Displacement Prediction Model for Concrete Dam Based on PSO-GBDT

Zhang Jun¹²*, Mo Jian³, Xie Jiemin¹², Meng Zuohong¹

¹ State Grid Hunan Electric Power Company Limited Research Institute, Changsha 410007, China
² Hunan Xiangdian Test & Research Institute, Changsha 410007, China
³ State Grid Hunan Electric Power Company Limited, Changsha 410007, China
*Corresponding author’s e-mail: sghnzzj1988@163.com

Abstract: Displacement plays a vital role in concrete dam safety monitoring data. It is affected by a multitude of factors such as water pressure, temperature, structure stiffness, and bottom bedrock settlement. A displacement prediction model based on gradient boosting decision tree (GBDT) is proposed to make accurate predictions. Meanwhile, particle swarm optimization (PSO) algorithm is used to optimize the model parameters, including learning rate, the trees’ number, the maximum optional variables, and the maximum depth of each tree. Finally, a real-world radial displacement dataset is used to compare the performance of the PSO-GBDT model, the multiple linear regression model (MLR), and the stepwise regression model (SR). The results indicate that the PSO-GBDT model achieves the best performance. Moreover, the model can also analyse the importance of each variable, which has substantial practical value.

1. Introduction
Dam monitoring is an efficient way to the safety supervises of the dam. The dam monitoring data can be obtained in time and used to detect anomaly by modelling and prediction. As an essential part of dam monitoring data, displacement is frequently predicted by linear regression models [1], such as multiple linear regression (MLR), stepwise regression (SR). Due to the influence of various internal and external factors, such as structural stiffness, water pressure, temperature, and bedrock settlement, the dam displacement is complicated, random and uncertain, and shows highly nonlinear behaviour. In recent years, some improvements are developed in nonlinear displacement prediction models based on neural network, support vector machine, and ensemble learning [2-4]. In this paper, Gradient Boosting Decision Tree (GBDT) algorithm from ensemble learning is adopted to predict the displacement of a concrete dam. Particle Swarm Optimization (PSO) algorithm is used to optimize the parameters of the GBDT model, which finally compared with MLR and SR models.

2. PSO - GBDT model

2.1. GBDT model
GBDT is an iterative decision tree algorithm. Each decision tree’s calculation result is sum up to form the final result. One decision tree learns the residual between the former trees’ summed results and output. The algorithm assumes that every decision tree has a dependency and generated by the serialization method. The decision tree uses the Classification and Regression tree (CART) model.
For the regression problem, CART model divides each region of the input space into two sub-regions recursively and determines the output value of each sub-region, and then constructs a binary decision tree. The steps are shown as follows [5,6]:

Step 1: Select the optimal shard variable \( j \) and shard point \( s \), and solve:

\[
f(x) = \min_{j,s} \left[ \min_{c_1} \sum_{x \in R_1(j,s)} (y - c_1)^2 + \min_{c_2} \sum_{x \in R_2(j,s)} (y - c_2)^2 \right]
\]  

(1)

Traverse the variable \( j \), scan its shard point \( s \), and select the (\( j, s \)) which makes the above equation reach the minimum value.

Step 2: Determine the output value of each divided sub-region by the selected (\( j, s \)):

\[
R_1(j, s) = \{ x | x(j) \leq s \}, \quad R_2(j, s) = \{ x | x(j) > s \}
\]  

(2)

\[
c_m(j, s) = \frac{1}{N_m} \sum_{x \in R_m(j, s)} y_i, \quad x \in R_m, \quad m = 1, 2
\]  

(3)

Where \( R_1 \) and \( R_2 \) are the two sub-regions generated by the shard point \( s \) of variable \( j \); \( c_m(j, s) \) is the mean value of the output in the \( m \)-th region; \( N_m \) is the samples’ number in the \( m \)-th region.

Step 3: Continue to perform steps 1 and 2 in the two sub-regions until specified conditions are met.

