients, since we agreed that the current disproportion in the existing male-to-female brain image data necessarily leads to biased predictions against the minority side: focusing on this open problem, although not expecting to make any breakthrough, we hoped to:

1. Train and test ML classifiers in a balanced male-to-female context
2. Acquire new insights
3. Improve our initial performances

Methods

The following Machine Learning methods were applied in our project:

- Logistic Regression
- Lasso Regularizer (L1)
- Support Vector Machine
- (k-fold) Cross Validation
- Hyperparameter Grid Search

Estimation of Logistic Regression parameters:

$$\hat{w}_0, \hat{w} = arg_w min - \sum_{i=1}^{N} y_i log\sigma(w_0 + w^T x_i) + (1 - y_i) log(1 - \sigma(w_0 + w^T x_i)) \quad (1)$$

L1 Regularizer Function:

$$R(w) = \sum_{i=1}^{D} |w_i| \quad (2)$$

Initial Results

We carried out our work over multi-site data obtained through the **ABIDE-II project**, focusing on resting-state fMRI scans of a 403 subjects filtered subset (balanced M/F and ASD/Control ratio, IQ and Age outliers removed).

Initially, we applied a pre-processing step to the images by using the **Configurable Pipeline for the Analysis of Connectomes (CPAC)**, obtaining for each subject higher quality brain scans, and analytical data such as the connectivity matrixes between Regions of Interest (ROIs) associated with several atlases.

An initial 5-fold cross-validated training over the Craddock matrixes resulted in overfitted models:

Final Results

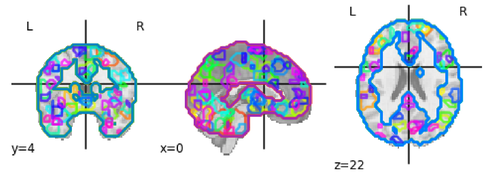

Figure 1: ROIs highlighted on Craddock atlas

1. Generated correlation matrices from preprocessed fMRIs
2. Trained different LR and SVM models using Nested and Repeated 5-fold Cross Validation
3. Used Grid Search to find optimal regularization strength parameter (C)
4. Tested separately on female-only and male-only samples
5. Generated connectivity matrices
6. Repeated from 3. with connectivity matrices

We were able to reach an average accuracy of around 0.57.

Model	Train Acc.	Test Acc.
Log. Regression	0.80	0.57
SVM	0.91	0.57

Table 1: Final model performances obtained

Conclusions

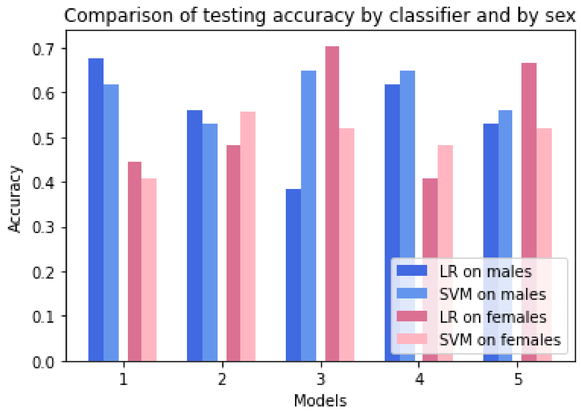

Figure 2: Final comparisons results

- We tried both with the correlation matrices and the connectivity matrices as features, but we couldn't see significant differences between the two.
- Having a small quantity of data for our testing set influenced our performance: we could see that from the testing results over the respective sexes varying significantly.

References

[1] Pegah Kassraian-Fard et alii. "promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example". *Frontiers in Psychiatry*, 7, 2016.

Future Work

As ASD diagnosis via Machine Learning remains an open problem, much is still to be done from a research point of view. Several are the classifiers we could try, but improvements in the future will highly depend on the amount of data that will be able to be collected from the research centers.

Code

Repo https://github.com/GiuliaLo/MALIS-autism-diagnosis