Transactions on Cancer (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/tranc.2023.040119
ISSN 2523-6482 Vol. 4 Num. 1

Research on Prediction of Breast Cancer Based on BP Neural Network

Wenjing Li[1,*], Weihan Wang[2]

[1]Software College, Taiyuan University of Technology, Taiyuan, Shanxi, 030000, China
[2]Maynooth International Engineering College, Fuzhou University, Fuzhou, Fujian, 350112, China
[*]Corresponding author: 2041310644@qq.com

Keywords: Breast cancer, Disease prediction, BP neural network, machine learning

Abstract: Breast cancer, as the most common cancer in women worldwide, poses a great threat to human life. Therefore, the prediction of benign and malignant breast cancer is particularly important. In this paper, from the perspective of machine learning model, BP neural network model is used to predict the incidence of breast cancer. In the fitting of the neural network, the regression R value is selected as an observation index, and the R value is 0.99126. The fitting degree of the model is good, which proves that the model has better availability. The model can be applied to the prediction of benign and malignant breast cancer in hospitals, and the idea of the model can also be used for the prediction of other diseases.

1. Introduction

Breast Cancer is a phenomenon of uncontrolled proliferation of breast epithelial cells under the influence of various carcinogens. In the late stage of cancer, with the metastasis of cancer cells, multiple organs will appear lesions, which directly threatens the life of patients. It is mainly divided into two categories: carcinoma in situ and invasive carcinoma, and there are a few rare cancer types. Breast cancer is more common in women, so it is also known as the "pink killer", and less common in men. According to the 2020 Global Cancer Statistics Report[1] , breast cancer has become the most common cancer in women worldwide. In 2017, a total of 6 972 new cases of female breast cancer were diagnosed in Guangdong cancer registration areas, with a crude incidence rate of 51.32/100 000 [2]. Breast cancer has also been ranked the first in the incidence of female cancer in China [3], and the fourth in the death of female cancer in China [4]. Predicting benign and malignant tumors may help to treat the disease in time and decrease the death rate. Therefore, the prediction of benign and malignant breast cancer is particularly important.

At present, whether domestic or international, the research on breast cancer mainly focuses on the intervention of various drugs on breast cancer cells and the proliferation of breast cancer cells. There is still room for exploration in the field of building data models through machine learning to predict the disease of breast cancer.

In existing machine learning, neural networks are applied in various fields. As a commonly used artificial neural network at present, BP neural network has multiple structures such as input layer, output layer and hidden layer. The key lies in the selection of input layer and the setting of output

layer [5]. The prediction of benign or malignant breast cancer involves multiple features, and the data types are huge. BP neural network just compares the results with the expected value through the classification and processing of each feature, and correcates the error through backtracking [6]. Therefore, this paper uses BP neural network to predict the prevalence of breast cancer. This paper makes a prediction of breast cancer through the analysis of the relevant characteristics of breast cancer data. The prediction results can provide a basis for the hospital to judge the patient's breast cancer, and also can treat serious patients in time to reduce mortality. At the same time, the idea of the BP neural network prediction model for breast cancer can also provide a constructive basis for the prediction of other types of tumors.

2. Data introduction

2.1 Data sources and characteristics

The dataset for this article comes from the Breast Cancer Wisconsin diagnosis by Dr. William, which has 17,070 entries.

The prevalence of breast cancer is closely related to the morphology of cancer cells, so the data in this paper are calculated from digital images of fine needle aspiration of breast masses, and they describe the nuclear features present in the images. The data were given an ID number for each image, which was diagnosed using a multi-surface method tree to test whether it was benign or malignant. Radius (median distance from centre to circumference), texture (grey standard deviation), Perimeter, area, smoothness (local change in radius length) and density (circumference ^ 2) have been calculated for core/area 1.0, Concavity (depth of concave section of outline), concavity (number of recess in outline), Symmetry, fractal dimension ("approximate approximation" -1), and the mean, standard error, and the "worst" or the largest (mean of the three highest values) were computed for each image, resulting in 30 features.

