CATEGORIZING THE RISK LEVEL OF AUTISTIC CHILDREN USING DATA MINING TECHNIQUES

Mohana E[1], Poonkuzhali.S[2]

[1]PG Scholar, [2]Professor, Department of Information Technology

Rajalakshmi Engineering College, chennai (India)

ABSTRACT

Autism spectrum disorders (ASD) are enclosure of several complex neurodevelopmental disorders characterized by impairments in communication skills and social skills with repetitive behaviors. It is widely recognized for many decades, yet there are no definitive or universally accepted diagnostic criteria. Studies indicate early intervention services, for young children with ASD significantly improve the children's prognosis and should begin as early as 18 months of age. The Modified Checklist for Autism in toddlers, better known as the M-CHAT, is a free screening tool that can be administered to children between 16 and 30 months-of-age. Hence by using the M-CHAT, this paper focuses on finding the best classifier with reduced features for predicting the risk level of autism. The four feature selection algorithms such as Fisher filtering, ReliefF, Runs filtering and Stepdisc are used to filter relevant feature from the dataset, and then several classification algorithms are applied on this reduced features. Finally performance evaluation is done on all the classifier results.

Keywords: Accuracy, Autism Spectrum Disorder, Classification, M-CHAT Screening Tool, Feature Selection

I. INTRODUCTION

Autism is a neurodevelopmental disorder. The autistic children are characterized by lack of social interaction, communication and behavior. ASD is a spectrum disorder as its impact on every child varies. ASDs affect one out of every 68 children in the U.S. They occur more often among boys than girls. The causes of autism has different source. It cannot be confined to a particular factor. For example it may be due to medical reason or genetically induced. It has a unique feature and severity in each child. It varies from level to level. The children with autism are attached to things; possess repetitive behavior like arranging things in a row. Even though they lack these factors, some children master in a particular area. The children may seem to appear normally, but still they might have some risk level of autism which is very difficult to predict at the early stage of autism. M-CHAT is the Modified Checklist for Autism in Toddlers. It is a screening tool for diagnosing autism in age between 16 to 30 months. In case of timely diagnosis of autism, the early intervention program can be started. Data mining is the process of analyzing through large amounts of data for useful information. It uses artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns, and relationships, which might otherwise have remained undetected. It is the step of the knowledge discovery in databases (KDD) process concerned with the algorithmic means by which patterns or structures are enumerated from the data. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. The use of computers with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result data

mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets [13].Feature selection algorithm used to find the subset of input variables by eliminating the features with less or no predicting information. It significantly improves the accuracy of the future classifier models formed by different classification algorithms.

And then several classification algorithms such as BVM, C4.5, C-RT, CS-MC4, CS-CRT, C-SVC, CVM, ID3, K-NN, Rnd Tree etc., is applied on the reduced datasets produced by feature selection algorithms. Finally performance evaluation is done to find a best classifier. So that with minimum attributes toddler children's autism level can be found.

II. RELATED WORK

JyotiSoni et.al [1] compared predictive data miningtechniques such as Decision tree, Naïve Bayes, K-NN, and classification based on clustering for analyzing the heart disease dataset. The classified data is evaluated using 10 fold cross validation and the results are compared. Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictivemethods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Carloz Ordonez et al., [4] applied association rule mining on heart disease data. Search constraints and test data validation reduces the number of association rules with high predictive accuracy.In the survey of [5] the author proposed the minimal subset of attributes for predicting heart disease. In future this work can be expanded and enhanced for the automation of heart disease prediction. Real data should be collected from health care organizations and agencies are taken to compare the optimum accuracy with all data mining technique.

G. Parthiban et al., [6] applied Naïve Bayes classification through WEKA ("Waikato Environment for Knowledge Analysis") tool to diagnose heart disease of diabetic patient. 10 folds cross validation is used to avoid any bias in the process and improve efficiency of the process.AbdelghaniBellaachi et al., [7] analysed breast cancer data with three data mining techniques such as Naïve Bayes, Back-Propagated Neural Network and C4.5. In that, C4.5 produces more accuracy of about 86.7%.

GeethaRamani et al., [8], applied feature relevance algorithm and then different classification algorithm on the selected features. Error rate and accuracy of the different classification tool is calculated using Tanagra. In that, Rnd tree produced 100% accuracy.

