Application of RS-RF Model in Deformation Prediction of Concrete Dam

Zhangjun Guo¹, Huadong Huang²

¹Power Science Research Institute of Shaanxi electric power company, Xi’an, Shaanxi, 710019, China
²NARI Group Corporation, Nanjing, Jiangsu, 211106, China

*Corresponding author’s e-mail: 376573199@qq.com

Abstract. The high-performance concrete dam deformation prediction model serves as an important reference for structural safety behavior diagnosis, early warning, and scientific decision-making, and it is also one of the guarantee measures to fully exert the benefits of the project. This paper aims at the subjectivity of factor selection, multicollinearity among factors, and poor generalization of the concrete dam deformation monitoring model. It combines rough set and random forest theory to achieve feature attribute reduction, importance evaluation, and high-precision prediction. In terms of rough set and random forest advantages, a concrete dam deformation prediction model based on RS-RF was established. The application of engineering examples shows that the deformation monitoring model of concrete dams based on RS-RF can reduce the influence factor set, give the importance of each factor, and do well than commonly used models based on SVM and RF in prediction accuracy. Therefore, the deformation prediction model of concrete dam based on RS-RF achieves optimization of influence factors, which makes up for the shortcomings of intelligent prediction model in quantitative analysis and prediction generalization, and has strong engineering practicability.

1. Introduction

Concrete dams are currently the preferred type of dam construction. They have the advantages of strong adaptability, high safety factor, and convenient operation and maintenance. However, there are many factors affecting the working behavior of concrete dams, and the service behavior a non-linear dynamic evolution process influenced by the interaction of materials and structures under the cooperation of multiple factors [1]. Deformation is a comprehensive effect that directly reflects the safety behavior of concrete dams, and it can be used as an important indicator for structural changes in trend or even change. Therefore, strengthening the research on the deformation safety monitoring model of concrete dams to ensure its longevity is one of the important means of effective service safety [2].

Various factors such as design, construction, and operation management affect its working behavior comprehensively. The prediction model that reflects the relationship between the amount of deformation effect and the influence factors is a complex non-linear function. When using a complete and reliable monitoring data for a fixed-length sequence to forecast the working behavior of concrete dams, the selection of influence factors and modeling methods has a great impact on the forecast results. With regard to the selection of influence factors: the influence factor set of the deformation prediction model is based on the simplified physical model of the dam and its foundation, the embedment of monitoring equipment, and the prototype monitoring data, the influence factor set in the statistical model is used. There is multicollinearity among the water pressure component factors, temperature component factors,
and aging component factors selected by the statistical model, and this feature may reduce the accuracy of the model during modeling and adversely affect the prediction results. At the same time, in the actual application process, the predictive model did not take into account the effects of quantifiable influencing factors such as seepage flow, fracture opening, and lift pressure, and difficult-to-quantify influencing factors such as dam materials, construction quality, and geological conditions on deformation, and the importance of impact factors is also not distinguished [3]. At present, the most commonly used optimization methods for the influence factors of concrete dam deformation prediction models include traditional linear regression methods, dimensionality reduction analysis methods, and fuzzy mathematical analysis methods [4-8]. However, in the practical application process, it is known a priori knowledge method dependent experience heavily and the error is large. The dimensionality reduction analysis principal component analysis requires that variables must have a high linear correlation, it can be obtained higher accuracy analysis results based on this analysis. In fact, the factors that affect the deformation of concrete dams are complex and have strong non-linearity. The gray correlation analysis method in the fuzzy mathematical analysis does not give a criterion for factor selection, it can only give relevance ranking. Therefore, there are problems such as incomplete selection criteria and indiscriminate importance of the influence factors in the process of selecting the influence factors of the concrete dam deformation prediction model, which seriously affect the predictive performance of the model. With regard to the establishment of prediction models: At present, the deformation prediction models of concrete dams mainly include statistical models, deterministic models, mixed models, combined models, and intelligent algorithms models. With the rapid development of dam engineering theory, finite element theory, artificial intelligence technology, etc., the monitoring models established by various scholars have achieved fruitful practical results in scientific research and production. Due to the complexity of water conservancy projects, structural fluctuations and uncertainties in the working conditions, all current prediction model research and applications have not been perfected. Poor robustness, large calculations, low prediction accuracy, overfitting and other deficiencies affecting the reliability and usefulness of the model still exist in the deformation prediction models of concrete dams [9-11].

