Prediction of concrete compressive strength based on principal component analysis and radial basis function neural network

Zongjun Sun¹, Kai Li¹*, and Zhongyi Li²

¹College of Transportation, Shandong University of Science and Technology, Qingdao, China
²Ningxia Communications Construction Co., Ltd, Yinchuan, China

*Corresponding author e-mail: 825840462@qq.com

Abstract. Aiming at the concrete compressive strength data measured in the field laboratory, the data samples are standardized and the missing and abnormal values of the data samples are detected and processed. Then, the principal component factors of several influencing factors are extracted with the idea of "dimensionality reduction" by principal component analysis, and a new data sample set is established. Finally, the radial basis function neural network model is constructed to simulate the concrete compressive strength. The strength is predicted and the predicted value is obtained. The results show that three principal component influencing factors are extracted by principal component analysis, and the RBF network structure with 50 neuron nodes in the hidden layer is established. The error of the predicted value of concrete compressive strength is less than 4%, which meets the control requirements of engineering test accuracy. The predicted model can be used to predict the compressive strength of concrete.

1. Introduction
It is necessary to establish a prediction model of concrete strength in practical engineering test. It can not only provide reference for concrete mix ratio test, but also infer concrete strength in advance, and provide help for later construction arrangement. The conventional concrete strength prediction model adopts the linear regression method, but now the composition of concrete is more complex, and there are many factors affecting the strength of concrete. Simple linear regression can not adapt to the current concrete compressive strength prediction, so the concrete strength prediction model should adopt the non-linear form. The main prediction methods include SPSS regression analysis, genetic operation tree and neural network prediction. The genetic operation tree can construct the strength equation, but the programming is complex and not suitable for generalization; the SPSS regression analysis method is simple but not accurate enough; the neural network prediction method has high accuracy and simple programming, which is suitable for the construction of concrete strength prediction model [1]. Many experts and scholars have studied the prediction model based on neural network. Piwenshan [2] has established an eight-factor BP neural network prediction model for concrete compressive strength, Guo Yaodong [3] has established a BP neural network-based prediction model for the compressive strength of recycled thermal insulation concrete, and Li Hong [4] has established a BP neural network-based
compressive strength. Degree prediction model. According to the test data of concrete mix ratio with 10 influencing factors, the principal component analysis is used to extract the principal component of these influencing factors. As a new influencing factor of concrete strength, the prediction model is constructed by using radial basis function neural network, which provides a theoretical reference for the practical application of Engineering mix ratio test.

2. Analytical methods and processes

2.1. Data Quality Analysis

Data quality analysis is the premise of data preprocessing. The main task is to check whether there is dirty data in the original data samples. The so-called dirty data refers to the data that does not meet the requirements and cannot be analyzed directly. The common dirty data are mainly missing values and outliers [5].

2.1.1. Missing value analysis. The missing of data mainly includes the missing of records and the missing of some information in records, which will result in the inaccurate analysis results. The missing value of the original data sample will lead to a series of problems, such as loss of some important information, inaccurate grasp of the rules contained in the model, and unreliable output. Common methods for dealing with missing values can be divided into three categories: deletion of records, data interpolation and non-processing [6]. By simply deleting missing values, there are great limitations. By reducing historical data for the integrity of the original data set, a lot of information hidden in the records will be discarded. Lagrange interpolation and Newton interpolation are commonly used in interpolation methods, but interpolation methods are often used in sequential data sets, that is, missing interpolation of regular data sets. Without dealing with missing values, it is directly regarded as a special value, and can be directly analyzed on the data set with missing values.

2.1.2. Abnormal Value Analysis. Outlier value analysis is to check whether there are recording errors and unreasonable data in the original data samples [7]. Ignoring the existence of outliers leads to very serious results. Putting outliers into the calculation and analysis process of the original data sets without deletion will bring adverse effects on the results. Outliers, also known as outliers, refer to individual values in the sample data, whose values deviate significantly from the rest of the normal data values.

2.2. Principal component analysis

2.2.1. Principle of Principal Component Analysis. There are many factors affecting the compressive strength of modern concrete, and there is a certain correlation among them. Therefore, the selection of influencing factors variables needed to establish the model by traditional methods may lead to low prediction accuracy and complex calculation process of the model [8]. Principal Component Analysis (PCA) uses the idea of dimensionality reduction to transform the interrelated numerical variables into a few irrelevant comprehensive indicators. These comprehensive indicators are the principal components of the original multiple influencing factors, each principal component is a linear combination of the original influencing factors, and each principal component is not mutually related. Relevant [9]. Before establishing the prediction model of concrete compressive strength, the principal components of influencing factors of concrete compressive strength are extracted by principal component analysis, and then the principal components are used as influencing factors variables of the prediction model to establish the prediction model.

