MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190477

Detecting Child Autism Using Classification Techniques

Md Delowar Hossain, Muhammad Ashad Kabir

School of Computing and Mathematics, Charles Sturt University, NSW, Australia

Abstract

Autism spectrum disorder (ASD) is a brain development disorder that restricts a person's communication abilities and social interaction capabilities from natural growth. In this paper, we have applied various supervised classification techniques to detect the presence of child autism. Our findings show that the Sequential Minimal Optimization (SMO) classifier performs best to detect ASD cases with the highest accuracy and minimum execution time and error rate. We also identify the most dominant features in dectecting child autism.

Keywords:

Autism Spectrum Disorder, Supervised Machine Learning, Child

Introduction

Autism spectrum disorder (ASD) is considered as a neurodevelopment disorder that hinders daily communication and social behavior from natural growth [1]. Data mining classification techniques can be applied to detect ASD cases. The main objectives are to reduce the diagnosis time in order to get quicker access to healthcare services, improve diagnosis accuracy, and finding the highest ranked features of ASD.

There are number of prior studies that have applied classification techniques on adult ASD data. For example, K. Basu *et al* [2] analyzed the Adult Autism screening data set using supervised data mining techniques and showed that Support Vector Machine (SVM) classification technique outperforms other classifiers. Later, Brian McNamara *et al* [3] classified the same dataset by applying decision tree and random forest classifiers. They pre-processed the dataset by removing the records with missing categorical instances and less significant variables, before applying the classifiers. They found that random forest outperforms over decision tree classifier.

Compared to the aforementioned research, in this paper, we apply supervised learning techniques on child ASD data. To the best of our knowledge, our study is the first one that applied supervised learning techniques on ASD dataset of children aged 4 to 11 years. Our goal is to analyze the dataset using existing classification techniques and classify them in one of the two categories: "children having ASD" or "children not having ASD". We compare the performance of various methodologies to determine the best classification technique for this dataset. Our results show that Sequential Minimal Optimization (SMO) classifier performs best amongst all the supervised classifiers. We also analyze the dataset to find out dominant features that cause ASD based on the answers given to the ASD questionnaire.

Methods

In this section, we present our approach for detecting ASD cases using various classification techniques. It includes several steps such as dataset exploration, data preprocessing, and classification.

Data Set Exploration

We used the dataset from UCI Machine Learning Repository [4]. The child dataset contains ten binary features (A1_Score to A10_Score), two numeric features - age, results, and categorical variables such as gender, ethnicity, jaundice status, family member having PDD (Pervasive Developmental Disorder), country of residence of the person who answered the survey, used the screening app before, age description, and ASD class. The dataset contains 292 records and 19 attributes after selecting the screening type.

Data Preprocessing

In order to simplify our analysis, we discarded less significant variables such as used_app_before, country_of_res, and age_desc [2, 3]. By analysing the "results" feature, we found that result score =>7 indicates ASD positive and score <7 indicates non-ASD. So, we excluded this attribute before classification to avoid predefined situation where the output is already known. We removed all of the records containing missing values. We found 43 ethnicity entries missing in the dataset, so we removed those records. We also noticed "age" was missing for one record. So, we replaced this missing value with the median value of age. Finally, our observations on final datasets count are shown in Table 1.

Table 1- Final Child Dataset

Female	Male	ASD_Class
38	85	No (123)
36	90	YES (126)
Total number of cases:		249

Results

We analyzed the child dataset and applied 28 supervised classification techniques of different groups such as Rules, Bayes, Function, Lazy, Meta (Decision Tree used a base algorithm) and Tree. We applied pruning for Tree base algorithm. We used WEKA data mining tool and applied 10-fold cross validation. We found the classifiers from "Function" such as Multilayer Perceptron, Simple Logistic, SMO, classifiers from "Meta" group such as Iterative Classifier Optimizer, LogitBoost, and Real Adaboost, and LMT (classifier from Tree group) result in 100% accuracy, precision, recall, and F-measure (Table 2). We also determined the dominant features for child Autism by analyzing the LogitBoost supervised classifier and performing weight-based analysis. We found the following are the most dominant features in detecting child autism:

- *A4_Score*: S/he finds it easy to go back and forth between different activities
- *A10_Score*: S/he finds it hard to make new friends
- *A8_Score*: When s/he was in preschool, s/he enjoys playing games involving pretending with other children

Here, *A4_Score* is related to child's intellectual disability. *A10_Score and A8_Score* are related to social interaction abilities. Autistic child find it difficult to make new friends and sometimes they enjoy playing games involving pretending with other children. This implies that difficulty with social interaction is a key symptom of ASD. Answering "yes" to these questions are dominant contributors to the model prediction.

