International Journal of Advanced Smart Convergence Vol.7 No.1 24-32 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.1.4

IJASC 18-1-4

Dam Sensor Outlier Detection using Mixed Prediction Model and Supervised Learning

Chang-Mok Park†

Department of Industrial Management Engineering, INDUK University, Korea
cmpark@induk.ac.kr

Abstract

An outlier detection method using mixed prediction model has been described in this paper. The mixed prediction model consists of time-series model and regression model. The parameter estimation of the prediction model was performed using supervised learning and a genetic algorithm is adopted for a learning method. The experiments were performed in artificial and real data set. The prediction performance is compared with the existing prediction methods using artificial data. Outlier detection is conducted using the real sensor measurements in a dam. The validity of the proposed method was shown in the experiments.

Key words: *Sensor Measurement, Outlier Detection, Supervised Learning, Genetic Algorithm.*

1. Introduction

To monitor the state of the dam, various measuring instruments are embedded and used. Pore water pressure to measure water pressure in the soil, displacement sensor to measure physical movement, and so on. In order to make an appropriate judgment on the state of these dams, it is necessary to detect the sensor error in advance. Such a judgment mathematically predicts the sensor measurement value and compares it with the actual measured value, and if the difference is statistically very large, it can be judged as a sensor error.

Dam sensors measure the changes in the surrounding physical environment and the same sensors are installed in multiple places. The pore water pressure is influenced by the water level of the dam, and the pore water pressure measurements in neighboring sensors shows much more correlation characteristic. Also, natural phenomena have time series characteristics, so the measured values will have the same time series characteristics. In consideration of these characteristics, finding a mathematical model for prediction of sensor signal in a dam is an important tool for fault detection of the dam measurement sensor. In the conventional method, the relationship between the measurement sensors is utilized using a regression model analysis[1][2]. Several studies used a time series prediction model[3]. In some studies, a supervised learning method, such as a neural network, was used without using a mathematical prediction model[4][5][6]. Dam measurements have complex characteristics of time series and linear regression characteristics, so it is difficult to estimate the coefficients of the mathematical model. Alternatively, a prediction model based on supervised learning such as a neural network can be used, but the mixing characteristic of the time series and the regression model could

Manuscript Received: Feb. 3, 2018 / Revised: Feb. 21, 2018 / Accepted: Feb. 23, 2018
Corresponding Author: cmpark@induk.ac.kr
Tel: +82-2-950-7607, Fax: +82-2-950-7619
Department of Industrial Management Engineering, INDUK University, Korea

interfere with each other's learning and lower the prediction performance

In this paper, we will present a prediction method suitable for such a mixed system. Our method consists of a mixed model construction and an optimizing the model coefficients by supervised learning. We are trying to detect an error in the sensor measurement of the actual dam sensors with these predictions. The genetic algorithm is adopted for supervised learning. In order to test the performance of proposed method, we conducted experiment using an artificial sensor data and real sensor measured data of actual dam.

2. Existing Method

The method of predicting sensor measurement values can be summarized by mathematical methods and model-based learning methods. The mathematical method can be divided a time series model approach and a linear regression model approach. A typical model based learning method is a neural network model approach. Since the target system has a mixture of time series and regression characteristics, prediction using one mathematical model may degrade the performance. So two methods can be applied in order. That is, time series prediction is performed first and linear regression prediction is executed by the signal which is removed the time series component from the original signal. Alternatively, it can be applied in reverse. In this paper, to compare with the proposed methodology, three existing methodologies were presented and used for comparative experiments.

2.1 Regression model first and ARIMA second prediction (Regression-ARIMA)

The specific prediction procedure is as follows.

(1) Predict the sensor signal using the linear regression model by setting the neighboring sensors as an independent variable.
(2) Remove the linear model component from the original signal
(3) Predict using the time series model
(4) The final prediction value is obtained by combining predicted values of regression model components and predicted values of time series components.

2.2 Time series model first and regression second prediction (ARIMA- Regression)

The specific prediction procedure is as follows.

(1) Predict the sensor signal using the time series model
(2) Remove the time series model component from the original signal
(3) Predict using the linear regression model by setting the neighboring sensors as an independent variable.
(4) The final prediction value is obtained by combining predicted values of regression model components and predicted values of time series components.

