CLASSIFYING AUTISM SPECTRUM DISORDER USING MACHINE LEARNING MODELS

Tingyan Deng

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

ABSTRACT

Autistic Spectrum Disorder (ASD) is a very common and serious developmental disability, which impairs the ability to communicate and interact, causing significant social, communication, and behavior challenges. From a rare childhood disorder, ASD has evolved into a disorder that is found, according to the National Institute of Health, in 1% to 2% of the population in high income countries. A potential early and accurate diagnosis can not only help doctors to find the disease early, leading to a more on time treatment to the patient, but also can save significant healthcare costs for the patients. With the rapid growth of ASD cases, many open-source ASD related datasets were created for scientists and doctors to investigate this disease. Autistic Spectrum Disorder Screening Data for Adult is a well-known dataset, which contains 20 features to be utilized for further analysis on the potential cause and prediction of ASD. In this paper, we developed an Autism classification algorithm based on logistic regression model. Our model starts with featuring engineering to extract deep information from the dataset and then applied a modified logistic regression classifier to the data. The model predicts the ASD well in an average F1 score of 0.92.

KEYWORDS

ASD, Classification, Machine Learning, Neurodiversity.

1. INTRODUCTION

Autistic Spectrum Disorder (ASD) is a mental disorder that can affect cognitive, communication and social abilities. Autism is the fastest growing development disability in the world. ASD has been reported to affect as many as 1 in 88 children in the US. Epidemiologic surveys of adult populations suggest that the apparent rise in number of affected children may not represent a true increase in prevalence rates. Nevertheless, there is speculation that broadened definitions, growing awareness, and diagnostic substitution may be contributing to the apparent rise. Regardless of the cause, the current prevalence estimates suggest that there will be more than 2 million individuals in the US with ASD. Up to now, no preventive strategies have demonstrated consistent benefits and no treatments have proven widely efficacious in treating the core symptoms of ASD. Consequently, ASD causes lifelong disabilities for affected individuals and significant burdens on their families, schools, and society.

Raising awareness helps people understand and not be frightened by the disability. Many related researches were investigated to recognize it, prevent it or treat it. In work [1], the authors presented a method using the screening trying to diagnosis this disorder. In work [2], LT Curtis and his partners gave several approaches including nutritional and environmental approaches to prevent the Autism. In work [3], music therapy was used to enable communication and expression and thus can solve some problems of this disorder. In an effort to make easier and

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 59-65, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110306

earlier detection possible, we are using a modified logistic regression model to construct an ASD classification algorithm and the work is discussed in this paper.

1.1. Statement of the Problem

As discussed in the introduction above, the problem is the increasingly ubiquitous occurrence of the ASD symptom, yet little recognitions and efforts were put into solving this issue. With more and more cases of teenager ASDs, we have to come up with a way to ease the pain of ASD. We are aiming to push the barrier of ASD detection knowledge in this study and help future scientists by providing our results.

1.2. Aims/Goals of the Research

In this study, our goal is to provide a faster, and easier machine-learning based approach that can be implemented in future ASD detections and find the related attributes that are causing the ASD. Machine learning has immense potential to enhance diagnostic and intervention research in the behavioral sciences and may be especially useful in investigation involving the highly prevalent and heterogenous syndrome of ASD. In recent years, with more and more advanced computational and engineering methodologies being employed to meet the needs of cross-subject applications, machine learning showed promise in detecting many medical symptoms, which greatly increased the chance of being cured for millions of patients.

Applying the state-of-the-art model is crucial because more and more teenagers have the symptom and there will be a huge amount of middle-age autism community in future and we must detect the symptom earlier so that doctors can cure them earlier.

In our study, we are using a method based on logistic regression model for ASD classification. The specific design of our experiment is listed in the section 2 and section 3.

1.3. Workplan

The work plan, including the experiment/approach, the data we used, and findings are discussed in section 2 and 3.

