An Application Support Vector Machine Model (SVM) Technique for Biochemical Oxygen Demand (BOD) Prediction

Ali Najah Ahmed [1], A. El-Shafie[2]
[1] Senior Lecturer, School of Ocean Engineering, Universiti Malaysia Terengganu, UMT, Malaysia
[2] Associate Professor, Dept. Civil & Structural Eng, Universiti Kebangsaan Malaysia, UKM, Malaysia

ABSTRACT

In this study, Support Vector Machine (SVM) technique has been investigated in prediction of Biochemical Oxygen Demand (BOD). To assess the effect of input parameters on the model, the sensitivity analysis was adopted. To evaluate the performance of the proposed model, three statistical indexes were used, namely; Correlation Coefficient (CC), Mean Square Error (MSE) and Correlation of Efficiency (CE). The principle aim of this study is to develop a computationally efficient and robust approach for predict of BOD which could reduce the cost and labour for measuring these parameters. This research concentrates on the Johor River in Johor State, Malaysia where the dynamics of river water quality are significantly altered.

Keywords: Support Vector Machine, Water quality prediction, Johor River

INTRODUCTION

Malaysia is a developing country that moves towards the vision 2020.Unfortunately the development that had been carried throughout the country gives a bad impact to the environment especially about water quality. This has become sensitive issue, which not only affects human health, but also the entire environment .The development not only affects the water quality, but also the aquatic lives that live in it. Most acceptable ecological and social decisions are difficult to make without careful modelling, prediction, and analysis of river water quality for typical development scenarios. Water quality prediction enables a manager to choose an option that satisfies large number of identified conditions.

Recently, applications of Artificial Intelligence (AI) in the areas of water engineering, ecological sciences, and environmental sciences have been reported since the beginning of the 1990s. In recent years, ANNs have been used intensively for prediction and forecasting in a number of engineering and water-related areas, including water resource study [1-10], oceanography [11], and environmental science [12]. The use of data-driven techniques for modeling the quality of both freshwater [13] and seawater [14] has met with success in the past decade. Reckhow (1999) studied Bayesian probability network models for guiding decision making regarding water quality in the Neuse River in North Carolina. Limited water quality data and the high cost of water quality monitoring often pose serious problems for process-based modelling approaches. AI provide a particularly good option, because they are computationally very fast and require many fewer input parameters and input conditions than deterministic models. AI does, however, require a large pool of representative data for training. Therefore, this paper demonstrates the application of One of AI techniques, namely Support Vector Machine (SVM) to model the values of BOD, having the dynamic and complex processes hidden in the monitored data itself.

MATERIALS AND METHODS

2.1. Study area and water quality data

Johor is the second largest state in the Malaysia Peninsular, with an area of 18,941 km2. The Johor River and its tributaries are important sources of water supply, not only for the state of Johor but also for Singapore. The river comprises 122.7 km long drains, covering an area of 2,636 km2. The station's location map is provided in Figure (1). This station includes four locations along the main stream of the river, which are near to the mouth of the major tributaries and the two largest point sources. Most organic materials such as those from waste water treatment plants, industrial effluents and agricultural run-off are biodegradable. The amount of oxygen used in the metabolism of biodegradable organics is termed biochemical oxygen demand. When organic matter decomposes, microorganisms such as bacteria and fungi feed upon it and eventually it becomes oxidized. Biochemical oxygen demand (BOD) is a measure of the quantity of oxygen used by these microorganisms in the aerobic oxidation of organic matter.

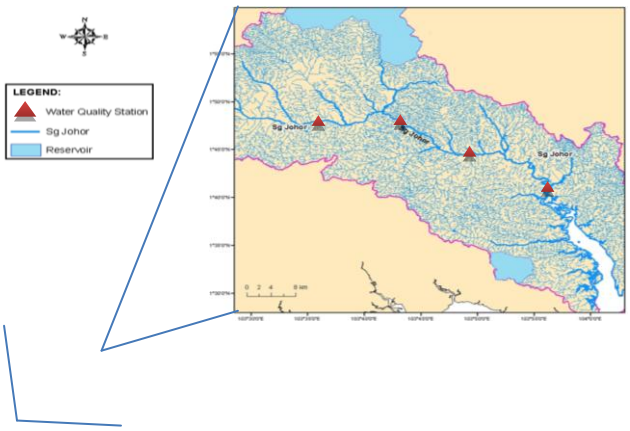

Figure 1 Map showing the geographical setting of the survey area with four field monitoring stations on the main stream.

