World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:8, No:1, 2014

Dissolved Oxygen Prediction Using Support Vector Machine

Sorayya Malek, Mogeeb Mosleh, Sharifah M. Syed

Open Science Index, Bioengineering and Life Sciences Vol:8, No:1, 2014 publications.waset.org/9997153/pdf

***Abstract*—**In this study, Support Vector Machine (SVM) technique was applied to predict the dichotomized value of Dissolved oxygen (DO) from two freshwater lakes namely Chini and Bera Lake (Malaysia). Data sample contained 11 parameters for water quality features from year 2005 until 2009. All data parameters were used to predicate the dissolved oxygen concentration which was dichotomized into 3 different levels (High, Medium, and Low). The input parameters were ranked, and forward selection method was applied to determine the optimum parameters that yield the lowest errors, and highest accuracy. Initial results showed that pH, Water Temperature, and Conductivity are the most important parameters that significantly affect the predication of DO. Then, SVM model was applied using the Anova kernel with those parameters yielded 74% accuracy rate. We concluded that using SVM models to predicate the DO is feasible, and using dichotomized value of DO yields higher prediction accuracy than using precise DO value.

***Keywords*—**Dissolved oxygen, Water quality,predication DO, Support Vector Machine.

I. Introduction

EUTROPHICATION is a common global concern in lakes and reservoir. Current status of eutrophication in Malaysian lakes and reservoirs are indicated to be more than 60% [1],[2]. The effects of eutrophication are deterioration of water quality for human consumption, limitation of recreational usage and exhaustion of dissolved oxygen (DO) below tolerable level which induces reductions in specific fish and other animal populations [3]. The concentration of DO is important for the healthy functioning of aquatic ecosystems and is a significant indicator of the state of aquatic ecosystems. DO concentration is a parameter commonly used to assess the water quality in different reservoirs and watersheds. DO concentration is strongly influenced by a combination of the physical, chemical, and biological characteristics of the streams' oxygen-demanding substances, including algal biomass, dissolved organic matter, ammonia, volatile suspended solids, and sediment oxygen demand [4]-[9].

Machine learning methods such as artificial neural networks (ANNs) has been successfully implemented for DO estimation [10]-[12]. ANN is a well-suited method with self-adaptability, self-organization, and error tolerance for nonlinear simulation. However, this method has limitationsdue to its complex structure that requires a great amount of training data, difficulty in tuning the structure parameter that is mainly based on experience, not sensitive to the noise produce and its "black box" nature that makes it difficult to understand and interpret the data [13], [14].

Considering the drawbacks of ANN, recently support vector machine (SVM) are become widely used for predicting the DO concentration [15], [16]. It is a new machine-learning technology based on statistical theory and derived from instruction risk minimization, which can enhance the generalization ability and minimize the upper limit of generalization error. Compared to ANN, SVM has advantages of dealing with complex and highly nonlinear data, only requiring a small amount of samples, high degree of prediction accuracy, and long prediction period if kernel function was used to solve the nonlinear problems of water quality modeling [13].

In this study, we attempted to develop an SVM-based predictive model to predict dissolved oxygen concentration using data from selected freshwater lake in Malaysia. The data samples for two lakes from 2005 to 2009 were used to train and test the model. Dissolved oxygen concentrations used are dichotomized because different threshold value of DO concentration can represent either acceptable or unacceptable water quality in a lake. This article is organized as follows; section II discusses the methodology which includes the study area, data collection, analysis, and SVM model development. Section III represents the results followed by the discussion and conclusion.

II. Methodology

A. Study Area

Water quality data were collected from two fresh water lakes in peninsular Malaysia. They are Bera and Chini Lakes as illustrated in Figs. 1 and 2. Lake Bera consists of 26,000 hectares of its core zone and 27,500 hectares of the buffer zone all has been preserved as RAMSAR sites and it is coordinate at 3°49′00″N102°25′00″E.Lake Chini consists of about 5026 hectares and is located in the coordinate 3°26′N102°55′E. Lake Chini and Lake Bera both serves as wetlands to prevent the floods and erosion of riverbank and serve as an important water reservoir in Malaysia. However due recent increase in pollution level the water qualities at these lakes are deteriorating [17].

