ISSN 2618-1584 Vol. 4, Issue 7: 48-51, DOI: 10.25236/FMSR.2022.040708

# Study on Quantitative Prediction of Anticancer Compound Activity Based on Neural Network Model

**Songhao Lu, Xinxin Fu, Mingzhou Chen, Xuezheng Yue***

*School of Materials and Chemistry, University of Shanghai for Science and Technology, Shanghai, 200093, China*
**Corresponding author*

***Abstract:*** *Breast cancer poses a serious threat to global women's health. Around this problem, a quantitative prediction model of biological activity based on neural network and a quantitative prediction model of multiple linear regression are established in this paper. 20 kinds of molecular descriptor data with the most significant effect on biological activity of 1974 compounds were selected as input layer and 1 biological activity data as output layer, and a quantitative prediction model of biological activity based on neural network was established. Based on the quantitative prediction model based on parameter setting, the single hidden layer BP neural network structure is used to test the model, and the fitting error R values of the training set and test set of the model are close to 1, and the trend in the scatter diagram is obvious, which shows that the prediction value of the model is more accurate.*

***Keywords:*** *Cancer, BP neural network, Model checking*

## 1. Introduction

In recent years, the incidence and mortality of breast cancer in women are increasing year by year, which seriously threatens the health of women worldwide. Therefore, it is necessary to study and analyze the candidate drugs for breast cancer treatment, especially for ERα. In patients with positive breast cancer, the type of patients is large and the mechanism is complex [1]. Erα therapy can become the candidate drugs for positive breast cancer patients can not only ER α It also has good pharmacokinetic properties and safety. The common analytical method is to establish the quantitative structure-activity relationship of compounds, and screen the possible active compounds or optimize the structure of known compounds.

## 2. Data processing

In this paper, 729 molecular descriptors of 1974 compounds were searched and analyzed, and it was found that all the molecular descriptors were 0, which caused some interference to the subsequent model establishment [2]. Therefore, it is necessary to eliminate these molecular descriptors. Through the preprocessing of data elimination, each compound has a total of 504 corresponding molecular descriptors, which are used for the following variable screening, elimination and variation of residual molecular descriptors which have the most significant effect on biological activity.

## 3. Quantitative prediction model based on BP neural network

### *3.1. Determine training set and test set*

The variables are maxaan, minaan, maxaaac, MINDO, Maxdo, maxhcsatu, alogp, maxsch3, minsch3 and ETA_ dEpsilon_C, nF10Ring, nT10Ring, n5Ring, nT5Ring, minHCsatu, minHBa, ETA_dAlpha_B, BCUTc-1h, ETA_ dEpsilon_A, ETA_dEpsilon_ D. It is used as the input layer data of neural network [3]. The output layer data is 1974 rows and 2 columns, and the variables are IC50 respectively_ nM, pIC50. In this question, the first 1580 lines of data are set as training data, and the last 394 lines of data are test data, that is, under the screening of 20 molecular descriptor variable indicators, for the 1974 group of data provided by the file "molecular_descriptor. Xlsx", the data numbered 1-1580 are BP neural network training data (80%), and the data numbered 1581-1974 are test data (20%). According to the above setting, the dimensionless processing of data is carried out according to formula 31, that is

ISSN 2618-1584 Vol. 4, Issue 7: 48-51, DOI: 10.25236/FMSR.2022.040708

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Where, $x_{min}$ is the minimum value of the sample and $x_{max}$ is the maximum value of the sample

### *3.2. Setting of model parameters*

Using the neural network toolbox neural net fitting of MATLAB software to predict, it is necessary to set the important parameters in the quantitative prediction model of biological activity based on BP neural network, which are the number of hidden layer neurons in the training set, the rules of training algorithm, the number of iterations and the target of training error [4]. In view of the small amount of data provided in this question, the number of iterations of model parameters is set to 1000 and the training error target is 0.5 0000001.

