Abnormal Diagnosis of Dam Safety Monitoring Data Based on Ensemble Learning

Abnormal Diagnosis of Dam Safety Monitoring Data Based on Ensemble Learning

Zhang Jun[1,2*], Xie Jiemin[1,2], Kou Pangao[1]

[1] State Grid Hunan Electric Power Company Limited Research Institute, Changsha 410007, China

[2] Hunan Xiangdian Test & Research Institute, Changsha 410007, China

[*]Corresponding author's e-mail: sghnzj1988@163.com

Abstract: Screening out the gross errors and systematic errors of dam safety monitoring data by theoretical hypothesis will lead to the risk of misjudgment of abnormal data. In order to reduce this risk, based on the ensemble learning method in machine learning, this article extracts and integrates multiple base learners from the stepwise regression model, and proposes a matrix of abnormal indexes based on real-time data update, and analyzes the abnormal diagnosis of the measured data subsequently. The results show that the abnormal indexes have a strong practicability, which don't need to screen out the data with systematic errors and gross errors, and can effectively identify the abnormal time points and the degree of interference between the measured values.

1. Introduction

Abnormal diagnosis of dam safety monitoring data includes a series of quantitative and qualitative analysis methods [1], among which the quantitative analysis often adopts the random error theory of mathematical model to judge the data anomaly. The mathematical model has higher requirements for its rationality and the accuracy of monitoring data. The rationality of the former lies in the selection of fitting methods, while the latter mainly depends on the identification and processing of systematic errors and gross errors.

The identification and treatment of systematic errors and gross errors are always a hot issue in dam safety monitoring [2]. For dam safety monitoring data, systematic errors and gross errors are the important basis to reflect whether the dam has abnormal signs. In the past, step function is often used to fit the known systematic error and gross error in the mathematical model of dam safety monitoring [3], while the unknown systematic error and gross error are more dependent on the experience of researchers and some theoretical assumptions. In addition, in the current research on the abnormal analysis and diagnosis of dam safety monitoring data, only the abnormal analysis and diagnosis focus on the measured values, and there is no research on the mutual interference between measured values. Therefore, this paper proposes a new method of anomaly diagnosis and analysis based on homogeneous ensemble learning method, and presents an index matrix of abnormal value to identify the degree of interference between measured values and reduce the risk of abnormal value discrimination error.

2. Homogeneous ensemble learning

The mathematical model of dam safety monitoring is often designated as a certain mathematical model based on the engineering experience of researchers, which is generally considered as a learner with strong performance. Ensemble learning, on the other hand, complete the learning task through the combination of components and multiple learners, and can generally obtain significantly superior generalization performance than a single learner [4, 5].

There have been some studies on the ensemble learning of dam safety monitoring data, such as [6-9]. Considering the nonlinear models such as neural network algorithm and support vector machine (SVM) algorithm are difficult to intuitively reflect the influence of system errors and gross errors as regression models. This paper take the stepwise regression model of the commonly used as a base learning algorithm consequently, and by adding a step function of system errors and gross errors in the form of effective recognition and processing, to reduce the risk of abnormal value discriminant error.

The advantages of ensemble learning are as follows [10] :(1) it reduces the risk of poor generalization performance caused by misjudgment of individual learners; (2) it reduces the risk of falling into a bad local minimum; (3) it improves the range of hypothesis space, which is possible to get a better approximation. Therefore, the mathematical model of dam safety monitoring adopts the method of individual learner combination, which can further reduce the risk of outlier error. Since it is difficult to determine whether there are anomalies in dam safety monitoring data, and the step functions represent its anomalies, it is not appropriate to improve the performance of individual learners without considering feature selection. In ensemble learning, the individual learner should meet the requirement of "good but different"[5], and satisfy the good in "good but different" by taking advantage of the fact that each factor in the stepwise regression algorithm satisfies the significance hypothesis test, and at the same time, all combination models of the significant factors in stepwise regression are selected. Therefore, the differences in "good but different" are mainly reflected in the different processing of systematic errors and gross errors and the different characteristics that lead to the different complexity of consistent hypothesis. In addition, it can also meet the requirement of as many individual classifiers as possible according to the Hoeffding inequation when the error rates of individual learners are assumed to be independent of each other.

