Research on automobile sales prediction model based on BP neural network

Yan Zhou[a,*], Jun Zhou[b]
[a]School of Mechanical and Electrical Engineering, Wuhan Business University, Wuhan 430056, China; [b]China Automotive Technology and Research Center (Wuhan), Wuhan 430056, China

ABSTRACT

This paper established a theory framework of the correlation between consumers' web information search and related products' sales. Taking the Chinese customers' search behavior using Baidu search engine and the search data left during the decision-making process, this paper built up and filtered a search keyword thesaurus with high correlation of automobile sales and leading time difference. Then a brand automobile monthly sales prediction model was set up based on BP Neutral Network. On this basis, this paper took a specific automobile model for example and predicted its monthly sales for a month. The predicted results showed that the absolute average percentage error was 5.6%, which was 0.5% lower than the MAPE model by improved principal component analysis, and the prediction accuracy was improved. The validity and rationality of the model were verified. This paper provides a new idea for product sales forecasting of automobile enterprises, and also can be used as a reference for other industries.

Keywords: BP (back propagation) neutral network, automobile sales, prediction model

1. INTRODUCTION

There have been abundant research results on market demand predictions and some traditional prediction methods have been gradually improved and perfected after being put forward one after another. In general, current prediction methods are separated into two kinds: qualitative prediction methods and quantitative prediction methods[1-3]. Qualitative prediction methods include empirical judgment method, expert judgements method (Delphi method), subjective probability method, et al. And there are more types of quantitative prediction methods, mainly including classic econometric model method and methods using intelligent algorithms rising in recent years such as ANN (Artificial Neural Networks) prediction method, SVM (Support Vector Machine) prediction model[4-5].

In 1985, Rumelhart et al. developed a EBP (Error Back Propagation, in short BP) algorithm, which solved the calculation issue of the connection weight between multi-layer connectors and greatly improved the availability of Neutral Network model[6]. In practical applications, 80%-90% artificial neural network models are the variation forms based on the core idea of BP Neutral Network. The training process of BP Neutral Network is the process of continuously adjusting the connection weight between neurons based on sample set, in which, the learning is trained with a teacher and the sample set is composed by vector pairs in the form of (input vector, ideal output vector). All the vector pairs should be the actual operation results of the system to be simulated in the internet. And they can be collected by real operation system[7].

2. BP NEUTRAL NETWORK MODEL

2.1. BP neutral network structure

BP Neutral Network usually contains at least three layers: input layer, hidden layer, and output layer as shown in Figure 1. Each layer consists of a certain number of neurons, which is the smallest information processing unit. Input layer provides information from the outside and passes the information to the middle layer through neurons. The information is processed in this layer. The structure of the middle layer can be increased or decreased according to the operation requirements, at least one layer. Processed information will be transferred to the output layer. The whole information transmission process is called forward propagation. When the error between actual output and expected output does not satisfy the requirements, neutral network will proceed error back propagation based on the error gradient descent training

*yan974248850@163.com; #1165586311@qq.com

International Conference on Computer Application and Information Security (ICCAIS 2021), edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122601J · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637682

method, adjusting the connection weights between neurons layer by layer till the input layer. This is the learning method of neutral network: by information flowing forward and backward repeatedly until the error meets the requirements or the cycle reaches the maximum learning times[8-9].

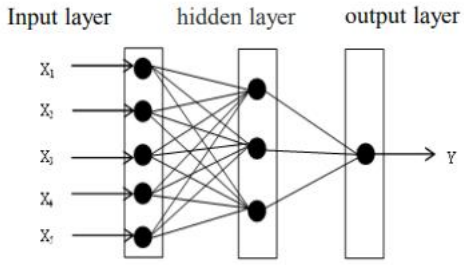

Figure 1. BP Neutral Network model structure.

2.2. BP neutral network algorithm

The main learning steps of the BP Neutral Network algorithm are[10]:

(1) Network initialization

We suppose the node numbers of each layer are *n, l, m*. The weight from input layer to hidden layer is w_{ij}, while the weight from hidden layer to output layer is w_{jk}. The bias between input layer and hidden layer is a_j, while the bias between hidden layer and output layer is b_k. Learning rate is η. Error function is set as e, as well as operation error ε and cycle upper limit M. The initial value of the connection weight of neurons is set between (-1, 1).

