International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Identification of Best Fit Learning Models Based on Calibration for Better Classification of Autism

Roopa B. S., R. Manjunatha Prasad

Abstract*: This paper is intended to exhibit the novel approach to improve the efficiency of the supervised learning models towards the accuracy of the predictions made to classify the autism from that of the normal subject. The state of the art is about 60-75% of Autism classification accuracy. The early prediction of autism plays a vital role as the rise of autism is alarming. The invasive way to analyze the problem at the earliest would render much support to the Autism Spectrum Disorder (ASD) community. In this work, various supervised learning models are first tested on 1101 subjects with 530 ASD subjects and 571 Normal subjects. The Datasets worked are collected from Autism Brain Imaging Data Exchange (ABIDE) repository. The performance measure is calibrated in terms of Brier score which is an accuracy measure of the predictions in probabilistic way. After assessing in probabilistic way, the statistically emphasized models are then evaluated for the same set of data to validate the prediction model efficiency with their statistical measures made and hence developing the confidence of the model selection for better classification based on probability calibration (CAL).

The performance evaluation of the model is tested with probability calibrated assessment and found that for given dataset the SVM and Logistic Regression provided better accuracy measure compared to other considered learning models. It is necessary to frame a hypothesis measure on the dataset before any model is deployed. This approach helps to identify the desired and validated supervised model for the given data samples.

Keywords: ASD, SVM, Calibration (CAL), Supervised Learning Models (SLM).

I. INTRODUCTION

In 1970s and 1980s, about one out of every 2,000 children had autism. Currently, about 1 in 59 children has been identified with autism spectrum disorder (ASD) according to estimates from CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network. From WebMD sources (an American corporation, human health related publisher online) also, Autism Cases are on the Rise. According to WHO fact sheet detailed in 2019, there are around 60 countries supporting ASD community with strong resolution made, so as to comprehend and coordinate the effects of managing ASD.

In this paper [1], they have tried to introduce the most popular and public consortium ABIDE ,which allows to access their repositories for research. The data base contributed from various renowned organizations with the motto of serving and supporting the ASD study further is an open opportunity to the further studies to make. The random SVM approach is applied in this paper [2] and achieved up to 93% on the optimal dataset and also arrived at extracting the best features to distinguish ASD from the normal subject. Also abnormal parts of the brain were identified. In [3],the greater change is addressed in white matter of the brain more than the grey matter. The author tried to explain the importance of diagnosing ASD at the earliest for addressing the issue and suppress the extremity of anomalies in future.

In [4] the physiological indicators involved are electrocardiogram, respiration, conductance and temperature of skin. With the help of these indicators three assessments were executed as arousal state, valence state and dominant state. Around 1386 pictures from IAPS and GAPED were used in consent with clinicians to assess the behavioral conduct in the above three states. Random Forest classifier in [5] is shown to have better accuracy of classification of Alzheimer's disease over the typical control with few overcoming on over fitting, ability to handle data which are nonlinear and also on multi-modality imaging of Alzheimer's disease. It is said in [6] -[7]that the trained clinicians find the apt therapies to treat Autism kids by diagnosing the relevance of it within two years. This clearly depicts us that early diagnose will help the community. This article finds that it is easy and best to diagnose from 2 years to 4 years and take corrective measures to below the intensity of the consequences. SVM classifier is shown to be one of the best classifier and AUC is the performance indicator used [8]. Here the test collections were done with 19 processes and 4 procedural reviews and then that balanced data sets are sampled and compared with other balanced data sheet. With this procedure they proved that an automatic and high quality classifier can be helpful to the experts. The Bayes and Naïve Bayes classifier [9]-[10] is showcased well; also the model efficiency is highlighted despite its inefficacy in dealing with independent features. In [11] the logistic regression model is used in combination with artificial neural network model for the biomedical classification.

Revised Manuscript Received on February 05, 2020.
* Correspondence Author
Roopa B. S.*, Department of ECE, DBIT, Bengaluru, India. Email: roopa.0303@gmail.com
Dr. R. Manjunatha Prasad, Department of ECE, DSATM, Bengaluru, India. Email: rmptiptur@yahoo.co.in

It showed that artificial neural network model is a generalized model of nonlinear nature of Logistic Regression. In [12] an good comparison is given between supervised models. Also the importance of calibrating the model for better efficiency is shown. Isotonic method is used for calibration to take care of monotonic disturbances, hence said to be one of the powerful calibration tool. The only disadvantage of isotonic is sometimes they over fit the model.

