Comparison of ANN and SVM for Prediction of Biochemical Oxygen Demand in Chaophraya River

Weeris Treeratanajaru[1], Supawin Watcharamul[2*] and Rajalida Lipikorn[3*]
[1, 3] Machine Intelligence and Multimedia Information Technology Lab,
Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University
[2] Department of Environmental Science, Faculty of Science, Chulalongkorn University
Phayathai Road, Pathumwan, Bangkok 10330, Thailand
E-mail: [1]Weeris.T@gmail.com, [2]Supawin.W@chula.ac.th, [3]Rajalida.L@chula.ac.th

Abstract: Artificial Neural Network (ANN) and Support Vector Machine (SVM) models are used increasingly to predict, monitor and forecast water quality. In this paper, two methods were implemented to predict biochemical oxygen demand (BOD) of Chaophraya River, Thailand using a set of simple measurable surface water quality variables including water temperature, dissolved oxygen (DO), electrical conductivity (EC), pH, nitrate, ammonia, total phosphate (TP), monitoring time, and monitoring location as input variables. The data set consists of 1248 water samples represent 18 different monitoring stations along the Chaophaya, which has been monitored for 17 years. The associated parameters for optimum ANN and SVM model were obtained using grid search technique. The ANN and SVM models can predict BOD in training and testing data sets with reasonably high correlation. The overall results showed that both models could be used as one of the fast, reliable and cost-effective methods for predicting BOD in environments.

Keywords-- **ANN, SVM, Water quality model**

1. Introduction

The water quality is a subject of ongoing concern. Deterioration of water quality has initiated serious management efforts in many countries. Most acceptable ecological and water related decisions are somewhat difficult to make without careful modelling, prediction, and analysis of river water quality for typical development scenarios. Accurate predictions of future phenomena are the lifeblood of optimum water resource management in a watershed.

Water quality modelling is the basis of water pollution control project. It predicts the water quality tendency according to the current water quality condition, transfer and transformation rules of the pollutions in the river basin. In addition, several water quality models, such as physicochemical-based models, have been developed to manage the best practices for conserving water quality [1-4]. Most of these models are very complex and require significant amount of field data to support the analysis. Furthermore, many statistical-based models assume that the relationship between response variables and prediction variables are linear and are distributed normally. However, as water quality can be affected by so many factors, traditional data processing methods are no longer efficient enough for solving the problem [5-6], as such factors show a complicated non-linear relation to the variables of water quality prediction. Therefore, utilizing statistical approaches usually does not possess high precision.

Recently, ANNs and SVM approaches have been applied to many fields of science, such as water engineering, ecological science, and environmental science, have been reported [7-12]. Artificial neural network and support vector machine models show that they are able to accurately approximate complicated non-linear input–output relationships. The ANN and SVM models are flexible enough to accommodate additional constraints that may arise during its application. Moreover, both models can reveal hidden relationships in historical data, thus facilitating the prediction of water quality.

Hence, motivating by many successful applications in modelling non-linear system behaviors in a wide range of areas, ANN and SVM models are used to predict BOD in this study. The main objective of this study is to analyze and compare the performance of ANN and SVM models in BOD prediction along Chaophraya River.

2. Data and Monitoring Stations

In this study, water quality data are provided by the Thai Pollution Control Department, Ministry of Natural Resources and Environment during 1996-2013. There are 1248 records of data. Each record consists of 10 attributes including monitoring time, monitoring location, water temperature, dissolved oxygen (DO), electrical conductivity (EC), pH, nitrate, ammonia, total phosphate (TP), and BOD.

Eighteen monitoring station along the Chaophraya belong to the Department of Pollution Control. The Chaophraya basin is important for the daily life of the people in Central Thailand. This river is used for consumption, transportation and recreation. The rapid growth in industry, agricultrue, high-rise and low-rise buildings, and other infrastructures, has had a significant effect on the river water quality. Biochemical oxygen demand (BOD) is an important parameter for interpreting the condition of surface water. The prediction of BOD, then can be utilized in water management and treatment systems.

* corresponding author

3. Methodology

3.1 Artificial neural network

ANN is a proper mathematical structure having an inter-connected assembly of simple processing elements or nodes. In this study, ANN customary architecture is composed of three main layers where is sufficient for ANNs to approximate any complex non-linear function [13]. A major reason is that intermediate cells do not directly connect to output cells. Hence, they will have very small changes in their weight and learn very slowly [14]. Therefore, an ANN model based on a feedforward neural network with a single hidden layer is used. The backpropagation algorithm is used to train the network. Also, the chosen activation function is sigmoidol function. Suitable number of hidden neurons are tested as trial and error approach.

