Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240643

Development and validation of an Adaboost model for breast cancer classification

Heqing Wang

Zhengde High School, Shenzhen, China

Wanghq2024@163.com

Abstract. The mutation of breast cells leads to breast cancer, a lump causing illnesses. A large amount of women who have the concern of cancer are suffering from mutation in their breast cells. Early diagnosis of breast cancer can significantly improve the survival rate of patients by allowing them to receive timely clinical treatment. Accurate classification from tumor data prevents patients from receiving unnecessary treatment. Therefore, correctly diagnosing breast cancer and classifying patients into malignant or benign groups is a hot topic. In this study, the Adaboost algorithm is used to sort breast cancer depending on whether it is benign or malignant by combining the one-hot coding approach and imbalance learning. The one-hot coding approach overcomes the bias caused by the LabelEncoder coding approach, and the synthetic minority over-sampling technique, which improves the classification performance of Adaboost by augmenting the categorization balanced data structure with more minority class samples, so that its diagnosis of breast cancer reaches 99% accuracy, which is better than other machine learning models, we propose this model can be used for the diagnosis of breast cancer, providing a more accurate and faster way to classify breast cancer.

Keywords: Breast cancer, Machine learning, One-hot coding, Synthetic minority over-sampling technique, Adaboost.

1. Introduction

A large number of persons are afflicted with breast cancer. In women, it is the most prevalent kind of cancer. Every year, approximately 276,480 women receive confirmation of breast cancer. Due to the varied nature of breast cancer symptoms, patients are likely to undergo various tests, encompassing but not restricted to mammograms, ultrasounds and biopsies, in order to clarify if they have breast cancer. The data reveals that the survival rate is 88% five years following the diagnosis and 80% ten years thereafter [1-2]. Therefore, patients can live longer and have a far higher probability of a successful outcome from early detection of breast cancer. The current way in which breast cancer is classified is for doctors to make a diagnosis based on the results of a patient's tests. This takes a long time and has a high risk of misses and misjudgments. This is where machine learning as an automated system can be of great help. It can speed up the diagnostic process and improve diagnostic accuracy.

Shruthi S et al. used the random forest algorithm and introduced random features based on bagging and the acc reached 97% [3]. However, the University of Wisconsin dataset they used was unbalanced and malignant cases. With smaller numbers, the handling of unbalanced datasets can be further optimised. In such datasets, commonly used classifier evaluation methods will fail. However, in this

DOI: 10.54254/2755-2721/67/20240643

study we use the Synthetic Minority Over-sampling Technique to balance the data distribution and solve the problem of data imbalance. The aim is to combine this technique with Adaboost to achieve better classification of breast cancer.

2. Methods

2.1. Data sets and pre-processing

2.1.1. Using MSM-T to extract characteristic datasets

This study used a collection of breast cancer diagnostic datasets from the University of Wisconsin System, characterized by calculations from digitized Fine Needle Aspiration breast images. They describe the properties of the core in the image. The slicing planes were determined using the Multisurface Method Tree (MSM-T), which uses an extensive search in the space of 1-4 features and 1-3 slicing planes to select the relevant features, and the actual linear program to obtain the slicing plane in 3D uses a robust linear programming method with two linear, non-separable sets. A total of 569 breast cancer patients'examples (357 benign and 212 malignant) were included in the dataset. The dataset contains 30 features, the remaining two features are ID and Diagnostic, primarily for core features, and the 10 true value features calculated for each core are: Radius, Texture, perimeter, area, smoothness, compactness, sag (heaviness of the contoured depression section), concave points, symmetry, fractal dimension and no missing values were found, indicating that the dataset had sufficient sample size and various dimensions and came from real cases.

2.1.2. Data analysis and pre-processing

In order for each model to achieve its best results, an attempt is made to analyses and pre-process the data. While reading the data information, it was found that the data contains a column Unnamed:32 which contains 569 error values. These are deleted. The characters in the data were encoded using error value filtering, LabelEncoder and one-hot encodings. The malicious codes were 1 and the benign codes were 0. Features were then screened to determine the relevance and importance of each feature in assessing breast cancer characteristics using Pearson product-moment correlation coefficients, logistic regression, decision trees and random structures. These feature selection methods can remove redundant and irrelevant features in the training data, resulting in model simplification, reduced training time and reduced risk of model over-fit. Figure 1 and Figure 2 show the results of feature filtering.

