Classification and Regression Trees and MLP Neural Network to Classify Water Quality of Canals in Bangkok, Thailand

Sirilak Areerachakul, Siripun Sanguansintukul
Chulalongkorn University, Thailand

Abstract

Water quality is one of the major concerns of countries around the world. This study endeavors to automatically classify water quality. The water quality classes are evaluated using 6 factor indices. These factors are pH value (pH), Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Nitrate Nitrogen (NO_3N), Ammonia Nitrogen (NH_3N) and Total Coliform (T-Coliform). The methodology involves applying data mining techniques using classification and regression tree (CART) compared with multilayer perceptron (MLP) neural network models. The data consisted of 288 canals in Bangkok, Thailand. The data is obtained from the Department of Drainage and Sewerage Bangkok Metropolitan Administration during 2003-2007. The results of classification trees perform better than multilayer perceptron neural network. Classification trees exhibit a high accuracy rate at 99.96% in classifying the water quality of canals in Bangkok. Subsequently, this encouraging result could be applied with plan and management source of water quality.

1. Introduction

Water quality is a major concern around the world. In Thailand, Bangkok has been known as "the Venice of the East" from its many canals and rivers. Bangkok is the capital city, as well as, the economic center of Thailand. Its activities, which include commercial, industrial and service have caused the expansion of the city and its population to accumulate environmental pollution to the point that nature cannot cope with the pollution loading, especially for water quality. Presently, the water quality is below standard due to population increase and industrial growth. Additionally, the canals and rivers around Bangkok are used for transportation, tourism, and consumption. In order to monitor water quality, people, lab instruments and sensors have been used, but the cost and time are expensive.

The aim of this investigation is to find an automated methodology that can quickly and efficiently classify the water quality of canals in Bangkok. Recently, several machine learning algorithms have been used to find patterns to classify water quality such as decision tree and artificial neural networks (ANNs). Classification and regression tree (CART) is a type of decision tree methodology. Classification and regression tree have the advantage of expressing regularities explicitly and thus being convenience to inspect for water quality validity [17]. Artificial Neural Networks have become the central focus of many scientific disciplines, such as ecology [8], analytical chemistry [9], and water quality. Literature on modeling water quality using ANNs include [1][2][12][13]. In this study, using classification and regression tree method compared with multilayer perceptron neural network and applied to k-foldss cross validation to find efficiently model classify water quality of canals in Bangkok.

This paper is organized as follows: Section 2 describes the materials used in the experiments. Section 3 demonstrated the methodologies used in the experiments. Section 4 contains the simulation in the experiments. The results and discussion are shown in Section 5. Finally, Section 6 concludes the paper.

2. Materials

Data and surface water quality standards are described in this section.

2.1. Site Description and Data

During the years 2003-2007, Bangkok was comprised of 288 canals. Figure 1 [16] shows the network of the canals, important for the daily life of the people in Bangkok. These canals are used for consumption, transportation and recreation. Especially a result of Bangkok being the capital of Thailand, the rapid growth of industry, condominiums, high-rise and low-rise buildings, and other infrastructures, have had a significant effect on

