SNRU Journal of Science and Technology

Journal home page: snrujst.snru.ac.th

Forecasting water levels using data mining techniques: a case study at Nong Han Lake, Sakon Nakhon, Thailand

Supamit Boonta[1], Rujira Chongklaiklang[2]

[1]Program of Business Computer, Faculty of Management Science, Sakon Nakhon Rajabhat University, Sakon Nakhon, 47000 Thailand

[2]Program of Business Computer, Faculty of Management Science, Sakon Nakhon Rajabhat University, Sakon Nakhon, 47000 Thailand

[*]**Corresponding Author**: supamit.bo@snru.ac.th

Received: 13 November 2018; **Revised**: 02 April 2019; **Accepted**: 17 April 2019; **Available online**: 01 May 2019

Abstract

Nong Han is the largest lake in the northeast region of Thailand. Looking at the flood crisis in 2017, Sakon Nakhon province was a significant impact area. This research aims to compare the efficiency of data mining techniques for predicting changes in water level in Nong Han by three methods, as follows: 1) support vector machine, 2) linear regression, and 3) multi-layer perceptron to create a suitable model for predicting changes in water levels in Nong Han. Mean absolute error, was employed to assess the model's efficiency for selecting the best techniques through analyzing daily data collected from 2011 to 2016 for 1,961 items.

Comparing the efficiency of models forecasting changing in water level for the Nong Han lake, it was found that the multi-layer perceptron was the highest performing model with MAE of 7.921. Then the linear regression and the support vector machine with MAE of 8.343 and 10.824, respectively.

Keywords: support vector machine; linear regression; multi-layer perceptron; water level changes forecasting

1. Introduction

Nong Han is the largest freshwater lake in the northeast of Thailand. It is the second largest lake in Thailand [1]. The water sources of Nong Han lake are from Phu Phan Mountain and other surrounding areas. Most of the water routes in the city of Sakon Nakhon are flowed into Nong Han lake and then drain out at a floodgate into Lam Nam Kam river before flowing into the Mekong river. During 2017, Sakon Nakhon province was damaged from the worst flooding in 30 years. More than 10,000 hectares of affected farmland. Therefore, better water management is needed urgently in order to prepare for similar future disasters.

The purpose of this study was to establish a model for predicting the changing in water level in Nong Han Lake, using secondary data of Sakon Nakhon Freshwater Fisheries Research and Development Center, which shows the changing in the water levels in Nong Han using daily data from 2011 to 2019, total 1,961 records, to get the forecasting changes in water level in advance 1 day in order to properly manage drainage and mitigate the impact of flooding in the community areas. This includes protecting surrounding agriculture during the rainy season, maintaining sufficient freshwater for fish, and also storing potable water for consumption during the dry season. This experiment looked at different data mining techniques, such as a support vector machine, linear regression, and multi-layer perceptron. To analyze the most appropriate technique for the Nong Han Water Level Forecasting, the performance of these techniques was compared, to find the most accurate predictive technique by measuring the model's predicted water yield by using the mean absolute error (MAE).

2. Materials and Methods

Related Content

Artificial Neural Networks is a model based on Neuron in the human brain, which has the component of cells that each cell have a bias. The cells are linked together by weights; the learning process of the model can modify their weight and bias [2].

Multi-layer perceptron neural network is one of the most popular data mining techniques. It is flexible to learn independently adaptable data to find data models and develop nonlinear system models to predict reliable data. Thus, this can be used to solve complex real-world problems [3].

The multi-layer perceptron neural network architecture consists of an input layer, a hidden layer, and output layer as shown in Fig. 1.

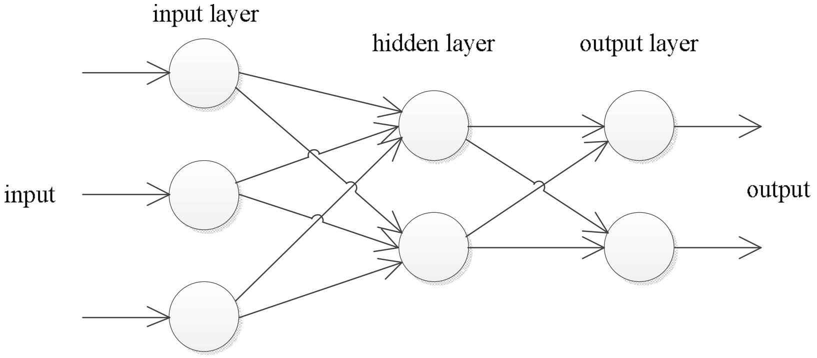

Fig. 1 multi-layer perceptron neural network architecture.