3.2 Support vector machine

SVM is formulated from the principles of statistical learning theory [15] which can be categorized into two types, regression and classification. In this study, regression SVM called ε-SVM is used for BOD prediction. The basic idea of ε-SVM is considered as a training set where each attribute represents the input space of each record and has a corresponding scalar measure output value. The goal is to find a function that predicts BOD in the best possible way. We implement SVM model based on Smola and Scholkopf (2004) [16].

3.3 BOD prediction models and performance evaluation

BOD prediction model can be divided into two parts: feature selection and prediction. In the first part, forward selection (FS) and genetic algorithm (GA) are used as feature selection techniques to select subset of nine water quality variables to feed into the second part. In the second part, ANN and SVM models are implemented to predict BOD of Chaophraya River. Consequently, six different models are generated from the combination of two parts (as shown in Figure 1).

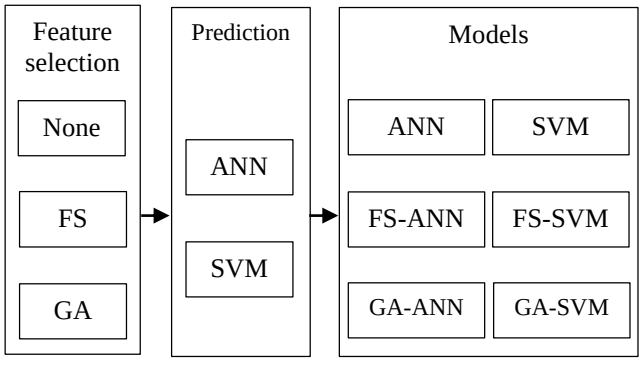

Figure 1. Six BOD prediction models

Nine surface water quality variables including water temperature, dissolved oxygen, electrical conductivity, pH, nitrate, ammonia, total phosphate, monitoring time, and monitoring location are used as input variables. The data set are divided into 70% training set and 30% testing set. The performance of each model is evaluated according to two statistical criteria consisting of correlation coefficient (R), root mean square error (RMSE).

4. Results and Discussions

4.1 ANN model performance

Neural networks were trained using learning rate = 0.3 and momentum = 0.2. The training iteration (epoch) were optimized between 100-1000 epochs. The number of experimental investigations were conducted to find optimum results. The best stopping criteria for training was 200 epochs for GA-ANN. The correlation coefficient, which measured the strength and direction of linear relation between actual BOD and predicted BOD, is R = 0.730. The RMSE of the model is equal to 1.198 as shown in Figure 2.

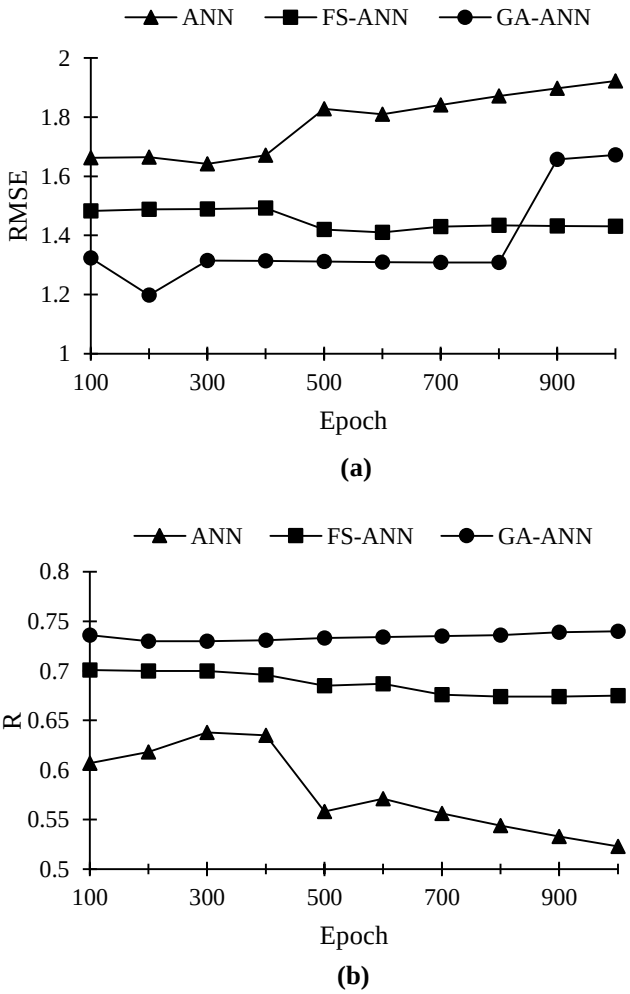

Figure 2. Performance of ANN models indicate by two criteria: **(a)** root mean square error and **(b)** correlation coefficient

4.2 SVM models performance

The SVM model performance was evaluated using the two criteria (R, RMSE) as shown in Table 1 where the best result came from GA-SVM (R and RMSE were 1.285 and 0.731, respectively).