Step 4: Finally, the input space is divided into \( K \) units \( R_1, R_2, \ldots, R_K \), and each unit \( R_i \) has a fixed output value \( c_i \). Thus, the decision tree can be expressed as:

\[
T(x; \theta) = \sum_{k=1}^{K} c_k I(x \in R_k)
\]  

(4)

Where \( T(x; \theta) \) represents the decision tree; \( \theta \) represent the decision tree’s parameters; \( I(x \in R_k) \) is the indicator function, which is 1 when \( x \in R_k \), otherwise is 0.

The GBDT model can be expressed as the additive model of the decision trees:

\[
f_M(x) = \sum_{m=1}^{M} T(x; \theta_m)
\]  

(5)

Where \( M \) denotes the decision trees’ number.

The GBDT model adopted the forward distribution algorithm. The initial decision tree \( f_0(x)=0 \) is specified, and the \( m \)-th decision tree model is:

\[
f_m(x) = f_{m-1}(x) + T(x; \theta_m)
\]  

(6)

Where \( f_{m-1}(x) \) is the current model, and the next decision tree’s parameters are determined through empirical risk minimization:

\[
\hat{\theta}_m = \arg \min_{\theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))
\]  

(7)

Where \( \hat{\theta}_m \) denote the parameters of the \( m \)-th decision tree.

For the regression problem, the squared error loss function \( S \) is generally adopted:

\[
S = \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m)) = \sum_{i=1}^{N} (y_i - f_{m-1}(x_i) - T(x_i; \theta_m))^2
\]  

(8)

Assuming \( r_{mi} = y_i - f_{m-1}(x_i) \), thus \( r_{mi} \) is the \( i \)-th sample residual error between the current model and output, so the decision tree only needs to fit the residual error.

To avoid overfitting, the learning rate, also known as the regularization parameter, is added to the equation (6):

\[
f_m(x) = f_{m-1}(x) + \nu T(x; \theta_m), \quad \nu \in (0, 1]
\]  

(9)

For the same training set, smaller \( \nu \) needs more decision trees to learn. Therefore, it is vital to optimizing the learning rate and the decision trees’ number.

The GBDT model can also calculate the importance degree of each variable. The global importance of variable \( j \) is the average importance degree of all the decision trees:

\[
\bar{I}_j = \frac{1}{M} \sum_{m=1}^{M} I_j(T(x; \theta_m))
\]  

(10)

\[
I_j(T(x; \theta)) = \sum_{t=1}^{Q} \Delta S_t I(\phi_t = j)
\]  

(11)

Where \( Q \) is the leaf nodes’ number of each decision tree; for a binary decision tree, \( Q - 1 \) is the non-leaf nodes’ number; \( \phi_t \) are the variables associated with node \( t \); \( \Delta S_t \) is the reduced value of the loss function \( S \) by the splitting of node \( t \).

The relative global importance index is to facilitate the comparison of each variable:
\[ \hat{j}_j = \frac{j_j}{\sum_{j=1}^{J} j_j} \]  

(12)

Where \( J \) is the variables’ number.

2.2 PSO - GBDT model

For PSO algorithm, each particle is a point in the n-dimensional space and has different individual fitness corresponding to the objective function. Each particle adjusts its flight path to the best point, and its speed and position are adjusted as follows [7]:

\[
v^{t+1}(i, k) = v^t(i, k) + c_1 \text{rand}[p^t_{\text{best}}(i, k) - x^t(i, k)] + c_2 \text{rand}[g^t_{\text{best}}(i, k) - x^t(i, k)]
\]

(13)

\[
x^{t+1}(i, k) = x^t(i, k) + v^t(i, k)
\]

(14)

Where \( i \) represents the i-th particle; \( k \) represents component coordinates; \( t \) represents iteration times; \( c_1 \) and \( c_2 \) are the acceleration factors with values between 1 and 2; \( \text{rand} \) is a random function with values between 0 and 1; \( p^t_{\text{best}}(i, k) \) and \( g^t_{\text{best}}(i, k) \) are the best positions for individuals and groups respectively.