II. METHODOLOGY

The methodology is applied as shown in the Fig. 1. Firstly the dataset is imported and then the features are selected with preprocessing technique to make it suitable for applying to various supervised learning models.

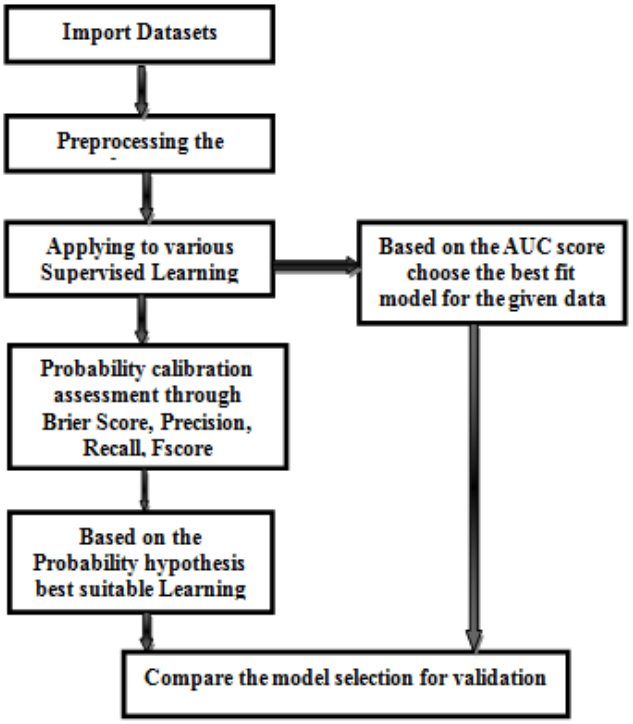

Fig. 1 System Architecture

A. Learning Models considered

The learning models considered for performance evaluation are 1) Decision Tree 2) Random Forest 3) SVM 4) K-NN 5) Naïve Bayes 6) Logistic Regression.

- *Decision Tree:* This model is good only when bias and variances are handled at high value else it would be a non-desirable model. The accuracy of distinguishing between two classes is done through the performance indicator called Gini impurity which indicates the purity of classifying correctly.

If a set is considered

$$S = \{(a_1, b_1), (a_2, b_2), \ldots (a_n, b_n)\} \quad (1)$$

Then the prediction is on the basis of maximum likelihood

$$P_k = \frac{|S_k|}{|S|} \quad (2)$$

Where

$$S_k = \{(a,b) \subset S \mid b = k\} \quad (3)$$

If a set is a combination of various subset, then

$$S = S_1 \cup S_2 \cup S_3 \ldots S_K \quad (4)$$

The Gini impurity measure is then given by

$$G(S) = \sum_{k=1}^{K} P_k (1 - P_k) \quad (5)$$

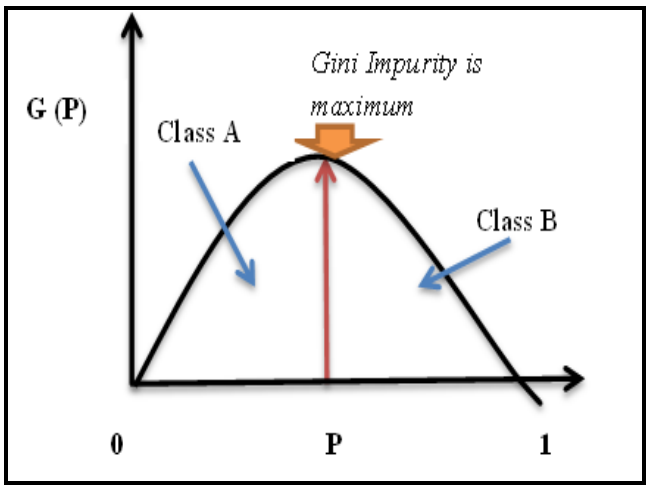

Fig. 2 Gini function curve (*purity measure*)

