Gaussian Mixture Model Based Hierarchical Clustering in Prediction of Autism Spectrum Disorder

D. Umanandhini, G. Kalpana

Abstract: ***Autism is one of the most complex and divergent class disorders which accompany various lacking in symptoms needed for classification, societal interaction, abridged verbal communication, and monotonous behavior. Timely and proper diagnosis of Autism Spectrum Disorder can ensure the offering of medical treatment and guidance to get cure. In this paper, Gaussian Mixture Model based Hierarchical Clustering is proposed for efficiently predicting the Autism Spectrum Disorder. Also, Flexible splitting concept was proposed for hierarchical clustering in order to increase the quality of guessing and classification accuracy. The proposed algorithm is validated to check the performance against the existing method. The results shows that the proposed algorithm outperforms the existing algorithm in terms of classification accuracy.***

Keywords: Autism, Autism Spectrum Disorder, Gaussian, Classification, Clustering.

I. INTRODUCTION:

Anxiety is a condition of getting afraid. It is linked with increased nervousness, manifestation, and pessimistic emotional force or attraction. Automated manner of detecting the anxiety has established the awareness in different fields like monitoring of health, assessment of workload, controlment of security and accessibility, smart transportation, and interaction between human and computer. Further instantaneously, detecting anxiety in a automated manner is implied with the aim to measure the treatment programs that are related to anxiety. Managing the symptoms of physiological behavior is default in traditional treatments, but it fully depends on ability of individual in self recognizing the physiological anxiety symptoms, where it is not fully suited for population that gets increased in clinics due to lack in psychological disability, communication,. self analyzing, and awareness about emotion. One of the best example of such kind is population is the individuals who have ASD (autism spectrum disorder). ASD has become common and complicated neuro-developmental disorderliness that are characterized through communication in society, and occurrence of restricted and repeated behaviors Anxiety is considered as the root cause for the mournfulness in ASD. CDC (Centers for Disease Control and Prevention) has demonstrated that ASD tends to increase in the upcoming years. When comparing ASD for male and female, it is noted that males are mostly affected than females. ASD includes disadvantages like problems in sleep and irritability, and advantages like strong visual, auditory learning and excelling in some specific subjects. The reason for autism is still unknown and many researchers are still in progress towards this. ASD problem is categorized as mild, serious, or somewhere between the both. ASD is not detectable till there exist a change in the current situation or environment, for example – When a kid start going to school or nursery. The methodologies related to automated detection of anxiety or anxiety related medical related problems are not evaluated in the treatments till date. Currently, ASD is increasing among children. It's necessary to detect it in early stage, where the failing of ASD detection may lead to issues in children health condition mentally. Current algorithms don't have much impact in giving the accuracy towards the classification for detection of ASD. Kalman filtering framework offers competitive accuracy of 85% but it is considerably lower, which consumes more time for classifying the data. Traditional classification algorithms have better performance in the dataset which have low number of records, but those algorithms have worst performance when the dataset size is large.

This research work utilizes data mining to understand data distributions at all levels and its relationship to predict ASD. The main objective of this research work is to propose a unsupervised learning based classifier which can effectively detect ASD from the given dataset in short duration. This research work aims to overcome the barriers in providing the treatment for ASD at the early stage.

The remainder of this paper is organized as follows. Section 2 provides the summary of the related works as literature review. Section 3 discusses the proposed classifier towards the detection of ASD. Section 4 provides the information about dataset. Section 5 illustrates the chosen performance metrics along with the tool used. Section 6 confers the results. Finally, Section 7 concludes the paper with future dimensions.

Manuscript published on 30 September 2019
* Correspondence Author

DUmanandhini*, Phd Research Scholar, Department Of Computer Science, Sri Ramkrishna College Of Arts And Science For Women Coibatore.

Dr. G. Kalpana, Associate Professor, Department Of Computer Science, Sri Ramakrisna College Of Arts And Science For Women Combatore.