S. Poonkuzhali et al., [10], taken TP53 germline database for classification. First feature construction done by converting all input to disc to cont function. Then different filtering algorithm is applied to reduce the number of features. Different classification algorithm produced on the reduced dataset. Finally performance evaluation is done. Rnd tree produces 100% accuracy using ReliefF filtering.

Christina Schweikert et al., [14], applied Combinatorial Fusion Analysis(CFA) and Association Rule Mining(ARM) to autism, lead, and mercury data. CFA revealed that autism prevalence has strong correlation with rank combination of mercury and lead than individual. ARM discovered a trend where increase in mercury strongly related to increase in autism prevalence.

Gondy Leroy et al.,[15], autism children are videotaped before, during and after therapy applied to them. Four conditions are taken to monitor child’s inappropriate and appropriate behaviour like when alone, accompanied

with parent, stranger and therapist. Decision tree and rule mining algorithm are applied on the above noted data to find out their level of behaviour.

M.S. Mythili et al., [16], taken autism children's learning skills dataset and the decision tree classifier (J48), Normalized PolyKernel based classifier (SVM) were enforced in weka tool. Visual image of decision trees are formed and accuracy of both the algorithms are calculated. In that SVM having high accuracy(95%) , correctly classified the dataset.

M.S. Mythili et al., [17], analysed the dataset of autism containing three attributes such as language, social and behaviour. Values of these attributes are represented with three discrete values like mild, moderate and heavy and the level of autism is detected from these attributes. Neural Network, Support Vector Machine and Fuzzy logic algorithms are used to produce classification model.

III. ARCHITECTURAL DESIGN

The architectural design of the proposed system is given in Fig 1 and each block is explained in the following sections.

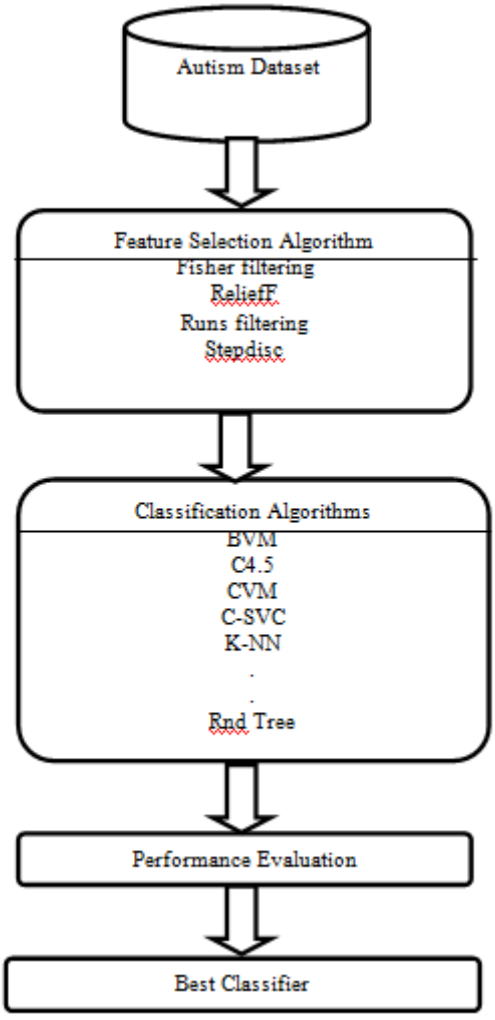

Fig. 1 Artichitecture Design of the Proposed System

3.1Autism Dataset

The Autism Dataset is formed with M-CHAT-R tool (Modified Checklist for Autism in Toddlers Revised tool) which is valid only for children in the age of 16 to 30 months. This tool contains 20 Yes/No questions, and

produce risk number of the child as output. The autism dataset contains 20input attributes which represent yes/no answers for 20 questions in a tool, one class attributewhich denotes the corresponding risk value which range from 0 to 20 and 438 instances. The description of the attributes of this autism dataset is given in Table I.

