2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9389902

Fetal Health Classification Based on Machine Learning

[1st] Jiaming Li
School of Intelligent Systems Science and Engineering/
JNU-Industry School of Artificial Intelligence
Jinan University
Zhuhai, China
lijiaming@stu2018.jnu.edu.cn

[2nd] Xiaoxiang Liu*
School of Intelligent Systems Science and Engineering/
JNU-Industry School of Artificial Intelligence
Jinan University
Zhuhai, China
tlxx@jnu.edu.cn

***Abstract*—Cardiotocogram (CTG) is the most widely used in the clinical routine evaluation of the main approach to detect fetal state. In this paper, twelve machine learning single models have firstly experimented on CTG dataset. Secondly, the soft voting integration method is used to integrate the four best models to build the Blender Model, and compared with the stacking integration method. Compared with the traditional machine learning models, the model proposed in this paper performed excellently in various Classification Model evaluations, with an accuracy rate of 0.959, an AUC of 0.988, a recall rate of 0.916, a precision rate of 0.959, a F1 of 0.958 and a MCC of 0.886.**

Keywords—CTG, Fetal Health, Machine Learning, Ensemble Learning

I. INTRODUCTION

Cardiotocogram (CTG) is the most widely used in the clinical routine evaluation of the main approach to detect fetal state. Prenatal monitoring of CTG contains two core physiological signals: fetal heart rate (FHR) and uterine contractions (UC). Among them, fetal heart rate refers to the number of fetal heart beats per minute (BPM), fetal distress will lead to low or high fetal heart rate abnormal phenomenon. In clinical prenatal health diagnosis of fetus, CTG as fetal anomaly "early detection, early diagnosis" of the important technology, through the study of the monitoring of heart rate, contractions, combined with heart rate curve and contractions pressure curve evaluation in the intrauterine fetal development condition, provide clinicians with the fetus in pregnant women palace of important physiological and pathological information, effectively prevent premature birth, is to reduce perinatal mortality, and one of the main measures of improving the quality of the population. In recent years, with the rapid development of signal processing technology and artificial intelligence, many researchers began to try to apply machine learning algorithm to fetal state intelligent assessment.

In 2012, Huang et al. [1] constructed three models to predict fetal distress and avoided the threat of fetal hypoxia to fetal health, among which the accuracy of Discriminant Analysis was 82.1%, the accuracy of Decision Tree was 86.36%, and the accuracy of Artificial Neural Network was 97.78%.

In 2013, Ocak et al. [2] proposed the ANFIS to predict the fetal state according to the features extracted from fetal heart rate and contractions.

In 2014, Peterek et al. [3] proposed a Random Forest (RF) method to classify CTG signals, and provided a performance comparison with Classification And Regression Tree and Self-Organizing Map methods.

In 2015, Sindhu et al. [4] constructed a system based on IAGA, which adopted Sigma scaling and achieved 94% accuracy. In the same year, Shah et al. [5] constructed a Bagging ensemble learning method based on three traditional decision tree algorithms, which used CTG signals to identify fetal normal and pathological states.

In 2016, Yilmaz et al. [6] constructed three artificial neural network models, MPNN, PNN and GRNN, of which the precision rate of MPNN was 84.44%, PNN was 87.63%, and GRNN was 85.81%.

In 2017, Subha et al. [7] proposed a hybrid filtering-inclusion method combining IG and OBFA to select the most relevant characteristic parameters, and then used SVM to judge the fetal state.

In 2018, M. Ramla et al. [8] built the CART decision tree model on CTG dataset, in which the calculation accuracy of the entropy method and Gini coefficient method can reach 88.87% and 90.12% respectively.

In 2019, Jayashree Piri et al. [9] used MOGA-CD for important feature extraction, and then compared multiple models, and finally reached the highest accuracy rate of 94% in XGBoost.

In 2020, Yue Fei et al. [10] proposed a fuzzy C-means clustering based Adaptive Neuro-Fuzzy Inference System (FCM-ANFIS), which used the fuzzy Cmeans clustering algorithm to divide the fuzzy space, and adjusted the parameters through the self-learning mechanism of the neural network and the least-squares algorithm.

To sum up, researchers have explored different methods on CTG dataset in recent years, but there is still room for improvement in the accuracy of CTG dataset. Moreover, no researchers have evaluated the performance of different machine learning models from multiple indicators. Therefore, this paper evaluates the performance of fourteen different machine learning models in terms of Accuracy, AUC, Recall, Precision, F1, MCC. In Section II, materials and methods are presented. Section III shows the results. In Section IV, conclusion and discussion are presented.