Sorayya Malek is with the Institute of Biological Sciences, Faculty of Science University Malaya (phone: +603-79676748; fax: +603-79674208; e-mail: sorayya@um.edu.my).

Mogeeb Mosleh was with Taiz University, Yemen. He is now with the Department of A.I, Faculty of Computer Science & Information Technology, University of Malaya, multivariate statistics, KualaLumpur, Malaysia (e-mail: MogeebMosleh@um.edu.my).

Sharifah M. Syed is with Department of Computer & Communication System Engineering University Putra Malaysia (e-mail: s_mumtazah@upm.edu.my).

Open Science Index, Bioengineering and Life Sciences Vol:8, No:1, 2014 publications.waset.org/9997153/pdf

B. Data Collection and Analysis

Water quality data from six monitoring stations from Bera lakes, and nine monitoring stations from Chini Lake were used in this study. Bimonthly water quality data from February 2005 until October 2009 was used.



Fig. 1 Lake Bera Map



Fig. 2 Lake Chini Map

The water quality data include 147 row samples for 11 parameters. Sampling procedures including preservation for water quality parameters were carried out in accordance with WHO[18] and the analytical methods for the measured parameters were adopted from manual published APHA [19]. Table I list the summary statistics of the parameters used in this study.

TABLE I
STATISTICS SUMMARY OF DATA PARAMETERS FROM LAKE CHINI AND LAKE BERA (2005 – 2009)

No.	Limnological Variables	Mean	Minimum	Maximum
1.	Water temperature ^{o}C	30.245	26.12	34.29
2.	Turbidity (NTU)	17.554	0.4	282.2
3.	pH	6.596	4.97	7.94
4.	DOmg/L	M	L	H
5.	Conductivity uS/cm	40.17	14	190
6.	Salinityppt	0.019	0.01	0.09
7.	Nitrate (NO3-N) mg/L	0.232	0.01	2.04
8.	Phosphate PO4mg/L	0.063	0	0.36
9.	Ammonia(NH3) mg/L	0.085	0.01	0.81
10.	*E-coli*mg/L	733.12	0	5400
11.	Total Coliformmg/L	8885.7	20	284000

The parameter "DO" was classified into three levels ranges from low to high. This is because water quality can be classified into different classes depending on DO concentration range. The Interim National Water Quality Standard, Malaysia (INWQS) and Department of Environment have classified water quality based on DO concentration. Table II, illustrates the DO concentration and water quality classes.This classification of DO range has been adapted in this study.

TABLE II
CLASSIFICATION OF WATER QUALITY BASED ON DO CONCENTRATION (INWQS)

DO Value	INWQS water Classification
DO < = 5	Class III – IV Extensive treatment required
5 > DO <7	Class II - Suitable for recreational use and conventional treatment required.
Do > = 7	Class I -Conservation of natural environment, Water Supply practically no treatment necessary.

The Do used in this study for SVM model development is classified into three categorizations (high, medium, and low) similar with the classification of the Interim National Water Quality Standard, Malaysia (INWQS) and Department of Environment as shown in Table III. In this classification the high class of DO range from value 7 and above, while the medium class ranges between 5 until 7, and lastly for the DO that had value below 5 is in class low.

TABLE III
OUR CLASSIFICATION FOR THE "DO" CONCENTRATIONS

DO Value	DO Categorizations
DO < = 5	Low
5 > DO <7	Medium
Do > = 7	High

C. SVM Model Data Development

The SVM was implemented in this study using the R Package software. Kernel-based Machine Learning Lab (Kernlab) package was used to develop our SVM model. It is an extensible package for kernel-based machine learning methods in R. This package contains dot product primitives (kernels), implementations of support vector machines and the relevance vector machine with many other kernel methods and

Open Science Index, Bioengineering and Life Sciences Vol:8, No:1, 2014 publications.waset.org/9997153/pdf

algorithms [20].