2.3 Neural network prediction

We use a neural network model to predict sensor measurements using neighboring sensors and past measurements as show in Figure 1. The learning of the neural network model use a nonlinear optimization method to improve the weights between the input node, the hidden layer node, the hidden layer node and the output node to predict the sensor measurement accurately.

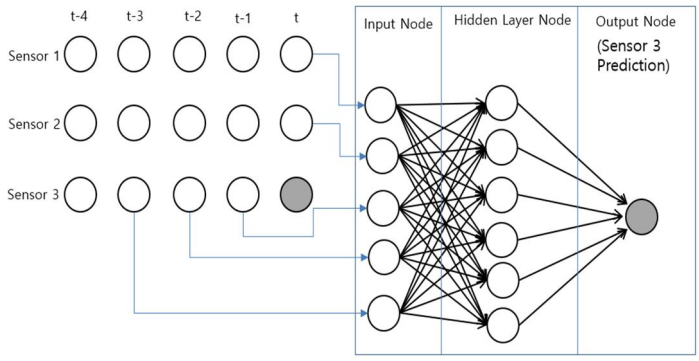

Figure 1. Neural network prediction model

3. Proposed Method

The proposed methodology is divided into two stages as showin in Figure 1. The first step is a procedure for recognizing the prediction system and the second step is checking outliers using the prediction system. The model identification stage consists of analyzing observations, building a mixed system of mathematical models, optimizing mixed model coefficients and calculating deviations of prediction errors. If the prediction error becomes stochastically large, it is judged by an abnormal value in outlier checking stage . Assuming that the prediction error follows the normal distribution, when the prediction error exists outside 5% on both sides, it is judged to be an abnormal value. When the standard deviation of the prediction error Se is taken, the outlier determination logic uses equation (1).

$$|e| > 1.645Se \tag{1}$$

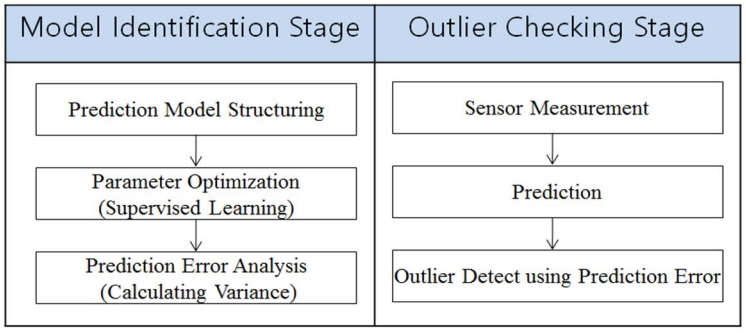

Figure 2. Overall Procedures of Proposed Method

3.1 Construction of mixed prediction model

As a result of analyzing the sensor measurement in a dam, it was confirmed that the measurement has the regression characteristic by neighboring sensors and also has the time series characteristic affected by past sensor signal. The mixed model proposed in this study is the equation (2).

$$y_{i,t} = y_{i,t}^{s} + y_{i,t}^{r} \tag{2}$$

$y_{i,t}$: Measurement value at the time t of sensor i

$y_{i,t}^{s}$: Time series component at the time t of sensor i

$y_{i,t}^{r}$: Regression model component at the time t of sensor i

Assuming the above mixed model, the model structure is determined by the following procedure.

(1) The neighboring sensors having significant correlation with the corresponding sensor are selected to determine a multiple regression analysis model.
(2) Determine the order of the AR(Auto Regressive) & MA(Moving Average) by analyzing sensor signals.

As a result, when the mixed prediction model of the measurement signal is assumed to be [AR = 3, MA = 3], it can be defined as Equation (3)

$$\hat{y}_{i,t} = \hat{y}^{s}_{i,t} + \hat{y}^{r}_{i,t} \tag{3}$$
$$\hat{y}^{s}_{i,t} = \phi_0 + \phi_1 y^{s}_{i,t-1} + \phi_2 y^{s}_{i,t-2} + \phi_3 y^{s}_{i,t-3} + \Theta_1 e_{t-1} + \Theta_2 e_{t-2} + \Theta_3 e_{t-3}$$
$$\hat{y}^{r}_{i,t} = b + a_1 y_{1,t} + \cdots + a_{i-1} y_{i-1,t} + a_{i+1} y_{i+1,t} + \cdots + a_n y_{n,t}$$

In the above equation, $\hat{y}^{r}_{i,t}$ means a multiple regression model prediction using the sensors which the *i* sensor is excluded.