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

II. Materials and Methods

A. Dataset Description

This dataset [11] used in this paper comprises 2126 CTG data classified by experts, including Normal, Suspect and Pathological. In the dataset, the three classes are dumb, Normal is represented by 1, Suspect is represented by 2, and Pathological is represented by 3.

B. Machine Learning Models

1) Gradient Boosting Classifier

Gradient Boosting Classifier is based on GBDT [12], and the algorithm of data classification or regression is achieved by using the addition model (that is, linear combination of basis functions) and continuously reducing the residual generated in the training process. The training process is shown in the Figure. 1 below:

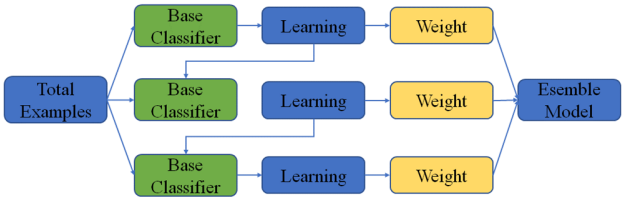

Figure.1 The training process of GBDT

2) CatBoost Classifier

CatBoost [13] is a machine learning library created by Russian search giant Yandex in 2017, which is a kind of Boosting algorithm. CatBoost is a kind of decision tree based on oblivious trees as the base to study and realize fewer parameters, supports category type variables and high accuracy GBDT framework, effective and reasonable processing handles type characteristic, but also solves the problem of deviation and predicts the gradient shift, thus reducing the occurrence of a fitting, and then improves the accuracy and generalization ability of the algorithm.

3) Light Gradient Boosting Machine

LightGBM [14] was proposed mainly to solve the problem of GBDT in mass data, so that GBDT can be better and faster used in industrial practice. It includes two new techniques: GOSS which can deal with a large number of data instances and EFB which can deal with a large number of features.

4) Cascade Forest Classifier

Cascade Forest Classifier is based on the Deep Forest proposed by Zhou et al. [15], which is the ancestor of the non-neural network depth model. Deep forest algorithm is a supervised machine ensemble learning algorithm based on random forest algorithm under the inspiration of deep learning theory and deep neural network. The structure of Cascade Forest is shown in Figure. 2.

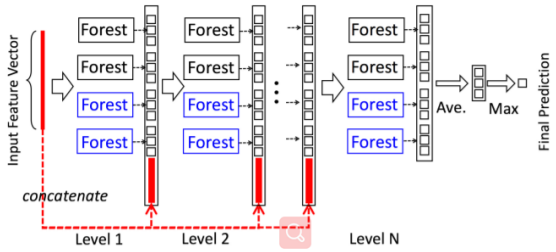

Figure.2 The structure of Cascade Forest

5) Ada Boost Classifier

Ada Boost Classifier is the most successful representative of Boosting, rated as one of the top 10 data mining algorithms [16]. Ada Boost Classifier adopts the idea of iteration. Only one weak classifier will be trained in each iteration, and then the calculated weak classifier will be used in the next iteration. This means in the Nth iteration, there are a total of N weak classifiers. The first N-1 classifiers have been trained before, and their various parameters will no longer be changed. This iteration only trains the Nth classifier, and the final classification output depends on the comprehensive effect of these N classifiers.

III. Results

A. Classification Model Evaluation

The accuracy rate refers to the accuracy of model prediction, which is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples, and can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The precision rate refers to the number of true positive cases in the predicted positive cases, reflecting the accuracy of the predicted model. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The recall rate is the proportion of correctly predicted positive to all actually positive, and the formula is:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

In order to evaluate the merits and demerits of different algorithms, the concept of F1 is proposed on the basis of Precision and Recall to make an overall evaluation of Precision and Recall. The formula is as follows:

$$F1 = \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

The MCC is generally considered as a more balanced index to measure the classification performance of dichotomies, which can be applied when the sample content of two categories is very different. The formula is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} (5)$$

The ROC curve is also known as receiver operation characteristic curve. By traversing all classification thresholds, FPR and TPR corresponding to each threshold are calculated. The horizontal axis is FPR, the vertical axis is TPR, and all points are connected successively to get a curve, which visually displays the performance of the classifier on the image. AUC represents the total area covered under the ROC curve, which can reflect the quality of the classifier numerically. Its measurement criteria are as follows:

(1) AUC=1, perfect classifier.

(2) 0.5<AUC<1, better than random guess. When thresholds are set properly, they can have predictive value.