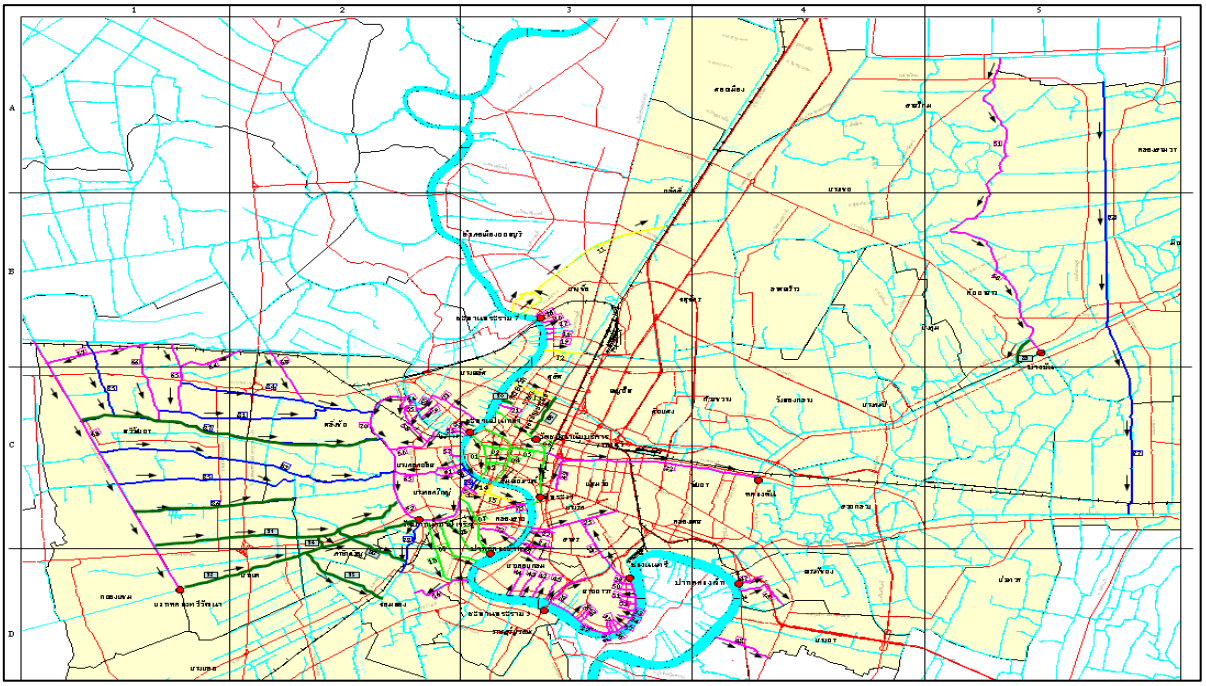

Figure 1. A map of canals in Bangkok

the canals' water quality. In order to improve canal water quality, classification different levels of water quality becomes a major concern. The understanding of different levels of water quality can be utilized in water management and treatment systems.

In this study, water quality data are provided by Department of Drainage and Sewerage Bangkok Metropolitan Administration during 2003-2007. There are 11,820 records of data. Each record consists of 6 parameters namely; pH value (pH), Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Nitrate Nitrogen (NO_3N), Ammonia Nitrogen (NH_3N) and Total Coliform (T-Coliform). The classifications of canal water quality are based on surface water standards [15].The lower the number of class, the better the quality of water quality.

2.2. Surface Water Quality Standards

Many parameters can influence the surface water quality. Six parameters are selected for the investigations.

In Thailand, the surface water quality can be classified as in Table 1[16]. Generally, surface water quality can be divided into five classes; class I, extra clean fresh surface water resources use for conservation that are not necessary to pass through water treatment processes and require only ordinary processes for pathogenic destruction and ecosystem conservation where basic organisms can breed naturally; class II, very clean fresh surface water resources use for consumption that require ordinary water treatment processes before use by aquatic organisms in conservation, fisheries and recreation; class III, medium clean fresh surface water resources use for consumption, but are passed through an ordinary treatment process before use; class IV, fairly clean fresh surface water resources use for consumption, but requires special water treatment processes before use; and class V, the sources which are not within class I to class IV and are used for navigation.

Table1. Surface Water Quality Standards [16]

Pollutants Index	Class				
	I	II	III	IV	V
pH (mg/l)	<5	5-9	5-9	5-9	>9
DO (mg/l)	>6	6	4	2	<2
BOD (mg/l)	<1.5	1.5	2	4	>4
NO_3N (mg/l)	<5	5	5	5	>5
NH_3N (mg/l)	<0.5	0.5	0.5	0.5	>0.5
T-Coliform (MPN)	<50	50	200	>200	>200

3. Methodology

In this section, we demonstrated methodology in this experiment. Classification and Regression tree and Multilayer perceptron neural network.

3.1. Classification and Regression Tree

Classification and regression tree (CART) is a type of decision tree methodology. CART analysis is a form of binary recursive partitioning [2]. The term binary implies that each group of data, represented by a node in a decision tree can only be split into two groups. Thus, each node can be split into two child nodes which case the original node is called parent node. The term recursive refers to the fact that the binary partitioning process can be applied over and over again. Therefore, each parent node can give rise to two child nodes and in turn each of these child nodes may themselves be split forming additional children. The term partitioning refers to the fact that the dataset is split into sections or partitioned.

CART analysis consists of four basic steps [2].

