Prediction and Comparison using AdaBoost and ML Algorithms with Autistic Children Dataset

Mary Stella. J
Department of CSE
Cambridge Institute of Technology
Bangalore, India

Dr. Shashi Kumar
Department of CSE
Cambridge Institute of Technology
Bangalore, India

Abstract- **Autism Spectrum Disorder in children is a neural disorder behavior which could be detected and predicted using Machine Learning techniques. Autism in children is a kind of inability in socio-behavioral and in communicative behavior. If not detected between 20 to 60 months and treated, the treatment would be exceedingly difficult. The early the detection, early would be the treatment and therapy, which still would be challenging. Many different ML techniques are imposed but still its beneficial prediction is inadequate in predicting autism of small age groups. In this paper, we imposed three algorithms svm, random forest and AdaBoost algorithms to detect and predict autism in children. AdaBoost classifier is our proposed algorithm that combines weak classifiers to club and boost to strong classifier. For performance comparisons we calculate accuracy, precision, F-score, and confusion matrix. Best accuracy yielded algorithm is implemented to predict autism in children.**

Keywords- Machine Learning, Autism, svm, Random Forest, AdaBoost.

I INTRODUCTION

Autism is social, communication disability, and behavioral disability. It is based on mainly the response of children to cognitive functions. It is characterized by the impairments of verbal and nonverbal communication and repetition of stereotypical behaviors. Unfortunately, autism disorder outgrows faster, even if diagnosed in any age of a human life, its symptoms generally appears in first 2years of human life [3]. Autism patients would face many challenges like not responding properly, learning disabilities, difficulty in concentration, sensory problems, anxiety and depression, motor difficulties etc. ASD affected children exhibits many symptoms in terms of family backgrounds, morbidity, and the cost which may vary from one child to the other. Researchers state that autism could be cause of genetic, nongenetic influences and environmental situations in children lives. Symptoms in children of early age can be identified if the children do not react with the parent's response, friends, and other children interactivity [5].

To overcome difficulties in children affected by autism, we proposed some techniques in Machine learning using algorithms to diagnose and predict whether autism exist in children or not, effectively. Machine learning techniques are useful in finding needful and useful information from data stored in long terms [6]. Machine Learning mines the hidden relationship scattered in a large database and retrieve categorical data for implementation. ML algorithms implement the acquired meaningful data for processing to predict any disease and treating measures. Machine Learning is widely beneficial in healthcare domain where the human lives are in risk and the doctors could diagnose the disease very easily and effectively treatments could be done to the patients in early stages.

II LITERATURE REVIEW

Smart Autism is an automated framework which is cloud based used to confirm and screen autism. Due to the lack of resources, in developing countries, where expertise is not improved, autism detection in later ages, delays timely intervention. Therefore, our proposed work is an interactive and integrated framework, using a mobile to confirm and screen for autism of different age groups from 0 to 17 years. It requires three levels of assessment process. Firstly, through the mobile screening is done by evaluating questionnaire in response of pictorial representation. If autism is detected in this stage, next the virtual assessment process is undertaken. In this, the child is admitted watching a video, its response and reaction towards the video is recorded and uploaded in cloud for expert assessment remotely. If autism is suspected in this stage, the child parents or autistic person now is advised to an Autism Resource center for a proper assessment. This proposed work of integrated framework can thereby reduce the users ARC visit which can also create an awareness [4].

Clinical instruments that are used currently in measuring the ASD symptoms tend to be time consuming, and strongly influenced with subjective observations. In cases, it tends to be delayed diagnosis and intervention towards it. And therefore, a gaze movement is invented by the scientists which is the biomarkers for detecting ASD. In this paperwork, we tend to speedup the autism diagnosis by combining Machine Learning with the gaze-based screening as a transformative process for tracing autism in early ages. In gaze screening, the three key features are data collected using eye tracking, feature extraction and building the predictive model. As a machine learning technique, we impose support vector machine as a technique and finding the performance measures, that are specificity, sensitivity, area under curve, and accuracy. It is found that SVM accomplishes the high-performance classification applying on eye movement dataset [2].

