ISSN:2790-1661 Volume-5-(2023)

Research on the influencing factors of customer rating based on PCA and RBF neural network

Yiting Yu[1, *, #] and Wanyi Zeng[2, #]

[1]Evergrande School of Management, Wuhan University of Science and Technology, Wuhan, China

[2]School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China

*Corresponding author: wyt3146779315@163.com

[#]These authors contributed equally to this work

Abstract. In the past four decades, China's mobile communication technology has made rapid development and application, changing the way of life of human beings and making it possible for everything to be connected. As the network continues to be built, the network coverage is getting better and better. Therefore, it is particularly important to study the factors that customers influence on the operator's product services and thus further improve the quality of network services. First, cleaning operations such as coding and missing value processing are performed on the data, and specific data are filled based on information gain to fully utilize customer information. Secondly, based on principal component analysis, the features of each influencing factor are dimensioned down, and the cumulative influence contribution rate of each principal component is calculated. The influencing factors that have a cumulative influence contribution rate of more than 80% in the two data affecting the scores are selected, and mathematical models of customer scoring based on relevant influencing factors are established for customer voice service and Internet access service respectively, and other customer scores are predicted accordingly. The models stabilized after 8 and 6 iterations for voice service and Internet service, respectively, and their mean square error was kept at about 2.

Keywords: PCA; Customer scoring; RBF neural network; Particle swarm algorithm.

1. Introduction

In the past four decades, China's mobile communication technology has achieved rapid development and application [1], from the initial comprehensive reliance on European and American standards and technologies, to the development of TD-SCDMA independent standards, to the full penetration of 5G networks in China now, which can be described as continuous innovation all the way to breakthroughs.

China Mobile Communications Group Beijing has asked customers to rate their satisfaction with voice service (network coverage and signal strength, voice call clarity, voice call stability and overall satisfaction) and Internet service (network coverage and signal strength, cell phone Internet speed, cell phone Internet stability and overall satisfaction) respectively, and to compile statistics on the factors affecting customers' voice call and Internet experience, hoping that this The main factors affecting customers' voice and Internet service experience can be analyzed to improve customers' Internet experience.

In this paper, we design and analyze the factors influencing customer ratings based on the idea of dimensionality reduction, and identify the main factors influencing customer ratings. Based on the quantitative analysis, a neural network model is built to predict the customer rating of the attached data.

2. Analysis of the degree of influence of customer scoring

2.1 Data Cleaning

We have coded the data for some common discrete text data. Depending on different cases, we have rounded off, directly filled and filled with mean or plurality for some missing values in the sample data, respectively. For some missing data, we will fill in the missing values in each column of the data based on the principle of information gain. To prevent the results from being affected by the subjective nature of the customer input, we reclassify them according to keywords.

2.2 Construction of a quantitative model of the degree of influence

Having subjected the raw data to data cleaning, we can now build the model on this basis. First, we will introduce some variables as follows.

The degree of influence of each indicator on the score is noted as:

$$A_{m,g_{k,t}} = \begin{bmatrix} |a_{11}| \\ |a_{21}| \\ \vdots \\ |a_{p1}| \end{bmatrix} \tag{1}$$

$$R_{F_{mk},g_{k,t}} = \left|r_{F_{mk},g_{k,t}}\right| \tag{2}$$

$$I_{g_{ij}} = \sum_{m=1}^{M} R_{F_{mk},g_{k,t}} * A_m = \begin{bmatrix} I_{FE_{1,k},g_{k,t}} \\ I_{FE_{2,k},g_{k,t}} \\ \vdots \\ I_{FE_{3,k},g_{k,t}} \end{bmatrix} \tag{3}$$

Where a is the eigenvalue in the construction of principal components; $r_{F_{mk},g_{k,t}}$ is the correlation coefficient of principal component F_{mk} on score g_{ij}; and $I_{FE_{j,k},g_{k,t}}$ is the degree of influence of the j-th indicator on the t-th score value in data k.

