Prediction of impermeability of the concrete structure based on random forest and support vector machine

Lei Wang¹, Xianguo Wu¹, Hongyu Chen² and Tiemei Zeng³

¹ School of Civil Engineering & Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
² School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, 6397898, Singapore
³ Wuhan Metro Group Co Ltd, Wuhan, Hubei, 430030, China

Abstract. The durability of concrete has a significant impact on the service life. Impermeability is one of the important indicators of concrete durability. It is of great significance to quickly and reasonably predict the impermeability of concrete. This paper combines random forest and support vector machine (RF-SVM) methods. Taking a highway project as the research background, 11 factors were selected as the impact index of concrete impermeability, and the chloride permeability coefficient was used as the evaluation index of concrete impermeability. After random forest index screening, six factors including water-binder ratio, cement dosage, cement strength, fine aggregate, water-reducing agent and coarse aggregate were selected to construct a support vector machine model to predict the impermeability of concrete. The prediction results of the RF-SVM model are compared with the BP neural network model and the support vector machine model without index screening. The results show that the RF-SVM model has higher prediction accuracy and better fitting effect, which provides an effective method for the prediction of concrete impermeability.

1. Introduction

The durability of concrete refers to the ability of the concrete structure to maintain its safety, normal use and acceptable appearance under various environmental conditions without the need for additional maintenance costs for reinforcement treatment within the specified service life. Nowadays, a large number of concrete buildings suffer from serious deterioration due to insufficient durability design, which also leads to huge maintenance, reinforcement and other costs[1]. Concrete permeability is closely related to durability. If the permeability of the concrete is low, the aggressive ions will penetrate into the concrete, causing damage to the concrete. Therefore, it is particularly important to predict the permeability coefficient of the aggressive ions of the concrete[2, 3].

In recent years, scholars at home and abroad have conducted relevant research on the impermeability of concrete based on Fick's law, and made great progress. Xiao et al. [4] studied the effect of aggregate gradation on the impermeability of concrete. Jiang et al. [5] studied the factors affecting the impermeability of concrete sealants. Uysal et al. [6] studied the effects of mineral admixtures on the mechanical properties, chloride ion permeability and impermeability of self-compacting concrete. Wawrzeńczyk et al. [7] studied the effect of ground blast furnace slag and polymer microspheres on the impermeability and freeze-thaw resistance of concrete. However, the influencing factors of concrete affect each other, resulting in complex nonlinear laws of chloride ion in the process of infiltrating...
concrete. Therefore, the traditional method of establishing analytical formulas to predict the impermeability of concrete through mechanism analysis has limitations[8].

The support vector machine model can overcome the non-linear problem. It is widely used in classification, regression, fitting and other problems. For small sample problems, the optimal solution of the original problem can be obtained from the global optimal solution of the quadratic programming problem to effectively reduce the sample point error[9]. Random forest is a multi-classifier integration algorithm, which has the advantages of few model parameters and strong generalization ability. For multi-feature problems, the random forest classifier can sort and optimize the importance of the features, achieve dimensionality reduction, and reduce the phenomenon of overfitting[10].

Therefore, this paper introduces a prediction model based on random forest and support vector machine. Combined with the actual concrete test data, 11 factors were selected as the impact indexes of concrete impermeability to predict the chloride ion permeability coefficient. The model is compared with the BP neural network model and the support vector machine model without index screening, which verifies the reliability of the model and provides a new method for prediction of impermeability of concrete.

2. Preliminaries

2.1. Random Forest
Random Forest is a statistical learning theory based on Classification Tree[11]. It uses the Bootstrap resampling method to extract multiple independent samples from the original sample set, model and construct a decision tree for each sample, and then average the predicted values of the multiple decision trees to obtain the final prediction result. When extracting the sample set, the probability that each sample is not drawn is \( \left(1 - \frac{1}{N}\right)^N \), called out-of-bag data (OBB). The generalization error used to calculate the model can be expressed by formula (1)[12]:

\[
E^* = P_{X,Y} (M(X,Y) < 0)
\]

Among them, the subscripts \( X \) and \( Y \) indicate that the probability \( P \) covers the \( X \) and \( Y \) spaces.

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:

\[
I = \frac{\sum_{i=1}^{N} (r_1 - r_2)}{N}
\]

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.

2.2. Support Vector Machine
Support vector machines are divided into linear regression support machines and nonlinear regression support machines according to regression types. When applying support vector machines to solve linear regression problems, setting the samples to \( n \)-dimensional vectors can transform the optimization problem into a maximization problem under the constraint of formula (3)[13].

\[
\sum_{i=1}^{k} (a_i - a_i^*) = 0, \quad 0 \leq a_i, a_i^* \leq C, i = 1,2,...,k
\]

The expression of the regression function is:

\[
f(x) = \sum_{i=1}^{n} (a_i - a_i^*) (x_i \cdot x) + b
\]
Where \( a_i, a_i^* \) is the Lagrange multiplier and \( \langle x_i, x_j \rangle \) is the linear kernel function.

