Hindawi
Scientific Programming
Volume 2018, Article ID 3287065, 7 pages
https://doi.org/10.1155/2018/3287065

*Research Article*

# Research on the Prewarning Method for the Safety of South-to-North Water Transfer Project Driven by Monitoring Data

**Yang Liu, Yaoling Fan, Xinqing Yan, Xuemei Liu, and Bin Yang**

*School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450011, China*

Correspondence should be addressed to Yang Liu; yangliu@ncwu.edu.cn

Received 8 November 2017; Accepted 4 March 2018; Published 4 June 2018

Academic Editor: Shangguang Wang



In order to solve the prewarning problem of South-to-North Water Transfer Project safety, an intelligent cooperative prewarning method based on machine learning was proposed under the framework of intelligent information processing. Driven by the monitoring data of the South-to-North Water Transfer Project, the single sensor in typical scenes was studied, and the security threshold was predicted along the vertical axis of time, firstly. With the support of the data correlation calculation, the sensors in the typical scene were intelligently grouped, and the study objectives were changed into sensor grouping, secondly. Then, the nonlinear regression model between the single sensor and the multisensors was built on the time cross section, and the model was used to dynamically calculate the safety threshold of the current sensor for the second time. Finally, in the framework of intelligent information processing, a double verification mechanism was proposed to support the construction of the intelligent prewarning method for the safety of South-to-North Water Transfer Project. The paper collected the monitoring data from November 2015 to September 2016 in the typical scenarios. The experimental results showed that the methods constructed in the paper can be able to identify the abnormal causes of data sudden jump effectively and give the different level prewarning. The method provides a strong theoretical support for further manual investigation work.

## 1. Research Background

The middle route of the South-to-North Water Transfer Project diverted water from the Dan Jiang Kou reservoir to Henan Province, Hebei Province, Tianjin, and Beijing. The project length was 1432 km. The complexity of the geological and meteorological conditions along the project made the South-to-North Water Transfer Project face serious challenges. The engineering safety referred to the safety problems of the middle route of the South-to-North Water Transfer Project mainly including the buildings, channels, and the important engineering facilities. In the actual monitoring process, the number of sensors was very large; at the same time, a lot of sensors were often installed in the bottom of the canal or were embedded in the channel projects, so the routine maintenance and maintenance work for these sensors was very difficult to implement. When the monitoring data was abnormal, we cannot judge whether the data anomaly was caused by sensors or by the channel engineering failure; thus, the staff could not grasp the overall situation of the channel security. In view of this problem, under the driving of the safety monitoring data of the South-to-North Water Transfer Project, the paper regarded a single sensor as research object and predicted the safety threshold along the time axis by Kalman filter method based on its historical monitoring data, firstly; then, the paper expanded the research object from single sensor to sensor grouping by using the data correlation analysis methods, which can reduce the computational complexity and improve the accuracy of prediction algorithm. Secondly, the nonlinear regression model between single sensor and sensor grouping was built to predict and check the sensor's data on the time cross section. Finally, the intelligent prewarning method was constructed under the framework of intelligent information processing, which provided scientific theoretical support and effective decision-making for the emergency troubleshooting and emergency countermeasures.

The first part of paper introduced the research background. The second part introduced the basic principle of several typical machine learning methods. The third part introduced the data prediction based on machine learning methods. The fourth part introduced the basic principle and processing flow of intelligent prewarning method. In the fifth part, the algorithm was validated and the results were analyzed.

## 2. Principles of Machine Learning Methods

*2.1. The Basic Principles of Radom Forest (RF).* The Random Forest (RF) algorithm was a kind of machine learning model proposed by Leo Breiman in 2001 [1, 2]. The RF method generated a lot of classification trees by randomly using attributes (columns) and data (rows) of the sample set and finally summarizes these classification trees to form the final Random Forest. Each tree in a RF is a binary tree, and its generation followed the top-down recursive splitting principle; in other words, the training set is divided from the root node in turn. In the binary tree, the root node contained all the training data and was divided into left node and right node which contained a subset of training data, in accordance with the principle of minimum node purity. Then, the left and right nodes continued to be split according to the same rule, until the stop rule was satisfied. If the classification data on one node $n$ all came from the same class, then the purity of this node was $I(n) = 0$. The specific implementation process of RF was as follows:

(i) The original training set is $N$, RF method randomly extracted $k$ new samples and constructed $k$ classification trees by using the bootstrap method; each time the samples which were not extracted would form the out of pocket data set.

