Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230929

Breast cancer prediction and feature importance estimation leveraging logistic regression model

Ruining Han

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

Ruining.Han21@student.xjtlu.edu.cn

Abstract. In 2022, breast cancer become the largest cancer among the world. It is necessary to help people know how to prevent breast cancer. Machine learning use algorithms and statistic models to help the computer find some rules based on mass data. It can rapidly analyse, predict, and process the data in a timely enough manner. By significantly reducing errors caused by data duplication and other inaccuracies, it can simplify time-intensive documentation in data entry. In medicine, it can facilitate accurate medical prediction and diagnosis by quickly identifying high-risk patients, recommending a range of drugs, and predicting readmission rates. This paper will use Logistic Regression to build a model to predict the whether a patient will develop cancer and find the dominant factors from the characteristic values. The research found that 'Bare Nuclei', 'Clump Thickness', and 'Single Epithelial Cell Size' are the key factors that will create a large impact to the result of the predication.

Keywords: breast cancer prediction, logistic regression, machine learning.

1. Introduction

Based on data statistics from International Agency for Research on Cancer (IARC) in 2022, global increasement of new breast cancer patients is about 2.26 million, which took up almost 11.7 percent of the population of the cancer case that was 19.293 million [1,2]. With such number, breast cancer has surpassed lung cancer's 2.2 million cases to become the dominate cancer type of the world. Although the morbidity of breast cancer in China is low compared to it among the whole world, it still increases 420,000 new breast cancer patients every year, with an annual increase of 3 to 4 percent [3,4]. Hence, there exist the demand of the prediction of the breast cancer patient's probability.

Diagnosis depending merely on doctors is expensive and laborious. To achieve large-scale medical screening among people, an automatic solution is desperately required. Machine learning models, algorithms mapping input features to output predictions, is the major solution towards it [5,6]. This paper will build a model based on a dataset involving the breast cancer by using Logistic Regression to predict whether the patient has breast cancer when some characteristic values of the patient has been known, and find one or a few of crucial characteristic values then offer some advices.

DOI: 10.54254/2755-2721/17/20230929

2. Method

2.1. Dataset

The dataset used in this paper is from Machine Learning Repository of UCI. It is an aggregation of databases, data generators and domain theories [7]. This dataset consists of ten characteristic values, containing bare nuclei, single epithelial cell size, sample code number, normal nucleoli, clump thickness, uniformity of cell size, marginal adhesion, uniformity of cell shape, mitoses, and bland chromatin. Besides such characteristic values, this dataset contains a label value named 'class', which indicates whether the individual has the cancer. Note that in this dataset, some individuals do not have all characteristic values.

2.2. Logistic regression

This paper will use Logistic Regression to predict the result and analyse the problems.

Logistic Regression is a kind of generalised linear regression analysis model. It is able to deal with the binary classifications, which means that the outcomes of these problems can be returned into 1 and 0 or some other pairs of number, and it can be called binary logistic regression [8,9]. In some special cases, Logistic Regression can be used to solve the problems with more than two outcomes.

Sigmoid function is a basic function in logistic regression, whose formula is:

$$S(x) = \frac{e^x}{e^x + 1} \tag{1}$$

where e means the Euler's number, and x is the sum of all characteristic values in a linear combination. And Figure 1 illustrates the curve of Sigmoid.

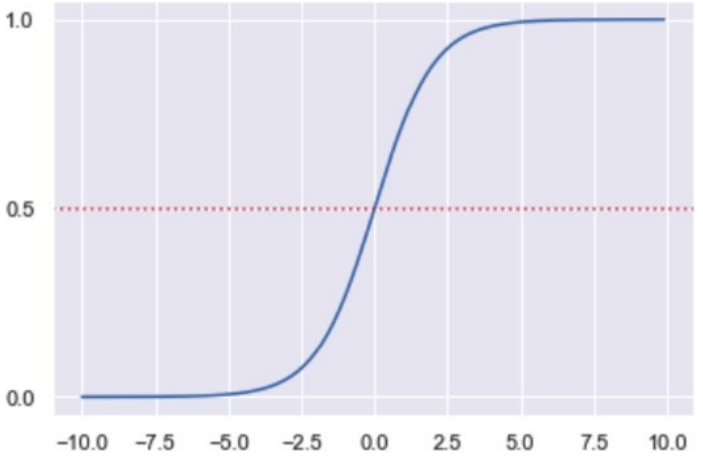

Figure 1. Curve of Sigmoid.