Excitation function $g(x)$ takes the Sigmoid function in the form of

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

(2) Output information from hidden layer

$$H_j = g\left(\sum_{i=1}^{n} \omega_{ij} x_i + a_j\right) \tag{2}$$

(3) Output information from output layer

$$O_k = \sum_{j=1}^{l} H_j \omega_{jk} + b_k \tag{3}$$

(4) Error calculation

Error equation is

$$E = \frac{1}{2}\sum_{k=1}^{m}\left(Y_k - O_k\right)^2 \tag{4}$$

where Y_k is the expected output; O_k is the actual output; the error between them is

Partial derivative is calculated by

$$e_k = Y_K - O_K \tag{5}$$

$$E = \frac{1}{2}\sum_{k=1}^{m} e_k^2 \tag{6}$$

where i=1,2…n, j=1,2…l, k=1,2…m.

(5) Weight update

The update equation for weight is

$$\begin{cases} \omega_{ij} = \omega_{ij} + \eta H_j (1 - H_j) x_i \sum_{k=1}^{m} \omega_{jk} e_k \\ \omega_{jk} = \omega_{jk} + \eta H_j e_k \end{cases} \tag{7}$$

(6) Bias update

The update equation for bias is

$$\begin{cases} a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^{m} \omega_{jk} e_k \\ b_k = b_k + \eta H_j e_k \end{cases} \tag{8}$$

The bias from hidden layer to output layer is improved by

$$\frac{\partial E}{\partial b_k} = (Y_k - O_k)\left(-\frac{\partial O_k}{\partial b_k}\right) = -e_k \tag{9}$$

So the update equation for bias is

$$\frac{\partial E}{\partial a_j} = \frac{\partial E}{\partial H_j} \cdot \frac{\partial H_j}{\partial a_j} \tag{10}$$

3. SAMPLE DATA SELECTION

3.1. Automobile sales data selection

Automobile sales data used in the paper comes from the statistical zone of China Association of Automobile Manufacturer (http://www.caam.org.cn/). The 36 months' monthly sales data from Jan. 2018 to Dec. 2020 of Langyi model of SAIC VOLKSWAGEN was selected for model prediction analysis and afterwards tests as shown in Table 1. Due to the COVID-19 pandemic, automobile sales during Feb. 2020 and Mar. 2020 have been affected and decreased suddenly. Therefore, instead of the actual sales of the two months, the expected actual sales have been calculated by quadratic exponential smoothing method, which are 34858 and 35678 respectively, taking these replacements data for neural network training[11].

Table 1. Langyi model's automobile monthly sales during 2018-2020 (unit: vehicle).

Year	Jan.	Feb.	Mar.	Apr.	May	Jun.
2018	44202	25460	38914	30449	40570	41653
2019	58537	33900	49039	36112	35517	33817
2020	35898	34858	35678	35505	36317	34907
Year	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
2018	33484	38946	44579	43974	46241	40320
2019	35629	31882	48226	40322	51567	62644
2020	35925	45549	39595	40984	48278	36369

3.2. Keyword selection

According to Baidu's investigation and research report on netizens, consumers will use search engine for related products' information after they get the intention of buying an automobile and the most commonly used search keyword is the title of the model. Besides, consumer's focuses are mainly on brand, manufacturer, price, performance and other information[12]. This paper takes factors like China's automobile consumption environment, usage environment and consumer's recognition on automobile products, and expands the keywords by Group discussion method and Baidu searching index recommendation on the basis of the above-mentioned keyword categories. Part of the selected initial keywords is shown in Table 2.

The historical search data towards the keyword in the initial keyword thesaurus is collected but the keywords with data losing and not included in the thesaurus are eliminated. The Cross Correlation function of SPSS software is used for keywords Baidu index and Langyi monthly sales data for correlation test and time difference correlation analysis. To enhance the quality of the research data and simplify the model complexity, keywords with the correlation coefficient less than 0.5 are eliminated. Thirteen keywords are reserved for afterwards model training as shown in Table 3.

The results of correlation analysis and time difference analysis have verified the previous idea that most of the keywords are ahead of the sales index in correlation analysis, only part of the keywords fall behind the sales index. Besides, search data of the internet keywords not only positively but also negatively relates with automobile's monthly sales. That is to say, with the increase of the internet keywords with positive correlation, the afterwards sales will increase correspondingly, so as the keywords with negative correlation.