(ii) Suppose there were $m_{\text{all}}$ variables. $m_{\text{try}}$ variables were randomly extracted at each node of the classification tree, and the most powerful variable is selected in the $m_{\text{try}}$. The threshold of variable classification is determined by checking each classification point.

(iii) Each tree would grow for the maximum without any pruning.

(iv) Finally, Random Forest was composed of the decision trees which were produced by (a), and the new data was discriminated and classified by the Random Forest classifier. The classification result depended on the number of votes of the tree classifier.

*2.2. The Basic Principles of Ada-Boost.* Ada-Boost was an acronym for "Adaptive Boosting" in English presented by Robert Schapire in 1995 [3, 4]. Its adaptation was that the sample which was classified inaccurately by the previous basic classifier would be strengthened. Then, the whole sample would be used again to train the next basic classifier. At the same time, a new weak classifier would be added to the classifier set each circle until a predetermined small error rate was reached or the maximum number of iterations specified in advance was reached.

Ada-Boost algorithm implementation process was as follows:

(1) Initialize the weight of each training sample; the number of the total training examples was $N$. The calculation of initialization was as formula (1):

$$W = (w_{11}, w_{12}, \ldots, w_{1N}), \quad w_{1i} = \frac{1}{N}, \ i = 1, 2, \ldots, N \tag{1}$$

(2) The algorithm was trained in M circles; the m circle of learning process was as follows:

(a) Use the training samples with weight distribution $W_m$ to get the base classifier $G_m$.

(b) Calculate the error rate of the base classifier obtained in the previous step:

$$\begin{aligned} e_m &= P\left(G_m(x_i) \neq y_i\right) = \frac{\sum_{i=1}^{N} w_{mi} I\left(G_m(x_i) \neq y_i\right)}{\sum_{i=1}^{N} w_{mi}} \\ &= \sum_{i=1}^{N} w_{mi} I\left(G_m(x_i) \neq y_i\right) \quad \sum_{i=1}^{N} w_{mi} = 1 \end{aligned} \tag{2}$$

(c) Calculate the weighting factor in front of $G_m$:

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{3}$$

(d) Updated the weight coefficient of the training sample:

$$\begin{aligned} W_{m+1,i} &= \frac{w_{mi}}{Z_m} \exp\left(-\alpha_m y_i G_m(x_i)\right) \\ Z_m &= \sum_{i=1}^{N} w_{mi} \exp\left(-\alpha_m y_i G_m(x_i)\right) \end{aligned} \tag{4}$$

(e) Repeat (a) to (d) to obtain a series of weight parameters $\alpha_m$ and the base classifier $G_m$.

(3) The base classifier obtained in the previous step would be combined linearly according to the weight parameters to obtain the final classifier:

$$\begin{aligned} f(x) &= \sum_{m=1}^{M} \alpha_m G_m(x) \\ G(x) &= \operatorname{sign}\left(f(x)\right) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_m G_m(x)\right) \end{aligned} \tag{5}$$

There were four machine learning algorithms in the third section of the paper. In addition to the Ada-Boost method and Random Forest method, the Bagging algorithm and the support vector machine (SVR) method were used as the contrast method in the paper. Because of the limitation of space, they were no longer detailed here. The basic principles and applications of the methods could be seen in the literature [5–7].

*2.3. The Basic Principles of SVR.* SVR is a machine learning method which can be used for time series prediction. Through a nonlinear kernel function, the multidimensional input is mapped onto the feature space of higher dimension and then the regression operation is performed to obtain the nonlinear mapping relation with the output index. Due to space limitations, the implementation details of SVR method are shown in the literature [8].