The first step consists of tree building, during which a tree is built using recursive splitting of nodes. Each resulting node is assigned a predicted class based on the distribution of classes in the learning dataset which would occur in that node and the decision cost matrix. The assignment of a predict class to each node occurs whether or not that node is subsequently split into child nodes.

The second step consists of stopping the tree building process. At this point a maximal tree has been produced which probably greatly overfits the information contained within the learning data set.

The third step consists of tree pruning which results in the creation of a sequence of simpler and simpler trees through the cutting off of increasingly important nodes.

The fourth step consists of optimal tree selection during which the tree which fits the information in the learning dataset but does not overfit the information, is selected from among the sequence of pruned trees.

3.2. Multilayer perceptron Network

The artificial neural network (ANN), or neural network in short, is inspired by simulating the function of a human brain. A neural network can be used to represent a nonlinear mapping between input and output vectors. Neural networks are among the popular signal-processing technologies. In 0engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters [5] [6]. A general network consists of a layered architecture, an input layer, one or more hidden layers and an output layer [12]. Figure 2 shows a typical architecture of a multilayer perceptron network. The Multilayer perceptron (MLP) is an example of an artificial neural network that is used extensively to solve a number of different problems, including pattern recognition and interpolation [4][7]. Each layer is composed of neurons, which are interconnected with each other by weights. In each neuron, a specific mathematical function called the activation function accepts input from previous layers and generates output for the next layer. In the experiment, the activation function used is the hyperbolic tangent sigmoid transfer function [13] which is defined as in equation (1):

$$f(n) = \frac{1 - e^{-2s}}{1 + e^{-2s}} \qquad (1)$$

where $s_i = \sum_{i=1}^{n} w_i x_i$, in which w_i are weights and x_i are input values.

The MLP is trained using the Levenberg–Marquardt technique as this technique is more powerful than the conventional gradient descent techniques [4].

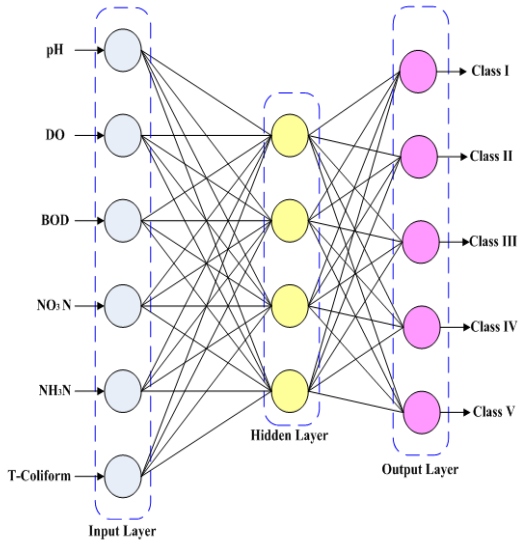

Figure 2. A typical Multilayer Perceptron ANN Architecture

The Levenberg-Marquardt (LM) algorithm [14] is the most widely used optimization algorithm. It outperforms simple gradient descent and other conjugate gradient methods in a wide variety of problems. If a function $V(x)$ is to be minimized with respect to the parameter vector $\underline{x}$, then Newton's method would be:

$$\Delta \underline{x} = -[\nabla^2 v(\underline{x})]^{-1} \nabla v(\underline{x}) \qquad (2)$$

where $\nabla^2 v(\underline{x})$ is the Hessian matrix and $\nabla v(\underline{x})$ is the gradient. If $v(\underline{x})$ reads:

$$v(\underline{x}) = \sum_{i=1}^{N} e_i^2(x) \tag{3}$$

then it can be shown that:

$$\nabla v(\underline{x}) = J^T(\underline{x})\underline{e}(\underline{x}) \tag{4}$$

$$\nabla^2 v(\underline{x}) = J^T(\underline{x})J(\underline{x}) + S(\underline{x}) \tag{5}$$