- *Random Forest:* Random Forest classifier in [5] is shown to have better accuracy of classification of Alzheimer's disease over the typical control with few overcoming on over fitting, ability to handle data which are nonlinear and also on multi-modality imaging of Alzheimer's disease. The RF has provided promising result as an ensemble learning model. These classifiers are statistical base models

Gini index for Random Forest is given by:

$$G(s) = 1 - \sum_{k=1}^{2} P_k^2 \quad (6)$$

- *SVM Classifier:* It resulted with better AUC for the high quality article classifier with constraints on inclusion and exclusion of certain task. The optimal hyper plane is calculated to distinguish into two classes by proper assessing. The decision boundary is given by

$$\min_{\theta} c \sum_{i=1}^{m} [y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_o(\theta^T x^{(i)})] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2 \quad (7)$$

For larger c value, the bias is lower and variance is high and for smaller c value, the bias is higher and variance is low in (7).

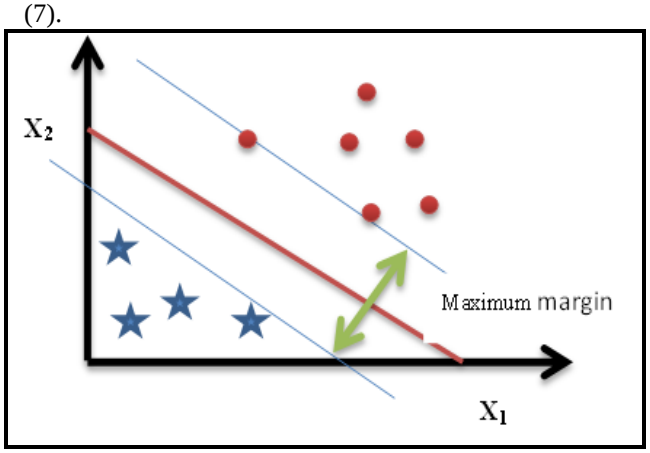

Fig. 3 setting optimal hyper plane

- *K-NN:* It's a supervised learning model which does not assume the data points underlying in the database. It checks for feature similarities to find the new set of data points.

$$Dist[(x1, y1),(x2, y2)] = \sqrt{(x1-x2)^2 + (y1-y2)^2} \quad (8)$$

The Euclidean distance is calculated using (8) to every time re-calculate and update the new centroid and hence the new data points to arrive at the final clusters

- *Naïve Bayes:* Naïve Bayes model is a statistical model oriented on Bayes theorem. This model handles the features for classification which are independent to each other but in real application the features are mostly dependent to each other hence this model is not appropriate for the applications. The Bayes rule is given in (9).

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)} \quad (9)$$

Where, *P(c/x) is posterior prob of target class Given predictors*
P(c) is prior prob of class
P(x) is prior prob of predictors
P(x/c) is likelihood which is the Prob of Predictor given class

$$\frac{P(L1/\ features)}{P(L2/\ features)} \equiv \frac{P(\ features/L1)P(L1)}{P(\ features/L2)P(L2)} \quad (10)$$

The probability representation shown in (10) is Naïve Bayes formula when two classes (L1 & L2) are considered for distinguishing the task from the normal.

B. Probability Calibration assessment

Firstly the predictive models are framed to predict for the given database and then the probability model can be calibrated.

The performance indicators used to indicate probability measures are easy in Logistic model.

For any other models, as they don't produce predictions based on the probability and hence approximated. Hence the one of the popular way of achieving calibration probability is to use Isotonic.

Some of the performance indicators used is Brier Score, Precision, Recall and F1-score.

Table-I: Confusion Matrix [14]

CONFUSION MATRIX (*As Per Scikit And Tensor flow Tools*)		
PREDICTED / **TRUE**	**0(*Negative*)**	**1(*Positive*)**
0(*Negative*)	**TN(0,0)**	**FP(0,1)**
1(*Positive*)	**FN(1,0)**	**TP(1,1)**

$$BrierScore = \frac{1}{N}\sum_{t=1}^{n}(f_t - O_t)^2$$

(9)

Where, f_t is the probability that was predicted
O_t is the actual outcome of the event at t
N is the number of predicting instances

