ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

Predictive Analysis of Breast Cancer Based on Stacking Algorithm

Kaiyao Tan[1], Zhikun Luo[2]

[1]GuiZhou University, School of Computer Science, Guizhou 550025, China
[2]Hunan University of Science and Technology, School of Resource & Environment and Safety Engineering, Xiangtan 411201, China

Abstract: *With the development of computers, machine learning algorithms can be applied in the medical field to solve many classification and prediction problems, thus assisting professionals to quickly judge and diagnose the disease. In this paper, we propose a breast cancer prediction model based on stacking algorithm, which integrates several traditional machine learning algorithms and compares with Adaboosting, SVM and other algorithms in terms of accuracy, ROC curve, PR curve, F1 value index, etc. The experiments show that the accuracy of the breast cancer classification model based on stacking algorithm can reach 97.23%, which is 6% higher than the classification accuracy of SVM, Adaboosting and other algorithms, and the AUC value of ROC curve can be improved by up to 0.26, which provides a certain reference value in breast cancer prediction examination and so on.*

Keywords: *Stacking, Ensemble Learning*

1. Introduction

The global incidence of breast cancer has been increasing since the 1970s, with World Health Organization cancer experts stating in February 2021 that as many as 19.3 million patients will be diagnosed with cancer in 2020, and breast cancer will account for 11.7% of these people [1-2].

With the development of computers, the accumulation of medical data and the improvement of medical technology, machine learning, ensemble learning and other algorithms are gradually used in the medical field, and a large number of researchers have done many related studies. For example, Liu [3] *et al*. studied the application of support vector machine algorithm for breast cancer classification decision, which reflects the good effect of SVM algorithm in diagnosing breast cancer; Chun Cai et al. used deep learning for predicting the magnitude of change in subsequent blood glucose level in diabetes, which provides great reference value for treating diabetes. Zhang *et al* [4] studied the application of ensemble learning in diabetes prediction, in which the Random Forest algorithm was mainly used to compare with traditional machine learning algorithms, showing that ensemble learning has some superiority. Li *et al* [5] studied breast cancer based on the C-AdaBoost model and concluded by comparing multiple datasets that using ensemble algorithms instead of individual machine learning classifiers can yield higher prediction accuracy on the disease prediction problem. However, it has been a research challenge for researchers to improve the prediction accuracy of the model, which is directly related to life safety.

Based on this, this paper uses the UCI Machine-Learning DataRepository to construct a breast cancer classification model using the integrated machine learning algorithm stacking, which can diagnose breast cancer patients and determine whether they are benign or malignant. By comparing with single machine learning algorithms SVM, KNN, DecisionTree and integrated learning algorithm Adaboosting, it can be concluded that Stacking algorithm outperforms SVM, KNN and other models in terms of accuracy, ROC curve, PR curve and other metrics.

2. Research method

2.1. Single machine learning model

2.1.1. SVM

Support vector machine (SVM) is an algorithm based on the principle of structural risk minimization. The classification principle is to find the hyperplane that maximizes the sum of distances between

ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

positive and negative sample data in an infinite number of interfaces that can partition the positive and negative samples.SVM has good results in problems such as nonlinear and small samples and can find optimal solutions with small sample size [6].

2.1.2. KNN

KNN (K-Nearest-Neighbor) algorithm is a common machine learning method. The basic method is to extract samples and calculate the K samples that are closest to the detection sample as a reference based on some distance metric to classify which class the detection sample belongs to. The token with the most occurrences among these K sample values is usually used as the result of the classification.

For calculating the sample distance, this paper uses Euclidean distance with K value set to 5 for training as well as prediction of features.

2.1.3. Logistic Regression

Logistic regression algorithm is a common machine learning method currently used to estimate the likelihood of something happening. It is a multivariate analysis method that studies the relationship between a target variable Y and a set of influences X1,X2,... ,Xm a multivariate analysis method that investigates the relationship between... The target variable Y of logistic regression indicates the occurrence or non-occurrence of an event and takes the values of 1 and 0.

