Identification of Autism from Functional Connectivity Analysis using Support Vector Machine and K Nearest Neighbour Classifier

Utkarsh Patel
IIT Kharagpur
utkarshpatel@iitkgp.ac.in

Abstract – **The goal of this report is to apply machine learning algorithms to classify autism spectrum disorder (ASD) patients and typically developing (TD) participants using fMRI data from ABIDE dataset. SVM and KNN were used for classification purpose. Multi-layer perceptron classifier was also used for comparison. I used a cross-validation grid search to fine-tune the hyperparameters for each classifier. Finally, a stacked ensembled model was used with the tuned hyperparameters of the classifiers.**

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a brain disorder that is characterized by social and communication impairments as well as restricted interests and repetitive behaviours. One in every 68 children in USA is affected by ASD. Early diagnosis of ASD is critical for the implementation of the early intervention and providing a proper treatment plan. Although ASD has been identified since the early 1960s, its exact cause is still unknown. Generally, the symptom-based diagnosis of ASD requires a very significant amount of time behavioural assessment under the guidance of a highly experienced multi-disciplinary team. However, symptoms-based diagnosis often results in poor treatment due to lack of knowledge of neuropathology. In the past years, an increasing number of neuroscience research studies have used machine learning and deep learning to implement data-driven diagnosis of ASD, which would lead to more effective treatment outcomes. One promising candidate for the data-driven diagnosis is resting state functional connectivity MRI. In past research, extensive brain imaging studies have reported that ASD is associated with brain connectivity. Despite extensive research evidence that ASD is a brain connectivity disorder, it lacks a distributed framework of brain abnormalities. It is still unclear whether brain abnormalities are associated with specific brain regions in ASD. In this study, we implemented a data-driven approach to classify ASD patients and typically developing (TD) participants by using the rs-fcMRI features extracted from resting-state functional MRI (rs-fMRI) data.

2 DATASETS

The pre-processed fMRI data with ASD and ID are downloaded from a large multisite data repository ABIDE (Autism Brain Imaging Data Exchange). ABIDE is a multisite platform that has aggregated functional and structural brain imaging data collected from 17 different labs around the world. The pre-processed connectomes project (PCP) from ABIDE has openly released 539 individuals who have ASD and 573 TD to public. These 1112 datasets consist of structural and pre-processed resting state fMRI data along with phenotypic description. The rs-fMRI data are slice time corrected, motion corrected and normalized. For this task, all rs-fMRI data are selected from the CPAC pre-processing pipeline and band-pass filtered (0.01-0.1Hz).

TABLE 1
ABIDE DATA PHENOTYPICAL INFORMATION

Site	ASD	TD	M	F	Age
Caltech	19	18	29	8	17 – 56
CMU	14	13	21	6	19 – 40
KKI	20	28	36	12	8 – 13
Leuven	29	34	55	8	12 – 32
Maxmun	24	28	48	4	7 – 58
NYU	75	100	139	36	6 – 39
OHSU	12	14	26	0	8 – 15
OLIN	19	15	29	5	10 – 24
PITT	29	27	48	8	9 – 35
SBL	15	15	30	0	20 – 64
SDSU	14	22	29	7	9 – 17
Stanford	19	20	31	8	8 – 13
Trinity	22	25	47	0	12 – 26
UCLA	54	44	86	12	8 – 18
UM	66	74	113	27	8 – 29
USM	46	25	71	0	9 – 50
Yale	28	28	40	16	7 - 18

In spite we have 1112 datasets, one cannot use them all to train any model. Recent literatures point out that the configuration of instruments used in different labs are subjected to be different, moreover, the time-series length is different for different sources. So as to remove this difficulty, I have used the datasets from NYU only for this task owing to the fact that it has the largest number of fMRI samples. For this task, I have used Automated Anatomical Labelling (AAL) brain parcellation which contains 116 regions of interest (ROIs).

3 METHODS USED

As we are using AAL brain parcellation, we have 116 ROIs. For each pair of ROIs, I computed the functional connectivity between them via correlation of their time-series signal extracted from fMRI data. This is quite intuitive that a high correlation factor guarantees a strong functional connectivity, and low correlation factor guarantees a weak functional connectivity. Thus, after this step, we have a correlation matrix of order 116.