When using support vector machine to solve the nonlinear regression problem, the solution effect is poor in the low-dimensional space, and the deviation of the solution result is large. It is necessary to apply linear regression in the high-dimensional feature space. Therefore, it is necessary to first map each data into a high-dimensional feature space with a non-linear map, and then solve it, so as to obtain the effect of non-linear regression in the original space. After mapping the sample data into the high-dimensional space, the original problem is transformed into solving the maximization problem under the constraint of formula (3).

The expression of the regression function is [14]:

\[
f(x) = \sum_{i=1}^{k} (a_i - a_i^*) K(x_i, x) + b
\]

Where \( K(x_i, x_j) = \phi(x_i) \phi(x_j) \) represents the kernel function.

3. Establishment of RF-SVM prediction model

There are many factors that affect the permeability of concrete, mainly including cement, additives, water-to-binder ratio, etc. The relationship between these factors and the chloride ion permeability coefficient is complex and non-linear. It is very difficult to predict with expressions. In this paper, the method of random forest and support vector machine is used to predict the chloride permeability coefficient of concrete. The model construction process is as follows:

First, establish the original training set. Based on a large amount of literature and engineering practice, the concrete penetration mechanism is analyzed to obtain factors that have a greater impact on the concrete's impermeability. Considering the influence of materials, mix ratio and concrete properties on impermeability, water-binder ratio, sand ratio, cement strength, cement dosage, fly ash dosage, fine aggregate, coarse aggregate, water reducing agent, air entraining agent, expansion agent and silica fume are chosen as independent variables, chloride ion permeability coefficient as dependent variable. A set of evaluation index system for concrete impermeability is constructed.

Second, select feature based on random forest variables:

- Establish a regression decision tree for each sample set to predict OOB (out-of-bag data). Noise interference is randomly added to each feature, and the importance of the feature is measured by the reduction in model accuracy. The importance degree of each characteristic variable can be calculated by formula (2).
- The feature selection is performed on the training data set based on the sequence backward elimination method. By comparing the changes of the mean square error of the new model after removing the variables, the model with the smallest mean square error is selected, and the optimal number of features is determined. The screened index set is used as the input variables of the support vector machine model.

Third, establish support vector machine model:

- Kernel function selection. For nonlinear problems, the kernel function can transform the problem into a linear problem, thereby generalizing the function with good performance. The kernel function has a great influence on the prediction accuracy of the support vector machine. Different prediction models should choose the appropriate kernel function. The Gaussian kernel function has both the advantages of the radial basis kernel function and good anti-interference ability. In this paper, the Gaussian kernel function is used as the kernel function of the prediction model. Its expression is

\[
K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)
\]

- Kernel function parameters optimization. After the kernel function is selected, the relevant parameters include the width parameter \( \sigma^2 \) and the penalty coefficient \( C \), which will greatly
affect the generalization level of the support vector machine. In this paper, the grid search method is used to optimize the parameters. All possible combinations of the parameters are listed to generate a "grid". The optimal solution is determined by searching all the results in the grid, and the support vector machine is constructed by K-fold cross-validation. The performance of the model is verified, and the parameter with the highest model accuracy is selected as the optimal parameter.

Fourth, forecast result evaluation. Two parameters, root mean square error (RMSE) and goodness-of-fit ($R^2$), are used to judge the prediction accuracy of the model. RMSE can reflect the degree of dispersion between the predicted value and the true value, $R^2$ can verify the degree of fit between the predicted value and the true value.

\begin{equation}
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\end{equation}

\begin{equation}
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\end{equation}

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

4. Case study

4.1. Original training set establishment

The data samples obtained in this paper are from seven bidding sections of a highway project. 11 indicators, such as water-binder ratio, sand ratio, and cement strength, are selected as the influencing factors of concrete impermeability and used as the model input variables. The chloride ion permeability coefficient is selected as the output variable, and 100 sets of data corresponding to each index are selected as the original training set. The specific sample data is shown in Table 1:

