Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Empirical Evaluation of Data Mining Classification Methods for Autistic Children

[1]Sumi Simon, [2]Chandra J and [3]Saravanan N
Department of Computer Science, Christ University, Bangalore, India

Abstract: Autism is a mental neural development disorder that is present from early childhood. Autism is characterized by difficulty in verbal and nonverbal communication, impaired social interaction and repetitive and restricted patterns of behavior. Parents notice these unusual behavior in the first two years for their toddler's age.

Autism is also known as autism spectrum disorder, the term is coined as spectrum due to its wide range of symptoms, levels of impairment or disability which vary from each toddler. Everyone with autism is unique. Autism can be classified as low, medium and high level of autism based on the scale used to detect autism. Surveys conducted by various child development organization across the world proves that boys are at a higher risk of autism than girls. The approximate ratio is four to five times higher in boys than in girls. The foremost and primary objective of the paper is to perform an empirical evaluation to compare the existing methods for data collection, preprocessing and classification methods for predicting autism.

Keywords: *Autism Spectrum Disorder (Asd), Data Mining, Preprocessing and Classification*

I. Introduction

Autism Spectrum Disorder (ASD) is a group of complex disorders of brain development that causes social, communication and behavioral challenges that continues lifelong. ASD is associated with various characteristics like repetitive behavioral patterns and activities, lacking social interaction or communication, difficulty in verbal and nonverbal actions and these symptoms of autism emerges between two to three years of age. According statistical analysis, one in sixty eight children are affected with autism. Studies have proved that autism is seen almost five times more commonly among boys than among girls. Diagnosing ASD is difficult at an early stage as there is no medical test or blood test to diagnose the disorder. Doctors look at the child's behavior and development to make an assessment. ASD can be detected at eighteen months or younger age based on the child's behavior. By age two, an assessment by an experienced professional can be considered very reliable. However, many children do not receive a final diagnosis until much older. This delay leads to delayed diagnosis of the disorder and the children doesn't get the early help they need. The work shows a broad empirical evaluation on classification and prediction of autism. The rest of the paper is organized as following. Section II summarizes the related work carried out related to autism classification using data mining classification methods. Section III and IV speaks about various classification methods followed by conclusion in section V.

II. Related Work

Numerous work is done in the area of autism for classification, prediction and behavioral analysis for categorizing the level of autism, to provide awareness to the caretakers and parents, to provide the right kind of therapy. Some of the work done earlier related to autism classification using data mining methods are defined in this section. Data mining is a promising field for the current researches that leverages information to train and test the underlying data to classify and predict the disorder that is of great value to the caregivers and the autistic children to improve the quality of their life. Data mining techniques are applicable for description, classification and prediction or optimization. The goal of data mining is to prove human understandable knowledge. Margaret H Dunham [1] describes the data mining concepts in a very simple lay man understanding and her work covers pretty much of data mining techniques, its usages and limitations. Data mining is used to predict various disorders in human, plants, animals and even pest diseases. Shika et al. [2] used a fuzzy neural network based computing approach for predicting pest disease in plants. The work included classification stage to identify the disease criticality in plants. Milan Kumari et al. [3] did a comparison on data mining classification methods for cardiovascular disease prediction. The work aimed at comparing the performance through accuracy, error rate and specificity of different data mining methods like RIPPER, Decision tree, ANN and SVM and concluded that SVM yielded the most accurate results to predict the cardiovascular disorder. Heart disease prediction using Bayesian classifier, KNN, decision tree and neural networks are done by Jyothi Soni et al. [4] and concluded that Bayesian and Decision tree accuracy is further improved by applying the generic algorithm that reduces the data size and creates a subset of the data for the disease prediction. Data mining is well applicable for predicting ASD. Mohana E et al. [5] used data mining techniques to categorize the risk level of autistic children. To categorize the autistic children, M-CHAT-R (Modified checklist for autism in toddler's revised tool) dataset is used and different feature selection algorithms-Fisher filtering, ReliefF, Runs filtering and stepdisc are used to extract the required features to characterize the risk level of autism. The classification algorithms like BVM, C4.5, C-RT, CS-MC4, C-SVC, CVM, ID3, K-NN, Linear discriminant, Naïve Bayes, PLS-LDA and Rnd tree are applied to the feature selection algorithms and the accuracy and error rate are calculated and work proves that BVM, CVM and MLR provides the best classification. Cheol Hong Min et al. [6] uses a pattern detection in the accelerometer data using iterative subspace detection algorithm that automatically detects stereotypical behavioral patterns in the sensor data. With these methods ninety percent classification accuracy is achieved. A novel approach to predict the learning skills of autistic children is suggested by Mythili et al. [7] that uses SVM and decision tree. Weka tool is used to classify the learning skills of the autistic kids. Weka (Waikato Environment for Knowledge Analysis) is a very easy yet powerful tool for knowledge analysis [8], most researchers use weka for analysis related to data mining. Mythili et al. [9] detects the level of autism with data mining classification algorithms like neural network, SVM and fuzzy logic to classify the autistic children. Mythili et al. [10] proposed an improved feature selection algorithm to predict the learning skills of the autism affected kids. SVM, J48, multilayer perceptron and IB1 classifiers along with the filters and wrappers are used to build the improved feature selection model to achieve better accuracy and to detect the learning skills of the autistic children. Siriwan Sunsirikul et al. [11] used associative classification mining to detect the behavioral patterns of the autism affected kids. The dataset used is categorized by the doctors as – autism and pervasive development disorder. The methodology included data preprocessing, rules generation, building the classifier and cross validation of the test data with the trained data. The classifier was effectively able to classify the behavioral patterns between the autism and pervasive developed disorder children with relatively high accuracy. How thinking pictures can explain the behaviors of the autism was proposed by Maithilee Kunda et al. [12]. They developed a cognitive account of pictorial representations for the autistic children and put forth the hypothesis "think in pictures". The methodology adopted includes classification, visual attention and reasoning and neurobiological evidence to detect the thinking capabilities of the autistic children. Joao F. Santos et al. [13] proposed an early prediction of ASD based on the analysis of pre-verbal vocalizations of the autistic children. The acoustic prosodic features are extracted and provided as a training data to the SVM and probabilistic neural network classifiers and it is concluded that classification of autism is achieved by analyzing the pre-vocalization of the autism affected children. The SVM and neural network classifier achieves ninety seven percent accuracy.