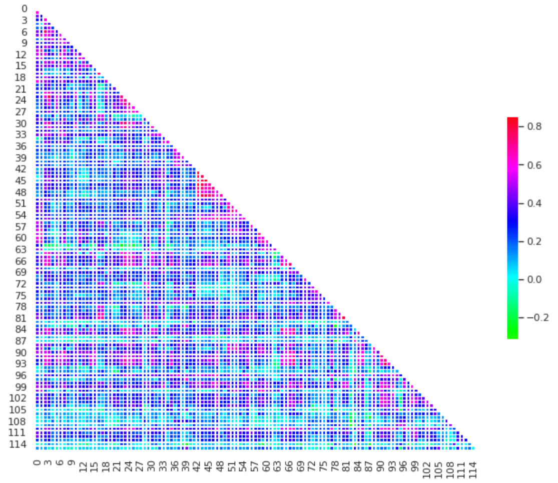

Fig. 1. Correlation matrix of a random subject

This matrix is then flattened to a one-dimensional vector which will be used as feature vector for classification purpose. The class labels are extracted from the phenotypic description. Having obtained the feature vectors and class labels, state-of-the-art techniques like SVM, KNN, etc. can be used for classification. For each classifier, its hyperparameters are tuned via a grid search using 5-fold cross-validation.

4 SUPPORT VECTOR MACHINES

For the purpose of classification, I used SVM. Now for SVM, various kernel functions can be used and it also has tunable hyperparameters C and γ. For this task, I used linear, quadratic, radial and sigmoid basis functions as kernels and varied C from 2^{-5} to 2^{15}, and γ from 2^{-15} to 2^{3}.

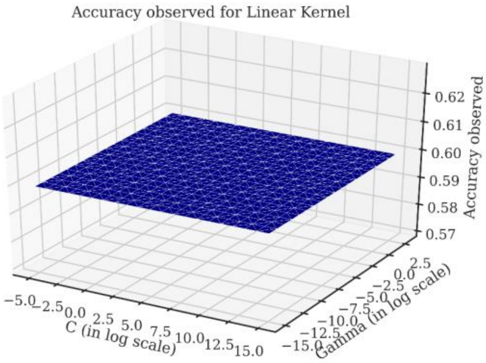

Fig. 2. Accuracy observed for linear kernel

It can be observed from Fig. 2. that while using linear basis function as kernel, the accuracy observed is same for all the values of hyperparameters. The observed accuracy was 59.92%.

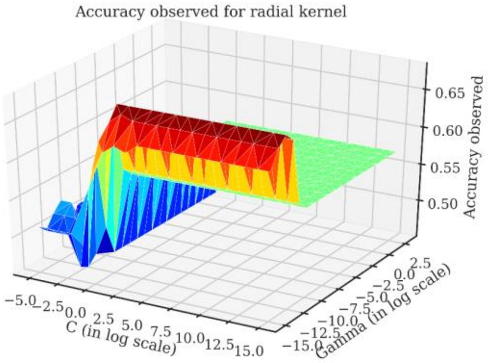

Fig. 3. Accuracy observed for radial kernel

From Fig. 3. we can observe that accuracy of the model using a radial basis function is quite high for lower values of γ and higher values of C. The best accuracy was observed to be 67.97% for $\gamma = 2^{-13}$ and $C \geq 2^{1}$.

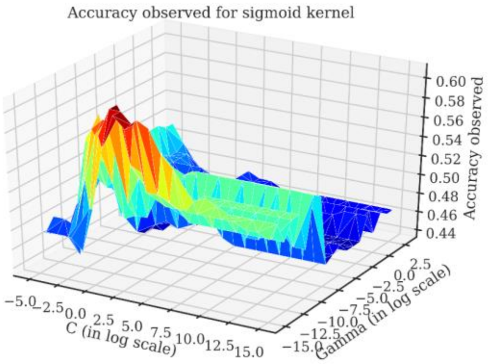

Fig. 4. Accuracy observed for sigmoid kernel

It can be concluded from Fig. 4. that model with sigmoid kernel function gives best accuracy of 61.04% for $\gamma = 2^{-13}$ and $C = 2^{1}$.