ASD is the disability in human that separates them from normal humans, compared to the behavioral analysis and communicative disorientation. ASD cases found to be increasing in number of cases in world which needs to develop various screening methods. In this proposed work machine learning algorithms are compared with its performance that consists of many classifications like random forest, naïve Bayes, IBK (K-nearest neighbor), Radial Basis function network. The performance measures were imposed on UCI dataset of 2017. The result is derived by analyzing the algorithms, found that, Random Forest produces successive measure compared to IBk, Naïve Bayes and others [1].

Having high dimensionality and non-linearity of the data, false negative rates are high to be negligible. In this paper, we combine the existing classifiers, say simple classifiers to form an ensemble classifier for data expression. The classifier used are naïve Bayes, KNN, and decision trees (DT). Two main key issues in ensemble learning are the integrating multiple classifiers and the diversity of base classifiers. In this work, a decision group, a special structure of classifier, is designed in increasing the performance. For the weights that are assigned for each classifier the genetic algorithm is imposed. This work introduces an ensemble algorithm which is based on AdaBoost. AdaBoost and genetic algorithm are the proposed algorithm for diagnosing cancer with the classification of the gene data. The two challenges in this work is how the integration and the base classifiers could improve the performance. The base classifiers include KNN, naïve Bayes, and the decision trees. The work results that AdaBoost and GA (genetic algorithm) improves the performance very effectively with selected classifiers in our work [8].

III BACKGROUND INFORMATION

Machine Learning is a part of AI application which trains the system by providing training data and makes the system learn automatically and gathers experience without human intervention. With this enough experience, it predicts the output for any given input without using human intervention, just relying on the patterns and the inferences. ML is generally categorized into three main learning. Supervised, unsupervised and reinforced learning.

Supervised Learning is applying learned pattern, parameters of the past in the input instance with labelled data. The supervised learning can be employed with many ML algorithms. Some supervised algorithms are SVM, Random Forest, naïve Bayes etc. The algorithms generate an inferred function starting with the analysis with the training data, to predict the outcome. Labelled data of the instances provides time consuming and provides targets for input instances of any numbers with multiple training. Finally, the algorithm can now compare the trained result with the intend output and calculate the accuracy, sensitivity, specificity, and the errors. The errors are adjusted in model to make it null and to produce the high accuracy.

A. Dataset

In predicting autism for the children, in our work we used Autism Screening Data for Children dataset (Toddler Dataset). The dataset consists of 1054 records of the children from 12 months to 36 months. Each record has been selected with 15 features containing binary values and string values as well. Feature Engineering is imposed on the data, such that the string values are converted to binary values, which can be now used for training and classification. Our dataset can be used on text classification and can be imposed on algorithms that works with text classification.

B. Algorithms

Random Forest is a supervised learning model, which uses labelled data and learns to classify any input unlabeled data. RF is used to solve classification as well as regression problems, by making it a diverse model. It is amazingly fast to train the test data. Random Forest is composed of huge number of DT (decision trees) and can be used as ensemble methods also. RF supports bagging where the building of each tree is created the uncorrelated forests in feature randomness. RF finds difference from decision trees, like, DT is built with all features on the entire dataset whereas random forest selects the records randomly to build the tree and to result the outcome.

SVM, like random forest, is used to solve challenges in classification and also regression. Support Vector Machine would not need depth knowledge of mathematics But in Random Forest once the model is made, it slows down in creating predictions for the outcome., rather is uses the hyperplane where we locate each data in the n-dimensional space with feature value coordinate. The hyperplane maximizes the margin of the two classes, the vectors which defines the hyperplane is said to be the support vectors. Kernel Trick is used in SVM, to operate in its feature space instead to compute the coordinates of data in higher dimensional plane. Therefore, it offers less expensive way and more efficient to convert data into higher dimensions. But when comes to larger dataset it does not perform well as it requires more training time.

AdaBoost algorithm is an ensemble learning with modern technique to solve complex classifications in integrating simple weak classifiers to strong classifiers. Ensemble learning models are designed on two approach, Boosting and Bagging. AdaBoost is one among the boosting algorithm where the weights of the individual instance is iteratively determined on relying the accuracy of the last classification outcome. It is a high accurate classifier which offers the error rate to close to zero. AdaBoost seems to be sensitive with noisy data and outliners. AdaBoost must undergo two actions, first, the classifier is trained iteratively on various weighed training data. Second, in each iteration it provides good fit for the instances by minimizing the training error. We can use many base classifiers with the AdaBoost, and it is found that its not prone to overfitting.