In recent years, mature rough set theory and random forest algorithm in the field of artificial intelligence have achieved fruitful research results in reducing feature attributes, eliminating redundant information, learning complex non-linear relationships, and improving model generalization. Therefore, in view of the shortcomings in the concrete dam deformation monitoring model and the unique advantages of rough set and random forest, this paper establishes a concrete dam deformation monitoring model based on RS-RF, which realizes the optimization and important measurement of the influence factor set and improves the model's fitting effect and prediction accuracy.

2. RS-RF theory

2.1. Rough Set Theory

The core objectives are the mining and refining of essential information under the premise of maintaining equivalence relations. The main tasks in this approach are attribute reduction, correlation analysis and importance evaluation for uncertain information systems.

2.1.1. Information System. To describe the samples that encompass the necessary information in RS theory, a quaternary information system S is established, and it can be expressed as follows:

\[ S = (U, R, V, f) \]  

where \( U \) is a non-empty finite set of all samples; \( R \) is a set of attributes, including a set of conditional attributes \( C \) and a set of decision attributes \( D \); \( V \) is the attribute value set; and \( f \) is the information function, also known as the decision table.

2.1.2. Attribute Reduction. For arbitrary \( P \subseteq R \) and \( P \neq 0 \), the indistinguishable relationship between \( P \) and \( U \) is defined as follows:

\[ IND(P) = \{(x, y) \in U^2 | \forall \alpha \in P, \alpha(x) = \alpha(y)\} \]  

where \( IND(P) \) is the indistinguishable relationship between \( P \) and \( U \).
For an arbitrary set of objects $X \subseteq U$ and attributes $B \subseteq C$ in a given information system $S$, the approximation of $X$ is defined as $BX = \{x | [x]_B \subseteq X\}$; the approximate definition of $X$ is defined as $BX = \{x | [x]_B \cup X \neq \emptyset\}$; the boundary area of $x$ is defined as $BN_\delta(X) = BX - BX$. In this case, $[x]_B$ represents the set of indistinguishable relations for the division of $U$ by $B$.

If $BN_\delta(X)$ is not empty, then $X$ is called a rough set of $B$. The positive region of $B$ relative to $D$ is as follows:

$$POS_\delta(D) = \{BX | X \in U / IND(D)\}$$

When $SIM = POS_C(D) - POS_{C \setminus \{a\}}(D) = 0$, where $a \in C$, $a$ can be omitted. Additionally, when each element in $C$ is not omissible from $D$, it can be concluded that $C$ is independent of $D$. When $C^* = C \setminus C^*$, where $C^*$ is independent of $D$ and all the elements in $C^*$ can be omitted, then $C^*$ is called the relative reduction of $D$.

2.1.3. Importance Evaluation. In attribute reduction, the importance of the attribute can be defined by the degree of interdependence between the attribute sets $B$ and $D$. The degree of interdependence between $P$ and $R$ is defined as follows:

$$\gamma_\delta(D) = \frac{|POS_\delta(D)|}{|U|}$$

$$Sig(\alpha, B, D) = \gamma_\delta(D) - \gamma_{\delta - \{a\}}(D)$$

2.2. Random Forest theory. RF using the CART decision tree as weak learners [16], and ensemble learning is implemented based on all weak learners. Bootstrap is used to resample [17], and the final result is determined by voting based on the results of the tree analysis. The RF model can improve operation efficiency, and it is robust to multiple collinearities between variables [18].

2.2.1. Sample set selection. The original sample sets of the model include a total of $N$ samples with $M$ features. These data contain complex interactions between independent and dependent variables, which can be linear or non-linear. The original data sets were resampled by the Bootstrap method, and $n$ sample sets were randomly selected from the original sample sets as the training set of the model, with a total of $n$ times. Each taken sample may contain duplicate samples, but not all samples at the same time, which avoids overfitting caused by inputting all samples into the model. In the sampling process, samples haven’t been taken is called out-of-bag (OOB) data, and it can assess the independent variables' impact on the dependent variables and the model performance because OOB data does not appear in model training. Sampling amount $n$ (which the number of decision trees) and feature split nodes $m$ determine the prediction ability of the random forest model, the number of decision trees can be determined through experiments on its relationship with OOB error, and the number of feature split nodes is selected as $M/3$ according to the recommended value [19].

2.2.2. Model training and prediction. Each decision tree is trained with independent and identically distributed training samples, and the final prediction result of RF is determined based on the prediction results of all decision trees. RF does not need to specifically set cross-validation, and usually uses OOB data samples to optimize the optimal parameters for model testing [20].