III. METHODOLOGY

The collected dataset is usually associated with a high level of noise. Noisy data includes missing values, inappropriate vales, null values or redundant values. It is essential to take care of these noisy data to make it consistent and complete so as to yield the best outcome. Data preprocessing technique is used to remove the noisy or unwanted data in the dataset. Post preprocessing, the data will be clean, clear, noiseless data, free from missing values and redundant values. Data pre-processing also includes converting the qualitative data into quantitative data which is converting the data into numerical format.

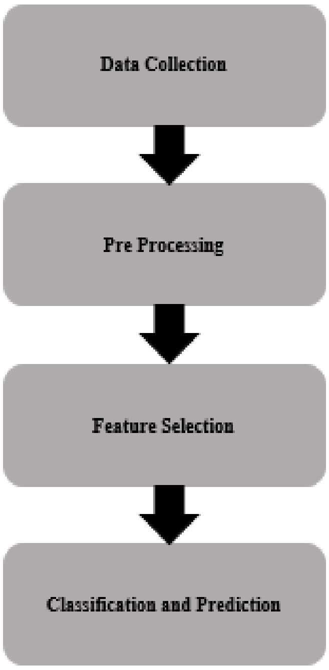

Figure 1: Process Flow

Predicting the autism level of autistic children using data mining classification methods includes three main steps as shown in figure 1. They are preprocessing, selected feature extraction and classification. Table 1 shows the data features that are the behavioral features used for evaluating autism.

Table 1: Dataset Features

Sl No	Attribute Group	Attribute Subgroup
01	Social Relationship And Reciprocity	Has poor eye contact
		Lack social smile
		Remain aloof
		Does not reach out to others
		Unable to relate to people
		Unable to respond to social-environmental cues
		Engages in solitary and repetitive play activities
		Unable to take turns in social interaction
		Does not maintain peer relationships
02	Emotional Responsiveness	Shows inappropriate emotional response
		Shows exaggerated emotions
		Engages in self-simulating emotions
		Lack fear of danger
		Excited or agitated for no apparent reason
03	Speech-Language and Communication	Acquired speech and lost it
		Has difficulty in using non-verbal language or gestures to communicate
		Engage in stereotyped and repetitive use of language
		Engages in echolalic speech
		Produces infantile squeals/unusual noises
		Unable to initiate or sustain conversation with others
		Uses jargon or meaningless words
		Unable to grasp pragmatics of communication (real meaning)
04	Behavior Patterns	Engage in stereotyped and repetitive motor mannerism
		Shows attachment to inanimate objects
		Shows hyperactivity/restlessness
		Exhibits aggressive behavior
		Throws temper tantrums
		Engage in self-injurious behaviour
		Insists on sameness
05	Cognitive Components	Inconsistent attention and concentration
		Shows delay in responding
		Shows unusual memory of some kind
		Shows 'savant' ability
06	Sensory Aspects	Unusually sensitive to sensory stimuli
		Stares into space for long period of time
		Shows difficulty in tracking objects
		Shows unusual vision
		Insensitive to pain

		Responds to objects/people unusually by smelling, touching or tasting

A. *Feature Selection*

Feature selection is an important step post preprocessing for dimensionality reduction that helps in improving the learning accuracy and getting higher learning accuracy. Feature selection is a step of choosing only the relevant attributes of the dataset forming a subset of the original features. Feature selection involves different steps as shown in figure 2.Various feature selection algorithms like fisher filtering, relief, runs filtering and stepdisc are compared in [5] and it is seen that relief and runs filtering are best suited approach for feature selection for autism behavioral data. Relief and runs filtering picks the most relevant attributes and can reduce up to half the original behavior attributes.