3.2 Parameter estimation by genetic algorithm

In order to estimate the parameter [φ, θ, b, a] using the mixed prediction model, a genetic algorithm applied to the supervised learning. In order to apply the genetic algorithm to the optimization problem, the solution of the problem must be represented by a chromosome. When the size of the AR model is N, the MA model size is M and the regression model size is Z, the chromosome representation is shown Figure 3. Each parameter is expressed as a real value between -1 and 1.

ϕ_0	ϕ_1	...	ϕ_N	θ_1	...	θ_M	b	a_1	...	a_Z

Figure 3. Genetic Individuals

The sequential parameter search procedure is follows.

(1) Initially, generate a set. The set size is determined according to the problem size. (In this research, 100 pieces are used)
(2) Using each solution, calculate prediction accuracy and evaluate each fitness of solution.
(3) Using a crossover and a mutation, create a new solution and complete the set for the next generation.
(4) Until a satisfactory optimal solution is acquired, three procedures are repeated.

The procedure for calculating fitness of solution is as follows.

(1) The predicted value of the regression analysis component $\hat{y}^{r}_{i,t}$ is calculated with the regression parameters of the solution.
(2) Calculate $y^{\sim r}_{i,t} = y_{i,t} - \hat{y}^{r}_{i,t}$, $y^{\sim r}_{i,t}$ means the signal which a regression model component has been removed.
(3) The time series prediction $\hat{y}^{s}_{i,t}$ is calculated using the time series parameters of solution.
(4) Combining the prediction value of the regression model component and the prediction value of the time series component, that is, the final predicted value $\hat{y}_{i,t}$ is obtained using $\hat{y}^{s}_{i,t} + \hat{y}^{r}_{i,t}$.
(5) The fitness is obtained using the reciprocal of the prediction error.

The genetic algorithm searches an optimal solution by repeating the process of creating a new improved set. Crossover and mutation are used for generation of new solutin set. Crossover provides a way of local search by mixing two solutions to make better solution. Crossover is a major searching parameter that should be colser to 1. The mutation replaces the value of a particular position of the solution with a random value and provides opportunities for global search. The mutation rate should be samll so that it does not spoil the crossover search[7]. We used 0.8 for crossover rate and 0.2 for mutation rate.

4. Test Signal

4.1 Virtual signal

The virtual signal generation model is the equation (4).

$$s_{i,t} = ts_{i,t} + x_t a_i + b_i + e_{i,t} \quad (4)$$

$s_{i,t}$: Measurement value at the time *t* of virtual sensor *i*

$ts_{i,t}$: Time series component at the time *t* of virtual sensor *i*

x_t : Independent variable at the time *t*

a_i : Slope of regression model on virtual sensor *i*

b_i : Intercept of regression model on virtual sensor *i*

$e_{i,t}$: Noise at the time t of virtual sensor *i*

$ts_{i,t}$ is generated by three virtual signals using ARMA (1, 0, 1), ARMA (2, 0, 2) and ARMA (3, 0, 3) models. x_t represents a random number of 0 to 1. The virtual signal was generated in the form of i = 1~4, t = 1~128, that is, the number of sensors is four and the number of time indexes is 128. Using the first half data is used for learning and the second half is used for testing the prediction performance.

4.2 Pore water pressure in actual dam

Prediction performance and outlier detection is tested using the pore water pressure sensor data of a real dam. Figure 4 shows the daily measured values for 4 months of four pore water pressure sensors. The reliability of the S4 sensor is lower than other sensors in the opinion of the field expert. The S4 signal shows a different aspect from the other three sensors.