The flowchart of homogeneous ensemble learning finally adopted is as follows:

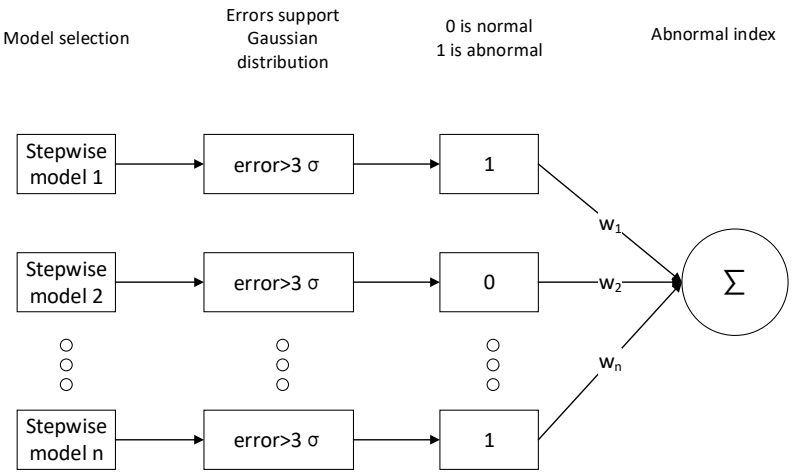


Figure 1. Flow chart of homogeneous ensemble learning

3. Non-aging model and abnormal indexes of measured values for dam safety monitoring

3.1. Non-aging mathematical monitoring model

The previous monitoring statistical models mainly consider that the dam displacement is composed of water pressure component, temperature component and aging component, and the general expression of the models is:

$$\delta(t)=\delta_H(t)+\delta_T(t)+\delta_\theta(t) \tag{1}$$

where $\delta(t),\delta_H(t),\delta_T(t),\delta_\theta(t)$ are the displacement, hydraulic component, temperature component and aging component at time *t*, respectively.

It is known from practical engineering experience that unstable aging factor is often difficult to predict, while stable aging factor can be completely replaced by step function. The difference is that the aging takes into account the continuous change of the sample, and the step function does not force this continuity. At the same time, the step function can be used to identify and deal with unknown systematic errors and gross errors.

The non-aging mathematical model is recommended as follows:

$$\delta(t_l)=\delta_H(t_l)+\delta_T(t_l)+\sum_{i=1}^{l}f(t_i) \tag{2}$$

where *l* is the total number of model data, t_i is the time point of the *i*-th data, $f(t_i)$ is the step function at time t_i which represents the systematic error and gross error.

3.2. Abnormal index

As shown in figure 1, the weighted average method is used for the combined strategy of abnormal indexes as follows:

$$\mathbf{H}_{1\times n}=[\mathbf{w}^1_{1\times m_1}\cdot\mathbf{h}^1_{1\times m_1},\mathbf{w}^2_{1\times m_2}\cdot\mathbf{h}^2_{1\times m_2},...,\mathbf{w}^n_{1\times m_n}\cdot\mathbf{h}^n_{1\times m_n}] \tag{3}$$

where $\mathbf{H}_{1\times n}$ is the monitoring data abnormal index matrix of *n* predictive samples, and the subscript is the number of rows and columns of the matrix; $\mathbf{w}^n_{1\times m_n}$ is the weight vector of the *n*-th prediction sample in m_n models; $\mathbf{h}^n_{1\times m_n}$ is the abnormal index vector of the *n*-th prediction sample in m_n models. It is assumed that 0 represents normal and 1 represents abnormal.

The weight can be selected according to the idea of literature [11]. The difference is that the loss function is the mean square error, and the weight is allocated by taking the minimum mean square error of the calculated value and predicted data of the model as the optimization objective. However, in the actual process, it is often difficult to determine whether there are systematic errors and gross errors in the prediction data themselves. The better the model fitting effect is, the worse the prediction effect may be, and there may be overfitting, etc. Therefore, it is still suggested to adopt the method of equal processing.

Due to the time sequence, it is difficult to judge the interference degree of the measured value. Therefore, it is suggested to update the mathematical model step by step, that is, to add the latest measured value constantly, to calculate the abnormal index of measured value as follows:

$$\begin{cases} a_{ij}=h_{ij}^{i-1}, i\le j \\ a_{ij}\ is\ null, i>j \end{cases}, i,j=1,2,...,n \tag{4}$$

where a_{ij} is the element of total matrix of abnormal indexes A_{ij}; h_{ij}^{i-1} is the element of the exception metrics matrix after the *i*-1 th update, when the predictive sample is reduced to *n*-*i*+1; for the predictive sample with serial number *i*, the effective anomaly indexes are the first *i* row of the *i*-th column of the total matrix element.

