Autism Screening Adult Data Set: A Machine Learning Approach

Dr. Kanad Basu

February 8, 2018

1 Domain Background

Autistic Spectrum Disorder (ASD) is the name for a group of developmental disorders that impacts the nervous system and results in a range of severity from mild to severe of three main symptoms which include language impairment, challenges in social interaction, and repetitive behaviors along with many other possible symptoms such as anxiety, mood disorders and Attention-Deficit/Hyperactivity Disorder (ADHD).

ASD has a significant economic impact in the healthcare domain both due to the increase in the number of ASD cases and because of the time and costs involved in diagnosing a patient. Early detection of ASD can help both patients and the healthcare sector by prescribing patients the therapy and/or medication they need and thereby reducing the long term costs associated with delayed diagnosis. Thus, health care professional across the globe have an urgent need for the development of easy, time-efficient, robust and accessible ASD screening method that can accurately predict whether a patient with certain measured characteristics has ASD or not and inform individuals whether they should pursue formal clinical diagnosis.

However, challenges remain. Pursuing such research necessitates working with datasets that record information related to behavioral traits and other factors such as gender, age, ethnicity, etc. Such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature.

2 Problem Statement

This work aims to explore several supervised machine learning classification techniques such as Logistic Regression, Decision Trees, Support Vector Machines (SVM, together with Ensemble Learning and Multi-Layer Perceptron (MLP) to solve the classification problem of predicting whether an adult individual with certain characteristics has Autistic Spectrum Disorder (ASD).

3 Data Sets and Inputs

Our **data set** involves ten behavioral features (AQ-10-Adult) (binary data) and ten individual characteristics such as Gender, Ethnicity and Age (categorical data), and we would

like to be able to identify the most influential autistic traits. This algorithm can prove to be invaluable by helping to identify individuals who have high chance to be diagnosed with ASD and provide them with relevant treatment, therapy and counseling in a time sensitive fashion.

Variable Name	Description
Age	age in years
Gender	male or female
Ethnicity	list of common ethnicities in text format
Born with Jaundice	whether case was born with jaundice
Family member with PDD	whether any immediate family member has a PDD
Who is completing the test	parent, self, caregiver, medical staff, clinician, etc
Country of Residence	list of countries in text format
Used the screening app before	whether the user has used screening app
Screening Method Type	type of screening method chosen based on age category
Question 1 Answer	the answer code of the question based on the screening method used
Question 2 Answer	the answer code of the question based on the screening method used
Question 3 Answer	the answer code of the question based on the screening method used
Question 4 Answer	the answer code of the question based on the screening method used
Question 5 Answer	the answer code of the question based on the screening method used
Question 6 Answer	the answer code of the question based on the screening method used
Question 7 Answer	the answer code of the question based on the screening method used
Question 8 Answer	the answer code of the question based on the screening method used
Question 9 Answer	the answer code of the question based on the screening method used
Question 10 Answer	the answer code of the question based on the screening method used
Screening Score	the final score obtained based on the scoring algorithm of the screening method used.this was computed in an automated manner.

Table 1: List of Attributes

4 Solution Statement

With the data on 704 individuals my goal is to make a prediction regarding each patient and classify them into one of two categories: "patient has ASD" and "patient does not have ASD ". In other words, we are working on a binary classification problem using *supervised machine learning* with the ultimate goal of being able to classify new instances, i.e. when we have a new adult patient we wold like to able to predict whether or not that individual has high probability of having ASD. We will use supervised machine learning to refer to creating and using models that are learned from data i.e., there is a set of data labeled with the correct answer for the model to learn from.

I will also apply *Principle Component Analysis (PCA)* to figure out which of the 21 variables are most important in determining whether an individual has ASD or not.

5 Benchmark Model

Our **data set** was released on the **UCI repository** on the 24th of December, 2017. Very little work has been done with this particular data. I am unaware of any previous data classification problems related to ASD using a machine learning approach.

6 Evaluation Metrics

No one model is perfect or universally applicable and so the question that naturally arises is, how to determine which machine learning algorithm to choose for a particular classification problem such as predicting whether a person has ASD. This work will study several competing classifiers such as Logistic Regression, Decision Trees, Support Vector Machines

(SVM) and implement the one that proves most effective (both in terms of correctness and CPU time) or a combination of classifiers (*Ensemble learning*) to arrive at a decision i.e., to identify which patient has ASD. We will discuss performance statistics, and the strengths and weaknesses of each model. Our goal with this implementation is to construct a model that accurately predicts whether an individual with certain characteristics has ASD or not.

In order to choose the appropriate model that avoids *underfitting* or *overfitting* to the data we will analyze the *Bias-Variance Trade-Off*, *Model Complexity Graph, Learning Curves* and *ROC curves*. To measure the effectiveness of each model we will study the *accuracy score* along with the *precision, recall* and *F-Beta Score* and *confusion matrix*.

7 Project Design

Assuming that the available data for analytics fairly represents the real world process that we wish to model and that the this process is expected to remain relatively stable, then the data we currently have should be a reasonable representation of the data we expect to see in the future. As a result, withholding some of the current data for testing is a fair and justified way to provide an honest assessment of our model. Thus we split the given data into three parts. 70% of the data will be used to *train* the model and this data will be referred to as the *training data set* and 15% of the data will be reserved for *testing* the accuracy and effectiveness of the model on data that the model has never seen before and will be referred to it as the *testing data set*.

The random partitioning of data into testing and training data also helps us determine whether our model is *underfitting* (too simple, high bias, low variance) or *overfitting* (too complicated, high variance, low bias). We also reserve 15% of the data called the *cross validation data set* for model selection.

I will apply the following supervised machine learning algorithms and a comparative study of the methods. All algorithms have been coded using Python and its various packages(scikit learn and tensorflow with Keras).

	Run Time	Storage	Evaluation Metrics
Logistics Regression			
Decision Trees			
Random Forests			
Support Vector Machines			
Multi-Layer Perceptron			

References

[1] Brian Godsey, Think Like a Data Scientist *Manning*, ISBN: 9781633430273

[2] H. Brink, J. Richards, M. Fetherolf, Real World Machine Learning, *Manning*, ISBN: 9781617291920

[3] D. Cielen, A. Meysman, M. Ali, Introducing Data Science, *Manning* ISBN: 9781633430037.

[4] J. Grus, Data Science From Scratch First Principles With Python, *O'Reilly* ISBN: 9781491901427

[5] A. Géron, Hands-On Machine Learning with Scikit-Learn & Tensor Flow, *O'Reilly* ISBN: 9781491962299

[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second Edition, *Springer*

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, *Springer*, ISBN 9781461471370

[8] Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. *Proceedings of the 1st International Conference on Medical and Health Informatics 2017*, pp.1-6. Taichung City, Taiwan, ACM.

[9] Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

[10] Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. *Informatics for Health and Social Care Journal. December, 2017* (in press)