SVM works by predicting the labels of training data as$D = \{(\overrightarrow{x_i}, y_i), i = 1..N\}$ with $y_i \in \{-1, +1\}$ is separable by a hyperplane. Where if the data is positive it belongs to class 1 and if the data is negative it belongs to class -1. When data is linearly separable and support vector existed it can be derived as:

$$f(\vec{x}) = \vec{w}^T\vec{x} + b(= w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b)$$

W and b are determined during training process where it locates the support vector. W is defined as weight vector while b is term for bias and is a scalar. T stands for transpose operator.

In SVM kernel function is introduced as it offer a better performance when dealing with a nonlinear data. Here data is mapped and derived as below where K is the kernel function:

$$\Phi(x)^T\Phi(y) = K(x, y)$$

By introducing the kernel it can be derived as

$$\sum_i a_i - {}^1\!/_2 \sum_i \sum_j a_i a_j y_i\, y_j K(x_i x_j)$$

subject to:

$$\sum_{i=1}^{N} a_i y_i = 0,$$
$$0 \le a_i \le C, \ \ \forall i$$

The equation produce is as below

$$f(x) = w^T\Phi(x_i) + b = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b$$

The kernel selection depends on the data distribution but kernel selection also generally done through trial and error [21], [22].

Prior to model development data was divided into training and testing data set; 80% for training data and 20% for testing data. All data was converted into comma delimited (CSV) format before it can be used to run in the R software. Input parameters were ranked using linear kernel in ascending order based on the cross validation errors. This is carried out to determine effects of each input parameter on the output parameter DO. Once the effect of each parameter has been determined, SVM prediction models using RBF, ANOVA and Laplace kernel was developed. The default value was used for each kernel for the SVM prediction model development.Forward elimination method was used for each SVM model to eliminate less significance parameters.

Manual calculations are also carried out to determine the accuracy for each class of DO predication using the formula below for training and testing data.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

True positive (TP) is represented the correct prediction state, where false negative (FN) is represented the wrong predication state. True negative (TN) is class which not wrongly predicted as target class and false positive (FP) is class that is wrongly predict and is predicted on target class.

III. Results

Forward selection is used to determine optimum number of parameter for DO SVM prediction model. Results of SVM model that yields the highest accuracy rate using various kernels with optimum number of parameters is shown in Table IV. It illustrates the results of the input parameters for the selection process. This table is yield from the process of selection, forwarding, and ranking based on the least errors value obtained.

TABLE IV
Input Parameter Ranking

Ranking	Parameter	Cross validation Error
1	Temperature	0.458333
2	pH	0.516667
3	Conductivity	0.575
4	Total Coliform	0.608333
5	Turbidity	0.625
6	Ammonia	0.675
7	Salinity	0.691667
8	Phosphate	0.7
9	*E.coli*	0.708333
10	Nitrate	0.8

Selection procedure was repeated, until the SVM model with kernel that has the highest accuracy rate was determined automatically using a built in function in R. The results of SVM prediction model with highest accuracy rate and optimum number of parameters for each kernel is presented in Table V.

TABLE V
SVM Prediction Model Result for "DO"

Input Parameter	**Kernel**	**Accuracy (%)**
Water Temperature, pH	Radial Basis	44.44
Water Temperature, pH, Conductivity.	Anova	74.07
Water Temperature, pH, Conductivity., Coliform	Anova	62.96
Temperature, pH, Conductivity., Coliform, Turbidity	Anova	70.37
Temperature, pH, Conductivity., Coliform, Turbidity, Ammonia	Laplacian	66.67
Temperature, pH, Conductivity., Coliform, Turbidity, Ammonia, Salinity	Anova	42
Temperature, pH, Conductivity., Coliform, Turbidity, Ammonia, Salinity, Phosphate	Anova	48. 14
Temperature, pH, Conductivity., Coliform, Turbidity, Ammonia, Salinity, Phosphate, *E.coli*	Anova	44.44
Temperature, pH, Conductivity., Coliform, Turbidity, Ammonia, Salinity, Phosphate, *E.coli*, Nitrate	Radial Basis	51.85

After we determined the appropriate parameters that have the most effects on the predication results of DO, we isolated those parameters (Water Temperature, pH, and Conductivity)

in individual comma separated file. Then we used the test data as input to the SVM model. 27 test data samples were run to predicate the DO level using the SVM. Finally, we calculate the accuracy of predication manually as shown in Table VI.

TABLE VI
SVM ACCURACY FOR DO PREDICATION RESULTS

	DO Level Accuracy (%)		
Data Classification	Low	Medium	High
Training	78.33	72.5	75.83
Testing	81.48	77.78	88.89

As shown in Table VI, we calculated the accuracy of predication for the both data sets, training and testing data. Then we calculated manually the accuracy of each DO classification.

IV. DISCUSSION

In the SVM theory, there are different types of kernel functions can be used, such as linear, polynomial, RBF, ANOVA and sigmoid. Different base functions are applicable for dealing with different types of data. ANOVA kernels have been shown to work rather well in multi-dimensional regression problems [21],[23]. The SVM accuracy using the ANOVA kernel gave more accurate predication results if compared with other kernels which used in this study such as RBF and laplacian kernels. The highest accuracy obtained using ANOVA kernel with the three selected parameters was 74%.Forward selection method was used to select parameters that yield the optimum results. The parameters were pH, conductivity and water temperature. DO in water is primarily determined by water temperature [6].

Water temperature also has a reverse relationship with DO in water and proportional relationship with hydrogen ion concentration (pH). Conductivity meanwhile is related to salinity which affects the DO level in water. Water that has high salinity has high conductivity which decreases DO level in water. Salinity also decreases the DO saturation value Salinity enables conductivity affects DO level in water.

Similar studies on DO prediction using SVM has reported accuracy level of 70% [13],[24]-[27]. The success rate of DO prediction in this study is reported to be 74% using dichotomous values of DO instead of precise values as used in other studies. Many ecological responses are difficult to measure accurately and definitely. Furthermore there are many factors that need to be considered when modeling water quality such as climate factors and development activities in the surrounding areas. Therefore characterizing responses that are dichotomous such DO for lakeeutrophication level increases the accuracy of prediction in this study.

V. CONCLUSION

SVM is considered recently one of the most suitable approaches for predication and forecasting. The preparation approaches such as selection, forwarding, and ranking that applied to determine the significant water quality parameters can play essential roles in the prediction of DO. In addition, results showed that *ANOVA* is the most appropriate kernel to obtain the highest accuracy in predication results because of its ability in reducing data redundancy of nonlinear data. In this study the prediction of DO level by using the SVM for the Lake Bera and Lake Chini produced accuracy about 74.07%. The acceptable accuracy results showed that predication of the DO level using dichotomized value isfeasible, and more accurate than predicating exact value of DO.

ACKNOWLEDGMENT

The authors would like to thank National Hydraulic Research Institute of Malaysia (Nahrim) for their data and UMRG grant RG241-12AFR.