4. MODEL STRUCTURE AND PREDICTION RESULT ANALYSIS

4.1. BP neutral network structure

BP Neutral Network has three layers, with 13 input layer nodes, 1 output layer node. Based on empirical formula f = 1.5mn (f is the node number of hidden layer, n and m are the numbers of input neurons and output neurons). The nodes in hidden layer are preliminarily set as 5. By using Trial and Error Method, different numbers of nodes in hidden layer have been reset for network training. And the results show that 10 nodes in hidden layer are the best. Node transfer function uses the tangent sigmoid transfer function tansi, logarithmic sigmoid transfer function losi and linear transfer function purelin. Training function uses Momentum Back Propagation and dynamic adaptive learning rate on integrated stochastic gradient descent BP algorithm trainlm. And the BP neutral network is shown in Figure 2.

Table 2. Part of the selected initial keywords.

Perspective	Category	Initial keywords
In micro level	Automobile model	Langyi, Langyi model of SAIC VOLKSWAGEN, New Langyi, Langyi 1.6l, Langyi 1.4t, New model of Langyi
	Automobile brand/ manufacturer	VOLKSWAGEN Automobile, SAIC VOLKSWAGEN, Shanghai VOLKSWAGEN, SAIC VOLKSWAGEN Automotive Ltd. Co.
	Price and description	Langyi price, How about Langyi, Langyi pictures, Langyi configuration, Langyi performance, Langyi quality, Which car to choose with 120 thousand yuan
	Competing models	Xuanyi, Suteng, Yinglang, Baolai, Langxing, Family car, Compact car
	Automobile website	Autohome, xcar, pcauto, yiche, official website of SAIC VOLKSWAGEN
	Sales and services	VOLKSWAGEN 4S shop, VOLKSWAGEN automobile's after-sales service, automobile maintenance, automobile repair, autocare service
	Others' evaluation	Langyi forum, Langyi Bar, Langyi car club, ifxtx forum
In macro level	Macro influencing factors	Tax on automobile purchase, gasoline price, automobile licensing, automobile loan, automobile policy, preferential policies for automobile purchase

Table 3. Keywords' correlation coefficient and leading order.

Keywords	Correlation coefficient	Leading order
VOLKSWAGEN Langyi	0.623	1
Langyi 1.4t	0.534	9
Langyi	0.557	6
Langyi price	-0.537	4
How about Langyi	0.504	1
Which car to choose with 150 thousand yuan	0.555	1
VOLKSWAGEN 4S shop	0.507	5
Autohome	0.598	4
xcar	0.532	7
yiche	0.521	10
Tax on automobile purchase	0.588	2
Xuanyi	-0.527	0
Suteng	-0.521	8

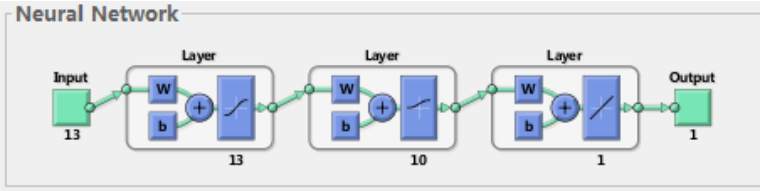

Figure 2. BP Neutral Network structure.

4.2. Prediction result analysis

The monthly search data of each of the 13 selected keywords is dislocated and aligned with monthly sales data according to the time difference. Sales data from Jan. 2018 to Sept. 2020 is aligned with 33 months' keywords sales data, both taken as the training data and input into Matlab. In the 5th training, the set error range is reached out as shown in Figure 3. And the total training results are shown in Figure 4, in which the curves show the actual sales, square point means the neural network fitting value and the network training error is small.

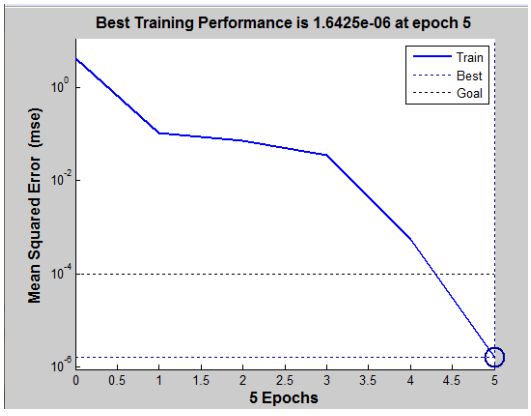

Figure 3. Training process and error performance curve.