2.2.2. Principal Component Analysis. The main task of principal component analysis is to calculate the principal component. The calculation steps are as follows:

(1) Standardizing the original variables, eliminating the influence of dimension, and standardizing the original variables;
(2) Calculate the correlation matrix between variables, the eigenvalues and eigenvectors of the matrix;
(3) The characteristic roots are arranged in order from big to small, and the corresponding principal components are calculated respectively.

Another task of principal component analysis is to determine the number of principal components and determine the method:
(1) Accumulated contribution rate: The current cumulative contribution rate of the principal component reaches a specific value (generally more than 89%) and the former principal component is retained;
(2) Characteristic roots: Generally, the principal components of the feature roots are selected.

2.2.3. Relevant Indicators in Principal Component Analysis. (1) Characteristic roots: Indicators of the influence of principal components, that is to say, how many original variables can be carried by introducing the principal components. If the eigenvalue is less than 1, it means that the interpretation degree of principal component is not as good as the average interpretation degree of introducing a primitive variable directly. When determining the number of principal components, the principal component whose eigenvalue is greater than 1 is usually selected.
(2) The variance contribution rate of principal components:

\[ Z_i = \frac{\lambda_i}{\sum_{i=1}^{p} x_i} \]  

In formula: \( \lambda_i \) represents the proportion of variance of principal component \( Z_i \) in all variances. the value, the stronger the ability of principal component \( Z_i \) to synthesize the information of original variables.
(3) Cumulative contribution rate: the cumulative contribution rate of the first \( k \) principal components \( \xi \) is:

\[ \xi = \sum_{i=1}^{k} \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \]  

In formula: \( \xi \) denotes how much information the first \( k \) principal component accumulatively extracts from the original variables.

2.3. Radial Basis Function Neural Network Model
Radial Basis Function Neural Network has good pattern classification and function fitting ability. It is a three-layer forward network. It consists of three-layer network [9]: the first layer is the input layer, and the number of nodes equals the input dimension; the second layer is the hidden layer, and the number of nodes is determined according to the complexity of the problem; the third layer is the transmission layer. Out of the layer, the number of nodes is equal to the dimension of output data [10]. The different layers of RBF neural network have different functions. The hidden layer is non-linear. The RBF function is used as the basis function to transform the input vector space into the hidden layer space, which makes the original linear inseparable problem linear and the output layer linear.
3. Prediction Model of Concrete Compressive Strength and Its Application

3.1. Prediction Model of Concrete Compressive Strength and Its Application

This paper takes the experimental data of concrete mix ratio provided by the construction site laboratory of a Yellow River Bridge Project of Ningxia Jiaotong Construction Company as an example, as shown in Table 1.

### Table 1. Samples of raw data.

| Serial number | X1   | X2   | X3   | X4   | X5   | X6   | X7   | X8   | X9   | X10  |
|---------------|------|------|------|------|------|------|------|------|------|------|
| 1             | 0.45 | 43   | 1.8  | 193  | 58   | 38   | 130  | 5.2  | 843.0| 1118 |
| 2             | 0.43 | 42   | 1.8  | 200  | 60   | 40   | 130  | 5.4  | 819.0| 1131 |
| 3             | 0.41 | 41   | 1.8  | 212  | 63   | 42   | 130  | 5.71 | 793.0| 1140 |
| 4             | 0.43 | 42   | 1.8  | 230  | 58   | 38   | 140  | 5.87 | 812.0| 1122 |
| 5             | 0.41 | 41   | 1.8  | 240  | 60   | 40   | 140  | 6.12 | 787.0| 1133 |
| …            | …    | …    | …    | …    | …    | …    | …    | …    | …    | …    |
| 35           | 0.41 | 40   | 1.8  | 230  | 60   | 40   | 140  | 5.94 | 754.0| 1131 |
| …            | …    | …    | …    | …    | …    | …    | …    | …    | …    | …    |
| 69           | 0.32 | 40   | 3.1  | 400  | 40   | 40   | 139  | 15   | 726  | 1090 |

Among them: X1-X10 denotes water-binder ratio, sand ratio water reducing agent content, cement content, fly ash, slag powder, water, water reducing agent, sand, gravel weight in Xinjing Coal Mine. Standardized processing of sample data. The minimum-maximum normalization method, also known as deviation normalization, is a linear transformation of the original data and mapping the values to intervals \([0, 1]\). The standardized sample data are shown in Table 2.