IV PROPOSED WORK AND RESULT ANALYSIS

Our proposed work uses three algorithms SVM, AdaBoost, and Random Forest algorithms on comparison and predict the best accuracy and produce an outcome for any given input. The Proposed architecture is given below in Fig 1.

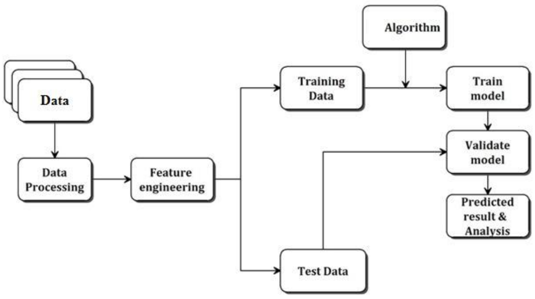


Fig1. Proposed architecture

The basic criteria in comparing the performance of the classifiers is to measure the effectiveness of the algorithms.

A. Precision

Precision gives the output quality of the model by evaluating the below mentioned formula.

$$Precision = \frac{TP}{TP + FP} \qquad (1)$$

Precision can be calculated by dividing the true positive to the summation of true positive and false positive values. It is a measure of result relevancy[12].

B. Recall

Recall is also another metric to find the output quality to find how many true relevant results are obtained. Recall is sensitivity.

$$Recall = \frac{TP}{TP + FN} \qquad (2)$$

In mathematical form, the true positive values are divided by the summation of true positive and false negative values of the instances, both of which are correctly classified.

C. F1 score

F1score is the weighted average of recall and precision. It gives the single score that balances of precision and recall.

$$F1\ score = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \qquad (3)$$

D. Accuracy

Accuracy is the overall classification validation with overall classification ratio

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+ FN}} \qquad (4)$$

The confusion matrix is the prescribed general tool to measure the classification performance. It is measured against the true cases and the predicted cases with positive and negative outcomes [10].

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Fig 2. Confusion matrix

In Fig 2. The true cases against the predicted cases are intersected with four possible outcomes. True Positive means the children who detected to be autistic children (disease) with classification also. False Positive means children who are not actually sick and classified not sick by the classifier. False Negative means the children who detected as positive to autistic but classified as not autistic children. True negative means the children is detected not autistic and classifier results as not autistic children. The confusion matrix is calculated for three algorithms and the autistic children with true positive are detected.

Table 1. Performance measures obtained using confusion matrix for three algorithms

Algorithms	Accuracy	Precision	Recall	F1 score
Random Forest	96.20	0.97	0.95	0.96
SVM	96.68	0.97	0.96	0.96
AdaBoost	100	1.00	1.00	1.00

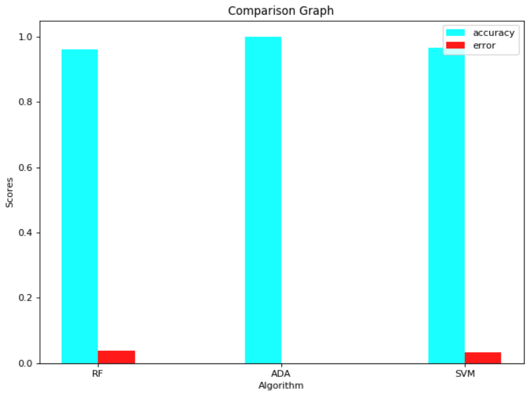
In Table 1. the performance measures of accuracy, precision, recall, and f1 score are calculated for RF, SVM, and AdaBoost algorithms, respectively. In this table, the accuracy of random forest is calculated as 96.20%, and SVM is achieved to 96.68% whereas in AdaBoost classifier the accuracy is achieved to 100% which is termed to the best accuracy compared to SVM and RF classifiers. Also, considering the precision, recall and the f1 score it is compared and studied that the recall value of both Random Forest and SVM is 0.95 and 0.96 which is less than the AdaBoost algorithm which scored 1.00 comparatively. The f1-score achieved by RF and SVM are same like 0.96 but the f1-score of AdaBoost is gained to 1.00 and overall based on accuracy, recall and f1-score it is observed that AdaBoost works best than the Random Forest and SVM classifiers and to predict autism for any given input of children data.

