A Comparative Study of Customer Complaint Prediction Model of Time Series, Multiple Linear Regression and BP Neural Network

Xin Xu¹²³, Zhijie Sun¹²³, Li Wang¹²³, Jun Fu¹²³, Chao Wang¹
¹ State Grid Jibei Electric Power Company Limited.
² State Grid Jibei Electric Power Company Limited Electric Power Research Institute
³ Country Huadian Electric Power Research Institute Co., Ltd.
Corresponding author’s e-mail: 676269930@qq.com

Abstract. The current prediction algorithm is mainly the “time series”, the “multiple linear regression” and the “BP neural network”. This article studies and compares the three algorithms in the field of customer complaint prediction. By using SPSS, this article predicts taking “Jibei Electric Power Customer Complaints” as the target. By comparing and analyzing the actual prediction results, the “BP neural network” algorithm is described as the most suitable complaint prediction for customer.

1. Introduction
The customer complaint is a mirror of the quality of the enterprise. How to prevent complaints is an unavoidable problem for enterprises with customer demand-oriented today. The prevention is better than the disaster relief. The enterprise should pay attention to the prevention of the complaint. The enterprise is required to eliminate the customer's dissatisfaction in the initial stage, so as to avoid the deterioration of problem and the cost of the enterprise investment [1]. The premise of prevention is to be predicted. At present, the prediction of customer complaints constructs prediction model often based on “ARIMA (Autoregressive Integrated Moving Average Model) Time Series” [2], “Multiple Linear Regression” [3], and “BP (Back Propagation) Neural Network” [4]. The three algorithms each have their own characteristics and disadvantages. However, when constructing the complaint prediction model, the selection of the appropriate core algorithm will essentially determine the success of the complaint prediction model. In this article, a comparative study of three kinds of algorithm prediction is carried out based on the “Number of complaints” of the power customers in the Jibei Region as the prediction target to explore the algorithms that apply to the customer's complaint prediction model.

2. The principle of algorithm

2.1. Arima time series algorithm
The time series refers to a set of statistical data in the order of time. The ARIMA time series algorithm includes a self-regression process AR, a moving average process MA, and a differential process \( DX = \text{diff}(y, i) \).

The concrete implementation process is divided into three steps.
(1) Time series differential/stationary processing.
The stability of the sequence is checked by a scatter diagram, an auto-correlation function and a partial auto-correlation function, and it is determined whether to perform the differential processing and the differential order according to the stationarity characteristics.

(2) Model parameter order recognition.
The parameter $p$ in the ARIMA (p, d, q) model is a self-regression term, the parameter $q$ is the number of moving average items, and the parameter $d$ is the number of times that the time series becomes stationary. [5] ARIMA (p, d, q) model parameters are set according to the truncated and trailing characteristics of the autocorrelation function and the offset function of the data sequence.

(3) Model test
The residual sequence white noise test is performed by constructing the correction statistic $Q$ of the box-pierce.

$$Q = \left(N - D - \max(p, q)\right) \sum_{k=1}^{N} \hat{a}_k^2.$$ (1)

2.2. Multiple linear regression
Regression is a study of the relationship between the variable and the independent variable, and the relationship between the independent variable and the dependent variable is expressed by means of the regression equation. The multivariate linear regression model is a correlation between a variable and a plurality of independent variables.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$ (2)

The concrete implementation process is divided into three steps:

(1) Independent variable selection
At the time of modeling, the selection of the independent variable is first made. The method is commonly used as stepwise regression and grey relational degree.

(2) Coefficient estimation of multivariate linear regression model
The estimation of the parameter vector $B$ is performed using the least square method. The residual error is $E$:

$$E = Y - \hat{Y}$$

$$\hat{Y} = XB$$

The least square method is used:

$$E'E = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - XB)'(Y - XB) = \min$$ (3)

According to the extremum principle, the above formula is for $B$ and $B=0$:

$$\frac{\partial E'E}{\partial B} = \frac{\partial (Y - XB)'(Y - XB)}{\partial B} = -2(YX)' + 2(XX)B = 0$$ (4)

It is therefore available:

$$\hat{B} = (XX)^{-1}XY$$

(3) Model test of multiple linear regression
Commonly used as the complex correlation coefficient test (R), and F test

$$R = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{Y})^2}}$$ (5)

$R$ describes the degree of linear correlation between the independent variable and the dependent variable.
\[ F = \frac{\sum (\hat{y}_j - \bar{y})^2 / (n - 1)}{\sum (y_j - \hat{y})^2 / (n - m)} \]  
\[ (6) \]

F test is to verify hypothesis whether \( \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_s = 0 \) is right. [6]

2.3. BP neural network

The full name of BP neural network is Back Propagation Neural Network, which was proposed by a group of scientists led by Rumelhart and McClelland in 1986. The BP neural network is a kind of “multi-layer feedforward neural network”, which uses the “error inverse propagation algorithm” to train the neural network. It is also one of the widely used neural network algorithms. The neurons of the BP neural network have three basic functions, namely, “modified weight value”, “summation” and “transfer”. [7]

The implementation steps are as follows:

1. Neural network initialization.
   Randomly give the random number of each neuron connection weight value (Wji) and a threshold value (-1, 1).