2.2. Support Vector Machine model SVM

SVM possess great potential and superior performance as is appeared in many previous researches. This is largely due to the structural risk minimization (SRM) principle in SVM that has greater generalization ability and is superior to the empirical risk minimization (ERM) principle as adopted in neural networks. In SVM, the results guarantee global minima whereas ERM can only locate local minima. For example, the training process in neural networks, the results give out any number of local minima that are not promised to include global minima. Furthermore, SVM is adaptive to complex systems and robust in dealing with corrupted data. This feature offers SVM a greater generalization ability that is the bottleneck of its predecessor, the neural network approach.

Considering a set of training data $\{(x_1, y_1),....,(x_\ell, y_\ell)\}$, where each $x_i \subset R^n$ denotes the input space of the sample and has a corresponding target value $y_i \subset R$ for i=1,…, l where l corresponds to the size of the training data [16,17]. The idea of the regression problem is to determine a function that can approximate future values accurately.

The generic SVR estimating function takes the form:

$$f(x) = (w \cdot \Phi(x)) + b \tag{1}$$

where $w \subset R^n$, $b \subset R$ and Φ denotes a non-linear transformation from R^n to high dimensional space. Our goal is to find the value of w and b such that values of x can be determined by minimizing the regression risk:

$$R_{reg}(f) = C\sum_{i=0}^{\ell}\Gamma(f(xi) - yi) + \frac{1}{2}\|w\|^2 \tag{2}$$

where $\Gamma(\cdot)$ is a cost function, C is a constant and vector w can be written in terms of data points as:

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\Phi(x_i) \tag{3}$$

By substituting equation (3) into equation (1), the generic equation can be rewritten as:

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)(\Phi(x_i)\cdot\Phi(x)) + b$$
$$= \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{4}$$

In equation (4) the dot product can be replaced with function $k(x_i, x)$, known as the kernel function. Kernel functions enable dot product to be performed in high-dimensional feature space using low dimensional space data input without knowing the transformation Φ. All kernel functions must satisfy Mercer's condition that corresponds to the inner product of some feature space. The radial basis function (RBF) is commonly used as the kernel for regression

Performance Criteria

Due to the fact that water parameters had been truthfully monitored over the 5-year period, the performances of the proposed models could be examined and evaluated. The performances of the models were evaluated according to three statistical indexes. Coefficient of Efficiency (CE) is often used to evaluate the model performance, introduced by [18].

$$CE = 1 - \frac{\sum_{i=1}^{n}(X_m - X_p)^2}{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2} \tag{5}$$

where n is the number of observations, X_p and X_m are the predicted and measured parameter, respectively, and $\overline{X}_m$ is the average of measured parameter. The Mean Square Error (MSE) can be used to determine how well the network output fits the desired output. The smaller values of MSE ensure better performance. It is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_m - X_p)^2 \tag{6}$$

The coefficient of correlation (CC) is often used to evaluate the linear relationship between the predicted and measured parameter. It is defined as follows:

$$CC = \frac{\sum_{i=1}^{n}(X_m - \overline{X}_m)(X_p - \overline{X_p})}{\sqrt{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2 \sum_{i=1}^{n}(X_p - \overline{X_p})^2}} \quad (7)$$

RESULTS AND DISCUSSION

It is essential to determine approximate values of optimal hyper parameters C, ε and γ. In fact, there is no formal and/or mathematical method for determining the appropriate "optimal set" of the key parameters of Neural Network. Therefore, it was decided to perform this task utilizing the trial and error method. Different ranges of kernel hyper parameters c, ε and γ were adopted. For instance, the range of C was set to [1-10] at increment 0f 1.0 and [0.1 - 0.5] at increment of o.1 for ε and γ. The optimal values of hyper parameters are selected based on 10-fold cross-validation repeated ten times until reach the optimal result. Table (1) demonstrates the best result for the training and prediction data sets. Apparently, the results show that the performance of SVM is sensitive to the hyper parameters, where SVM model that used to predict the BOD it reach the best results when C equal to 8, ε equal to 0.3 and γ equal to 0.16 respectively. In the SVM theory, there are different types of kernel functions can be used, such as linear, polynomial, RBF, and sigmoid. In this paper we study the result of the use of compactly supported RBF kernels. RBF kernels are frequently used in nonlinear function estimation problems [19].