## 3. The Data Prediction Based on the Machine Learning

*3.1. K-Fold Cross-Validation.* Cross-validation was a statistical analysis method to verify the performance of the algorithm. The basic idea of cross-validation is that the raw data was grouped into two subsets in some sense, one of the subset was used as the training set (train set), and the other subset was the validation set (validation set). The training set is used to train the model, and then the validation set is used to test the performance of the model which was obtained by the first step.

$K$-fold cross-validation is one of the most commonly used methods for data validation in cross-validation. The original data was divided into $K$ groups (generally equal); each subset of data was a verification set, and the rest of the $K - 1$ subset data as training set. The process of $K$-fold cross-validation was as follows.

*Step 1.* The whole sample set $S$ was divided into $k$ disjoint subsets, assuming that the number of samples in $S$ was $m$; then each subset has $m/k$ training samples, and the corresponding subset is called $S'\{s_1, s_2, \ldots, s_k\}$.

*Step 2.* Each subset in $S'$ would be picked out as the test set, and the other $k - 1$ as the training set.

*Step 3.* Obtain the model or hypothesis according to the training set.

*Step 4.* The training model was used to classify on the test set, and the accuracy of classification would be calculated.

*Step 5.* The mean value of the correct classification rate calculated by $K$ times was used as the true classification rate of the model or the assumed function.

*3.2. Data Prediction under the 6-Fold Cross-Validation.* When the monitoring data of current sensor was abnormal, the method would calculate the number of the sensors whose monitoring data was abnormal at that moment. The cooperative prewarning algorithm would send out an engineering red warning that mean the project was danger, if the number of sensors in the same group of sensors is higher than 60%. If the ratio was less than 60%, a nonlinear regression model between the current sensor and the residual sensors within the same group would be established, and the model was used to predict the current sensor monitoring data on the time cross section for the second time. The paper used the 6-fold cross-validation method, 16% of the total samples were randomly selected as the test samples, and the remaining 84% samples were used as the training samples. The machine learning algorithms investigated in the paper included Bagging, SVM, Ada-Boost, and Random Forest. Because of the limited space, only the prediction result of sensor R1_4 is shown here. In the prediction process, Bagging algorithm, Ada-Boosting algorithm, and Random Forest method all use the regression trees as basic modes, and the number of regression trees was 50, and the depth is 30. The constant $C = 10^{-3}$, gamma = 1 in SVR algorithm.

Figure 1 showed the prediction result curves of the sensor R1_4 under the four machine learning methods. The overall trend of monitoring data of sensor R1_4 was relatively stable, as we can see that the Ada-Boosting, SVR, and Random Forest methods showed good tracking performance. When the monitoring data had abrupt jump, the prediction curve of Ada-Boost method showed good convergence and can track the data jump in time. At the same time, the prediction result of SVR method showed a large fluctuation, and the prediction curve had obvious deviation. From the data prediction curve, we can see that the SVR algorithm is too sensitive to the data fluctuation. Figure 6 showed the prediction error curve of sensor R1_4 under various methods. From the error curve, we can see that, in the prediction process of sensor R1_4, the SVR method has good prediction accuracy when the data is stationary. When the data fluctuates slightly, the SVR prediction produces a larger prediction deviation. It can be seen from the whole prediction process that the Ada-Boost method has better performance in prediction accuracy and algorithm stability. Figure 2 showed the lines of mean value of the prediction error belonging to every method. We can see that the error line of SVR emerged sudden jump when the monitoring data of the sensor R1_4 has bigger change.

The prediction error of each algorithm was shown in Table 1. It can be seen from the mean value of error that the data prediction accuracy of Ada-Boost algorithm was best under the typical scene, and the worst accuracy was the Bagging algorithm. The statistic results of data prediction were consistent with the curve results; they had the same conclusion. At the same time, the error variance and standard deviation of each algorithm showed that the error variance and standard deviation of Ada-Boost method were the smallest. The Ada-Boost method and Random Forest method had better prediction accuracy and prediction stability in the data prediction. Therefore, they were used to construct the intelligent prewarning method for the South-to-North Water Transfer Project.

## 4. Design of the Intelligent Prewarning Method

This section focuses on how to construct the time-space cooperative intelligent prewarning information processing method based on the use of machine learning algorithms. When the monitoring data was abnormal, the method would find it in time and judge abnormal course whether the abnormal data was caused by the sensor fault or by the engineering itself faults. Then, the method could send out the different levels prewarning message according to the abnormal course.

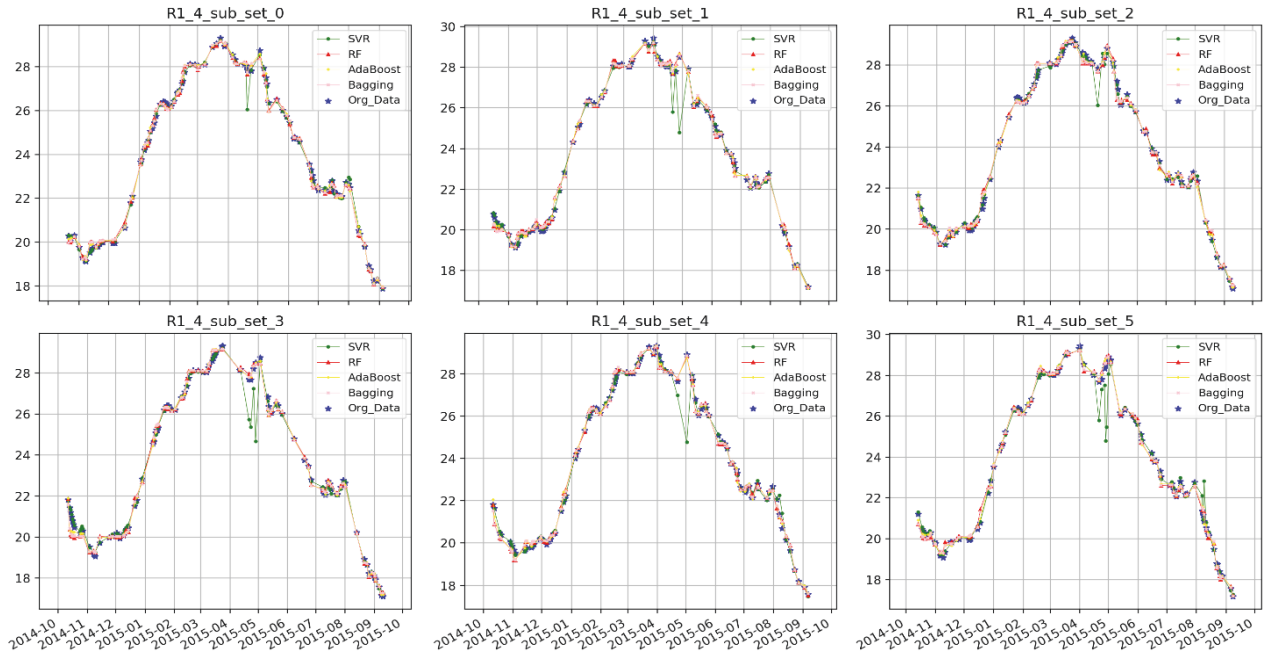


Figure 1: The data prediction curve of sensor R1_4.

Table 1: The error rate of SVM, Forest, Ada-Boost, and Bagging algorithms in the typical scene.

| Algorithm Name | Error Mean | Error Variance | Error (max) | Error (min) |
|---|---|---|---|---|
| SVM | 0.212 | 11.16 | 9.474 | −6.046 |
| Forest | −0.006 | 6.81 | 6.739 | −6.381 |
| Ada-Boost | −0.003 | 1.77 | 4.551 | −4.990 |
| Bagging | 0.626 | 14.02 | 10.198 | −9.229 |

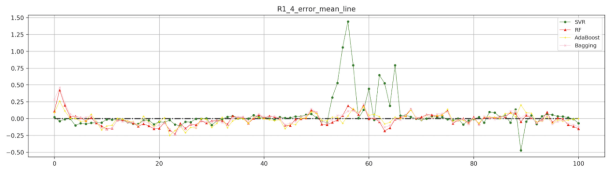


Figure 2: The prediction error curve of sensor R1_4.