2. Calculate the output of the input layer.
   Randomly select an input vector \( X_p = (x_1, x_2, \ldots, x_n) \), and calculate expected output vectors \( Y_p = (y_1, y_2, \ldots, y_s) \).

3. Calculate the hidden layer output.
   \[ O_{pl} = f_j(\text{net}_{pl}) = f_j(\sum W_{ji}O_{pi}) \]  
   \[ (7) \]

4. The error of each node is calculated in the reverse direction, and the weight of each neuron is corrected.
   \[ W_{ji}(t + 1) = W_{ji}(t) + \eta \delta_{pi}O_{pi} \]  
   \[ (8) \]

3. Test procedure

3.1. Data description

In this article, the predicted target is a total amount of single-week of customer complaint of the State Grid Jibei Electric Power Company. The used date time interval is from July 2015 to September 2016. The ARIMA time series prediction only needs the statistical data of the “Complaints” category.

The multiple linear regressions and the BP neural network will use the total amount of the single-week complaints as the dependent variable. Through the data screening, 25 secondary work orders from the “information query”, “service consultation”, “fault report”, “report”, “opinions”, “suggestions”, “praise” and “service application” as the independent variable of the prediction model. The time interval is also from July 2015 to September 2016 and the data is about 1444248.

3.2. ARIMA time series algorithm

Because the predicted target is total number of complaints in one week, the study needs to divide the original data into 7 groups of time series according to the period of 7. 7 ARIMA time series prediction is respectively carried out to obtain all the predicted values. Then the original time series is divided into seven different time series, and the time series prediction model is established. The model parameters and model evaluation are shown in Table 1 below:

Table 1 ARIMA Model Evaluation Form
The prediction value is evaluated as shown in Table 2 below:

| Model                      | Model parameters | Model fitting statistic | Ljung-Box Q(18) | Outlier |
|----------------------------|------------------|-------------------------|-----------------|---------|
| Complaint model _ Monday   | (2,0,1)          | .659                    | .680            | 17.175  | .376    | 0      |
| Complaint model _ Tuesday  | (1,0,0)          | .624                    | .633            | 17.384  | .429    | 0      |
| Complaint model _ Wednesday| (1,1,0)         | .633                    | .643            | 13.240  | .720    | 0      |
| Complaint model _ Thursday | (1,0,0)          | .639                    | .628            | 12.244  | .785    | 0      |
| Complaint model _ Friday   | (1,1,0)          | .552                    | .655            | 13.201  | .723    | 0      |
| Complaint model _ Saturday | (1,0,3)          | .645                    | .674            | 14.181  | .585    | 0      |
| Complaint model _ Sunday   | (1,0,3)          | .626                    | .662            | 14.410  | .568    | 0      |

The prediction value is evaluated as shown in Table 2 below:

| The relative error is less than 1 0% | The relative error is less than 2 0% | The relative error is less than 3 0% | Relative error is less than 4 0% |
|-------------------------------------|-------------------------------------|-------------------------------------|----------------------------------|
| 22.48%                              | 43.56%                              | 60.21%                              | 71.12%                           |

3.3. Multiple linear regression

The “multiple linear regression” uses the “complaint amount/week” as the dependent variable. When the complaint amount/week of t day is to be predicted, the t-1 day other secondary work order other than the complaint work order will be used as the argument. And the variable selection is carried out by adopting the stepwise regression method.

The multivariate linear expression is as follows:

\[ y = 0.113x_1 + 0.88x_2 + 0.207x_3 + 0.183x_4 - 1.804x_5 + 4.890x_6 - 0.81x_7 - 10.336x_8 - 0.179x_9 + 0.718x_{10} - 1.394 \]  

\( (9) \)

Table 3 Parameter Schematic Form of the Multiple Linear Regression Equation

| System parameters | Business order class                                      |
|-------------------|---------------------------------------------------------|
| y                 | Complaint amount/week                                    |
| x_1               | User information                                        |
| x_2               | Electric energy quality fault                            |
| x_3               | Customer internal fault                                  |
| x_4               | Charge of electricity and electricity                    |
| x_5               | Metering device                                         |
| x_6               | Urgency complaints                                      |
| x_7               | Call-up failure report                                  |
| x_8               | Change of power utilization information                  |
$x_9$

Demand for electric service

$x_{10}$

Request for the service of the call-up service

To be abandoned

High-voltage fault

To be abandoned

Model evaluation:

| R  | R Square | R square after adjustment |
|----|----------|--------------------------|
| 0.705 | 0.498 | 0.486 |