Table 1 The optimal SVM parameters for training and testing data sets

Par.	Input	SVM parameters			Trn		Tst	
		C	ε	γ	CC	CE	CC	CE
BOD	5	8	0.3	0.16	0.971	0.942	0.956	0.913

Plots of residuals versus predicted BOD could be more useful regarding model fitting to a data set. If the residuals appear to behave randomly it suggests that the model fit the data well. In contrast, if non-random distribution is evident in the residuals, the model does not fit the data effectively (Singh et al, 2009). Residuals versus predicted BOD are shown in Figure (2). The observed relationships between residuals and model predicted for BOD shows complete independence and random distribution. It obvious from the figure that the markers are well distributed on both sides of the horizontal line of zero coordinates which it represents the mean of the residuals and the respective correlation BOD equal to 0.019 which is slight small.

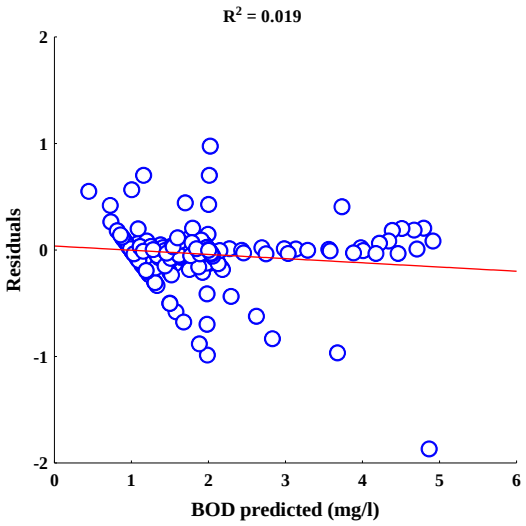

Figure 2: demonstrates of the residuals versus predicted BOD

In the systematic analysis, the performance evaluation of various possible combinations of the parameters was investigated utilizing Coefficient of Efficiency (CE) and Mean Square Error (MSE) approaches to determine the most effective parameters on the output. In the analysis, performance evaluation of various possible combinations of variables was investigated. Overall, all networks for BOD was compared, as shown in Table (2). Findings of the sensitivity analysis showed that nitrate was found to be the effective parameter when add to the model for all parameters. The MSE value becomes smaller when more combination variables were used together. On the basis of the performance evaluation of combinations of input variables, best group performances according to number of parameters equal to five. The respective MSE values, as given in Table (2) show that MSE values decrease as the number of variables in the group increases.

Table 2 Evaluation of possible combinations of input variables

Combination					BOD	
					MSE	CE
COND					20	0.32
COND	TEMP				0.69	0.69
COND	TEMP	NO3			0.22	0.71
COND	TEMP	NO3	CI		0.24	0.76
COND	TEMP	NO3	CI	Na	**0.07**	**0.91**

For further visualization of the proposed SVM model performance, a demonstration of the comparison of SVM model versus the measured BOD during testing data set is shown in Figure (3). It is obvious that the proposed SVM model able to mimic the pattern (dynamics) in the measured values, in addition, for those extreme and low values experienced during this set. The proposed model showed efficiency in predicting the concentration of BOD in the Johor

River, and it was compatible with the results of other researchers/authors. The results also indicated that the proposed model was basically an attractive alternative, offering a relatively fast algorithm with good theoretical properties to predict the water quality parameters and can be extended to predict different water quality parameters.

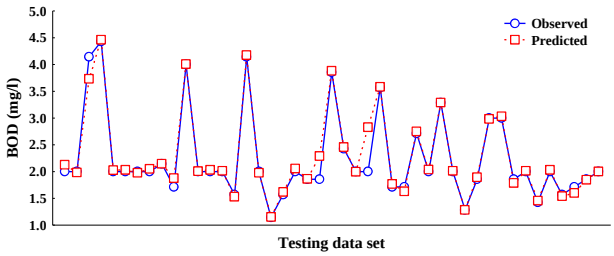

Figure 3: demonstration of the comparison of SVM model versus the measured BOD during testing data set

CONCLUSION

The Artificial Intelligence (AI) is a new technique with a flexible mathematical structure that is capable of identifying complex non-linear relationships between input and output data when compared to other classical modeling techniques. In this study, Support Vector Machine (SVM) was performed to identify the optimal prediction for water quality parameter along the Johor River basin. SVM improved the precision of BOD prediction with minimal computation. In general, this research work has managed to adopt SVM method that would prove to be useful for various institutions that are directly involved in the management of river basins in Malaysia. Moreover, the tools used in this work could form a basis for a more effective decision making process on the part of the policy makers in order to help maintain and improve the management of river basins.