### Table 2. Standardized data.

| Serial number | X1   | X2   | X3   | X4   | X5   | X6   | X7   | X8   | X9   | X10  |
|---------------|------|------|------|------|------|------|------|------|------|------|
| 1             | 0.95 | 0.833| 0.133| 0    | 0.246| 0.584| 0    | 0    | 1    | 0.666|
| 2             | 0.85 | 0.666| 0.133| 0.028| 0.276| 0.615| 0    | 0.020| 0.856| 0.790|
| 3             | 0.75 | 0.5  | 0.133| 0.076| 0.323| 0.646| 0    | 0.052| 0.700| 0.876|
| 4             | 0.85 | 0.666| 0.133| 0.148| 0.246| 0.584| 0.769| 0.068| 0.814| 0.704|
| 5             | 0.75 | 0.5  | 0.133| 0.188| 0.276| 0.615| 0.769| 0.093| 0.664| 0.809|
| …            | …    | …    | …    | …    | …    | …    | …    | …    | …    | …    |
| 35           | 0.85 | 0.5  | 0.133| 0.104| 0.230| 0.584| 0.384| 0.045| 0.616| 0.704|
| …            | …    | …    | …    | …    | …    | …    | …    | …    | …    | …    |
| 69           | 0.3  | 0.333| 1    | 0.911| 0.276| 0    | 0.692| 1    | 0.299| 0.381|

3.2. Data preprocessing

First of all, the data quality of the original experimental data is analyzed, and the original data set is loaded on the MATLAB platform for processing. After detection, there is no missing value in the original data. The box diagram of the original data sample cloth is shown in Figure 1.

**Figure 1.** Box Diagram.
From the box diagram, it can be seen that the overall quality of the data meets the engineering requirements. Among them, the true values of \( x_3, x_5, x_6, x_8, x_{10} \) are 1, 3, 1, 1 and 8, and the missing values account for 11% of the total data. In the actual engineering test work, the environment is complex, so we can not effectively determine what causes the occurrence of abnormal values, so the abnormal values of these factors are all reduced to 0. The compressive strength data of concrete in 7-day and 28-day age are real and reliable.

3.3. Principal Component Analysis of Influencing Factors

The principal component analysis was used to extract the principal component from the influencing factors, and a new influencing factor of concrete strength was constructed. The data samples were analyzed by SPSS according to the basic steps of principal component analysis, and Bartlett spherical test, principal component results and component score coefficient matrix were obtained.

(1) Bartlett spherical test: The test results are shown in Table 3.

Table 3. Bartlett spherical test.

| KMO Sampling Quantity | 0.377 |
|-----------------------|-------|
| Bartlett sphericity test | Approximate Chi Square | 1437.828 |
| Freedom | 45 |
| Significance | 0 |

Bartlett's spherical test rejects the original assumption of the unit correlation matrix, and its significance is less than 0.001. It is suitable for principal component analysis.

(2) Principal component analysis: The results are shown in Table 4.

Table 4. Statistical Information Result of Principal Components.

| Component | Total | Percentage of variance of initial eigenvalue | Cumulative% |
|-----------|-------|---------------------------------------------|-------------|
| 1         | 4.547 | 45.47                                       | 45.47       |
| 2         | 2.595 | 25.95                                       | 71.42       |
| 3         | 1.12  | 11.2                                        | 82.62       |
| 4         | 0.753 | 7.53                                        | 89.52       |
| 5         | 0.69  | 6.9                                         | 96.42       |
| 6         | 0.205 | 2.05                                        | 98.47       |
| 7         | 0.081 | 0.81                                        | 99.28       |
| 8         | 0.009 | 0.09                                        | 99.37       |
| 9         | 0.001 | 0.01                                        | 99.38       |
| 10        | 0     | 0                                           | 99.38       |

The contribution rate of each principal component and the cumulative contribution rate: the first, second and third principal components explain 45.47%, 25.95% and 11.2% of the total variation respectively, and the cumulative contribution rate of the first three principal components has reached 82.617%, so it can be roughly determined that the extractable principal components can be extracted. Three principal component influencing factors.

(3) Score Coefficient Matrix: Factor Score Coefficient Matrix is shown in Table 5, which is the final result of Principal Component Analysis.

