Optimize BP Neural Network Structure on Car Sales Forecasts Based on Genetic Algorithm

Tang Jun [1,a], Wu Qing [2,b,*]

1 China Tobacco Yunnan Industry Co., Ltd. Information Management Division

2 Kunming Health Information Center, Yunnan Province

[a] tang_min213@163.com, [b] 408537183@qq.com

Keywords: prediction, (BP) neural network, linear correlation, genetic algorithm

Abstract. In order to improve the ability of BP neural network to fit complex functions, we improve the structure of the BP neural network and optimize the weights and thresholds of structure of the BP neural network based on genetic algorithm, then, training the BP neural network model to improve its capability, so, we can apply the model to the automobile sales forecasting system. We compare the prediction accuracy with the traditional BP neural algorithm, it shows that this method obviously fits the data better and has higher prediction accuracy to dates with significant linear correlation.

Introduction

Currently, in the field of automotive sales, car sales forecasting methods represented by time series prediction, such as linear regression， seasonal prediction method, etc. Linear regression method can reflect the trend of the sales data changes over time [1]. Seasonal prediction method can effectively reflect the characteristics of sales volume fluctuate with the seasons. However, car sales are also affected by external factors such as market environment change、the economic crisis、the rise in oil prices, policies of purchaser tax concessions of small-engine car and so on [2]. When thinking of the external factors bringing greater market volatility, simple time-series models are often difficult to make accurate predictions. So, on the basis of collection、 analysis and mastering information of external factors efficiently, reflecting the identified and quantified factors in the whole forecasting process, then increasing the amount of information of predictive models that can be used, and improving the prediction quality of prediction model, the sales forecast rationality and accuracy will get a larger improvement. This study intends to establish an integrated of multiple regression algorithm based on GA, BP neural network. The proposed method makes full use of the advantages of genetic algorithm、BP neural network and multiple regression [3], abandon their defects when used alone [4].

The general idea of modelling

First of all, the analysis of correlation divides the data that to be processed into two parts which include the linear correlation and non-linear correlation, then, dealing with them on the basis of the characteristics of BP neural network and multiple regression , so as to maximize the use of their respective advantages. Finally, mixing the BP neural network and multiple regression together then take advantage of the genetic algorithm to optimize connection weights, weight thresholds and multiple regression values of BP neural network.

Correlation analysis

The correlation coefficient of Pearson is used to measure whether the two data sets on a line, and measure the linear relationship between the variables of fixed distance [5], when two variables are normal continuous variables and a linear relationship between the two, we can denote the degree of

correlation between the two variables as product-moment correlation coefficient, it is mainly Pearson correlation coefficient. The formula is:

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^{\,2} - (\sum x_i)^2}\sqrt{N\sum y_i^{\,2} - (\sum y_i)^2}} \quad (1)$$

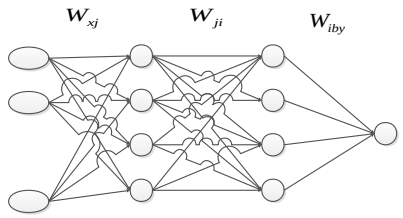
The greater the absolute value of correlation coefficient, the stronger the correlation, the closer to 1 or -1, the stronger the correlation, the closer to 0, the weaker the correlation.

BP neural network

BP network consists of an input layer, an output layer and one hidden layer at least [6]. When the signal is input, it transmits to the hidden layer nodes first, after the effect of function, transmitting the signal of output of hidden layer nodes to the output layer nodes. Then processing，the type of output always choose “s” function, such as:

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

In the following example，there are M inputs of the input layer、 an output of the output layer and two hidden layers each contain J and I neurons. Here, we detail the principle of BP neural network algorithm [7].



The first layer X is the input layer, nonlinear relevant data X_{i} and $i \in \{N+1, N+2...N+M\}$ as inputs, the second J and the third layer I are hidden layers, the forth layer BY is the output layer of Bp network. Layers between each node connection, Interlayers do not connect, connection has a weight of W.. If using $f(x) = \frac{1}{1+e^{-x}}$ as action function of nodes, the output of the node is:

$$Y_i^I = f(\sum_j W_{JI}^{ji} Y_j^J - B_i^I) \quad (3)$$

Where, W_{JI}^{ji} is the weight of the node and each node of the upper layer. Y_j^J is the output of each node in the upper layer, B_i^I is the threshold of the node.