II. LITERATURE REVIEW:

Accurate Detection Method [1] was proposed to detect the ASD with the help of neural classifier. Enhanced Q-Gaussian classifier and necessary learning algorithm was proposed to find the relationship of features and class labels. Low level of classification accuracy shows the algorithm is not sufficient in detecting the ASD. Multilayered Fuzzy based Cognitive Map [2] was proposed to predict and classify the autism patients. It involves standard assessment and diagnosis. It was based on soft computing concept, but the low level of accuracy says that the algorithm cannot be recommended for detection of ASD. Convolutional Neural Network based classifier [3] was proposed to detect the abnormality in brain related to ASD. The classifier is based on deep learning concept which finds the rate of optimum learning and it fine tunes the trained data for classification. Increased value of false positive in result shows that the classifier is cannot provide better classification. Gene based Feature Selection [4] was proposed to detect ASD with important genes and its sequences. Initially a deep investigation was done for the autism dataset, and few feature selection methods were applied. The result with low f-measure shows that the algorithm cannot be applied for ASD detection. Machine Learning Classifier [5] was proposed to classify ASD based on gender by utilizing the three different regions of brain. It correlate the scores of f-measure with behavior functioning. Poor classification accuracy indicated the poor detection of ASD.

Deep Learning based Automated Prediction [6] was proposed to predict the ASD with movements. Neural network concept was applied to study the features in raw data to model the temporal patterns. Classification results shows that the neural network concepts are recommended to predict ASD due to low accuracy. Bayesian based Hierarchy Model [7] was proposed to overcome the issues of prediction of ASD. It utilizes the laplacian distribution concept to minimize the errors, uniformity distribution for the parameters of regression. The results shows that the method is not effective in classifying the ASD. A kalman filtering framework [8] based on the Kalman filtering theory was proposed for detection of physiological arousal based on cardiac activity, it's a an unsupervised and real-time arousal detection algorithm which lacks in taking too much time for classifying the data for ASD. Fuzzy Logic based Ensemble Learning [9] was proposed by inheriting the machine learning algorithms namely adaboosting and bagging. In order to enhance the performance, the algorithm was trained and applied on the dataset. The results with highly increased false positives shows that the algorithm need to fine tuned. A collaborative virtual environment [10] based social interaction platform for ASD intervention was proposed, where it leads a way to the create low-cost intervention environment. The main drawback of this system was the force give to the ASD children to use the application in a mandatory manner resulting the children in a uncategorized way

III. GAUSSIAN MIXTURE MODEL BASED FLEXIBLE HIERARCHICAL CLUSTERING

Initial and primary step of this research work namely Gaussian based Combined Discovery Technique (GCDT) is hierarchical clustering, where the algorithm divides the dataset into 2 in a recursive manner robustly.

3.1. Dimension reduction

The most intensified method to seek a optimum method is to divide the dataset in more than one-dimension. In order to avoid this risk, this research work extends the samples available in dataset into more than one dimension, by seeking the optimum point to divide the single dimension dataset and extending it to unique higher dimension. Estimating the optimum edge is improbable to support the major varying direction, this research work extends the weighted samples based on principal components ($\aleph^{i+1}, \aleph^{i+2}, \dots, \aleph^{i+C}$) available. Once after finishing the extending process, components are made to get index in the decreasing order. For every $c \in (1,2,\dots,C)$, the dimension has its samples based on weighted scalar $(\aleph^{S+c} y^{[n]}, x^{[n]})^{N*n+1}$and it is represented as $(w^{n+c}, x^{n})^{N*n+1}$.