The third part of the paper is the basis of the fourth part. The intelligent prewarning method uses the prediction results belonging to the third part; at the same time, it can generate the security range dynamically. This is a standard to judge whether the data is abnormal.

In order to solve the correct prewarning of abnormal data, the paper adopted time-space cooperative verification method based on machine learning methods to forecast and verify the monitoring data of the sensor from time and space dimension, respectively. Generally, the sensor monitoring data changed in linear mode in a relatively short period of time. Here, the paper used the traditional Kalman filter method [8–10] to predict the single-sensor monitoring data on historical monitoring data and, at the same time, generate the security interval of monitoring data based on the prediction results. The whole processing process is shown in Figure 3.

*Step 1.* Based on the Kalman filter, the current monitoring data was predicted by using the historical data of a single sensor in the given time slice. Then the time domain security range was generated by using the predicted data points. If the current monitoring data was in the safe range, we would regard it as a normal data and record it down; otherwise, the method would jump to Step 2.

*Step 2.* Got the sensor group where the current sensor was located.

*Step 3.* Calculate the number of sensors that occur abnormally in the current sensor group. If the abnormal ratio of sensor grouping was greater than the predecision threshold, a high level red prewarning message would be carried out; otherwise, jump to Step 4.

*Step 4.* The nonlinear regression model between the current sensor and the other sensors in the sensor grouping was built based on machine learning methods, and the model was used to predict and check the current monitoring data of the sensor.

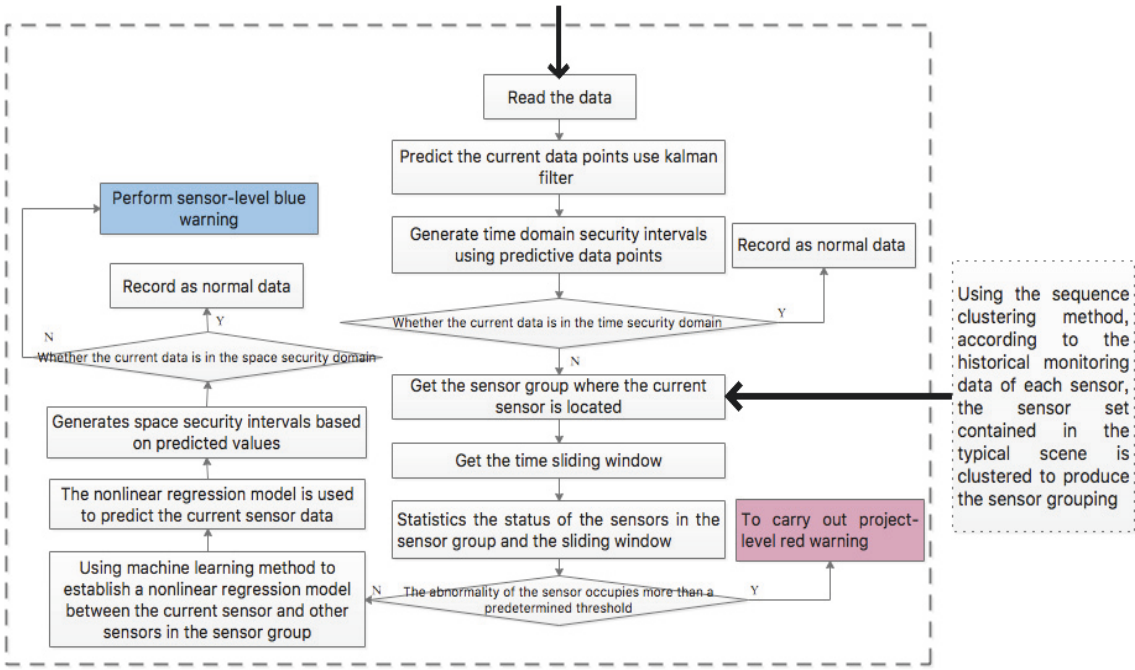


Figure 3: The flowchart of time-space coordination intelligent warning method for South-to-North Water Diversion Project.

