ASD Case Prediction and Risk Factor Identification CISC 6930- Team 10

Debarshi Dutta
Graduate School of Arts and Sciences
Fordham University
New York, USA
ddutta@fordham.edu

Henry Gorelick
Graduate School of Arts and Sciences
Fordham University
New York, USA
hgorelick@fordham.edu

***Abstract*—The paper proposes an algorithm outlining a technique to predict Autism Spectrum Disorder for adults, adolescents and children and identify the influential traits from behavioral features and individual characteristics.**

***Index Terms*—Autism, data mining, classification, feature selection**

I. Introduction

A. Autism Spectrum Disorder

Autism is a *spectrum disorder*, which means the symptoms can present in a wide variety of combinations, from mild to severe. Autism can make it difficult for an individual, most commonly children to communicate and interact with others. It can also cause individuals to perform repetitive activities and movements, become upset at changes in daily routine, and have unusual responses to certain situations.

Certain important facts about autism need to be taken taken into consideration before proceeding further-

- Autism can affect any child
- There is no known cause of autism
- There is no cure for autism

These facts highlight the need to identify the factors that cause autism.

B. Current diagnosis

Early diagnosis is key in helping a child reach important milestones. A child's first screen for autism should ideally be between the ages of 18 and 24 months. Research shows that with the early intervention and continued relevant therapies, a child suffering from the disorder can lead a productive, independent and happy life.

C. Approach to Isolate Factors

Both the causes and remedies of the autism spectrum disorder (ASD) remain unknown. Extensive research is going on to isolate these factors that cause autism and ways to prevent the disorder. The approach presented here is a classification algorithm, that has been devised using multiple data mining techniques and algorithms in ensemble and implemented using the scripting language **Python**. The algorithms used include-

- K Nearest Neighbor (KNN)
- Random Forest
- Support Vector Machine (SVM)

Besides these, conventional methods to pre-process, train and test the data have been applied as and when necessary. The aim is to identify the most important factors in ASD prediction and accurately predict ASD cases.

II. Datasets

Three datasets have been used to execute the algorithm- the Autism Spectrum Disorder (ASD) screening for adults, adolescents and children. All three are available from the UCI machine learning tab. Each dataset comes with a description and a data file. Each data file in turn contains **twenty** features that need to be analyzed to find the most important one(s). These twenty features are of two types - ten are behavioral features (based on answers provided to ten questions) and the other ten are individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science. For this approach, the datasets were used both individually and merged together

III. Pre-Processing the Data

Pre-processing was done in the following manner-

A. Explicit Manipulation

- There was inconsistency in the data with regarding the capitalization. This was corrected by making everything in the lower case.
- The first step was to rectify the misspelled columns. So,
  austim was made to **autism**
  contry_ of _ res' became **country_of_res** and
  jundice was changed to **jaundice**
- Secondly, an assessment has been designed where the total number of missing values (NaNs) is calculated with respect to the total size of the dataset. If the former is less than ten percent of the latter, then the respective missing values are simply ignored and the algorithm moves on to the next stage.
  The rationale behind this decision is that the loss of data represented by ignoring less than ten percentage of the overall dataset is too insignificant to affect the performance of the classification algorithm.

- If however, the dataset does not satisfy the above criteria then separate methods were designed and applied, according to the data.
- The **sklearn LabelEncoder** library available from Scikit was used for One-Hot encding - that is convert nominative values to corresponding numerical values. For example, the country names were converted to corresponding numeric values.

B. Questionnaires

The questionnaires are not identical. The adult is supposed to self administer, while the child and adolescent do not. Consequently, the questions do not align. This flaw was taken into account the following ways

- Similar questions were mapped by category

1) Pattern Recognition- Question 1 of child and adolescent mapped with Question 8 of Adult
2) Macro v/s Micro Thinking- Question 2 of each of the three
3) Multitasking- Question 3
4) Focus Retention- Question 4
5) Social Competance - Question 6

- Questions not occurring in each of the questionnaires were omitted.

*Note- The process described here only applies to the **merged** method- where the three datasets were combined. The questionnaires were left intact for the individual datasets.

C. Imputation

- To account for the inconsistent and sparsity of data, the assumption has been made that **mode is considered equal to the median**, that is, the item with the highest frequency is considered equal to the average of the items for the attribute.
- The **ethnicity** attribute had the highest number of missing values (NaNs). To implement imputation, the mode of the country of residence was filled in for the missing data. For the cases, where the data for the country of residence was not sufficient, **continent** column was created. The mode was calculated based on the continent, when there was not sufficient data to obtain the mode of the country of residence and the missing data was populated with the mode.
- The mode is chosen to replace the missing data in the **relation** attribute.