Each layer consists of a number of neurons. The data is continuously flowing forward, and the network errors go backward. The connection weight and network criteria are initially randomized. Then the training process is used to compare the error value of the result with the actual value. Furthermore, it will be used to determine the Mean Square Error (MSE) until getting the result. [4]

The linear regression analysis is widely used to predict for a single variable (Output variables) with one or more independent variables (Predictive variables) [5]. linear regression with a single independent variable is called simple linear regression, and the linear regression with more independent variables is called Multiple linear regression [6]

The simple linear regression equation is the estimation of the coefficients for each predictor variable (X). The general equation for regression analysis is shown in equation (1)

$$y = b + ax \tag{1}$$

where: y represents the dependent variable, x is the explanatory variable, a is the slope of the line, and b is the intercept of the regression line.

The support vector machine (SVM) is an instructor-led learning environment. That is one of the most popular techniques, especially with nonlinear modeling for complex systems and processes [7] which can apply to classify and regression. The support vector machine (SVM) has fewer parameters than the artificial neural network because it uses only kernel function and some coefficients while the artificial neural network is more difficult to implement [8].

The operation of support vector machine divides data in multidimensional planes into two groups. There is one unit that simulates neuronal characteristics by using Kernel Function to transform data into another dimension that has the dividing margin between groups of data. Features and variables are used to define the structure of Feature Space that is to define a multi-dimensional. Then feature selection chooses the most appropriate value. The set is divided into groups of data, called vector. Straight lines need to be close to the top of the data, separate the data completely, and have a high resolution of the data, resulting in the best fit and the best possible one. However, the performance of the support vector machine depends on the selection of the appropriate parameters [9] the support vector machine (SVM) in Fig. 2.

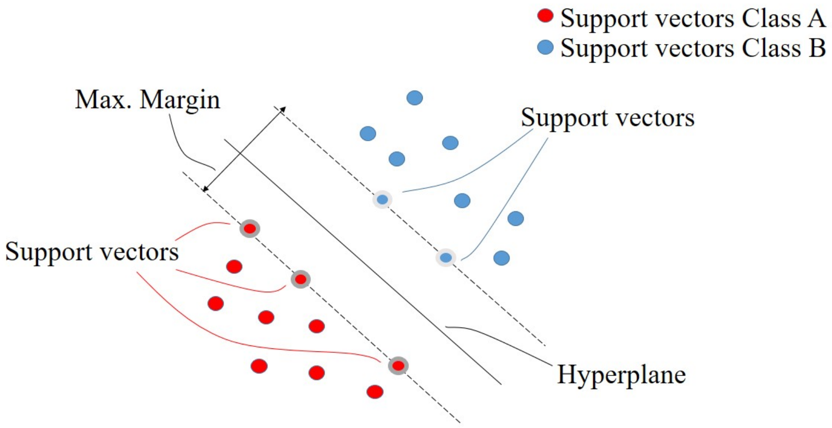

Fig. 2 Principles of support vector machine (SVM)

Literature Review

In the study of research on water prediction, There are related researches as the following. Masoud Bakhtyari Kia [10] studied the development of flood modeling by comparing the causes of flooding by using artificial neural networks and geographic information to create the model and simulate flooded areas in the southern part of the Malaysian Peninsula. The results showed that height is the most essential factor for flood susceptibility (R^2 = 0.931), followed by slope (R^2 = 0.963) and land use (R^2 = 0.986) respectively. Chih-Chiang Wei [11] conducted a study on tides in the Tanshui River basin during the typhoon seasons in Taiwan by using Wavelet SVMs were compared to the normal support vector machine. As a result of the performance comparison, it was found that the Wavelet SVMs had a more accurate prediction than the normal support vector machine with an RMSE value. 0.205 m and 0.160 m and at the Taipei Bridge Station and 0.154 m and 0.092 m, respectively, at Tudigong Station. Thus, the Wavelet SVM can be used to solve practical problems for predicting water levels during attacks Typhoon. Munin Wanatada and Punnee Sittidech [12] studied the water level forecasting in

Chaiyaphum municipality with Back-propagation neural network technique by analyzing the amount of Rainfall-Runoff from the gauge stations which are the major sources in flooding of the municipal area. The selection of gauge stations, variables as well as time lags was used to developing an appropriate forecasting model. The results of this research showed that the model using 15 input variables has Mean Absolute Error of 1.008.