2.3. Deformation prediction model of concrete dam based on RS-RF. The traditional concrete dam deformation influencing factor set has the characteristics of redundancy, collinearity, etc., which will not only lead to the reduction of the prediction accuracy of the model but also cause the model to be ill-conditioned or wrong results. Therefore, in view of the shortcomings of the RF algorithm that does not distinguish the sample attributes and the quantitative interpretation of the model, RS theory with the advantage of the reduction of property relations and importance evaluation...
is applied to establish concrete dam deformation prediction model based on RS-RF. The specific modeling steps are as follows:

2.3.1. Data Preprocessing. The statistical method is used to process the gross error of the concrete dam monitoring data to provide reliable data for the establishment of prediction models. RS theory is used to reduce the attributes of water level, air temperature, seepage, cracks, and lift pressure to accurately obtain the representative factors that affect the deformation of the concrete dam. Samples of representative impact factors and their derived variables, deformation monitoring data corresponding to measuring point were used as model data sets.

2.3.2. Model training. Taking the pre-processed standardized training set samples as the model input, the error backpropagation through the gradient descent algorithm drives the model's loss function to converge, and the optimal parameters of the model are obtained.

2.3.3. Model predictions. The independent factors of the test set is input to the prediction model with optimal parameters to obtain the corresponding deformation prediction results.

2.3.4. Model Performance Evaluation. To accurately measure the predictive performance of the model, considering the statistical multiple regression theory and the overlap of the evaluation indicators, root mean squared error (RMSE) and mean absolute percentage error (MAPE) are used as the evaluation indicators of prediction models.

\[
\text{RMSE}_p = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

\[
\text{MAPE}_p = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]

3. Project examples

3.1. Introduction to the project

The Shuikou Hydropower Station is located in the middle reaches of the mainstream of the Minjiang River in Fujian Province. It consists of a concrete dam, a hydropower house, a ship lock, and other buildings. The height of the concrete gravity dam is 74 m, and the length of the dam crest is 783 m. There is a total of 42 dam blocks. Deformation monitoring of the dam included the horizontal displacement, deflection, vertical displacement, crack opening, and other monitoring items. The horizontal displacement was monitored via tension wire alignment. Figure 1 is the plan layout of the measuring points of the tension wire of the concrete dam.

![Figure 1. Plan layout of the measuring points of the tension wire of the concrete dam](image)

3.2. Selection and Optimization of Influential Factors in the Prediction Model.

According to theoretical knowledge, monitoring data, expert experience, etc., the initial selection of empirical influence factors is as follows:
where $H$ is the water depth on a day when observations are collected, $H_{0}$ is water depth on the base day; $T_{i}$ is the mean reservoir region temperature $i$ days ago, and $T$ is the annual mean temperature. Additionally, $\theta=(t-t_{0})/100$, where $t$ is the observation date and $t_{0}$ is the date of the base day.

$J$ is the average fracture aperture at measurement points, $Q$ is the seepage flow, and $U$ is the average uplift pressure at measurement points.

The initial empirical influential factors are selected as the conditional attributes $X$, and the horizontal displacements obtained by the dam crest extension wire at point EX1 is set as the decision attributes $Y$ in the prediction model. Samples of horizontal displacement and influential factors were selected as the sample set $U$. The attribute range $V$ was determined based on the K-means clustering algorithm with adaptive discretization. To eliminate irrelevant or weakly informative input variables and keep only the representative factors that affect concrete dam deformation, RS theory is used to conduct an attribute reduction and importance evaluation and obtain an initial information table $S=\{U, X \cup Y, V, f\}$. The attribute reduction and importance evaluation results for the single-point and multipoint prediction models are shown in Table 1.

| Experience impact factors | Component name | SIM | Reduction | Importance evaluation |
|---------------------------|----------------|-----|-----------|-----------------------|
| $H-H_{0}$                 | Water pressure | -4  | No        | 0.14                  |
| $(H-H_{0})^{2}$           |                | -2  | No        | 0.10                  |
| $(H-H_{0})^{3}$           |                | -4  | No        | 0.05                  |
| $(T_{5}-T)$               | Temperature    | -7  | No        | 0.17                  |
| $(T_{20}-T)$              |                | -2  | No        | 0.34                  |
| $(T_{60}-T)$              |                | 0   | Yes       | 0.00                  |
| $(T_{90}-T)$              |                | 0   | Yes       | 0.00                  |
| $\theta$                 | Aging          | -2  | No        | 0.04                  |
| $\ln(1+\theta)$          |                | -3  | No        | 0.05                  |
| $J$                       | Fracture       | -2  | No        | 0.03                  |
| $Q$                       | Seepage        | -4  | No        | 0.05                  |
| $U$                       | Uplift pressure| -3  | No        | 0.03                  |

According to the attribute reduction and importance evaluation results: The importance evaluation values for each component are: 0.14, 0.10, 0.05, 0.17, 0.31, 0.00, 0.00, 0.04, 0.05, 0.03, 0.05, 0.03. Therefore, the horizontal displacement of the extension wire is greatly affected by changes in water level and temperature. Taking into account of $(T_{20}-T)$ on horizontal displacement at EX1 measurement point, so the lag period of the water level is approximately 20 days.