Table 2. Accuracy and Error Rate predictions of three algorithms

Algorithms	Accuracy %	Error rate %
Random Forest	96.20	3.79
SVM	96.68	0.33
AdaBoost	100	0

Viewing the Table 2. The accuracy achieved in AdaBoost is higher than the Random Forest and SVM. Also, the error rate is achieved as the ratio of incorrectly predicted to correctly predicted instances. In this, the Random Forest calculated is 3.79% error rate and the SVM calculated is 0.33%, whereas the AdaBoost classifier has achieved zero error rate, which is very negligible to predict the error. Finally, the AdaBoost reduces the error rate to zero and achieves good accuracy classifier.

Fig 3. Comparison Graph of the three classifications



In Fig 3. The comparison is made and reflected in the table for the three respective algorithms and the graph is plotted to represent the best classifier to achieve best accuracy and good for autism prediction accurately and precisely.

V CONCLUSION

In recent years, the boosting algorithms have gained the massive popularity in Machine Learning and Data Science. In the precise accuracy generating competitions, boosting algorithms are used to achieve the high accuracy. The experimental outcomes reveal that the proposed algorithm yields good accuracy and comparatively better performance and it is used to predict the autism traits for any input data of children. AdaBoost algorithm works efficient with large dataset and high feature selection. AdaBoost and other boosting algorithms are less affected by the overfitting problems. In our conclusion, its clearly detected and predicted that AdaBoost classifier gains 100% accuracy compared to SVM classifier and Random Forest classifier which achieved 96%. For any given input data of children, the presence or absence of autism associated with the trained model is predicted with the best classifier, here say, the AdaBoost Classifier. In future achievements, the algorithms can be analyzed with very large dataset and further boosting algorithms can be implemented in further studies.

REFERENCES

[1] Fatiha Nur Büyükoflaz, Ali Öztürk, “Early Autism Diagnosis of Children with Machine Learning Algorithms,” Bilgisayar Mühendisli˘gi Bölümü, KTO Karatay Üniversitesi Konya, Türkiye, 2018 IEEE.

[2] Alanoud Bin Dris, Abdulmalik Alsalman, Areej Al-Wabil, Mohammed Aldosari, Centre for Cyber Security Technology, King Abdulaziz City for Science and Technology (KACST) Riyadh, Saudi Arabia, 2019 IEEE.

[3] Kazi Omar, Nabila Khan, Prodipta Mondal, Md. Rezaul, Md Nazrul Islam, “A Machine Learning Approach to Predict Autism Spectrum Disorder,” International Conference on ECCE, February 2019.

[4] Khondaker Abdullah Al Mamun et al, “Smart Autism – A mobile, Interactive and integrated framework for screening and confirmation of autism,” Conference IEEE engineering in medicine and biology society, 2016.

[5] Osman Altay, Ulas, “Prediction of the Autism Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbor in Children,” Firat University, Turkey 2018.

[6] Sushama Rani Dutta, Sujoy Datta, Monideepa Roy, “Using Cogency and Machine Learning for autism detection from a preliminary symptom,” KIIT University, Odisha, India, 2019 IEEE.

[7] Simone khalifeh, Walid Yassin, Silva Kourtian,Rose mary Boustany, “Autism in Review,” http://www.lebanesemedicaljopurnal.org/articles/review1.pdf

[8] Huijuan Lu, Ke Yan, “A Hybrid Ensemble Algorithm combining AdaBoost Algorithm with Gene Expression Data,” 9th - International Conference in IT on Medicine & Education, 2018.

[9] Elizabeth Stevens, Abigail Atchison, Laura Stevens, Esther Hong, Doreen Granpeesheh, Dennis Dixon, Erik Linstead, “A Cluster Analysis of Challenging Behaviors in Autism Spectrum Disorder” 2017 16th IEEE International Conference on Machine Learning and Applications.

[10] Hardi Talabani, Engin AVCI,” Performance Comparison of SVM Kernel Types on Child Autism Disease Database,” Faculty of Technology, Firat University, Software Engineering Department, Elazig, Turkey, 2018 IEEE.

[11] Kajaree Das, Rabi Narayan Behera, “A Survey on Machine Learning: Concept, Algorithms and Applications,” Institute of Engineering and Management, Kolkata, India.

[12] Alrence Santiago Halibas, Leslyn Bonachita Reazol, Erbeth Gerald Tongco Delvo, Jannette Cui Tibudan, “Performance Analysis of Machine Learning Classifiers for ASD Screening,” 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies.