WATER QUALITY MONITORING BASED ON SUPPORT VECTOR MACHINE AT PUTRAJAYA LAKE AND WETLAND

ZAKIRAH BINTI MOHD AZAMLI
(SGJ 100013)

SUBMITTED TO
INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA

IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF BIOINFORMATICS
2012

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Zakirah Binti Mohd Azamli (I.C/Passport No:) 861021-29-5544

Registration/Matric No: SGJ 10013

Name of Degree: MASTER OF BIOINFORMATICS

TITLE("this Work"): Water quality monitoring based on support vector machine at Putrajaya Lake and Wetland.

Field of Study:Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any Copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:22 June 2012

Subscribed and solemnly declared before,

Witness's Signature Date: 22 June 2012

Name:
Designation:

ACKNOWLEDGEMENT

All praises and thanks be to Allah (S.W.T), who has guided and giving me this opportunity to accomplished the project. Words cannot express my gratitude towards my supervisor, Dr. Sorraya Bibi Malek and my co-supervisor Dr. Sharifah Mumtazah for their patience, humble supervision advice I received from them in the course of his project. I would like to thank you my family and friends that always there for me when I need them. My acknowledgment will be incomplete if I keep mum on support and help I received from my all other colleagues, Siti Fatihah and Awanis. Lastly, I would like to thank to all the lectures of the Department of Bioinformatics, University Malaya for the guidance while I pursued this degree.

ABSTRAK

Model ‘support vector machine’ (SVM) adalah salah satu teknologi ‘machine learning’. Ia digunakan untuk mengklasifikasikan data dengan mengenal pasti paten data tersebut. Ia digunakan dengan menjalankan ‘train’ ke atas data dari stesen persampelan data untuk meramalkan kualiti air dari Tasik Putrajaya dan Wetlandnya berdasarkan 29 stesen persampelan di kawasan kajian. Berikut adalah tujuh parameter yang digunakan untuk klasifikasi SVM iaitu oksigen terlarut (DO), suhu air, pH air, saliniti air, keperluan oksigen biokimia (BOD) dan *Escherichia coli* (*E. coli*). Oksigen terlarut digunakan sebagai indikasi kepada standard pengukuran kualiti air. Tujuan kajian ini untuk menentukan oksigen terlarut berdasarkan tiga paras kelas kualiti iaitu, ‘High’, ‘Medium’ dan ‘Low’. Semua data akan menjalani ‘training’ dan juga ‘testing’. Fungsi kernel RBF akan digunakan untuk ‘train’ data. Keputusan menunjukan ‘cross validation error’ adalah 0.304762. Manakala keputusan untuk ‘sensitivity’ dan ‘specificity’ juga ditentukan.

ABSTRACT

A support vector machine (SVM) model is a machine learning technology. It is used for the proposed to classify the data by recognize the patterns of data. It is by work by train the data of sampling area to predict the water quality of Putrajaya Lake and Wetland based from data of water quality parameters of 29 sampling stations in study area. The following seven water quality parameters were used for the proposed of SVM classification, namely; dissolved oxygen (D.O), temperature (Temp), water pH, salinity, Biochemical Oxygen Demand dan *Escherichia coli* (*E. coli*). The dissolved oxygen variable is being used as indicator of Putrajaya Lake and Wetland water quality measurements standard. The propose of this study is to predict the dissolve oxygen based on three level of class of water quality namely; High, Medium and Low. The data were undergoing training and testing. RBF kernel function is employed to train the data. The result shows that cross validation error is 0.304762. The optimal cost C and sigma value are 0.7 and 0.5. Meanwhile, the numbers of support vector on the other hand are about 90. The resulting from sensitivity and specificity is also determined.

TABLE OF CONTENTS

TITLE	PAGE
ACKNOWLEDGEMENT	ii
ABSTRAK	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix

TABLE OF CONTENTS

	TITLE	PAGE
1.0	INTRODUCTION	
1.1	OBJECTIVES	1
1.2	PROJECT SCOPE	3
1.3	BACKGROUND OF PUTRAJAYA LAKE AND WETLAND	6
1.4	SUPPORT VECTOR MACHINE (SVM)	7
2.0	LITERATURE REVIEW	
2.1	DISSOLVED OXYGEN AND ITS DEPENDENCY	14
2.2	APPLICATION IN CLASSIFICATION AND PREDICTION OF WATER QUALITY	18
3.0	MATERIALS AND METHODOLOGY	
3.1	STUDY AREA AND SAMPLING STATIONS	19
3.2	WATER QUALITY PARAMETER	20
3.3	SOFTWARE	21
3.4	METHODOLOGY	
3.4.1	CATALOGUE DATA	22
3.4.2	CONVERSION OF DATA TO SVM'S SOFTWARE FORMAT	22
3.4.3	SVM PROCEDURE	23

4.0 RESULT 27

5.0 DISCUSSION 31

6.0 CONCLUSION 36

7.0 REFERENCES 37

8.0 APPENDIX

8.1 CODING CODE 38

8.2 DATA RECORDS 42

LIST OF TABLES

Table 1: The ranking of the cross-validation error rate.

Table 2: A result to decide the input of parameters.

Table 3: Summary of the result obtained from the analysis.

Table 4: The total Prediction and actual water quality for each class.

Table 5: Predicted result in water quality for each class.

Table 6 False prediction and Actual prediction on data

Table 7: Sensitivity and specificity of the result.

LIST OF FIGURES

Figure 1: Sampling stations of water quality in Putrajaya Lake and Wetland

Figure 2: Putrajaya Layout including the lake and wetland.

Figure 3: Mapping the data from input space to high dimension space

Figure 4: Hyperplane through two linearly separable classes

Figure 5: Mapping linearly separable data into high dimension feature space ($x \rightarrow \phi(x)$)

Figure 6: Support vector machine architecture

Figure 7: Dissolved oxygen concentration.

Figure 8: Water temperature

Figure 9: The dissolved oxygen interaction in a water body, showing the decay (satisfaction) of carbonaceous, nitrogenous and sediment oxygen demands and water body re-aeration or de-aeration.

Figure 10: Nonlinear Mapping from sample space to high dimension future space.

Figure 11: SVM prediction and actual result.