Prediction of the antifreeze of the concrete structure based on random forest and wavelet neural network

Yangyi Tang¹, Xianguo Wu¹, Hongyu Chen² and Tiemei Zeng³

¹ School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
² School of Civil Engineering and Environment, Nanyang Technological University, Singapore
³ Wuhan Metro Group Co., LTD., Wuhan, Hubei, 430030, China

Abstract. The durability of concrete is the focus of research in the field of engineering, and the resistance of concrete is one of the important indicators of concrete durability. Based on the prediction of early resistance to concrete based on the random forest combined with the wavelet neural network algorithm, 12 factors affecting concrete antifreeze were selected from the ratio level of concrete material, and the relative dynamic elastic mode was used as the evaluation index of concrete resistance, and the importance evaluation and feature variable selection of influencing factors were used by random forest, and a concrete antifreeze prediction model based on RF-WNN (random forest-wave neural network) was established. The example analysis is carried out by taking the project project as an example, and the calculation of the wavelet neural network and BP neural network prediction model without the influence factor screening is compared to obtain the mean square root error and the fit superiority respectively. The results show that the prediction results of RF-WNN model are closer to the actual value and the prediction accuracy is higher, and the proposed RF-WNN prediction model provides an effective method for achieving concrete antifreeze prediction.

1. Introduction

Concrete is widely used in civil engineering construction, and its durability has a direct impact on the safety of engineering structures. In practical engineering, the case of structural degradation due to the durability of concrete has been common, and the resulting economic losses are also very alarming[1]. Dynamic elastic modulus is an important parameter of the performance of concrete materials, and by measuring the relative dynamic elastic modulus, it can well reflect the amount of change and damage in the internal microstructure of concrete[2]. Therefore, it has become an urgent need for engineering construction and concrete development to predict the relative dynamic elastic modulus of concrete.

At present, many scholars at home and abroad have carried out a series of related research on the dynamic modulus of concrete, and have obtained good research results. The workability, mechanical properties and damage behaviors of self-compacting concrete (SCC) under coupling effect of salt freeze-thaw and flexural load were experimentally investigated by Tian. The results showed that with increasing salt freeze-thaw cycles, the weight loss increased and the relative dynamic elastic modulus declined[3]. Chen studied The long-term performance of concrete subjected to freezing and thawing damage, and the relationships between the number of F/T cycles and damage parameter for different probabilities of reliability are established[4]. Gao proposed the original stress damage and the freeze-thaw damage involved mechanical damage evolution equation of lightweight aggregate
concrete[5]. Wang studied the durability of C50 high performance concrete under the coupled action of chloride salt erosion and freeze-thaw cycle through the rapid freezing and thawing test method of concrete[6]. From the present research, most of the research is mainly based on experiments, theoretical analysis and other methods, for the concrete dynamic elastic modulus research has not yet established a set of practical and reasonable complete evaluation methods, how to collect the information for effective data mining, so as to the durability of concrete for scientific risk identification, early warning and prevention and control has become an urgent problem to be solved.

With the development of computer, machine learning algorithm is gradually widely used in civil engineering field, among which artificial neural network is most commonly used, it has a good nonlinear mapping function, and provides an effective method to solve the nonlinear problem stoic in concrete field[7, 8]. The wavelet neural network is the combination of wavelet transformation and artificial neural network, which can freely choose the nonlinear wavelet base function, can handle the nonlinear data well, and has good wavelet transformation characteristics[9, 10]. Random forest is a multi-classer integrated algorithm, for the multi-feature set, random forest classifier can preferably reduce the feature dimension, reduce over-fitting[11].

Therefore, this paper introduces a prediction model based on random forest and wavelet neural network, combined with the actual fly ash concrete test data, selects the optimal variable feature set, predicts the relative dynamic elastic mode of concrete, and compares the prediction accuracy with the wavelet neural network and BP neural network prediction method without screening, and demonstrates the applicability and rationality of the model.