2.3 PCA-based data transformation and dimensionality reduction

Let the number of users be n and the number of indicators be p, then a sample matrix x of size $n \times p$ can be constructed, denoted as:

$$x = \begin{bmatrix} x_{1,1,k} & x_{1,2,k} & \cdots & x_{1,p,k} \\ x_{2,1,k} & x_{2,2,k} & \cdots & x_{2,p,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1,k} & x_{n,2,k} & \cdots & x_{n,p,k} \end{bmatrix} = \left(x_{1,k}, x_{2,k}, \cdots, x_{p,k}\right) \tag{4}$$

And then, construct a new set of variables $F_{1,k}, F_{2,k}, ..., F_{M,k} (M \leq p)$ and make them satisfy the following equation.

$$\begin{cases} F_{1,k} = a_{1,1,k}x_{1,k} + a_{1,2,k}x_{2,k} + \cdots + a_{1,p,k}x_{p,k} \\ F_{2,k} = a_{2,1,k}x_{1,k} + a_{2,2,k}x_{2,k} + \cdots + a_{2,p,k}x_{p,k} \\ \vdots \\ F_{m,k} = a_{m,1,k}x_{1,k} + a_{m,2,k}x_{2,k} + \cdots + a_{m,p,k}x_{p,k} \end{cases} \tag{5}$$

Where $a_{m,j,k}$ satisfies: $F_{m_1,k} \neq F_{m_2,k}(m_1 \neq m_2; m_1, m_2 = 1,2,\cdots,M)$ are mutually uncorrelated; $F_{1,k}$ is the one with the largest variance among all linear combinations of $x_1, x_2, \cdots, x_p$, further, that is, for $F_{\mathrm{M},k}$ is the one with the largest variance among all linear combinations of $x_1, x_2, \cdots, x_p$ that are uncorrelated with $F_1, F_2, ..., F_{m-1}$ [5].

To eliminate the effect of the magnitude, we need to normalize the sample matrix, which yields:

$$X = \begin{bmatrix} X_{1,1,k} & X_{1,2,k} & \cdots & X_{1,p,k} \\ X_{2,1,k} & X_{2,2,k} & \cdots & X_{2,p,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1,k} & X_{n,2,k} & \cdots & X_{n,p,k} \end{bmatrix} = \left(X_{1,k}, X_{2,k}, \cdots, X_{p,k}\right) \tag{6}$$

Where, $X_{i,j,k} = \frac{x_{i,j,k} - \bar{x}_{j,k}}{S_{j,k}}$, $\bar{x}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j,k}$, $S_{j,k} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{i,j,k} - \bar{x}_{j,k}\right)^2}{n-1}}$.

Meanwhile, from the standardized sample matrix, its corresponding covariance matrix can be calculated as follows.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{7}$$

Where, $r_{ij} = \frac{1}{n-1}\sum_{q=1}^{n}\left(X_{qi} - \bar{X_i}\right)\left(X_{qj} - \bar{X_j}\right) = \frac{1}{n-1}\sum_{q=1}^{n} X_{qi}X_{qj}$, $R = \frac{\sum_{q=1}^{n}\left(x_{qi} - \bar{x_i}\right)\left(x_{qj} - \bar{x_j}\right)}{\sqrt{\sum_{q=1}^{n}\left(x_{qi} - \bar{x_i}\right)^2 \sum_{q=1}^{n}\left(x_{qj} - \bar{x_j}\right)^2}}$.

Based on this, the eigenvalues λ and eigenvectors a of R can be calculated and expressed as follows.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \tag{8}$$

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \tag{9}$$

Then the mth principal component $F_{m,k}$ of Annex k can be expressed in the following form.

$$F_{m,k} = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \tag{10}$$

In order to avoid the redundancy of the extracted features and reduce the subsequent computation, some components are selected for subsequent analysis based on the cumulative contribution rate in this paper. The contribution rates can be calculated as follows.

$$c = \frac{\lambda_i}{\sum_{q=1}^{p} \lambda_q} (i = 1,2,\cdots,p) \tag{11}$$

Further, the cumulative contribution rate can be expressed as follows.

$$acc = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k} (i = 1,2,\cdots,p) \tag{12}$$

3. Build customer scoring model

3.1 Selection of data features

Based on the results of the quantitative analysis above, we ranked the influencing factors of each rating in descending order according to their degree of influence, and then selected the top n influencing factors with a cumulative influence contribution rate of more than 80% for the following modeling analysis. After calculation, the number of the top n influencing factors with a cumulative impact contribution rate of more than 80% in both data is 5, and their specific factor characteristics are also consistent, as shown in the following table.

Table 1. Factors influencing the selection of the two data

Influencing Factors	Data I	Data 2
Factor 1	Have you encountered network problems	Slow to open web pages or APP pictures
Factor 2	Residential neighborhoods	Slow cell phone Internet access
Factor 3	Subway	Dragon Valley
Factor 4	One party cannot hear during the call	Fantasy Exorcist
Factor 5	Sudden interruption during a call	Watching video lag

3.2 BP Neural Network

BP (BackPropagation) algorithm is one of the most important algorithms in neural network deep learning, a multilayer feedforward neural network trained according to the error backpropagation algorithm, and is one of the most widely used neural network models. Its basic structure is shown as follows.

