CHAPTER 1

INTRODUCTION

1.1 Objectives

The vital part of water resource management is monitoring the water quality. Water quality may refer to the state of the water, including its chemical, physical and biological characteristics.It is fundamental to sustain the ecological processes and also in people's daily life where the purpose of using it for irrigation, watering stock, drinking, fishing and recreation, and to meet cultural and spiritual needs.

Proper water management will enable the everlasting use of water as well as allowing us to keep track anthropogenic influences. The fast paced industrialization and the ever increasing economic activities together with the need of more housing projects, environment pollution, especially water is getting more serious. Trying to analyze the pattern, study and predict the changes of water quality is of utmost important in trying to resolve and monitor water pollution in water reservoir such as lake. Since all those attributes directly to the people daily life, it is not only proper for the people but also the environment protection division to monitor the water quality of the local area and to find and solve pollution problem effectively.

The studies of the Putrajaya catchment area showed that alarming levels of pollutant are being channeled from upstream sources as well as outside the Putrajaya development

boundary. Understandably with the future and ongoing development of Sg. Chuau catchment will for sure increase run-off and pollutant concentration which will eventually drain into Putrajaya Lake.

In Putrajaya man made wetland is being constructed as a natural treatment system to treat upstream inflow to the lake. But even that, to check the water quality standard, monitoring is still required as well as to identifying sources and loads of pollutants that are causing these declines.

With this problem, the project is needed to predict the classification of the water quality of the Putrajaya Lake and Wetland based on dissolved oxygen (D.O.) level. The D.O. is categorized into three groups; High, Medium and Low. Besides, the accuracy of the prediction is needed to determine. The constructed SVM model can be test for its accuracy on dissolved oxygen prediction. By then the safety water quality of the lake and wetland can be identify at safe level as classified by The Putrajaya Lake Water Quality Standards.

1.2 Project Scope

This water quality monitoring project is using Support Vector Machine (SVM) model in R programming language. A Support Vector Machine is a Kernel-based technique that used to perform classification and regression analysis. The theory of SVM method is to analyze data and recognize the patterns of data. The suitability concept of SVM is important in water resource management to study and identify the pattern of water quality of the Putrajaya Lake and Wetland.

A set of feasible water quality parameter are measured for SVM model for comprehensive assessment. The physical, chemical and biological variables were temperature, dissolved oxygen (DO), conductivity, pH, chlorophyll-a, biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammonia-N, nitrate, T-phosphorous and *E.coli*.

The data of the water quality of Putrajaya Lake and Wetland was collected between years 2006 to 2009 which begin in October 2006 until December 2009. The predicting of the Lake water quality is not based on time taken but used dissolved oxygen as the indicator for main purpose. Twenty nine sampling stations were selected representing the open water body in the Putrajaya Lake and wetland. The samples data are collected monthly and statistically assessed in an attempt to make an evaluation and classification of the river system by referring to the Putrajaya Lake Water Quality Standards. The Putrajaya Lake Water Quality Standards is the water quality standard that own by Putrajaya Lake which the only one lake that have their own water quality standard [2]. Meanwhile other water reservoirs rely on National Water Quality Standards for Malaysia (NWQSM).