2. Preliminaries

2.1. Random Forest

Random Forest is a statistical learning theory based on the Classification Tree. Random forest stoically uses Bootstrap resampling method to extract multiple sample sets from the original sample set, and models the decision tree separately for each sample set, and each decision tree randomly selects features to divide the internal nodes, which eventually constitutes a random forest, and the final prediction results are combined with the results of each decision tree.

When the population of the random forest was taken with a generalized error set, about 36.8% of the samples in each original sample set were not drawn, and these samples were called out-of-bag data (OBB), which can be used to calculate the generalization error of the model. The generalization error of a random forest can be expressed as: Formatting author names

$$E' = P_{x,y}(M(x,y) < 0)$$

(1)

Among them, the subscript X, Y indicates the probability P covers X, Y space. In random forests, when the number of decision trees is sufficient, $E'$ converge as the number of trees increases:

$$P_{x,y}(P_{h}(h(x,\theta) = y) - \max_{j\neq i} P_{h}(h(x,\theta) = j) < 0)$$

(2)

Explain that generalization errors do not overfit as the number of decision trees increases, but rather move closer to a finite upper bound. When the random forest sorts the importance of the model, it uses the corresponding out-of-bag data to calculate its out-of-bag error r1, then randomly transforms the order of a feature in the out-of-pocket data, and again calculates the out-of-bag error r2, assuming that the random forest has n decision trees, then the importance I of a feature is:

$$I = \sum_{i=1}^{N} \frac{(r_i - r_{i})}{N}$$

(3)

After obtaining the importance of each feature, the least important feature is eliminated by recursive feature elimination (RFE) until the optimal number of features is reached and the feature selection is realized.
2.2. Wavelet neural network

The neural network that combines wavelet theory and neural network is the wavelet neural network (WNN for short), which has both efficient self-learning ability and good localization characteristics, and its network model is similar to THE BP neural network, including the input layer, the implied layer and the output layer of the three-layer network structure. Unlike BP neural networks, the excitation function of the implicit layer of a smallwave neural network uses a smallwave-based function with the expression:

$$y = \cos(1.75x)e^{-\frac{x^2}{2}}$$  \hspace{1cm} (4)

When the input layer sequence is $x = \{x_1 \ x_2 \ x_3\}^T$, the output formula for the implied layer node is calculated as

$$h(i) = h\left(\frac{\omega^T_j x - b_i}{a_i}\right), i = 1, 2, 3...$$  \hspace{1cm} (5)

In the formula, $\omega^T_j$ is weight for input layer j to implied layer i, $b_i$ is the translation factor of the ith implicit layer node, and $a_i$ is the scaling scale factor of the ith implied layer node. By connecting the weights, the calculation formula for the output layer of the wavelet neural network is:

$$y(k) = \sum_{i=1}^{l} \omega_k h(i), k = 1, 2, 3...m$$  \hspace{1cm} (6)

In the formula, $\omega_k$ represents the weight between the implied layer i and the output layer k, m is the number of nodes in the output layer, l is the number of nodes in the implied layer, and h(i) represents the output of the ith implied layer node.

3. The establishment of a wavelet neural network prediction model for random forests

There are many factors affecting the resistance of concrete, including cement, sand, water and glue ratio, the relationship between these factors and dynamic elastic modulus is complex and nonlinear, it is very difficult to predict modeling by establishing mathematical expressions by analyzing the mechanism. Therefore, the relative dynamic elastic modulus of reinforced concrete structure is predicted by random forest and wave-wave neural networks, and the construction process of the prediction model is as follows:

Firstly, create original training set. Based on a large number of literature and the summary of the actual experience of the project, the influence factors of concrete resistance to freeze are selected from the ratio level of concrete material to construct the index system, which is used as the input variable of random forest, and the concrete test data is used as the original training set.

Secondly, select variable feature based on random forest:

- Noise interference is added to the characteristics of all samples of OOB of out-of-bag data, and the importance of each variable is calculated according to the equation(3) and ordered in descending order to arrive at the importance of different influencing factors. The text should be set to single line spacing.
- Using recursive feature elimination (RFE) to select the training data set, compare the change of the mean square error in different variable combinations, and find the optimal number of features according to the principle of the minimum mean square error.
- Remove the least important indicator from the current indicator set one by one until the optimal number of indicators is reached. The filtered indicator set is used as an input variable for the wavelet neural network model.