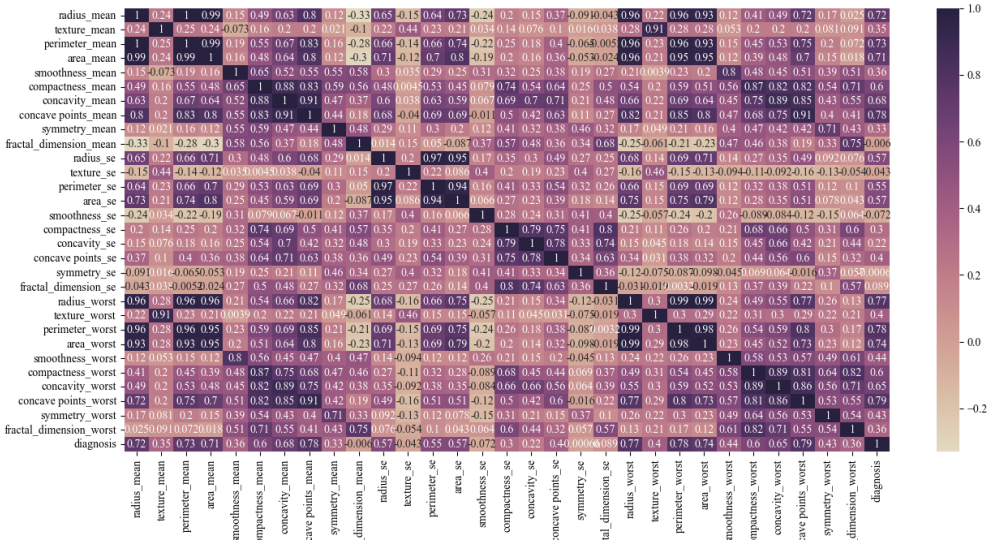

Figure 1. Heat map based on Pearson correlation coefficient feature selection.

DOI: 10.54254/2755-2721/67/20240643

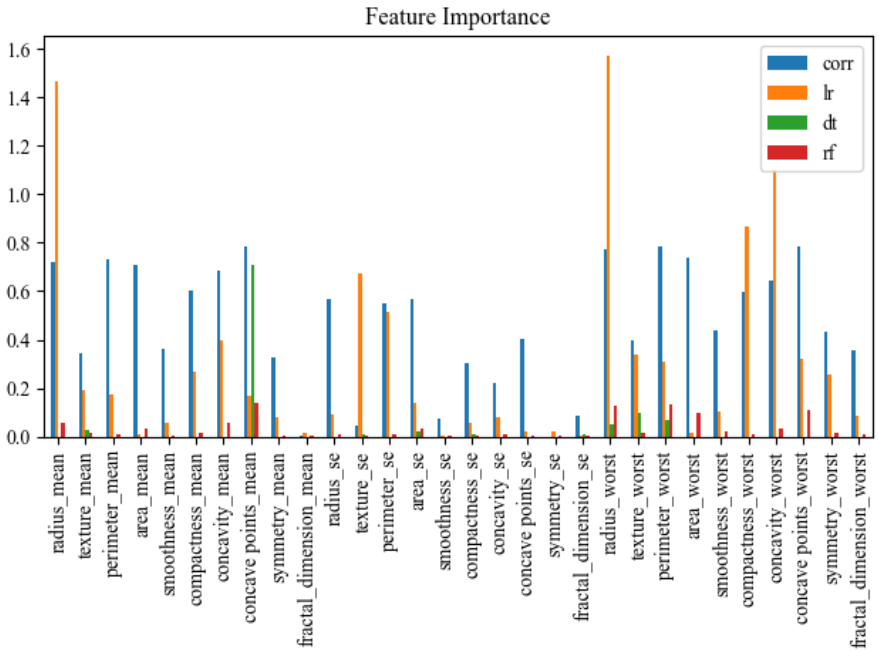

Figure 2. Bar chart of processing results using four methods.

Furthermore, the most significant characteristic that is evaluated is the worst radius, which corresponds to the average radius and has a strong correlation with the type of cancer present in the marine region. Furthermore, thermography reveals that 10 out of 30 features exhibit a stronger correlation with the type of crab (Pearson Superior correlation coefficient of 0.6 or higher), indicating that these features may reflect specific information in certain areas due to their poor performance in the model. The data is visualized to assess its balance, and the results are presented in Figure 3.