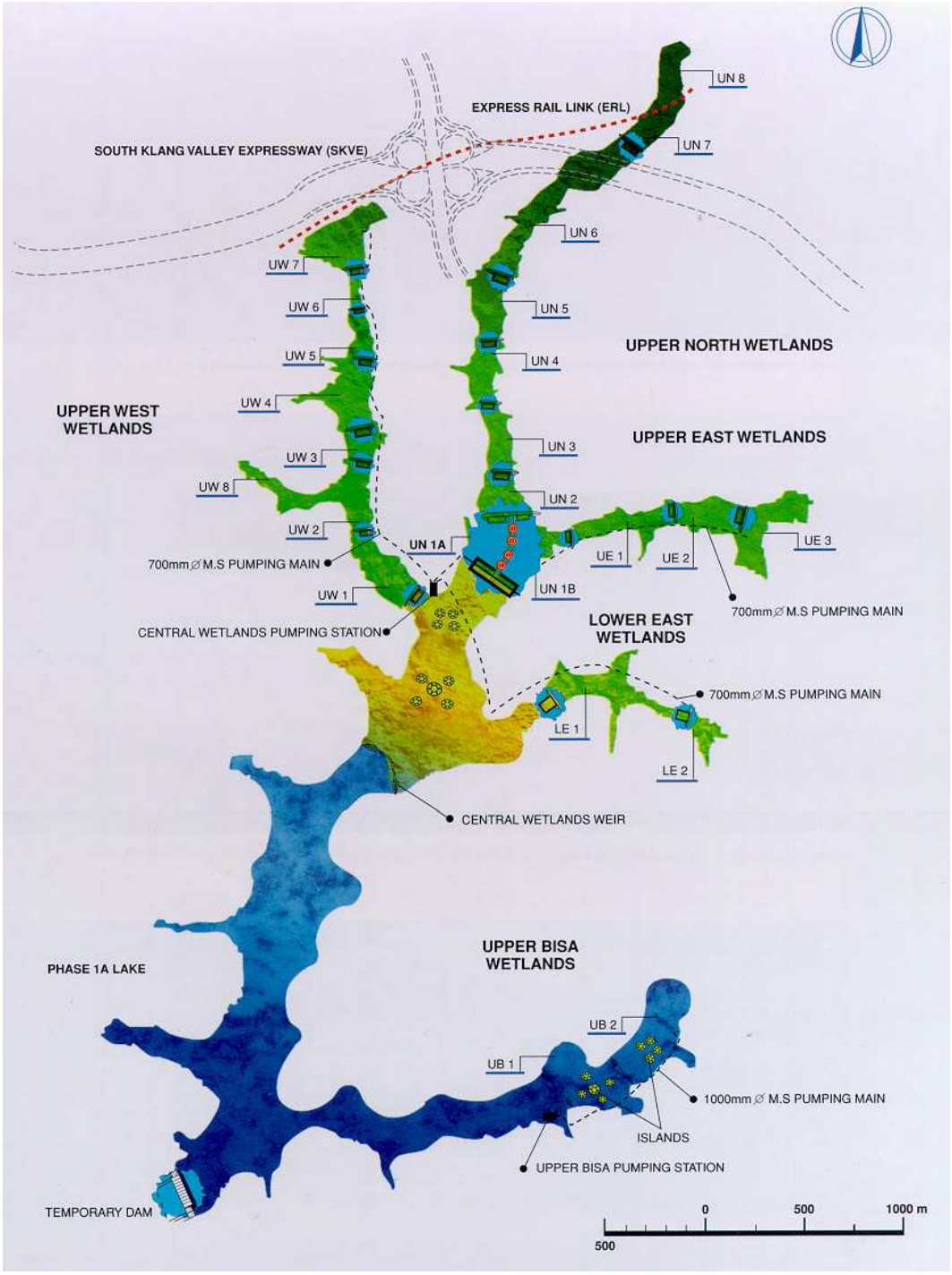

Figure 1: Sampling stations of water quality in Putrajaya Lake and Wetland

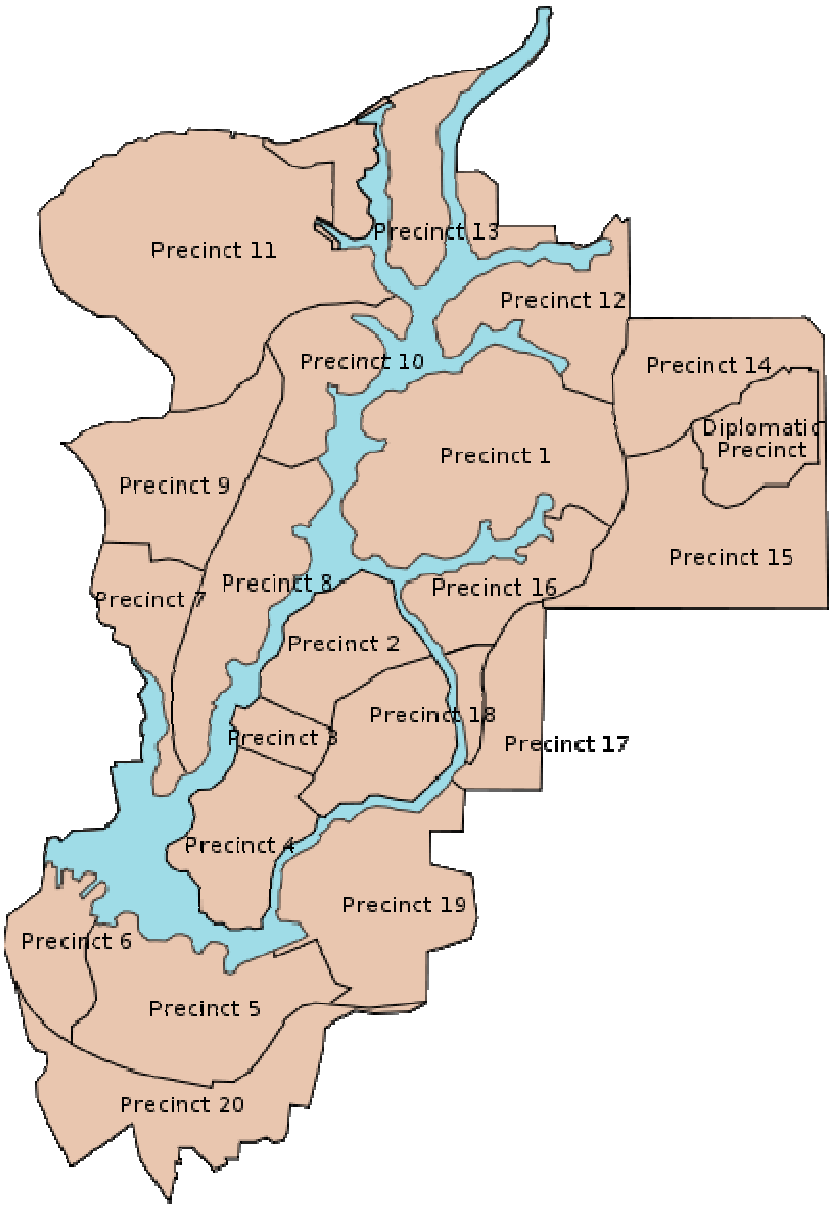

Figure 2: Putrajaya Layout including the lake and wetland.

1.3 Background of Putrajaya Lake and Wetland

Putrajaya is being and constructed as an intelligent city in a garden to serve as Malaysia Government Administrative Centre. A man made 400 acre lake is being constructed by damming the Sg. Chuau and Sg. Bisa. The centrally located lake (Fig. 1) gives the city a unique and distinctive feature and identity. Perbadanan Putrajaya, the agency responsible for the overall management of Putrajaya constructed a series of wetland to guarantee that the water going into the lake is clean and pollutant free. This environmental friendly approach by the agency is most appreciated.

The construction of wetland as a natural treatment system is to deal more with the upstream inflow into the lake. A wetland is a land area that is saturated with water, either permanently or seasonally, such that it takes on characteristics that distinguish it as a distinct ecosystem. The wetland is to be accompaniment by riparian parks and gross pollutant traps.

With current technology and advance ecological methods in design and construction, the formally palm oil estate are being transform into the first man made wetland in the tropics. Their major function is to make sure that the water course of Sg. Chuau and Sg. Bisa that enters the lake meets the standard set by Perbadanan Putrajaya.

So as to maximize the role of the wetland as a natural filtration system, a variety of aquatic plant had been planted. These plants also help in get rid of organic substance and pollutant from the catchment area and take care of natural run-off from the Sg. Chuau catchment.

1.4 Support Vector Machine (SVM)

Support vector machine (SVM) is popular in a wide variety areas including in biological applications which is an up-and-rising learning technology that is being used as a classification tool. A support vector machine (SVM) is a computer algorithm that learns by example to assign labels to the objects. The SVM algorithm is based on the statistical learning theory and the Vapnik–Chervonenkis. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vapnik and Lerner.

Support vector machines (SVM) are a group of supervised learning methods that can be used in classification or regression. The development of SVM could be found in numerous areas of application such as in chemistry and biology. In the chemistry field, the application is widely used in drug design, chemical engineering, forecasting molecule concentration from spectral data, forecasting qualitative and quantitative sensor from sensor records, and text mining.

The application of SVM has been broadly used in research area such as computational biology. The support vector modeling was used to recognize translation start sites, pattern recognition problems including protein remote homology detection, prediction of protein-protein interactions, microarray gene expression analysis, functional classification of promoter regions, and solved many other cases.

Linear and separable data sets

In the beginning, SVM is build to classify for two classification problems. Classification is the model for pattern recognition which tries to allocate each input value to one of a given set of classes. For example to decide whether the student's gender whether 'boy' or 'girl' in the school. The hyperplane is constructed to identify the resulting margins between the data points of these two groups. The SVM classification using the techniques based on the theory of the best possible separation of classes. The maximum separating plane (SP) will be built when it mapping the original data from input space into a high dimensional space using a right kernel function.(Figure 3)