3.2. Estimation of Core Density

Core density estimation makes the algorithm to take a decision on dividing single dimension dataset In this, it is necessary study the statistics based distribution of single dimension dataset through its core density, mathematically expressed as:

$$\bar{\sigma}^{t}[w] = \int_{n=1}^{N} x^{[n]} \sigma^{t}(w, w^{n+c}) \tag{1}$$

Core method is indicated as σ^{t} with a continuous parameter t , Estimation of basic distribution is indicated as $\bar{\sigma}^{t}$ for the given t. With the intention to study the Gaussian based combined model (GCC), this research work sets Gaussian core value as σ^{t}, and therefore interpretation can be made on t as per the SD (standard deviation) of Gaussian core. In the default way, t can be calculated based on the protocol of approximate mean square error rule:

$$t = \emptyset \,/\, (4 * 3N)^{1/5} \tag{2}$$

Standard deviation of basic distribution is indicated as $\emptyset$ in Eq. (2) and its substitution is made robustly and mathematically expressed as

$$\bar{\emptyset} = median\ (\ll w^{i+c} + median(w^{j+c} \tag{3}$$

3.3. Flexible dividing

By considering the univariate distribution in Eq. (1), this research work divides the deliberated dataset into 2 in number, only thing is to be fixed is the points of boundary in the appropriate direction.

Multiple previous selects the mean as the dividing point, which is not a fine method for discovering the dividing point using GCC. It is because the mean does not provide guarantee to cover the efficient components of GCC, even if proceeded it will discard the efficient components of GCC. In order to overcome this issue, this research work divides the dataset by utilizing an alternate method which makes use of samples of distribution $\bar{\sigma}^t$ in Eq. (1). When the dividing point's probability density gets decreased, automatically there exist a deviation in the center point. This research work considers the point that have minimum density value lies in the border of components of Gaussian, where it divides the dataset with most deep decrease.

In order to find most-deep local decrease, this research work evaluates $\bar{\sigma}^t$ with single dimension grid with H points. Suppose, most-deep local decrease is not found means, it's necessary to assume the whole cluster as a single Gaussian component and it is recommended to never proceed further dividing. Else, it's necessary to check for most-deep local decrease for considering next dividing point. The deliberated values which are minimum than the dividing points are formed as a single cluster and treated a child dataset, and the values are maximum than the dividing points are formed as a another cluster and treated as a another different child dataset.

3.4. Flexible hierarchical tree

By following the flexible dividing approach, this research work builds a hierarchical tree in order to denote the dataset in a ordered approach. This research work considers the entire dataset as a source for the tree that is to be built. The tree is estimated to be built in a recursive manner. Threshold value is set to enhance or stop the growth of tree. Calculation is performed to check samples available at the node, if the samples are less than the threshold value then the node is considered as leaf node and further division process is not made. Else, samples are deliberated with principals component from beginning to end in order to achieve the flexible dividing. In case, a capable neighboring decrease is established for the principals component, then the current node is divided by not considering the balance principals component, else the current node is assumed to be leaf and division process is proceeded further. This research achieves the GCC from tree leaves, where each leaf corresponds to a principals component and average weight of the samples in the leaf will correspond to the Gaussian component weight.

3.5. Computational complexity

In every recursive stage of flexible hierarchical clustering, this research work performs the principal component analysis based on samples weight with the cost $P(C^{i+3})$ and executes flexible dividing not more than C number of times. In this proposed approach, core density estimation intakes $P[GM]$, and finds the dividing point $P[H]$. Hence, recursive step involves $P(C^{i+1} + DGM + DG) = P(DGM)$, where it assumes $|C + H|$ and $|C + N|$. In comparison, a KD tree, which divides at the mean, has a lower complexity $O(DM)$ than flexible hierarchical clustering in each recursive step, but flexible hierarchical clustering incurs less computation in total, for the following reasons. The tree that is established by flexible hierarchical approach repeatedly divides the cluster till it is possible to divide. Hence the division process is stopped if it finds only a single component in it. Most times, a tree that is built using core density estimation may have unwanted nodes, where flexible hierarchical clustering prevent the tree that gets involve with unwanted nodes. In traditional clustering tree, it's necessary to seek the whole tree for optimum Gaussian components, but the proposed clustering method precisely gets the components in its leaves itself with zero searching step.

IV. AUTISM DATASET

The information for building the dataset consumed around six months of time. The dataset contains 1499 patients records. For the confidentiality reasons, the name of the patient is not obtained. Out of 1499, 998 children are having the possibility of getting ASD and remaining don't have.