1.4. Our Contribution and the Flow of the Paper

This paper applies logistic regression to ADS diagnosis. More specifically, the data columns and meanings are described in the first step. Then the data imputation method and feature engineering methods are introduced to further proceed the original data. Data visualization technique was also utilized and the paper contains many easy-to-understand figures. The experiments show the metrics like accuracy, recall, and F1 score. We got an average F1-score of 0.92, which proves the feasibility of our model.

The remainder of this paper is organized as follows. Section II introduces data structure and feature engineering methods. Section III gives a brief introduction to logistic regression models, and the experimental results and analysis. Finally, Section IV gives the summary of whole paper and discuss the ethical concerns of the project.

1.5. Ethics of the Project

Ethics are broadly the set of rules that govern our expectations and of our own and other's behavior. As discussed in [5], research ethnics are important for a numerous of reasons. Firstly, it supports the values required for collaborative work, such as mutual respect and fairness. Secondly, it means that researchers can be held accountable for their actions. Thirdly, good research ethnics support important social and moral values, including the quality of not harming others. If our team members have an internal conflict due to the fact our members have a different opinion towards neurodiversity, I will suggest them to reconcile by taking a look at NISE (Frist Center for Autism and Innovation) website where there are a plethora amount of information about the idea and inspiration behind neurodiversity and also the inspired engineering in this field.

In this project, the dataset we are using is directly downloaded from Kaggle and it's an anonymous, open source dataset, which guarantees the privacy of the research participants. The whole purpose of this project is to accelerate the studies of autism and make the ASD community more inclusive and welcoming. If any ethical dilemmas occur in our project, we will first talk to the people who believes our project violates any computer ethics or ethics in general and then solve the issue by consulting one of the professionals in the NISE community or other experts in the field.

2. DATA AND FEATURE ENGINEERING

The data set can provide a lot of information. In order to explain the dataset intuitively, we list the features of dataset in the table 1.

Table 1. Data Structure

A1_Score to A10_Score
age
gender
ethnicity
jundice
austim
contry_of_res
Used_app_before
result
relation
Class/ASD

As shown in the table 1, the features age, gender, ethnicity are attributes of the ASD testers and easy to figure out the meaning of them. The A1_Score to A10_Score are the answer code of the question based on the screening method used. In particular, they are binary values, either equal to 0 or 1. Statistically, the data has 704 entries and the memory size is 115.6KB.

From the figure 1, we can find the columns ethnicity, relation and age columns have missing values. Since the ratio of the missing values are small, we can drop these missing values or impute the missing values. In this paper, I impute the missing value of age column with the averaged age. And drop missing values of other columns like relation. These steps are necessary for later visualization or building machine learning model.