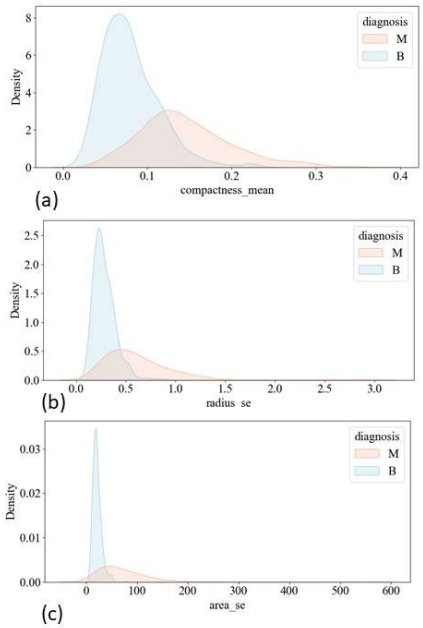

Figure 3. Data visualization results of data. (a) Compactness_mean for the distribution of Diagnosis. (b) Radius_se for the distribution of Diagnosis. (c) Area_se for the distribution of Diagnosis

DOI: 10.54254/2755-2721/67/20240643

2.1.3. Synthetic Minority Over-Sampling Technique

For a minority of samples, a simple random oversample procedure was used, which found the K nearest neighbors in dimension P for a minority of samples, multiplying them by a random number between 0 and 1 and forming a new minority sample. Sample numbers were increased in several categories to balance data structure.

The AdaBoost method prioritizes more complex samples with greater information and is less susceptible to overfitting than many other learning algorithms. In each iteration, AdaBoost balances the weights of each sample, selects sample points based on these weights, and trains the classifier Ck. Misclassified samples have their weight increased, while correctly classified samples have their weight decreased. The updated example set is then used to train the next classifier Ck. The training process is iterative and follows a logical progression [4].

2.2. The Adaboost Model

First, is the weight distribution of the training data. Every breast cancer sample, initially, is given the same weight.

$$D_1(i) = (w_1, w_2, \dots w_{569}) = \left(\frac{1}{569}, \dots \frac{1}{569}\right) \tag{1}$$

During the construction of the subsequent training set, the corresponding weight should be decreased. Conversely, If the training sample does not classify an item correctly, the item must be weighted more heavily. The revised sample is then used to train the next classifier, and the whole process of training is repeated in this way until the next classifier is trained. Eventually, the weak classifiers are combined into strong classifiers, which are the result of the training process. After training each weak classifier, the importance of classifiers which have low error rates is increased to give them a greater role in the classification function at the end, whereas the importance of classifiers which are likely to give false results is decreased, aimed at giving them a less significant role in the classification function at the end. In essence, the final classifier assigns greater weight to weak classifiers with low error rates, while assigning lower weight to other classifiers [5].

2.3. Other Machine Learning Models

The logistic regression algorithm is generally used in scenarios where unambiguous results are required. Typically, the function used in logistic regression compresses probability values into a particular range [6].

The decision tree algorithm is a method for approximating the value of a discrete function. This is a typical classification method where you first process the data, then create rules and a decision tree using an inductive algorithm and then analyse the new data based on the decisions. The data is classified according to a set of rules. Decision trees have a single result. This algorithm is commonly used to solve classification problems [7].

A random forest is a collection of decision trees; each decision tree in a random forest estimates a ranking, a procedure known as 'voting'. Ideally, with each vote coming from a decision tree, the classification with the most votes is chosen [8].

Gradient Boosting Decision Trees (GBDT) use an additive model to train multiple decision trees, and the predicted outcome is the sum of these decision trees. The training target of each decision tree is the residuals of the previous model [9].

XGBoost's algorithm is similar to GBDT's: it continually performs a feature split to generate a tree. As the tree learns in each round, it adjusts for the difference between the prediction and training values of the previous round of the model [10].

The neural network algorithm is a computer model developed to simulate the intuitive thinking of the human brain. This algorithm connects multiple layers of neurons through Folding Neural Networks, Repetitive Neural Networks, etc. to classify different data information and learn from multiple training

DOI: 10.54254/2755-2721/67/20240643

data. In order to improve the accuracy of the algorithm, the connection weights between the neurons are continuously optimized [11].