Thirdly, establish wavelet neural network model:

- Determine the parameters related to the network structure. The number of input layer nodes of the wavelet neural network is determined according to the number of samples, and the number of output layer nodes is generally between 1 and 10 constants. For the wavelet neural network model, too many implied layers tend to cause the convergence speed to decrease significantly and make the error larger, so the number of implied layers is generally determined to be one
The number of nodes in the implied layer is determined according to the empirical formula \( n = j + i + k \), where \( j \) is the number of input layer nodes, \( i \) is the number of output layer nodes, and \( k \) is the number of implied layers.

- Select the training method and the incentive function. The training of the sample strains in this paper adopts the fast gradient drop method with fast convergence speed and high convergence precision, selects morlet wavelet function with stable performance and good robustness as the excitation function of its implied layer, and the expression shows the model as shown in the form (4), so that the model has good anti-jamming ability.

Fourthly, Evaluation of prediction results. The prediction accuracy of the model was judged by two parameters: the mean square root error (RMSE) and the fit suit \( (R^2) \), and the prediction results of the wavelet neural network and the BP neural network model were compared with those that were not selected by feature selection. The mean square error and the fit superiority expressions are such as equation (7) and (8).

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2 
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left( y_i - \bar{y} \right)^2}
\]

Where, \( n \) is the number of test sets samples; \( y_i (i=1,2,3...l) \) is the true value of the /sample; \( \hat{y}_i (i=1,2,3...l) \) is the predicted value for sample. Performance evaluation of the established regression model using standard error.

### 4. Case study

#### 4.1. Create original training set

Concrete antifreeze influence factors as input variables, relative motion elastic modulus as output variables, select each indicator corresponding to 100 sets of data as the original training set, of which 10 sets of data as specific as Table 1:

| Water-glue ratio | Sand ratio (%) | Cement strength (Mpa) | Fly ash aggregate (kg/m3) | Fine aggregate (kg/m3) | Coarse aggregate (kg/m3) | Water reducer (%) | Air-entraining agent (kg/m3) | Swelling agent (kg/m3) | Steel fiber (kg/m3) | Silica fume (kg/m3) | Relative dynamic modulus of elasticity(%) |
|------------------|----------------|-----------------------|---------------------------|------------------------|--------------------------|------------------|-----------------------------|----------------------|------------------|------------------|----------------------------------|
| 0.29             | 38             | 61.4                  | 417                       | 49                     | 691                      | 112              | 1.2                         | 0                    | 24               | 0                 | 24                 |
| 0.32             | 40             | 57.1                  | 376                       | 70                     | 732                      | 1097             | 1.2                         | 0                    | 0                | 0                 | 24                 |
| 0.33             | 40             | 59.7                  | 375                       | 47                     | 730                      | 1095             | 1.1                         | 0                    | 24               | 0                 | 23                 |
| 0.31             | 40             | 57.4                  | 366                       | 69                     | 740                      | 1110             | 1.0                         | 0                    | 0                | 0                 | 80                 |
| 0.31             | 38             | 58.5                  | 387                       | 73                     | 690                      | 1126             | 1.2                         | 0                    | 0                | 0                 | 24                 |
| 0.3              | 38             | 60.1                  | 380                       | 49                     | 706                      | 1153             | 1.3                         | 0                    | 39               | 0                 | 25                 |
| 0.34             | 38             | 60.3                  | 340                       | 44                     | 730                      | 1192             | 1.3                         | 0                    | 35               | 0                 | 22                 |
| 0.33             | 39             | 61.4                  | 359                       | 45                     | 723                      | 1131             | 1.6                         | 0                    | 1.792            | 22                | 93.81              |
| 0.33             | 40             | 60.6                  | 423                       | 47                     | 718                      | 1077             | 1.2                         | 0                    | 0                | 0                 | 97.14              |
| 0.33             | 40             | 59.2                  | 424                       | 47                     | 718                      | 1076             | 1.2                         | 0                    | 0                | 0                 | 96.72              |