Table 5. Component Score Coefficient Matrix.

|       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| x1    | 0.205 | 0.056 | 0.038 | x6   | -0.51 | -0.218 | 0.477 |
| x2    | 0.142 | 0.256 | 0.303 | x7   | 0     | 0.256 | -0.2 |
| x3    | -0.168 | 0.196 | 0.115 | x8   | -0.202 | 0.13  | 0.068 |
| x4    | -0.205 | 0.093 | -0.094 | x9   | 0.183 | 0.157 | 0.282 |
| x5    | 0.9   | 0.069 | -0.634 | x10  | 0.047 | -0.332 | -0.155 |
Through this coefficient matrix, the principal component can be expressed as a linear combination of variables. The new input variables are calculated according to the principal component factors as shown in Table 6. The three principal components extracted by principal component analysis instead of the nine original influence factors are used as input samples of the prediction model, which reduces the calculation dimension of the model and improves the operation efficiency. The information carried by the input variables is retained to the greatest extent, and the prediction accuracy can be guaranteed.

**Table 6. Principal Component Factor Data.**

| Serial number | Z1         | Z2         | Z3         |
|--------------|------------|------------|------------|
| 1            | 0.5131694  | 0.0959737  | 0.5777483  |
| 2            | 0.4325546  | -0.0114118 | 0.4540638  |
| 3            | 0.3446289  | -0.1030166 | 0.3232786  |
| 4            | 0.3941938  | 0.2285956  | 0.3081739  |
| 5            | 0.3089558  | 0.1285469  | 0.185595   |
| ...          | ...        | ...        | ...        |
| 35           | 0.5144694  | 0.3627772  | -0.2484476 |
| ...          | ...        | ...        | ...        |
| 68           | -0.3193791 | 0.6363475  | -0.0689075 |

3.4. Establishment of Radial Basis Function Neural Network Prediction Model

The input variable vectors of RBF neural networks are multiplied by the weight vectors, and then input into the hidden layer to calculate the distance between the sample and the center of the node. The distance between the sample and the center of the node is mapped by radial basis function to form the output of the hidden layer, and then output to the output layer. The linear combination of the nodes of each hidden layer constitutes the final network prediction output. The radial basis function (RBF) neural network is created by newrb function in MATLAB, and the concrete strength prediction model is established. The steps of predicting concrete compressive strength using radial basis function network are as follows:

1. Define sample data. As shown in Table 7, the input vectors and their target output values of each sample are defined, the input variables are defined as $3 \times 68$ matrix, and the predicted output values are $3 \times 68$ row vectors. The compressive strength of concrete aged 7 days and 28 days is predicted respectively.

2. Dividing training and testing samples. Sample data from 1 to 57 were used as training data set, and sample data from 58 to 68 (a total of 10 sets of data sequences) were selected as test data set. The prediction model is obtained by training data sets, and then the test data sets are checked.

**Table 7. Sample Data.**

| sequence | Input sample | Output sample |
|----------|--------------|---------------|
|          | Z1           | Z3            | 7d            | 28d           |
| 1        | 0.5131694    | 0.0959737     | 0.5777483     | -0.431963656  | -0.167303628 |
| 2        | 0.4325546    | -0.0114118    | 0.4540638     | -0.434443691  | -0.169327925 |
| 3        | 0.3446289    | -0.1030166    | 0.3232786     | -0.436559275  | -0.171468435 |
| ...      | ...          | ...           | ...           | ...           | ...           |
| 38       | 0.292972     | -0.0236283    | 0.4002823     | -0.434725827  | -0.170208146 |
| ...      | ...          | ...           | ...           | ...           | ...           |
| 68       | -0.3193791   | 0.6363475     | -0.0689075    | 0.872979215   | 0.923076923  |

3. Radial basis function (RBF) neural network is created by using newrb function. Radial Basis Function (RBF) neural network requires several parameters. The tolerance error is set to 1E-4, the
diffusion factor is 22, and the maximum number of neurons is 101. By calling this function, the system gradually increases the number of neurons and reduces the training error until the error is less than the tolerance error. At the same time, view (net) is used to display the network structure. The structure of RBF neural network is shown in Figure 2. The structure diagram shows that the hidden layer contains 50 neuron nodes, and the error descent curve is shown in Figure 3.

Figure 2. RBF Neural Network Structure Diagram

Figure 3. Residual descent curve

(4) Testing. The test samples are tested with the established radial basis function network prediction model, and the results are displayed. The curves and residual diagrams of the predicted and real values are shown in Figs. 4 to 5. From the figure, we can see that the predicted value is very close to the test sample value. From the angle of error, the maximum error is not more than 4%, which meets the basic requirements. Therefore, it is reasonable to conclude that radial basis function can accurately predict the compressive strength of concrete in the sample data set.