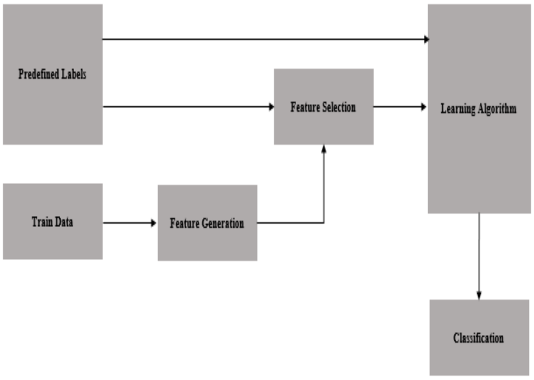

Figure 2: Generic Framework for Feature Selection

Figure 3 shows the proposed architectural design for the effective classification of autistic children. Autistic data is collected from National Institute for the Mentally Handicapped who provide an Indian scale for assessment of Autism. The form is a collection of various statements to observe and assess autism. Relevant feature is extracted using feature selection algorithms like relief and runs filtering which as these algorithms provide high accuracy for feature selection. Classification methods are discussed briefly in the next section.

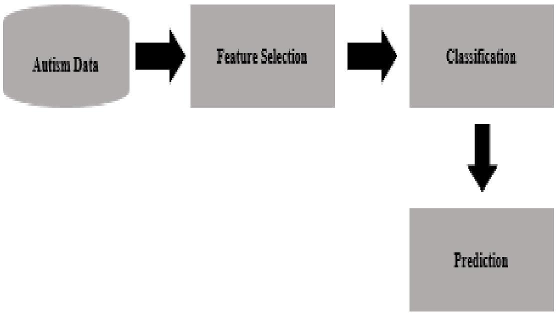

Figure 3: Proposed Architectural Design

IV. CLASSIFICATION METHODS

Classification is a process of predicting the class labels of the underlying dataset. These classification are used for predicting the resultant dataset. Classification algorithms helps in differentiating between different features. Classification has various steps like making the machine learn the patterns, basically training data. Next step is the generation of the classifier to correctly classify the dataset into different class labels and finally evaluating the classification using the test data. Generally approached step for classification of autistic data is through supervised learning. Classification includes dividing the dataset into train and test data where train data is used to train the classifier and test data is used to test the performance and accuracy of the classifier built. It is advisable to keep 70:30 partitioning of test and train data respectively to achieve higher accuracy of the classifier. Classification is a methodology of supervised learning.

SVM, J48, Multilayer Perceptron and IB1 classification algorithm are compared to check the classification accuracy on behavioral data and it is concluded that Support Vector Machine (SVM) provides highest classification accuracy as tabulated in table 2. SVM is a simple and effective method of classification and regression analysis. Given a set of trained data with class labels assigned then SVM builds a model that can classify the test data into the respective class labels. SVM and decision tree is used to classify the learning skills of the autistic children. Combination of neural networks and fuzzy logic is used to classify the levels of autism in autistic children as tabulated in table 3.

Table 2: Behavioral Analysis

Sl No	Classifier	Accuracy
01	SVM	95-97%
02	J48	90-92%
03	Multilayer Perceptron	90-91%
04	IB1	88-90%

Table 3: Learning Skill Analysis

Sl No	Classifier	Accuracy
01	SVM	95-97%
02	J48	90-92%

Classification methods used are BVM, C-RT, C4.5, CS-CRT, C-SVC, ID3, and KNN, Linear discriminant, Multilayer perceptron, Naïve Bayes, Multinomial Logistic Regression, PLS-DA, PLS-LDA and Rnd Tree are compared based on M-CHAT R dataset which is a collection of behavioral data as tabulated in table 1. Among these classification methods, it is concluded that BVM, CVM and MLR yields the best results and most accurately classifies the autistic data as tabulated in table 4.