Sigmoid function maps the combination of characteristic to a number bounded from 0 to 1, which can be assumed as probability under some conditions. Under further assumptions, the outcome can only be 0 and 1. For instance, 0.5 can be chosen as a threshold, that is, assume the outcome is 1 if the value of Sigmoid function is larger than or equal to 0.5, and the outcome is 0 if the value of Sigmoid function is smaller than 0.5. Another threshold can be chosen in diverse situation.

Logistic Regression has its features. Firstly, Logistic Regression is a straightforward model. It can be expressed and implemented in a simple way. The calculated amount of Logistic Regression is little. It has high computing efficiency. And Logistic Regression only needs to store the characteristic values. At last, the functions involving in Logistic Regression are differentiable of any order. They have good mathematical analysis property.

DOI: 10.54254/2755-2721/17/20230929

Conversely, Logistic Regression also has some disadvantages. Logistic Regression is not sensitive to non-linear relationships. when the characteristic space is large, Logistic Regression does not have good performance. In general, Logistic Regression have low accuracy. It is prone to underfitting.

2.3. Evaluation index

This paper will use accuracy, precision, recall rate and F1-score (also called F1-value) to evaluate the results [10].

The combination of the reality and the predicted results exist four situations, denoted as TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative), where True and False mean whether the predicted result is the same as the reality, and Positive and Negative means the predicted result itself, corresponding to the 1 and 0. Table 1 shows four situations.

Table 1. Confusion matrix.

Reality\Predicted result	True	False
Positive	TP	FP
Negative	TN	FN

Accuracy: Accuracy counts the proportion of the correct prediction among the whole sample. In formula,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{TN} + \text{FN}} \tag{2}$$

To evaluate the unbalanced samples, precision and recall rate are introduced.

Precision: Precision counts the proportion of the correct prediction in the positive prediction. In formula,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall rate: Recall rate counts the proportion of the correct prediction in the true individuals. In formula,

$$\text{Recall rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Under some situation, there exists a contradiction to increase both precision and recall rate at the same time. It is needed to weigh precision and recall index. One possible way is to consider F1-score.

F1-score: It is calculated from the precision and recall rate, which can indicate comprehensive performance of the model in both sides. In formula,

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

From the formula, a larger F1-score indicates better performance. If one of these two values are closer to 0, which means the model is not good at this side, F1-score will be low.

3. Result

3.1. Experiment setting

This paper will do some experiment to serach for the characteristic values which have a great influence to the indicators.

The first step is to establish the dataframe, define the name of the characteristic values and delete the wrong data. Then the characteristic value and target value will be chosen.

This paper will take the whole set into a training and a testing set. And the testing one will include 25 percent of the whole dataset. Due to the differences of units, it is necessary to standardize the data to

DOI: 10.54254/2755-2721/17/20230929

cancel the negative effects before training the model. Next step is to use Logistic Regression to train the set and do some predication. Finally, use the formula to compute the indicators and evaluate the model.

Except for the normal group with all characteristic values, one of the characteristic values will be removed in one experiment.

The experiment would be repeated ten times and averaged to reduce the errors under the same condition.

3.2. Performance comparison

Table 2 is the result of the experiment. The first row of the table is the name of the removed characteristic value except for the column named 'No removed value', which means no characteristic value has been removed. And the second to the fourth rows display the Accuracy, Precision, Recall and F1-score separately. In every table cell the former number is coming from the patients with benign tumor, and the latter one is coming from the patients with malignant tumor.

Table 2. Performacnes after removing some features.

Removed feature	No	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion
Acc	0.97	0.95	0.94	0.96	0.96
P (0/1)	0.98/0.97	0.97/0.89	0.96/0.92	0.97/0.95	0.98/0.93
R (0/1)	0.97/0.96	0.95/0.94	0.94/0.94	0.97/0.95	0.97/0.96
F1 (0/1)	0.98/0.96	0.96/0.91	0.95/0.93	0.97/0.95	0.97/0.95
Removed feature	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
Acc	0.94	0.94	0.94	0.96	0.94
P (0/1)	0.94/0.95	0.95/0.92	0.96/0.91	0.98/0.93	0.96/0.92
R (0/1)	0.97/0.88	0.96/0.90	0.94/0.94	0.95/0.97	0.95/0.92
F1 (0/1)	0.96/0.91	0.95/0.91	0.95/0.92	0.97/0.95	0.96/0.92

4. Discussion

This paper will compare the indicators from the experiment, then gives some possible causes and the suggestions. The normal group has the highest Accuracy which is 0.97, but several groups share the lowest Accuracy which is 0.94. To get more precise information, the other indicator will be considered.

In consideration of that in the real life, the predication of the malignant tumor will be emphasised (because the correctness of the predication of the benign tumor will not change the result that the patient does not have cancer, even if this patient has been diagnosed by mistakes, he can also do other examinations to check), this paper will mainly focus on the indicators coming from the patients with malignant tumor.

