Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230878

Diagnosis of breast cancer based on logical regression and random forest algorithms

Junhao Chen

Department of Computer, Central South University, Changsha, Hunan, 410017, China

8208200523@csu.edu.cn

Abstract. Currently, breast cancer possesses a dominate position among all kinds of cancers. Besides, it is also the major cause of the cancer-related death in China. Considering the existence of various cancer types, it is hard to be accurately diagnosed. After early diagnosis and treatment, the 5-year survival rate of female breast cancer reached 97%. Although early diagnosis is curable, about one third of female breast cancer patients still die of the disease. However, despite early detection and selection of new treatment methods, as many as 50% of women will still have metastasis. At present, as the cause of breast cancer has not been determined, accurate early detection is crucial to reduce mortality. This paper aims to leverage machine learning methods for achieving accurate cancer diagnosis. This paper explains the principal knowledge of logistic regression in detail, and classifies the data set by combining the logistic regression and random forest. It is a binary classification problem, which annotations are made up of malignant and benign. It can be well applied to the logical regression model. The classification results are more than 90% accurate, so machine learning-based solutions have broad application prospects and can produce certain value.

Keywords: machine learning, logistic regression, random forest, breast cancer.

1. Introduction

As one of the major types of malignant tumor, breast cancer is witnessed increasingly happened. Among all cancer patients, 25% to 30% are suffering from breast cancer [1,2]. Every year, around 1.3 million people are newly diagnosed breast cancer patients globally. Besides, about 400000 people will die from the disease. According to Chinese statistics, the incidence rate of breast cancer is the second among female malignant tumors. In some big cities, the incidence rate of breast cancer has risen to the first, and the rate in rural women is the fifth [3,4]. At present, breast cancer has become the biggest problem threatening women's health. It is crucial to early diagnose breast cancer and conduct medical interference on time [5,6]. To tackle this problem many classification methods have been applied to such diagnosis. These methods are used to analyze and compare the performance with each other, among which support vector machine is better [7,8]. Support vector machine has a better generalization capacity, especially when applied to small sample sets, compared with other conventional models [9]. When adopted to large data, it often requires a long training time. KNN method is a case-based learning, which can generate arbitrary shape decision boundaries without establishing a model. However, its classification costs a lot and needs to calculate the similarity one by one. Besides, with small k value, the model could be sensitive to noise. Given these deficiencies, corresponding improvements are implemented, but there is not yet a

DOI: 10.54254/2755-2721/16/20230878

classification method that could overcome these deficiencies at the same time. The logical regression classification method used in this paper is a linear regression after the normalization of logistic equation. This normalization method is often more reasonable, and can suppress too large and too small results (often noise), To ensure that the mainstream results will not be ignored. Moreover, the model is easy to interpret, easy to extract rules, and has good robustness against noise interference and redundant attributes.

2. Method

2.1. Dataset

The breast cancer dataset is achieved from Kaggle, including 569 samples [10]. Each sample contains 29 features and a diagnostic result. As for data preprocessing, highly correlated features are removed and PCA is applied for data dimension reduction.

2.2. Logistic regression

Logistic regression is widely applied for tackling binary classification. Considering its simplicity, parallelism, and interpretability, it is welcomed by the industry fields. This model assumes that all samples follow the same distribution, and the weights could be estimated by maximum likelihood.

2.2.1. Model. Taking binary classification as an example, assuming the existence of such a hyperplane capable of linearly separating all samples. The decision boundary could be expressed as$\omega_1 x_1 + \omega_2 x_2 + b = 0$. If a sample point is assumed as $h_\omega(x) = \omega_1 x_1 + \omega_2 x_2 + b > 0$, its category can be determined as 1. Logistic regression requires an additional layer to construct relationships between the P (Y=1), probability of positive classification, and x, the input feature. Then the prediction could be decided by a probability comparison.

Consider the binary classification problem, given the dataset:

$$D = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_i \subseteq R^n, y_i \in 0,1, i = 1,2, \dots, N \tag{1}$$

Considering that the values $w^T x + b$ are continuous, discrete values are not supported. It can be considered to fit conditional probability $P(Y = 1|x)$, because the value of probability is continuous.