Figure 4. Comparison of 7d and 28d predicted values with measured samples.
8

Figure 5. Residual graph.

(5) The predicted values are inversely standardized to obtain the ultimate compressive strength of concrete at 7d and 28d ages, as shown in Table 8.

| sequence | 7d     | 28d     | Residual     | 7d     | 28d     | Residual     |
|----------|--------|--------|--------------|--------|--------|--------------|
| 1        | 50.5   | 49.47289305 | 1.027106953  | 60     | 58.07311463 | 1.926885366  |
| 2        | 45.6   | 44.27017805 | 1.329821954  | 56.6   | 54.47369866 | 2.126301345  |
| 3        | 36.4   | 35.44896302 | 0.951036977  | 48.4   | 47.486087  | 0.913913     |
| 4        | 37.8   | 35.5059845  | 2.294015497  | 49.6   | 46.73034087 | 2.869659134  |
| 5        | 41.8   | 40.8171279  | 0.982872096  | 55.5   | 53.7164746 | 1.783525403  |
| 6        | 37     | 35.97623239 | 1.023767606  | 49.6   | 47.69965142 | 1.900348577  |
| 7        | 47.1   | 46.11055265 | 0.989447352  | 58.9   | 57.71935746 | 1.180642541  |
| 8        | 50.7   | 49.7620032  | 0.937996797  | 63.3   | 62.20453985 | 1.095460154  |
| 9        | 48.7   | 48.15413233 | 0.545867667  | 62.6   | 61.22954823 | 1.370451768  |
| 10       | 62.1   | 60.684741   | 1.415259     | 71.9   | 70.182309   | 1.717691     |

Table 8. Prediction results of radial basis function neural network.

4. Conclusion
In view of the complexity of influencing factors of concrete in mix proportion test of practical engineering and the problem that the quality of experimental data can not be guaranteed, this paper uses data mining methods such as missing values and outliers test to improve the quality of experimental data, and uses the idea of "dimensionality reduction" in principal component analysis to transform 10 factors affecting concrete strength into The three linear independent principal component factors further reduce the interference of the relevant information contained in each influence factor, further improve the calculation efficiency of the model, and help to improve the accuracy of the prediction value. The prediction error of RBF neural network model based on data cleaning and principal component analysis can be controlled at about 4%, which meets the basic requirements of practical engineering testing and is also applicable to the statistics of samples with less data. It provides a feasible method for predicting the compressive strength of concrete and has certain practical application parameters. Examination value.

References
[1] Zhang Jing, Liu Xiangdong. Prediction of concrete strength based on least squares support vector machine optimized by chaotic particle swarm optimization [J]. Journal of Jilin University (Engineering Edition), 2016, 46 (04): 1097-1102.
[2] Piwenshan, Zhou Hongbiao, Hu Jinping. Prediction of concrete compressive strength based on BP neural network [J]. Low temperature building technology, 2011, 33 (04): 14-16.
[3] Guo Yaodong, Liu Yuanzhen, Wang Wenyi, Qin Xiaochao. Prediction of compressive strength of recycled thermal insulation concrete based on BP neural network [J]. Concrete, 2018 (10): 33-35+39.

[4] Li Hong, Peng Tao. Prediction of concrete compressive strength based on BP and RBF neural networks [J]. Journal of Wuhan University of Technology, 2009, 31 (08): 33-36.

[5] Han Bing, Yuan Ying. GA-SVM prediction model of land subsidence based on principal component analysis [J]. Chinese Science and Technology Paper, 2018, 13 (09): 1045-1049.

[6] Li Wei, Meng Lingpei. Prediction of 4G User Loss Based on Big Data [J]. Information and Computer (Theoretical Edition), 2018 (03): 132-134+137.

[7] Huang Hang, Minfan, Ren Jie. Active learning missing value filling algorithm based on weighted prediction of collaborative filtering [J]. Journal of Nanjing University (Natural Science), 2018, 54 (04): 758-765.

[8] Zhang Liangjun, Yang Tan, Xiao Gang, Xu Shengbing. MATLAB Data Analysis and Mining Practice [M]. Beijing: Machinery Industry Press, 2017.

[9] Xu Kaicheng, Biliping, Chen Mengcheng. Prediction model of compressive strength of lithium slag concrete based on SPSS regression analysis [J]. Journal of Architectural Science and Engineering, 2017, 34 (01): 15-24.

[10] WEN H, XIE W, PEI J.A structure-adaptive hybrid RBF-BP classifier with an optimized learning strategy [J]. Plos One, 2016, 11 (10).