#### 4.2. Feature selection based on random forest

Direct selection of the above 12 influencing factors to establish a predictive model is easy to lead to overfitting, therefore, it is necessary to first filter the influence factors, remove the redundant variables of weak importance, preferably the new influence set for the last wavelet neural network modeling

- Sample diversity. The original sample was randomly divided into 5, of which 4 were the original training set and one was the original test set, which was used to determine the parameters of the random forest, to construct a model of the random forest, to screen the influence factors and to
train the wavelet neural network, and the test set to evaluate the predictive performance of the final model.

- Importance Calculation. The parameters of the random forest model are mainly the number of random features mtry and the tree ntree of the decision tree, the number of random features in regression analysis is generally defaulted to the input influence renter 1/3, in this case, select mtry 4, select ntree 600. The importance of different influencing factors is sequenced in descending order to get the importance evaluation of the variables in the training model as shown in Figure 1. The greater the change in the purity of the node (InNodePurity), the greater the effect of the variable on the evaluation indicator, the greater the importance of the variable. Figure 1 can be seen, water-glue ratio, cement dosage, cement strength, coarse aggregate, water reduction agent, fine aggregate variable importance measure is relatively large.

- RFE feature selection. The characteristic selection of the training data set is used to find the optimal number of features by using recursive feature elimination (RFE). The RMSE change trend graph, shown in Figure 2, is obtained after a 5fold cross-validation of different combinations of feature variables. Figure 2 shows that in order to minimize the average square root error of the model, the number of variables of the influence factor combination should be 6, combined with the factor series in Figure 1 for recursive feature elimination, and then reject the least important variable features until the optimal number of features is reached, then when the combined variables are hydrogel ratio, cement dosage, cement strength, fine aggregate, water reduction agent, coarse aggregate, the model accuracy prediction is highest.

![Figure 1 Order the importance of antifreeze influencers](image1)

![Figure 2 The trend graph of RMSE when different variables are combined](image2)
4.3. Wavelet Neural Network Model Prediction

In this paper, 80 sets of data from the original training set were randomly selected as the training sample of the wavelet neural network, and the remaining 20 sets of data were used as test samples, based on the hydrogel ratio, cement dosage, cement strength, fine aggregate, water reduction agent, coarse aggregate stake a total of 6 characteristics, using MATLAB software to establish a wavelet neural network model to predict concrete resistance, first through training samples for network training, and then use the trained wavelet neural network model to predict the test set. The predicted value of the output is subjected to the final prediction value after the reunification processing, and the final prediction result is Figure 3, which shows that the prediction curve of the wavelet neural network model on the test set is very close to the real value.

![Figure 3 Test result prediction graph](image)

4.4. Evaluation of regression prediction results

In order to test the superiority of the wavelet neural network model (RF-WNN) based on random forest, the smallwave neural network and BP neural network model without feature selection were selected to predict the resistance of concrete, and the prediction results were compared with the results of the RF-WNN model, and the prediction effect of the model was measured by the mean square root error RMSE and deterministic coefficient $R^2$. The proportion of variability in the set of data explained by the statistical model, $R^2$, provides an indicator of how good a model might be to predict future results. $R^2$ ranges from 0 to 1, and closer to 1 means that the more accurate the observation data is. The RMSE value is the sum of the individual differences between the estimated and actual observations. The value of RMSE is equal to or greater than 0, and closer to 0 means that the observation data is statistically perfect. Different prediction models get error results for the pair, such as Table 2:

| Model   | RMSE         | $R^2$         |
|---------|--------------|---------------|
| RF-WNN  | $3.174 \times 10^{-7}$ | 0.9431        |
| WNN     | $2.561 \times 10^{-3}$  | 0.8931        |
| BPNN    | 0.093        | 0.7356        |

It can be seen that the average square root error of RF-WNN prediction results is significantly smaller than the other two, and its deterministic coefficient is closest to 1, indicating that the RF-WNN model prediction results are closer to the actual value, the prediction accuracy is higher, and the prediction effect is better.
5. Conclusion

- In view of the many influence factors of concrete resistance, complex noise interference, and the characteristics of nonlinearity and complexity of sample data, this paper introduces the method of random forest combined with wavelet neural network to predict the relative dynamic elastic mode of concrete, the prediction model, by obtaining the order of the importance of influencing factors, reduces the dimension of the training model, accelerates the training speed and improves the prediction accuracy. The proposed RF-WNN prediction model provides an effective tool for the prediction of concrete's relative dynamic elastic modulus.