2.4. Model implementation

Divide the training and test sets, set the proportion of the test set to 0.3, and set the random number to 0. When training a neural network, set the number of neurons in the first fully connected layer at 64 and use the Relu activation function; 32 neurons existed in the second fully connected layer, with the utility of Relu activation function; there is only 1 neuron in the third fully connected layer, implementing the sigmoid activation function. The Adam optimizer is used in training, with binary cross-entropy loss and precision metrics. The number of samples for each gradient update is set to 100 and the number of iterations to train the model is 50.

2.5. Assessment indicators

The question of whether breast cancer is malignant or not was formulated as a dichotomous problem in this study. Given this issue, first of all, accuracy is an essential evaluation indicator, usually expressed as a percentage of the number of correctly classified samples and the number of all samples used for diagnosis. In addition, failure to recognize illnesses during diagnosis usually results in higher costs than misdiagnosis. Therefore, the recall score (RS) is used as an important evaluation indicator for the performance of these classification models. Recall1 is defined as the probability of predicting malignant breast cancer samples as malignant. In addition, the area undercurve (AUC) is a commonly used measure of model performance to measure the generalization ability of the model. Precision and recall are the two components of the F1 score. The goal of the F1 score is to combine precision and recall into one metric. At the same time, the F1 score handles unbalanced data well [12].

3. Results and discussion

3.1. Training results

After 10 training sessions, the accuracy stabilized and then increased more slowly. At 50 drives, the accuracy approached 1.0, the loss curve showed a decreasing trend, and at 15 times the training frequency, the rate of decline slowed down. When the training frequency reached 50 times, the loss was about 0.1 (Figure 4).

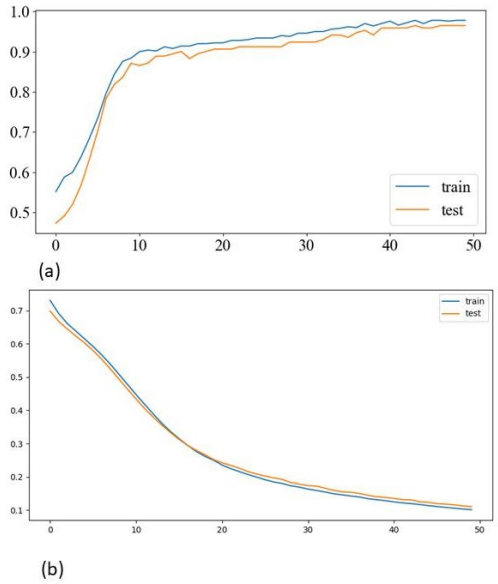

Figure 4. Training results of the neural networks used for classification.

DOI: 10.54254/2755-2721/67/20240643

3.2. Model comparison

As shown in Figure 4, the performances of the models were compared in terms of accuracy, callback 1, precision 1 and AUC. The graph shows that XGBoost surpasses AdaBoost in accuracy, accuracy 1 and AUC performance and that AdaBoost1 has a significant callback advantage (Table 1). Since false-negative results are usually more expensive than false-positive results, more attention should be paid to recall rates when diagnosing diseases. A high recall rate1 indicates a greater application potential of AdaBoost, which is why methods were used to optimize the model.

Table 1. The result of models with LabelEncoder

	ACC	Recall0	Precision0	F1-score0	Auc	Recall1	Precision1	F1-score1
LR	0.9359	0.8743	0.9487	0.9105	0.9913	0.9720	0.9286	0.9500
DT	0.9332	0.9375	0.8953	0.9153	0.9358	0.9335	0.9615	0.9476
RF	0.9707	0.9371	0.9833	0.9598	0.9808	0.9911	0.9636	0.9768
GBDT	0.9474	0.9061	0.9515	0.9279	0.9888	0.9717	0.9455	0.9585
Adaboost	0.9529	0.8904	0.9814	0.9341	0.9895	0.9908	0.9379	0.9636
Xgboost	0.9645	0.9368	0.9614	0.9520	0.9932	0.9806	0.9631	0.9721

Table 2. The result of models with One-Hot label

	ACC	Recall0	Precision0	F1-score0	Auc	Recall1	Precision1	F1-score1
LR	0.9416	0.9625	0.9450	0.9538	0.9930	0.9062	0.9354	0.9204
DT	0.9471	0.9720	0.9457	0.9585	0.9392	0.9062	0.9507	0.9276
RF	0.9648	0.9904	0.9548	0.9726	0.9822	0.9220	0.9829	0.9515
GBDT	0.9651	0.9809	0.9629	0.9721	0.9910	0.9370	0.9676	0.9518
Adaboost	0.9879	1.0000	0.9816	0.9911	0.9967	0.9691	1.0000	0.9840
Xgboost	0.9824	1.0000	0.9728	0.9859	0.9950	0.9527	1.0000	0.9761

The results of comparing the two encoding methods show that the performance of AdaBoost has significantly improved after switching the encoding method from LabelEncoder to One-Hot (Table 2). AdaBoost is superior to XGBoost in Acc, Recall1, and AUC, while Precision1 has no noticeable advantages, as Figure 5 shows. One-hot encoding solved the problem of discrete values of categorical data and avoids bias during the encoding process.

