Ensemble of bagged regression trees for concrete dam deformation predicting

Jiemin Xie\textsuperscript{1,2*}, Jun Zhang\textsuperscript{1,2}, Xuan Xie\textsuperscript{3}, Zhiwei Bi\textsuperscript{2}, Zhuoheng Li\textsuperscript{4}

\textsuperscript{1}State Grid Hunan Electric Power Company Limited Research Institute, Changsha 410007, China
\textsuperscript{2}Hunan Xiangdian Test & Research Institute, Changsha 410007, China
\textsuperscript{3}College of Hydraulic & Environmental Engineering Institute, China Three Gorges University, Yichang 443002, China
\textsuperscript{4}State Grid Hunan Electric Power Company Limited Material Company (Branch), Changsha 410004, China
\*Corresponding author’s e-mail: jemmy_1989@126.com

Abstract. Tree based ensemble learning algorithms involve generating multiple trees that together produce a consistent prediction and often significantly improving the accuracy of the prediction. In this paper, the technique is further explored, and a dam displacement prediction model based on bootstrap aggregating (bagged) regression trees (BART) is proposed. The case study of concrete dam indicates that the performance of bagged regression trees is better than the conventional regression tree (RT) model in predicting dam displacement. Moreover, the model can also analyze the importance of each variable, which has strong practical value.

1. Introduction

With the continuous development of the electric power industry, hydropower has entered a new era of large units, ultra-high voltage transmission and intelligent management. The proportion of hydraulic energy generation in the power supply of China is gradually increasing. As the most important civil structure of hydropower plant, the safety of dam appear to be particularly important. Yet due to long operation and service time, a number of safety issues occur in dams, such as serious aging of equipment, low level of observation automation. Which may lead to failure or partial failure of the dams, ultimately bring great harm to people's life and property [1]. Therefore it’s important to monitor the behavior of each dam during its operation to assess its safety on a continual basis.

Dam deformation prediction is a hot topic in dam monitoring research. By modeling and predicting the monitoring data of the dam, the safe operation state of the dam can be obtained in time. As an important part of dam monitoring data, displacement is often predicted by linear regression model, such as multiple linear regression, stepwise regression etc. Due to the comprehensive influence of various internal and external factors, such as the structural stiffness, water pressure, temperature and bottom bedrock settlement, the displacement response of the dam is complicated, random and uncertain, thus showing highly nonlinear behavior [2,3]. In recent years, tree based ensemble learning algorithms are very popular in machine learning applications [4,5]. These algorithms make the prediction model more accurate, stable and easy to interpret. In order to further explore the information reflected by the dam deformation data, a prediction model based on ensemble of bagged
regression trees (BART) is adopted and compared to regression tree (RT) model in predicting dam displacement. The performance differences between both methods are well discussed.

2. Bagged regression trees

The observed values are denoted as \( D = \{X_i, Y_i\}, i = 1, 2, \ldots, N \), where \( N \) is the number of observations. The goal of a predictive model is to estimate the value of an output variable \( Y_i \) (i.e. radial displacement of dam in our case), based on a set of predictors \( X_i \) (water level, air temperature, etc.), i.e. \( \hat{Y}_i = f(X_i) \).

2.1. Regression Trees

Decision trees are divided into two categories, including classification trees and regression trees (CART), which are statistical models first proposed by Breiman et al. [6]. Each step of the evaluation involves examining the value of an input variable and posing a binary problem that divides the node into two child nodes.

The basic idea of regression analysis is to make the smallest difference between the sample value and the fitting value [7]. For regression problem, CART model recursively divides each region into two sub-regions and determines the response of each sub-region in the input space where the training data set is located, and constructs binary decision tree. Specific steps are as follows. First, select the optimal splitting variable \( j \) and splitting point \( s \), and solve:

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

where, \( c_1, c_2 \) are the fixed responses for each sub-region. Traverse the variable \( j \), scan the splitting point \( s \) for the fixed splitting variable \( j \), and select \((j,s)\) that makes the above equation reach the minimum value.