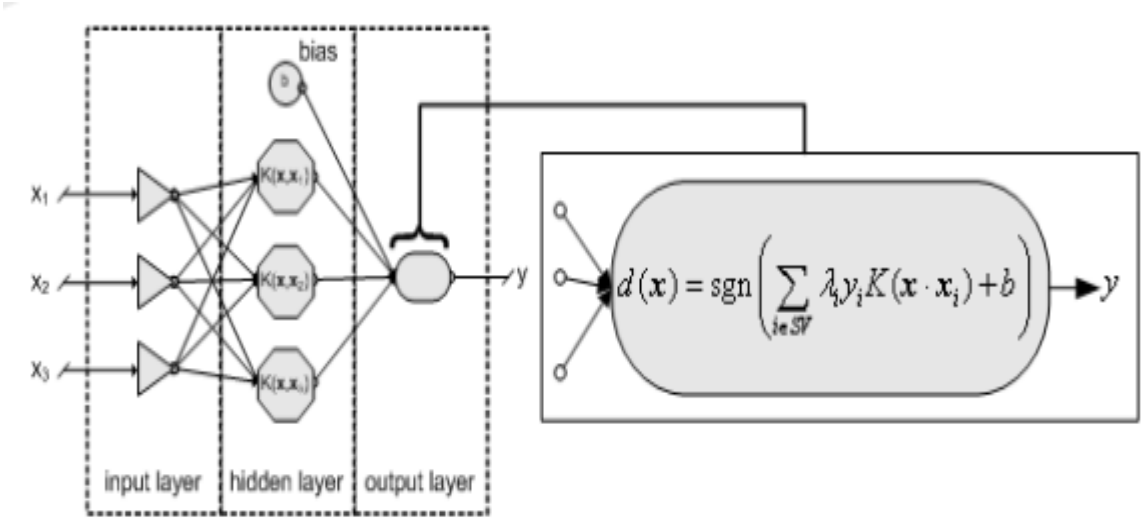

Figure 3: Mapping the data from input space to high dimension space

By mapping the data, two parallel hyperplanes are constructed, one on each side of the SP that divides the data (Figure 4). The space in between the two parallel hyperplanes is undergoing maximization by separating plane. If the information or data are separable, the hyperplane is selected from among the number of linear classifiers. SVM approach will choose the hyperplane that has lowest generalization error that from the process of structural risk minimization. Hence, the hyperlane that be constructed that have the maximum boundary between the two classes.

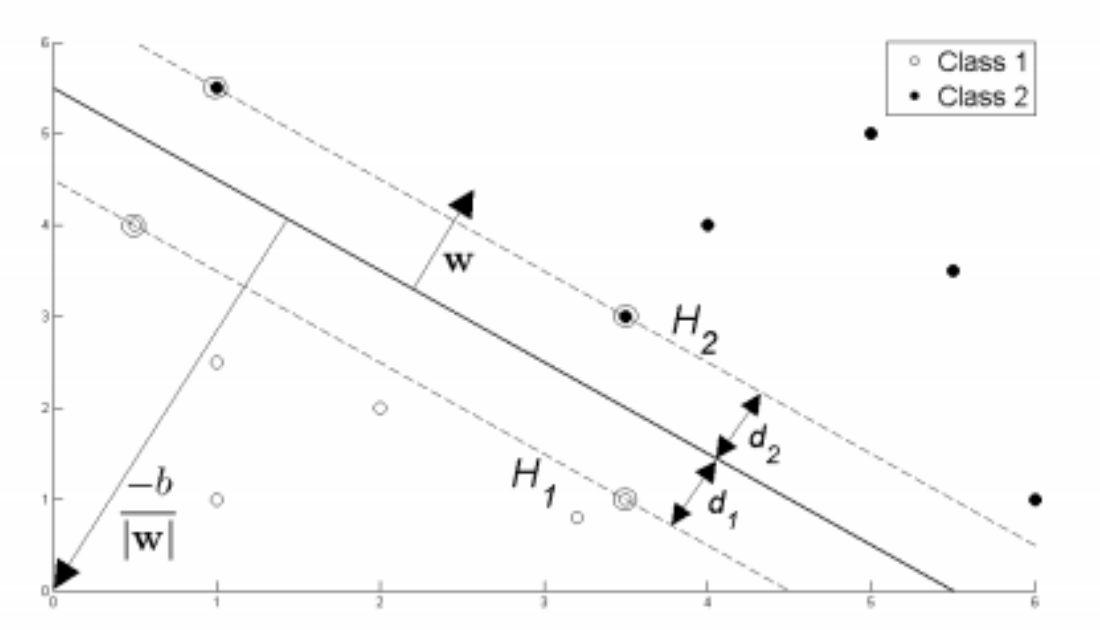

Figure 4: Hyperplane through two linearly separable classes

SVM is a kernel-based technique which mapping the input data (x) into a high dimensional feature space $\phi(x)$. [8] For two data points x, y, the function give the output $\langle\phi(x), \phi(y)\rangle$ in the feature space. The "kernel trick" works by learning the samples that takes place in the feature space and the final product will show the data points inside dot products with other points.

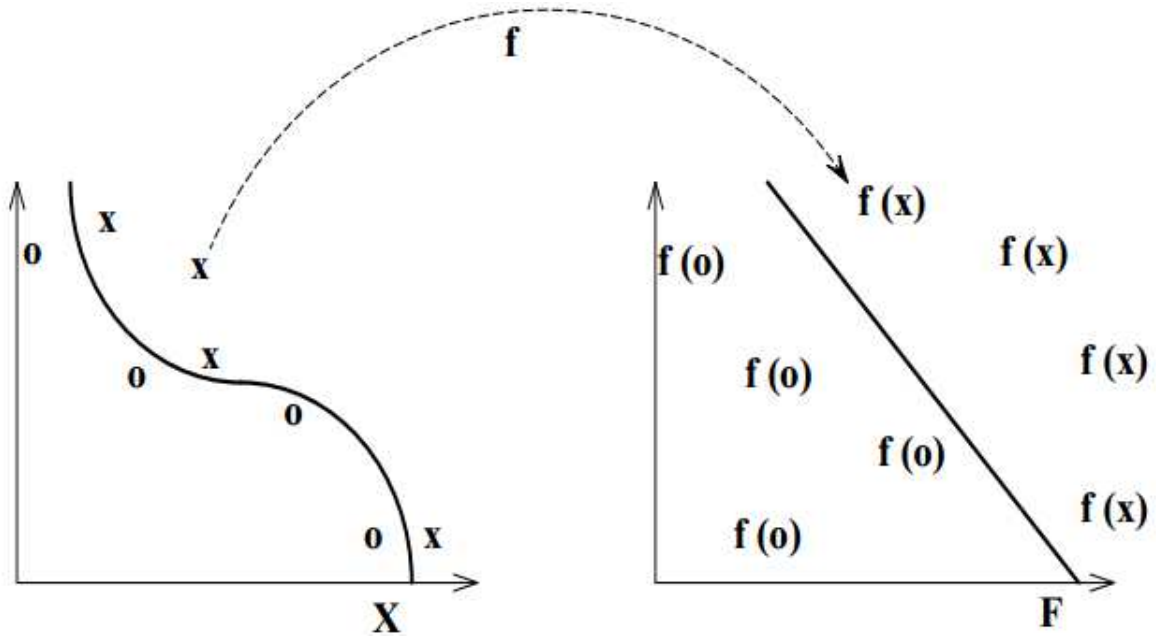

Figure 5: Mapping linearly separable data into high dimension feature space (x → $\phi(x)$)

For example, consider separating the set of *n* training vectors samples belonging to two separate classes. *y* is belonging to the two classes.

$\{x_i, y_i\}$, i = 1, . . ., *n*, $x_i \in R^n$,$y_i \in \{-1, 1\}$

We need to find a hyperplane which can separate the training set into two classes with where; $w \in R^n$, $b \in R$., The function of hyperplane is:

$\langle w, x \rangle + b = 0$ (1)

w and b are parameters for the hyperplanes. w defined as weight vector and b is the scalar.