where $J(x)$ is the Jacobian matrix

$$J(x) = \begin{bmatrix} \frac{\partial_{e1}(\underline{x})}{\partial x_1} & \frac{\partial_{e1}(\underline{x})}{\partial x_2} & \cdots & \frac{\partial_{e1}(\underline{x})}{\partial x_N} \\ \frac{\partial_{e2}(\underline{x})}{\partial x_1} & \frac{\partial_{e2}(\underline{x})}{\partial x_2} & \cdots & \frac{\partial_{e2}(\underline{x})}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial_{eN}(\underline{x})}{\partial x_1} & \frac{\partial_{eN}(\underline{x})}{\partial x_2} & \cdots & \frac{\partial_{eN}(\underline{x})}{\partial x_N} \end{bmatrix} \tag{6}$$

and

$$s(\underline{x}) = \sum_{i=1}^{N} e_i \nabla^2 e_i(\underline{x}) \tag{7}$$

For the Gauss-Newton method it is assumed that $s(\underline{x}) \approx 0$, and equation (2) becomes:

$$\Delta \underline{x} = \left[J^T(\underline{x})J(\underline{x})\right]^{-1} J^T(\underline{x})\underline{e}(\underline{x}) \tag{8}$$

The Lavenberg-Marquardt modification to the Gauss-Newton method is:

$$\Delta \underline{x} = \left[J^T(\underline{x})J(\underline{x}) + \mu I\right]^{-1} J^T(\underline{x})\underline{e}(\underline{x}) \tag{9}$$

The parameter μ is multiplied by some factor (β) whenever a step would result in an increased $V(\underline{x})$.

when a step reduces $V(\underline{x})$, μ is divided by β. When the scalar μ is very large the Levenberg-Marquardt algorithm approximates the steepest descent method. However, when μ is small, it is the same as the Gauss-Newton method. Since the Gauss-Newton method converges faster and more accurately towards an error minimum, the goal is to shift towards the Gauss-Newton method as quickly as possible. The value of μ is decreased after each step unless the change in error is positive; i.e. the error increases. For the neural network-mapping problem, the terms in the Jacobian matrix can be computed by a simple modification to the back-propagation algorithm [3].

4. Simulations

This section discusses data preprocessing, experimental data and model in experiment.

4.1. Preprocessing Data

At the initial stage of the experiment, data was scaled or normalized using equation (10)

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

Where x is the original data point, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values in the data set, respectively. This is done in order to ensure that the minimum value in the data set is scaled to zero, and that the maximum value is scaled to one [11].

4.2. Experimental Data

In this study, we use the water quality data of canal in Bangkok, over a period of five years from 2003 to 2007. The main water quality indices include. According to the above indices, the water quality can be classified into 5 categories based on surface water quality standards in Thailand. 11,820 samples are available for the analysis in water quality classification.

First step, the ratio of the train and test set employed in the experiment is 60:40 randomly. This means that with 11,820 data record, there are 7,092 records for the train set and 4,728 records for the test set.

Second step, we make several different divisions of the observed data into training set and testing set. K-folds cross validation is used to measure the performance of CART and MLP neural network. K-folds cross validation is one of the most adopted criteria for assessing the performance of a model and for selecting a hypothesis within a class. An advantage of this method, over the simple training and testing data splitting, is the repeated use of the whole available data for both building a learning machine and for testing it [4]. In this study, a 3 fold partition of the data set was created. Split data to 3 fold; hold out successive blocks of observation as test sets. Each fold is held out in turn and learning scheme trained on the remaining second-third, then the error rate is calculated on the hold out set. Thus, the learning procedure is executed a total 3 times on different training sets. This means that with 11,820

data record, there are 7,880 records for the train set and 3,940 records for the test set.

4.3. CART model and MLP Neural Network Model

As in this section, shown classification and regression tree model and MLP neural network model.

4.3.1. Classification and regression tree Model

An example of the decision tree that generated from the classification and regression tree algorithm is shown in figure 3. Classification and regression tree start with parent node which is BOD. The independents parameters contain of 5 parameters. These are pH, DO, NO_3N, NH_3N and T-Coliform.
The CART procedure examines all possible independent, variables and selects one that results in binary group. In figure 3, Node 1 (parent node, BOD) split into node 2, 3 which is NH_3N and class 5 of surface water standard. Within these 2nodes, node 2 (NH_3N) become parent node and split into 2 nodes which are node 4 (DO) and node 5 (surface water standard class 5). Similarly, the tree growing

4.3.2. MLP Neural Network Model

The Levenberg-Marquardt algorithm uses input vectors and corresponding target vectors to train neural networks. All the training records were fed into the network to make it learn the potential relationships between water quality indices and their corresponding categories. Accordingly, the 6 input layer nodes represent 6 water quality indices, while the 5 output layer nodes represent the 5 different class categories. The trained neural networks can provide an output representing the specific class for each of water quality indices. The testing samples are used to verify its classification ability.

