ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

The Application of Data Mining Technology Base on BP Neural Network for Forecasting the Share Price

Meng Yi

Business School/Lingnan Normal University, Zhanjiang 524048, China

ABSTRACT. *This paper analyze the topology structure of BP network and research how to advance the generalization ability and rationalization of the model. Then, we give some concrete suggestion. Finally, unifying the analysis above, we have carried on the forecast of stock price using the BP network. The result indicated that so long as we choose the reasonable network model and the structure, its forecasting ability is quite considerable.*

KEYWORDS: *BP Neural Network; Generalization ability; Over-Fitting; Share Price*

1. Introduction

Data mining is a step in the process of knowledge discovery. It mainly utilizes some specific knowledge discovery algorithms to mine valuable knowledge from data under certain operational efficiency constraints[1]. In principle, data mining can be applied to any type of information source. This includes relational database, data warehouse and transaction database. Among these data sets, time series is a time relationship between the data of one kind of data set. In the process of data mining for time series, we must consider the time relationship between data in data set. This kind of data mining is called time series data mining(TSDM).

It has always been a dream for investors to predict the stock market vicissitudes, grasp its changing rules and forecast its trend. Stock prices involve many uncertainties, and the correlations among them are complex. “Random Walk Theory” [2] holds that stock price volatility is completely random, but a large number of facts show that there is some implicit regularity in stock price volatility. There are many factors affecting the stock market, so it is very difficult to understand the mechanism of stock market change thoroughly in theory. This paper regards the stock market as a deterministic non-linear dynamic system, i.e. the internal dynamic mechanism is deterministic. The historical data and other information of stock prices contain information that can be used to predict future stock prices. Using equation (1):

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

$$P(t+1) = f(P_{t-k}, \cdots, P_t, X_{t-l}, \cdots, X_t, \cdots, Y_{t-m}, \cdots, Y_t \cdots) \quad (1)$$

Where F is the stock price, the external variable is X, Y. If only considered the internal relationship of stock price series, then equation (1) can be expressed as equation (2).

$$P(t+1) = f(P_{t-k}, \cdots, P_t) \quad (2)$$

The key to prediction is to construct sample data or appropriate approximation of function f. However, f is very complex. It is obviously inappropriate for traditional linear prediction method to use weighted sum of observation samples as prediction result. The neural network has a strong ability of approximating non-linear functions. The functional relationship between input and output variables can be obtained by training the sample data. That is to say, the coupling weights of each neuron can be determined by learning the neural network, so that the network has the function of approximating functions. In this way, we can predict the future behavior of the market without knowing the intrinsic dynamic mechanism of the market. Therefore, neural network is widely used in time series analysis and financial forecasting.

This paper focuses on the determination of BP network structure, and uses BP network to forecast the stock price. The results show that its prediction ability is considerable.

2. Sample Date

2.1 Collect and organize groupings

The primary precondition of using neural network to model is that there are enough typical and high precision samples. Moreover, in order to monitor the training process so that it does not occur "over-fitting" and evaluate the performance and generalization ability of the established network model, the collected data must be randomly divided into training samples, test samples (more than 10%) and test samples (more than 30%). When grouping data, the balance between sample patterns should be considered as much as possible.

2.2 Determination of Input and Output Variables and Data Preprocessing

Generally, the input variables of BP network are endogenous variables. If there are many input variables, principal component analysis can be used to reduce the input variables, and the input variables can also be reduced according to the ratio of the system error caused by eliminating a variable to the original system error. The output variable is the exogenous variable, it can be one or more. Generally, it is better to transform a network model with multiple inputs with one output, and the training is more convenient.

There are many methods of pretreatment. However, it must be noted that after

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

the pre-processing data training is completed, the network output results need to be inversely transformed to get the actual value. In order to ensure the extrapolation ability of the model, it is better to make the data pretreated between 0.2 and 0.8.

3. Determination of Network Topology Structure

About the network structure, we need to study how to determine the network structure to ensure good promotion ability when the number and quality of training samples are fixed.

The network structure involves the following aspects: the number of hidden layers, the number of nodes in the hidden layer, the initial weight, the excitation function, the learning rate, the momentum factor and the error accuracy.

3.1 Hidden Layer Number

It is generally accepted that increasing the number of hidden layers can reduce the network error and improve the accuracy, but it also complicates the network, thus increasing the training time of the network and the tendency of "over-fitting".