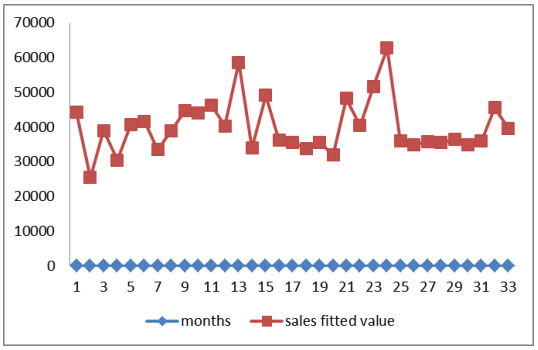

Figure 4. Comparison between sample and network output.

By inputting 13 keywords and the corresponding internet search data between Oct. 2020 and Dec. 2020, the predicted sales for the three months using already trained neutral network are shown in Table 4. And the error analysis shows that the absolute mean error of sales is 5.6% using this model predicting 3 months' sales. And the prediction subject is specific model, which means the research is valuable in practical application. One more thing should be pointed out is that the timeliness of the prediction model depends on the number of leading periods of network keywords to actual sales and due to the keywords' minimum leading period is 1, so the model's prediction leading time is one month.

Absolute Percent Error (APE) and Mean Absolute Percent Error (MAPE) are the average values of absolute error, which can reflect the real situation of the predicted errors. MAPE shows the model's prediction accuracy. The smaller of the MAPE, the higher the prediction accuracy. The definitions are

$$\mathrm{APE} = \left|\frac{\hat{x}_t - x_t}{x_t}\right| \tag{11}$$

$$\mathrm{MAPE} = \frac{1}{n}\left|\frac{\hat{x}_t - x_t}{x_t}\right| \tag{12}$$

Table 4. Prediction result and error analysis.

	Month			**MAPE**
	Oct. 2020	**Nov. 2020**	**Dec. 2020**	
Actual sales	40984	48278	36369	
Predicted sales	41935	47515	38121	
APE	8.3%	4.2%	4.2%	5.6%

4.3. Model improvement based on principal components analysis

The 13 selected keywords have linear or non-linear relations with each other, which interferes the model's training and prediction process with the repeated information. Therefore, by principal components analysis, collinearity indices are synthesized by transferring multiple indices into a few comprehensive indices with more information and stronger interpretation ability to simplify the model and improve the prediction accuracy.

The factors affecting the automobile sales are calculated by SPSS and analyzed by principal components analysis method. After the analysis of the original internet search data, two principal components have the characteristic roots bigger than 1. And as to the default settings of SPSS, the first two main components' accumulative variance contribution exceeds 73%, which can explain the information contained in the original data. And the other components containing little information have been abandoned.

On the basis of the above-mentioned BP Neutral Network prediction model, this paper further conducts the principal components analysis on the keyword search indices affecting automobile sales to eliminate the collinearity between different search indices. Taking the selected two main components as the new network input, the input layer nodes are modified to 2. Based on empirical formula and comparative analysis, hidden layer nodes are modified to 2 and other structure and key parameters are kept still. The improved neutral network is shown in Figure 5.

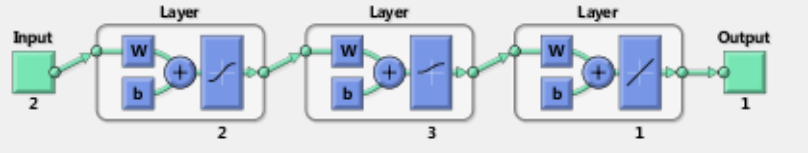

Figure 5. Improved neural network diagram.

Then the reinput data of the model is trained, and the results show that with less data dimension, the structure of the model is simpler and the efficiency of the training is improved. The data fitting effect can be found in Figure 6.

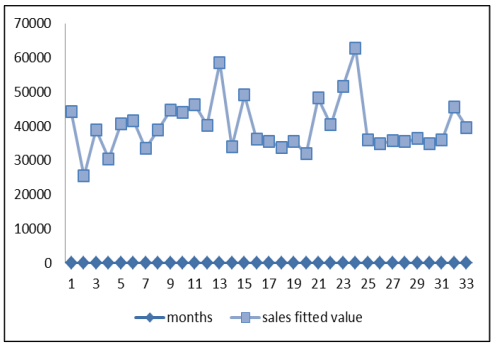

Figure 6. Improved fitting effect.

