Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 154 (2016) 1110 – 1115

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Hydroinformatics, HIC 2016

Forecasting Water Quality Parameters by ANN Model using Pre-processing Technique at The Downstream of Cheongpyeong Dam

Il won Seo[a], Se Hun Yun[a], Soo Yeon Choi[a]

[a]*Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*

Abstract

In this study, the water quality parameters (Temperature, DO, pH, Electric Conductivity, TN, TP, Turbidity and Chlorophyll-a) at the downstream of Cheongpyeong dam are predicted using artificial neural network. The artificial neural network(ANN) is a powerful computational technique for modeling complex relationship between input and output data. Typically, Time series generally consists of a linear combination of trend, periodicity and stochastic component. In this study, to reduce the influence of trend, periodicity and stochastic component and to enhance the performance of ANN model, developed the Ensemble ANN model with stratified sampling method. 7 parameters (Temperature, DO, pH, Electric Conductivity, TN, TP, and Chlorophyll-a) have the higher than 0.85 R squared. And 5 parameters (Temperature, DO, pH, TN, and TP shows than 1.0 RMSE

Keywords: : Artificial Neural Network; Ensemble modeling; stratified sampling; water quality forecasting; North Han River

Peer-review under responsibility of the organizing committee of HIC 2016

1. Introduction

Because of the increased demand for recreation activity and re-development of the waterfront infrastructure, Interest of water leisure activity has been increasing substantially. However, by the human body is in contact with the water pollution, there is a possibility of infecting with a waterborne diseases and eye and skin diseases. For this reason, it is important to predict the water quality of the region where water sports are doing actively.

Many researchers using ANN technique to predict water quality parameters in river systems. But unbalanced input data set and effect of initial weight parameter makes it difficult to predict the water quality parameters

Peer-review under responsibility of the organizing committee of HIC 2016
doi:10.1016/j.proeng.2016.07.519

accurately. In addition, it gives different results for the same input data set. So this study aims to lessen the influence of the biased data set and delete the effect of initial weight parameters by using ANN Ensemble model and stratified sampling method.

2. Model and methods

2.1 ANN ensemble modeling

Artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs. In general, ANN are presented as systems of interconnected "neurons". Neurons are fundamental to the operation of a neural network. It is a highly interconnected processor and exchange messages between each other. Each neuron consists of a weight parameter and an activation function. When the input data is passed to the input layer, it is binding with the weight and was granted a non-linearity by activation function. The process of being transmitted to the next neuron is repeated until derive a final result.

Development of ANN model consists of two parts, one is pre-processing part and the other is modelling part. Data collecting, data analysing and selecting input, output data is pre-processing part. And selecting number of parameter like hidden layers, hidden neurons and ensemble is modelling part. ANN model calculate the weight parameters through an optimization algorithm in the modelling part. Depending on the initial weight parameter, model gives different results for the same original inputs. For this reason, ensemble technique has been applied.

Ensemble techniques have been applied with considerable success in hydrology and environmental science, as an approach to enhance the skill of forecasts (Krogh and Vedelsby, 1995; Araghinejad et al., 2011). The motivation for this procedure is based on the idea that one might improve the performance of a single generic predictor, by combining the outputs of several individual predictors (Krogh and Vedelsby, 1995).

Laucelli et al. (2007) applied ensemble modelling to hydrological forecasts, and showed the error due to the variance is effectively eliminated. This result shows that ensemble modelling could reduce the uncertainty.

In this study, the ensemble modelling technique was applied to estimate the performance of the ANN, without the influence of initial weight parameters on the model results. Randomly generated initial weight are applied in each ensemble members on the training process.

2.2 Stratified sampling method

As described above, ANN model is data-driven model so most important thing is the quality of data. the next important thing is the pre-treatment process of the data. For this reason, in this study stratified sampling method was used to reduce the sampling error. Stratified sampling is the process of dividing members input data into homogeneous subgroups before sampling. The strata should be mutually exclusive and also be collectively exhaustive. This often improves the representativeness of the sample by reducing sampling error. also It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population. To avoid the biased sampling in training, validation and test data set and obtain the appropriate model, training, validation, test data set was sampled by a stratified sampling method according to the distribution ratio of each parameters have. As shown in Table. 1.