The characteristic is "backhaul error" for BP network. So if the number of hidden layers increases, the mapping expressed by the network will become more complex. In addition to increasing the amount of computation, it will also easily cause data distortion. Especially when the input data is not "normalized".

Hornik have also proved that MLP networks with a hidden layer can approximate any rational function with arbitrary accuracy if the input and output layers use linear transformation functions and the hidden layer uses Sigmoid transformation functions[3]. Obviously, this is an existential conclusion. Three layers of BP network (i.e. one hidden layer) should be given priority in the design of BP network. Generally, the training effect is easier to achieve by increasing the number of hidden layers to obtain lower error. For the neural network model without hidden layer, it is actually a linear or non-linear regression model (depending on the form of linear or non-linear transformation function in the output layer).

3.1 Hidden Layer Node Number

The selection of hidden layer nodes is very important in BP network. It not only has a great impact on the performance of the established neural network model, but also is the direct cause of "over-fitting" in training. However, there is no scientific and universal method to determine it in theory.

The selection of hidden layer nodes is often related to the characteristic factors implicit in the input data. Before the initial training of the network, the weights of all parameters are preliminary, so far there is no uniform specification. At present, most of the formulas for determining the number of hidden layer nodes proposed in the

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

literature are for arbitrary number of training samples, and most of them are for the most disadvantageous situation, which is difficult to meet in general engineering practice and should not be adopted. In fact, the number of hidden layer nodes obtained by various calculation formulas sometimes differs several times or even tens of times. In order to avoid the phenomenon of "over-fitting" and ensure high enough network performance, the basic principle of determining the number of hidden layer nodes is as follows: On the premise of satisfying the accuracy requirement, the number of hidden layer nodes should be as few as possible. The research shows that the number of hidden layer nodes is not only related to the number of input and output layer nodes, but also related to the complexity of the problem to be solved, the form of conversion function and the characteristics of sample data.

In a word, if the number of hidden layer nodes is too small, the network may not be trained at all or the network performance is poor. If there are too many hidden layer nodes, the system error of the network can be reduced, but on the one hand, the training time is prolonged. On the other hand, the training is easy to fall into the local minimum and can not get the optimal point, which is the internal reason of "over-fitting". Therefore, the reasonable number of hidden layer nodes should be determined by gradual increase or pruning of nodes when the complexity and error of network structure are considered comprehensively.

4. Network Training

4.1 Training

The training of BP network is to make the sum of squared errors between the output value of the network model and the output value of the known training sample reach the minimum or less than a certain expected value by applying the principle of error back propagation. Although it has been proved that BP network with one hidden layer (using Sigmoid transform function) can approximate any function arbitrarily. Unfortunately, there is no constructive conclusion so far, that is, how to design a reasonable BP network model and approximate satisfactorily the rules contained in the samples by training from the limited samples given. Therefore, the process of establishing a reasonable BP neural network model through training samples learning (training) is called "the process of artistic creation" in foreign countries, which is a complex and very cumbersome and difficult process.

Because BP network adopts error back propagation algorithm, its essence is an unconstrained non-linear optimization calculation process. When the network structure is large, not only the calculation time is long, but also it is easy to fall into local minimum and get the optimal result. Now many optimization methods, such as improved BP method, genetic algorithm (GA) and simulated annealing algorithm, have been applied to the training of BP network (these methods can obtain global minimum by adjusting some parameters). But in application, the adjustment of these parameters often varies with different problems, and it is difficult to find the global

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

minimum. The most widely used of these methods is the improved BP algorithm which adds momentum terms.

4.2 Learning Rate and Impulse Coefficient

Learning rate affects the stability of system learning process. Large learning rate may lead to excessive modification of network weights at each time, and even lead to irregular jump and non-convergence of weights exceeding the minimum of a certain error in the process of modification. And too little educational background leads to too long learning time, but it can ensure convergence to a certain minimum. Therefore, the general tendency is to choose a smaller learning rate to ensure the stability of the learning process, usually between 0.01 and 0.8.

The purpose of adding impulse items is to avoid the network training trapped in a shallow local minimum. In theory, the size of the value should be related to the size weight correction, but in practical application, the constant is usually between 0 and 1, and the learning rate is generally higher than that.

5. Stock price prediction

Stock market is a moving and special system, and its prediction has always been divided into basic analysis and technical analysis. Through basic analysis to predict the long-term trend and know what kind of stocks we should buy, while technical analysis lets us grasp the timing of specific purchases.