(3) AUC=0.5, just like random guessing, the model has no predictive value.

B. Experimental Results

In this paper, firstly, the CTG dataset is randomly divided into training set: test set = 7:3, in which there are 1488 samples in the training set and 638 samples in the test set.

TABLE I. EVALUATION RESULTS OF DIFFERENT MACHINE LEARNING MODELS IN CTG TEST SET

Model	Accuracy	AUC	Recall	Precision	F1	MCC
Gradient Boosting Classifier	0.955	0.985	0.911	0.956	0.954	0.876
CatBoost Classifier	0.955	0.986	0.903	0.955	0.954	0.874
Light Gradient Boosting Machine	0.953	0.988	0.904	0.953	0.952	0.869
Extreme Gradient Boosting	0.951	0.988	0.899	0.951	0.950	0.863
Cascade Forest Classifier	0.947	0.981	0.871	0.934	0.900	0.857
Random Forest Classifier	0.940	0.984	0.864	0.940	0.938	0.832
Extra Trees Classifier	0.936	0.985	0.846	0.935	0.933	0.818
Decision Tree Classifier	0.911	0.875	0.854	0.912	0.911	0.756
Logistic Regression	0.885	0.924	0.752	0.878	0.878	0.664
K Neighbors Classifier	0.885	0.913	0.753	0.880	0.879	0.665
Linear Discriminant Analysis	0.882	0.955	0.741	0.881	0.880	0.669
Ada Boost Classifier	0.875	0.874	0.742	0.874	0.870	0.640
Stacker Model	0.952	0.987	0.913	0.954	0.951	0.866
Blender Model	0.959	0.988	0.916	0.959	0.958	0.886

In this paper, K-fold cross validation is used to train the model [17], where K =10. During the training process, each fold divides the training set into 10 parts, of which 9 parts are used for training and 1 part is used for verification. Input the test set into the trained model to get the evaluation results of the model.

We first carry out single model training and get model performance, integrating the four best single models, Gradient Boosting Classifier, CatBoost Classifier, Light Gradient Boosting Machine and Extreme Gradient Boosting. The Stacker Model integrates the above four models using the stacking method. The Blender Model uses the soft voting method to test the different models with different class labels based on the argmax of the sums of the predicted probabilities.

TABLE I. shows the evaluation results of different machine learning models in CTG test set, in which it can be seen that the Blender Model performs best in various classification model evaluations. Compared with the best performance single model, Gradient Boosting Classifier or stacking integration method, all classification model evaluations have been improved, in which MCC is 1 percent higher than Gradient Boosting Classifier, 2 percent higher than the Stacker Model.

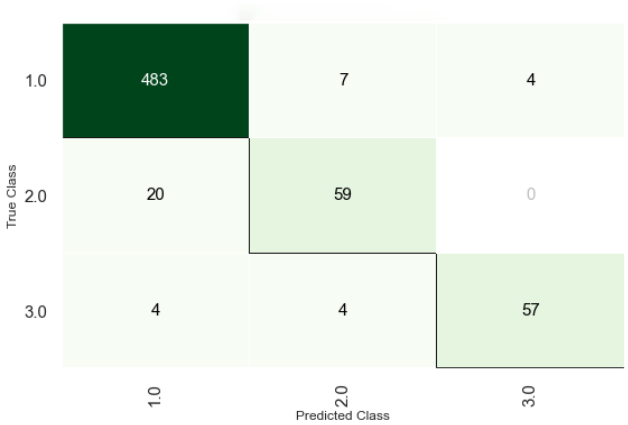

Figure.3 The confusion matrix of the Blender Model

Figure.3 shows the confusion matrix of the Blender Model, It can be seen that the vast majority of samples are correctly predicted. In addition, only 4 samples of class 3 are predicted by the Model as class 1, and none are mistaken for class 2.