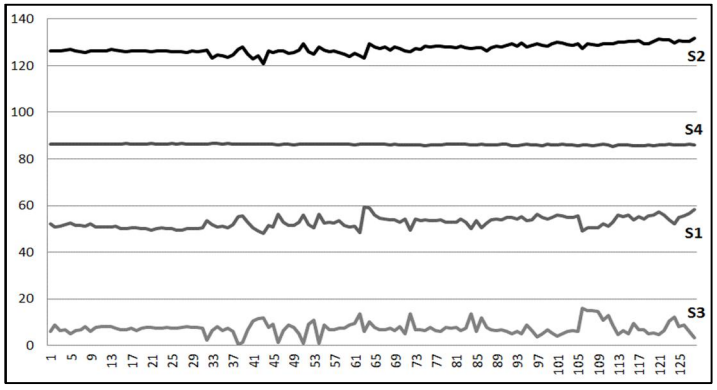

Figure 4. Real Sensor Data

The difference signal is used for the prediction and the outlier analysis as shown in Figure 5. Using the first half data is used for learning and the second half is used for testing the prediction performance.

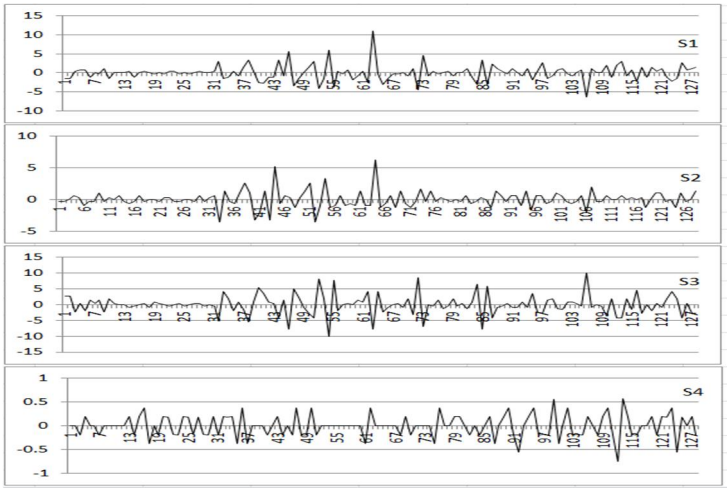

Figure 5. Difference Value of Sensor Data

5. Experimental Results

5.1 Experimental result of virtual signal prediction

Prediction accuracy was measured by MSE (Mean Squared Error). The MSE of three virtual signal prediction is in Table 1. As shown in Figure 6, it was analyzed that the prediction accuracy of the proposed method was the highest. Among the existing methodologies, regression-ARIMA method shows relatively good performance.

Table 1. Prediction error of artificial sensor data

	Proposed method	*Neural network*	*Regression-ARIMA*	*ARIMA-Regression*
ARIMA(1,0,1)+Regression	1.328	1.895	1.683	3.697
ARIMA(2,0,2)+Regression	1.338	1.835	1.928	2.152
ARIMA(3,0,3)+Regression	1.093	1.757	1.388	3.103

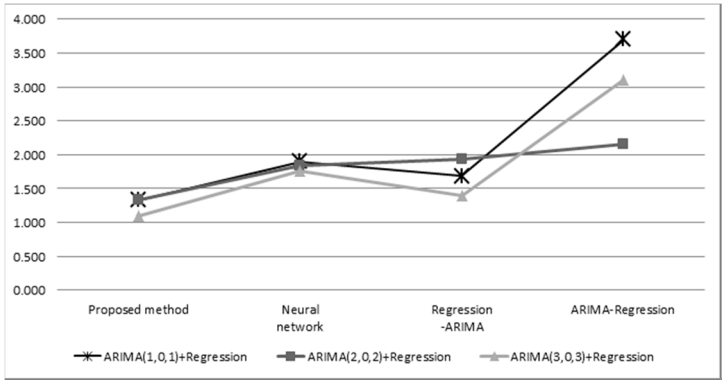

Figure 6. Prediction performance of artificial sensors

5.2 Experimental result of actual signal prediction

The MSE of four prediction experiments on pore water pressure sensor is in Table 2.

