Research on prediction of concrete frost resistance based on random forest

Lei Wang^{1,*}, Tiemei Zeng², Yanyan Tao³, Xianguo Wu¹, and Hongyu Chen⁴

¹ School of Civil Engineering & Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

² Wuhan Metro Group Co Ltd, Wuhan, Hubei, 430030, China

³ School of Urban Construction, Wuchang Shouyi University, Wuhan, Hubei, 430074, China

⁴ School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, 6397898, Singapore

Abstract. Poor frost resistance of concrete will accelerate the deterioration of concrete structure and shorten the service life of concrete. This paper predicts the frost resistance of concrete and selects the initial index of factors that affect the frost resistance of concrete. The random forest algorithm is introduced to remove the unimportant indicators in the initial indicator system of the concrete mix ratio and determine the optimal indicator combination. The relative dynamic elastic modulus is used as the output index of random forest to predict the frost resistance of concrete. The proposed random forest model provides a reasonable and effective method for concrete frost resistance prediction.

1 Introduction

Compared with other building materials, concrete is economical, easy to manufacture on site, and has good durability. It is one of the most widely used materials in construction[1]. Under the conditions of freezing and thawing, the frost resistance of concrete is a consideration that cannot be ignored in engineering practice. Improving the frost resistance of concrete has very important engineering value. In order to ensure the safe service life of buildings, accurate and rapid prediction of concrete frost resistance is of great significance to actual projects.

At present, many domestic and foreign scholars are conducting research on concrete prediction. Amini K proposed statistical univariate and multivariate regression models to predict concrete durability[2]. Lin et al. used the relative dynamic elastic modulus to calculate the performance degradation index and showed that the relative dynamic elastic modulus has a linear relationship with the number of freeze-thaw cycles[3]. Using formula fitting to perform regression prediction, the prediction process is relatively simple, but the biggest flaw is the poor fitting accuracy. Ji established a concrete strength and slump prediction model based on artificial neural networks (ANNs)[4]. Yan et al. used regression technology to introduce a precision-insensitive loss function, and studied the application of SVM in the relative dynamic elastic modulus of concrete[5]. Behforouz B et al. used ANN prediction model to predict the durability of concrete such as frost resistance[6]. The above traditional machine learning algorithms are easy to deviate from the global optimal situation, making the prediction results inaccurate.

Therefore, aiming at the deficiencies of the above

methods, this paper proposes a random forest frost resistance prediction model based on the ratio of raw materials. This paper establishes initial indicators of factors affecting concrete frost resistance, introduce random forest algorithm (RF) to remove unimportant indicators in the initial indicator system of concrete mix ratio, and determine the optimal factor indicators. The relative dynamic elastic modulus is used as the output index of the random forest to predict the frost resistance.

2 Preliminaries

Random Forest (Random Forest) model is based on the theoretical basis of Classification Tree[7]. Random forest adopts Bootstrap re-sampling method to establish a decision tree model for each sample set. When modeling, each decision tree randomly selects features to split the attributes of internal nodes, thereby forming a random forest. The final output result is the result of the comprehensive decision tree. For the regression algorithm, the final prediction result is the mean value of the output results of all decision trees, and the regression results are expressed as[8]:

$$\bar{H}(x) = \frac{1}{k} \sum_{i=1}^k (h_i(x = Y)) \quad (1)$$

Among them, $\bar{H}(x)$ represents the combined random forest model, $h_i(x)$ represents the single decision tree model, Y represents the output variable (or called the target variable), and k is the number of decision trees in the random forest model.

One of the biggest advantages of the random forest prediction model is its good generalization ability, which can effectively prevent overfitting when predicting data.

* Corresponding author: 1063609358@qq.com

The generalization error of the model will not cause over-fitting as the number of decision trees increases, but will approach a finite upper bound[9]. This finite upper bound is:

$$E^* = \frac{\bar{\rho}(1-s^2)}{s^2} \quad (2)$$

Among them, ρ is the average correlation coefficient of the decision tree, and s is the average strength of the decision tree.

Constructing a random forest model can be used to rank the characteristics of indicators. Assuming that the random forest has N decision trees, use the corresponding out-of-bag data to calculate its out-of-bag error r_1 , and randomly calculate the out-of-bag error r_2 after randomly changing the order of a feature in the out-of-bag data, then the importance of a certain feature I is:

$$I = \frac{\sum_{i=1}^N (r_1 - r_2)}{N} \quad (3)$$

After obtaining the importance degree of each feature, the features are sorted reasonably, and the feature of the least importance is sequentially removed from the feature set using the sequence backward method until the optimal number of features is reached, and feature selection is realized.

3 Equations and mathematics

3.1 Random forest parameter settings

About 4/5 of the data in the original data set is used as the training sample set of the RF model, and the rest contains 1/5 of the data in the original data set as the test sample set of the RF model. In the modeling process, two parameters are mainly involved: the number of regression trees n_{tree} and the number of random features m_{try} . The value of n_{tree} will affect the training degree and accuracy of the random forest model. Generally, the larger the value of n_{tree} , the higher the accuracy of the model; m_{try} controls the degree of disturbance of the random forest model attributes. m_{try} is generally determined according to empirical formulas. The number of variables is P , by default $m_{try}=P$ (classification model) or $m_{try}= P/3$ (regression model). After the two parameters are determined, the RF algorithm can be used to establish a random forest regression model for the training sample set and the test sample set.