Both the training process consists of One-Hot encoding to compare the application and non-application of the unbalanced learning method. After changing the encoding method and using unbalanced learning, AdaBoost's performance improved. Neural network was added into the modes compared and its training results were shown in Figure 6. The f0 score of AdaBoost increased from 0.9341 to almost 0.99, showing the effectiveness of unbalanced learning for this breast cancer dataset, as in Figure 7 can be seen.

DOI: 10.54254/2755-2721/67/20240643

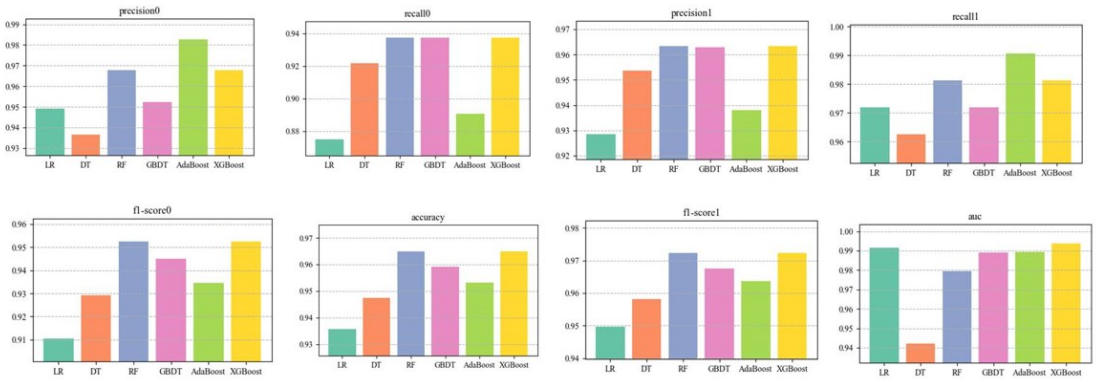

Figure 5. Performance comparison of model training using LabelEncoder encoding method.

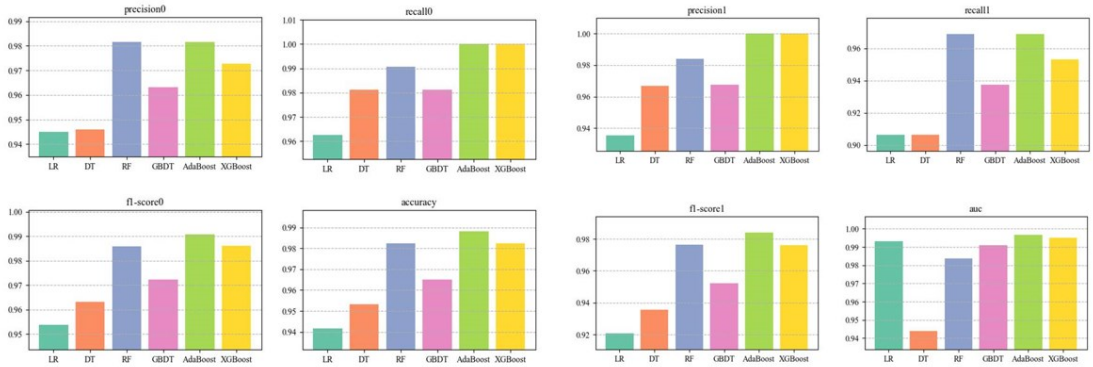

Figure 6. Performance comparison of training models using one-hot encoding method