According to formula (4), the initial abnormal index of the predicting sample is located in the first row of the total matrix of abnormal indexes, while the final abnormal index is located on the diagonal of the matrix, and the triangle area data under the diagonal are null. With the update of calculation steps, the abnormal index of individual sample also fluctuates, which can reflect the process that the

sample is affected by the previous measurement to some extent. For samples with large abnormal indexes, the entire updating process should be analyzed for comprehensive judgment. The flow chart of specific abnormal determination is as follows:

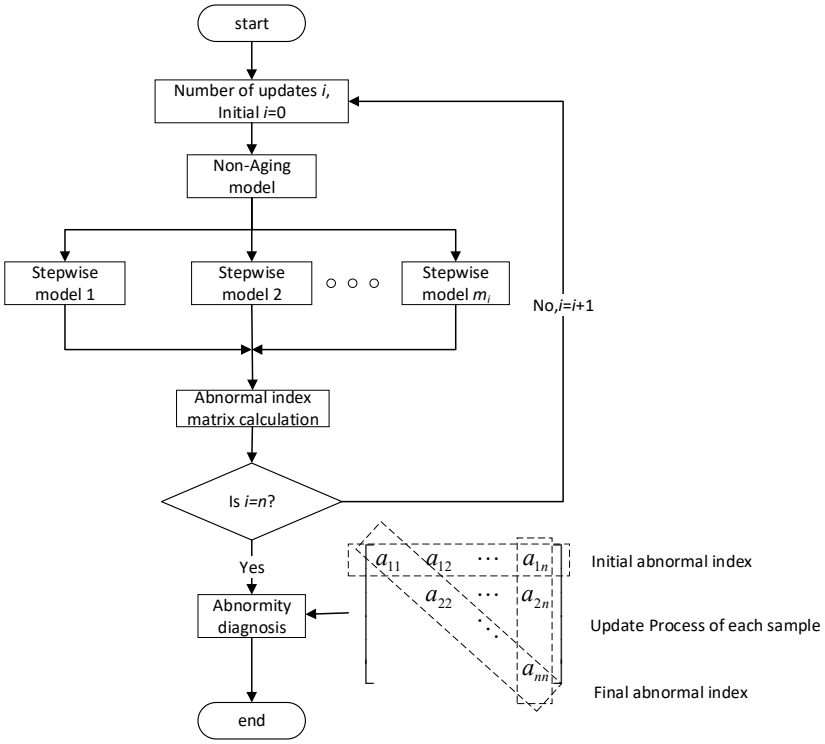


Figure 2. Flow chart of abnormal diagnosis and analysis

4. Example analysis of displacement monitoring data

Taking Dongjiang hydropower plant arch dam as an example, the displacement mathematical model is adopted as follows:

$$\delta(t_n)=a_0+\sum_{i=1}^{4}a_iH^i+\sum_{j=1}^{7}a_jT_j+\sum_{k=1}^{n}a_kf(t_k) \tag{5}$$

where H is the water head, and the first 4 degree polynomial is taken; T is the temperature, taking the average temperature of the current day, 1-7 days, 8-15 days, 16-30 days, 31-60 days, 61-90 days and 91-120 days before the current day; f is the unit step function; a is the coefficient; n is the sample number of mathematical model.

Taking 291m elevation radial horizontal displacement (no. L1-291R) as an example, a total of 96 samples from January 2010 to January 2014 were selected as the initial model samples, and 59 samples from January 2014 to June 2016 were taken as the initial prediction samples for analysis. See figure 3 for the process diagram. The initial and final abnormal indexes are shown in figure 4.