This paper constructs a model with logistic regression that $p(Y = 1) > 0.5$ as malignant, $p(Y = 1) < 0.5$ for benign, and training as well as prediction of the features.

2.1.4. Decision Tree

Decision tree model is a tree structure model that uses the known probability of various situations to occur for decision classification [7]. The decision tree consists of two parts, nodes and edges. The decision tree is divided into two branches at each node based on the set attributes and is repeated until the leaf node is reached and the category result is obtained. The schematic diagram of the decision tree is shown in Figure 1.

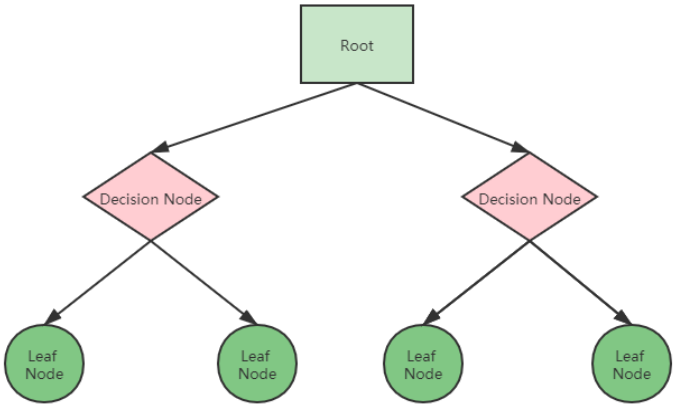

Figure 1: DecisionTree

The uncertainty indicator (criterion) of the random variables of the decision tree used in this paper is the Gini index (gini) of the CART decision tree for selecting the optimal features when used in classification problems; the minimum value of the number of samples contained in each leaf node (min_sample_leaf) is specified as 7, the minimum value of the number of samples contained in each decision node (min_ samples_split) is 2, and the maximum depth (max_depth) is 5, and the model is constructed accordingly.

2.2. Ensemble machine learning model

2.2.1. Stacking

Stacking algorithm is an ensemble learning model that constructs multiple models on the data and fuses the modeling results of all models according to a certain principle [8]. In this paper, Stacking algorithm is used to construct a prediction model, complete a five-fold cross-validation on the training set, randomly divide the training set into five subsets, use four of them for training, make predictions for the remaining set, and also make predictions for the test set, and then repeat this process four times using

ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

different four subsets until finally for each training subset there is a corresponding prediction value, and combine each training subset The output of each training subset is combined into the training set of the second layer model. The mean of the prediction results of the test set during the training process is then taken as the feature of the test set of the second layer model. The bottom model will choose the model with higher prediction accuracy, in this paper, we choose KNN, SVM, Decision Tree, Adaboosting as the bottom model, the top model will usually choose the model with good stability and explainable ability, in this paper, and we choose Logistic Regression as the top model. The schematic diagram is shown in Figure 2.