However, for $w \neq 0$, the value $w^T x + b$ is R, and the probability of nonconformity is 0 to 1, so the generalized linear model is considered. The most ideal is the unit step function:

$$P(y = 1|x) = \begin{cases} 0, & z = 0 \\ 0.5, & z = 0, z = w^T x + b \\ 1, & z > 0 \end{cases} \tag{2}$$

However, considering it is nondifferentiable, and the Sigmoid function is widely used:

$$y = \frac{1}{1+e^{-(w^T x+b)}} \tag{3}$$

So, there are:

$$y = \ln \frac{y}{1-y} = w^T x + b \tag{4}$$

y is considered as the probability of positive prediction, and 1-y is the negative one. Their ratio is named odds. Given the probability of an event at p, the logarithmic probability is:

$$\ln(odds) = \ln \frac{y}{1-y} \tag{5}$$

DOI: 10.54254/2755-2721/16/20230878

Regarding y as a quasi posterior probability estimation, it could be reformulated as:

$$w^T x + b = \ln \frac{P(Y=1|x)}{1-P(Y=1|x)} \tag{6}$$

$$P(Y=1|x) = \frac{1}{1+e^{-(w^T x+b)}} \tag{7}$$

In this model, the larger value of $w^T x + b$ represents the larger likelihood that $P(Y=1|x)$ equals 1. As a result, the essence of logical regression is to learn a decision boundary, and measures the distance between a sample and the boundary, for achieving the predictive probability.

2.2.2. Cost function. This part elaborates the parameter optimizing, leveraging the maximum likelihood estimation. It is designed for seeking parameter combinations that maximize the likelihood of samples. Given $P(Y=1|x) = p(x)$ *and* $P(Y=0|x) = 1-p(x)$ *the likelihood function is:*

$$L(w) = \prod [p(x_i)]^{y_i} [1-p(x_i)]^{1-y_i} \tag{8}$$

To facilitate the solution, logarithm operation is conducted on both sides of the equation and forms:

$$L(w) = \sum [y_i \ln p(x_i) + (1-y_i) \ln(1-p(x_i))] = \sum [y_i (w x_i) - \ln(1+e^{w x_i})] \tag{9}$$

It is recognized as loss function in machine learning. If the average logarithmic likelihood loss is taken across the entire dataset, it can be obtained:

$$J(w) = -\frac{1}{N} \ln L(w) \tag{10}$$

It means maximizing the likelihood function is equivalent to minimizing the loss.

2.2.3. Random gradient descent. There are many methods for solving logistic regression, and the most typical one is gradient descent. Its main objective is to identify the direction in feature space that could rapidly decrease the loss values. This direction is usually obtained by various combinations of the first order partial derivative or the second order partial derivative.

2.3. Random forest

As a typical ensemble learning method, it is one of the bagging models. Via integrating various weak classifiers, the results could be further improved.

2.3.1. Model. It leverages the CART decision tree as the basic block. Considering various trees could grasp different information, the variation of the entire model could be diminished. To lower the repeatability of the trees, pruning could be leveraged to distil informative features for learning.

2.3.2. Feature selection. Currently, information gain, Gini coefficient, and chi square test are commonly observed solution for selecting feature. This work exemplifies the Gini coefficient (GINI). During training, when each sub node reaches its highest purity, the Gini coefficient reaches its smallest value. Given a total of K classes probability p_k, then it is:

$$Gini(p) = \sum_{k=1}^{K} p_k (1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2 \tag{11}$$

2.3.3. Anti overfitting. First of all, as mentioned in Bagging's introduction, features are randomly selected from all features. It is helpful for reducing happening of overfitting. Secondly, unlike decision

DOI: 10.54254/2755-2721/16/20230878

trees, it has made improvements to the construction of decision trees. For a regular one, the optimal feature will be selected from features on the node to partition the subtrees. However, each tree actually selects a subset of features. Among these few features, selecting the optimal feature for the subtree partitioning expands the randomness effect and enhances the model's generalization capacity.

