MALIS Project Progress Report

.

Machine Learning for Autism Spectrum Disorder Diagnosis in Women

Giulia Lorini
Gaganjot Shan
Dario Ferrero

December 16, 2021

1 Motivation

Following exploration of the Autism Spectrum Diagnosis research's current state, we decided to focus on ASD Diagnosis for 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's side: focusing on this open problem, although not expecting to make any breakthrough, we hope to acquire new insights and aim at thinking of ways to improve our initial performances.

After a careful consideration, we finally decided to work on the Abide II dataset, opting for a higher quality dataset instead of the former one (albeit pre-processed). In order to make the dataset equally representative of both sexes, our decision was to remove enough of the male samples to reach a number equal to the female ones, while maintaining an agreeable balance of diagnosed and control subjects' data. Furthermore, following the advice found in[7], we removed samples presenting attribute values to be considered as outliers, such as more than 40 years of age and IQ lower than 80.

2 Method

Our top-level objectives regarding the learning model and techniques to be implemented haven't changed: once obtained the pre-processed fMRI data, we plan on comparing training and testing performance of Linear Regression and Support Vector Machines by applying k-fold Cross Validation. To that end, we have already identified the scikit-learn libraries and wrote template code requiring only minor configurations. Additionally, given the unbalanced feature to sample size ratio proper of fMRI data, we plan on applying L2 (Ridge Regression) Regularization.

3 Preliminary Experiments

The time required by the images pre-processing phase (identifying the proper tools, setting up the working environment and being able to run them) has unfortunately prevented us from actually executing the training for the model.

Despite (and also thanks to) this, we were able to get a deeper grasp of the problem and to discuss future strategies, so that we feel confident that once we'll finally get the processed data to work, we will be able to proceed quickly.

Choosing to work on medical images has initially provided us with the challenge of handling large amounts of data (close to 100GB), which we were finally able to reduce by applying the filters and selections described in the above section. Looking for an adequate pre-processing pipeline, we initially found and attempted to use the tool *fmriprep*[8], following the advice of[9]. We ran into a

roadblock when we realized it required our data's directory structure to conform with the "Brain Imaging Data Structure" standard (BIDS)[10], and we couldn't find a way to convert it.

Subsequently, we found a solution that also supported inputs in custom formats, namely the "Configurable Pipeline for the Analysis of Connectomes" (C-PAC)[11], and after a fair amount of time to understand and find the right settings for the configuration files, we were able to run this pre-processing pipeline on what is for now a tiny sample of the images (which still required considerable time and resources).

4 Next Steps

- We still need to decide whether to include the repeated observations we have for certain data subjects (from multiple session recordings of rs-fMRI), which we need to determine how much they might influence the balance in the dataset.
- We plan on debugging our CPAC pipeline configuration to be able to run the tool on laptops other than the one which we've utilized so far. Including possibly additional computing power could be obviously beneficial to this task, we are currently evaluating our options to make it possible.
- We will try to see how much the "outliers" influence the performance.
- Question: should we also add as features some of the diagnostic data (age, IQ, handedness, etc.)? This approach is the one we found in the literature[7], but we are worried it could make the model too complicated for us to handle.
- Our main final objective is to determine whether we can obtain a decent testing accuracy for respectively female and male test data on a model built with both female and male samples. We expect to have a significantly lower testing accuracy for males with respect to existing research (which decided to build their model exclusively on male data[7], because of the significantly higher number of data samples of that type). Nonetheless, our model will hopefully give us more insight on female testing samples with respect to existing solutions.

5 Contributions

- **Common:** Data collection and project planning.
- **Giulia Lorini:** Project management, main research analysis, baseline code preparation, data selection and preprocessing.
- **Gaganjot Shan:** Dataset analysis, exploration of alternative pathways.
- **Dario Ferrero:** BIDS format (attempted) conversion, image preprocessing configuration, debugging, report composition.

References

[1] Gene Blatt, "autism", Encyclopedia Britannica, 9 Sep. 2021, https://www.britannica.com/science/autism, Accessed 27 October 2021.

[2] Child Mind Institute, "Autism Brain Imaging Data Exchange", ABIDE, Child Mind Institute, http://fcon_1000.projects.nitrc.org/indi/abide/, Accessed 27 October 2021.

[3] Kaat Alaerts, Stephan P. Swinnen, and Nicole Wenderoth., "Sex Differences in Autism: A Resting-State Fmri Investigation of Functional Brain Connectivity in Males and Females.", Social cognitive and affective neuroscience., U.S. National Library of Medicine., Accessed October 27, 2021., https://pubmed.ncbi.nlm.nih.gov/26989195/.

[4] Cameron Craddock, Pierre Bellec., "ABIDE Preprocessed.", Preprocessed Connectomes Project., Accessed October 27, 2021., http://preprocessed-connectomes-project.org/abide/.

[5] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, Chaogan Yan, Pierre Bellec, "The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives.", In Neuroinformatics 2013, Stockholm, Sweden.

[6] Child Mind Institute., "Autism Brain Imaging Data Exchange II", ABIDE. Child Mind Institute., http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html., Accessed 27 October 2021.

[7] Pegah Kassraian-Fard, Caroline Matthis, Joshua H. Balsters, Marloes H. Maathuis, Nicole Wenderoth, "Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism as an Example", Frontiers in Psychiatry, vol. 7, 2016, ISSN-1664-0640, doi:10.3389/fpsyt.2016.00177

[8] "fMRIPrep: A Robust Preprocessing Pipeline for fMRI Data", https://fmriprep.org/en/stable/

[9] https://andysbrainbook.readthedocs.io/

[10] "Brain Image Data Structure standard", https://bids.neuroimaging.io/

[11] "Configurable Pipeline for the Analysis of Connectomes (C-PAC)", https://fcp-indi.github.io/