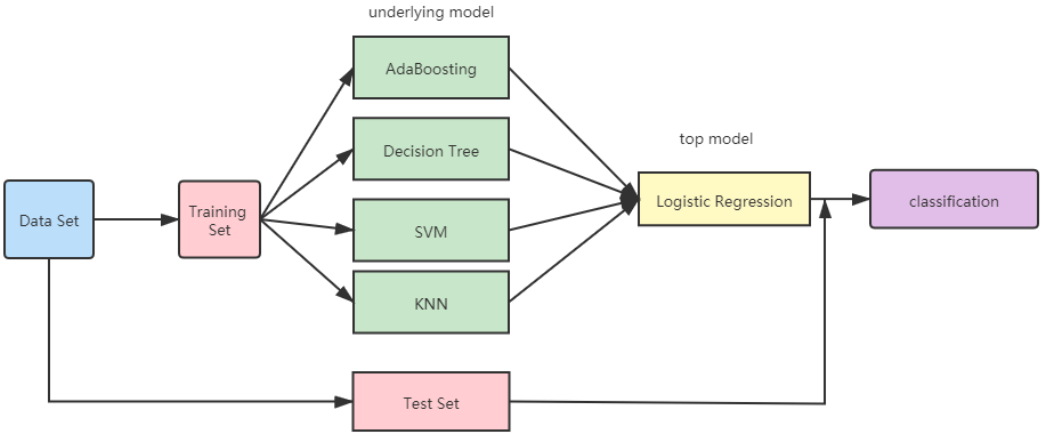

Figure 2: Stacking algorithm flow chart

2.2.2. Boosting

Boosting is a machine learning model which can be used in a cascading manner to continuously go through and adjust the distribution of samples based on the effect of the previously generated learners and then generate new base learners based on that [8]. Figure 3 shows the structural schematic of Boosting algorithm, the first step is to train on the original dataset with the initial weights to get the weak learner 1 and update the sample weights according to the difference between its predicted value and the actual value to get the training set 2. In this process, if the sample predicted value is different from the actual label, then increase its weight. The above process is repeated until the condition is satisfied.The Boosting algorithm trains a series of learners, focusing on samples that were misclassified by previous learners, enhancing the learning of previously misclassified samples, and then combining the individual learners to obtain a strong learner with satisfactory results.

In this paper, we adopt the typical representative of Boosting algorithm-Adaboosting algorithm, which takes decision tree as weak learner, the maximum number of iterations (n_estimators) of weak learner is 150, the algorithm (algorithm) is specified as "ASMME.R" and the learning_rate is 0.001, according to which the model is constructed.

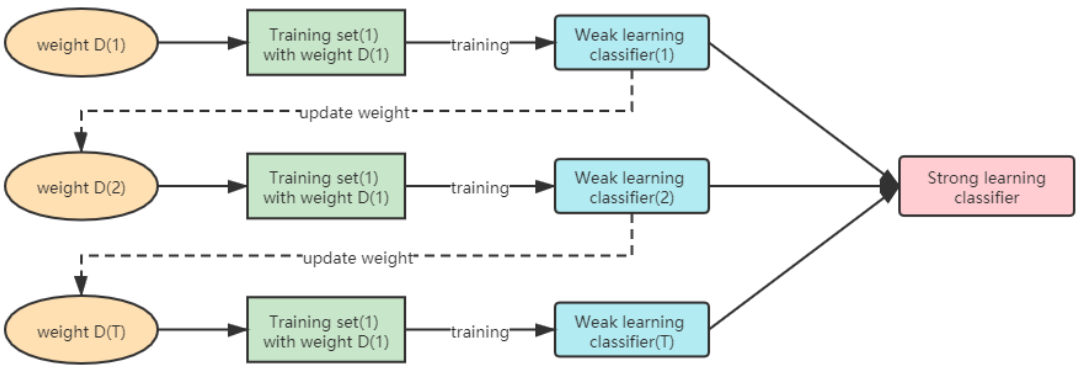

Figure 3: Schematic diagram of the Boosting model

ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

3. Results and Analysis

3.1. Statistical analysis of sample data

Table 1 shows the statistical information of some of the Conceivable Breast Cancer data. The dataset has two major categories (0 for benign and 1 for malignant) and 8 features for a total of 699 samples, the features are Clump Thickness, Uniformity of Cell Size, Uniformity of CellShape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses.

Table 1: Statistical results for the UCI dataset on the 8 features

Clump Thickness	Uniformity ofCell Size	Uniformity ofCellShape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	class
5	4	4	5	7	10	3	2	1	0
3	1	1	1	2	2	3	1	1	0
6	8	8	1	3	4	3	7	1	0
4	1	1	3	2	1	3	1	1	0
8	10	10	8	7	10	3	7	1	1

3.2. Experimental results

3.2.1. Single model experiments

In this paper, 80% of the dataset is selected as the training set and 20% as the test set, and the training set is input into SVM, KNN, and LR for training, and then the test set is input into each model, and the model accuracy shown in Figure 4 can be obtained.