*Step 5.* The security range of the data was generated on Step 4 prediction results. If the monitoring value of the current sensor was in the new security range, even if the data exceeds the security range produced by Kalman filter in Step 1, it would be treated as normal data. Otherwise, the sensor level alarm would be send out, which indicted the current sensor was wrong. It was suggested that the monitoring data of this equipment should be ignored; otherwise, the overall judgment of the subsequent safety of the project would be affected.

As introduced in "Step 2", the sensor group was the result based on the Pearson correlation coefficient. Pearson correlation coefficient is a linear correlation coefficient, which is used to reflect the linear correlation of two variables. The correlation coefficient is denoted by r, and it is a value between 1 and −1, where 1 indicates that the variable is completely positive, 0 is independent, and −1 means completely negative correlation. Pearson correlation coefficients are calculated as shown in formula (6).

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}\left((X-u_X)(Y-u_Y)\right)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(X)}{\sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2}\sqrt{\mathrm{E}(Y^2) - \mathrm{E}(Y)^2}} \tag{6}$$

With the sensors R1_4, R1_7, R1_8, R1_16, R1_18, and R1_19, the correlation coefficients between the two sensors were calculated. Table 2 shows the real-time grouping results of sensor sets. In the table, the sensor group with strong correlation with R1_4 contains {R1_7, R1_18, R1_19}, and the order in the sensor group is arranged according to its correlation with R1_4.

Table 2: The grouping results of sensor sets.

| Current Sensor | Sensor Group |
|---|---|
| R1_4 | R1_7, R1_18, R1_19 |
| R1_5 | R1_16, R1_18, R1_7 |
| R1_7 | R1_18, R1_4, R1_19 |
| R1_8 | R1_19, R1_4, R1_18 |
| R1_14 | R1_8, R1_19, R1_4 |
| R1_16 | R1_5, R1_18, R1_7 |
| R1_18 | R1_7, R1_4, R1_19 |
| R1_19 | R1_4, R1_7, R1_18 |

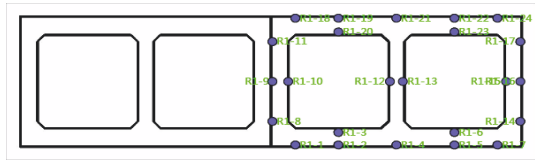

Figure 4: Location of sensors embedded in typical scenes.

## 5. Experiment and Analysis

*5.1. The Description of Typical Scene.* The safety monitoring data of channel is from the South-to-North Water Transfer Project Construction Administration. A typical scene of the South-to-North Water Transfer Project was showed in Figure 4, in which a number of steel bars ($R_{1\_1} \sim R_{1\_22}$) were regularly embedded in the project. Their concrete position

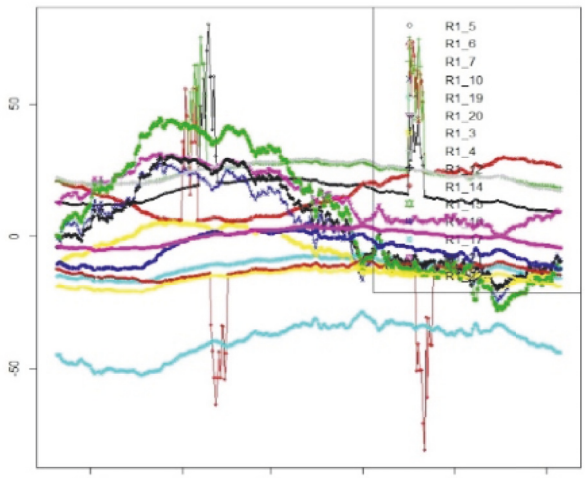


Figure 5: The original data curve of the typical scene.

is shown in Figure 4, and the safety parameters of the canal engineering are monitored in real-time. In the process of research, the monitoring data were collected from October 2014 to October 2015 for about one year. The original monitoring data of the steel bars in the typical scene was showed in Figure 5. Take the monitoring data of R1_4 as an example. There are 335 batches data in all, and the average value of temperature is −13.06 degrees centigrade. The mean value of stress is 23.90 MPa, the variance of stress is 3.51, the minimum value is 17.11 MPa, the maximum value is 29.45 MPa, the median value is 23.76 MPa, and the quartiles value is 20.43 MPa.