2.2 Data preprocessing

Since the magnitude difference between different data is too large, the results may be affected by extreme data, so the data should be normalized first. The data were filtered and divided into two categories: benign cells (denoted by B) and malignant cells (denoted by M). The benign cells and malignant cells were processed separately, so that the input data were mapped into the interval [0,1]. The normalization formula used in this paper is as follows.

$$\hat{x}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (i = 1,2,3,\dots,212) \tag{1}$$

$$\hat{x}_j = \frac{x_j - x_{min}}{x_{max} - x_{min}} \quad (j = 1,2,3,\dots,357) \tag{2}$$

Here, x_i, x_jrepresent the original input data, i, jrepresent the numbers of M and B, $\hat{x}_i, \hat{x}_j$represent the normalized input data, x_{min}, x_{max}represent the minimum and maximum values of the original input data, respectively.

3. Introduction to the basic concepts of BP neural network

3.1 Principle of BP neural network

BP neural network is a multilayer feed-forward network trained according to the error back propagation, and its algorithm is called BP algorithm. By Object To be studied, this paper establishes a neural network with one hidden layer, The BP neural network structure is shown in Figure 1.Its

learning rule is to continuously adjust the weights and thresholds of the network through back propagation using gradient descent to minimize the mean square error between the actual output value and the desired output value of the network.

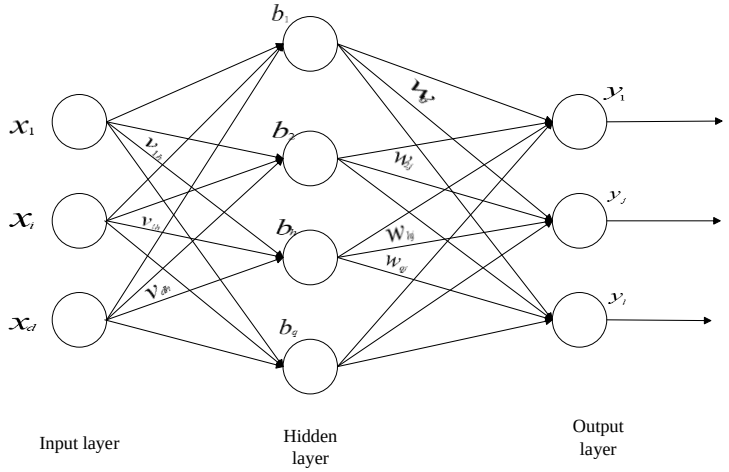

Figure 1: Structure diagram of BP neural network model

In the BP neural network model, each neuron receives input signals from other neurons, and each signal is passed through a weight. The neuron adds up all the received signals to obtain a total input value, which is then compared with the threshold of the neuron, and processed through an "activation function" to obtain the final output. This output will serve as an input for the next neuron.

In this paper, the activation function relu is selected when building a neural network model to predict the benign and malignant of breast cancer, and its formula is as follows.

$$f_{(x)} = \max(0, x) \tag{3}$$

The input of the HTH hidden layer neurons in Figure 1 is$\mu_h = \sum_{i=1}^{d} v_{ih} x_i$, and the output of the JTH neuron is$\varepsilon_j = \sum_{h=1}^{q} w_{hj}\, b_h$. The input layer reads the input data, the hidden layer processes the data, and the output layer outputs the results. Where v, w are the weights from input layer to hidden layer and hidden layer to output layer, respectively.

We implement forward propagation by multiplying the input values of each layer by the weights plus the activation function (θ), which can be calculated using the following two formulas:

Input layer to hidden layer: $\mu_h = \sum_{i=1}^{d} v_{ih}\, x_i + \theta_h$

Hidden layer to output layer: $\varepsilon_j = \sum_{h=1}^{q} w_{hj} b_h + \theta_j$

Because the parameters are random, there is a large error between the results of one calculation and the real results. We need to adjust the parameters according to the error to make the parameters better fit, until the error reaches the minimum value. Backpropagation of the model is performed at this point.

