Optimal Modeling of Anti-Breast Cancer Candidate Drugs Based on SSA-BP

Biao Xu

School of Electronics and Control Engineering, Chang'an University, Xi'an, Shaanxi, 710064, China

Keywords: Molecular descriptor, Dimension reduction, Qsar model, Bp neural network, Sparrow optimization algorithm

Abstract: At present, in the research of candidate drugs of anti-breast cancer, the method of establishing quantitative structure activity relationship (QSAR) of compound activity is usually used to screen important compound molecular descriptors in order to save time and cost. Considering the nonlinear relationship between molecular descriptors and biological activities of compounds, neural network prediction models with strong nonlinear mapping ability and high accuracy have been widely used. However, there is no consensus on the best algorithm for QSAR modeling. In this paper, 11 molecular descriptors were firstly screened by nonlinear dimension reduction technique, and 11 molecular descriptors which have significant influence on the biological activity of compounds were selected. What's more, QSAR models were constructed based on three traditional neural network models (BP neural network, Elman neural network and wavelet neural network) and neural network model improved by optimization algorithm (SSA-BP neural network). The results showed that the prediction error of SSA-BP neural network was the lowest, the mean square error was only 0.04898, and R2 reached 0.94572. Compared with the suboptimal BP neural network, MAE value decreased by 15.3%, MSE value decreased by 25.3%, MAPE value decreased by 37.4%, which indicates that BP neural network can predict the biological activity of compounds more accurately, and the optimization model can further improve the prediction performance of BP neural network. It is helpful to better screen efficient compound molecules and guide the structural optimization of existing active compounds and development of drugs for treating breast cancer.

1. Introduction

Breast cancer is one of the most common cancers with high mortality in the world, and it is also the cancer with the highest incidence among Chinese women. According to statistics, in 2014, China accounted for 12.2% of newly diagnosed breast cancer cases and 9.6% of global breast cancer cases, and the incidence and death of breast cancer were on the rise [1]. Recent clinical medical experiments have shown that molecular analysis and target recognition can effectively select anti-breast cancer drugs. Estrogen receptor alpha (Erα) is expressed in 50%-80% of breast tumor cells, so it is regarded as an important target for the treatment of breast cancer.

In the research and development of anti-breast cancer drugs, the molecular structure descriptor is usually used as independent variable and the biological activity value of ERα is used as dependent variable, and the QSAR model of compounds is established [3]. It can save time and development costs and improve economic benefits by predicting the biological activity of new compounds or guiding the structure optimization of existing compounds by models. With the continuous development of information technology such as data mining, various machine learning algorithms have been widely used in QSAR modeling. For example, Zekri et al. [4] used multiple linear regression (MLR) to predict pharmacokinetic properties, and the results showed that they had obvious advantages in computational efficiency and interpretability. However, MLR has a preassumed linear assumption, that is, the relationship between independent variables and dependent variables is linear, which may not be applicable to QSAR models with strong nonlinear relationship. Therefore, some nonlinear machine learning algorithms have attracted the attention of scholars. In recent years, support vector machine (SVM) is widely used in QSAR modeling [5, 6]. Because it can produce quite stable prediction effect for low-dimensional data, it is easier to

  DOI: 10.25236/icbcme.2021.032

implement than neural network model. When applied to high-dimensional data, the prediction performance would be greatly reduced. Artificial neural network was used in QSAR model by scholars in the early 1990s [7], and its powerful nonlinear mapping ability and efficient training speed once became very popular. For example, Ghasemi et al. [8] proved that the deep neural network after parameter initialization provides high-quality model prediction ability, but the hidden units and layers of the network are artificially adjusted, and there are some defects in efficiency.

As a result, there is still no unified standard for the modeling of high-dimensional nonlinear molecular descriptor-bioactive system. Although some existing machine learning models have obvious advantages in performance, the parameter optimization of the models is still a complex process. In order to screen molecular descriptors of important compounds more effectively and build an accurate QSAR quantitative prediction model, and achieve maximum economic benefits, Xgboost model is used to reduce the dimension of nonlinear high-dimensional data, then QSAR model is constructed based on several neural network models with better performance at present, and an improved BP neural network model based on sparrow optimization algorithm (SSA-BP) is proposed to predict the biological activity of compounds, which provides decision-making suggestions for the research and development of clinical anticancer drugs according to the diagnostic breast cancer data set provided by Owen Knowledge Base of the University of California.

2. Organization of the Text

2.1 Establishment of Quantitative Prediction Model of Biological Activity

The experimental data used the diagnostic breast cancer data set provided by Owen Knowledge Base of the University of California, which includes the biological activity data of 1974 compounds to ERα (the biological activity value is expressed by pIC_{50}, which is usually positively correlated with biological activity, that is, the higher the pIC_{50} value, the higher the biological activity) and 729 molecular descriptor information corresponding to each compound. Firstly, 729 molecular descriptors were non-linearly reduced based on Xgboost model. According to the importance of features, 20 molecular descriptors with high contribution were selected. Secondly, based on the variance expansion factor method (VIF>10 indicates significant correlation with other variables), the multicollinearity between molecular descriptors was eliminated. Finally, 11 molecular descriptors which have significant influence on the biological activity of compounds were obtained including MDEN-22, VCH-5, nAtomP, MDEO-12, ETA_BetaP_ns_d, maxssO, MDEO-11, nHsOH, ETA_Eta_B_RC, nHBAcc and nRotB. What's more, a prediction model was established based on artificial neural network. BP, Elman and Wavelet neural networks were used to fit the relationship between molecular descriptors of compounds and biological activity. The optimal neural network model was selected by comparing the prediction effects of neural network models, and the parameters of neural network model were optimized by intelligent optimization algorithm. Finally, the quantitative prediction model of biological activity was obtained.