The calculation flow chart of PSO-GBDT model parameters optimization algorithm is as follows:

![Flow chart of parameters optimization algorithm of PSO-GBDT model](image)

Figure 1. Flow chart of parameters optimization algorithm of PSO-GBDT model

3. The example analysis

The displacement prediction models’ variables were selected according to the reference [1], including water level, temperature, and aging factor. These models’ variables are \( \phi = [H, H^2, H^3, H^4, T_0, T_{1.7}, T_{8.15}, T_{16.30}, T_{31.60}, T_{61.90}, \theta, \ln(\theta + 1)] \). Where \( H \) is the upstream water level; \( T \) is the air temperature; \( \theta \) is the number of days from the initial sample date and divided by 100; the subscript of \( T \) represents the average value of the period in terms of the day.

Considering the dam prediction model such as SR model often needs to screen important variables to alleviate the multicollinearity problem among variables, the maximum number of optional variables of each tree in GBDT model is an important parameter for variable selection. The maximum depth of each tree controls the splitting of the tree, as well as an important parameter that controls overfitting and the maximum node number. The GBDT model parameters finally to be optimized are set as follows: the maximum number of optional variables \( p \) for each tree; the maximum depth \( d \) for each tree; the learning rate \( \nu \); the trees’ number \( i \). These parameters are positive integers except for the learning rate. Therefore,
restrictions should be added in the PSO optimization process: \( p \in [1, 12] \); \( d \in [1, 10] \); \( v \in [0.001, 1] \); \( i \in [10, 500] \). For PSO algorithm, parameters should be normalized and continuous. Therefore, segmentation values are used to correspond to discrete integer parameters. Taking \( d \) as an example, 1 can correspond to the random number [0,0.1), and 10 corresponds to the random number (0.9, 1]. For PSO, the parameters are set as follows: the particles’ number is 30; the iterations’ number is 50; the acceleration factors are all 1.

The training sample is divided into a training set, validation set, and test set according to the reference [8]. The training set is used to fit the model parameters. The validation set is used to select the appropriate model parameters to prevent overfitting. Because for GBDT model, the trees’ number and the fitting precision are related, which grows infinitely if not limited, and will lead to overfitting. The test set is used to assess the generalization ability of the model.

A real-world radial horizontal displacement of the arch crest dam section is used for this study. 237 samples in the period from 2000/6/11 to 2010/2/19 are used as the training set, 52 samples from 2010/3/6~2012/4/9 are used as the validation set, and 52 samples from 2012/4/24~2014/5/29 are used as the test set. The coefficients of MLR and SR models are shown in table 1. The GBDT model parameters after PSO optimization are as follows: the maximum number of optional variables \( p=8 \), the maximum depth \( d=2 \); the learning rate \( v=0.181 \); the trees’ number \( i=410 \). The training process of GBDT model and the relative importance of each variable are shown in figure 2. The measured and predicted displacement of each model are shown in figure 3. The residual errors of the models are shown in figure 4. The performance indicators of each model are shown in table 2, which are formulated as follows:

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \\
\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\
R = \frac{\sum_{i=1}^{N} (y_i - \frac{1}{N} \sum_{i=1}^{N} y_i)(\hat{y}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i)}{\sqrt{\sum_{i=1}^{N} (y_i - \frac{1}{N} \sum_{i=1}^{N} y_i)^2 \sum_{i=1}^{N} (\hat{y}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i)^2}}
\]

Where MAE represents the mean absolute error; MSE represents the mean square error; R represents the correlation coefficient; \( \hat{y}_i \) represents the predicted value of the model; \( y_i \) represents the measured value; \( i \) represents the sample mark; \( N \) represents the samples’ number.