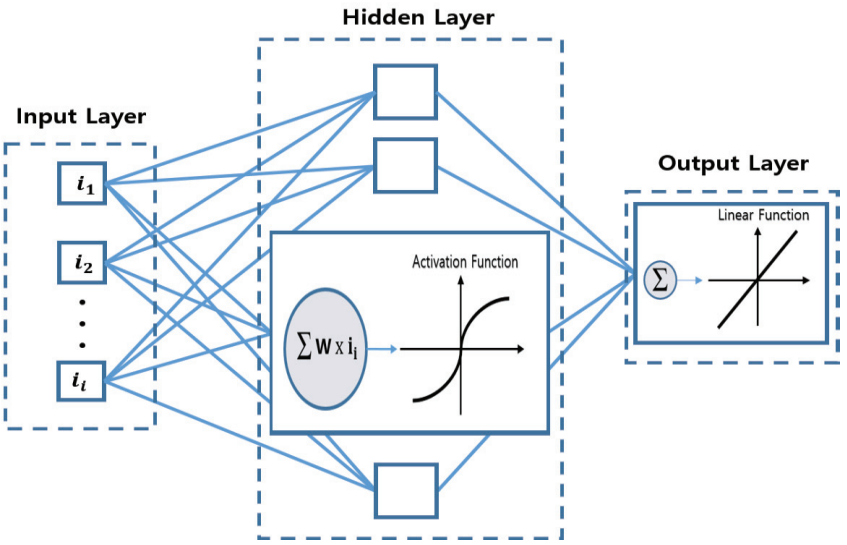

Figure 1 Schematic diagram of ANN model

To the estimate the result, the root-mean-square error (RMSE), interquartile range (IQR)and R squared were used. RMSE was used to measure the whole bias error between the ensemble means and the observed values, and IQR was used to measure the variance error of the ANN ensemble model itself. R squared is a number that indicates how well data fit a statistical model. So it was used to estimate the performance of this models.

$$\text{RMSE} = \sqrt{\frac{SS_{rss}}{n}} \quad (1)$$

where, n is the number of the observed data.

$$IQR_i = Q_{75}(y_i) - Q_{25}(y_i) \quad (2)$$

Where $Q_{75}(y_i)$ and $Q_{25}(y_i)$ are the 25th and 75th percentile values of the ANN ensemble model result for the ith data set.

$$R^2 = 1 - \frac{SS_{rss}}{SS_{tss}} \quad (3)$$

$$SS_{tss} = \sum_i (x_i - \bar{x})^2 \quad (4)$$

$$SS_{rss} = \sum_i (x_i - \bar{y_i})^2 \quad (5)$$

SS_{tss} is the total sum of squared, and SS_{rss} is the sum of squares of residuals. x_i is the ith observed value or target value. $\bar{x}$ is the mean value of x_i for all the observed data set. $\bar{y_i}$ is the ensemble mean of Network for the ith data set.

Table 1 stratified sampling result for temperature

Boundary		Distribution of data		Number of sampling data set		
lower	upper	# of data	Ratio(%)	Training data	Validation data	Test data
1.40	3.96	190	19.19	152	19	19

3.96	6.52	93	9.39	75	9	9
6.52	9.08	85	8.59	67	9	9
9.08	11.64	69	6.97	55	7	7
11.64	14.20	73	7.37	59	7	7
14.20	16.76	92	9.29	74	9	9
16.76	19.32	87	8.79	69	9	9
19.32	21.88	169	17.07	135	17	17
21.88	24.44	114	11.52	92	11	11
24.44	27.00	18	1.82	14	2	2
Total		990	100	792	99	99

3. Model application

The Cheongpyeong dam is concrete gravity dam height of 31m, length of the bank is 470m, total poundage capacity is 185 million tons, Basin area 9,921 ㎢, sleeping area at the time of full water level is 12.5 ㎢. Dam is completed in 1944, Cheongpyeong pumped storage power plant, which was completed in 1980 is equipped with the facilities to produce the power of a day 2.4 million ㎾, facility power generation capacity of 80,000 ㎾ and it supplies the majority of the power in Seoul as. Downstream of Cheongpyeong dam is One of the most famous sites to enjoy the water sports near the city Seoul. In this study, Supervised-Feedforward Neural Network is adapted which has been widely used to predict a variety of water quality parameters and use the error -backpropagation algorithm, which is based on the optimization algorithm. Among various error-backpropagation methods, the LevenbergeMarquardt (LM) algorithm has been very successfully applied to the training of ANN to predict streamflow and water quality, providing significant speedup and faster convergence than the steepest descent-based algorithm, and conjugate gradient-based algorithms (e.g. Zamani et al., 2009). So LevenbergeMarquardt (LM) algorithm was applied to train the network.