| Water-binder ratio | Sand ratio (%) | Cement strength (Mpa) | Cement dosage (kg/m³) | Fly ash dosage (kg/m³) | Fine aggregate (kg/m³) | Coarse aggregate (kg/m³) | Water reducing agent (%) | Air entraining agent (kg/m³) | Expansion agent (kg/m³) | Silica fume (kg/m³) | Chloride ion permeability coefficient ($10^{-12}$ m²/s) |
|---------------------|----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------------------|---------------------------|--------------------------|----------------|-----------------------------|
| 0.31                | 37             | 61.3                  | 365                   | 85                    | 658                   | 1121                  | 1.0                      | 0.948                     | 0                        | 24             | 3.66                        |
| 0.32                | 40             | 58                    | 422                   | 47                    | 732                   | 1099                  | 1.2                      | 0                         | 0                        | 0               | 4.15                        |
| 0.29                | 40             | 55.8                  | 407                   | 48                    | 733                   | 1099                  | 1.1                      | 0                         | 24                       | 0               | 4                           |
| 0.3                 | 41             | 58.9                  | 416                   | 43                    | 747                   | 1075                  | 1.2                      | 0                         | 24                       | 0               | 4.1                         |
| 0.3                 | 41             | 56.2                  | 407                   | 56                    | 758                   | 1090                  | 1.2                      | 0                         | 0                        | 0               | 4                           |
| 0.3                 | 38             | 56.3                  | 388                   | 70                    | 687                   | 1122                  | 1.0                      | 0                         | 0                        | 0               | 22                          |
| 0.37                | 37             | 58.4                  | 352                   | 52                    | 676                   | 1140                  | 1.0                      | 0                         | 47                       | 0               | 3.52                        |
| 0.33                | 37             | 57                    | 344                   | 61                    | 667                   | 1156                  | 0.8                      | 0                         | 48                       | 0               | 3.44                        |
| 0.36                | 37             | 58.4                  | 385                   | 45                    | 681                   | 1136                  | 1.0                      | 0.894                     | 0                        | 0               | 3.83                        |
| 0.34                | 37             | 59.6                  | 415                   | 47                    | 675                   | 1146                  | 1.0                      | 0.894                     | 24                       | 0               | 4.1                         |

4.2. Feature selection based on random forest

If the 11 influencing factors of concrete impermeability in Table 1 are directly used for model construction, it is easy for the model to overfit. By screening the influencing factors, the redundant variables with weaker influence on the permeability coefficient of chloride ion are eliminated, and an appropriate number of influencing factors are selected to construct the support vector machine prediction model.

- Initial sample set classification. The original data set is randomly divided into five parts, four parts are selected as the training set samples, which are used to construct a random forest model, determine the model parameters, sort the importance of the impact indicators, and filter out the
appropriate number of impact indicators. The remaining one part is used as a test set sample for testing the constructed model.

- Index importance calculation. The parameters of the random forest model include the number of random features mtry and the decision tree ntree. In regression analysis, mtry is generally 1/3 of the number of input influencing factors. In this paper, mtry = 4 and ntree = 600. Obtain the importance of each influencing factor and arrange it in descending order to obtain the importance evaluation results of each influencing factor in the constructed random forest model, as shown in Figure 1. It can be found that water-binder ratio, cement dosage, cement strength and other influencing factors have a larger measure of importance.

- Feature selection of sequence backward elimination method. The sequence backward elimination method is used to select features from the training data set to find the optimal number of features. Through cross-validation, we can get the RMSE change trend of the random forest model constructed with different number of characteristic variables, as shown in Figure 2. It can be seen from Figure 2 that when the number of selected feature variables is six, the RMSE of the random forest model reaches the minimum, and the constructed model is optimal. With reference to the ranking results of the importance of the influencing factors in Fig. 1, the influencing factors with weak importance ranking are eliminated one by one until the remaining influencing factors reach the optimal number of characteristic variables. The final impermeability index of concrete determined by the random forest model is water-binder ratio, cement dosage, cement strength, fine aggregate, water reducer and coarse aggregate.
4.3. Support vector machine model prediction
This paper randomly selects 80 groups from 100 sets of original data for training the support vector machine model, and the remaining 20 sets of data are used to test the support vector machine model. The input variables selected in this paper include water-binder ratio, cement dosage, cement strength, fine aggregate, water-reducing agent and coarse aggregate. The output variable of the constructed support vector machine model is the chloride ion permeability coefficient. The results are shown in Figure 3. It is found that the predicted value of the constructed support vector machine model is very close to the actual value when predicting the chloride ion permeability coefficient, indicating that the model has a good prediction effect.
4.4. Evaluation of regression prediction results

In order to verify the superiority of the random forest and the support vector machine model (RF-SVM) constructed in this paper, it is compared with the BP neural network model and the support vector machine model that do not use random forest for feature selection. Three models are used to predict the chloride ion permeability coefficient of concrete, and the root mean square error (RMSE) and the goodness of fit (R^2) are used to measure the prediction effect of the model. The RMSE value represents the sum of individual differences between the estimated value and the actual observed value. The closer to 0, the closer the predicted data to the observed data. R^2 is the ratio of variability in the data set explained by the model. The closer to 1, the better the prediction data fit. The results of the three prediction models are shown in Table 2:

| Model                  | Performance |
|------------------------|-------------|
|                        | RMSE        | R^2         |
| RF-SVM                 | 2.846 × 10^-7 | 0.9513      |
| SVM                    | 3.689 × 10^-5 | 0.9186      |
| BP neural network      | 0.087       | 0.7175      |

It can be seen from Table 2 that the RMSE and R^2 of the random forest and support vector machine model constructed in this paper is much better than the prediction results of the BP neural network model and the support vector machine model without index screening, indicating that the RF-SVM model has higher prediction accuracy and better effect.

