Mining resting-state fMRI Signal to Develop Functional Connectivity and Classification

Team 5

Anamul Haque, Reek Majumder

CPSC8650 Data Mining, Spring 2020

Mining resting-state fMRI Signal to Develop Functional Connectivity and Classification

Team 5

Anamul Haque[†]
School of Computing
Clemson University
Clemson, South Carolina, USA
ahaque@g.clemson.edu

Reek Majumder
School of Computing
Clemson University
Clemson, South Carolina, USA
rmajumd@g.clemson.edu

ABSTRACT

Autism Spectrum Disorder (ASD) is a developmental and may develop at the early stage of life. The diagnosis of ASD has become easier with the development of non-invasive techniques such as Functional Magnetic Resonance Imaging (fMRI). fMRI generates 4D images of the brain and mining those images are considered a big data problem. We mined the Autism Brain Imaging Data Exchange (ABIDE) dataset to build functional connectivity in the brain of the study participants. We identified the brain regions of interest of the brain deemed important for ASD, time series extracted each of those regions, estimated the connectivity matrix among pairs of brain regions, and finally classified those regions. We applied different methods in each of these steps and compared them to determine which performed best. Our result showed no significant change in the prediction accuracy between single- and multisite data.

KEYWORDS

ASD, rs-fMRI, ABIDE, BOLD, ROI, Functional Connectivity

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is characterized as an altered growth of the developmental system. The prevalence of ASD is increasing with the development of non-invasive diagnostic technologies. Therefore, within the last 10 years, the prevalence of ASD jumps to 1 out of 59 from 1 out 88. Researchers are trying to develop molecular and physiological biomarkers for ASD which will help them, to separate normal individuals from ASD.

MRI is a brain-scanning technology that collects spectral information of the slices of the brain from different angles. Technically MRI can be categorized into three different types: structural MRI (sMRI), functional MRI (fMRI), diffusion MRI (dMRI). fMRI is based on the method known as the BOLD (Blood Oxygenated Level Dependence) to produce the time-series image of the brain. During BOLD fMRI, also known as the resting-state fMRI(rs-fMRI), individuals with resting-state were stimulated by external stimuli [7]. During those stimulations, specific regions of the brain show different patterns of blood oxygenated level (BOLD) for a short period. Since Oxygen is only supported from the outside of the brain, neurons that are activated only presented the active part of the brain. Therefore, the inactive and inactive parts of the brain can be identified using the fMRI scan. Further analysis was performed to extract those active and inactive regions. These regions are generally considered as the Region of Interest (ROI) for specific stimuli. These raw fMRI images need to heavily preprocess to discard any artifacts and to find valuable information grouping or classifying them.

The scarcity of reliable fMRI datasets to classify ASD urges large studies like Autism Brain Imaging Data Exchange (ABIDE) [6]. The ABIDE dataset contains resting-state fMRI (rs-FMRI) from 1112 participants collected from 16 sites in the USA, Netherlands, Germany, Belgium, and Irelands. Among these 1112 participants, 539 participants are labeled as ASD case and 573 participants are typical controls (ages 7-64 years, median 14.7 years across groups). For each participant, two types of data collected: phenotypic data (age, sex, height, IQ score, etc.) and imaging data (4D rs-fMRI time-series data).

The connectome is a functional connectivity matrix between a set of brain regions of interests (ROIs). Those ROIs are known as the atlas and can be derived either from the dataset or any predefined brain atlas. Functional connectivity is measured when the different part of the brain is activated together and cause these regions to fire[3]. The

[†] **Team Leader**

@ CPSC8650 Data Mining, Spring 2020, Clemson University

connectivity matrix contains connectivity information of all the regions of the interests present in the Atlas and can be used to classify different ROIs of the brain.