When carrying out back propagation, the error should first be calculated. The formula of error is $E = \frac{1}{2}\sum_{k=1}^{2}(y_k - T_k)^2$, and then the gradient descent method is used to update the parameters of the model. The inverse weight update is performed using$\Delta\omega_{ij} = (l) E_{y_k}, \omega_{ij} = \Delta\omega_{ij}$, where l represents the learning rate, and a suitable learning rate can make the objective function converge to a local minimum in a suitable time.

The above completes a training of the BP neural network model, and the specific flow chart is shown in Figure 2.

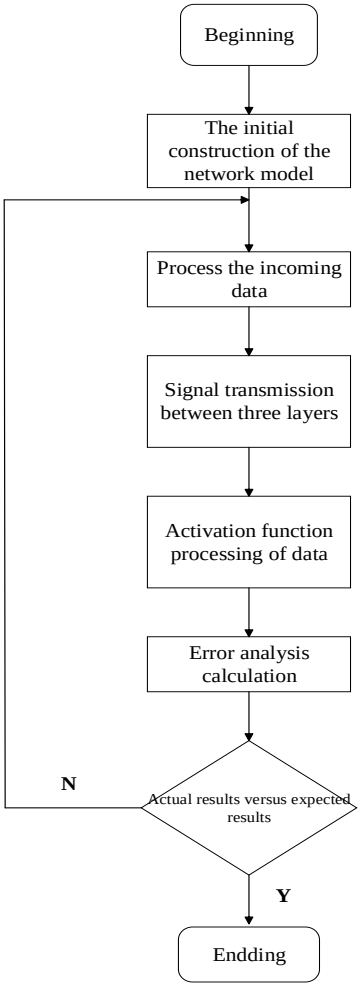

Figure 2: Flow chart of BP neural network algorithm

3.2 Selection of model parameters

BP neural network is divided into input layer, hidden layer and output layer, and the data of these three levels should be determined separately. The size of the input layer is 30, with 200 nodes; The size of the hidden layer is 10 and there are 40 nodes. The size of the output layer is 30 with 200 nodes. The training dataset was 200*0.7=140, and the validation and testing dataset was 200*0.15=30.

In this paper, matlab is used to implement the model. First, the function is used to calculate the weight of input layer to hidden layer, the threshold of input layer to hidden layer, the weight of hidden layer to output layer, and the threshold of hidden layer to output layer. Then the transfer function of the input layer to the hidden layer and the transfer function of the hidden layer to the output layer are used to calculate the transfer of data between the three layers.

The iteration of this round of testing finally stopped at the 13th.

3.3 Construction of neural network model

According to the characteristics of the model and the selection of breast cancer data, the process of the model is shown in Figure 3.

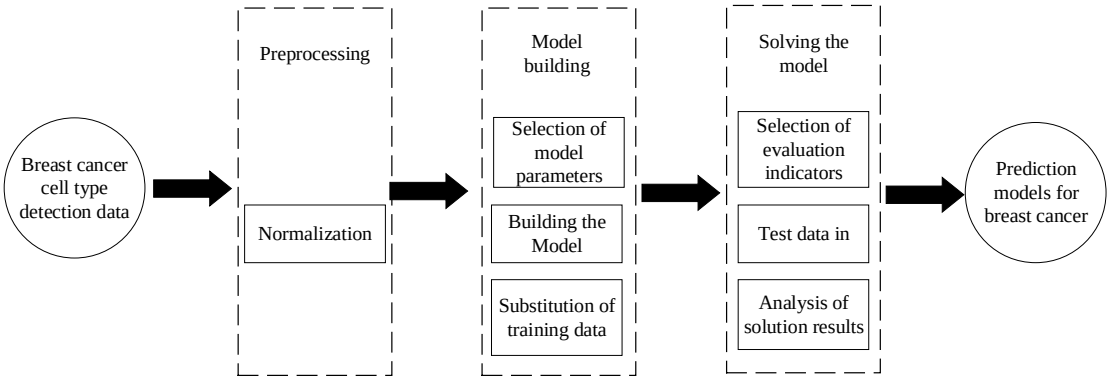

Figure 3: Flowchart of model building

BP neural network is divided into three parts: input layer, output layer and hidden layer. It is a multi-layer feedforward network. According to the need of breast cancer prediction, the simplest three-layer network model is selected.