3.2 Importance evaluation

Variable importance measure can evaluate the degree of influence of characteristic variables on dependent variables, so as to determine the key variables that need to be controlled. This paper calculates the importance of feature variables according to formula (3), that is, uses OBB error to evaluate the importance of feature variables. It evaluates the importance of feature variables by measuring the reduction in mean square residual error (%Inc MSE) and the reduction in model accuracy (Inc

Node Purity) after random replacement[10].

3.3 Evaluation of prediction results

The prediction result of the model is obtained by taking the average of all decision trees according to formula (2). In order to evaluate the prediction accuracy of random forests, two evaluation parameters, root mean square error (RMSE)[11] and goodness of fit (R^2)[12] can be used. It can be used to describe the deviation between the predicted value and the measured value, and it can be used to describe the degree of correlation between the predicted value and the measured value. The expressions of mean square error and goodness of fit are as equations (4) and (5).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (5)$$

Where n is the number of samples in the test set; y_i ($i = 1, 2, \dots, n$) is the true value of the i -th sample; $\hat{y}_i$ ($i = 1, 2, \dots, n$) is the predicted value of the i -th sample.

4 Case study

4.1 Background

Taking a civil engineering project as an example, this paper selects relative dynamic elastic modulus as the output index of random forest prediction. With reference to a large number of related documents[13] and the actual situation of construction engineering[14], 11 indicators are selected as the initial input indicators of the random forest, as shown in Table 1:

Table 1. Monitored sample data.

Variables	Data		
	Min	Max	Ave
Cement dosage (kg/m ³)	358	457	407.5
Fly ash dosage (kg/m ³)	27.6	127.9	73.6
Fine aggregate (kg/m ³)	675	701	687.2
Coarse aggregate (kg/m ³)	1102	1143	1121
Water reducing agent (%)	0.8	1	0.9
Air entraining agent (%)	0	1.79	0.56
Water-binder ratio	0.31	0.35	0.33
Cement strength (Mpa)	55.5	61.5	58.5
Steel fiber(kg/m ³)	0	48	16
Expansion agent(kg/m ³)	0	80	5.3
Silica fume (kg/m ³)	0	25	9.4
Relative dynamic elastic modulus %	91.0	96.1	92.5

4.2 Random forest for index selection

In this case, there are 100 sets of sample data, the training sample set is 80 sets, and the test sample set is the remaining 20 sets. The generalization error of the random forest regression model will gradually converge as *ntree* increases. Therefore, a large enough *ntree* value should be selected to build the model to ensure that the error of the training set can stabilize. Use OOB (out of bag data) error rate estimation method to determine the value of *ntree*, as shown in Fig. 1. In this study, the optimal *ntree* is determined to be 600, which can not only ensure the sufficient amount of decision trees, but also control the calculation amount of the algorithm within a relatively reasonable range.

Fig. 1. Selection of the number of decision trees

Through the importance evaluation, the importance of the influencing factors to the relative dynamic elastic modulus can be determined, as shown in Fig. 2. Variables such as water-binder ratio, cement strength, cement dosage, fly ash dosage, coarse aggregate and fine aggregate are of high importance, which should be given priority in actual engineering projects.

Fig. 2. Importance ranking

In order to screen out the number of variable combinations of concrete raw material mix ratios based on random forest, recursive feature elimination (RFE) can be used to select the characteristics of the training data set, and finally obtain the optimal number of variable combinations. Fig. 3 shows the trend diagram of the relative dynamic elastic modulus variable combination. When the number of deformation combinations is 9, the root mean square error RMSE reaches the lowest value and the model has the highest accuracy. The top nine concrete raw material mix ratio parameters are selected as

the key factors selected by the random forest and will be used in the construction of the random forest prediction model later.

Fig. 3. Relative dynamic elastic modulus variable combination trend chart

4.3 Frost resistance prediction based on random forest

The input characteristic indexes are water-binder ratio, cement strength, cement dosage, fine aggregate, fly ash dosage, coarse aggregate, silica fume, expansion agent and air-entraining agent. The Randomforest package of R software was used to establish a random forest model of concrete frost resistance prediction, and the two parameters of the model were set to *mtry*=3 and *ntree*=600. The regression fitting curve of the training sample set is shown in Fig. 4. The trained random forest model is used to perform regression fitting on the test sample set, and the prediction result of the test sample set regression fitting is shown in Fig. 5.

It can be seen intuitively from Fig. 4 and 5 that the established random forest prediction model for concrete frost resistance is very consistent with the actual value of the test set samples. After calculation, the root mean square error (RMSE) between the actual value and the predicted value of the random forest model on the test sample set is 0.001, and the goodness of fit is 0.907. The closer the RMSE is to 0, the closer the goodness of fit is to 1, indicating that the prediction effect is better, which can indicate that the model has a good effect on the prediction of concrete frost durability and has good generalization ability.