Table. I Attribute of Autism Dataset

Attribute No.	Attribute Description (Questions)
1	If you point at something across the room, does your child look at it?
2	Have you ever wondered if your child might be deaf?
3	Does your child play pretend or make-believe?
4	Does your child like climbing on things
5	Does your child make unusual finger movements near his or her eyes?
6	Does your child point with one finger to ask for something or to get help?
7	Does your child point with one finger to show you something interesting?
8	Is your child interested in other children?
9	Does your child show you things by bringing them to you or holding them up for you to see – not to get help, but just to share?
10	Does your child respond when you call his or her name?
11	When you smile at your child, does he or she smile back at you?
12	Does your child get upset by everyday noises?
13	Does your child walk?
14	Does your child look you in the eye when you are talking to him or her, playing with him or her, or dressing him or her?
15	Does your child try to copy what you do?
16	If you turn your head to look at something, does your child look around to see what you are looking at?
17	Does your child try to get you to watch him or her?
18	Does your child understand when you tell him or her to do something?
19	If something new happens, does your child look at your face to see how you feel about it?
20	Does your child like movement activities?
21	Risk level of child

3.2 Feature Selection Algorithm

The autism dataset contains 21 attributes of which 20 input attributes are discrete and the target attribute risk is continous attribute (0-20) . In order to apply filtering algorithm, target attribute has to be transformed into discrete attribute.Then the filtering algorithms such as Fisher Filtering, ReliefF, Runs Filtering and Stepwise Discriminant Analysis are applied to the feature constructed dataset and the results are given in Table II.

Table. II Feature Selection

S.No	Feature selection algorithm	No. of attributes Before Filtering	No. of attributes After Filtering	Attribute No. After Filtering
1	Fisher Filtering	20	20	All attributes
2	ReliefF	20	9	17,16,18,10,15,9,8, 14,11
3	Runs Filtering	20	13	3,5,7,9,11,12,14,15,16, 17,18,19,20
4	Stepdisc	20	20	All attributes

3.3 Classification Algorithm

Classification algorithms such as BVM, C4.5, C-RT, CS-CRT, CS-MC4, C-SVC, CVM, ID3, K-NN, Linear Discriminant Analysis, Multilayer perceptron, Naïve bayes continuous, Multinomial Logistic Regression, PLS-DA, PLS-LDA and Rnd Tree are applied to each of the above filtering algorithms and the results are given in Table III.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

Different classification algorithms are compared in terms of error rate, accuracy, recall and precision. Each are discussed below.

4.1. Error rate

Error rate of a classifier was defined as the percentage of the dataset incorrectly classified by the method. It is theprobability of misclassification of a classifier. Error rate of different classifier before filtering and after filtering is represented in table III.

$$\textit{Error rate} = \frac{\textit{No.of incorrectly classified samples}}{\textit{Total no of Sample in the class}}$$

Table. III Error Rate of Different Classifiers

Classification Algorithm	Error Rate Before Filtering	Error rate after filtering	
		ReliefF	Runs Filtering
BVM	**0.0160**	**0.1073**	**0.0479**
C4.5	**0.0160**	**0.1187**	**0.0685**
C-RT	0.1073	0.1370	0.1187
CS-CRT	0.1073	0.1370	0.1187
CS-MC4	0.1073	0.1256	0.0868
C-SVC	**0.0183**	**0.1096**	**0.0502**
CVM	**0.0114**	**0.1073**	**0.0479**
ID3	0.2237	0.2237	0.2237
K-NN	**0.0731**	**0.1370**	**0.0685**
LDA	**0.0388**	**0.1142**	**0.0639**
MP	0.0548	0.1096	0.0548

MLR	**0.9795**	**0.1164**	**0.0479**
NBC	0.0457	0.1256	0.0982
PLS-DA	0.0662	0.1233	0.1119
PLS-LDA	0.0502	0.1233	0.0776
Rnd- tree	**0.0639**	**0.1073**	**0.0502**

4.2 Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method. The accuracy of all the classifiers used for classifying this autism dataset are represented in Table IV.

$$Accuracy = \frac{No\ of\ correctly\ Classified\ Samples}{Total\ no\ of\ Sample\ in\ the\ class}$$

Table. IV Accuracy of Classifiers

Classification Algorithm	Accuracy(%)
BVM	**95.2**
C4.5	93.2
C-SVC	94.97
CVM	**95.2**
K-NN	93.2
LDA	93.6
Rnd Tree	94.97
MLR	**95.2**

4.3 Recall

Recall of the classifier was defined as the percentage of errors correctly predicted out of all the errors that actually occurred. The recall of the best classifiers for three levels of autism is represented in Table V and graphically represented in fig. 2.