The function is resultant to this decision function:

$$f(x) = \sin(\langle w, x \rangle + b) \quad (2)$$

The linear decision function is can be stated as:

$$f(x) = \sin\left(\sum_{i=1}^{k} \alpha i, yi(xi.yj) + b\right) \quad (3)$$

A kernel function ϕ is used for mapping the training data for nonlinear classification case into high dimension space. The separating plane will make sure the training data will linearly separable. From the equation 1, the hyperplane can be stated as

$$\langle w, \phi(x) \rangle + b = 0$$

Thus the nonlinear decision function is :

$$f(x) = \sin\left(\sum_{i=1}^{k} \alpha i, yi(\phi(xi).\phi(yj)) + b\right)$$

Non-linear and non-separable data sets

The linear data sets theory on the training data is a very strong presumption. Not many real-world data sets are linearly separable and therefore the current setting is somewhat unrealistic. For this problem, SVM can replace the linear setting of SVM model easily by extended to the non-linear setting through taking into consideration of kernel functions. If the data is categorized in non-separable group, the SVM approach will try to locate the hyperplane that maximizes the margin and at the same time minimizes a quantity proportional to the number of misclassification errors.

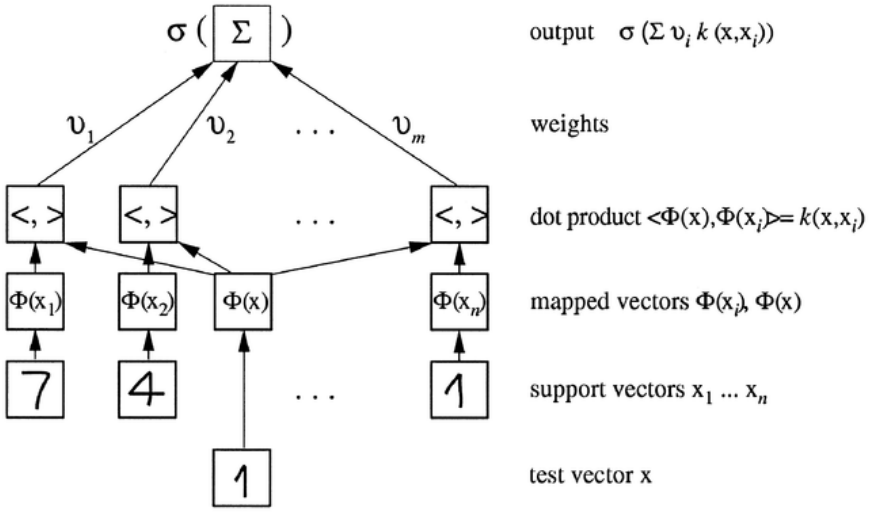

Figure 6: Support vector machine architecture

Support vector machine kernel

By using SVM, we need considered several issues in this application such feature selection, kernel function selection, and the penalty and inner parameters of kernel function selection. The kernel function $K(x, y)$ returns the inner product between two points in a suitable feature space. The radial basis function (RBF) is the most widely used of several kinds of kernel function and performs very well in most cases. Kernels commonly used with kernel methods and SVMs in particular include the following:

• Linear kernel

$k(x, y) = \langle x, y \rangle$

• Gaussian Radial Basis Function (RBF) kernel

$k(x, y) = \exp(-\sigma \|x - y\|^2)$

• Polynomial kernel

$k(x, x') = (\text{scale}.\langle x, x' \rangle + \text{offset})^{\text{degree}}$

• Hyperbolic tangent kernel

$k(x, x') = \tanh(\text{scale}.\langle x, x' \rangle + \text{offset})$

CHAPTER 2

LITERATURE REVIEW

Support Vector Machine (SVM) are one of machine learning technology that prove to have guarantee in water quality classification, prediction and pattern recognition of water reservoir; lake and river. Water quality and its improvement have a close connection with the presence of dissolved oxygen. Dissolved oxygen concentration should move toward equilibrium concentration with partial pressure of atmospheric oxygen for continuance of aquatic health.

2.1 Dissolved oxygen and its dependency

Dissolved oxygen is selected as indicator for water quality prediction because of it component can be influence by other parameters of surrounding. DO is determined by the concentration oxygen that present in water. The concentration can be indicates as percentage of saturation of oxygen in water. In other hand, it told how much concentration of oxygen can be hold by water. For example, dissolved oxygen is really dependent by temperature of the water. The water can hold more oxygen in the colder temperature compared to warmer temperature.

The dependency of dissolved oxygen toward temperature was proven by the study that had been done in Chesapeake Bay, United State. The data were collected from continuous monitoring at the sampling site. All water quality information was collected at the same sampling station and in the same period of time for a week. The data were then been illustrate in graphical representation to show the connection between them (Figure 7 and 8). The first graph shows dissolved oxygen concentration and the second graph shows the water temperature in Fahrenheit. From the graph we noticed that how the fluctuations in oxygen level declined when water temperature increase.

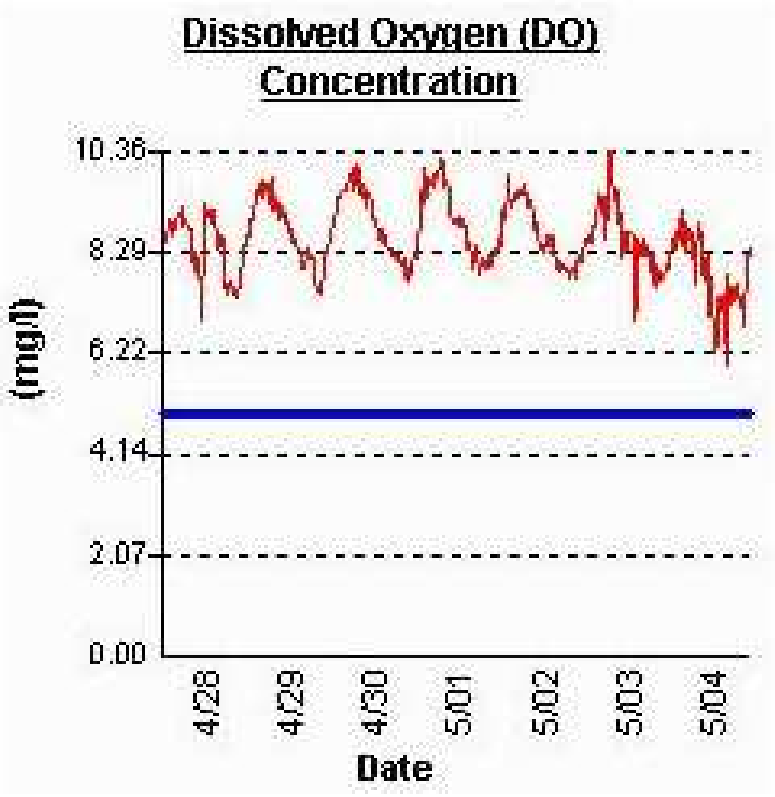

Figure 7: Dissolved oxygen concentration.