Table 2. Prediction error of real sensor data

	Proposed method	*Neural network*	*Regression -ARIMA*	*ARIMA -Regression*
S1	10.022	15.110	15.343	18.643
S2	11.205	21.040	16.205	19.594
S3	11.955	15.855	19.938	19.673
S4	20.955	25.422	24.097	26.493

Figure 7 is the prediction error analysis result using the four methods. The prediction accuracy of the proposed method is the highest. The prediction error of the sensor S4 is remarkably higher than the other sensors. Same with the results using virtual sensors, regression-ARIMAN prediction shows relatively good performance.

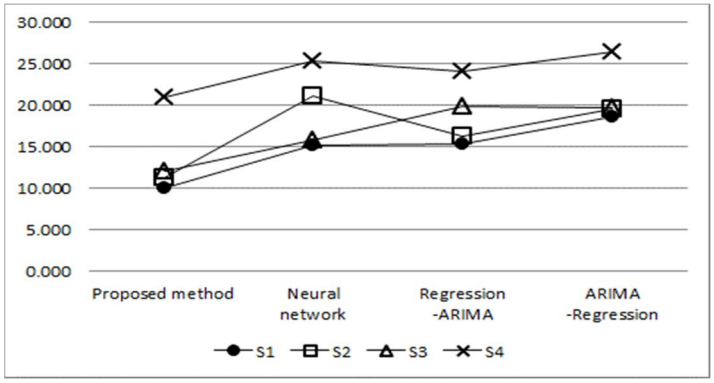

Figure 7. Prediction performance of real sensors

5.3 Outlier detection result of actual sensor signal

Figure 8 is the outlier detection result of the real sensor using the proposed method. The sensor measurement is accurately predicted and the outlier is well detected.

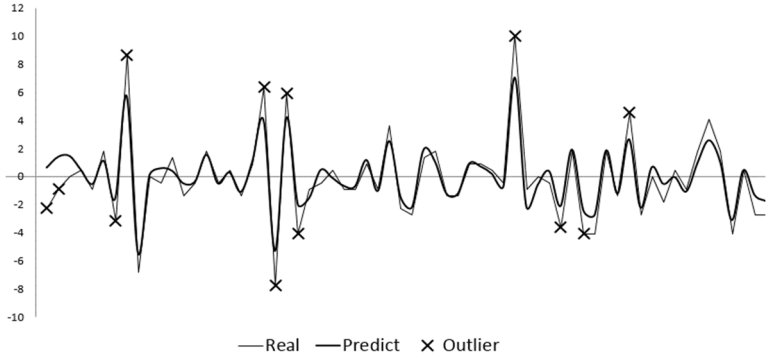

Figure 8. Outlier detection using proposed method

Figure 9 is the outlier detection result of the neural network. Some erroneous outlier detection are observed at the arrow portions.

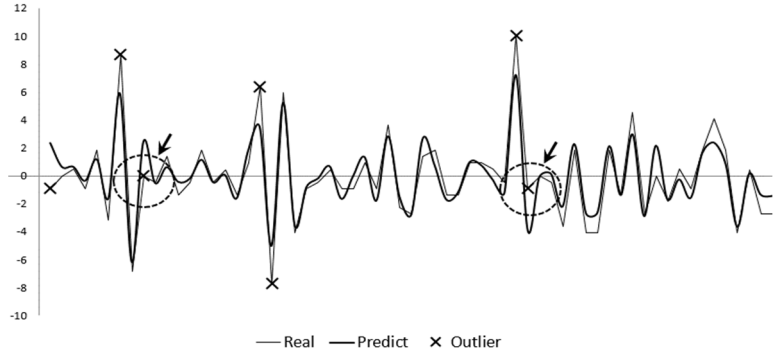

Figure 9. Outlier detection using neural network

Figure 10 is the outlier detection result of regression-ARIMA method. Some erroneous outlier detection are observed at the arrow portions.