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative}$$

4.4 Precision

Precision of the classifier was defined as the percentage of the actual errors among all the encounters that were classified as errors. Precision of BVM, CVM and MLR classifiers for three levels of autism is represented in Table V.

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive}$$

The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to classifier's expectation.

Table. V Precision and Recall of Classifiers

Classification algorithm	Class	Precision	Recall
BVM	Low	0.5714	0.0769
	Medium	0.9812	0.0457
	High	0.9388	0.0515
CVM	Low	0.5714	0.0769
	Medium	0.9812	0.0457
	High	0.9388	0.0515
MLR	Low	0.5714	0.0769
	Medium	0.9781	0.0429
	High	0.9490	0.0606

Recall for three classifiers BVM, CVM and MLR having high accuracy (95.21%) is represented in fig. 2.

Fig.2 Recall of BVM,CVM and MLR Classifiers

V. CONCLUSION

In this paper, autism affected children of age 16-30 months dataset is taken. The dataset is pre-processed and taken for feature selection. Feature Selection Algorithm such as Fisher Filtering, ReliefF, Runs Filtering and Stepdisc is applied. In that Fisher Filtering and Stepdisc does not filter any features. So Runs filtering and ReliefF is chosen. Then different classification algorithm is applied on the subset produced by both feature selection algorithm. Finally, performance evaluation is done on the results such as error rate, recall and accuracy. This paper helps in finding the best classifier for autism dataset through feature relevance analysis and classification algorithm. Among different classification algorithm applied , algorithms such as BVM, CVM AND MLR produced high accuracy of 95.21% using Runs Filtering and it also accurately classified the test dataset.

VI. ACKNOWLEDGEMENT

This research work is the part of the funded project titled Interactive Teaching Aid for Autistic Children (ITAAC),D.O. No.: SEED/TIDE/034/2013 under TIDE programme of SEED division from Department of Science & Technology, Ministry of Science and Technology, New Delhi.

REFERENCES

[1]. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[2] "Data mining: Introductory and Advanced Topics" Margaret H. Dunham

[3]. JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June 2011.

[4]. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[5] M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370- 5376.

[6] G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method" .

[7] BellaachiaAbdelghani and ErhanGuven, "Predicting Breast Cancer Survivability using Data Mining Techniques,"Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[8] R. GeethaRamani, G. Sivagami, Parkinson Disease Classification using Data Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 32– No.9, October 2011.

[9]ShomonaGracia Jacob, R.GeethaRamani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[10] S. Poonkuzhali, R. GeethaRamani, R. Kishore Kumar,Efficient Classifier for TP53 Mutants using Feature Relevance Analysis, in International Multiconference of Engineers and computer scienctists, Vol 1, 2012.

[11] Tanagra-Data Mining tutorials http://data-mining- tutorials.blogspot.com

[12] Arun K Pujari, Data Mining Techniques, University Press 2001

[13] ShwetaKharya, "International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", Vol.2, No.2, April 2012.

[14] Christina Schweikert, Yanjun Li, David Dayya, David Yens, Martin Torrents, D. Frank Hsu," Analysis of Autism Prevalence and Neurotoxins Using Combinatorial Fusion and Association Rule Mining", in Ninth IEEE International Conference on Bioinformatics and Bioengineering, 2009.

[15] Gondy Leroy, Annika Irmscher, Marjorie H. Charlop-Christy,"Data Mining Techniques to Study Therapy Success with Autistic Children".

[16] M.S. Mythili, A.R.MohamedShanavas," A Novel Approach to Predict the Learning Skills of Autistic Children using SVM and Decision Tree", in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014.

[17] M.S.Mythili, A. R. Mohamed Shanavas," A Study on Autism Spectrum Disorders using Classification Techniques", in International Journal of Soft Computing and Engineering (IJSCE) ISSN:2231-2307,Volume-4Issue-5,November2014.