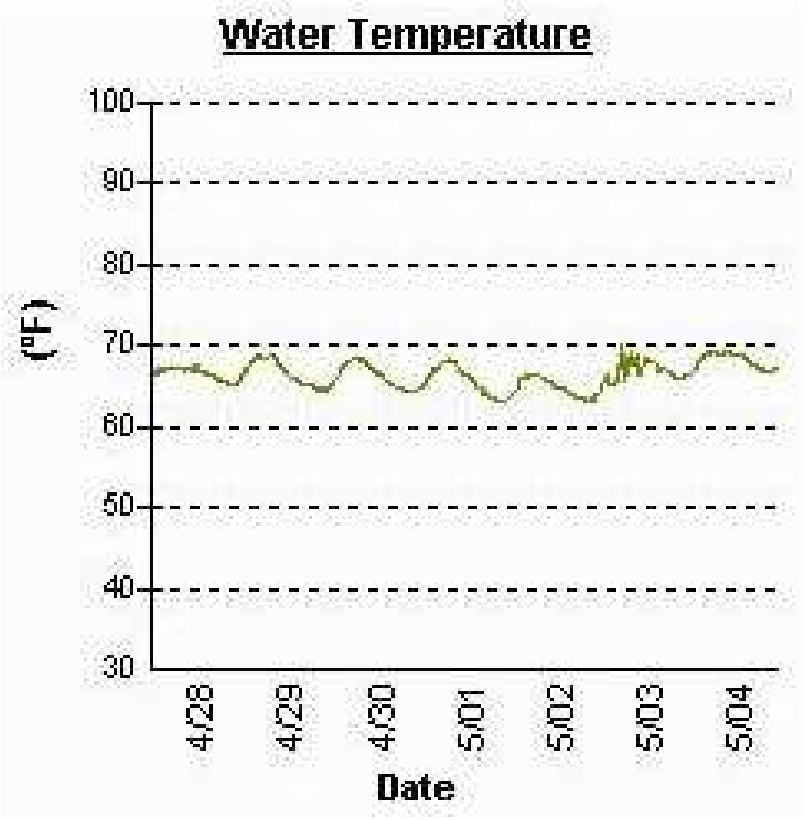

Figure 8: Water temperature

Another variable that has high potential in affecting the dissolved oxygen in water streams is Biochemical oxygen demand (BOD). BOD measures the quantity of oxygen used by microorganisms in decomposing organic material in the water. BOD also measures the chemical oxidation of inorganic substance. The greater the BOD, the more rapidly oxygen is depleted in the stream. This means less oxygen is available to higher forms of aquatic life.

Dissolved oxygen (DO) concentration is a common indicator of the health of the aquatic ecosystem. DO was originally modeled in the Ohio River (US) by Streeter and Phelps (1925). Since then a number of modifications and extensions of the model have been made

relating to the number of sinks and sources of DO being considered, and how processes involving the nitrogen cycle and phytoplankton are being modeled, as illustrated in Figure 3.1. The sources of DO in a water body include re-aeration from the atmosphere, photosynthetic oxygen production from aquatic plants, denitrification and DO inputs.

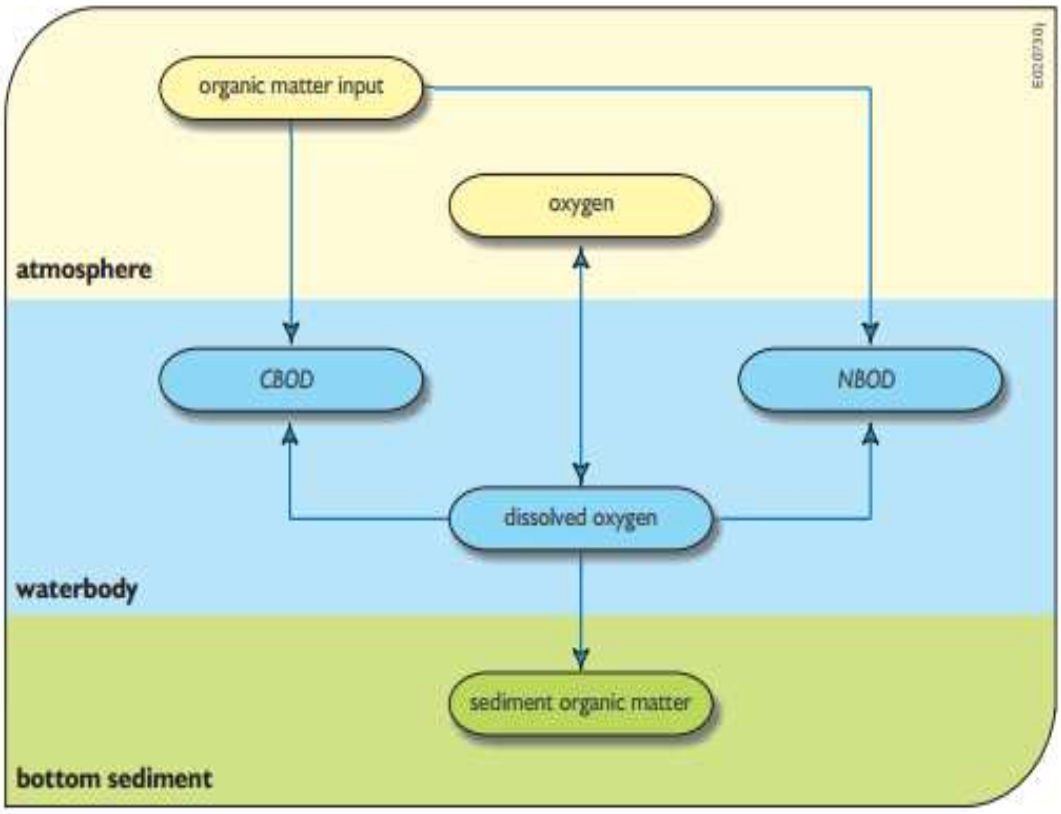

Figure 9: The dissolved oxygen interaction in a water body, showing the decay (satisfaction) of carbonaceous, nitrogenous and sediment oxygen demands and water body re-aeration or de-aeration.

2.2 Application in classification and prediction of water quality

Many research used SVM modeling as one of approaches to settle the problem in classification and regression case that may include water quality classification and forecast. The research had been done to forecast the biochemical oxygen demand (BOD) of water by built a fit Support Vector Regression (SVR) model. The classification of water quality was based on the sampling sites (spatial) and months (temporal) that which will cluster the data to the similar ones. The data records of water quality were split to three sub-sets namely; train, test and validation. The result showed that the classification produce high number of support vector for spatial and also in the temporal case. It can be concluded that support vector classification models and regression is able to do a future prediction because from the resulting showed that most of data points is used when the construction of support vector classification models and regression that suggest by regression modeling.

SVR model provided a tool for the prediction of the water BOD using set of a few measurable variables. Support vector regression model offered something like a tool for future forecast of water biochemical oxygen demand which also may be used in water quality prediction for future trends by using selected number of parameters.

The SVR predictions are accurate because the values of correlation coefficient (R) are high which are closer to unity. Other than that, evaluation parameters of performance resulting in the low values prove that it is closer to unity which proposed for a good-fit of the model to the data set and its predictive ability for the new future samples.

CHAPTER 3

MATERIALS AND METHODOLOGY

3.1 Study area and sampling stations

The focus of the study area is located around the city of Putrajaya, Putrajaya Lake and Wetland. Water samples were collected from twenty-nine water sampling stations that representing the open water body in the Putrajaya Lake and wetland. The samples data are collected monthly and statistically assessed in an attempt to make an evaluation and classification of the river system by referring to the Putrajaya Lake Water Quality Standards.

