Kill the Outlier

Ahmet Agah Karagöz
Istanbul Technical University
Istanbul/Turkey

Anıl Zeybek
Istanbul Technical University
Istanbul/Turkey

Mehmet Ataberk Atan
Istanbul Technical University
Istanbul/Turkey

Tuğba Durman
Istanbul Technical University
Istanbul/Turkey

***Abstract*—Brain is very complex and complicated network. To estimate how it is going to act in the future is hard. Machine learning methods have to used. So, we use support vector regression and outlier reducing methods to predict how brain connectivities between different regions will change. With this information we can detect illnesses long before they showed up.**

I. Introduction

Our main purpose in this project is to predict the brain connectivity at a given time t1 from brain connectivity measured 6 months ago at timepoint t0. By measuring the brain connectivity, we can track the changes in the brain. As of 13.06.2020, our Kaggle score is 0.00213 and we are in the 3th place.

II. Dataset and Preprocessing

The only preprocessing we are doing is removing the outliers. We also tried scaling the features but that did not help us, instead mean squared error only increased. We tried to investigate it, but we could not find why.

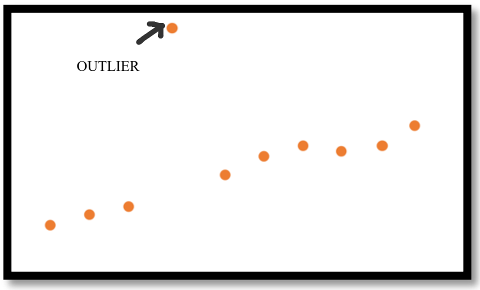

Figure 1: Outlier in a dataset without any treatment

To remove outliers (Fig.1), we used 2 different methods: Neighbors method and mean method. In neighbours method, first we find all the sample features that has a z value greater than 3 and we eliminate them by taking the mean of this sample feature's neighbours (Fig.2).

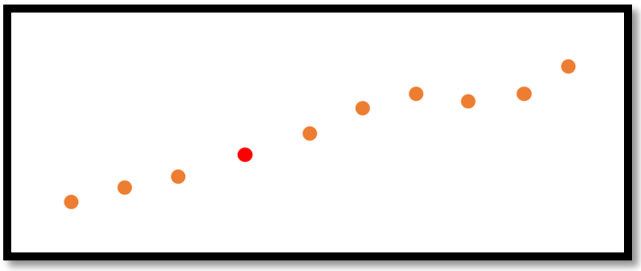

Figure 2: After using neighbour outlier method

For example, if our value is 143.0 and if it has a z value greater than 3 compared to other values in the same feature, and if the neighbors are 5 and 3, our value will become 4. The mean method does a very similar job that instead of taking the means of our value's neighbors, this method takes the mean of the whole feature (Fig.3).

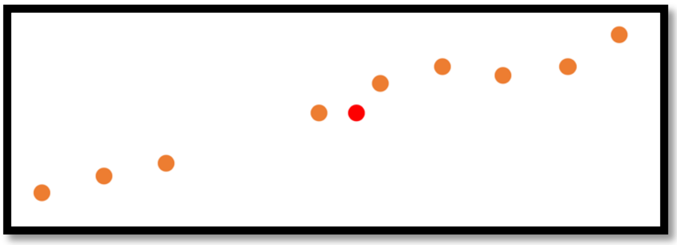

Figure 3: After using mean outlier method

III. Method

For learning part of our project, we are using support vector regression (SVR). After preprocessing the data, we are using 5-Fold for training and testing stages. For each fold of our data, we train it and we calculate the mean squared error of the prediction with our test data. For training part, we use Gaussian RBF kernel and we set the regularization parameter as 1000 for SVR module in the scikit learn to increase the regularization for our learning. Also, we set the gamma value as 0.1 and we set the epsilon to 0.01 just for increasing the number of training loops to acquire better results.

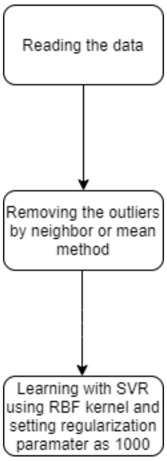
Reading the data
Removing the outliers by neighbor or mean method
Learning with SVR using RBF kernel and setting regularization paramater as 1000

Figure 4: System pipeline diagram

Instead of SVR, we also tried to use the Linear Regression and Random Forest Regression. Firstly, we used the simple linear regression and the results were not too bad, but we were one of the last in the competition. After that, we tried to use the Random Forest Regression and the results was better compared to linear regression, but we did not want to stop and lastly tried the support vector regression. The results were much better compared to linear regression and random forest regression. So, we selected to use this algorithm. After we deal a lot of with the parameters of SVR, and outlier reduction methods(threshold) we found the parameters I gave was the best.

Result and conclusions

Our 5-Fold cross-validation results are: 0.00240, 0.00191, 0.00230, 0.00214 and 0.00225 with a mean of 0.00220. All the results are mean squared error results, we did not use any other evaluation metrics. Our values are not so bad, and they are generally close to each other. As of 13.06.2020, our Kaggle score is 0.00213 and we are in the 3th place.

References

[1] Simon Tong and Daphne Koller, " Support vector machine active learning with applications to text classification",2001

[2] Stefan Wagnet, "Cross-Validation,Risk Estimation and Model Selection",2019

[3] Basak, Debasish & Pal, Srimanta & Patranabis, Dipak. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews. 11.

[4] Bettinger, Ross. (2020). Outlier Detection and Treatment.

[5] Quirk, Thomas. (2020). Correlation and Simple Linear Regression. 10.1007/978-3-030-39261-1_6.