3. Result

3.1. Effectiveness of hyper-parameters in logistic regression

In this experiment, the effectiveness of the parameter 'C' and 'tol' in logistic regression are demonstrated in Table 1. It could be observed that with the increasing of 'C', the performances are getting better. However, the 'tol' has limited influences on logistic regression.

Table 1. Effectiveness parameter C and tol in logistic regression.

C	tol	mean_f1	std_f1	mean_auc	std_auc
0.01	0.0001	0.739	0.070	0.987	0.010
0.01	0.001	0.739	0.070	0.987	0.010
0.1	0.0001	0.909	0.052	0.9878	0.009
0.1	0.001	0.909	0.052	0.988	0.009
1	0.0001	0.926	0.042	0.988	0.008
1	0.001	0.926	0.042	0.988	0.008
10	0.0001	0.931	0.029	0.989	0.007
10	0.001	0.931	0.029	0.989	0.007

3.2. Effectiveness of hyper-parameters in random forest

The influences of the parameter 'max_depth' and 'n_estimators' in random forest are displayed in Table 2. It could be observed that the performances reach their optimal when the max depth is larger than 7.

Table 2. Effectiveness parameter max_depth and n_estimators in random forest.

max_depth	n_estimators	mean_f1	std_f1	mean_auc	std_auc
3	100	0.903	0.409	0.982	0.017
3	300	0.910	0.447	0.984	0.017
3	500	0.896	0.049	0.984	0.017
5	100	0.937	0.363	0.987	0.018
5	300	0.941	0.029	0.989	0.013
5	500	0.940	0.038	0.986	0.019
7	100	0.937	0.035	0.985	0.023
7	300	0.940	0.381	0.988	0.014
7	500	0.940	0.381	0.989	0.016
9	100	0.952	0.408	0.988	0.018
9	300	0.945	0.029	0.987	0.018
9	500	0.944	0.039	0.988	0.015
11	100	0.941	0.037	0.983	0.021
11	300	0.944	0.039	0.987	0.016
11	500	0.944	0.039	0.987	0.018

3.3. Result comparison

To further demonstrate the superiority of the logistic regression and random forest algorithms, more evaluation matrixes are introduced. In Table 3, indexes including precision, recall F1 and accuracy are leveraged for evaluation. Besides, in Figure 1, the ROC curve of the two models are displayed.

DOI: 10.54254/2755-2721/16/20230878

Table 3. Comparison of logistic regression and random forest.

	precision	recall	F1	accuracy
Logistic regression	0.897	0.968	0.931	0.947
Random forest	0.938	0.952	0.945	0.959

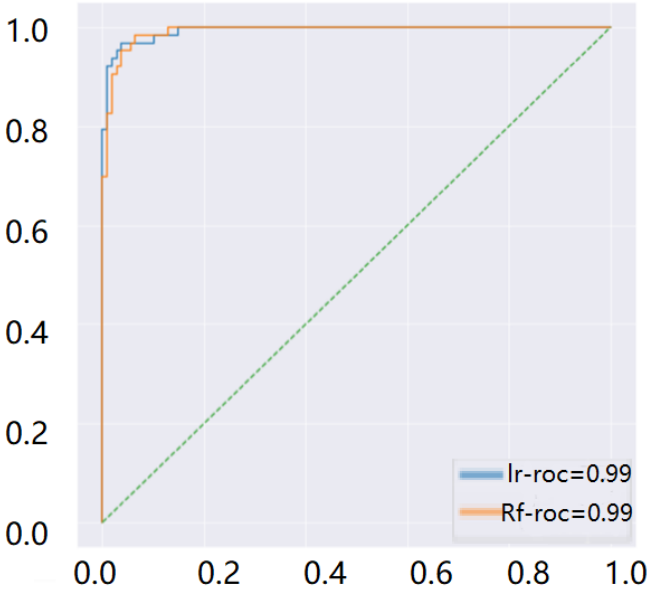

Figure 1. ROC curve of the logistic regression (blue line) and random forest (orange line).