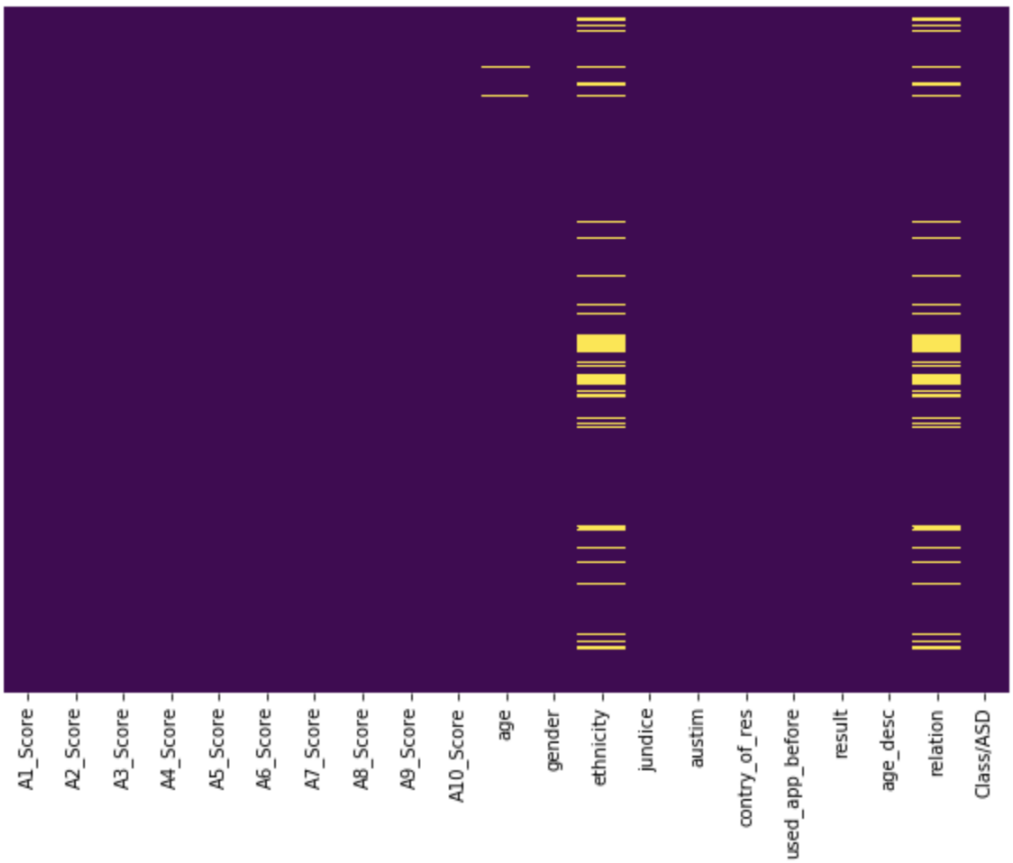

Figure 1. ASD Missing Values Distribution

Besides, we also did some visualization of different features. In the Figure 2, it shows the age distribution. We can find that most of recognized cases are between 0 and 40 years old. This result shows that the ASD mainly occur in young generation other than old people. In figure 3, we can see the ADS Case distribution. The positive sample are higher than negative sample, which means the data is unbalanced data.