Methodology

Research tools

In order to create a predictive model, researchers have created a prediction model using IBM SPSS Modeler version 14.1 for analyzing by using data mining technique.

Data Preparation

The researcher divided the test data into seven sets. Water volume in the river above the floodgate in the previous 1 – 7 days, average rainfall in the previous 1 – 7 days, average daily water discharge volume in the previous 1 – 7 days and water level in 1 day were collected by Freshwater Fisheries Research and Development Center of Sakon Nakhon, in a total of 1,961 records. Then eliminated the missing data and got in a total of 1,797 records and then converted the data into numerical data to analyze which used to create the model of 13 variables, this is shown in Table 1.

To create a model for forecasting the changing in water level in order to predict water level changes in advance for one day. The researchers divided the data into two parts which are training data set with the record number of 1,257 records and the rest of 540 records for information in testing data set.

Table 1 Variables used in model building.

Variable	Explanation
AvgStreamflow1	average rainfall for 1 day
AvgStreamflow2	average rainfall for 2 days
AvgStreamflow3	average rainfall for 3 days
AvgStreamflow4	average rainfall for 4 days
AvgStreamflow5	average rainfall for 5 days
AvgStreamflow6	average rainfall for 6 days
AvgStreamflow7	average rainfall for 7 days
AvgDrainage1	average amount of drained water per day for 1 day
AvgDrainage2	average amount of drained water per day for 2 day
AvgDrainage3	average amount of drained water per day for 3 days
AvgDrainage4	average amount of drained water per day for 4 days
AvgDrainage5	average amount of drained water per day for 5 days
AvgDrainage6	average amount of drained water per day for 6 days
AvgDrainage7	average amount of drained water per day for 7 days
AvgStreamflow1	average water volume in the river above the floodgate for 1 day
AvgStreamflow2	average water volume in the river above the floodgate for 2 days
AvgStreamflow3	average water volume in the river above the floodgate for 3 days
AvgStreamflow4	average water volume in the river above the floodgate for 4 days
AvgStreamflow5	average water volume in the river above the floodgate for 5 days
AvgStreamflow6	average water volume in the river above the floodgate for 6 days

Table 1 (cont.)

Variable	Explanation
AvgStreamflow7	average water volume in the river above the floodgate for 7 days
WaterLevel1	Water level for 1 day
Change	Water level change

Data Analysis

The researchers were analyzed data by using 3 data mining techniques including support vector machine, linear regression analysis, and multi-layer perceptron. The mean absolute error (MAE) is shown in equation (2).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| f_i - y_i \right| \tag{2}$$

where: MAE represents the mean absolute error, n is the total amount of data, f_i is the predicted water level, and y_i is the actual water level.

3. Results and Discussion

The research results of the development of suitable models for water level forecasting of Nong Han lake in Sakon Nakhon province, by using data mining techniques which are: multi-layer perceptron networks, linear regression analysis and support vector machine. Experiment with the training data set and the test data set and evaluate the efficiency of the model with the mean absolute error (MAE). The multi-layer perceptron network is the most efficient method with the minimum MAE of 8.245 and 7.921 for the training data set and the test data set, respectively, when using predictive variables which are: the water level in the previous 1 day, the average rainfall in the previous 2 days, the average amount of drained water per day in the previous 2 day and the average water volume in the river above the floodgate in the previous 2 days. Then the linear regression with the minimum MAE of 8.520 and 8.343 for the training data set and the test data set, respectively, when using predictive variables which are: the water level in the previous 1 day, the average rainfall in the previous 2 days, the average amount of drained water per day in the previous 2 day and the average water volume in the river above the floodgate in the previous 2 days. And the support vector machine with the minimum MAE of 10.627 and 10.824 for the training data set and the test data set, respectively, when using predictive variables which are: the water level in the previous 1 day, the average rainfall in the previous 3 days, the average amount of drained water per day in the previous 3 day and the average water volume in the river above the floodgate in the previous 3 days, presented in the Table 2.