ACKNOWLEDGMENT

The authors wish to thank the Department of Environment for providing the required data for this research. This research was supported by the research grants Universiti Malaysia Terengganu GGP 68007/2013/119.

REFERENCES

[1] El-shafie AE, Noureldin M, Taha R, Basri H (2008). Neural network model for Nile River inflow forecasting analysis of historical inflow data. Journal of Applied Sciences 8 (24), 4487-4499.

[2] El-shafie A, Mukhlisin M, Najah AA, Taha MR (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. International Journal of the Physical Sciences Vol. 6(8), pp. 1997-2003, 18 April,

[3] Liong SY, Lim WH, Paudyal G (1999). Real time river stage forecasting for flood Bangladesh: neural network approach. Journal of Computing in Civil Engineering ASCE 14 (1), 1 8.

[4] Najah A, Elshafie A, Karim OA, Jaffar O (2009) Prediction of Johor River water quality parameters using artificial neural networks. Eur J Scienti Res 28(3):422–435 ISSN 1450-216X

[5] Najah, A., El-Shafie, A, Karim OA. and Jaffer, O. (2010a). Water Quality Prediction Model Utilizing Integrated Wavelet-ANFIS Model with Cross Validation. Neural Computing and Applications journal.DOI: 10.1007/s00521-010-0486-1

[6]Najah AA, El-Shafie A, Karim OA, Jaafar O (2010b): Water Quality Prediction Model Utilizing Integrated Wavelet-ANFIS Model with Cross Validation, Neural. Comput. Appl., in press, doi:10.1007/s00521-010-0486-1.

[7] Najah A, Elshafie A, Karim OA, Jaffar O (2010c) Evaluation the efficiency of radial basis function neural network for prediction of water quality parameters.Eng Int Syst 4: 221–231

[8] Muttil, N, Chau, KW (2006). Neural network and genetic programming for modelling coastal algal blooms. International Journal of Environment and Pollution 28 (3/4), 223–238.

[9] Razavi SV, Jumaat MZ, Ahmed H EI-Shafie, Mohammadi P (2011). General regression neural network (GRNN) for the first crack analysis prediction of strengthened RC one-way slab by CFRP International Journal of the Physical Sciences Vol. 6(10), pp. 2439-2446

[10] Razavi SV, Jumaat MZ, Ahmed H EI-Shafie (2011). Using feed-forward back propagation (FFBP) neural networks for compressive strength prediction of lightweight concrete made with different percentage of scoria instead of sand. International Journal of the Physical Sciences Vol. 6(6), pp. 1325-1331

[11] Kasabov NK (1996) Foundations of neural networks, fuzzy systems, and knowledge engineering MIT Press Cambridge.

[12] Grubert JP (2003). Acid deposition in the eastern United States and neural network predictions for the future. Journal of Environmental Engineering and Science 2 (2), 99–109.

[13] Chen Q, Mynett AE (2003). Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecological Modelling 162 (1/2), 55–67.

[14] Lee JHW, Wong KTM, Huang Y, Jayawardena AW (2000). A real time early warning and modeling system for red tides in Hong Kong. In: Wang, Z.Y. (Ed.), Proceedings of the Eighth International Symposium on Stochastic Hydraulics. Balkema, Beijing, pp. 659–669.

[15]Reckhow KH (1999). Water quality prediction and probability network models. Canadian Journal of Fisheries and Aquatic Sciences 56, 1150–1158.

[16] Muller KR, Smola A, Ratch G, Scholkopf B, Kohlmorgen J, Vapnik V (2000). Using Support Vector Support Machines for Time Series Prediction, Image Processing Services Research Lab, AT&T Labs

[17] Vapnik VN (1995). The Nature of Statistical Learning Theory. Springer, New York

[18] Nash JE, Sutcliffe JV (1970). River flow forecasting through conceptual models. Part 1: a discussion of principles. Journal of Hydrology 10(3): 282–290.

[19]Evgeniou T, Pontil M, Poggio T (2000). Regularization networks and support vector machines", Advances in Computational Mathematics, 13(1), 1-50.