According to the learning phase of the traditional BP neural network algorithm, the network error is:

$$E = (t - BY) * BY(1 - BY) \quad (4)$$

And t is the output of expectations. We compare the network error with the maximum allowable error, if it is satisfied, then ends. Otherwise, correcting each node weights according to the following formula.

$$W_{JI}^{ji}(K+1) = W_{JI}^{ji}(K) + \eta E_j^{\ I} X_j \quad (5)$$

Where, $W_{JI}^{ji}(K+1)$ is the weight in the next time, $W_{JI}^{ji}(K)$ is the connection weights in this time, η represents the efficiency of learning, $E_j^{\ I}$ is the node error. The formula is:

$$E_j^{\ I} = \frac{\partial E}{\partial W_{JI}^{ji}(K)} \quad (6)$$

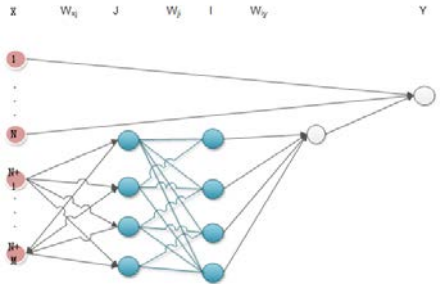
$$Y(k)' = \xi_0 + \xi_i X_i + \xi_{i+1} BY, i \in \{1, 2...N\} \quad (7)$$

The model error shows as follow:

$$e(k) = (Y(k)' - Y(k))^2 k \in \{1, 2...U\} \quad (8)$$

Optimization of the improved genetic algorithm model

The genetic algorithm is a searching method based on the natural selection and the natural genetic which are a kind of the biological evolution mechanism [8]. The genetic algorithm works like that it begin to search each individual in the process group from a set of randomly generated initial solutions, and the result is a solution of a question, and called as chromosomes. The chromosomes are evolving in the subsequent iteration, and this is called the genetic. The genetic generates the generation groups mainly through the selection mathematical operation of crossover and mutation. After several generations of evolution like that, the algorithm converges to the best chromosome, and called as the optimal solution of the problem.



The minimum network error be treated as the objective of optimization, by using the genetic algorithm to optimize the weights and the threshold of the BP network and the regression coefficient ξ. Here we will make a detailed presentation：

i. Detailed introduction

According to the formula $X^{'} = \frac{2(X - X_{\min})}{X_{\max} - X_{\min}} - 1$ and $Y^{'} = \frac{2(Y - Y_{\min})}{Y_{\max} - Y_{\min}} - 1$, the original sample values X , Y be normally processed as the model input data. The data after normalization processing will be $[-1,1]$. Where the $X^{'}$ 、 $Y^{'}$ are the data after normalization, $Y_{\min}$ and $X_{\min}$ are the minimum value of the original data, $Y_{\max}$ and $X_{\max}$ are the maximum value of the original data.

ii. Parameter set

W_{JI}^{ji} 、 B_i^I and ξ are the parameters of the model which consisting of a set of parameters denoting as $C = \{W_{JI}^{ji}, B_i^I, \xi\}$, C is chromosome and C_i is the single gene.

iii. Individual fitness

$f_i^t = E_{\max} - E(C_i^t)$ is an individual fitness function, f_i^t represents the fitness of the t generation of the i-th individuals, $E_{\max}$ is the maximum error of the system, $E(C_i^t)$ is the individual systematic error of C_i^t. The function of fitness ensures the fitness of individuals is positive. At the same time record the value of maximum individual calculated of all individuals.

iv. Selection

Calculate the probability of selecting according to individuals fitness value which is $P_i^t = \frac{f_i^t}{\sum_i f_i^t}$, randomly selecting a portion of an individual chromosome from $pop(t)$ to the next generation $pop(t+1)$ according to selection probabilities. The selected individual chromosomes tentatively called an intermediate generation $mespop(t)$ which regard as the object of the following genetic operations (crossover and mutation).

v. Cross algorithm

Assumed $P_i = (\mathrm{p}_1^i, \mathrm{p}_2^i, ..., \mathrm{p}_i^i)$ and $P_j = (\mathrm{p}_1^j, \mathrm{p}_2^j, ..., \mathrm{p}_j^j)$ were the two parent bodies to cross, $C_j = (c_1^j, c_2^j, ..., c_j^j)$ and $C_i = (c_1^i, c_2^i, ..., c_i^i)$ were hybridized individual. N independently of uniformly random vectors generated within（0, 1）, n is the chromosome length (dimension of individual variables), they are $\mathrm{A}_i^1 = (\mathrm{a}_1^1, \mathrm{a}_2^1, ..., \mathrm{a}_n^1)$ and $\mathrm{A}_i^2 = (\mathrm{a}_1^2, \mathrm{a}_2^2, ..., \mathrm{a}_n^2)$, $\mathrm{i} \in \{1, 2, ..., \mathrm{n}\}$.