Table 2 The Comparison of performance, forecasting model, and change in water level

	SVM		Linear Regression		Multilayer Perceptron	
	Training	Testing	Training	Testing	Training	Testing
Lag1	12.064	11.371	10.912	10.237	10.105	9.521
Lag2	11.115	10.982	**8.520**	**8.343**	8.894	**7.921**
Lag3	**10.627**	11.099	9.273	9.122	**8.245**	8.169
Lag4	10.804	11.161	9.217	8.784	8.303	8.305
Lag5	11.181	**10.824**	9.285	8.660	8.337	8.334
Lag6	11.420	10.863	9.459	8.827	8.699	8.442
Lag7	12.103	11.484	9.644	9.113	11.513	12.146

4. Conclusion

The results of this research show the findings on water level forecasting by using data mining techniques. The recorded data which are: the average rainfall in the previous, the average amount of drained water per day in the previous, and the average water volume in the river above the floodgate in the previous, suitable for The predicting changes in water level in Nong Han ranged from 2 – 5 days, depending on the technique used. The water level in the previous 1 day is the factor that forecasts the changing in water level more accurate. The most effective technique is the multi-layer perceptron neural network. The mean absolute error (MAE) was 8.245.

5. Suggestion

Suggestion to apply

Based on the results of the study, it was found that if needed to be developed in the form of an information system for simulating the water level in Nong Han lake, it was necessary to bring the water level data for the past one day. Forecast of rainfall from the Meteorological Department And calculating the amount of water drained per day obtained from The amount of water through the floodgate The amount of water through Spillway should be used as input data.

Suggestions for the next research

Researchers should also add other factors that affect the increasing or decreasing of water levels in Nong Han, such as the seasonal effects. For example, in the summer water levels may drop more but during the wet season, levels may increase more. The evaporation of water depends on the weather and surface area. For example, the discharge of water over the dam, rainfall and the release of water that flows from the rivers into Nong Han, etc. Therefore, we should promote the development of a data monitoring system. For example, the telemetry system in order to obtain sufficient information to make better forecasting possible.

6. Acknowledgments

The research team would like to thank Freshwater Fisheries Research and Development Center of Sakon Nakhon for the information provided in the analysis and also Sakon Nakhon Irrigation Project, the Fifth Irrigation Office for providing guidance and information technology for water management.

7. References

[1] S. Boonkirdram, Development of Water Quality Monitoring Wireless Comunication System Using Zigbee, KBEJ. 7(1) (2017) 92 – 104.

[2] T. Vangpaisal, J. Threenat, Factors Affecting the Accuracy of Water Level Forecasting at M.7 Station Using Artificial Neural Network Model, UBU Engineering Journal. 6(1) (2013) 50 – 60.

[3] S. Samarasinghe, Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition, Auerbach Publications, New York, 2007.

[4] L. Wang, Y. Zeng, T. Chen, Back propagation neural network with adaptive differential evolution algorithm for time series forecasting, Expert Syst Appl. 42(2) (2015) 855 – 863.

[5] S. Chiu, D. Tavella, Data Mining and Market Intelligence for Optimal Marketing Returns, Butterworth-Heinemann, Burlington, 2008.

[6] R. Nisbet, G. Miner, J. Elder, Handbook of Statistical Analysis and Data Mining Applications, Academic Press, California, 2009.

[7] D.L. Olson, D. Delen, Advanced Data Mining Techniques, Springer, Berlin, 2008.

[8] S. Pinyopan, B. Kijsirikul, Support Vector Machines for Derivatives Price Prediction, The 10th National Conference on Computing and Information Technology, Angsana Laguna Phuket. 8 – 9 May 2014, 466 – 471.

[9] N. Ployong, N. Porrawatpreyakorn, A Comparison of Decision Tree and Support Vector Machine Techniques for Classifying Students for e-Learning in Information Technology Course, The 9th National Conference on Computing and Information Technology, King Mongkut's University of Technology North Bangkok. 15 – 17 May 2013, 127 – 132.

[10] M.B. Kia, S. Pirasteh, B. Pradhan, A.R. Mahmud, W.N.A. Sulaiman, A. Moradi, An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia, Environ Earth Sci. 67(1) (2012) 251 – 264.

[11] C.C. Wei, Wavelet kernel support vector machines forecasting techniques: Case study on water-level predictions during typhoons, Expert Syst Appl. 39(5) (2012) 5189 – 5199.

[12] M. Wanatada, P. Sittidech, Runoff Forecasting Using Back-propagation neural network Technique: Case Study of Municipality of Chaiyaphum, The 9th National Conference on Computing and Information Technology, King Mongkut's University of Technology North Bangkok. 15 – 17 May 2013, 179 – 184.