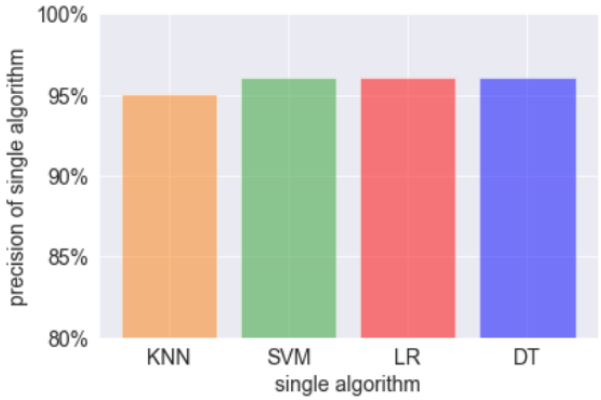

Figure 4: Precision of single models

From the figure, it can be found that the accuracy of the three models does not differ much, and the accuracy of KNN, SVM, LR, and DecisionTree are 95.33%, 96.11%, 96.03%, and 96.11%, respectively. Therefore, in this paper, we consider the Adaboosting algorithm and Stacking algorithm in the ensemble learning method to further improve the accuracy rate.

3.2.2. Ensemble model experiments

The overall comprehensive performance of a single machine learning model needs to be improved and the features learned may not be comprehensive enough. Therefore, this can be avoided to some extent by adopting the use of ensemble learning methods. By integrating multiple machine learning models, models with better generalization ability and robustness can be constructed.In this paper, the performance of the model is evaluated comprehensively by using the ensemble model Adaboosting algorithm and the Stacking algorithm model that incorporates KNN, SVM, LR, and Adaboosting algorithms using accuracy, ROC curve, F1 value, and PR curve. The results are shown in the figure below.

ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

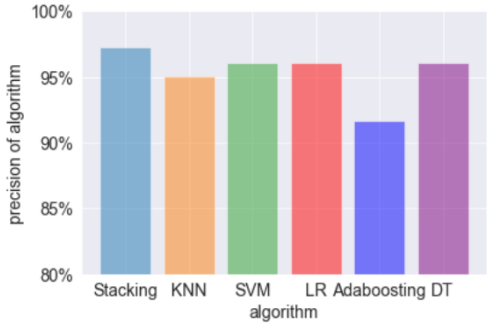

Figure 5: Precision of models

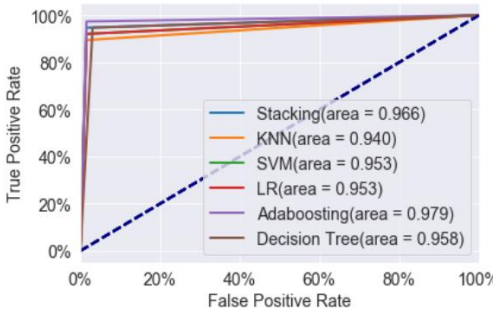

Figure 6: ROC of models

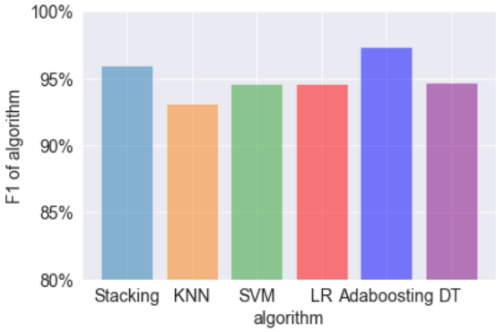

Figure 7: F1 of models

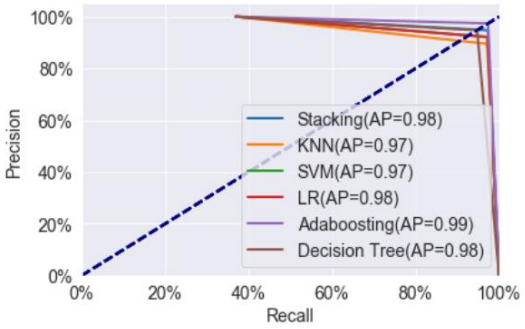

Figure 8: PR of models

Table 2: Comparison of precision, F1, AUC, AP

	Precision	F1	AUC	AP
Stacking	97.23%	0.960	0.966	0.98
KNN	95.33%	0.931	0.940	0.97
SVM	96.11%	0.945	0.953	0.97
LR	96.03%	0.946	0.953	0.98
Adaboosting	91.65%	0.973	0.979	0.99
DecisionTree	96.11%	0.947	0.958	0.98

The results in Figure 5,6,7,8 show that the Stacking breast cancer prediction model built in this paper achieves an accuracy of 97.23%, which is higher than that of the single machine learning model, as well as the integrated learning model Adaboosting, improving the accuracy by up to about 6%. The AUC value of the ROC curve is 0.966, which is second only to the integrated learning model Adaboosting, improving it by up to 0.26 compared to the single machine learning model. The F1 value of 0.960 is slightly weaker than the Adaboosting algorithm and at most 0.29 higher than the single machine learning model. The AP value of 0.98 is less different from the other algorithms.

As can be seen from Table 2, the Stacking algorithm outperforms both the single machine learning algorithm and the integrated learning algorithm Adaboosting when the metrics of precision, F1, AUC, and AP are taken into account.

4. Conclusion

In this paper, we use integrated learning Stacking algorithm fused with single machine learning algorithm to train and test based on breast cancer dataset, and compare with single machine learning model as well as integrated learning Adaboosting algorithm.we can conclude that:

(1) Using integrated algorithm instead of single machine learning algorithm can achieve better results in the field of breast cancer prediction.

(2) Using Stacking algorithm that integrates a single machine learning model can further improve the accuracy rate. These conclusions can assist physicians to perform better examinations and diagnoses.

Although the model in this paper achieved good results and demonstrated the usefulness of Stacking algorithm and integrated learning for breast cancer prediction analysis, breast cancer data collection is

ISSN 2616-5791 Vol.2, Issue 1: 36-41, DOI: 10.25236/AJMHS.2021.020107

more difficult and the available dataset is small, and more data can be collected as much as possible in future studies to improve the model and further improve the performance.

References

[1] Siegal R L, Miller K D, Jemal A. C Cancer statistics [J]. CA: A Caner Journal for Clinicians, 2016, 66(1): 7-30.
[2] Wang SH, Shi HY, Kong WW, Wang L, Li F. Recent advances in risk factors for high-incidence breast cancer in China [J]. Journal of Clinical Nursing, 2017, 16(01): 72-75.
[3] Liu H X, Zhang R S, Luan F, Yao X J, Liu M C, Hu Z D, Fan B T. Diagnosing breast cancer based on support vector machines. [J]. Journal of chemical information and computer sciences, 2003, 43(3):
[4] Zhang YX, He S, Y ou SM. Application of integrated learning in diabetes prediction [J]. Intelligent Computers and Applications, 2019, 9(05): 176-179.
[5] Li Y, Chen S-Xuan, Jia H, Wang X. Research on breast cancer prediction based on C-AdaBoost model [J]. Computer Engineering and Science, 2020, 42(08): 1414-1422.
[6] Xiong Ting. Research on multi-classifier integration method for disease diagnosis [D]. East China Jiaotong University, 2018.
[7] Bi Xuehua, Wu Miao, Wu Jing. Analysis of data mining technology in the field of Chinese medicine [J]. Computer Knowledge and Technology, 2012, 8(10): 2175-2176.
[8] Zhang Y ao. Integrated learning based pathological image analysis of breast cancer [D]. Shandong University, 2021.