Many experimental investigations are conducted. The number of hidden nodes that provided the optimal result is 4 hidden nodes. Therefore, the architecture of the network is 6-4-5. The target mean square error (MSE) is 0.001 after 5000 iterations. Equation (11) shows the mean square error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \qquad (11)$$

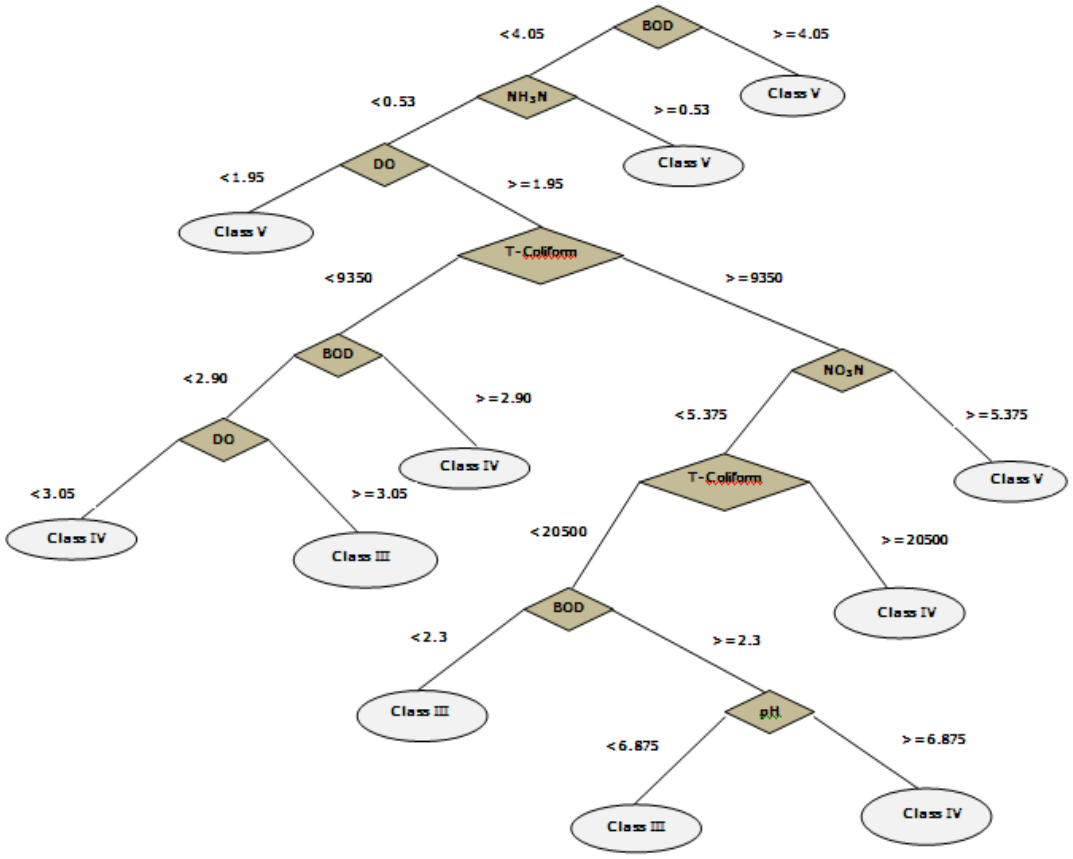

Figure 3. Example of Classification and Regression Tree Model

5. Results and Discussion

Comparing with the using, CART as a methodology lead to better result than MLP neural net work methodology of confusion matrix as following figure 4 and figure 5.