The Lake which is centrally located in the covers an area of 4,581 hectares. The fast growing it will eventually accommodate a population of some 330,000. This artificial lake is surrounded by 20 planning precincts consisting the government precinct, core island, sports and recreational precincts and residential precincts.

The development of Putrajaya involved the transformation of 4581 hectares of mainly agriculture land to modern urban centre. With the massive development, it has large impact on the hydrological system of that region. The effect of urban development is in any part of the world will surely affect the run-off any flow rate of water. There is also potential increase in pollution loads entering the Putrajaya Lake.

The lake drains a total catchment area of 51.9km2. But the runoff entering the lake is too huge to just totally depend on the wetland to clearance the runoff entering the lake. The wetland is only able to intercept only 60% of the runoff. So on to avoid overloading the wetland beyond it pollution retention capacity it is critical to control the point source of the catchment.

3.2 Water quality parameter

The data of the water quality of Putrajaya Lake and Wetland was collected between years 2006 to 2009 which begin in October 2006 until December 2009. Six water quality parameters were measured for SVM model for comprehensive analysis. The physical, chemical and biological variables selected for the proposed SVM classification, namely; dissolved oxygen (D.O), temperature (Temp), water pH, electrical conductivity (COND), Chemical Oxygen Demand *(*COD*),* Biochemical Oxygen Demand, *Escherichia coli* (*E. coli*),

The dissolved oxygen variable is being used as indicator of Putrajaya Lake and Wetland water quality measurements standard. The dissolved oxygen parameter is categorized to three groups; high, medium and Low. High indicating that the dissolved oxygen has value more than 7, Medium has value more than 5 but cannot exceed 7 and lastly the Low group will indicated the value is lower than 5.

Dissolved oxygen is influence by several variables such as temperature, pH, nitrogenous material and salinity. In general, oxygen concentrations are beneath saturation due to the presence and oxidation of decaying organic substance.

3.3 Software

This water quality monitoring project is using Support Vector Machine (SVM) model in R programming language. R is 'GNU S', an open source programming language and software environment for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. It can run on a various platforms such UNIX, Windows and MacOS. The R user can used the nearest CRAN mirror to minimize network load. CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.

R is an integrated suite of software facilities for data manipulation where it is an effective data handling and storage facility. The software also fit for a large, coherent, integrated collection of intermediate tools for data analysis. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of *packages*.

In this study, the use of computer with Intel Core i3 2.93 GHz central processing unit (CPU), 2 GB DDR3 1333 memory, and Windows XP operating system (OS).

3.4 Methodology

3.4.1 Catalogue data

A database (Excel spreadsheet) was created to catalogue general operational and environmental information of water quality from Department of Environment documents and other sources as well as information on lake and wetland of Putrajaya for future analysis. The data including all the parameters selected.

3.4.2 Conversion the data to SVM's software format

The first step in analysis the data by using the SVM is, the data must be transform to the format of an SVM's software where R programming language can read the data in CSV format. CSV is term for Comma Separated Values or also stand for Comma Delimited. A CSV file is a specially formatted plain text file which stores spreadsheet or basic database-style information in a very simple format, with one record on each line, and each field within that record separated by a comma.

Creating a CSV file from spreadsheet data can be done just by using Microsoft Excel. Firstly, open the spreadsheet document and go to the 'File' pull-down menu and choose 'Save As'. Change the "Save as type" or "Format" field to read: "CSV (Comma delimited)". Enter a name for the document and click 'Save'.

3.4.3 SVM procedure

1. Partitioned data into training and testing

   For the purpose of classification, the data were partitioned into two subsets; training and testing. Separating data into training and testing sets is an important part of evaluating data in support vector machine model. Typically, when separating a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. The data is divided to 80% for training data and 20% for testing data records in the study. For this matter, the total of 150 data records, there are 105 records for the training set and 45 records for the test set. The support vector machine uses input vectors and corresponding target vectors to train the model. So that a set of model can be train on one half and test them on unseen data.

2. Read data from file

   Before the SVM procedure was done, the data that store in the Excel spreadsheet had been converted its format from XLS to CSV for the compatibility format of R software. The initial stage in SVM method is to read the data from the CSV file using the `read.csv` function. Convenience functions `read.csv` provides arguments to `read.table` appropriate for CSV and tab-delimited files exported from spreadsheets. The function `read.table` is a way to read in a rectangular grid of data.

The three sets of data in CSV format have been separated by their own ratio of data training and testing. So, the data must be read into two separate functions.The training data is assign to object `xtrain` and `ytrain` and the test data is assign to object `xtest` and `ytest`. `xtrain` and `ytrain` object is assign to read the data from second column to the eleven column. Meanwhile the `xtest` and `ytest` object is assign to read the data from first column only namely dissolved oxygen (D.O.).

3. Generate the label

The label is generating for each classes. The label generate by using the `factor` function. The function `factor` is used to encode a vector. The three classes are High, Medium and Low.

4. Cross validation

Cross-validation is a method for evaluating how the results of a statistical analysis will simplify to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. Suppose we have a model with unknown parameters, and a data set to which the model can be fit (the training data set). The fitting process optimizes the model parameters to make the model fit the training data as well as possible. The performance of all methods developed in this study is evaluated using

five-fold cross validation. In five-fold cross validation dataset has been divided into five sets where each set have nearly equal number of High, Medium and Low level.

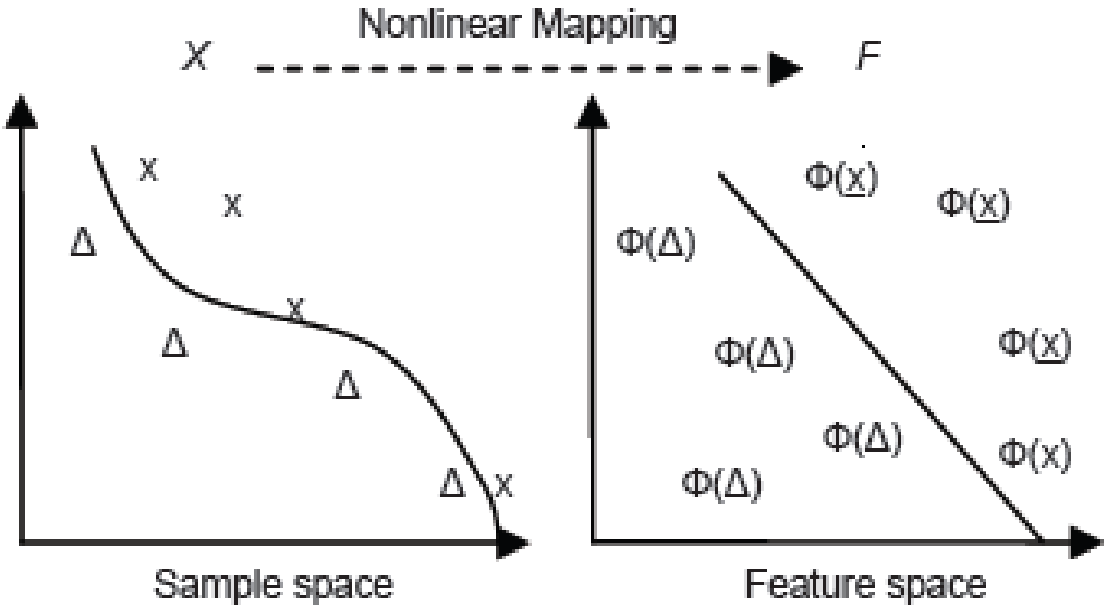

Figure 10: Nonlinear Mapping from sample space to high dimension future space.