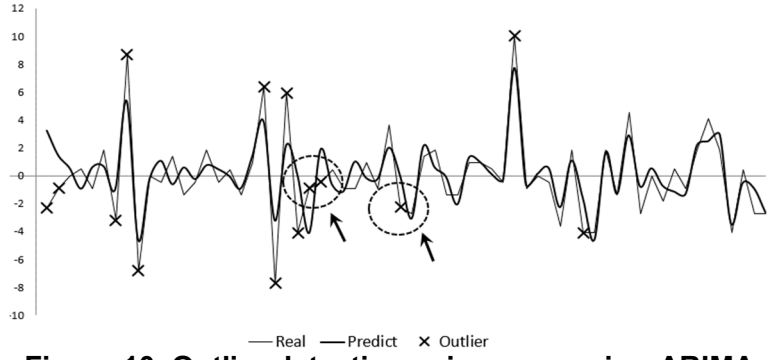

Figure 10. Outlier detection using regression-ARIMA

Figure 11 is the outlier detection result of ARIMA-regression method. Some erroneous outlier detection are observed at the arrow portions.

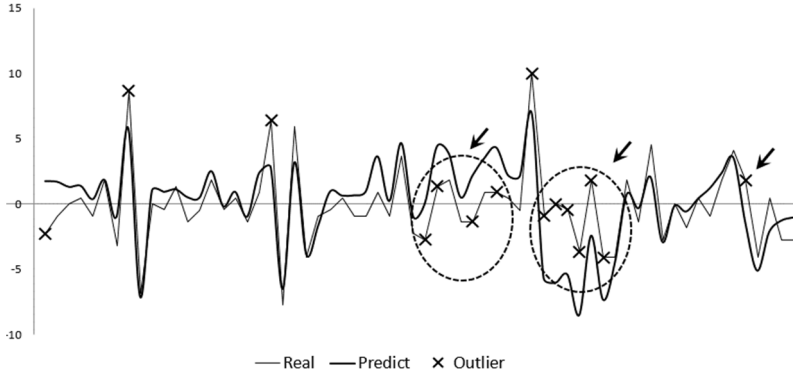

Figure 11. Outlier detection using ARIMA- regression

6. Discussion and Conclusion

In this research, a prediction method was proposed for detecting abnormal values of dam sensor. The dam sensor measurement has the characteristics of complex prediction models, and it is difficult to accurately

estimate the parameter of the model using an existing mathematical method. In such a case, a prediction method based on an artificial intelligence such as neural network can be useful. However, although the neural network based prediction model has advantages that it can be used generically for various systems, restrictions can occur in learning without reflecting the characteristics of various systems. In order to overcome these drawbacks, estimating the parameter of the mixed prediction model is optimized using GA based supervised learning.

A comparative analysis of the proposed method and the conventional method is presented in this paper. The virtual data was created and experimented for prediction comparison. The pore water pressure sensor of the actual dam was also used. In the experimental results, the prediction of the proposed method shows best performance for all data, and the proposed method shows a reliable result in outlier detection. As a result of these studies, we confirmed the effectiveness of the proposed method for outlier detection in dam measurement sensors.

References

[1] J. H. Park, M. K. Jang, G. H. Lee, E. K. Oh and S. W. Hur, "Forecasting Algorithm for Vessel Engine Failure", Journal of KIIT, 14(11), pp.109-117, 2016

[2] C. M. Park and J. S. Jeon, "Regression-based outlier detection of sensor measurements using independent variable synthesis", Journal of the Korean Institute of Plant Engineering, 20(3), pp. 87-93, 2015

[3] M. Mourad and J. L. Bertrand-Krajewski, "A method for automatic validation of long time series of data in urban hydrology", Water Science & Technology, 45(4), pp. 263-270, 2002

[4] I. S. Jung, S. C. Park and G. N. Wang, "Two phase reverse neural network based facilities failure prediction system", Journal of the Korean Institute of Plant Engineering, 11(2), pp. 145-154, 2006

[5] G. J. Williams, R. A. Baxter, H. X. He, S. Hawkins and L. Gu, "A comparative study of RNN for outlier detection in data mining", IEEE International Conference on Data-mining Technical Report 02(102), pp. 709, 2002

[6] A. B. Sharma, L. Golubchik, R. Govindan, "Sensor faults: Detection methods and prevalence in real-world datasets", ACM Transactions on Sensor Networks v.6 n.3, pp. 1–39, 2010

[7] M. Mitchell, "An introduction to genetic algorithms," The MIT Press, pp. 166, 1997