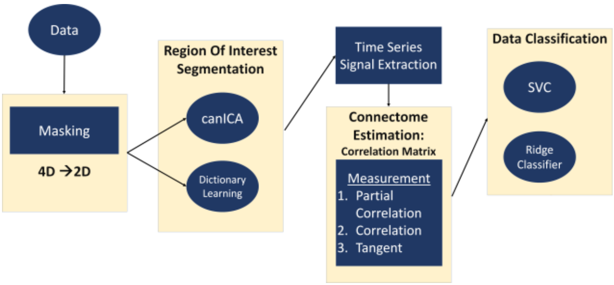

Figure 1: Outline of the Project

In the following data project, we mined the abide fMRI and phenotype dataset to find valuable information toward classifying ASD and non-ASD individuals. We followed presented a four data mining approach. In the first step, we estimated the ROIs. Then we estimated time-series information for those ROIs. In the third step, we calculated the connectivity matrix from the time series data. Finally, we extracted features from the third step to classify the data. For some of these steps, we have used multiple techniques as described in the methods section.

2 METHODOLOGY

2.1 Dataset & Tools

For this project, we have used the ABIDE I dataset from the International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/indi/abide/). We worked with the preprocessed version of the dataset preprocessed by the Configurable Pipeline for the Analysis of Connectome (C-PAC). This preprocessed connectome dataset includes 871 participants rather than 1112 participants of the original dataset. The C-PAC preprocessing includes slice time correction, image realignment to correct motion and intensity normalization. The C-PAC pipeline also includes Nuisance Regression which removes signal fluctuation induce by the head motion, respiration cardiac pulsation and scanner drift. Our next step was to use background masking to convert the 4D data to 2D data (features, samples).

For the data analysis we have used the Python package Nilearn (https://nilearn.github.io/) and for the classification part we have used the package scikit-learn (https://scikit-learn.org/stable/) [8]

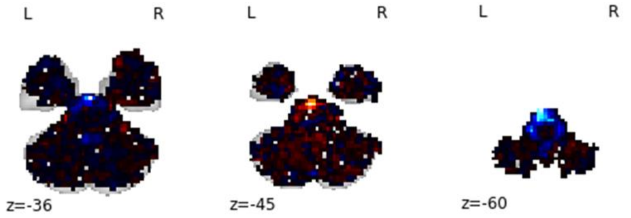

Figure 2: Set of 3D images we receive from each subject. Above are first few 3D image of subject 0050592 Outline of the Project

2.2 Region Definition Using ICA and Dictionary Learning

Independent Component Analysis is an increasingly used data-driven method to analyze functional Magnetic Resonance Imaging (fMRI) data. It is used to extract meaningful patterns without prior information. However, ICA is not robust to mild data variation and remains a parameter-sensitive algorithm [1,2]. We have used canonical ICA to overcome this issue, which can build a generative model on the group data to introduce the probabilistic ICA pattern-extraction algorithm. Although ICA is widely used to maximize statistical independence between spatial maps for extracting brain maps for rs-fMRI. We have also tested our performance with the Multi-subject Dictionary learning algorithm over initialized canonical ICA to produce more stable maps [5].

Finally, we used a predefined function from Nilearn to segment brain atlas maps into different sets of brain activated regions to show that each decomposed brain maps can be used to focus on a target-specific ROI analysis

2.3 Time Series Extraction

We have extracted one time-series from each of the ROIs. Here we have used the ICA and MSDL to create Fuzzy Images. The mask is computed by filtering and extracting data from the in-mask voxels by NiftiMapsMasker function from Nilearn. The advantage of using such tools restricts our analysis to mask specific voxel time data. Masking over the data helps in pre-processing steps like filtering, smoothing, and standardizing on voxels time-series signals. NiftiMapsMasker in nilearn helps us to extract signals from each 4th-dimensional map using least square regression.

2.4 Connectome Estimation using the correlation matrix from timeseries

The most frequently used method to assess the brain connectivity between different regions of the brain is to use the pairwise correlation between the BOLD time course from two different brain regions. Some studies suggest “partial correlation” as a better alternative of “full correlation” to measure direct connectivity between two nodes. It estimates their correlation by regressing out effects from other nodes. For our project we have used three methods – the correlation, partial correlation from the inverse covariance matrix [9], and tangent embedding parametrization of the covariance matrix[10].

2.5 Supervised Learning

ABIDE provides post hoc data aggregation of data from several sites [6]. Autism severity may vary from site to site. We used supervised learning from Nilearn and scikit-learn to classify the ASD. In our experiment, we have compared the supervised learning from a single site - NYU and the whole dataset. For supervised learning, connectivity measures between all pairs of the regions extracted in the previous step used as features to train the classifier in the previous stage. We used the l2-penalized support vector classification SVC at first. We also used l1-penalized sparse SVC due to our expectation of lower connection. Finally, we used stratified sampling to avoid bias while creating models for classification since we have two labels only (ASD-1, Control-2) [10]. It helps to improve the performance of our model at the later stage, by reducing bias during cross-validation.