Firstly, set the selection proportion of training set, verification set and test set in the training data to 0.5% respectively 7,0.15,0.15. The training algorithm is Levenberg – Marquardt. Then set the number of neurons in the hidden layer to 10, 15 and 20 respectively. On this basis, the test errors of the number of neurons in different hidden layers are discussed [5]. The number of neurons in the hidden layer is determined by comparing the error between the predicted value and the real value. The mean square error (MSE) is the average value of the square sum of the difference between the predicted value and the real value [6]. The calculation formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[ f\left(x_i\right) - y_i \right]^2 \tag{2}$$

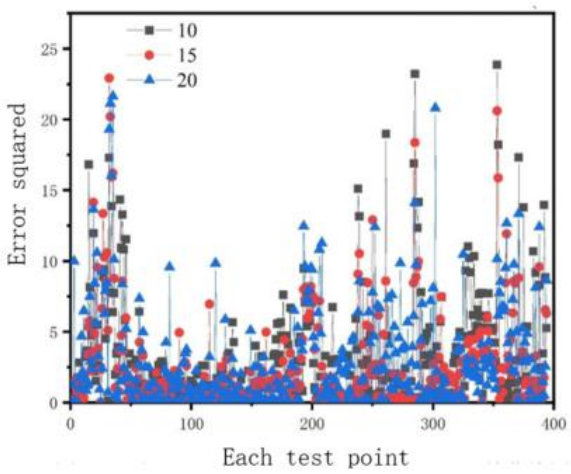


*Figure 1: Error curve of each test point for the number of neurons in different hidden layers*

It can be seen from the error curves of test points with different numbers of neurons in the hidden layer that the error curves of test points with different numbers of neurons are basically consistent, but the error curves with 15 and 20 neurons in the hidden layer are more consistent. It is obvious from table 7 that when the number of neurons in the hidden layer is 15, the mean square error MSE is the smallest. Therefore, the number of neurons in the hidden layer is set to 15.

On the basis of setting the number of hidden layer neurons to 15, Levenberg – Marquardt, Bayesian regularization and scaled converge gradient training algorithms are set respectively.

To sum up, the parameters of the biological activity quantitative prediction model based on neural network are as follows: the number of neurons in the hidden layer is set to 15, the training algorithm is set to scaled conflate gradient, the number of iterations is set to 1000, and the training goal is set to 0 0000001

## 4. Model test

Multicollinearity processing method: because we don't care about the specific regression coefficient, but only about the ability of the whole equation to predict the explained variables, we can usually ignore

ISSN 2618-1584 Vol. 4, Issue 7: 48-51, DOI: 10.25236/FMSR.2022.040708

multicollinearity. The multiple linear regression model constructed above is used to predict the compound samples with serial number of 1581-1974. The error curve of each test point is relatively flat, as shown in Figure 14. The curve between predicted value and true value of compound samples with serial number 1581-1974 is shown in Figure 3.

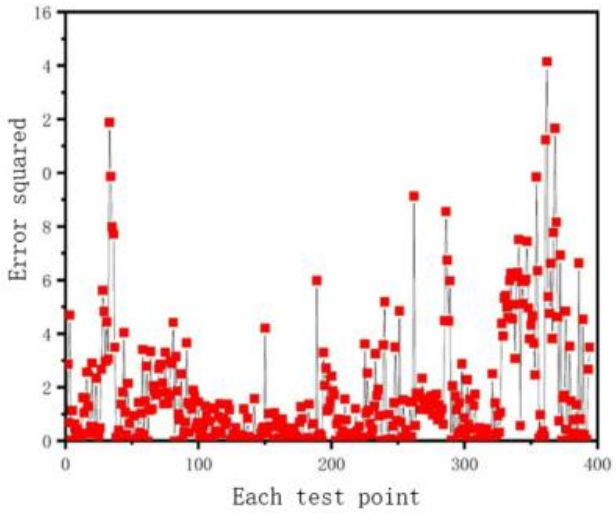


*Figure 2: Error curve of each test point under multiple linear regression analysis*

It can be seen that the error of each test point predicted in the multiple linear regression classification model is mostly within 2, indicating that the prediction effect of the model is very good.

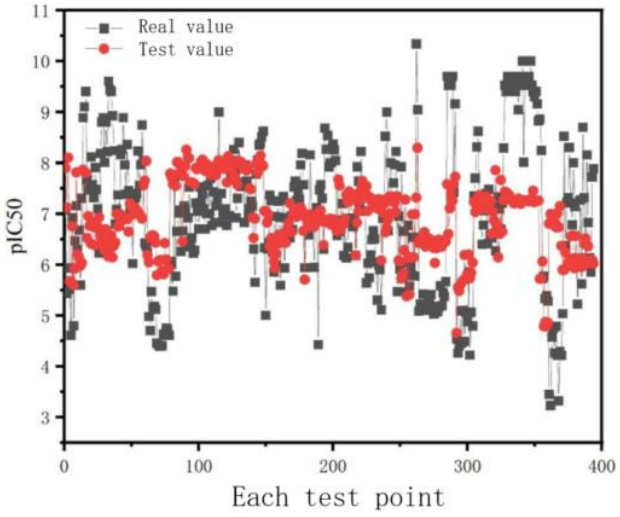


*Figure 3: Real value and predicted value curve*

In the multiple linear regression classification model, the predicted value of bioactivity index predicted is basically consistent with the real value, and the error is small, indicating that the fitting accuracy of the model is high.