The first is to preprocess the obtained data. There are many influencing factors of the data. In order to be more suitable for the prediction of the breast cancer model, the irrelevant factors are firstly excluded to make the selected data more relevant. In this paper, the selected data have different numbers of benign cells and malignant cells. In order to make the values of the two groups of observations equal, the two groups of data are normalized, and 6000 data are screened out for comparison with the area feature as the main reference standard.

Then the model was established. The establishment of the model is divided into three parts: the selection of model parameters, the construction of the model, and the substitution of the training data. The parameters are selected and normalized to the data, and then the 6000 data of the two groups with 30 features are processed, and the processed data are trained to obtain the results after training. The training of the model is reflected to some extent by the mean square error and regression R value. The mean square error represents the expected value of the difference between the predicted output and the target output, where a lower value is better and 0 indicates no error. The regression R value represents the correlation between the forecast output and the target output. The closer the R value is to 1, the closer the relationship between the forecast and the output data is, and the closer the R value is to 0, the greater the randomness of the relationship between the forecast and the output data is. Therefore, when solving the model in this paper, the mean square error and regression R value are selected for analysis, and substituted into the test data for training. The training progress is shown in Table 1. The analysis of the results is shown in Table 2.

Table 1: Analysis table of test results

Units Value	Initial Value	Stop Value	Target
Round	0	13	1000
Duration	-	00:00:05	-
Performance	177000	59.8	0
Gradient	617000	469	0.0000001
Mu	0.001	0.1	1000000000
Validation	0	6	6

Table 2: Analysis table of training results

	Observations	MSE	R
Training	140	919.9060	0.9956
Verify	30	4952.6	0.9730
Test	30	2765.5	0.9873

3.4 Model performance evaluation

The input data were divided into three sets, namely training set, validation set and test set. The training set was used to fit the model, adjust the parameters, and select the input variables. The validation set plays the role of verifying the stability, robustness and generalization error of the model after the training is completed. The test set is to evaluate the effect of the trained model and verify whether the model is overfitting or underfitting. The results obtained by training the filtered data using the Levenberg-Marquardt training algorithm are shown in Figure 4. It can be seen from the figure that the best verification performance is 4952.5724 in the 7th round, and the verification line coincides with the best line at this time.

Regression fit results for the three sets of the model are shown in Figure 5. The selected evaluation index is R, and the closer R is to 1, the better fit the model is. As can be seen from the figure, the R value of the training set is 0.99561, the R value of the validation set is 0.97295, the R value of the test set is 0.9873, and the R value of the whole data is 0.99126, indicating that the prediction model has a good tracking ability of the samples.

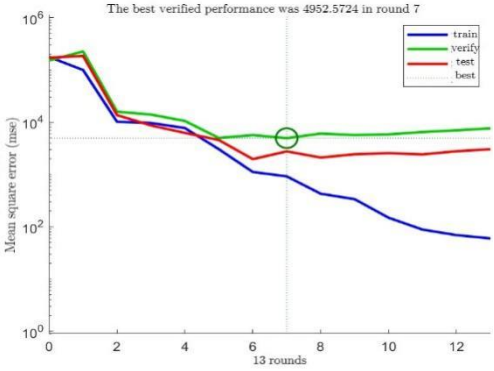

Figure 4: Performance curve of the model

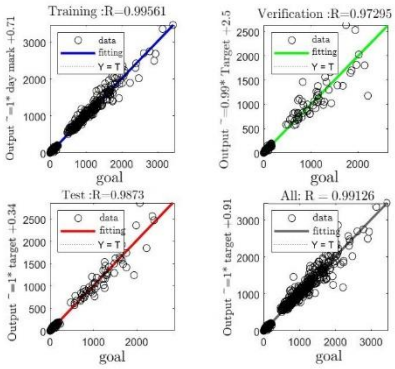

Figure 5: Model regression plot

3.5 Analysis of the results of solving the model

The resulting error histogram is shown in Figure 6. It can be seen from the figure that this training has an error of 20 bins. The errors for the training set are clustered at instance 4010, for the validation set at 4851, and for the training set at 5700. The record of the training state is shown in Figure 7. When the gradient is 469.1194, Mu=0.1, and the validation check is 6, the calculations are all carried out to the 13th round.