This paper forecasts the future trend of Thomson's stock price, a listed company in the United States, by using the stock data, the historical data of relevant economic and technical indicators.

This paper found a three-layer feedforward network, and using the back propagation algorithm in the MATLAB neural network toolbox. The input of the system is some technical and economic indicators, such as Thomson's stock price(X_1),trading volume(X_2), Standard & Poor's Index(X_3), Purchasing Manager Index(X_4), Federal Fund Interest Rate(X_5), American Consumer Price Index(X_6), Thomson's stock price-earnings ratio(X_7), and the output is a momentary share price. According to the previous network structure analysis, in order to reduce the impact of random fluctuations in data, each variable takes the week as the sampling frequency and takes the week mean. Before input, each data is normalized to a value between 0 and 1. The whole process is time-slipping, using the data of the first 30 weeks to predict the stock price of the next week. The first 30 weeks were learning samples, and the last 7 weeks were predictive comparisons to verify the accuracy of prediction. The original data is presented in Table 1.

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

Table 1 Raw data of each index variable

	X_1	X_2	X_3	X_4 (%)	X_5	X_6 (%)	X_7
1	24.925	182	1420.331	4.72	168.8	56.8	21.3
2	27.25625	328.4	1448.653	5.68	168.8	56.8	23.3
3	29.2578125	225.5	1449.497	5.59	168.8	56.8	25.0
4	35.725	481.6	1394.877	5.43	168.8	56.8	30.5
5	45.025	908.4	1412.446	5.66	169.8	56.7	38.5
6	59.55	582	1416.331	5.71	169.8	56.7	50.9
7	56.3	682.8	1382.808	5.75	169.8	56.7	48.1
8	55.3203125	290.5	1349.916	5.72	169.8	56.7	47.3
…	…	…	…	…	…	…	…
31	66.53	27.4	1444.630	6.49	172.8	49.9	56.9
32	64.66	48.0	1473.418	6.45	172.8	49.9	55.3
33	70.55	167.4	1488.722	6.53	172.8	49.9	60.3
34	66.65	103.0	1503.674	6.46	172.8	49.9	57.0
35	62.00	69.8	1512.999	6.54	172.8	49.9	53.0
36	60.19	91.0	1499.095	6.56	173.7	48.8	51.4
37	57.20	62.4	1480.573	6.50	173.7	48.8	48.9

First, we normalize all the data into values between 0 and 1. Then,using the conjugate gradient algorithm (traincgf) to complied the MATLAB program and calculated. The results are shown in Table 2.

serial number	actual value	predicted value	Error rate(%)
31	66.53	67.3799	1.2775
32	64.66	65.2466	0.9072
33	70.55	70.2543	-0.4191
34	66.65	66.7526	0.1539
35	62.00	62.1307	0.2108
36	60.19	60.6099	0.6976
37	57.20	57.9596	1.3280

The evaluation error rate is 0.5937, and the prediction effect is well.

6. Conclusion

The key to the application of neural network in forecasting is to determine the network topology. The network structure determines the validity of the prediction results, and a good and reasonable network structure prediction results should be realistic. The results of this paper show that the accuracy of BP neural network in forecasting stock price is quite high, and it also shows the rationality of the process of determining the network structure.

Although neural networks are widely used in forecasting and other fields, the network model itself does not have the ability of causal interpretation. For example, the application of back-propagation neural network model to predict stock prices, and the results are very accurate, which only means that there is a certain

ISSN 2522-6398 Vol. 2, Issue 12: 32-38, DOI: 10.25236/FER.2019.021206

relationship between the independent variables of the network model and the stock price index, but it does not mean that the independent variables of the model are the factors affecting the stock price index. If want to know whether the variables have causal relationship with each other, it must further carry out other statistical identification.

References

[1] Einstein A (1956).Investigations on the Theory of the Brownian Movement. Dover Publications.
[2] Kurt Hornik (1991). Approximationcapabilities of multilayer feedforward networks.Neural Networks.
[3] Friedrich Hubalek,Petra Posedel (2011). Joint analysis and estimation of stock prices and trading volume in Barndorff-Nielsen and Shephard stochastic volatility models.Quantitative Finance, no.6, pp.58-59.
[4] Yuan-Ming Lee,Kuan-Min Wang (2015). Dynamic heterogeneous panel analysis of the correlation between stock prices and exchange rates.Economic Research-Ekonomska Istra? Ivanja, no.1, p.1-2