## 5. Test of quantitative prediction model

The parameters of the biological activity quantitative prediction model based on neural network are set, that is, the number of neurons in the hidden layer is 15, the training algorithm is scaled conflate gradient, the number of iterations is set to 1000, and the training goal is 0 0000001, input 20 molecular descriptors 15811974 into the neural network model to predict the bioactivity index. Using the single hidden layer BP neural network structure, there are mainly 20 inputs (molecular descriptors that have the most significant impact on biological activity), one hidden layer (15 neurons) and one bioactive output.

According to the prediction value of single hidden layer BP neural network model, the fitting error of the model is analyzed

ISSN 2618-1584 Vol. 4, Issue 7: 48-51, DOI: 10.25236/FMSR.2022.040708

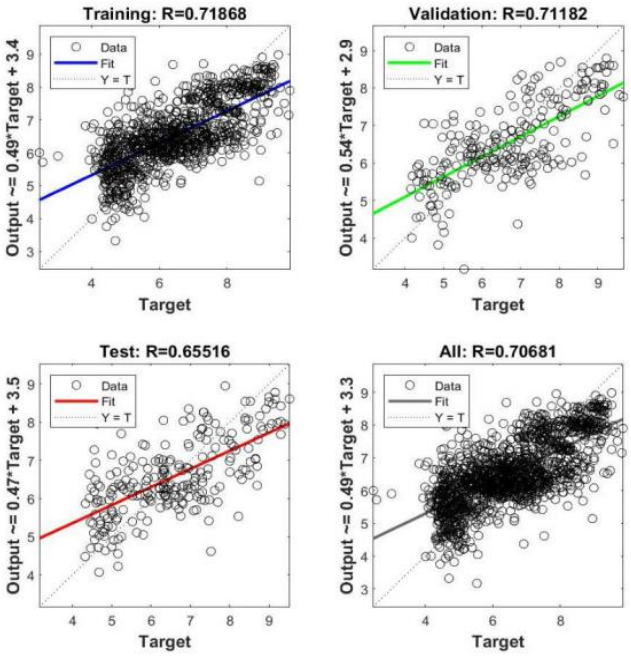


*Figure 4: Fitting error*

Fitting error diagram it can be seen from the fitting error diagram in Figure 17 that the training set and test of the model

## 6. Conclusion

In this paper, the quantitative prediction model of biological activity based on neural network is studied, and the quantitative prediction model of biological activity based on neural network is established. the parameters of the model are set as the number of hidden layer neurons is 15. Based on the quantitative prediction model based on parameter setting, the single hidden layer BP neural network structure is used to test the model, and the fitting error R values of the training set and test set of the model are close to 1, and the trend in the scatter diagram is obvious, which shows that the prediction value of the model is more accurate. Through the above prediction, we can clearly judge and understand the factors that affect the compounds that affect cancer.

## Acknowledgements

The authors are grateful for funding of Shanghai Sailing Program (Grant Number 19YF1434300) and Shanghai Engineering Research Center of High-Performance Medical Device Materials (No. 20DZ2255500)

## References

*[1] Liu Zuwang, Wang Yumei Meta analysis of dietary fiber and carbohydrate intake associated with breast cancer risk [J]. China health statistics, (03): 464-4672015.*
*[2] Zhang Cuifeng, Xie Haitang, pan Guoyu Absorption, distribution, metabolism, excretion and toxicity characteristics of macromolecular drugs and application of pharmacokinetic model [J]. Journal of pharmacy, (8): 1202-1208, 2016.*
*[3] Burden F R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix[J]. Quantitative Structure-Activity Relationships, 16(4): 309-314, 1997.*
*[4] Lin Wenfu. Econometrics [M]. Shanghai University of Finance and Economics Press, 2005.*
*[5] Rumelhart D E, Hinton G E, Williams R J. Learning Internal Representation by BackPropagation Errors [J]. Nature, 323: 533-536, 1986.*
*[6] Li Jie. Improved particle swarm optimization optimization support vector machine for project cost prediction [J]. Computer system application, 25 (006): 202-2062016.*