5. Conclusion

- There are many difficulties in the prediction of concrete impermeability, such as many influencing factors, complex sample data and non-linear relationship between data. This paper introduces the method of random forest combined with support vector machine to predict the impermeability of concrete. The index selection of random forest model can reduce the dimensionality and overfitting phenomenon, thereby improving the prediction accuracy of the support vector machine model. The proposed RF-SVM model provides an effective method for predicting the impermeability of concrete.

- Taking a highway project as research background, this paper proposes 11 concrete impermeability impact indicators based on the concrete impermeability mechanism. After random forest feature screening, water-binder ratio, cement dosage, cement strength, fine aggregate, water-reducing agent and coarse aggregates are used as support vector machine input variables, and concrete chloride ion permeability coefficient is used as model output variables. The reliability of the model is verified by comparing the model predicted value with the actual engineering value.

- This paper compares the prediction effect of random forest and support vector machine model with BP neural network model and the support vector machine model without feature selection. The results show that the prediction result obtained by the random forest combined with the support vector machine model is smaller in RMSE and R^2 is closer to 1, indicating that the RF-SVM prediction model has a greater advantage in the study of the impermeability of concrete structures.

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] Singh, A. and B. Pandey, A New Intelligent Medical Decision Support System Based on Enhanced Hierarchical Clustering and Random Decision Forest for the Classification of Alcoholic Liver Damage, Primary Hepatoma, Liver Cirrhosis, and Cholelithiasis. Journal of Healthcare Engineering, 2018. 2018: p. 1-9.
[2] Zhang, T. and O.E. Gjørv, Diffusion behavior of chloride ions in concrete. Cement and Concrete Research, 1996. 26(6): p. 907-917.
[3] Qingzhou, Z., et al., Leakage Zone Identification in Large-Scale Water Distribution Systems Using Multiclass Support Vector Machines. 2016.
[4] Xiao, H., et al., Role of nano-SiO2 in improving the microstructure and impermeability of concrete with different aggregate gradations. Construction and Building Materials, 2018. 188: p. 537-545.
[5] Jiang, L., et al., The investigation of factors affecting the water impermeability of inorganic sodium silicate-based concrete sealers. Construction and Building Materials, 2015. 93: p. 729-736.
[6] Uysal, M., K. Yilmaz, and M. Ipek, The effect of mineral admixtures on mechanical properties, chloride ion permeability and impermeability of self-compacting concrete. Construction and Building Materials, 2012. 27(1): p. 263-270.
[7] Wawrzeńczyk, J., A. Molendowska, and A. Klak, Effect of Ground Granulated Blast Furnace Slag and Polymer Microspheres on Impermeability and Freeze-thaw Resistance of Concrete. Procedia Engineering, 2016. 161: p. 79-84.
[8] Cheng, Y., et al., Durability of concrete incorporated with siliceous iron tailings. Construction and Building Materials, 2020. 242: p. 118147.
[9] Yan, H., et al., Predicting permeability changes with injecting CO2 in coal seams during CO2 geological sequestration: A comparative study among six SVM-based hybrid models. Science of The Total Environment, 2020. 705: p. 135941.
[10] Li, L., et al., Prediction of plant transpiration from environmental parameters and relative leaf area index using the random forest regression algorithm. Journal of Cleaner Production, 2020. 261: p. 121136.
[11] Lau, K., M. Salibian-Barrera, and L. Lampe, Modulation recognition in the 868MHz band using classification trees and random forests. AEU - International Journal of Electronics and Communications, 2016. 70(9): p. 1321-1328.
[12] Verikas, A., A. Gelzinis, and M. Bacauskiene, Mining data with random forests: A survey and results of new tests. Pattern Recognition, 2011. 44(2): p. 330-349.
[13] Shao, M., et al., Prediction of energy consumption in hotel buildings via support vector machines. Sustainable Cities and Society, 2020. 57: p. 102128.
[14] Prayogo, D., et al., Combining machine learning models via adaptive ensemble weighting for prediction of shear capacity of reinforced-concrete deep beams. Engineering with Computers, 2019.