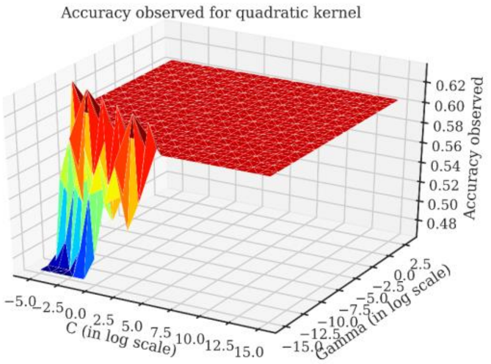

Fig. 5. Accuracy observed for quadratic kernel

For quadratic kernel, we observe, from Fig. 5. that the best accuracy is 63.41% for values of C and γ satisfying,

$$log_2(\gamma) + \frac{1}{2} log_2(C) + \frac{25}{2} = 0$$

TABLE 2
SUMMARY FOR SVM

Kernel	***Best Accuracy***	***C***	***γ***
Linear	0.5992	ℝ	ℝ
Radial	0.6797	$\geq 2^1$	2^{-13}
Sigmoid	0.6104	2^1	2^{-13}
Quadratic	0.6341	2^3	2^{-14}

From Table 2, it can be concluded that the best kernel to be used in SVM for this task is radial with corresponding hyperparameters.

5 K NEAREST NEIGHBOURS

In KNN, we need to tune the value of parameter k and distance function defined as *Minkowski* distance with parameter p. $p = 1$ corresponds to *Manhattan* distance and $p = 2$ corresponds to *Euclidean* distance. I varied the parameter k from 1 to 50 for both values of p.

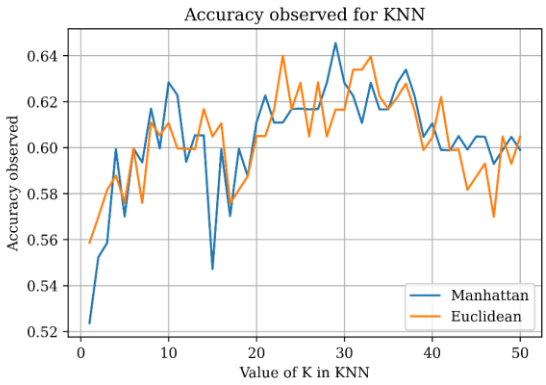

Fig. 6. Accuracy observed for KNN

TABLE 3
SUMMARY FOR KNN

p	***Best Accuracy***	***k***
1	0.6455	29
2	0.6399	23

By comparing Table 2 and Table 3, we can observe that a finely-tuned SVM model outperformed a finely-tuned KNN model. Another popular approach in machine learning is to use stacked ensemble models for classification. For this, we are using logistic regression, finely-tuned SVM and finely-tuned KNN as *base* models and another logistic regression model as *meta* model.

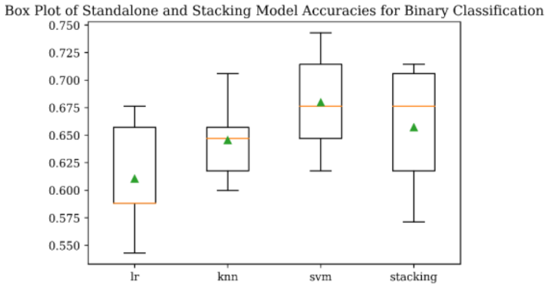

Fig. 7. Box Plot of Standalone and Stacking Model Accuracies

From Fig. 7. it can be concluded that though stacked model performs better than standalone finely-tuned KNN model, yet the standalone finely tuned SVM model outperforms the stacked model.

6 COMPARISON WITH NEURAL NETS

From the previous section, it is observed that a standalone finely-tuned SVM classifier performs the best for the given task. In this section, various neural net architectures are used to perform the same task. The results obtained is shown in Fig. 8.

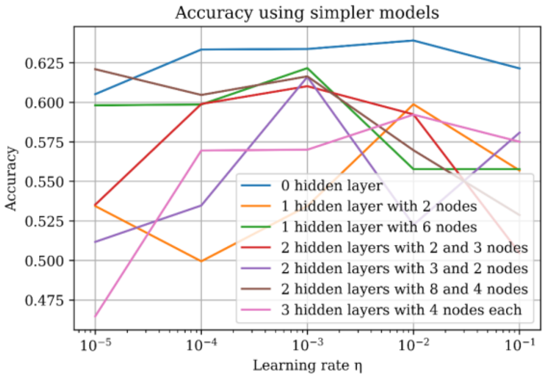

Fig. 8. Accuracy observed for different neural net architecture

7 CONCLUSIONS

From Table 2, Fig. 7. and Fig. 8. it can be concluded that a standalone finely-tuned SVM model is best fitted for this task and can provide accuracy as high as 68%.