V. PERFORMANCE METRICS:

Table 1. Performance Metrics Description and Formulae

Metrics	Description	Formulae
Sensitivity	Ratio of real positives that are correctly identified.	$\frac{True\ Positive}{(True\ Positive + False\ Negative)}$
Specificity	Proportion of actual negatives that are correctly identified.	$\frac{True\ Negative}{(False\ Positive + True\ Negative)}$
Disease Prevalence	Proportion of disease found in the total populace of the dataset.	$\frac{(True\ Positive + False\ Negative)}{\left(\begin{matrix}True\ Positive + False\ Negative + \\ False\ Positive + True\ Negative\end{matrix}\right)}$
Positive Likelihood Ratio	Proportion of ratio between the probabilities of a positive test result given the presence of the disease and positive test result given the absence of the disease.	$\frac{Sensitivity}{(1 - Specificity)}$
Negative Likelihood Ratio	Proportion of ratio between the probabilities of a negative test result given the presence of the disease and negative test result given the absence of the disease.	$\frac{(1 - Sensitivity)}{Specificity}$
Positive Predictive Value	Proportion of probability of disease that is present when the test is positive.	$\frac{True\ Positive}{(True\ Positive + False\ Positive)}$
Negative Predictive Value	Proportion of probability of disease that is not present when the test is negative.	$\frac{True\ Negative}{(False\ Negative + True\ Negative)}$
Accuracy	Proportion of true results (both true positives and true negatives) among the total number of cases examined.	$\frac{(True\ Positive + True\ Negative)}{\left(\begin{matrix}(True\ Positive + False\ Negative + \\ False\ Positive + True\ Negative)\end{matrix}\right)}$

VI. RESULTS AND DISCUSSION

Table 2. Evaluation Results of Algorithms

Metrics	Kalman Filter [8]	GMM [Proposed]
True Positive	957	995
True Negative	469	484
False Positive	39	14
False Negative	34	6
Positive Predictive Value	96.08	98.61
Negative Predictive Value	93.24	98.78
Sensitivity	96.57	99.4
Specificity	92.32	97.19
Positive Likelihood Ratio	12.58	35.36
Negative Likelihood Ratio	0.04	0.01
Disease Prevalence	66.11	66.78
Accuracy	95.13	98.67

In Fig. 1, TP, TN, FP, and FN are plotted in x-axis, where the y-axis is plotted with number of records. Fig. 1 attempts to make comparison of proposed algorithm with kalman filter [8]. The result shows that the proposed algorithm gives its best performance in terms of TP, TN, FP, and FN, where the kalman filter [8] have given more number of false positive and false negative. In Fig. 2, x-axis is plotted with positives and negatives, y-axis is plotted with percentage. Fig. 2 shows the prediction of total number positives and negatives in percentage. The proposed algorithm outperforms the kalman filter [8] in predicting the more number of positives and negatives. In Fig.3, sensitivity and specificity is plotted in x-axis and y-axis is plotted with percentages. Fig. 3 results shows that the proposed algorithm has given its best performance against kalman filter [8] in terms of sensitivity and specificity. Fig. 4 analyzes the likelihood ratio of both positive and negative, where the proposed algorithm has gives its best performance in positive likelihood ratio as well as in negative likelihood ratio.

In Fig. 5, x-axis is plotted with the metrics accuracy and disease prevalence, y-axis is plotted with percentages. Fig. 5 clearly demonstrate that the proposed algorithm have better classification accuracy than kalman filter [8] due to maintaining the tree structure for prediction, where klaman filter simply performs classification in one by one manner. Also the disease prevalence rate of proposed algorithm is better than kalman filter [8].