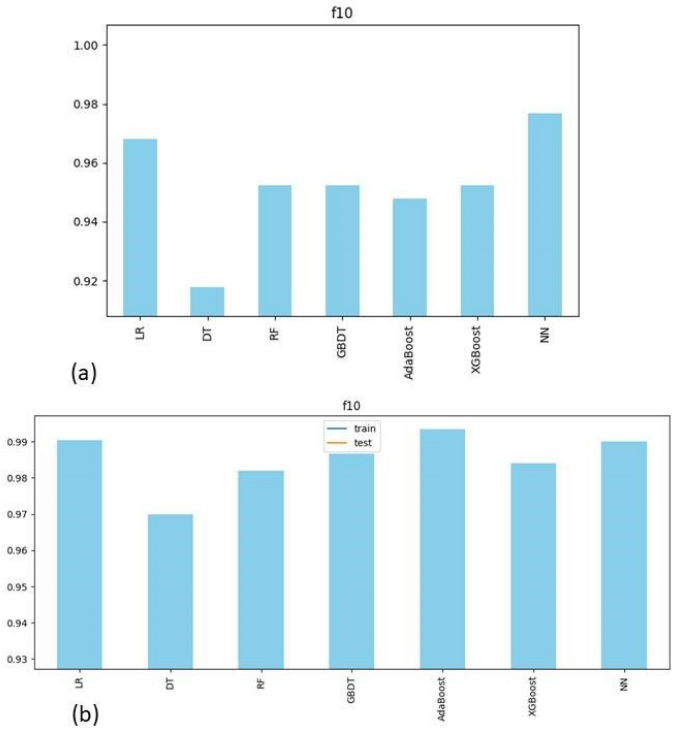

Figure 7. Performance comparison before and after using imbalanced learning. (a) Without using imbalanced learning. (b) Using imbalanced learning.

DOI: 10.54254/2755-2721/67/20240643

In machine learning, imbalanced learning is used to prevent bias attributed to the lack in the extreme values of data, solve the problem of data imbalance, and improve model performance. On grounds of this, better results were presented when training the model (Acc=98.9%) than has been reported so far (Acc=97.2%) [2].

4. Conclusion

Accurate and convenient diagnosis of whether breast cancer is malignant or not has been a challenge of concern. In this study, the Adaboost algorithm is used to distinguish the malignant samples from benign counterparts of breast cancer by combining the one-hot coding approach and imbalance learning with its diagnosis of breast cancer reaches 99% accuracy, which is better than other machine learning models. The main contributions of this study are as follows: first, it is shown that AdaBoost has better performance than XGBoost for screening malignant breast cancer based on this dataset. What's more, it is shown that one-hot coding and the application of the Synthetic Minority Over- Sampling Technique on unbalanced datasets can achieve better performance of AdaBoost-based breast cancer diagnosis. This model can be used for the diagnosis of breast cancer, providing a more accurate and faster way to classify breast cancer. In order to achieve better results in assisting doctors in sorting breast cancer, the model can be further developed into a real-time online learning system to learn the latest collected data to improve the generalization ability of the model to the latest data. Overall, this study implemented an Adaboost model for breast cancer classification and achieved some research results. In addition to mutation of breast cells, patients with malignant breast cancer also exhibit some changes in the color of the breast skin in the early stages of the disease. This suggests that combining data of the shades of breast skin color which has changed due to the development of tumor may improve the accuracy of breast cancer diagnosis.

References

[1] Shruthi S, Binu Xavier F, Ravi Kumar A, Yeshwanth S and Mahalinga V Mandi. 2020 Breast Cancer Classification using Python Programming in Machine Learning J Eng Res Tech 9
[2] Siegel R L, Miller K D and Jemal A 2020 Cancer statistics CA Cancer J Clin 70 7–30
[3] Shruthi S, Binu Xavier F, Ravi Kumar A, Yeshwanth S and Mahalinga V Mandi. 2020 Breast Cancer Classification using Python Programming in Machine Learning J Eng Res Tech 9
[4] Nakamura M, Kajiwara Y, Otsuka A and Kimura H 2013 LVQ-SMOTE-Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data BioData Min 6 16
[5] Hastie T, Rosset S, Zhu J, et al. 2009 Multi-class AdaBoost Stat Int 2 349-60
[6] LaValley M P 2008 Logistic regression Circulation 117 2395-9
[7] Song Y Y and Ying L U 2015 Decision tree methods: applications for classification and prediction Shanghai Arch Psy 27 130
[8] Rigatti S J 2017 Random Forest J Insur Med 47 31-9
[9] Ke G, Xu Z, Zhang J, et al 2019 DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 384-94
[10] Li Q, Yang H, Wang P, Liu X, Lv K and Ye M 2022 XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer J Transl Med 20 177
[11] Wu Yand Feng J 2018 Development and application of artificial neural network Wireless Per Com 102 1645-56