4. Discussion

There are many benefits of the logistic regression algorithm, which are demonstrated as follows. Firstly, the structure is simple, with strong interpretability, and the influence of various features could be estimated by the corresponding weights in the model. Secondly, the training speed is rapid, and during classification, the computational workload is merely related to the feature dimensions, occupying less computational resources. Thirdly, the output results are easy to adjust and are all probability values, making it easy to adjust the threshold for classification. Fourthly, the performance is good, as long as the feature engineering is done well, the effect will not be too bad.

As for the random forest, due to the use of integrated algorithms, the inherent accuracy could outperform these individual algorithms. Its good performances during testing may derived from the randomness introduced to the system and the strong capacity to overcome overfitting. It is capable of resist noise in industry. Moreover, it has a rapid training speed and good adaptability to large-scale datasets. Moreover, it can tackle missing data without the extra assistance of processing. When out-of-pocket data (OOP) presence, it could also fit them well during testing. Different from other conventional machine learning models, the dependences between features could be omitted, and to learn the importance of features, which makes it distinguished from other models. Due to the independent and simultaneous generation of each tree, it could be parallelized.

5. Conclusion

In the current era of advanced medical technology and technology, cancer has become an important disease that affects the lives and health of modern people. The diagnosis of cancer is a complex process in itself, and the treatment in the later stage also has iterations and uncertainties. With the amazing capacity of computer technology, the continuous follow-up and deepening of machine learning theory, as well as the emergence of artificial intelligence and deep learning algorithms, all of these have

DOI: 10.54254/2755-2721/16/20230878

accelerated the development of the medical industry. Based on such a large environment, this paper applies machine learning related classification algorithm to the diagnosis and prediction of disease. For the classification of breast cancer, this paper first loads data and cleans them, explores and counts various types, uses the correlation thermogram to find and remove strong correlation variables to reduce errors, and uses feature engineering to numerically diagnose, then uses logical regression and random forest algorithms respectively. The final result is the accuracy of random forest algorithm is 95.9%, and the accuracy of logical regression is 94.7%, It can be seen that the classification accuracy of the two algorithms after training is higher than 90%, so it is feasible to use logical regression and random forest algorithm for diagnosing breast cancer.

References

[1] A. G. Waks, and E. P. Winer, "Breast cancer treatment: a review," *Jama,* vol.321, no.3, pp.288-300, Jan. 2019.

[2] M. Akram, M. Iqbal, M. Daniyal, et al., "Awareness and current knowledge of breast cancer," *Biol. Res.*, vol. 50, no.33, pp.1-23, Oct. 2017.

[3] B.Yang, G.Ren, E.Song, "Current Status and Factors Influencing Surgical Options for Breast Cancer in China: A Nationwide Cross‐Sectional Survey of 110 Hospitals," *Oncologist,* vol. 25, no.10, pp.e1473-e1480. Oct. 2020.

[4] Y. L.Li, Y. C. Qin, L. Y. Tang, et al., "Patient and care delays of breast cancer in China," *Cancer Res. Treat.,* vol. 52, no.3,pp.1098-1106. Jul. 2019.

[5] L.Wang, "Early diagnosis of breast cancer," *Sensors,* vol. 17, no.7, pp.1572. Jul. 2017.

[6] Y. S.Younis, A. H. Ali, O. K. S.Alhafidhb,et al., " Early diagnosis of breast cancer using image processing techniques," *Journal of Nanomaterials,* vol. 2022, pp.1-6.2022.

[7] M.Amrane, S.Oukid, I. Gagaoua, et al., "Breast cancer classification using machine learning," *in Electric electronics, computer science, biomedical engineerings' meeting,Apr.* 2018,pp. 1-4.

[8] N. Fatima, L. Liu, S. Hong, et al.,"Prediction of breast cancer, comparative review of machine learning techniques, and their analysis,"*IEEE Access,* vol. 8, pp.150360-150376. Aug. 2020.

[9] J. Chen, J. Jiao, , S. He, et al.,"Few-shot breast cancer metastases classification via unsupervised cell ranking," *IEEE ACM Trans. Comput. Biol. Bioinform.,*vol. 18, no.5, pp. 1914-1923. 2019.

[10] M. Yasser, Breast Cancer Dataset, 2022. [online]. Available: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.