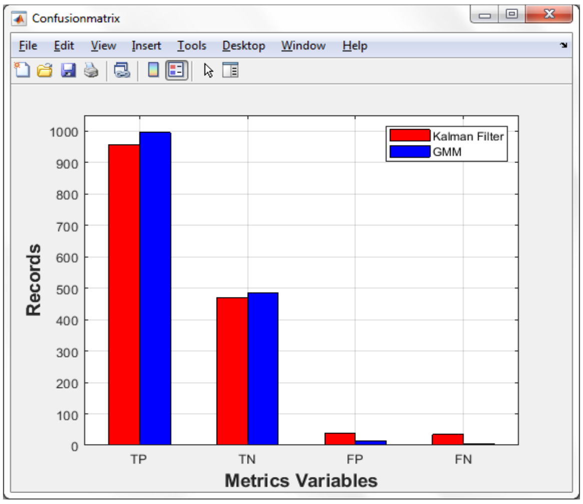

Fig 1. TP, TN, FP, FN Analysis

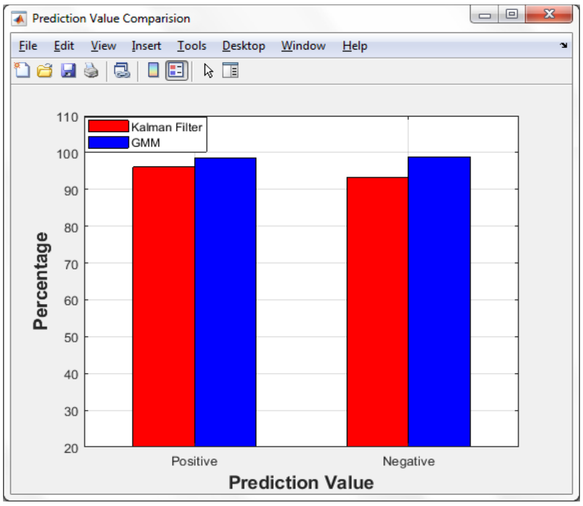

Fig 2. Predictive Analysis

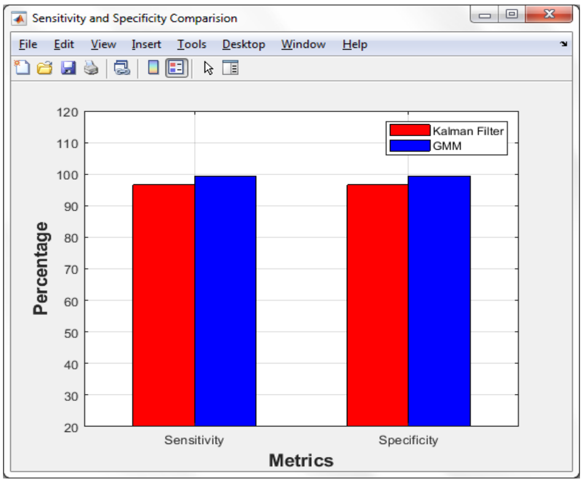

Fig 3. Sensitivity, Specificity Analysis

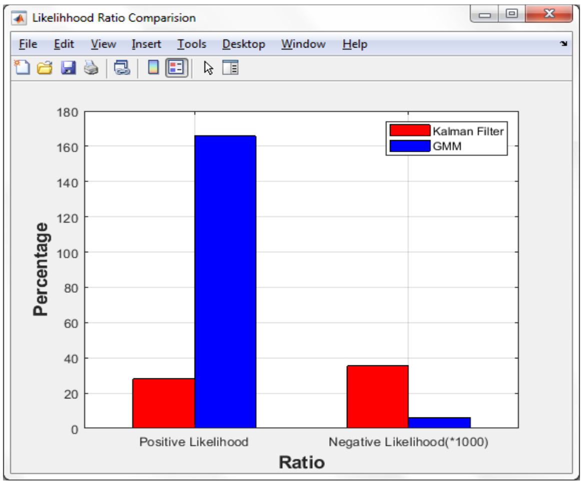

Fig 4. Likelihood Ratio Analysis

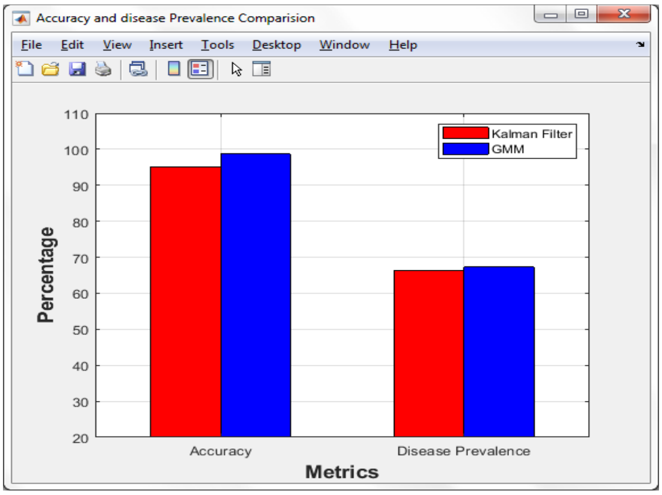

Fig 5. Accuracy and Disease Prevalence Analysis

VII. CONCLUSION

Currently ASD is getting increased all over the world. Timely diagnosis and treatment is very essential. Still now, ASD is detected only with the individuals behavior onlym where it takes much lengthy time to detect. This research work has attempted to predict the ASD by using patients medical cum records in terms of dataset holding 1499 records. In this research work Gaussian mixture model based hierarchical clustering was proposed to detect ASD. It splits the dataset into maximum number by using tree concept. The result shows that the proposed algorithm is able to detect ASD more efficiently than previous method with benchmark performance metrics.

REFERENCES:

1. Vigneshwaran. S., Suresh. S., Sundararajan. N., Mahanand. B. S. “Accurate detection of autism spectrum disorder from structural MRI using extended metacognitive radial basis function network,” Expert Systems with Applications, Volume 42, Issue 22, Pages 8775-8790, 2015.
2. Puerto. E., Aguilar. J., López. C., Chávez. D. “Using Multilayer Fuzzy Cognitive Maps to diagnose Autism Spectrum Disorder,” Applied Soft Computing, Volume 75, Pages 58-71, 2019.