3 RESULTS

3.1 ICA Decomposition

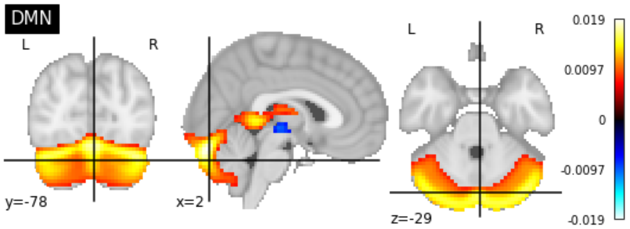

Figure 3: ICA Decomposition of the Brain

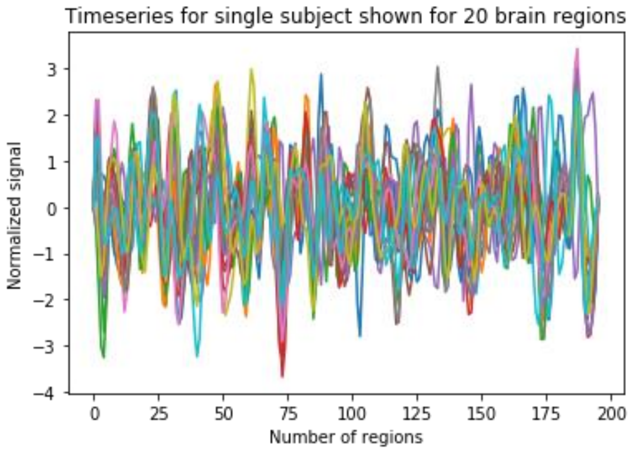

Figure 4: Time series analysis of a single subject

3.2 Region Extraction

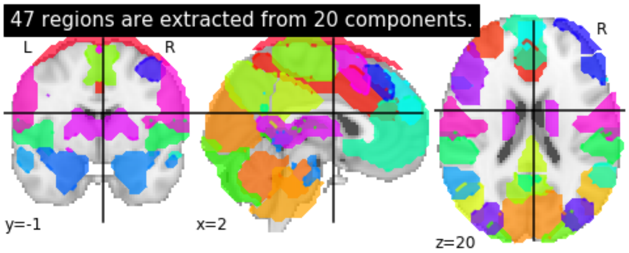

Figure 5: 47 ROIs extracted from 20 components

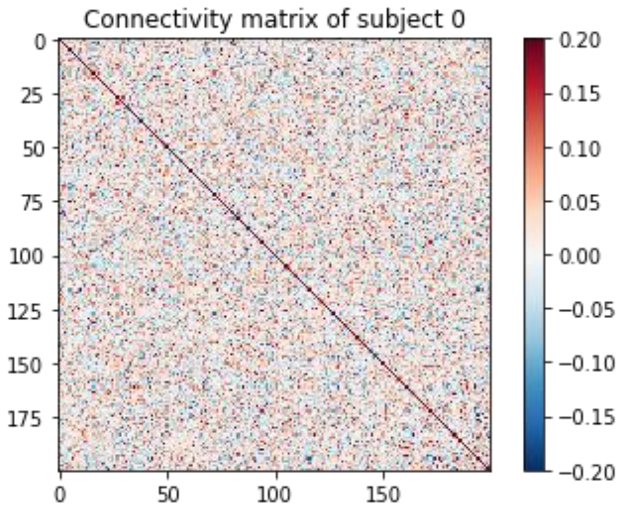

Figure 6: Connectivity Matrix of Subject 0

3.3 Classification

For the classification of the data we compared the whole datasets with the dataset from the New York University (NYU). The accuracy measure did not show any significant difference in single and multiple sites. For all data Ridge classifier performs better than other two classifiers. Also tangent measurement performs better in all of the cases.