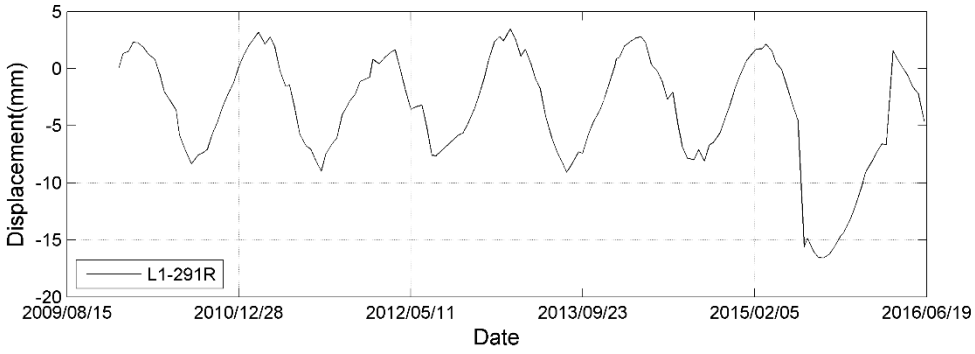


Figure 3. Process diagram of L1-291R

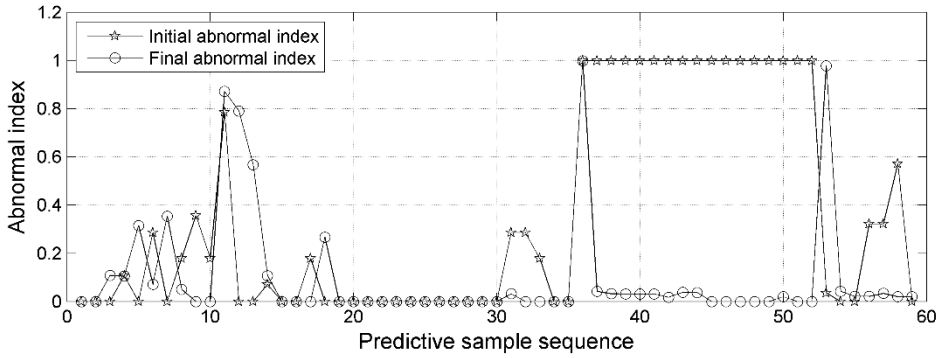
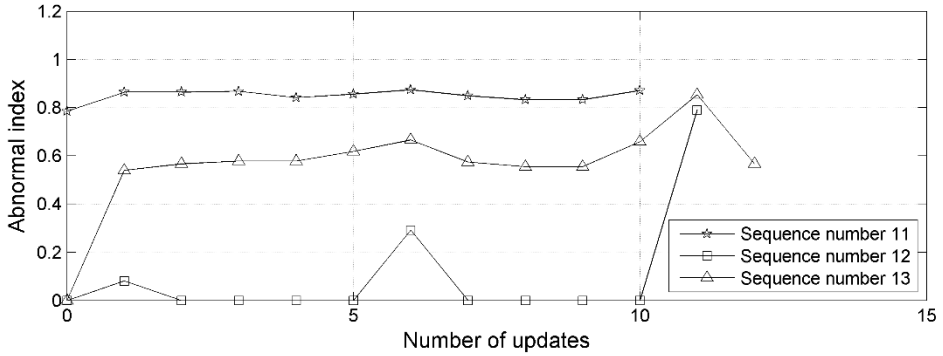


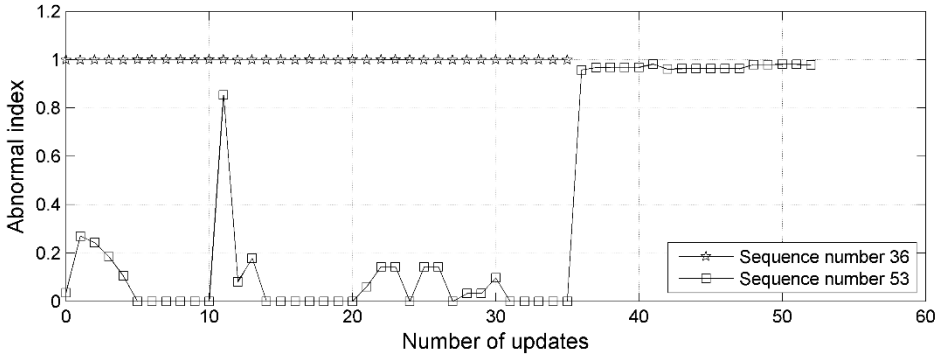
Figure 4. Comparison of abnormal indexes at time points of predictive samples

As can be seen from figure 4, there is a big difference between the measured initial abnormal index and the final abnormal index, especially after the 36th measured value (2015/6/29), the difference between the initial abnormal index and the final abnormal index of 37-53 is close to 1. According to the final abnormal index, it can be determined that there is obvious jump in the measured values (2016/3/14) 36 and 53. The field investigation found two reasons for the jump in the measured values: the former is due to the vertical line transformation and steel wire replacement, while the latter is to connect and adjust the data. Therefore, the final abnormal index reflects the abnormal of the measured time point.