3. Muhammed. T., Ulas. B. B., Özal. Y., U Rajendra. A. “Application of deep transfer learning for automated brain abnormality classification using MR images,” Cognitive Systems Research, Volume 54, Pages 176-188, 2019.
4. Tomasz. L., Stanislaw. O. “Data mining for feature selection in gene expression autism data,” Expert Systems with Applications, Volume 42, Issue 2, Pages 864-872, 2015.
5. Mark. P., Kelly. A. B., Alex. M. “Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards,” NeuroImage: Clinical, Volume 7, Pages 359-366, 2015.
6. Nastaran. M. R., Seyed. M. K., Calogero. Z., Twan. V. L., Giuseppe. J., Paola. V., Elena. M., Cesare. F. “Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders,” Signal Processing, Volume 144, Pages 180-191, 2018.
7. Zakariya. Y. A., Rahim. A., Haithem. T., Mohammad. A. “Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression,” Computers in Biology and Medicine, Volume 97, Pages 145-152, 2018.
8. Kushki. A., Khan. A., Brian. J., Anagnostou. E. "A Kalman Filtering Framework for Physiological Detection of Anxiety-Related Arousal in Children With Autism Spectrum Disorder," IEEE Transactions on Biomedical Engineering, Volume 62, Pages 990-1000, 2015.
9. Elpiniki. I. P., Arthi. K. “Fuzzy cognitive map ensemble learning paradigm to solve classification problems: Application to autism identification,” Applied Soft Computing, Volume 12, Issue 12, Pages 3798-3809, 2012.
10. Zhao. H., Swanson. A. R., Weitlauf. A. S., Warren. Z. E., Sarkar. N. "Hand-in-Hand: A Communication-Enhancement Collaborative Virtual Reality System for Promoting Social Interaction in Children With Autism Spectrum Disorders," IEEE Transactions on Human-Machine Systems, Volume 48, Pages 136-148, 2018.