The computer is configured as follows: processor is Intel core CPU i5-6500, CPU frequency is 3.20 GHz; memory is 4.00 GB; the operating system is Windows 10 (64-bit); the programming language is Python 3.5.2(64 bits); the integrated development environment is Pycharm Community Edition.

*5.2. Realization of Cooperative Intelligent Prewarning Method.* Figure 6 showed the prewarning results of steel bars such as $R_{1\_4}, R_{1\_5}, R_{1\_7}, R_{1\_8}$. The prewarning results were produced by the intelligent cooperative prewarning method for the safety of South-to-North Water Transfer Project, which was constructed based on the Ada-Boost method and the Random Forest method. In the figure, the horizontal axis was the monitoring time, and the longitudinal axis was the monitoring data of the steel bars from October 2014 to October 2015. At the same time, the prewarning points of the channel engineering safety produced by intelligent cooperative prewarning method also were marked. The dark blue "∗" was the project level prewarning points. Project level warning points mean that the channel project itself may have security risks. It was necessary to organize relevant persons go to the scene immediately for further security investigation. Green "•" was the sensor level warning points based on the Random Forest method, and the yellow "•" was the sensor level warning point based on the Ada-Boost method. These sensor level warning points mean that the data abnormal was caused by the sensor failure and had nothing to do with the channel engineering. In order to know the overall situation of the channel safety accurately in the follow-up work, the data of current sensor could be ignored temporarily, so as to reduce the interference to the results of the intelligent prewarning data processing.

## 6. Conclusion

Along the Middle Route Project of the South-to-North Water Transfer Project, the geological conditions were complicated, the number of sensors was large, and the position was special. Therefore, the regular maintenance and routine maintenance of these sensors were difficult. So, when the monitoring data was abnormal, it was impossible to judge whether the abnormal data was caused by sensor's fault or by the quality of the channel engineering. To solve this problem, the paper jumped out of the traditional theory research of water conservancy engineering and built the intelligent prewarning method, just from the perspective of data research. At the same time, the method was used in the typical scenario of South-to-North Water Transfer Project.

In this method, the data correlation analysis method was applied to realize the dynamic grouping of sensors. Then, a kind of double checking mechanism of abnormal data was constructed, which could check the monitoring data from the axis of time and on the time cross section. Finally, in the framework of intelligent information processing, data analysis and machine learning methods were organically combined to establish a complete data processing process and engineering safety warning mechanism. The experimental results showed that this method was a feasible and effective method for the abnormal data processing in the South-to-North Water Transfer Project.

## Disclosure

This article is original for the first author. It does not contain any conflict with any other article or project (project number: 51509090, project name: The Discovery and Inversion of Emergent Groundwater Contaminant Based on Sequence Mining and Intelligent Computing; project number: 16HASTIT034; project name: Research on Intelligent Computing Method for Large-Scale Space Temporal Sequence Data).

## Conflicts of Interest

The authors declare no conflicts of interest related to this work.

## Acknowledgments

The research of this paper is partly sponsored by the National Science Foundation of China (under Grants nos. 51509090 and U1604152) and Program for Scientific and Technological Innovation Talents of Colleges and Universities in Henan (under Grant no. 16HASTIT034).

