A Hybrid Forecasting Model based on ARIMA and BP-NN

Jingyang Li [a], Guangjun Huang [b]

College of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China

[a]lijingyang1990@126.com, [b]huangguangjun@126.com

Abstract

This paper analyzes several methods of time series forecasting and their advantages and disadvantages. The accuracy of the original ARIMA model is not enough in some cases, mainly due to the shortage of ARIMA description for nonlinear residuals. A hybrid ARIMA-BP model is proposed. The hybrid model includes ARIMA model for linear prediction of time series, and neural network for nonlinear residual prediction. Finally, we validate the hybrid model using the number of influenza cases in Henan province. The results show that the hybrid model proposed in this paper has higher accuracy and is suitable for the prediction of influenza.

Keywords

Influenza Forecasting, Neural Networks, ARIMA, Hybrid Model.

1. Introduction

With the development of technology and technology, the prediction of influenza incidence trends become more and more important. Influenza is an acute respiratory infectious disease caused by influenza virus, due to its wide range of spread, speed, social harm and special attention. In recent years, influenza epidemic situation is grim, especially when the influenza virus mutation will become a new virus, spread rapidly, the possibility of large-scale outbreak, so the dynamic prediction of influenza caused by the international community's high attention. Accurately predict the development trend of influenza, for the correct formulation of planning, rational allocation of health resources is of great significance. There are many methods for the prediction of infectious diseases. For the prediction of influenza, time series prediction is the main prediction method. In this method, the ARIMA (Autoregressive Integrated Moving Average) model is one of the most widely used prediction models because of its simplicity and feasibility [1,2].

However, the ARIMA model has a fundamental flaw [3,4]: in the ARIMA model, the future value of the sequence variable is assumed to satisfy the linear relationship between the past observations of the variables and the stochastic errors, ie the stationary time series. In reality, most of the time series are non-stationary, so before modeling the data need to be differential processing, although the difference can be regarded as a stable data sequence, but which still contains non-stationary factors, which resulted in ARIMA Predicted Error Increase of Non - stationary Time Series. In recent years, because of the strong learning and data processing ability of BPNN(Back Propagation Neural Network), it can excavate the nonlinear relation which is complicated or even difficult to describe by mathematical formula, and there is no special requirement on the number of samples [5]. A three-layer neural network can theoretically approximate at any precision.Neural network is more and more attention, the application is more and more extensive. From the above analysis can be drawn ARIMA method is based on linear time series prediction, and nonlinear data processing effect is not reasonable, the effect is poor. The nn technology is good at mining the implicit non-linear relationship in data, but in dealing with linear data is ineffective. Obviously, the actual prediction problem usually

includes both linear and non-linear components [6,7]. Therefore, some scholars will integrate the above method for time series prediction research. In addition, a large number of studies on prediction indicate that a single method or a single model can not optimize the prediction, and the method of combining the various models is better than the prediction of a single model [8,9]. Both theory and experiment show that the combination of multiple prediction methods is an effective way to improve the prediction performance, and the risk of using the combination model is lower than that of the single model.

At present, there are few researches on the prediction of influenza time series by combinatorial model, and the existing combinatorial model mainly focuses on the weighted average of the single forecasting method. The emphasis is on the determination of the weighting coefficient, which directly affects the prediction of the model. Effect, but the weighting coefficient is difficult to determine with a lot of subjectivity. In this paper, based on ARIMA and nn combination prediction model, the ARIMA model is first used to fit the linear part of the sequence, and then the nn is used to estimate the nonlinear residual part of the sequence. Finally, the final prediction result is formed. The experimental results show that the combined model improves the prediction accuracy of the model compared with the single model.

2. Basic principle of ARIMA and BPNN model

2.1 ARIMA Model

Arima model is the basic idea is to predict the object over time to form the data sequence as a random sequence, with a mathematical model to approximate the description of the sequence.

The expression of the ARIMA (p, d, q) model is denoted as:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-1} + \cdots + \varphi_p X_{x-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

Where p is the autoregressive order, $\varphi_{1,} \ldots \varphi_p$ is the autoregressive coefficient, q is the moving average term, $\theta_1, \ldots \theta_q$ is the moving average coefficient, d is the time series smoothing Difference times.