Firstly, the normal group has the highest precision which is 0.97. In contrast, the precision of the group without 'Clump Thickness' merely approach 0.89, which creates a gap from the normal group. This gives evidence that 'Clump Thickness' may be the greatest factor of influence of the value of precision. Besides that, the group without 'Bland Chromatin' approaches 0.91, which can also be considered as one of the

Then, from the line of recall rate, it is known that the normal group has the recall of 0.96. The lowest recall is 0.88, which is coming from the group without 'Single Epithelial Cell Size'. In addition, the group without 'Bare Nuclei' whose recall rate is 0.90 does not have a good performance in this indicator.

Finally, F1-score should be considered as well. The normal group's is equal to 0.96, and some groups share the same lowest value which is 0.91, including the groups without 'Bare Nuclei', 'Clump Thickness' and 'Single Epithelial Cell Size'. Although there is not much difference in numbers, this indicator is still valuable.

DOI: 10.54254/2755-2721/17/20230929

Under comprehensive consideration, this paper believes that 'Bare Nuclei', 'Clump Thickness', and 'Single Epithelial Cell Size' are the most important influencing factors. Based on the consequence, this paper will give some suggestions. First, people should check their diets. They need to control the absorb quantity of the food which contains too much estragon such as the bean products and take some food with high-quality protein. Secondly, people can have some regular exercise, if necessary, they can also get massages to boost their metabolism. For the people over 40 years old who is not at high risk for breast cancer, they are recommended to have the mammary gland molybdenum target once a year, and for the dense mammary gland (that is, these people have C-type breast or D-type breast by examination), combining with the B-ultrasound examination will be a good suggestion. And for the breast cancer high risk group, besides the yearly mammary gland molybdenum target, they are suggested to have Breast ultrasound every 6 to 12 months. They should take dynamic contrast-enhanced magnetic resonance imaging once a year if necessary.

What needs illustration is that, the dataset used in the experiment is coming from an academic website, so there exist some limitations in this paper. Primarily, the dataset only mentioned nine characteristic values. In the real application, other variables can also affect the result for instance the history of illness of the patient and genetic disease. The second one is that the size of the dataset is too small, it only contains 699 pieces of data, of which only 683 are valid. This number may be enough to train a model, but cannot satisfy the needs of real application. The other one is that this paper has mentioned that Logistic Regression has some problems itself, so it may exist some other models which is more suitable for this predication.

5. Conclusion

This paper employed Logistic Regression to build a model used in prediction of whether the patient will have breast cancer. Then one of the characteristic values was selected and removed away to contribute a new experimental group, which would be repeated ten times to get the average of the indicators. This paper compared diverse groups and the normal group which has not been changed to get more useful information, and found that 'Bare Nuclei', 'Clump Thickness', and 'Single Epithelial Cell Size' are the dominant factors in this predicted model. This paper used the logical indicators to evaluate the characteristic values, then depending on the results, put forward some advices such as to vary the recipe to adjust the intake of food, conditioning the body in several manners, or do the medicine examinations orderly, which will help people to prevent cancer. In future, the writer wants to select other datasets that include multitudinous characteristic values, then do different experiments to find other dominant factors involving breast cancer to expand the conclusion. In addition, the writer also wants to use more analysis models except Logistic Regression to do experiments and determine whether the result will be different under variety of analysis models.

References

[1] Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., et, al. (2021). Global patterns of breast cancer incidence and mortality: A population‐based cancer registry data analysis from 2000 to 2020. Cancer Communications, 41(11), 1183-1194.

[2] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 149(4), 778-789.

[3] Li, Y., Abudureheiyimu, N., Mo, H., Guan, X., Lin, S., et, al. (2022). In real life, low-level HER2 expression may Be associated with better outcome in HER2-negative breast cancer: a study of the national cancer center, China. Frontiers in oncology, 11, 5482.

[4] Zheng, Y., Dong, X., Li, J., Qin, C., Xu, Y., et, al. (2022). Use of Breast Cancer Risk Factors to Identify Risk-Adapted Starting Age of Screening in China. JAMA Network Open, 5(11), e2241441-e2241441.

[5] Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology, 23(1), 40-55.

DOI: 10.54254/2755-2721/17/20230929

[6] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR), 54(2), 1-36.
[7] Khan, M. M. R., Arif, R. B., Siddique, M. A. B., & Oishe, M. R. (2018). Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology, 124-129.
[8] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.
[9] Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301.
[10] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., et, al. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. American Journal of Roentgenology, 212(1), 38-43.