Table 1: Classification of Single Site (NYU)

Predictors	Accuracy Measure		
(For Single Site NYU)	Correlation	Partial Correlation	Tangent
SVC L1	0.61	0.58	0.64
SVC L2	0.60	0.60	0.65
Ridge Classifier	0.66	0.68	0.67

Table 2: Classification of all sites

Predictors	Accuracy Measure		
(For all the sites)	Correlation	Partial Correlation	Tangent
SVC L1	0.64	0.58	0.61
SVC L2	0.65	0.62	0.68
Ridge Classifier	0.67	0.60	0.67

4 DISCUSSION

The connectome is During the project, we have faced several problems in implementing our methods and tools. Initially, we planned to start the project from the raw image data. We faced several obstacles with that approach. First of all, working with raw fMRI images requires a lot of disk space because of the large size DICOM(.dcm) image format. We have also faced technical difficulties in downloading and managing those files from the Amazon web service. Therefore, we decided to use the C-PAC pre-processed data which is already pre-processed with steps such as slice-timing correction, image realignment, intensity normalization, and nuisance regression. We tried the Matlab tool Statistical Parametric Modeling (SPM12) at the beginning to work with the .dcm image format then moved to Python package Nilearn and used this throughout the project. The good thing about Nilearn is that it supports the built-in download function for downloading the ABIDE dataset. Moreover, this package also supports working on extensively with processed .nii datasets which are very smaller in size than the raw .dcm format. Still, the dataset is very large and requires

In our project, we have used the Configurable Pipeline for Analysis of Connectome (C-PAC) pre-processed dataset. Since the data has been collected from different sites (16 sites), a preprocessed dataset helps to maintain the quality of the data among sites. Initially, we were working to identify the important biomarker for ASD, we could not reach that goal because it requires much more expertise than we expected at the beginning of the project. In addition to these, we wanted to use some unsupervised and hybrid models but could not be succeeded to do them.

ACKNOWLEDGMENTS

We are grateful to Professor Wang for allowing us to learn new techniques in data mining. Working in this project helped us to learn data mining in the biomedical domain especially in the field of biomedical imaging analysis and this helped us to improve our skills.

REFERENCES

1. Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T. Liu. "A component based noise correction method (CompCor) for BOLD and perfusion based fMRI." *Neuroimage* 37, no. 1 (2007): 90-101.
2. Calhoun, Vince D., Jing Sui, Kent Kiehl, Jessica A. Turner, Elena A. Allen, and Godfrey Pearlson. "Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder." *Frontiers in psychiatry* 2 (2012): 75.
3. Craddock, R. Cameron, Paul E. Holtzheimer III, Xiaoping P. Hu, and Helen S. Mayberg. "Disease state prediction from resting state functional connectivity." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 62, no. 6 (2009): 1619-1628.
4. Craddock, R. Cameron, G. Andrew James, Paul E. Holtzheimer III, Xiaoping P. Hu, and Helen S. Mayberg. "A whole brain fMRI atlas generated via spatially constrained spectral clustering." *Human brain mapping* 33, no. 8 (2012): 1914-1928.
5. Craddock, Cameron, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S. Ghosh, Chaogan Yan, Qingyang Li et al. "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac)." *Front Neuroinform* 42 (2013).
6. Di Martino, Adriana, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X. Castellanos, Kaat Alaerts, Jeffrey S. Anderson et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic

brain architecture in autism." *Molecular psychiatry* 19, no. 6 (2014): 659-667.
7. Kelly, AM Clare, Lucina Q. Uddin, Bharat B. Biswal, F. Xavier Castellanos, and Michael P. Milham. "Competition between functional brain networks mediates behavioral variability." *Neuroimage* 39, no. 1 (2008): 527-537.
8. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12, no. Oct (2011): 2825-2830.
9. Varoquaux, Gaël, and R. Cameron Craddock. "Learning and comparing functional connectomes across subjects." *NeuroImage* 80 (2013): 405-415.
10. Varoquaux, Gaël, Sepideh Sadaghiani, Philippe Pinel, Andreas Kleinschmidt, Jean-Baptiste Poline, and Bertrand Thirion. "A group model for stable multi-subject ICA on fMRI datasets." *Neuroimage* 51, no. 1 (2010): 288-299.
11. Abraham, Alexandre, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example." NeuroImage 147 (2017): 736-745.