2.2 BP neural network.

BP neural network is a multi-layer feedforward neural network, is one of the most widely used neural network models, the main characteristics of the network is the signal forward transmission, error back propagation. In the forward transmission, the input signal from the input layer through the hidden layer by layer processing, until the output layer. The neuronal state of each layer affects only the next neuron state. If the output layer can not get the desired output, it goes backwards and adjusts the network weights and thresholds according to the prediction error. The topological structure of BP neural network is shown in Fig-1.

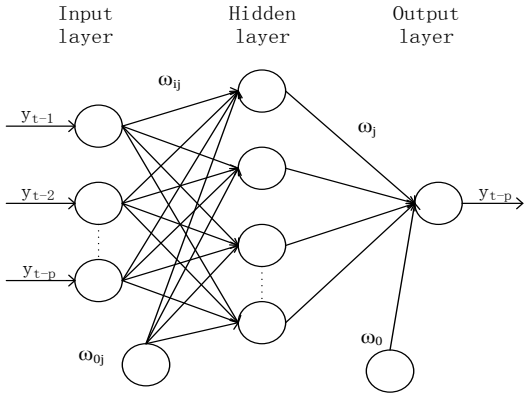

Fig. 1 Toopulogy of bp

In the neural network model, the relationship between the output y_t and the input $y_{t-1}, y_{t-2}, \ldots y_{t-p}$

$$y_t = \sum_{j=1}^{q} w_j g\left(\omega_{0j} + \sum_{i=1}^{p} \omega_{ij} y_{t-i}\right) + \varepsilon_t \quad (2)$$

In the formula, $\omega_{ij}$and $\omega_j$are model parameters (connection weight vector and threshold vector), p is the number of nodes in the input layer, q is the hidden layer node number formula, g (x) is the hidden layer excitation function. The function is

$$g(x) = \frac{1}{1+e^{-x}} \tag{3}$$

The neural network of the sub-description is actually a non-linear relationship between the previous observations of the sequence and the expected value of y_t:

$$y_t = f\left(y_{t-1,}y_{t-2}, \dots y_{t-p}, W\right) + \varepsilon_t \tag{4}$$

2.3 Hybrid model of ARIMA and BP-NN

In the hybrid model, a set of time series is decomposed into linear autocorrelation structure and non-linear structure two parts. Specifically, the ARIMA model is used to model the linear part of the time series, and the ARIMA modeling error is modeled using the neural network, that is, the error is corrected.

Assume that the time series y_t is decomposed into a linear autocorrelation subject column $L_t$and a non-linear residual column N_t:

$$y_t = L_t + N_t \tag{5}$$

In this paper, the following steps are used to build the model:

Assuming that the arima model predicts , the residual is

$$e_t = y_t - \widehat{L_t} \tag{6}$$

The error sequence with nonlinearity is

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t \tag{7}$$

f(x) is a nonlinear function determined by the neural network, ε_t is a random error. The estimated residuals estimated by the neural network are $\widehat{N_t}$

Then the combined model predictions are

$$\widehat{y_t} = \widehat{L_t} + \widehat{N_t} \tag{8}$$

In this paper, we use MAPE(Mean Absolute Percentage Error) ,RMSE(RootMean Square Error) to evaluate the predicted results. The formula is as follows:

$$E_{MAPE}, = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y_i}-y_i}{y_i}\right| \tag{9}$$

$$E_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2} \tag{10}$$

3. Experiments and results

Data from the public health scientific data center, including data from 2004 to 2012, Henan Province, the number of monthly reports of influenza disease [11]. The province's population base, and relatively stable, the incidence and percentage of the trend of development trend is basically the same, so the number of months the number of cases using the establishment of time series.

3.1 Establishing ARIMA forecasting model

According to the incidence of influenza in Henan Province in 2004 to 2012, the following trends.

The ARIMA model is established to estimate the linear part of the sequence, and the prediction result $\widehat{L_t}$is obtained.

From the above figure we can see that the time series is non-stationary series, and the original data are logarithmically separated, and the new sequence is stable. After comparing the parameters one by one, the optimal model is ARIMA (2,1,6). Using this model, the ARIMA estimate $\widehat{L_t}$ was obtained, and the number of cases from January to December 2012 was L_t.The results of the white noise test show that the Ljung-Box statistics correspond to p <0.05, indicating that the sequence is a non-white noise sequence.