Table 2- Classifier Performance Statistics for the child dataset

Classifier	Accuracy	Precision	Recall	F-measure
ZeroR	50.60	?*	0	?*
OneR	78.31	0.82	0.71	0.76
PART	89.95	0.90	0.88	0.89
ByesNet	96.78	0.97	0.95	0.96
Naïve Bayes	97.59	0.98	0.96	0.97
Naïve Byes Updateable	97.59	0.98	0.96	0.97
LibSVM	96.38	0.98	0.94	0.96
Multilayer Perceptron	**100**	**1**	**1**	**1**
Simple Logistic	**100**	**1**	**1**	**1**
SMO	**100**	**1**	**1**	**1**
IBK	91.96	0.972	0.86	0.91
KStar	90.76	0.99	0.82	0.89
LWL	76.70	0.788	0.72	0.75
Bagging	85.94	0.87	0.82	0.85
Classification Via Regression	92.77	0.92	0.93	0.92
Iterative Classifier Optimizer	**100**	**1**	**1**	**1**
LogitBoost	**100**	**1**	**1**	**1**
Multi class Classifier	98.79	0.99	0.98	0.98
Multi class Classifier Updateable	99.19	1	0.98	0.99
Random Committee	91.16	0.89	0.93	0.91
Real Adaboost	**100**	**1**	**1**	**1**
Hoeffding Tree	97.59	0.98	0.96	0.97
J48	90.76	0.91	0.89	0.90
LMT	**100**	**1**	**1**	**1**
NBTree	95.18	0.96	0.93	0.95
Random Forest	96.78	0.97	0.95	0.96
Random Tree	80.72	0.83	0.76	0.79
SysFor	86.34	0.85	0.87	0.86

Note. *no result produced

We observed different types of error during the classification time such as mean absolute error, root mean square error, relative absolute error, and root relative squared error. These errors are the outcome of the difference of our predicted model and observed data. We determined which classifier has the least error or is totally free of errors.

From Table 3, we can see that the best performing classifier is SMO which implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.

Table 3- Comparison of Classifier's Error and Execution Time

	Mean Absolute Error	Root Mean Square Error	Relative Absolute error	Root relative squared error	Execution Time (sec)
Multi-layer Perceptron	0.005	0.026	1.102	5.357	1.22
Simple Logistic	0.061	0.117	12.20	23.50	0.24
SMO	**0**	**0**	**0**	**0**	**0.06**
Iterative Classifier Optimizer	0.098	0.156	19.73	31.22	0.3
LogitBoost	0.098	0.156	19.73	31.22	0.1
Real Adaboost	0.069	0.13	13.90	25.99	0.03
LMT	0.061	0.117	12.20	23.50	0.14

It breaks the problem into a series of smallest possible sub problems, which are then solved analytically. As a result, this classifier has least execution error (0 in this case) and execution time (0.06 second).

Conclusions

In this study, we classified ASD child dataset using supervised classification techniques and found that SMO classifier performs best in terms of accurate prediction of child ASD. Additionally, SMO is also free from classification errors, suggesting that it aligns well with 10-Fold cross validation. By analyzing our experimental results, we also identified the most dominant features that significantly contribute to detecting child autism.

References

[1] P. Bolton, H. Macdonald, A. Pickles, P. Rios, S. Goode, M. Crowson, A. Bailey and M. Rutter, A Case-Control Family History Study of Autism, *The Journal of Child Psychology and psychiatry* 35(5) (1994), 877-900

[2] K. Basu, Machine learning approaches to the classification problem for autism spectrum disorder, *GitHub*, 2018. https://github.com/kbasu2016/Autism-Detection-in-Adults/blob/master/report.pdf (accessed February 23, 2018)

[3] B. McNamara, C. Lora, D. Yang, F. Flores and P. Daly, Machine Learning Classification of Adults with Autism Spectrum Disorder, *RPubs,* 2018. http://rpubs.com/brianmcnamara/ASD_Classification (accessed March 2, 2018)

[4] F. Thabtah, Autistic Spectrum Disorder Screening Data for Children Data Set, https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++# (accessed Feb 5, 2018)

Address for correspondence

Md Delowar Hossain,
Email: mailtodelowar@gmail.com