REFERENCES

[1] Nahrim: A desktop Study on the Status of Lake Eutrophication in Malaysia, *Final Report*, Malaysia; 2005.
[2] Z. Sharip, and Z. Yusop, "National overview: the status of lakes eutrophication in Malaysia." pp. 2-3.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
[3] S.O. Ryding, and W. Rast, "Control of eutrophication of lakes and reservoirs," *Manual the biosphere series*, vol. 1, 1989.
[4] M. Spanou, and D. Chen, "An object-oriented tool for the control of point-source pollution in river systems," *Environmental Modelling& Software*, vol. 15, no. 1, pp. 35-54, 2000.
[5] M. D. Williams, and M. Oostrom, "Oxygenation of anoxic water in a fluctuating water table system: an experimental and numerical study," *Journal of hydrology*, vol. 230, no. 1, pp. 70-85, 2000.
[6] J. Kalff, Limnology: inland water ecosystems: Prentice Hall New Jersey, 2002.
[7] B. Cox, "A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers," *Science of the Total Environment*, vol. 314, pp. 335-377, 2003.
[8] P. J. Mulholland, J. N. Houser, and K. O. Maloney, "Stream diurnal dissolved oxygen profiles as indicators of in-stream metabolism and disturbance effects: Fort Benning as a case study," *Ecological Indicators*, vol. 5, no. 3, pp. 243-252, 2005.
[9] N. W. T. Quinn, K. Jacobs, C. W. Chen, and W. T. Stringfellow, "Elements of a decision support system for real-time management of dissolved oxygen in the San Joaquin River Deep Water Ship Channel," *Environmental Modelling& Software*, vol. 20, no. 12, pp. 1495-1504, 2005.
[10] E. Dogan, B. Sengorur, and R. Koklu, "Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique," *Journal of Environmental Management*, vol. 90, no. 2, pp. 1229-1235, 2009.
[11] A. Akkoyunlu, and M. E. Akiner, "Pollution evaluation in streams using water quality indices: A case study from Turkey's Sapanca Lake Basin," *Ecological Indicators*, vol. 18, pp. 501-511, 2012.
[12] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water quality—a case study," *Ecological Modelling*, vol. 220, no. 6, pp. 888-895, 2009.
[13] M. Bouamar, and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification." pp. 1047-1050.
[14] X. Yunrong, and J. Liangzhong, "Water quality prediction using LS-SVM and particle swarm optimization." pp. 900-904
[15] S. Liu, J. Ren, and W. You, "A Study on Purification of the Eutrophic Water Body with Economical Plants Siollessly Cultivated on Artificial Substratum [J]," ActaScicentiarumNaturalumUniversitisPekinesis, vol. 4, 1999.
[16] A. Najah, A. El-Shafie, O. Karim, O. Jaafar, and A. H. El-Shafie, "An application of different artificial intelligences techniques for water quality prediction," *Int. J. Phy. Sci*, vol. 6, no. 22, pp. 5298-5308, 2011.
[17] M. Shuhaimi-Othman, E. C. Lim, and I. Mushrifah, "Water quality changes in Chini Lake, Pahang, West Malaysia," *Environmental Monitoring and Assessment*, vol. 131, no. 1-3, pp. 279-292, 2007.
[18] WHO:UNEP/WHO/UNESCO/WMO Project on Global Environmental Monitoring.*GEM Water Operational Guide* 1987.

Open Science Index, Bioengineering and Life Sciences Vol:8, No:1, 2014 publications.waset.org/9997153/pdf

[19] American Public Health Association (APHA): *Standard methods for the examination of water and waste water*. 19th edition. American Water Works Association (AWWA) and Water Environment Federation APHA, Washington, DC; 1995.
[20] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an S4 package for kernel methods in R," 2004.
[21] M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston, "Support vector regression with ANOVA decomposition kernels," *Advances in kernel methods—Support vector learning*, pp. 285-292, 1999.
[22] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller, "Engineering support vector machine kernels that recognize translation initiation sites," *Bioinformatics*, vol. 16, no. 9, pp. 799-807, 2000.
[23] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171-1220, 2008.
[24] R. V. Thomann, and J. A. Mueller, Principles of surface water quality modeling and control: Harper & Row, Publishers, 1987.
[25] A. Najah, A. El-Shafie, O. Karim, and O. Jaafar, "Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations," *Hydrology and Earth System Sciences Discussions,* vol. 8, no. 3, pp. 6069-6112, 2011.
[26] X.-S. Qin, G. H. Huang, G.-M. Zeng, A. Chakma, and Y. Huang, "An interval-parameter fuzzy nonlinear optimization model for stream water quality management under uncertainty," *European Journal of Operational Research*, vol. 180, no. 3, pp. 1331-1357, 2007.
[27] J. Liu, M. Chang, and X. Ma, "Groundwater Quality Assessment Based on Support Vector Machine." pp. 173-178.

Open Science Index, Bioengineering and Life Sciences Vol:8, No:1, 2014 publications.waset.org/9997153/pdf