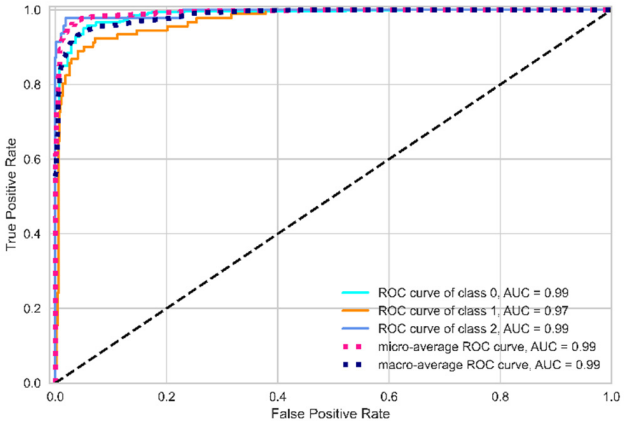

Figure.4 The ROC curves for the Blender Model

AUC is defined as the area under the ROC curve, which is a performance index to measure the performance of the learner [18]. When AUC is equal to 1, the model is a perfect model. The closer the AUC is to 1, the better the model performance will be Figure.4 shows the ROC Curves for the Blender Model, showing that the Model has high AUC for different classes, of which the AUC of class 1 is 0.99, that of class 2 is 0.97, and that of class 3 is 0.99. The results show that the Blender Model has excellent performance.

IV. Conclusion and Discussion

In this paper, twelve machine learning single models have firstly experimented on CTG dataset. Secondly, the soft voting integration method is used to integrate the four best models to build the Blender Model, and compared with the stacking integration method. Finally, various Classification model evaluations are used to evaluate and analyze all machine learning models, as well as the confusion matrix and ROC curve of the Blender Model. Experiments show that compared with the traditional machine learning models, the Blender Model performed excellently in various Classification Model evaluations, with an accuracy rate of 0.959, an AUC of 0.988, a recall rate of 0.916, a precision rate of 0.959, a F1 of 0.958 and a MCC of 0.886. Due to the limitation of the CTG dataset, this paper can be improved in the following directions:

(1) The dataset of this paper may not be rich enough, and the performance might have been better if there had been more data.

(2) Considering the influence of different features on the model, feature extraction was not carried out in this paper to comprehensively evaluate the performance of different machine learning models on CTG dataset. Further feature extraction will be carried out in the future to improve the performance of the model.

Acknowledgment

This work was supported in part by the Science and Technology Planning Project of Guangdong Province 2019B010137006.

References

[1] Mei-Ling Huang, Yung-Yan Hsu. Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network. 2012, 5(9):526-533.

[2] Hasan Ocak, Huseyin Metin Ertunc. Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems. 2013, 23(6):1583-1589.

[3] Peterek T, Gajdo P, Dohnalek P, et al. Human Fetus Health Classification on Cardiotocographic Data Using Random Forests: Advances in Intelligent Systems and Computing, 2014[C]. SpringerInternational Publishing.

[4] Sindhu R , Bahari J A , Hariharan M , et al. A Novel Clinical Decision Support System Using Improved Adaptive Genetic Algorithm for the Assessment of Fetal Well-Being[J]. Computational and Mathematical Methods in Medicine,2015,(2015-2-22), 2015, 2015:1-11.

[5] S. Shah, W. Aziz, M. Arif and M. Nadeem, "Decision Trees Based Classification of Cardiotocograms Using Bagging Approach," in 2015 13th International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2015 pp. 12-17.

[6] Ersen Yılmaz. Fetal State Assessment from Cardiotocogram Data Using Artificial Neural Networks. 2016, 36(6):820-832.

[7] Subha V , Murugan D , Boopathi A M . A Hybrid Filter-Wrapper Attribute Reduction Approach For Fetal Risk Anticipation[J]. 2017.

[8] M. Ramla, S. Sangeetha and S. Nickolas, "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1799-1803, doi: 10.1109/ICCONS.2018.8663047.

[9] J. Piri, P. Mohapatra and R. Dey, "Fetal Health Status Classification Using MOGA - CD Based Feature Selection Approach," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198377.

[10] Fei et al., "Automatic Classification of Antepartum Cardiotocography Using Fuzzy Clustering and Adaptive Neuro -Fuzzy Inference System," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 1938-1942, doi: 10.1109/BIBM49941.2020.9313143.

[11] Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318

[12] Friedman J H . Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[13] Dorogush, Anna & Ershov, Vasily & Gulin, Andrey. (2018). CatBoost: gradient boosting with categorical features support.

[14] Meng, Qi. (2018). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[15] Zhou, Zhi-Hua & Feng, Ji. (2019). Deep forest. National Science Review. 6. 74-86. 10.1093/nsr/nwy108.

[16] Zhou Z H, Yang Y, Wu X D, Kumar V.The Top Ten Algorithms in Data Mining.New York, USA:CRC Press, 2009,127149

[17] Salzberg,S.L..On comparing classifiers: Pitfalls to avoid and a recommended approach[J].Data Mining and Knowledge Discovery,1997,1(3):317-328.

[18] Zweig, M.H. & Campbell, Gregory. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical chemistry. 39. 561-77. 10.1093/clinchem/39.4.561.