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

Table-II: Probability Calibration of Logistic Model with Naïve Bayes Model

	BRIER SCORE	PRECISION	RECALL	F-SCORE
Logistic	0.002	1.000	1.000	1.000
Naïve Bayes	0.006	1.000	0.987	0.993
NB+Isotonic	0.006	1.000	0.987	0.993
NB+Sigmoid	0.006	1.000	0.987	0.993

Table-III: Probability Calibration of Logistic Model with SVM Model

	BRIER SCORE	PRECISION	RECALL	F-SCORE
Logistic	0.002	1.000	1.000	1.000
SVM	0.463	1.000	0.987	0.993
SVM+Isotonic	0.007	1.000	0.987	0.993
SVM+Sigmoid	0.218	1.000	0.347	0.515

F-score is given by

$$F-score = \frac{2(precision)(recall)}{precision + recall} \quad (13)$$

F- Score is harmonic mean relation of precision and recall.

As shown in the table, the performance indicators of Logistic Model which is a statistical model is taken as reference and compared with the other learning models like Naïve Bayes and SVM models. The smaller the Brier score better is the efficiency of the model.

The Calibration graph of logistic model in comparison with Naïve Bayes and SVM is shown in Fig. 4 and Fig. 5

The Calibration graph of logistic model in comparison with Naïve Bayes and SVM is shown in Fig. 4 and Fig. 5. The calibration plot depicts the nature of model with and without calibration referred to a logistic model. The SVM without calibration gives very high brier score, which means the mean square error is more.

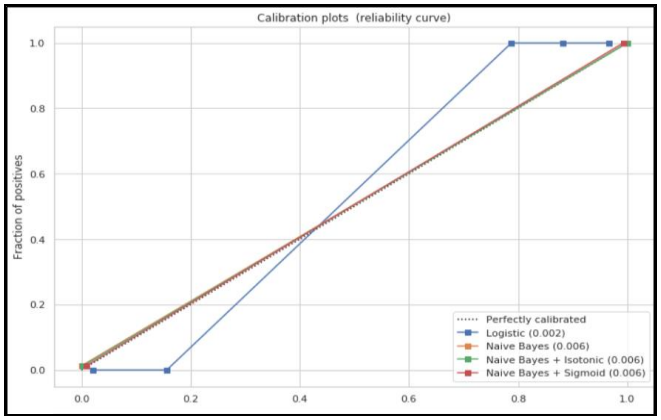

Fig. 4 Calibration Plot of Logistic in comparison with Naïve Bayes

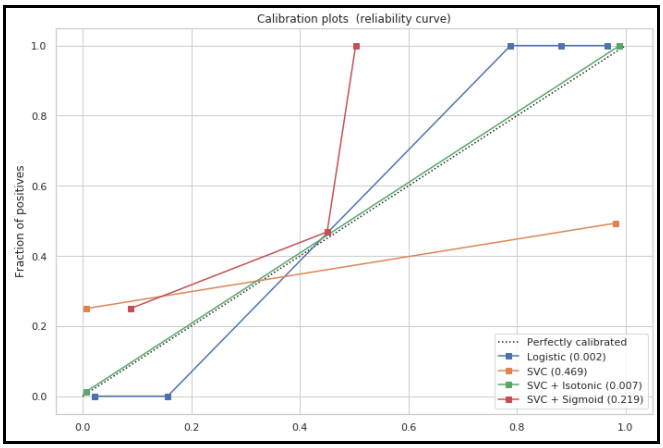

Fig. 5 Calibration Plot of Logistic in comparison with SVM

III. RESULT DISCUSSION

Based on the calibration made as above the model performance assessment is made to validate for the deployment of better classification.

Performance Score of learning models

Now, the In-sample data is used to find the performance evaluation of various models. The evaluation would suggest the best fit model for the given data.

Table-IV: Supervised Learning Models performance Score

	CV_SCORE	AUC	F-BETA
Decision Tree	0.81748	0.81744	0.8333
Random Forest	0.852	0.8989	0.8210
SVM	0.88	0.945	0.7582
K-NN	0.846	0.9266	0.8368
Naïve Bayes	0.867	0.9300	0.8710
Logistic Regression	0.8755	0.9452	0.8241

The performance indicators for the model performance assessment are Cross validation score, Area under the curve (AUC) and F-Beta. Among them the AUC would indicate the correctness of the classification. In the above table, it's very clearly shown that the AUC score is high in SVM and Logistic regression for the given in-sample data set.