Table 4: M-Chat R Analysis

Sl No	Classifier	Accuracy
01	BVM	95.2-95%
02	C4.5	93.5-95%
03	C-SVC	95-97%
04	CVM	95.2-96%
05	K-NN	93.2-94%
06	LDA	93.6-95%
07	RND TREE	94.9-98%
08	MLR	95.2-97%

Associative rule mining classification is an efficient approach to classify the behavior of ASD. Classification based on association (CBA- apriori algorithm), classification based on multiple rules (CMAR- FP growth algorithm) and classification based on predictive association rules (CPAR). PNN (Probabilistic Neural Networks) classifier is deployed as a supervised pattern learning recognition model to classify the pre-verbal vocalization of the autistic children and it is proved that PNN provides higher accuracy than SVM classifier in classifying acoustic prosodic data as tabulated in table 5.

Table 5: Acoustic Analysis

Sl No	Classifier	Accuracy
01	PNN	97-98%
02	SVM	69-72%

Accuracy of a classifier determines the count of correctly classified dataset. Higher the accuracy means more number of correctly identified data points. Accuracy is defined as number of correctly classified samples divided by the total number of sample in the class.

CONCLUSION AND FUTURE WORK

The goal of the paper was to provide an empirical evaluation on features and classification methods involved in classifying the autistic children. From the study, it is evident that behavioral data like social relationship, emotional responsiveness, language and communication, behavioral patterns, cognitive components and sensory aspects are assessed to classify the autistic children. Empirical evaluation proves that SVM, J48, BVM and decision tree are the most commonly used classifiers are best suited to classify the autistic data and it provides high accuracy and low error rate. As future work we propose to build a model that combine high accuracy classifiers to classify autistic children and predict the level of autism in autistic children.

Acknowledgements

The autism data for the research work is provided by the National Institute for the Mentally Handicapped.

References

[1] Margaret H. Dunham, Data Mining: Introductory and Advanced Topics, 2nd ed., aPearson Education, 2006.

[2] Shika & Shika Khera, “ A Fuzzy Improved Neural Based Soft Computing Approach for Pest Disease Prediction,” International Journal of Information & Computation Technology, vol.4, pp. 1335-1341, 2014.

[3] Milan Kumari & Sunila Godara, “ Comparative Study of Data Mining Classification Methods in Cardiovascular Disesase Prediction,” International Journal of Computer Science and Technology, vol.2, pp. 304-308, 2011.

[4] Jyothi Soni, Ujma Ansari, Dipesh sharma & Sunitha Soni, “ Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” International Journal of Computer Applications, vol.17, pp. 43-48, 2011.

[5] Mohana E & Poonkuzhali S, “ Categorizing the Risk Level of Autistic Children using Data Mining Techniques,” International Journal of Advance Research In Science and Engineering, vol.04, pp. 223-230, 2015.

[6] Cheol-Hong Min & Ahmed H. Tewfik, “ Novel Pattern Detection in Children with Autism Spectrum Disorder using Iterative Subspace,” IEEE International Conference on Acoustics Speech and Signal Processing, pp. 2266-2269, 2010.

[7] M. S. Mythili & A. R. Mohamed Shanavas, “ A Novel Approach to Predict the Learning Skills of Autistic Children using SVM and Decision Tree,” International Journal of Computer Science and information Technologies, vol.05, pp. 7288-7291, 2014.

[8] Mark Hall, Eibe Frank, Geoffery Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, vol.11, pp. 10-18, 2009.

[9] M. S. Mythili & A. R. Mohamed Shanavas, “ A Study on Autism Spectrum Disorder using Classification Techniques,” International Journal of Soft Computing and Engineering, vol.04, pp. 88-91, 2014.

[10] M. S. Mythili & A. R. Mohamed Shanavas, “ An Improved Feature Selection (IFS) Algorithm for Detecting Autistic Children Learning Skills,” Bioscience Biotechnology Research Asia, vol.12, pp. 499-505, 2015.

[11] Siriwan Sunsirikul & Tiranee Achalakul, “ Associative Classification Mining in the Behavior Study of Autism Spectrum Disorder,” IEEE International Conference on Computer and Automation Engineering, vol.03, pp. 279-283, 2010.s

[12] Maithilee Kunda & Ashok K. Goel, “ How Thinking in Pictures Can Explain Many Characteristic Behaviors of Autism,” IEEE International Conference on Development and Learning, pp. 304-309, 2008.

[13] Joao F. Santos, Nirit Brosh, Tiago H. Falk, Lonnie Zwaigenbaum, Susan E. Bryson, Wendy Roberts, Isabel M. Smith, Peter Szatmari and Jessica A. Brian, “ Very Early Detection of Autism Spectrum Disorders Based on Acoustic Analysis of Pre-Verbal Vocalizations of 18-Month Old Toddlers,” IEEE International Conference on Acoustics Speech and Signal Processing, pp. 7567-7571, 2013.