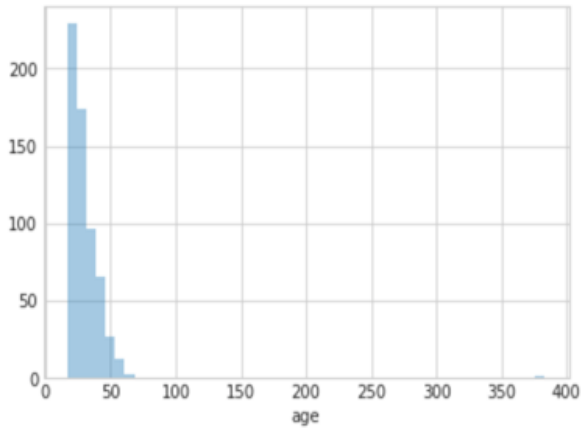

Figure 2. Age Distribution

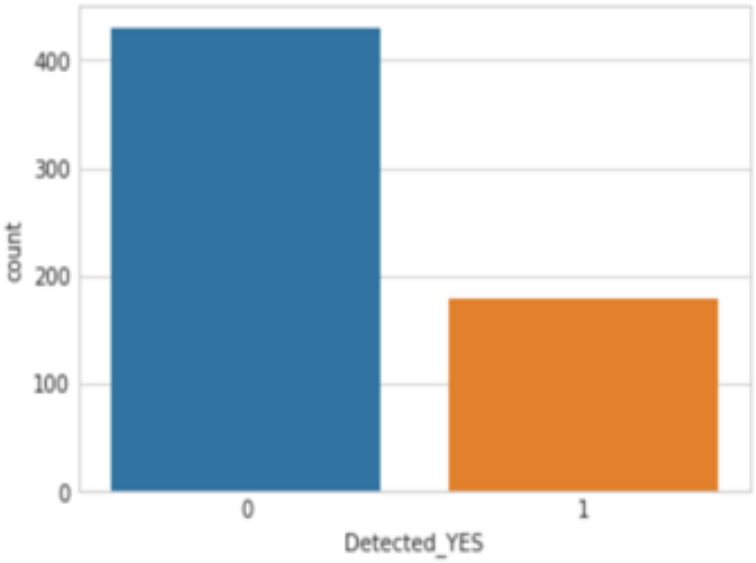

Figure 3. ADS Case Distribution

3. MODELS AND THE EXPERIMENTS

Logistic regression is a linear model which uses a logistic function for classification. The logistic function is as followed:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Equation1. logistic function

The *f(x)* represents the output of the function, L represents the curve's maximum value, k stands for logistic growth rate or steepness of the curve, x0 stands for x value of the sigmoid midpoint, x stands for the real number. Besides, we use the L2 Norm in this model to prevent overfitting.

In the experiment step, we split the data into training set and test set in a ratio of 7:3. Training set is used for fitting the model and test set is used to validate the performance of the model. In the table 2, we list the metrics and performance of our model.

Type	Precision	Recall	F1-score	Count
0	0.98	0.97	0.98	132
1	0.92	0.96	0.94	51
avg	0.97	0.97	0.97	183

Table 2. Performance of our model

As shown in the table 2, the average score of precision is 0.91 and F1-score is around 0.92, which means our model is very accuracy for Autism classification.

4. CONCLUSIONS

In this paper, we propose a method based on logistic regression model for Autism classification. We used the screening data together with meta data like age, gender to fit the model. In Section II, we introduce the data structure and the data size. In addition, the feature engineering is also covered in this part. In Section III, the logistic regression and its function is explained. Then we mentioned the details of the experiments step and the metrics of our model. The experiments show the power of our model since it has a high accuracy. In the future, we are planning on using more machine models like SVM, LightGBM and do a compare and contrast on different models on classification results as discussed in [8].

ACKNOWLEDGEMENTS

We thank the Kaggle platform for provide an ASD screening dataset for researchers to investigate artificial algorithms. Besides, we also appreciate the VUSE (Vanderbilt University School of Engineering) lab for providing the free computation resources like free GPU cards, Rtx 2080 Ti.

REFERENCES

[1] Filipek P A, Accardo P J, Baranek G T, et al. The screening and diagnosis of autistic spectrum disorders[J]. Journal of autism and developmental disorders, 1999, 29(6): 439-484.

[2] Curtis L T, Patel K. Nutritional and environmental approaches to preventing and treating autism and attention deficit hyperactivity disorder (ADHD): a review[J]. The Journal of Alternative and Complementary Medicine, 2008, 14(1): 79-85

[3] Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder[J]. Cochrane Database of Systematic Reviews, 2006 (2).

[4] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[5] David B. Resnik, J.D., Ph.D, What Is Ethics in Research & Why Is It Important? Retrieved December 10, 2020, from https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

[6] M. F. Misman et al., "Classification of Adults with Autism Spectrum Disorder using Deep Neural Network," 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Malaysia, 2019, pp. 29-34, doi: 10.1109/AiDAS47888.2019.8970823.

[7] M. Elbattah, R. Carette, G. Dequen, J. -L. Guérin and F. Cilia, "Learning Clusters in Autism Spectrum Disorder: Image-Based Clustering of Eye-Tracking Scanpaths with Deep Autoencoder," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 1417-1420, doi: 10.1109/EMBC.2019.8856904.

[8] T. Deng, Y. Zhao, S. Wang and H. Yu, "Sales Forecasting Based on LightGBM," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 383-386, doi: 10.1109/ICCECE51280.2021.9342445.

AUTHOR

Tingyan Deng is a junior student at Vanderbilt University studying computer science, mathematics and economics. He is passionate about using technology to make an impact.