The comparison of the prediction values from the improved model and the actual values is shown in Table 5. In the table, we can see that the MAPE decreases by 0.5% and the prediction accuracy is increased after the principal component analysis on the keyword internet search data.

Table 5. The prediction results of the principal component analysis on BP neutral network.

	Month			MAPE
	Oct. 2020	Nov. 2020	Dec. 2020	
Real sales	40984	48278	36369	
Predicted sales	41531	47520	36132	
APE	7.3%	4.2%	3.9%	5.1%

5. CONCLUSION

This paper establishes a web search key Thesaurus on the characteristics of automobile products, which can effectively help in selecting web search data for prediction and analysis. By the correlation analysis and time difference analysis on keyword web search amount and the actual sales, the paper certifies that there is a strong correlation and leading time difference between part of the web search data and the actual automobile sales. And this indeed shows the value of the web search data. The paper trains the 33 months' keyword search data and automobile sale data by the BP Neutral Network based automobile sales model, and the trained model is used to predict the afterwards three months' automobile sales. The prediction results show that the absolute mean percentage error is 5.6%, the model's MAPE decreases by 0.5%, and the prediction accuracy is improved. With the good model-fitting degree and prediction accuracy, the effectiveness and rationality of the model is certified.

ACKNOWLEDGMENTS

This research was funded by Scientific Research Project of Wuhan Business University in 2020, "Optimization Design of Operation mode of Spare Parts Reuse and Material Recovery of Retired New Energy Vehicles" (2020KY005).

REFERENCES

[1] Wu, J. and Deng, Y. H., "Intercity information diffusion and price discovery in housing markets: Evidence from google searches," The Journal of Real Estate Finance and Economics, 50(3), 289-306 (2015).

[2] Sharad, G., Hofman, J. M., Sébastien, L., et al., "Predicting consumer behavior with web search," Proceedings of the National Academy of Sciences of the United States of America, 107(41), 17486-17490 (2019).

[3] Sun, W., Teng, X. S. and Ma, N., "Fault diagnosis of relay protection device based on fuzzy Bp neutral network model," Proc. of 2011 13th IEEE Joint Inter. Computer Science and Information Technology Conf. (JICSIT 2011), 02, (2011).

[4] Cha, M. H., Lu, Z. H., Zhai, J. W. and Zhang, F. S., "Using double-suppressed BP neutral network model to predict water quality in Laoha River," Journal of Water Resources and Water Engineering, 29(2), 56-61 (2018). (in Chinese)

[5] Yu, Z., Qin, L., Chen, Y. J. and Parmar, M., "Stock price forecasting based on LLE-BP neural network model," Physica A: Statistical Mechanics and Its Applications, 553, 124197 (2020).

[6] Li, B., Zhang, Y. F., Zhang, S. H. and Li, W. Y., "Prediction of grain yield in Henan province based on grey BP neural network model," Discrete Dynamics in Nature and Society, 2021, (2021).

[7] Sun, X. and Lei, Y., "Research on financial early warning of mining listed companies based on BP neural network model," Resources Policy, 73, 102223 (2021).

[8] Cavalcante, E. S., Vasconcelos, L., de Farias Neto, G. W., Ramos, W. and Brito, R., "Automotive painting process: Minimizing energy consumption by using adjusted convective heat transfer coefficients," Progress in Organic Coatings, 140, 105479 (2020).

[9] Chen, C., Liu, Y., Sun, X. F., Cairano-Gilfedder, C. D. and Titmus, S., "An integrated deep learning-based approach for automobile maintenance prediction with GIS data," Reliability Engineering and System Safety, 216, (2021).

[10] Liang, Y., Jia, Y., Li, J., Chen, M., Hu, Y., Shi, Y. and Ma, F., "Online shop daily sale prediction using adaptive network-based fuzzy inference system," 12th IEEE Inter. Cong. on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), (2020).

[11] Meyer, A., Glock, K. and Radaschewski, F., "Planning profitable tours for field sales forces: A unified view on sales analytics and mathematical optimization," Omega, 105, 102518 (2021).

[12] Kato, T., "Demand prediction in the automobile industry independent of big data," Annals of Data Science, 2, (2020).