Then, use the selected \((j,s)\) to divide regions and determine the corresponding response:

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

\[
\hat{c}_m(j,s) = \frac{1}{N_m} \sum_{x \in R_{j,s}} y_j x \in R_m, \quad m = 1, 2
\]

where, \( R_1 \) and \( R_2 \) are the two regions generated by the splitting point \( s \), \( \hat{c}_m(j,s) \) is the mean response of the \( m \)-th region, and \( N_m \) is total number of the observations in the \( m \)-th region.

Generally, divide the input space into units \( R_1, R_2, \ldots, R_m \). Repeat above process for each region until the stop condition is met, one can get the regression tree model represented as:

\[
f(x) = \sum_{m=1}^M \hat{c}_m I(x \in R_m)
\]

where, \( I(x \in R_m) \) is indicator function, namely \( I = 1 \) when \( x \in R_m \), otherwise \( I = 0 \).

2.2. Bootstrap aggregating (bagging)

Decision trees are usually combined with bootstrap aggregating (bagging) approach to form stronger estimators [8], which is a general process to reduce the high variance of CART. Moreover, bagged decision trees combine the results of multiple decision trees to reduce the influence of over-fitting and improve the generalization ability. This combining process is called ensemble learning. Bagging trains multiple (\( n \)) models in different samples and creates an estimate by averaging the results of \( n \) models (figure 1). Since we cannot access multiple training datasets, we can bootstrap by repeatedly sampling from a single training dataset. In this way, we get \( T \) different bootstrapped training datasets. Then we train our method on the \( t \)-th bootstrapped training set to get a prediction \( \hat{f}^{(t)}(x) \), and thus obtain an aggregate prediction. Therefore bagging for regression can be expressed as:

\[
\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T \hat{f}^{(t)}(x)
\]
In general, bagging reduces the error by reducing variance in the results due to unstable learners, such as decision trees whose output may change greatly when the training data is slightly changed [8]. As can be seen that on average, each bagged tree takes advantage of about two-thirds of the observations. The remaining one third of the observations do not fit into a given bagged tree, and these observations are called out-of-bag (OOB) observations. Those OOB observations are used as validation data sets to estimate prediction errors and help avoid severe over-fitting. The resulting OOB error is a valid estimate of the bagged regression model test error.

\[
\hat{f}_{\text{bagging}} = \frac{1}{T} \sum_{i=1}^{T} \hat{f}_i^* (x)
\]  

Figure 1. (a) The bagging training procedure; (b) The prediction procedure with trained model.

3. Case study

3.1. Dataset and model variables
The data used for the study correspond to the dam of Dongjiang hydropower plant. It is a double curvature arch dam, with a height of 157 m, which entered into service in 1980. The available time
series data set cover 14 years (from 26/6/2000 to 29/5/2014 with 340 displacement samples). Figure 2 depicts the reservoir level variation in the period considered. All the calculations were performed on a training set containing 300 training samples to build and fit model, and the model accuracy was assessed for a validation set containing 40 validation samples. The number of trees in the bagging ensemble was 9.

For input variables selection, RT and BART adopt 10 variables according to the statistical model involving aging factor, reservoir level and air temperature $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}\} = \{\theta-\theta_0, \ln(\theta-\ln\theta_0), (H-H_0)^2, (H-H_0)^3, \sin(2\pi t/365)-\sin(2\pi t_0/365), \cos(2\pi t/365)-\cos(2\pi t_0/365), \sin(4\pi t/365)-\sin(4\pi t_0/365), \cos(4\pi t/365)-\cos(4\pi t_0/365)\}$, where $H_0, H$ stands for reservoir level of the starting observation day and the underway observation day respectively; the trigonometric function is used to simulate the temperature field of quasi-steady dam operation for many years, $t$ represents the number of days between the beginning observation day and the underway observation day, $t_0$ represents the number of days between the beginning observation day and the first day of modeling; aging factor $\theta = t/100, \theta_0 = t_0/100$.