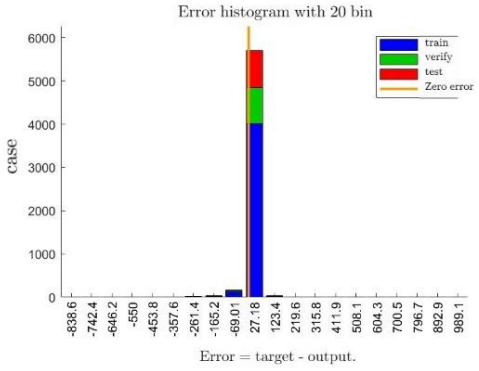

Figure 6: Histogram of model solution error

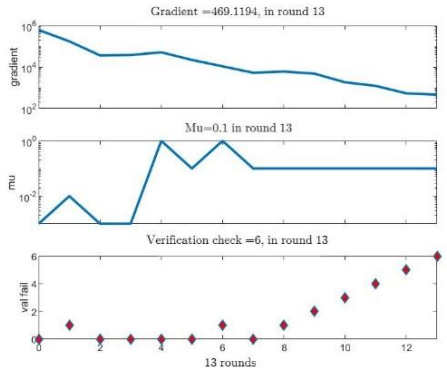

Figure 7: Diagram of the training state

4. Conclusion

4.1 Validity and Limitations of Conclusions

Based on the benign and malignant data of breast cancer cells, this paper builds a BP neural network model to fit the data, and explores whether it can be applied to the prediction of breast cancer. Through the performance analysis and solution of the model, it is found that the R value of the model is 0.99126, and the fitting degree is good. The R value is 0.9969 and the MSE is 101.9911, which indicates that the model has a good application in the field of breast cancer prediction.

4.2 Develop space for discussion

This model can be applied to the prediction of benign and malignant breast cancer in hospitals, which can reduce the mortality of human because of breast cancer to a certain extent, so that people with breast cancer can be treated in time. At the same time, this machine learning method also provides new ideas for the prediction of other diseases, which can be widely used in the prediction of other diseases.

However, this experiment also has the disadvantages of small data selection, and there are also other factors affecting the data, and there are certain limitations in the establishment of the model. In terms of model establishment, on the basis of this model, GA algorithm can be considered to optimize the model to obtain a better model, and the parameters can also be adjusted more finely to obtain the optimal solution.

References

[1] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2021, 71(3): 209-249.

[2] Li A J, OU M Y, Xu Y J, et al. Epidemiological characteristics of female breast cancer in cancer registration areas of Guangdong province in 2017 and its changing trend from 2013 to 2017 [J/OL]. Chin J Cancer: 1-6[2023-10-08].

[3] He J, Wei W Q. 2019 Chinese Cancer Registry Annual Report [M]. Beijing: People's Medical Press, 2021. (in Chinese with English abstract)

[4] Zhang Yacong, Lv Zhangyan, Song Fangfang, et al. Trend of incidence and mortality of breast cancer in China and in the world [J]. Journal of Oncology, 2021, 7(2): 14-20.

[5] Yu Meixia, Dong Gang, Xu Ruxun et al. Precision Time Base Source Calibration and Prediction Model Based on BP Neural Network [J/OL]. China Testing :1-7[2023-10-08].

[6] Zeng Qingyang, Ding Chuheng, Gu Zhanying et al. Oil-tea production prediction model based on BP neural network to build [J]. Journal of economic studies, 2022, 40 (03): 87-95. The DOI: 10.14067 / j.carol carroll nki. 1003-8981.2022.03.009.