Fig. 4. Training set sample fitting results

Fig. 5. Test set sample fitting results

In order to further verify the reliability of the random forest model, support vector machines and BP neural network models are selected to predict the frost resistance of concrete. The prediction results are compared with the results of the RF model, and the root mean square error RMSE and the goodness of fit R^2 are selected to measure the prediction effect of the model. The error comparison of the prediction results of different models is shown in Table 2.

Table 2. Comparison of prediction results of different models.

Model	Performance	
	RMSE	R^2
RF	0.0010	0.907
SVM	0.3216	0.842
BP neural network	0.4567	0.806

It can be seen from Table 2 that the root mean square error and goodness of fit of the prediction model based on the random forest relative dynamic elastic modulus are 0.0010 and 0.907, respectively. The root mean square error and goodness of fit of the support vector machine prediction model are 0.3216 and 0.842, respectively. The root mean square error and goodness of fit of the BP neural network prediction model are 0.4567 and 0.806, respectively. Compared with the support vector machine prediction model, the random forest prediction model reduced the root mean square error by 0.3206 and the goodness of fit increased by 0.065. Compared with the BP neural network prediction model, the root mean square error is reduced by 0.4557, and the goodness of fit is increased by 0.101. Therefore, in the random forest prediction model, support vector machine prediction model, and BP neural network prediction model, the random forest prediction model predicts the early frost resistance of concrete and can obtain prediction results with higher accuracy and reliability.

5 Conclusion

This paper proposes a random forest algorithm, which fully demonstrates the complex nonlinear relationship between the frost resistance and mix ratio of concrete, and makes high-precision prediction of concrete. Compared with other forecasting models, it reflects the reliability of the random forest forecasting model.

- (1) The random forest method is introduced to predict the frost resistance of concrete, and the corresponding modeling process and steps are

proposed. Use the variable importance measure of the random forest algorithm to rank the feature variables. After removing the variables of small importance, the dimensionality of the training model is reduced and the prediction accuracy is increased.

- (2) In this paper, taking a concrete project as an example, 9 factors are determined as important factors after selection of characteristic variables. Taking into account that the data comes from the project, it indicates that in the actual project, attention should be paid to strengthen the control of these factors. Taking these 9 factors as the optimal set of input variables, a random forest-based concrete frost resistance prediction model was constructed. The final output predicted value and actual value verify the accuracy and reliability of the model.
- (3) The prediction results of random forest model, support vector machine and artificial neural network model are compared and analyzed. The results show that the random forest feature selection has successfully improved the prediction and fitting accuracy of the model. The random forest model can obtain more accurate and stable prediction results, which further illustrates the good application prospects of the model.

References

1. Sun, Y., et al., A new mixture design methodology based on the Packing Density Theory for high performance concrete in bridge engineering. *Construction and Building Materials*, **182**(2018)
2. Amini, K., et al., Development of Prediction Models for Mechanical Properties and Durability of Concrete Using Combined Nondestructive Tests. *Journal of Materials in Civil Engineering*, **31**(2019)
3. Lin, S., H. Li, and H. Zhang, Experimental Study on Frost Resistance Durability and Service Life Prediction of Normal Cement Concrete. *Advanced Materials Research*, **368-373**: p. 2425-2429(2011)
4. Ji, T., T. Lin, and X. Lin, A mortar mix proportion design algorithm based on artificial neural networks. *Cement and Concrete Research*, **36**: p. 1399-1408(2006)
5. Yan, K. and C. Shi, Prediction of elastic modulus of normal and high strength concrete by support vector machine. *Construction and Building Materials - CONSTR BUILD MATER*, **24**: p. 1479-1485(2010)
6. Behforouz, B., P. Memarzadeh, and M. Eftekhari, Regression and ANN models for durability and mechanical characteristics of waste ceramic powder high performance sustainable concrete. *Computers and Concrete*, **25**(2020)
7. Zhang, P., A Novel Feature Selection Method based on Global Sensitivity Analysis with Application in Machine Learning-based Prediction Model. *Applied Soft Computing*, (2019)

8. Breiman, L., Random Forests. Machine Learning, **45**(1): p. 5-32(2001)
9. García-Pedrajas, N., C. Martínez, and J. Muñoz-Pérez, Multi-objective cooperative coevolution of artificial neural networks (multi-objective cooperative networks). Neural networks : the official journal of the International Neural Network Society, **15**: p. 1259-78(2003)
10. Genuer, R., J.M. Poggi, and C. Tuleau-Malot, Variable selection using random forests: Elsevier Science Inc(2010)
11. Hyndman, R. and A. Koehler, Another look at measures of forecast accuracy. International Journal of Forecasting, **22**: p. 679-688(2006)
12. Boddy, R. and G. Smith, Statistical Methods in Practice: for Scientists and Technologists(2009)
13. Kwan, A., Concrete mix design based on water film thickness and paste film thickness. Cement and Concrete Composites, **39**: p. 33–42(2013)
14. Kalla, P., et al., Mechanical and durability studies on concrete containing wollastonite–fly ash combination. Construction and Building Materials, **40**: p. 1142-1150(2013)