Table 1. Performance of SVM models

Model	#inputs	RMSE	R
SVM	9	1.539	0.574
FS-SVM	3	1.318	0.654
GA-SVM	6	1.285	0.731

4.3 ANN and SVM model performance comparison

The inputs selected for training ANN and SVM were different as shown in Table 2. GA-ANN used location, pH, DO, TP, nitrate, and ammonia as inputs whereas GA-SVM used time, temperature, pH, DO, TP, and ammonia as inputs. Both models showed the ability to generalize and reasonably predict the BOD.

Table 2. Best correlation coefficients and inputs of the two model.

Model	#inputs	R	Inputs
GA-ANN	6	0.730	location, pH, DO, TP, nitrate, and ammonia
GA-SVM	6	0.731	time, temperature, pH, DO, TP, and ammonia

5. Conclusions

In this study, six different models, including ANN, FS-ANN, GA-ANN, SVM, FS-SVM and GA-SVM were implemented to identify the optimal BOD prediction along the Chaophraya. The experimental results show that GA-ANN and GA-SVM models provide the highest correlation coefficients (0.730 and 0.731, respectively). By using GA feature selection, dissolved oxygen (DO), pH, total phosphate (TP) and ammonia are the most common variables used by ANN and SVM models for BOD prediction. Therefore, the proposed models could be used as an efficient tool for managing natural resources and environment, and maintaining compliance with water management regulations and policy.

6. Acknowledgement

The authors wish to thank the Thai Pollution Control Department, Ministry of Natural Resources and Environment for providing the water quality data in this research.

References

[1] J.Y. Park, A.P. Geun, and J.K. Seong, "Assessment of future climate change impact on water quality of Chungju Lake, South Korea, using WASP coupled with SWAT," *Journal of the American Water Resources Association*, vol. 49, no. 6, pp.1225-1238, 2013.

[2] A. Ekdal, M. Gürel, C. Guzel, A. Erturk, A. Tanik, and I.E. Gonenc, "Application of WASP and SWAT models for a mediterranean coastal lagoon with limited seawater exchange," *Journal of Coastal Research*, vol. 64, pp.1023-1027, 2011.

[3] B. Cao, C. Li, Y. Liu, Y. Zhao, J. Sha, and Y. Wang, "Estimation of contribution ratios of pollutant sources to a specific," *Environmental Science and Pollution Research*, vol. 22, no. 10, pp.7569-7581, 2015.

[4] M.F. Ali, M.H. Ahmad, K. Khalid, N.F. Abd Rahman, "Water quality measures using QUAL2E: a study on RoL project at upper Klang River," *Proceedings of the International Civil and Infrastructure Engineering Conference*, pp.757-767, January 2014.

[5] A. Rahman, and M.O. Chughtai, "Reginol interpretation of river Indus water quality data using regression model," *African Journal of Environmental Science and Technology*, vol. 8, no. 1, pp.86-90, 2014.

[6] S. Areerachakul, and S. Sanguansintukul, "A comparison between the multiple linear regression model and neural networks for biochemical oxygen demand estimations," *Eighth International Symposium on Natural Language Processing*, pp.11-14, 2009.

[7] C.G. Wen, and C.S. Lee, "A neural network approach to multiobjective optimization for water quality management in a river basin," *Water Resourses Research*, vol. 34, no. 3, pp.427-436, March 1998.

[8] S. Liu, H. Tai, Q. Ding, D. Li, L. Xu, and Y. Wei, "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction," *Mathematical and Computer Modelling*, vol. 58, pp.458-465, 2013.

[9] M. Liu, and J. Lu, "Support vector machine─an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?," *Environmental Science and Pollution Research*, vol. 21, pp.11036-11053, 2014.

[10] H. Seshan, M.K. Goyal, M.W. Falk, and S. Wuertz, "Support vector regression model of wastewater bioreactor performance using microbial community diversity indices: Effect of stress and bioaugmentation," *Water Research*, vol. 53, pp.282-296. 2014.

[11] A. Najah, A. El-Shafie, O.A. Karim, and A.H. El-Shafie, "Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring," *Environmental Science and Pollution Research*, vol. 21, pp.1658-1670, 2014.

[12] T. He, and P. Chen, "Prediction of water-quality based on wavelet transform using vector machine," *Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, pp.76-81, 2010.

[13] G. Cybenko, 1989. "Approximation by superposition of a sigmoidal function," *Mathematics of Control, Signals, and Sytems*, vol. 2, pp.303–314, 1989.

[14] S.I. Gallant, *Neural Network Learning and Expert Systems*, MIT Press, Cambridge, 1993.

[15] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 1995.

[16] A.J. Smola, and B. Scholkopf, *A tutorial on support vertor regression*, Kluwer Academic Publisher, Netherland, 2004.