- Taking a civil engineering project as an example, this paper evaluates the importance of the relevant factors of concrete resistance, selects the water-glue ratio, cement dosage, cement strength, fine aggregate, water reduction agent, coarse aggregate as input variables, constructs the concrete antifreeze training model based on the wave-wave path network of random forest, selects some of the actual engineering data as a test set, analyzes the prediction accuracy of the model, and verifies the accuracy of the model.

- The RF-WNN model was compared with the prediction results of the wavelet neural network model and the BP neural network model without feature selection. The results show that the RF-WNN model can obtain more accurate and stable prediction results than the single wavelet neural network prediction model, which further illustrates the good application prospect of the model.

Acknowledgments
The National Natural Science Foundation of China (Grant nos. 71571078, 51778262 and 51708241), and National Key Research Projects of China (Grant no. 2016YFC0800208) are acknowledged for their financial support of this research.

References
[1]. Hongxia, Q., H.E. Zhongmao, and L.I.U. Cuilan, Dynamic elastic modulus and microstructure study of concrete in sulfate environment. Journal of Harbin Institute of Technology, 2008. 40(8): p. 1302-1306.

[2]. Qiao, H.-x., et al., THE DYNAMIC ELASTIC MODULUS AND MICROSTRUCTURE STUDY OF HIGH PERFORMANCE CONCRETE IN SULFATE ENVIRONMENT. Advances in Concrete Structural Durability, Proceedings of Icdcs2008, Vols 1 and 2, ed. W.L. Jin, T. Ueda, and P.A.M. Basheer. 2008. 426-429.

[3]. Tian, J., W. Wang, and Y. Du, Damage behaviors of self-compacting concrete and prediction model under coupling effect of salt freeze-thaw and flexural load. Construction and Building Materials, 2016. 119: p. 241-250.

[4]. Chen, F. and P. Qiao, Probabilistic damage modeling and service-life prediction of concrete under freeze-thaw action. Materials and Structures, 2015. 48(8): p. 2697-2711.

[5]. Gao, C., et al., Freeze-thaw Resistance of Lightweight Aggregate Concrete with Stress Damage. Journal of the Chinese Ceramic Society, 2014. 42(10): p. 1247-1252.

[6]. Wang, Y., et al., Durability of C50High Performance Concrete under the Coupled Action of Chloride Salt Erosion and Freeze-Thaw Cycle. China Railway Science, 2014. 35(3): p. 41-46.

[7]. Viet-Linh, T., T. Duc-Kien, and N. Duy-Duan, Practical artificial neural network tool for predicting the axial compression capacity of circular concrete-filled steel tube columns with ultra-high-strength concrete. Thin-Walled Structures, 2020. 151.

[8]. Nguyen, M.-S.T., D.-K. Thai, and S.-E. Kim, Predicting the axial compressive capacity of circular concrete filled steel tube columns using an artificial neural network. Steel and Composite Structures, 2020. 35(3): p. 415-437.

[9]. Ye, R., et al., Short-term speed forecasting method based on wavelet packet decomposition and improved Elman neural network. 2016 International Conference on Probabilistic Methods Applied To Power Systems. 2016.
[10]. Wang, N. and H. Adeli, *Self-constructing wavelet neural network algorithm for nonlinear control of large structures*. Engineering Applications of Artificial Intelligence, 2015. 41: p. 249-258.

[11]. Sun, Y., et al., *Prediction of the Strength of Rubberized Concrete by an Evolved Random Forest Model*. Advances in Civil Engineering, 2019. 2019.