3.3. Sample Selection for Prediction Models.

Pre-process the environmental monitoring data (primary error removal and standardized processing) of environmental data such as original water level, air temperature, seepage, cracks, and lift pressure. According to the optimization of influential factors result, the dataset is established between June 2, 2016 and October 22, 2018 and has a total of 864 samples of data. The dataset of 700 samples selected from June 2, 2016 to May 10, 2018 is used as the training set, and the dataset of 164 samples selected from May 11, 2018 to October 22, 2018 is adopted as the testing set. Investigations of the concrete dam deformation prediction model based on the SVM, RF, and RS-RF methods are performed using the dataset with 864 samples of data.
3.4. Prediction model training and prediction

3.4.1. Model parameter setting. The number of decision trees in the RF model is the most important parameter, and its selection results play a key role in the model’s ability to fit and predict. To obtain the optimal model parameters, the number of decision trees is set to the range of 1~25, and the OOB errors for the random forest model under the parameters of each number of decision trees is calculated to get the optimal parameter, respectively. The relation curve between OOB error and the number of decision trees is shown in Figure 2. Based on the criterion that the model has the smallest prediction error under the optimal parameters, the number of decision trees under this data set is determined to be 20.

![Figure 2. The relation curve between OOB error and the number of decision trees](image)

3.4.2. Model predictive analysis. Based on pre-processed standardized monitoring data, concrete dam deformation prediction models based on SVM, RF, and RS-RF were established. Figure 3 shows the measured and predicted value process lines for the concrete dam deformation prediction models.

![Figure 3. Process line of the predicted and measured values of each model](image)

According to the analysis of Figure 3, it can be known that the deformation prediction model of the concrete dam based on RS-RF has the highest curve fit and the smallest target loss function, and the model training results are better. At the same time, its prediction performance is also significantly better than the deformation prediction model of the concrete dam based on SVM and RF. It also shows that RS-RF theory can accurately capture the impact factors affecting the horizontal displacement of EX1 measuring point, and it can mine internal features information of monitoring data to authenticity reflect the state structure.

3.4.3. Model evaluation. In order to evaluate the performance of the deformation prediction model of concrete dam based on RS-RF, the residual plots, root mean square error (RMSE) and mean absolute percentage error (MAPE) between the predicted values and measured values are used to evaluate prediction model accuracy. According to the analysis results of evaluation indicators of the concrete dam deformation prediction model based on SVM, RF and RS-RF are compared and analyzed. The
horizontal displacement residuals of each model are shown in Figure 4, and the calculation results of RMSE and MAPE of each model are shown in Table 2.

![Figure 4. Horizontal displacement residual plot of each model](image)

Table 2 Prediction model accuracy indicators

| Prediction model | SVM  | RF   | RS-RF |
|------------------|------|------|-------|
| RMSE             | 0.55 | 0.46 | 0.13  |
| MAPE             | 20.17| 19.20| 5.12  |

According to the analysis of Figure 5 and Table 2: SVM, RF Models can effectively improve the accuracy of the prediction model of the concrete dam, but the deformation prediction model concrete dam based on RS-RF has the best prediction result with high accuracy. Compared to on SVM, RF model indicators, concrete dam deformation prediction model based on RS-RF has minimum horizontal displacement residuals, RMSE is less than 0.15, MAPE is less than 10, all the indicators are at the low interval. Therefore, the deformation prediction model of the concrete dam based on RS-RF has better accuracy performance, and the prediction result is closer to the real data.

4. Conclusions and Discussion

Based on the RS theory, the improved RF model is introduced into the field of dam safety monitoring, and a deformation prediction model for concrete dams based on RS-RF is established. The following conclusions can be obtained through engineering examples.

(1) The successful application of machine learning technology in the establishment of dam safety prediction models can effectively improve the performance of the prediction model and predict the development trend of dam safety behavior more accurately.