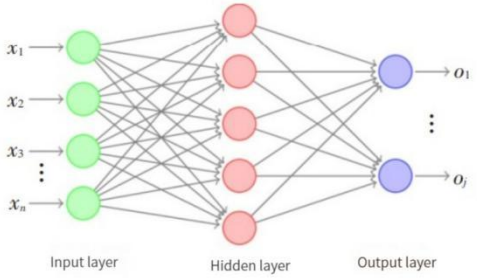

Fig. 1 BP neural network structure diagram

Which contains three main parts, the input layer, the hidden layer and the output layer. Through continuous forward propagation of information and backward propagation of error, the layer

weights are continuously adjusted, which is the learning training of the neural network. When the error of the output is reduced to the desired degree or a predetermined number of learning iterations, the training ends and the BP neural network finishes learning [3].

In this section, the model is trained and performs on the validation set with a total of eight ratings for voice and Internet services, respectively, as shown in the following table.

Table 2. Neural network validation set performance

Voice Services	MSE	Internet access	MSE
Overall satisfaction with voice calls	3.700131238	Overall satisfaction with cell phone Internet access	6.078615555
Network coverage and signal strength	4.086254162	Network coverage and signal strength	6.199173634
Voice call clarity	3.919772466	Mobile Internet speed	5.388426199
Voice call stability	4.108057712	Mobile Internet Stability	6.228937788

From the above table, it can be seen that the model performs poorly for the prediction of the four scores of Internet service. For this reason, we will use RBF neural network and RBF neural network based on particle swarm optimization algorithm for the model construction and prediction of each customer satisfaction score respectively on the basis of traditional BP neural network.

3.3 RBF neural network based on particle swarm optimization

3.3.1 RBF neural network principle

In 1985, Powell proposed the radial basis function (RBF) method for multivariate interpolation [2]. A radial basis function is a real-valued function that takes a value that depends only on the distance from the origin, or it can be the distance to any point c, which is called the centroid.

The RBF (Radial Basis Function) radial basis function network is a single hidden layer feedforward neural network that uses the radial basis function as the activation function of the hidden layer neurons, while the output layer is a linear combination of the outputs of the hidden layer neurons. Its basic structure is shown as follows.

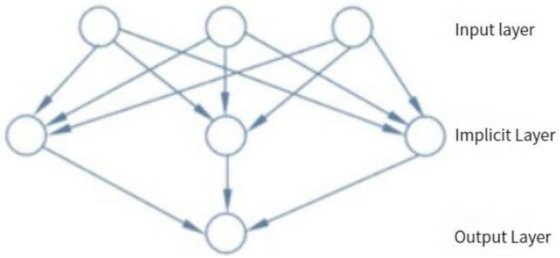

Fig. 2 RBF neural network structure

The general structure is the same as that of an ordinary neural network, but it should be noted that the role of the implicit layer is to map the vector from the low-dimensional p to the high-dimensional h, so that the low-dimensional linearly indistinguishable case to the high-dimensional can become linearly distinguishable, mainly the idea of the kernel function.

In this way, the mapping of the network from input to output is nonlinear; while the network output is linear with respect to the adjustable parameters, the power of the network can be directly solved by a linear system of equations, thus greatly speeding up the learning speed and avoiding the problem of local minima.

The activation function of a radial basis neural network can be expressed as follows.

$$R(x_p - c_i) = exp(-\frac{1}{2\sigma^2} \| x_p - c_i \|^2) \tag{13}$$

Where x_p is the p-nd input sample, c_i is the i-th centroid, h is the number of nodes in the hidden layer, and n is the number of samples or classifications in the output.

The structure of the radial basis neural network yields the output of the network as follows.

$$y_i = \sum_{i=1}^{h} \omega_{ij} exp(-\frac{1}{2\sigma^2} \| x_p - c_i \|^2),\ j = 1,2,\cdots,n \tag{14}$$

Using the least squares loss function representation.

$$\sigma = \frac{1}{p} \sum_{j}^{m} \| d_j - y_i c_i \|^2 \tag{15}$$

3.3.2 RBF neural network principle

Particle swarm algorithm (PSO) is a kind of swarm intelligence algorithm, which simulates a bird in a flock by designing a massless particle with only two attributes: velocity V and position X. Velocity represents the speed of movement and position represents the direction of movement [4].