Then: $C^1 = P^1 + A^1(P^2 - P^1)$ （9）

$C^2 = P^1 + A^2(P^2 - P^1)$ （10）

Simulation and results analysis

We choose MATLAB as a computer simulation platform to explore the algorithm ability of dealing with the linear and nonlinear data. We take the method of multiple linear regression、 the traditional BP neural network and the proposed algorithm to fit equations of multiple linear、multivariate nonlinear、 linear and nonlinear partial. Equations of multivariate nonlinear:

$$Y = 0.4x_1 + 0.8 * x_2 + 0.3x_3 + 0.7x_4 + 0.9 \quad (11)$$

Equations of multiple linear:

$$Y = \sin(0.3x_1) + \cos(0.8x_2)^2 + \sin(0.1x_1x_2) + \cos(x_1^2) \quad (12)$$

Equations of linear and nonlinear partial：

$$Y = \sin(0.3x_1) + \cos(0.8x_2)^2 + \sin(0.1x_1x_2) + \cos(x_1^2) + 0.4x_3 + 0.8 * x_4 + 0.3x_5 + 0.7x_6 + 0.9 \quad (13)$$

Sample data

Input data of Sample $x_i \in [-1,1]$, the total number of samples is 20, sample values are randomly generated twenty groups of data：

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
-0.8808	0.6363	0.9459	-0.8331	-0.8791	-0.4160	-0.2552	-0.8946	-0.1645	0.3962	-0.9348	-0.0785	-0.6182	-0.2308	0.6488	0.8126	-0.1495	0.1970	-0.8624	0.4367
0.3639	0.6351	0.2980	-0.7337	-0.2015	-0.1367	-0.6038	0.4757	0.9661	0.3331	0.1224	0.9633	-0.1435	0.1660	0.9653	0.7593	-0.3746	-0.0582	-0.3608	0.9373
-0.9151	0.4449	0.6007	-0.6532	0.0538	-0.9690	-0.0206	-0.4618	-0.3971	-0.6437	0.7637	-0.6872	-0.0360	-0.4964	0.4605	0.6355	-0.6770	0.3919	0.0617	0.0627
-0.8571	-0.7003	-0.0924	-0.2181	-0.1664	0.9681	-0.3210	-0.1543	0.4022	-0.7440	0.3384	0.7110	-0.7588	-0.4191	-0.31...	-0.47...	-0.6425	0.3998	0.3089	-0.3497
0.0433	0.3192	-0.1352	0.6628	0.3137	-0.6657	0.9033	0.0957	0.3327	0.9982	-0.6191	0.2895	0.1790	0.2342	0.1681	0.1887	-0.1542	0.2771	-0.1848	-0.7887
-0.8065	0.0372	0.6506	0.6067	0.2559	-0.7876	0.8407	0.8855	0.0783	-0.6578	-0.2622	-0.2475	-0.5476	-0.4694	-0.78...	-0.95...	-0.8115	-0.9328	0.6400	0.2219

Selecting the value of the first row to the fourth row as input of formula 11, Y is:

-0.0357	1.3059	1.6323	-0.3688	0.2868	1.0112	0.0840	0.6762	1.7695	0.6110	1.0900	1.9308	-0.0040	0.4982	1.8513	1.6882	-0.1123	1.3297	0.5011	1.5985

Selecting the value of the first row to the second row as input of formula 12, Y is:

1.3383	1.9126	1.8781	1.2758	1.4473	1.8543	1.7211	1.2507	1.4469	2.0502	1.3441	1.4836	1.7392	1.9081	1.6818	1.7670	1.8734	2.0550	1.4302	1.6890

Selecting the value of the first row to the second row as input of formula 13, Y is:

0.6350	2.5521	3.3593	2.3635	2.5089	2.3902	3.2155	2.4911	2.6644	1.9365	2.4510	2.5912	1.6882	2.0159	2.0176	1.9265	1.3743	2.8618	2.9946	2.2530

Simulation result

The following figure 1、2、3 are formula 11、12、13 Simulation results of multiple linear regression、BP neural Network and the improved algorithm.