(2) The deformation prediction model of concrete dams based on RS-RF can extract the core influencing factors that affect the deformation, improve the calculation efficiency, and effectively avoid overfitting. It has higher prediction accuracy and stronger extension, and these good properties make concrete dam deformation safety early warning more sensitive.

(3) With the development of large concrete dams’ construction, artificial intelligence, online dynamic learning, and the spatiotemporal prediction model is combined to establish a comprehensive real-time monitoring system for concrete dam deformation. The intelligent monitoring will be a concrete dam inevitable trend of the development of deformation safety monitoring.
References
[1] Xu, Z.K., Wang, F., Wei, B.W., et al. (2016) Research progress on the performance degradation mechanism of concrete dams under multi-factor synergy [J]. Advances in Science and Technology of Water Resources and Hydropower, 36 (04): 80-88.
[2] Gu, C.S., Su, H.Z., Wang, S.W. (2016) Research progress on calculation models and monitoring methods for long-term deformation characteristics of high concrete dams [J]. Journal of Hydroelectric Engineering, 35 (05): 1-14.
[3] Zhong, L. (2005) Research on RCC dam safety monitoring model and its application [D]. Hohai University.
[4] Lu, J., Dai, W.M., Zhang, Z.T. (2015) Regression Modeling of Dam Coefficient with Variation Coefficient [J]. Wuhan University Journal (Information Science Edition), 40 (01): 139-142.
[5] Qiu, L.T., Shen, Z.Z., Nie, B.S. (2014) Prediction of dam deformation based on stepwise regression analysis—Markov chain model [J]. Hydropower Energy Science, 32 (05): 51-55.
[6] Lv, K.Y. (2012) Research and analysis of dam deformation prediction method for the Xiaolangdi Water Control Project on the Yellow River [D]. China University of Mining and Technology (Beijing).
[7] Sun, W.W., Li, L., Zheng, H.Y. (2012) Comprehensive Evaluation of Dam Risk Consequences Based on Grey Correlation Degree Method [J]. People's Yellow River, 34 (04): 102-104.
[8] Li, G.C., Dai, W.Z., Yang, G.X., et al. (2015) Application of spatiotemporal autoregressive model in dam deformation analysis [J]. Wuhan University Journal (Information Science Edition), 40 (07): 877-881.
[9] Qu, X.D., Yang, J., Chang, M. (2019). A Deep Learning Model for Concrete Dam Deformation Prediction Based on RS-LSTM. Journal of Sensors.
[10] Gu, C.S, Wang, Y.C., Peng, Y., et al. (2011) Ill-conditioned problems of dam safety monitoring models and their treatment methods [J]. SCIENCE CHINA: Technical Sciences, 41 (12): 1574-1579.
[11] Gao, L. (1993) Study on the deterministic model of Liujiaxia main dam displacement [J]. Dam and Safety, (04): 51-55 + 70.
[12] Yu, H., Wang, G.Z., Yao, Y.Y. (2015) The Research Status and Prospect of Decision Rough Set Theory [J]. Chinese Journal of Computers, 38 (08): 1628-1639.
[13] Ye, H.C., Zhang, S.W., Huang, Y.F., et al. (2014) Application of rough set theory to the determination of soil fertility evaluation index weights [J]. Chinese Agricultural Science, 47 (04): 710-717.
[14] Zhang, M. (2012) Research on knowledge acquisition and reduction methods in rough set theory [D]. Nanjing University of Science and Technology.
[15] Zhang, L.W., Zhang, D.Y., Li, S.C., et al. (2012) Application of genetic-RBF neural network based on rough set theory to the prediction of rock burst [J]. Rock and Soil Mechanics, 33 (S1): 270-276.
[16] Liu, L., Liu, W.J., Liao, W.W., et al. (2018) Comparison of hotspot crime predictions in different periods based on random forest and space-time kernel density methods [J]. Progress in Geographical Sciences, 37 (06): 761-771.
[17] Kang, F., Liu, J., Li, J., et al., Concrete dam deformation prediction model for health monitoring based on extreme learning machine, Structural Control and Health Monitoring, vol. e1997, 2017.
[18] Zhou, ZH. (2017) Research on wind farm power prediction based on machine learning [D]. Nanjing Normal University.
[19] Wang, P.X., Qi, X., Li, L., et al. (2019) Estimation of corn yield based on random forest regression [J]. Journal of the Chinese Society of Agricultural Machinery, 50 (07): 237-245.
[20] Chen, D.L., Lu, X.H., Kuang, B. (2019) Construction and Application of Arable Land Use Efficiency Measurement Model Based on Random Forest [J]. Journal of Natural Resources, 34 (06): 1331-1344.