Because the data need to be analysis is non-linear data, the kernel function $K(x, y)$ performs the non-linear mapping samples of the input space into higher dimension space, unlike the linear kernel, it can handle the case when the relation between class labels and attributes is nonlinear. The radial basis function (RBF) is the most widely used of several kinds of kernel function and performs very well in most cases. It was used in this study because of its better ability to deal with the nonlinear relationship between the label set and the attribute set and because it has fewer parameters.

RBF kernel supports two parameters which are cost C and sigma (σ). The parameter search for optimal C and σ must be carry out beforehand for a given problem. The aim is to identify optimal C and σ so that the classier can accurately forecast unknown data or the testing data. The cross-validation process can avoid the over fitting problem in the classification of data.

The function formula is:

$$k(x, y) = \exp(-\sigma||x - y||^2)$$

For SVM model, there exists no standard procedure to determine the free parameters C and σ. Here, the technique of cross-validation and grid-search [19] was applied to obtain SVM optimal parameters of C and σ. When applied to a large data set, however, it requires a long time for training so the model selection task and its performance can be degraded a long time. Value pairs (C, σ), respectively was assessed using cross-validation and then choose the pair with highest precision. The value of C and σ are increasing exponentially.

5. Use the parameters to train the model.

After the best parameter model of RBF (C; σ) is found, the whole training set is trained again to generate the final classifier and classify observations from the test data. The water quality forecast, the predicting table and accuracy of the prediction will be obtain.

CHAPTER 4

RESULTS

In this experiment, water quality data are provided by the Department of Environment, Putrajaya, Malaysia during 2006-2009. There are 150 records of data for both training and testing data sets. The best six parameters were selected out of eleven parameters. So those, each of records consist of six parameters. Therefore, these six parameters are included in model development were dissolved oxygen (DO), temperature, pH, biochemical oxygen demand (BOD), salinity and lastly E.coli.

Initially, the eleven parameters going into the selection by ranking it based on the lowest cross validation error rate. The parameter selection method is based on the cross-validation error rate is derived from one cross-validation procedure by training in linear kernel SVM. The selection of parameters is considered necessary because of some of the parameters are affecting the performance of analysis. Cross-validation is valid approach to predict or estimate the accuracy of the study. The lower the cross validation error rate, the better the performance. The ranking of the cross-validation error rate is shown in Table 1.

Based on the lowest to high cross-validation error, the result show water pH has the lowest cross-validation error rate. Following by biochemical oxygen demand, salinity, water temperature, E.coli, NH3N, conductivity, phosphorus, COD, nitrogen and the highest cross-

validation error rate is chlorophyll-a. Each variable is append or bind to its own variable to train in the linear kernel function by using `cbind` function.

Parameters	Cross-validation error rate	Ranking
pH	0.47619	1
BOD	0.533333	2
Salinity	0.542857	3
Temperature	0.561905	4
E. coli	0.561905	5
NH_3N	0.609524	6
Cond	0.628571	7
T.Phosphorus	0.647619	8
COD	0.657143	9
T.Nitrogen	0.67619	10
Chlorophyll-a	0.733333	11

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

After training for each parameter, all data records were separated into training and testing for further analysis using the Gaussian RBF kernel function. Gaussian RBF kernel function can acquire better result compare to polynomial kernel function and sigmoid kernel function because of its capability of operation time and performance for modeling the support vector machine. In order to find the best input parameter for the pattern recognition, the backward training is used which each parameter will be remove at every training course. At lowest cross-validation error, the number and type of input parameters are selected based on its performance. The result to decide the input of parameters is shown in table 2.

Input of parameter	Type of parameters	Accuracy
Eleven	pH, BOD, Salinity, temperature, E.coli, NH_3N, Cond., T-phosphorous, COD, T-nitrrogen, Chlorophyll-a	0.5333
Ten	pH, BOD, Salinity, temperature, E.coli, NH_3N, Cond., T-phosphorous, COD, T-nitrrogen	0.6
Nine	pH, BOD, Salinity, temperature, E.coli, NH_3N, Cond., T-phosphorous, COD	0.644
Eight	pH, BOD, Salinity, temperature, E.coli, NH_3N, Cond., T-phosphorous	0.644
Seven	pH, BOD, Salinity, temperature, E.coli, NH_3N, Cond.,	0.6888889
Six	pH, BOD, Salinity, temperature, E.coli, NH_3N	0.6777
Five	pH, BOD, Salinity, temperature, E.coli	0.71111
Four	pH, BOD, Salinity, temperature	0.6222222
Three	pH, BOD, Salinity	0.6
Two	pH, BOD	0.6

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

Several kernel functions also been applied in the support vector classification. Other than RBF kernel, the ANOVA and polynomial kernel is used to find the optimal performance result. The summary of the result is illustrated in table 3. According to the support vector classification by using all three kernels, RBF kernel performance has the most optimal performance compare to other. RBF kernel is employed for the training the data set. The RBF kernel supports two parameters, sigma (σ) and the cost parameter C. Meanwhile for the Anova kernel employed two parameters; degree and sigma (σ) and polynomial kernel supports three parameters to control the kernel; degree, scale and offset.

Kernel	Degree	Scale	Offset	Sigma	Training Error	Cross Validation Error	Accuracy
Anova	1	9999	9999	0.1	0.380952	0.361905	0.6222222
Anova	1	9999	9999	1.2	0.314286	0.380952	0.4888889
Anova	1	9999	9999	0.2	0.190476	0.295238	0.6222222
Poly	1	1	1	9999	0.314286	0.390476	0.5777778
Poly	1	1	2	9999	0.314286	0.380952	0.5333333
Poly	1	1	3	9999	0.304762	0.380952	0.5111111
RBF	9999	9999	9999	0.7	0.152381	0.342857	0.6888889
RBF	9999	9999	9999	0.5	0.142857	0.304762	0.7333333
RBF	9999	9999	9999	0.2	0.190476	0.314286	0.7111111

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

	Prediction	Actual
High	15	15
Medium	9	15
Low	22	15

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

Predicted	Test		
	High	Medium	Low
High	12	1	1
Medium	0	8	1
Low	3	6	13

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

The training result give total predicted water quality each class is summarized in table 4, 5 and 6 respectively. The testing data which used to be predicted include the end month of

sampling of year 2006 which is in October until October 2009. The training accurately predicted data in High, Medium and Low group of DO is about 15, 8 and 13 data respectively. The SVM analysis shows that the accuracy of the prediction is about 0.73333 which is equal to 73% of accuracy. The result is plotted to show the comparison between predictive data and the original data as shown in Figure 4.1.