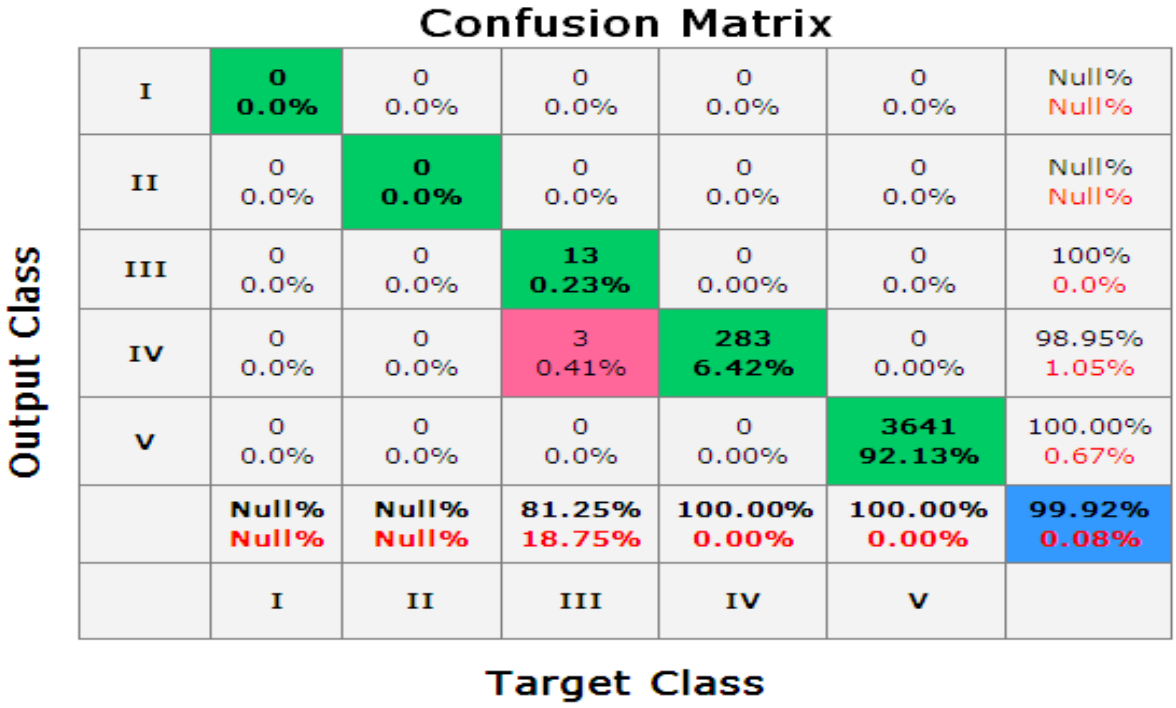

Figure 4. Example of Confusion Matrix of CART Model

Figure 4 illustrates example of the confusion matrix of CART model and figure 5 shown example of the confusion matrix of MLP neural network model for the experiment results. The confusion matrix demonstrates information about the target (actual class of surface water standard) and the output (predicted class by the network). In the matrix, each column of the matrix represents a target (actual) class, whereas, each row represents an output (predicted) class.

From figure 4 performance of the classification and regression tree approach can be evaluated using data in the matrix. The examples of interpretations include:

- Vertically reading from Target Class III, there are 13 records classified correctly. The accuracy percentage is 81.25%
- Vertically reading from Target Class IV, there are 283 records classified correctly. The accuracy percentage is 100.00%
- Vertically reading from Target Class V, there are 3641 records classified correctly. The accuracy percentage is 100.00%

Table 2. Result of Percentage Accuracy in the Test Set

Method / Accuracy Percentage (%)	CART	MLP
Without applied to K-foldss Cross Validation	99.75	97.31
With applied to K-foldss Cross Validation	99.96	98.82

Table 2 demonstrates the comparison percentage accuracy of canal water quality classification between CART methodology and MLP neural network methodology with and without K-foldss cross validation. It can be seen that classification and regression tree and multilayer perceptron neural network higher performance after applied K-foldss croess validation. The CART high correctly classified than multilayer perceptron neural network, accuracy percentage is 99.96%.

6. Conclusion

In this paper, k-folds cross validation applied to classification and regression tree (CART) and multilayer perceptron (MLP) neural network using the Levenberg-Marquardt algorithm are applied to classify the water quality of canals of Bangkok, Thailand. The results indicate that the CART performs with a high accuracy classification percentage of 99.96%, while the MLP neural network shows percent accuracy of 98.82. These encouraging results may be applied to automate water quality classifications.