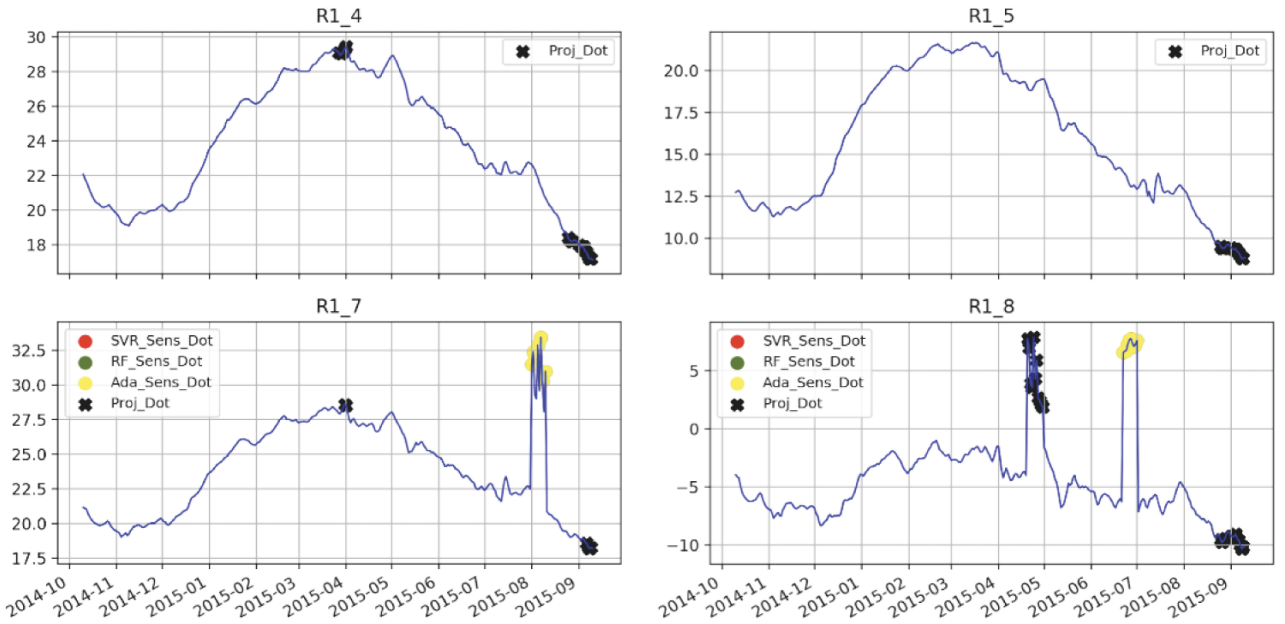


Figure 6: The results of intelligent prewarning method based on the RF and Ada-Boost method.

## References


[1] G. Biau, E. Scornet, J. Welbl et al., "Mondrian forests: efficient online random forests," *Neural Random Forests*, pp. 3140–3148, 2014.

[2] G. Biau, E. Scornet, and J. Welbl, "Neural random forests," *Machine Learning*, 2016.

[3] A. Beygelzimer, S. Kale, and H. Luo, "Optimal and adaptive algorithms for online boosting," in *Proceedings of the 32nd International Conference on Machine Learning (ICML '15)*, pp. 2313–2321, July 2015.

[4] Y. Freund, "An adaptive version of the boost by majority algorithm," *Machine Learning*, vol. 43, no. 3, pp. 293–318, 2001.

[5] Y. Zhong, T. Chu-dong, Y. Jie, and F. Xiao-qin, "Regional PM2.5 concentration prediction method of PSO-SVR model with weighting factors," *Application Research of Computers*, vol. 34, no. 2, pp. 405–408, 2017.

[6] N. H. Afdhal and L. Serfaty, "Effect of registries and cohort studies on HCV treatment," *Gastroenterology*, vol. 151, no. 3, pp. 387–390, 2016.

[7] W. Li-guo, Z. Liang, and L. Ting-ting, "Two improvements for least squares support vector machines," *Journal of Harbin Engineering University*, vol. 36, no. 6, pp. 847–850, 2015.

[8] A. Monfort, J.-P. Renne, and G. Roussellet, "A quadratic Kalman filter," *Journal of Econometrics*, vol. 187, no. 1, pp. 43–56, 2015.

[9] C. Paleologu, J. Benesty, S. Ciochină, and S. L. Grant, "A Kalman filter with individual control factors for echo cancellation," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)*, pp. 5974–5978, May 2014.

[10] H.-M. Kim, D. Ryu, B. K. Mallick, and M. G. Genton, "Mixtures of skewed Kalman filters," *Journal of Multivariate Analysis*, vol. 123, pp. 228–251, 2014.