Each particle individually searches for the optimal solution in the search space, and records it as the current individual extremum, and shares the individual extremum with the other particles in the whole swarm.

The idea of the particle swarm algorithm is relatively simple and is divided into:

1. Initialize the particle population.
2. Evaluate the particles, i.e., calculate the adaptation values.
3. Find the individual extremes P_{best}.
4. Find the global optimal solution G_{best}.
5.Modify the velocity and position of the particles.

With the introduction of the particle swarm optimization algorithm on top of the RBF neural network, we will be able to find the global optimal solution much faster. The following figure shows the performance of the model based on particle swarm optimization for 15 iterations.

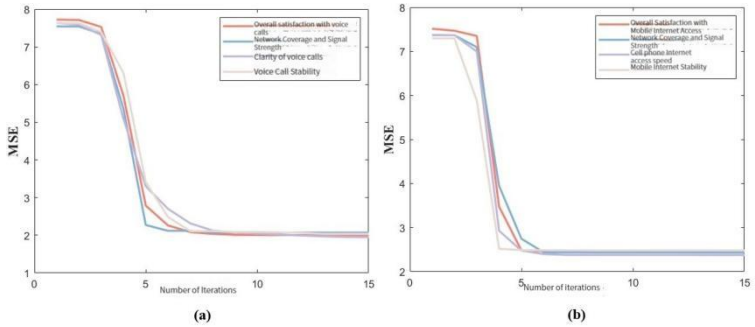

Fig. 3 Performance of RBF neural network validation set based on particle swarm optimization

Based on this, we find that the model stabilizes after 8 and 6 iterations for each user rating of voice service and Internet service, respectively, and its mean square error is kept around 2. Compared with the model performance in the previous section, the performance effect of the RBF neural network based on particle swarm optimization has been greatly improved.

3.4 Model Predictions

Based on the above analysis, we finally used models of optimal RBF neural networks (eight in total) found on the basis of particle swarm optimization to predict each satisfaction score of user data, and some of the prediction results are shown in the table below.

Table 3. Neural network validation set performance

User id	Overall satisfaction with voice calls	Network coverage and signal strength	Voice call clarity	Voice call stability
1	10	10	10	9
2	8	7	8	9
3	10	10	10	9
4	8	7	8	7
5	8	7	8	7
6	7	6	7	6
7	10	10	10	9
8	10	10	10	9
9	10	10	10	9
10	8	7	8	7

4. Summary

In this paper, the data interpolation process uses the information gain-based data grouping model and the plural filling method to interpolate the missing values with high interpolation accuracy. The RBF neural network can approximate any nonlinear function with arbitrary accuracy and has global approximation capability, which fundamentally solves the local optimum problem of BP neural network. At the same time, we applied the particle swarm optimization RBF neural network algorithm, and the optimized algorithm is faster than the traditional RBF neural network in finding the optimal solution and easier to find the global optimal solution.

However, when reclassifying user descriptions and remarks, for some of the statements appear unrecognizable, the statements in this paper are removed, which may have some impact on the accuracy of the data. When the number of training samples increases, the number of hidden layer neurons of RBF network is much higher than that of BP network, which makes the complexity of RBF network increase greatly and the structure is too large, and thus the amount of operations increases.

References

[1] Liu Bing, Tian Cheng, Chen Guiqin. Analysis of the development prospect of communication technology in the era of big data[J]. Network Security Technology and Application,2022(12):171-172.

[2] Huang Xionghua, Zheng Zhenliu, He Minghui, Yang Xiangsong. Health monitoring system based on RBF-BP neural network[J]. Information Systems Engineering,2022(12):43-46.

[3] Chen Z Z Z, He J P, Li F, Hua X M, Huang W R. Gap adaptive process parameter optimization of thin plate P-PAW lap based on BP neural network [J/OL]. Materials Science and Technology:1-7 [2023-03-10].

[4] Zhou, S. L., Li, G. L., Wang, Q. J., Zheng, C. B., Wen, Y.. Research on permanent magnet spherical motor drive strategy based on improved particle swarm optimization algorithm[J]. Journal of Electrotechnology,2023,38(01):166-176+189.

[5] YANG Shan,LI Wenwen,CHEN Jianhong. Study on the safety of spontaneous combustion of sulfide ore based on PCA-RBF network model[J]. Gold Science and Technology,2022,30(6):958-967.