As a result, the cost and time of water resource management could be minimized. Application of the CART and MLP neural network indicates that it is robust and remarkably improves the efficiency of the classification of water pollutions which are useful for planner and watershed management nutrient loading, sedimentation and also water treatment process.

7. Acknowledgment

The authors would like to thank Suan Sunandha Rajabhat University for scholarship support. Thanks to Department of Drainage and Sewerage Bangkok Metropolitan Administration for the provided data.

8. References

[1] A.Najah, A.Elshafie,O.Karim and O.Jaffar "Prediction of Johor River Water Quality Parameter Using Artificial Neural Networks", *Journal 0f Scientific Research*, EuroJournals Publishing, 2009, pp. 422-435.

[2] Breiman L, FriedmanJH, Olshen RA, and Stone CJ *"Classification and Regression Tree"* Champman &Hall (Wadsworth, Inc.), Newyork, 1984.

[3] Chi Zhou, Liang Gao and Chuanyong Peng, "Pattern Classification and Prediction of Water Quality by Neural Network with Particle Swarm Optimization", *Proceedings of the 6th World Congress on Intelligent Control and Automation*, China, June 2006, pp. 2864-2868.

[4] D. Anguita, S.Ridella and F.Rivieccio, "K-folds Generalization Capability Assessment for Support Vector Classifiers", *Proceeding of International Joint Conference on Neural Network*, Canada,2005,PP.855-858.

[5] D.Marquardt, *"An Algorithm for Least Squares Estimation of Non-Linear Parameter"*, J. Soc. Ind. Appl. Math., pp. 1963.

[6] L.Fausett, *"Fundamentals of Neural Networks Architecture.Algorithms and Applications"*, Pearson Prentice Hall, USA, 1994.

[7] Li-hua Chen, and Xiao-yun Zhang, "Application of Artificial Neural Network to Classify Water Quality of the Yellow River", *Journal 0f Fuzzy Information and Engineering*, Springer-Verlag, Jan 2009, pp. 15-23.

[8] Li, Y., Jiang, J.H., Chen, Z.P., Xu, C.J., Yu, R.Q.: *A New Method Based on Counter Propagation Network Algorithm for Chemical Pattern Recognition*, 1999, pp. 161–170.

[9] L.Khuan, N.Hamzah and R Jailani, "Prediction of Water Quality Index(WQI) Based on Artificial Neural Network(ANN)",*Conference on Research and Development Proceedings*, Malasia, 2002, pp. 157-161.

[10] L.Khuan, N.Hamzah and R Jailani, "Water Quality Prediction Using LS-SVM with Particle Swarm Optimization",*Second International Workshop on Knowledge Discovery and Data Mining*, China, 2009, pp. 900-904.

[11] Martin T.Hagen and Mohammad B.Menhaj, "Training Feedforward Networks with the Marquardt Algorithm", *IEEE Transactions on Neural Networks,*vol.5, no.6,Nov 1994, pp.989-993.

[12] M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail "The Use of a Neural Network Technique for the Prediction of Water Quality Parameters of Axios River in Northern Greece", *Journal 0f Operational Research*, Springer-Verlag, Jan 2005, pp. 115-125.

[13] S.Areerachakul and S.Sanguansintukul "Water Classification Using Neural Network: A Case Study of Canals in Bangokok, Thailand", *The 4th International Conference for Internet Technology and Secured Transactions (ICITST-2009)*, United Kingdom, 2009.

[14] S.H.Musavi and M.Golabi "Application of Artificial Neural Networks in the River Water Quality Modeling: Karoon River,Iran", *Journal 0f Applied Sciences*, Asian Network for Scientific Information, 2008, pp. 2324-2328.

[15] Simon Haykin, "*Neural Networks:A Comprehensive foundation second edition*", Pearson Prentice Hall, Delhi India, 2005.

[16] S.Lek, Guegan, J.F. (eds.): *Artificial Neural Networks: Application to Ecology and Evolution*. Springer, Berlin, 2000.

[17] Walley, W.J., DZeroski,S Biological monitoring "A Comparison between Bayesian, neural and machine learning methods of water quality classification",

International Symposium on Environmental Software System, 1996.

[18] Ministry of Natural Resources and Environment: http://www.mnre.go.th/

[19] Department of Drainage and Sewerage Bangkok Metropolitan Administration:http://dds.bangkok.go.th/ wqm/Thai/home.html