The observed daily data from 1st July 2012 to 31th August 2015 was used to develop the model. These data divided into input and output data. input data is (t-1) and (t) data and output data is (t+1) data. this ANN model predicted 8 the water quality parameters (Temperature, DO, pH, Electric Conductivity, TN, TP, Turbidity and Chlorophyll-a). Training data set accounting for 80% from the total available data in each parameter. And validation, test data set accounting 10% each. The number of ensemble members is 100 and The number of hidden neurons is 1,2,6,8 and 12 at every parameter. Training goal of algorithm is 0.001 in 500 epochs. Tangent sigmoid function was used as activation function and in output layer single monotonic increase function is used

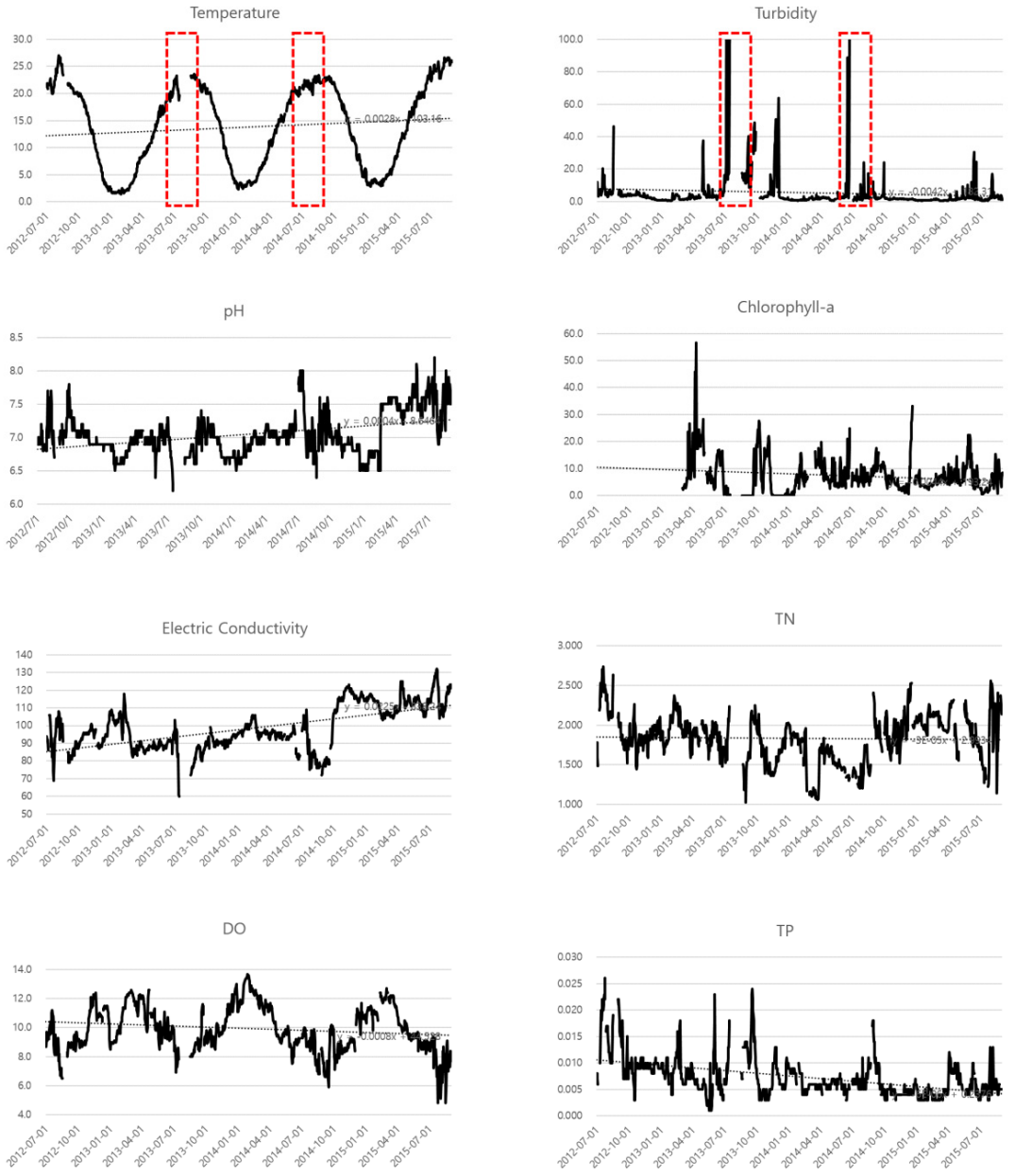

Figure 2 Data distribution in Gapyeong

4. Conclusion

As shown in figure2. The observed daily data from 1st July 2012 to 31th August 2015 tend to increase in Temperature, pH, Electric Conductivity and tend to decrease in DO, TN, TP, Turbidity and Chlorophyll-a. and have the periodicity in all the variances. In turbidity parameter, rare event is occurred frequently from July to August. This shows a tendency to coincide with the rainy period in Korea (from July to August) at summer season.

In this study ANN ensemble model with stratified sampling data was developed and was applied to Gapyeong station on the North Han River. Forecasting the one-step ahead water quality parameters ($Temperature_{t+1}, DO_{t+1}, pH_{t+1}, EC_{t+1}, TN_{t+1}, TP_{t+1}, Turbidity_{t+1}, Chlorophyll-a_{t+1}$) shows good performance in 7 parameters with R squared higher than 0.85 except Turbidity only 0.638. In RMSE, 5 parameters $Temperature_{t+1}, DO_{t+1}, pH_{t+1}, TN_{t+1}, TP_{t+1}$, show good performance lower than 1.0. EC_{t+1}, have 1.649 RMSE in best fitted model and $Chlorophyll-a_{t+1}$ 2.690 . In case of $Turbidity_{t+1}$ it shows 0.683 R squared and 10.071 RMSE. this show that despite the effort to reduce the sampling error by the stratified sampling method, the biased data set affected to result (96% data are belonging to 1 boundary) and this ANN ensemble model is hard to predict the rare event. So further research to lessen the influence of rare event is required.