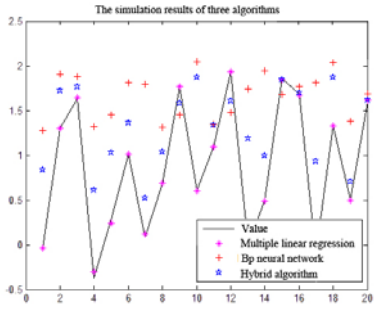


Figure 1

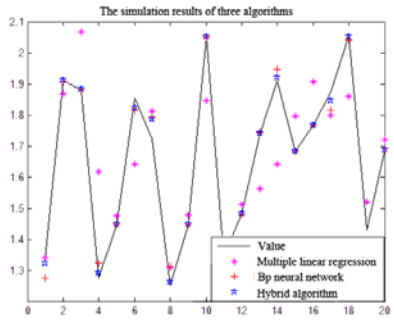


Figure 2

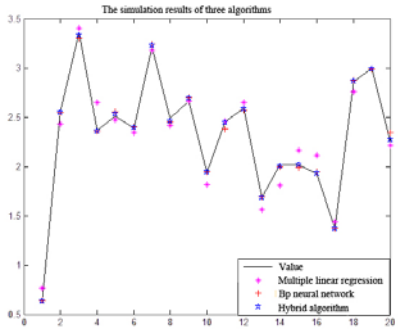


Figure 3

Simulation results analysis

We can draw the following conclusions according to the results of simulation:

When the fitting linear equations appear in multiple linear regression, the error is small, however, when fitting nonlinear equation, the error is large. When the fitting nonlinear equations appear in BP neural network, the error will be smaller, it is difficult to fitting linear equation. Improved algorithm combines the advantages of multiple linear regression and BP neural network algorithms，error of improved algorithm is much smaller than error of multiple linear regression when fitting a nonlinear equation, and the ability of fitting a linear equation is better than BP neural network

Conclusion

Improved algorithm is much better than multiple linear regression or BP neural network in the ability of fitting complex functions. The factors influencing the sales larger and other factors use weighted addition to obtain sales value, at last, optimizing the weights based on genetic algorithm, this method obviously fits the data better and has higher accuracy of prediction to dates with significant linear correlation. That is the purpose of combining the BP neural network and linear regression.

Acknowledgements

Thanks Wu Qing as corresponding author of this article, he made a lot of contributions to this article.

References

[1] M. Kumaresan,P. Riyazuddin. Factor analysis and linear regression model (LRM) of metal speciation and physico-chemical characters of groundwater samples[J]. Environmental Monitoring and Assessment . 2008 (1-3)

[2] Jinwei Cui, Prediction Of China's car market segmentation and development[D], Jinlin, JinlinUniversity,2004(In Chinese)

[3] Li Xiaofeng，Xu Jiuping ,Wang Yinqing ,et al. The establishment of self2adapting algorithm of BP neural network and its application[J] . Systems Engineering - Theory & Practice ,2004 , 24 (5) :1 - 8. [1]

[4] Schaffer J D, Whitley D. Proceedings of the Workshop on Combinations of Genetic Algorithms and Neural Networks 1992[M].Los Alamitos, CA: The IEEE Computer Society Press,1992.

[5] ZHANG Yu-lei，DANG Yan，HE Ping-an. Quantitative analysis of the relationship of biology species using pearson correlation coefficient [J].Computer Engineering and Applications，2005，33: 79-82．

[6] Chen Xiaoqian ,Luo Shibin ,Wang Zhenguo , et al. Research on preprocessing and post processing of the application of BP neural network[J]. Systems Engineering - Theory & Practice 2002 , 22 (1) :65 - 70.

[7] Harpham C et al (2004).A review of genetic algorithms applied to training radial basis function networks [J].Neural Compute App 13(3):193–201

[8] David J, Frenzel J F. Trainning Product Unit Neural Networks with Genetic Algorithms[J].IEEEExpert-Intelligent IEEE Expert-Intelligent Systems & Their Applications,1993,(05):26-33.