2.2 Solution of Quantitative Prediction Model of Biological Activity

2.2.1 Model Training Results

The prediction accuracy of the model can be evaluated by Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Mean Square Error (MSE), and its equation is as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \quad (3)$$

Before training the model, we need to preprocess the data set. There are 1974 data (compounds) in the sample set of this topic, 70% of the sample set is used as the training set, and the other 30% of the sample set is used as the test set. Eleven molecular descriptors were screened as characteristic variables, and the biological activity pIC_{50} of each data was used as tag data. After training BP, Elman and WNN respectively, the model prediction results are obtained, and the comparison of prediction error results is shown in Table 1.

Combined with Table 1, it can be seen that MAPE, MAE and MSE of BP neural network are the lowest among the three models, and the evaluation indexes of wavelet neural network are the largest. For example, in the performance of MSE index, Elman is about 1.6 times of BP and WNN is about 2.9 times of BP, which shows that BP neural network has the best performance in quantitative prediction of biological activity, followed by Elman and WNN. However, the parameters of BP neural network have not reached the optimal state at present, and its prediction effect needs to be improved. Therefore, it is necessary to adjust the parameters for BP to improve the prediction accuracy of BP neural network.

2.2.2 Sparrow Search Algorithm

Sparrow Search Algorithm (SSA) is a population intelligence optimization algorithm [9], and its principle is the foraging behavior and anti-predation behavior of sparrow population. Foraging behavior is that after a predator sparrow (discoverer) finds food, it will inform its followers and make the sparrows move towards the direction of food, which is embodied in the algorithm as approaching the optimal solution position. Anti-predator behavior is that when the discoverer detects the danger, it will inform other sparrows to give up their food, which is embodied in the algorithm as approaching the origin.

SSA has the advantages of good stability, strong global search ability and fast convergence speed. Although BP neural network has achieved good results in the quantitative prediction model of biological activity, there are many weight and threshold parameters in BP neural network, and it is difficult to adjust each parameter systematically. Therefore, in order to make BP neural network achieve better prediction effect, this paper uses SSA to optimize BP neural network. The optimization process is shown in Figure 1.

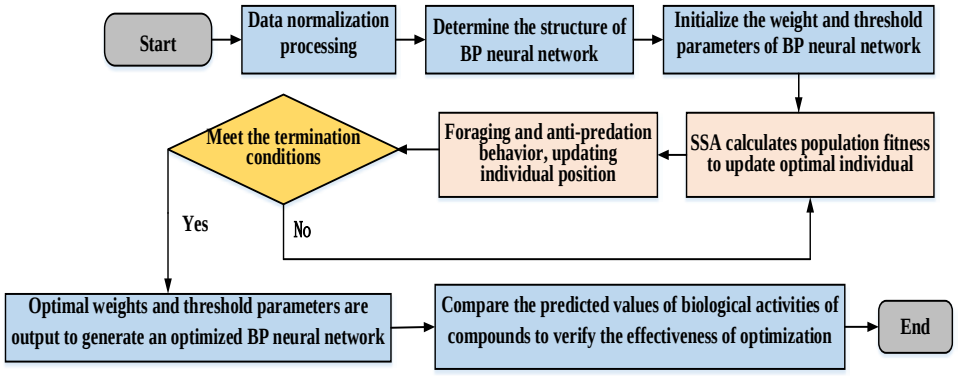

Fig.1 Ssa Optimized Bp Neural Network Flow Chart

2.2.3 Bp Neural Network Optimization Results

By optimizing the parameters of BP neural network with SSA, a new BP neural network SSA-BP can be obtained. The parameters of SSA are as follows: the population size is 20, the dimension is 8, the maximum iteration times is 30, the upper boundary of threshold is 5, the lower boundary of threshold is 5, the warning value is random number of [0, 1], and the safety value is 0.8. The prediction accuracy of SSA-BP model is compared with the other three models, and the results are shown in the following table.

Table 1 Comparison of Model Prediction Accuracy

	WNN	Elman	BP	SSA-BP
MAPE	0.21045	0.16283	0.03414	0.02136

MAE	0.30287	0.20142	0.15382	0.13024
MSE	0.16292	0.09185	0.06559	0.04898
R^2	0.90158	0.91425	0.93412	0.94572

Table 1 confirms that it is effective to optimize the network structure of BP by SSA. According to the index MSE, the error of WNN is 3.33 times that of SSA-BP, and that of Elman is 1.87 times that of SSA-BP. The prediction accuracy of WNN is better than that of WNN and Elman. The MSE of SSA-BP was 0.04898, which was 25.3% lower than that of BP (0.06559); Compared with BP, MAPE decreased by 37.4%; Compared with BP, MAE decreased by 15.3%.