IV. Algorithm

*Note- Data was divided into the 70-30 rule [70% for training and 30% for testing]

- Conventional Algorithms- Three algorithms were used with the following parameters:

1) SVM- linear SVC
2) K-nearest neighbors (KNN)- k = square root of the number of instances
3) Random Forest- Tree depth used was 2

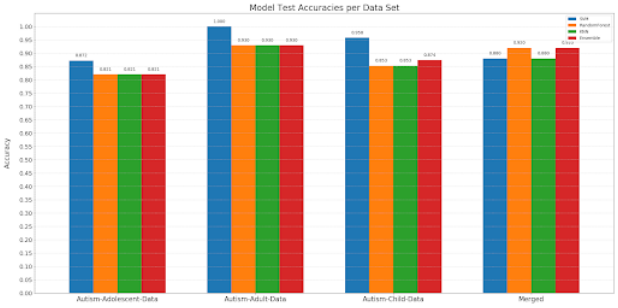

Fig. 1. Model Test Accuracy per Data Set- Iteration 1

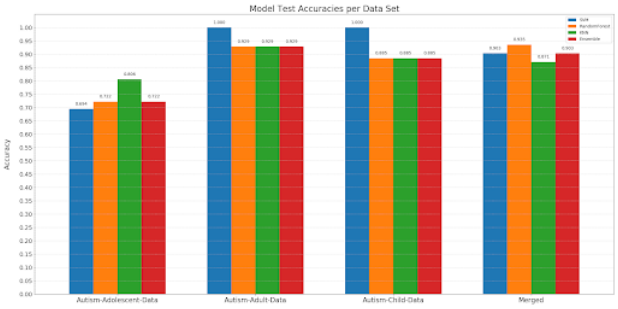

Fig. 2. Model Test Accuracy per Data Set- Iteration 2

- Ensemble method - The ensemble method of choice was the majority of the above three classifiers.
  For both the prediction and feature analysis, the algorithm was executed once and the results were tabulated. This is important because the results (both accuracy and feature analyses) will be different for every iteration.
  *Note- An inherent method of improving the algorithm is to run it multiple times and then find the mean.

V. Prediction Accuracy

The most significant discovery after the initial iterations were that the **results** attribute was far and away the most important factor. With an accuracy of almost 100% for each iteration compiled, that single attribute was affecting the algorithm negatively as such a high accuracy for a data of this size inevitably meant overfitting of the training data. So an innovative normalization technique has been devised that reduces the impact of the **results** attribute which has been described later on.

A. Analysis

Figures 1, 2 and 3 clearly show the effectiveness of the individual algorithms and the ensemble one across all three datasets and the merged one. Since random forest itself is an ensemble method, it is safe to assume that this ensemble method is a valid one.

VI. Identification of Important Features

It is to be taken into account that **results** attribute is unanimously the most important one. But since it impacted the model negatively, the impact of results was normalized.

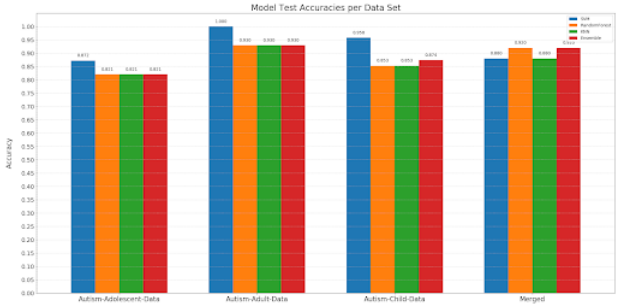

Fig. 3. Model Test Accuracy per Data Set- Iteration 3

Fig. 4. Feature Importance by Model per Data Set

A. *Normalize **results***

An innovative approach was designed to normalize the effects of the results attribute. The **results** was distributed defined by the following ranges. $\{0 \leq x < 2.499$: 0, $2.5 \leq x < 4.999$: 2.5, $5 \leq x < 7.499$: 5, $7.5 \leq x < 9.99$: 7.5, $9.99 \leq x$: 10$\}$ Fig:4 shows the important features as produced by SVM and Random Forest after **results** has been normalized.

VII. Analysis

As identified in Fig:4 and using the mapping technique described for the questionnaires earlier, the following common factors across all the three datasets are identified as the key factors

- Focus retention
- Age-description

while, the following non-common factors are also essential

- Results
- Relations
- Gender
- Age

Two key observations of note here

1) Both **Age-Description** and **Age** are important features. The former is the age range of the candidate and the latter is the explicit value. This is significant because autism is normally diagnosed in children, that is a particular age range, leading credence to the results.
2) **Results** is such an overwhelmingly important feature that in spite of normalizing it, it still appears as an important feature