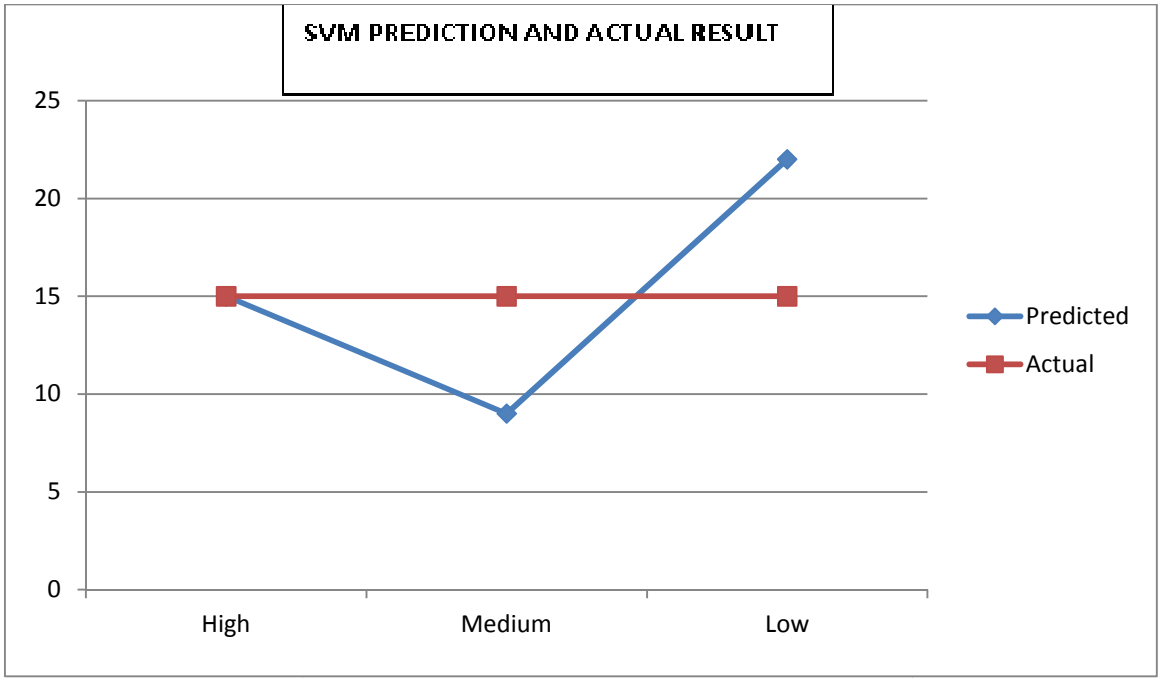

Figure 11: SVM prediction and actual result.

DATA NUMBER	PREDICTED	ACTUAL		DATA NUMBER	PREDICTED	ACTUAL
1.	Low	Low		**24.**	Low	Medium
2.	Low	Low		**25.**	High	Medium
3.	Low	Low		**26.**	Medium	Medium
4.	Low	Low		**27.**	High	Medium
5.	Low	Low		**28.**	Medium	Medium
6.	Low	Low		**29.**	Low	Medium
7.	Low	Low		**30.**	Low	Medium

8.	High	Low		**31.**	High	High
9.	Low	Low		**32.**	High	High
10.	Medium	Low		**33.**	High	High
11.	Low	Low		**34.**	High	High
12.	Low	Low		**35.**	High	High
13.	Low	Low		**36.**	High	High
14.	Low	Low		**37.**	High	High
15.	Low	Low		**38.**	Low	High
16.	Medium	Medium		**39.**	Low	High
17.	Low	Medium		**40.**	High	High
18.	Low	Medium		**41.**	High	High
19.	Medium	Medium		**42.**	Low	High
20.	Low	Medium		**43.**	High	High
21.	Medium	Medium		**44.**	High	High
22.	Medium	Medium		**45.**	High	High
23.	Medium	Medium				

Table 6 False prediction and Actual prediction on data

The sensitivity and specificity of data is calculated as shown in table 5 for each class.

	Sensitivity	**Specificity**
High	0.8	0.93
Medium	0.53	0.97
Low	0.87	0.7

Table 7: Sensitivity and specificity of the result.

CHAPTER 5

DISCUSSION

The classification and prediction of water quality of Putrajaya Lake and wetland experiments are conducted by using the Support Vector Machine (SVM) in R programming language. For the purpose of classification, the data were divided into two subsets; training and testing. The data is split to70% for training data and 30% for testing. So that a set of model can be train on one half and test them on unseen data.

Among the three kernels that have been used to train the data, RBF kernel has been selected because of high in accuracy. The summary of the result can be seen in table 3. In this study, 5 fold cross-validation is employed to decide the optimal value of different parameter C and σ. The best combination of these two parameters will give the best classification for accurate forecasting. From the SVM classification analysis of water quality data from Putrajaya, the result that produce from least cross validation error which is about 0.304762. The best cost C and sigma value are 0.7 for C and 0.5 for sigma. Meanwhile, the numbers of support vector on the other hand are about 90. The analysis show high number in support vectors for the classification that indicates that the classification modeling used most of the data points.

The evaluation of performance in support vector classification was based on sensitivity, specificity and accuracy of forecast. The sensitivity, specificity and accuracy are calculated based on the equation below (1), (2) and (3) below.

Sensitivity (1)

$$= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\sum \text{True Negtive}}{\sum \text{Condition Negative}}$$

Specificity (2)

$$= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$$

Accuracy (3)

$$= \frac{\sum \text{prediction data that equal to original data}}{\sum \text{Original data}}$$

Sensitivity is defined in this study as the ability of an indicator variable to correctly classify and predicted data to the actual data. Being able to detect water quality data in real time is also imperative. The compute result showed that the classification of sensitivity was mostly higher than the specificity. The classification has high sensitivity, it indicate that the test can accurately predict the High class of water quality. The rate value of High class sensitivity is 0.8 or 80%. If the amount the prediction of High class water quality is accurate, the water quality can be concluded as has good index in water quality and the water is safe to be used.

But in other hand, rate value of Medium class showed quiet low in sensitivity. This may be due to the fact that the variations water quality in a sampling site is largely due to the other parameters such as water temperature and pH. At different sites in a geographical area largely determines the water quality and largely account for the water quality variation.

The SVM analysis shows that the accuracy of the prediction is about 0.73333 which is equal to 73% of accuracy. A lot of approaches can be used to calculate the accuracy of the support vector model. In this study the accuracy is computed in terms of; the sum of predicted value- regarded as accurate if it is equal to the actual value and it is divided to the total length of actual value set.

CHAPTER 6

CONCLUSION

The support vector machine is a suitable method to analyze the water quality of Putrajaya Lake and Wetland. SVM can recognize the pattern of water quality based on the water quality parameters from the data records. The dissolved oxygen has been used as indicator for water quality forecasting. The suitability of dissolved oxygen as indicator because of the variable is influencing by many surrounding factors or parameters such water temperature and biochemical oxygen demand.

Because of the data is nonlinear data set, the RBF kernel has been selected to mapping the data to high dimension space. The analysis show high number in support vectors for the classification that indicates that the classification modeling used most of the data points. The percentage rate sensitivity of High class of dissolved oxygen shows that it is has the high value. It can be concluded that the prediction is accurate. The resulting proved that the support vector classification model can be constructed for future forecasting trend.