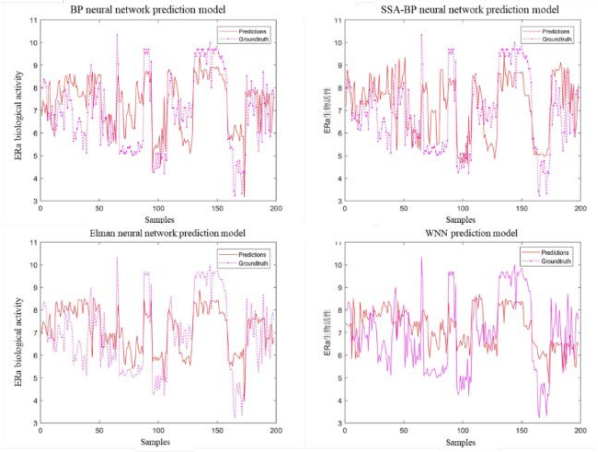

Fig.3 Bp, Ssa-Bp, Elman and Wnn Prediction Effect Diagram

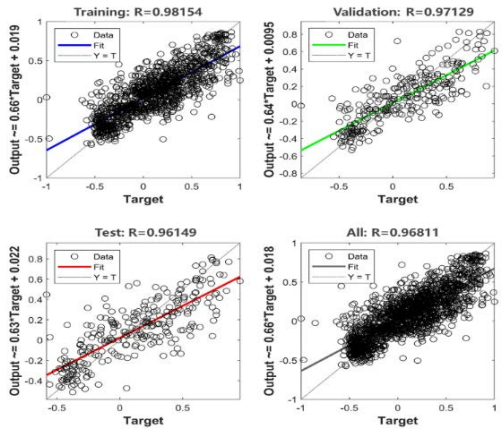

Fig.3 Goodness of Fit of Ssa-Bp Neural Network Model

It can be seen from Figure3 that the prediction value of SSA-BP is well fitted with the sample value, and the R^2 of the verification set reaches about 0.94. When the sample value changes drastically, other models have large errors, but SSA-BP can show good stability when the sample changes drastically, which shows that the method of improving BP neural network with SSA is practical and effective.

3. Summary

By comparing the performance of several machine learning models commonly used in QSAR modeling, this study aims to provide guidance for the selection of anticancer drugs. Firstly, important molecular descriptors are selected based on the nonlinear dimension reduction technology Xgboost model. Secondly, three different neural network models (BP neural network, wavelet neural network and Elman neural network) are used to construct quantitative prediction models of compound activity. The results show that the three machine learning models have good prediction performance. By comparing MSE, MAE, MAPE and R^2, it is proved that BP neural network has the best performance, followed by Elman neural network and wavelet neural network. In order to further improve the prediction performance of BP neural network, an improved BP neural network (SSA-BP) model based on sparrow optimization algorithm is proposed. It is found that the prediction error MSE is reduced to about 0.04, and the test set R^2 reaches 0.94, which is one of the best algorithms for QSAR modeling at present. It greatly improves the economic benefits of anticancer drug research and development. Future research can further optimize the best range of molecular descriptors to improve the biological activity of compounds.

4. Acknowledgement

This research was supported by National Natural Science Foundation of China (71971029) and Huo Yingdong Young Teachers Foundation (171069).

References

[1] Fan L, Strasser-Weippl K, Li J J, et al. Breast cancer in China[J]. The lancet oncology, 2014, 15(7): e279-e289.

[2] Cheng L, Schneider B P, Li L. A bioinformatics approach for precision medicine off-label drug drug selection among triple negative breast cancer patients[J]. Journal of the American Medical Informatics Association, 2016, 23(4): 741-749.

[3] Cherkasov A, Muratov E N, Fourches D, et al. QSAR modeling: where have you been? Where are you going to?[J]. Journal of medicinal chemistry, 2014, 57(12): 4977-5010.

[4] Zekri A, Harkati D, Kenouche S, et al. QSAR modeling, docking, ADME and reactivity of indazole derivatives as antagonizes of estrogen receptor alpha (ER-α) positive in breast cancer[J]. Journal of Molecular Structure, 2020, 1217: 128442.

[5] Akay M F. Support vector machines combined with feature selection for breast cancer diagnosis[J]. Expert systems with applications, 2009, 36(2): 3240-3247.

[6] Zheng B, Yoon S W, Lam S S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms[J]. Expert Systems with Applications, 2014, 41(4): 1476-1482.

[7] Gedeck P, Kramer C, Ertl P. Computational analysis of structure–activity relationships[J]. Progress in medicinal chemistry, 2010, 49: 113-160.

[8] Ghasemi F, Mehridehnavi A, Fassihi A, et al. Deep neural network in QSAR studies using deep belief network[J]. Applied soft computing, 2018, 62: 251-258.

[9] Xue J, Shen B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.