Predicted value evaluation:

| The relative error is less than 10% | The relative error is less than 20% | The relative error is less than 30% | Relative error is less than 40% |
|-----------------------------------|-----------------------------------|-----------------------------------|----------------------------------|
| 21.67%                            | 36.57%                            | 55.53%                            | 68.62%                           |

3.4. Neural network algorithm

Because the data volume is small, the training quantity and the calculation amount are not small, a lower learning rate is adopted to improve the matching degree of the model. Using the IBM SPSS Statistics as a tool to set up a neural network model, the relevant configuration information is shown in Table 6 below.

| Configuration item | Parameter |
|--------------------|-----------|
| Interval           | 70% training 30% test 0% detection |
| Network system     | Single layer hidden layer |
| Initial learning rate | 0.01 |
| Lower boundary of learning rate | 0.001 |
| Time-history learning rate reduction | 10 |
| Kinetic energy     | 0.9 |
| Interval center point | 0 |
| Interval offset    | ±0.5 |
| Minimum relative change in training errors | 0.0001 |
| Minimum relative change in training error rate | 0.001 |
| Hide layer activation function | Hyperbolic tangent function |
| Output layer activation function | Identity function |
Figure 1 Schematic Diagram of the Neural Network

Table 7 Evaluation Form of Neural Network Predicted Value

| Relative Error | Relative Error is Less Than 10% | Relative Error is Less Than 20% | Relative Error is Less Than 30% | Relative Error is Less Than 40% |
|----------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                | 31.38%                          | 58.47%                          | 76.52%                          | 90.29%                          |

4. Conclusion

Table 8 Comprehensive Prediction Value Evaluation Form

| Algorithm                    | The Relative Error is Less Than 10% | The Relative Error is Less Than 20% | The Relative Error is Less Than 30% | The Relative Error is Less Than 40% |
|------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| Multiple linear regression   | 24.83%                               | 45.15%                               | 62.75%                               | 74.27%                               |
| ARIMA time series            | 22.48%                               | 43.56%                               | 60.21%                               | 71.12%                               |
| Neural network algorithm     | 31.38%                               | 58.47%                               | 76.52%                               | 90.29%                               |

It can be concluded from the above table that the prediction model constructed by the neural network algorithm is much higher than the other two algorithms. From the essence of the algorithm and the prediction results, the essence of the time series is the prediction of the future value through the process of self-regression and moving average. Its prediction is based on the identification of the “long-term trend”, the “seasonal variation”, and the “cyclic variation”, and is poor for the “irregular variation”. The problem of multiple linear regression is not a linear relationship between the “Complaints” and the “Other customer claims” in the reality, and the way of regression fitting does not really embody the mapping relation between the independent variable and the dependent variable. Similarly, in the process of constructing the model, the choice of the independent variable is a difficult problem. Gray correlation in a large number of independent variables is not much different from each other. A large number of independent variables can not be well fitted, and a small number of selected independent variables can not be accurately predicted. The neural network algorithm is based on the non-linear model between the independent variable and the dependent variable, which reflects the more and more complex mapping relationships between independent variables and dependent variable through multi-layer network structure. At the same time, the error inverse propagation algorithm of the BP neural network gives it a better ability to study and train.

By the full text, the BP neural network is the optimal algorithm for constructing the customer complaint prediction model from the perspective of the accuracy of numerical example prediction, and the degree of fitting between the independent variable and the dependent variable.
References

[1] Liu Yingying, Tian Xue, Yao Jia, and so on. Customer complaint management for e-commerce[J]. E-commerce, 2015 (1): 54-54.

[2] Yan Wei, Cheng Chao, Xue Bin, and so on. Monthly sales forecast method combining X12 multiplication model and ARIMA model[J]. Power system and its automation journal, 2016, 28 (5): 74-80.

[3] Fu Qianrao. Study on the Method of the Prediction of the Haze Based on the Multiple Linear Regression[J]. Computer Science, 2016, 43 (S1): 526-528.

[4] Chen Xinyun, Jiang Yongkang, Li Muyuan, and so on. Research on the Demand for Single-Site Dispatch of Public Bicycle based on BP Neural Network[J]. Traffic Standardization, 2016, 2 (3): 30-35.

[5] Han Chao, Song Su and Wang Cheng-hong. Real-time Adaptive Prediction of Short-term Traffic Flow based on ARIMA Model[J]. Journal of System Simulation, 2004, 16 (7): 1530-1532.

[6] Chen Yongsheng and Song Lixin. Multiple Linear Regression Modeling and the Application of SPSS Software[J]. Journal of Tonghua Normal University, 2007, 28 (12): 8-9.

[7] Liu Xiaotong. BP Neural Network Input Layer Data Normalization Research[J]. Mechanical engineering and automation, 2010 (3): 122-123.