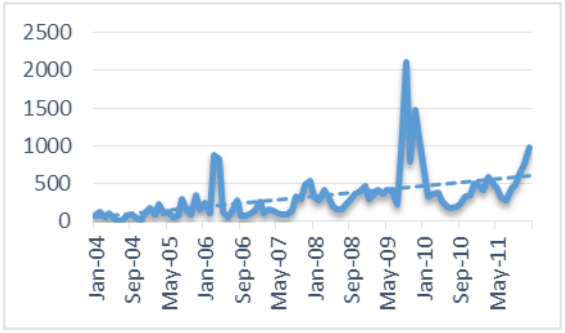

Fig. 2 Establishing ARIMA forecasting model

3.2 Establishing The neural network

In this paper, the BP-NN structure is 8-13-1, that is, 8 input elements. Considering the performance of Bp neural network in this paper, including the convergence efficiency, prediction accuracy and generalization ability, the number of nodes in the hidden layer is 13, and the fitting error of BP neural network is set to 10^{-5}, The maximum number of times is 50,000.The ARIMA model prediction error was only from January 2004 to December 2011, so the total sample size was 96. From January 2004 to December 2011, the ARIMA prediction error was the input to the network, from May 2004 to 2011 December incidence of the ideal output for the network, composed of neural networks. The neural network is used to predict the error forecast N_t from January to December. Table 1 shows the predicted values of the nonlinear part and the error values predicted using ARIMA.

Table 1. Predicted values of the nonlinear part and the error values predicted using ARIMA

Month	1	2	3	4	5	6	7	8	9	10	11	12
Real	-25	-43	136	-82	-223	-240	-177	-48	-30	-93	-168	-48
Predict	-7	-64	11	-9	-35	-56	-50	--45	22	-13	-29	-42

Since the actual value is equal to the ARIMA model linear predictive value plus the nonlinear prediction of the error part by the neural network. So that we can get the final results of the forecast. As a comparison we use ARIMA model, BP neural network model also predicted the number of cases in 2012, Table 2 shows the three different models of forecasting results and actual values, from the table we can see that these models have a certain ability to predict.

Table 2. Real value and the Predict values

month	Real	ARIMA	BP	ARIMA-BP
1	907	932	886	925
2	1155	1198	1189	1136
3	1349	1213	1490	1224
4	911	993	831	982
5	776	999	560	961
6	540	780	290	724
7	516	693	363	643
8	720	768	658	733
9	758	788	715	813
10	831	924	729	911
11	1062	1230	916	1199
12	1493	1541	1443	1499

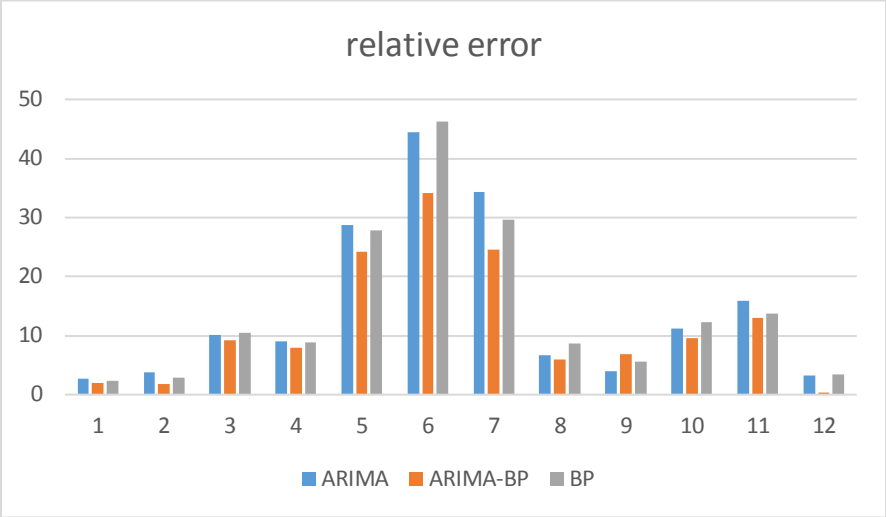

Fig. 3 relative error of the models

Figure 3 shows the relative error of the three different models of contrast. In addition to September, the other months are mixed model of the relative error of the smallest, indicating that the mixed model of the relative error range is smaller and more stable. Hybrid model, ARIMA model, and BP neural network, respectively.The average relative error of the hybrid model is 11.65% lower than the other two models 14.45% and another 14.32%, so the prediction of the hybrid model is more stable.