![Figure 2. Time series of the reservoir level at Dongjiang arch dam.](image)

### 3.2. Model Evaluation Index

The evaluation indexes of the models adopt mean absolute error (MAE), mean square error (MSE), mean relative error (MRE), and correlation coefficient (R):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$$  \hspace{1cm} (6)

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$  \hspace{1cm} (7)

$$\text{MRE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{|y_i|}$$  \hspace{1cm} (8)

$$R = \frac{\sum_{i=1}^{N} (y_i - \frac{1}{N} \sum_{i=1}^{N} y_i)(\hat{y}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i)}{\sqrt{\sum_{i=1}^{N} (y_i - \frac{1}{N} \sum_{i=1}^{N} y_i)^2 \sum_{i=1}^{N} (\hat{y}_i - \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i)^2}}$$  \hspace{1cm} (9)

where $y_i$ are the measured values and $\hat{y}_i$ are the predicted values, $N$ is the size of the training/validation set.
3.3. Model Performance Analysis

The relative importance of each predictor variable is shown in figure 3. It can be seen that temperature variables $v_7$, $v_8$ and $v_{10}$ are much more important than other variables. Among the rest variables, two aging variables, $v_3$ and $v_9$ have slightly higher influence on dam deformation than other variables.

![Figure 3. Predictor variable importance of BART model.](image)

It can be seen from figure 4 that the prediction of BART model is more stable than that of RT model. And table 1 indicates that the BART model performs well in both the training set and the validation set. Its MAE, MSE, MRE are smaller than RT model, and the correlation coefficient R also reaches the best value.

| Model | MAE Training | MAE Validation | MSE Training | MSE Validation | MRE Training | MRE Validation | R Training | R Validation |
|-------|--------------|----------------|--------------|----------------|--------------|----------------|------------|--------------|
| RT    | 1.347        | 1.766          | 3.246        | 4.972          | 0.887        | 0.607          | 0.981      | 0.967        |
| BRT   | **1.218**    | **1.390**      | **2.538**    | **2.953**      | **0.506**    | **0.330**      | **0.986**  | **0.982**    |
4. Conclusion
A concrete dam displacement prediction model based on BART is adopted to predicting Dongjiang dam deformation. The case study shows that the model has good prediction effect on the displacement monitoring data of concrete dam, and its performance is better than the traditional RT model. The model can also analyze the importance of each variable, showing strong practical value.

References
[1] Wu, J. (2009) Dam Safety Monitoring Theory and Test Technology. China Water Resources and Hydropower press, Beijing.
[2] Mata, J., Tavares de Castro, A., Sã da Costa, J. (2014) Constructing statistical models for arch dam deformation. Struct. Control Health Monit., 21: 423-437.
[3] De Sortis, A., Paoliani, P. (2007) Statistical analysis and structural identification in concrete dam monitoring. Eng. Struct., 29: 110–120.
[4] Kotsiantis, S.B., Kanellopoulos, D., Zaharakis, I.D. (2006) Bagged Averaging of Regression Models. In: IFIP International Federation for Information Processing, vol. 204, Artificial Intelligence Applications and Innovations, eds. Maglogiannis, I., Karpouzis, K., Bramer, M., Springer, Boston. pp. 53-60.
[5] Salazar, F., Toledo, M., Oñate, E., Suárez, B. (2016) Interpretation of dam deformation and leakage with boosted regression trees. Eng. Struct., 119: 230-251.
[6] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984) Classification and Regression trees. Wadsworth & Brooks, Monterey, CA.
[7] Zhou, Z. (2016) Machine learning. Tsinghua university press, Beijing.
[8] Breiman, L. (1996) Bagging Predictors. Machine Learning, 26: 123–140.