Table 2 Comparisons of each ANN ensemble model result

Parameters	# of best fitted neuron	training result			validation result			test result		
		RMSE	IQR	R^2	RMSE	IQR	R^2	RMSE	IQR	R^2
TN	1	0.080	0.000	0.917	0.077	0.000	0.924	0.074	0.000	0.932
Turbidity	1	10.071	0.000	0.761	2.132	0.000	0.84	3.208	0.000	0.638
DO	2	0.325	0.045	0.950	0.305	0.045	0.958	0.317	0.054	0.957
pH	6	0.101	0.019	0.889	0.078	0.020	0.923	0.096	0.020	0.881
Temperature	12	0.405	0.089	0.997	0.454	0.087	0.996	0.360	0.091	0.998
TP	12	0.001	0.000	0.932	0.001	0.000	0.878	0.001	0.000	0.891
EC	12	1.649	0.418	0.981	1.857	0.792	0.975	2.183	0.525	0.963
Chlorophyll-a	12	2.690	0.778	0.849	3.940	1.431	0.691	2.137	0.775	0.908

Acknowledgements

This research was supported by a grant (11-TI-C06) from Advanced Water Management Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

This work was conducted at the Research Institute of Engineering and Entrepreneurship and the Integrated Research Institute of Construction and Environment in Seoul National University, Seoul, Korea.

References

[1] 국립환경과학원 (2014). 종합적 친수활동 보호를 위한 수질예보 기법의 개발.

[2] Alejo, R., Garcia, V., Sotoca, J.M., Mollineda, R.A., Senchez, J.S., 2007.Improving the performance of the RBF neural networks trained with imbalanced samples. Lect. Notes Comput. Sci. 4507, 162e169

[3] Bishop, C.M. (1995). “Neural networks for pattern recognition.” Oxford. pp. 140-148.

[4] Berardi, V.L., Zhang, G.P., 1999. The effect of misclassification costs on neural network classifier. Decis. Sci. 30 (3), 659e683.

[5] Cho, Y.J., and Kim, J.M. (1998). “Water Supply forecast Using Multiple ARMA Model Based on the Analysis of Water

S.E. Kim, I.W. Seo / Journal of Hydro-environment Research 9 (2015) 325-339