The serial number of measured values with large abnormal indexes in figure 3 is selected for analysis, as shown in figure 5.



(a)



(b)

Figure 5. The update process of abnormal indexes with different measured values

From figure 5a, no. 11 measure value abnormal index is large and the change is relatively stable, no. 12 measure value abnormal index is most affected by update times of 11 (corresponding to 11 measurement). However, most of the abnormal indexes are low before the update of the no.11 value, so it can be considered that the abnormal degree of the no.12 value is not high, but is affected by the no.11 value with higher abnormal indexes. No. 13 measure value abnormal index is most affected by update times of 1 (corresponding to 1 measurement). As can be seen from the total matrix of abnormal indexes, the abnormal index of no. 1 measured value (2014/1/8) is 0. It can be seen that even the data considered normal will have an impact on the abnormal determination of the later measured value, so the influence between measured values cannot be ignored.

As can be seen from figure 5b, the measured value 36 is stable at 1, no. 53 measure value abnormal index is most affected by update times of 36 (corresponding to 36 measurement). Most of the abnormal indexes are low before the update of no. 36, and the no.11 measurement update had a certain impact on them. However, with the increase of update times, the abnormal index returns to a lower value, indicating that the degree of abnormal value is not high. Therefore, the final abnormal index is more expressed as the abnormal value at the time point of measurement rather than the abnormal value of measurement. This analysis is consistent with the initial abnormal index and the data from this time point (2016/3/14).

5. Conclusion

(1) By using the feature that multiple individual learners participate in decision making in ensemble learning, the regression model fully considers the possibility of systematic errors and gross errors of samples and the diversity of feature dimension selection, so as to reduce the risk of misjudgment of outliers caused by irrational screening of samples and mathematical models.

(2) An overall matrix of measured abnormal indexes is proposed to reasonably reflect the abnormal degree of dam monitoring data and the interference degree between measured values in a quantitative form. Case analysis of the diagonal element of the overall matrix, i.e. the final abnormal index, effectively reflects the actual abnormal time point of the measured value. The updating process of the row element of a column of the overall matrix, i.e. the specified sample abnormal index, can provide abundant information for the determination of abnormal value, thus reducing the risk of misjudgment of abnormal value.

(3) The weight allocation strategy needs to be further analyzed, including the reinforcement of the individual learner that can identify the known systematic errors and gross errors, so as to further optimize the performance of ensemble learning.

References

[1] DL/T 5209-2018, Data compilation code for concrete dam safety monitoring.
[2] Wu, J. (2009) Dam safety monitoring theory and test technology. China water resources and hydropower press, Beijing.

[3] He, J. (2010) Theory and application of dam safety monitoring. China water resources and hydropower press, Beijing.
[4] Zhou, Z. (2016) Machine learning. Tsinghua university press, Beijing.
[5] Zhou, Z. (2012) Ensemble Methods: Foundations and Algorithms. Chapman &Hall/CRC, Boca Raton, FL.
[6] Wang, C., Hu, T., Gu, Y. et al. (2018) ARIMA-BP combined forecasting model in dam safety monitoring. Journal of China Three Gorges University (Natural Sciences), 40:20-24.
[7] Huang, M., Yang, H. (2018) SVM**-**ARIMA Dam Safety Monitoring Model Based on Particle Swarm Optimization. Yellow River, 40:149-156.
[8] Du, H., Zhao, E., Guo, S. et al. (2018) Safety Monitoring Model of Dam Service Based on Genetic Algorithm and RBF Neural Network. Journal of China Three Gorges University (Natural Sciences), 40:11-14.
[9] Qian, C., Li, L., Zhou, Z. (2018) Dam Deformation Early Warning Model Based on ABCA-SVM Model. Yellow River, 40:124-127.
[10] Dietterich, T .G. (2000) Ensemble methods in machine learning. In Proceedings of the 1st International Workshop on Multiple Classifier Systems(MCS), Cagliari, Italy. pp. 1-15.
[11] Shen, J., Fang, B., Zheng, D. et al. (2018) Dam Deformation Monitoring by Radial Basis Function Model Optimized by Particle Swarm Optimization with Inertia Weight and AdaBoost. Journal of Yangtze River Scientific Research Institute, 35:57-62.