Table 3. Rmse and mape of the model

Model	RMSE	MAPE
ARIMA	300.5108	1.449132
BP	129.21	1.432752
ARIMA-BP	105.6567	1.126721

The MAPE of ARIMA, BPNN and Hybrid model are calculated as Table 3 respectively. Combined with Table 2 and Figure 2, we can see that the hybrid model has the best prediction effect, which is better than the other two single prediction models.

4. Conclusion

In this paper, hybrid neural network and ARIMA forecasting model, the incidence trend of influenza has good prediction accuracy. The model can describe both the linearity of historical data and the nonlinear law of time series. It is obvious from the results that the combined model is much better than the single model. Furthermore, for a complex time series, which includes both linear and nonlinear factors, ARIMA is used to predict the linear law, and then the neural network is used to predict the nonlinear law. Then the two-part method is general applicability.

Due to the problem of data availability, this paper does not consider other factors in neural network training part selection, which limits the prediction ability of the model to a certain extent. In addition BP neural network is not a very perfect network, because the initial weight of the network and the choice of the threshold value of the lack of basis, but also has great randomness, which also largely affect the network's generalization ability. In this paper, the reciprocal of errors is used to guide the learning process, which is essentially local optimization. When there are many local extreme, the residual part algorithm is easy to fall into the local optimum. As the improvement direction of the future model, we can use the heuristic algorithm such as genetic algorithm to optimize the weights and thresholds for neural network training. This is because the genetic algorithm and so on is to search from the question collection, the coverage is wide, is one kind of global optimization method, like this may the smaller fall into the local optimum.

References

[1] J.X.Shi, W.Z. Zhang, G. Q. Ji :Application of ARIMA model in forecasting and early warning of influenza (Capital Journal of Public Health,2010,2):Vol4,No1(In Chinese)
[2] Li Qi, Ge Li, Qin Li: Applications of ARIMA model on predictive incidence of influenza (Journal of the Third Military Medical University,2007,2):Vol29,No3
[3] Helfenstein: U. Box-Jenkins modelling in medical research. (Statistical Methods in Medical Research). 1996
[4] Y. Q. Yan, Ping Guo: Predicting resource consumption in a web server using ARIMA model (Journal of Beijing Institute of Technology, 2014, 04):502-510.
[5] Chao Ren Ning, An, Jianzhou Optimal parameters selection for BP neural network base on particle swarm optimization: A case study of wind speed forecasting (Knowledge-based systems, 2014, 1, Vol56):226-239.
[6] Al-Sakkaf, A., Jones, G..Comparison of time series models for predicting campylobacteriosis risk in New Zealand. [J].Zoonoses and Public Health, 2014, 61(3):167-174.
[7] X. J. Dong, W .N. Jia: Predictive efficiency comparison of ARIMA-time-series and BP neural net model on infectious diseases[J] MODERN PRACTICAL MEDICINE 2010,5:Vol22:No2
[8] Dilli R Aryal, Y. W. Wang. Time-series analysis with a hybrid Box-Jenkins ARIMA [J]. Journal of Harbin Institute of Technology, 2004, 04:413-421.
[9] Feng-Kuang Chuang, Chih-Young Hung, Chi-Ya Chang et al. Deploying Arima and Artificial Dedicated to all Aspects of Sensors in Science, Engineering, and Neural Networks Models to Predict Energy Consumption in Taiwan[J].Sensor Letters: A Journal Dedicated to all Aspects of Sensors in Science, Engineering, and Medicine,2013,11(12):2333-2340.
[10] Zhong-Da Tian, Xian-Wen Gao, Kun Li. A Hybrid Time-delay Prediction Method for Netw orked Control System [J]. International Journal of Automation and Computing, 2014, 01:19-24.
[11] Information on http://www.phsciencedata.cn/Share/index.jsp