![Training Set Deviance vs. Boosting Iterations](image)

(a) Training process of GBDT model

![Relative Variable Importance](image)

(b) Relative variable importance

Figure 2. The training process of GBDT model and relative variable importance
Figure 3. Process of the measured displacement and the predicted displacement of each model

Figure 4. Residual error process of the measured displacement and the predicted displacement of each model

Table 1. Coefficients of MLR and SR model

| Variables | MLR   | SR    | Variables | MLR   | SR    |
|-----------|-------|-------|-----------|-------|-------|
| $H$       | -0.136| 0.16  | $T_{36-30}$| -0.225| -0.282|
| $H^2$     | 6.090e-04| -0.320| $T_{31-60}$| -0.353|
| $H^3$     | 1.466e-06| -0.594| $T_{61-90}$| -0.571|
| $H^4$     | 3.966e-09| 1.013e-08| $\theta$ | -0.203e-03|
| $T_0$     | -0.049| ln($\theta+1$) | 2.206e+02|
| $T_{1.7}$ | -0.153| -0.239| const     | -39.913|
| $T_{8.15}$| -0.117|       |           | -33.868|

Table 2. Performance indexes of each model

| model    | MAE  | MSE  | R     |
|----------|------|------|-------|
|          | training | validation | test | training | validation | test | training | validation | test |
| MLR      | 1.29  | 3.32 | 4.39  | 2.76 | 14.25  | 22.12 | 0.982    | 0.981     | 0.980 |
| SR       | 1.45  | 1.82 | 1.59  | 4.35 | 4.82   | 3.91  | 0.978    | 0.978     | 0.981 |
| PSO-GBDT | 0.11  | 1.82 | 1.35  | 0.02 | 4.81   | 2.79  | 1.000    | 0.973     | 0.986 |

MAE = mean absolute error; MSE = mean square error; R = the correlation coefficient. The best results are highlighted in boldface.

As can be seen from figure 2 (a), when the MSE of the training set decreases, the MSEs of the validation set and test set also decrease and tend to be stable. Therefore, the GBDT model has a relatively
ideal training effect. Figure 2 (b) shows that $T_{31-60}$ and $T_{16-30}$ are more important than other variables. The $H^4$ is the most crucial variable of the water level variables. However, it has not many differences among water level variables. The aging variables and $T_{61-90}$ have little influence on displacement, while the influences of $T_0$, $T_{1-7}$, and $T_{8-15}$ are almost negligible. Table 1 shows that the aging component $\theta$ coefficient of MLR model is negative, which will lead to a downward trend. Since this trend does not originate from water level or temperature, it has strong assumption for future predictions. The SR model does not select the aging components. From figure 3 and figure 4, it can be seen that the trend forecast of MLR model deviates from the measured values at the validation and test periods, while the SR and PSO-GBDT models are relatively stable. Table 2 shows that the PSO-GBDT model has excellent performance in the training set, validation set, and test set. Its MAE and MSE are smaller than MLR and SR models, and the R also reaches the best value except for the validation set.

4. Conclusion
The displacement prediction model of concrete dam based on GBDT is established, and the PSO algorithm is used to optimize the parameters of GBDT model. The concrete dam example shows that the PSO-GBDT model has excellent prediction performance on the displacement monitoring data, and is better than the commonly used MLR and SR models. Meanwhile, the PSO-GBDT model can also analyse the importance of each variable, which shows strong substantial value.

References
[1] He Jinping. (2010) Theory and application of dam safety monitoring. China water resources and hydropower press. Beijing.
[2] Ranković, V., Grujović, N., Divac, D. et al. (2014) Development of support vector regression identification model for prediction of dam structural behaviour. Structural Safety, 48, 33-39.
[3] Stojanovic, B., Milivojevic, M., Milivojevic, N. et al. (2016) A self-tuning system for dam behavior modeling based on evolving artificial neural networks. Advances in Engineering Software, 97: 85-95.
[4] Salazar, F., Toledo, M. A., Oñate, E. et al. (2015) An empirical comparison of machine learning techniques for dam behaviour modelling. Structural Safety, 56, 9-17.
[5] Friedman J H. (2001) Greedy Function Approximation: A Gradient Boosting Machin. The Annals of Statistics, 29(5):1189-1232.
[6] Li Hang. (2012) Statistical learning method. Tsinghua university press. Beijing.
[7] Eberhart R, Kennedy J. (1995) A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, IEEE. Nagoya. pp. 39-43.
[8] Fulkerson, B., Ripley, B. D. (1997) Pattern recognition and neural networks. Technometrics, 39(2), 233.