The performance of the Model SVM and Logistic regression for given 1101 samples of ASD are tested. Hence proves that statistical analysis would support the model deployment for better accuracy of classification of about 87.5% to 88%.

IV. CONCLUSION

The results obtained for in-sample datasets through probabilistic calibration indicates that for the given dataset, Logistic Model and SVM model possess *least* brier score (which also describes the measure of the mean squared error). Hence through probability measure approach Logistic and SVM model proves to be the desired one for the given data. Also another approach of testing in-sample data was on different supervised model with cross-validation score performance, which validates the choice of model as most suitable for the given data which is the same significant model arrived in statistical approach. In both the approach, its observed that Logistic and SVM are more suitable to deploy compared to Decision tree, Random Forest, K-NN, Naïve Bayes. Logistic and SVM Model has indicated its significance with highest AUC score of 94.52% and having least brier score of 0.002 and 0.007 respectively.

As a future scope, the above result helps to know that any model can be first modeled with probability calibration for any data set and then can be deployed with the best fit model for that data analysis. It can be tried with other model like Deep Neural Network for efficient deployment of the model of any given data.

REFERENCES

1. F. X. Castellanos, K. Alaerts, D. Ph, J. S. Anderson, and D. Ph, "HHS Public Access," vol. 19, no. 6, pp. 659–667, 2014.
2. X. Bi, Y. Wang, Q. Shu, Q. Sun, and Q. Xu, "Classification of Autism Spectrum Disorder Using Random Support Vector Machine Cluster," *Front. Genet.*, no. 2011, pp. 1–17, 2018.
3. E. DiCicco-Bloom *et al.*, "The Developmental Neurobiology of Autism Spectrum Disorder," *J. Neurosci.*, vol. 26, no. 26, pp. 6897–6906, 2006.
4. S. Sarabadani, L. C. Schudlo, A.-A. Samadani, and A. Kushki, "Physiological Detection of Affective States in Children with Autism Spectrum Disorder," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2018.
5. A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review," *Front. Aging Neurosci.*, vol. 9, no. OCT, pp. 1–12, 2017.
6. L. Mertz, "New Quantitative Approach to Autism Diagnosis," *IEEE Pulse*, vol. 10, no. 2, pp. 34–36, 2019.
7. F. N. Buyukoflaz and A. Ozturk, "Early autism diagnosis of children with machine learning algorithms," *26th IEEE Signal Process. Commun. Appl. Conf. SIU 2018*, pp. 1–4, 2018.
8. S. Kim and J. Choi, "An SVM-based high-quality article classifier for systematic reviews," *J. Biomed. Inform.*, vol. 47, pp. 153–159, 2014.
9. F. Peng, D. Schuurmans, and S. Wang, "Augmenting naive Bayes classifiers with statistical language models," *Inf. Retr. Boston.*, vol. 7, no. 3–4, pp. 317–345, 2004.
10. Vikramkumar, V. B, and Trilochan, "Bayes and Naive Bayes Classifier," 2014.
11. S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002.

12. H. N. Peterson J. J., Yahyah M., Lief K., "Predictive Distributions for Constructing the ICH Q8 Design Space, pp. 55-70, In Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture," 2017.
13. Roopa B S and Dr. R Manjunatha Prasad, "Concatenating framework in ASD analysis towards research progress," pp. 3–5, ICATECE-2019.
14. [Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

AUTHORS PROFILE

Roopa B S, working as Associate Professor at DBIT, Bengaluru. She has completed her BE in Electronics & Communication Engineering Under Bangalore University, M Tech at BMSCE under VTU and Pursuing Ph.D. at DSATM under VTU. She has published 2 International Journal and 4 International Conference papers. Her research area is Medical Image Processing and Data analysis of Medical images.

Dr. R Manjunatha Prasad, Working as Professor and Head of ECE Dept., DSATM, Bengaluru. His research interest is in Image Processing and VLSI. He published 14+ papers in many journals and conferences nationally and internationally. He is an ISTE member. Best Paper Award for "Dermatotraction – The new and best way for wound closure", International conference on Emerging trends in Engineering and Technology at Bangalore, on April 26, 2015.