Integration of unsupervised and supervised neural networks to predict dissolved oxygen concentration in canals

Sirlak Areerachakul*, Peraphon Sophatsathit, Chidchanok Lursinsap
Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Patumwan, Bangkok 10330, Thailand

ARTICLE INFO
Article history:
Received 14 December 2012
Received in revised form 1 April 2013
Accepted 2 April 2013
Available online 4 May 2013

Keywords:
Sub-space clustering
Clustering techniques
Supervised and unsupervised neural networks
Water management

ABSTRACT

The main focus of this paper was to devise a method to accurately predict the amount of dissolved oxygen (DO) in Bangkok canals at the present month based on the following 13 water quality parameters collected the previous month: temperature, pH value (pH), hydrogen sulfide (H₂S) content, DO, biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), total kjeldahl nitrogen (TKN), ammonia nitrogen (NH₃), nitrite nitrogen (NO₂⁻N), nitrate nitrogen (NO₃⁻N), total phosphorous (T-P), and total coliform (TC). Accurately predicting the amount of DO in a canal via scientific deduction is an important step in efficient water management and health care planning. We proposed a new technique that enhances the prediction accuracy by constructing a set of sub-manifolds of the prediction function by deploying unsupervised and supervised neural networks. The data were obtained from the Bangkok Metropolitan Administration Department of Drainage and Sewerage during the years 2007–2011. Comparisons between our proposed technique and other techniques using the correlation coefficient (R), the mean absolute error (MAE), and the mean square error (MSE) showed that our proposed approach with the sub-space clustering technique yielded higher accuracy than did other approaches without the sub-space clustering technique.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Surface water in rivers, canals and basins are usually subject to anthropogenic contamination. These human-caused activities are one of the most significant causes of water quality degradation, which can pose risks for public health, especially in urban areas. At the river basin scale, there is a need to establish a systematic data monitoring methodology to characterize water quality and to accurately analyze the collected data (Oliveira et al., 2005).

Bangkok is the capital city and economic center of Thailand. The rapid expansion of the city and its population prompted an increase in various municipal activities including commercial, industrial, and labor services that are the cause of accumulated environmental pollutants. In particular, the water quality of the interconnecting canals in the city has deteriorated to a point that the ecological system can no longer absorb and dilute such an overwhelming amount of pollutants into the water (Department of Drainage and Sewerage Bangkok Metropolitan Administration, 2008).

At present, the pollution level of the canals in most areas of Bangkok is very severe because the canals are still used as sewers for direct discharge of wastewater. Although there are legal regulations for large buildings regarding wastewater treatment and septic tanks in residential areas, the problem still prevails. Despite the regulations, some untreated wastewater is still discharged into public sewers, which, in turn, drains into the canals (Department of Pollution Control Bangkok, 2009). To measure this wastewater discharge, the city uses DO, which is one of the most common indicators of the aquatic ecosystem (Streeter and Phelps, 1925). The DO values range from 0 to 18 parts per million, far exceeding the range of most natural water systems, which is 5–6 parts per million. As a consequence, the increased DO cause excessive algae growth, which draws oxygen from the water.

The purpose of this paper is to develop a new and more accurate procedure for predicting the current concentration of DO in Bangkok canals using data collected from the previous month. There are several methods previously proposed to compute the DO concentration based on the deoxidation process in streams (Butcher and Covington, 1995), rivers (Cox, 2003), and lakes (Garcia et al., 2002). However, these water quality models are often complex and costly and require extensive amounts of data (Palani et al., 2009). The artificial neural network (ANN) is another successful analysis tool used in the field of water quality prediction and forecasting (Rankovic et al., 2010). Palani et al. (2008) applied the neural network models for the prediction and forecasting of selected seawater quality variables. Soypuk et al. (2003) used a neural network approach to compute the pseudo steady state time- and
space-dependent DO concentrations in three separate reservoirs having different characteristics with a limited number of input variables. Sengor et al. (2006) used a feed-forward neural network (FNN) to estimate the DO from limited input data. Kuo et al. (2007) applied the ANN model to predict the DO in the Te-Chi reservoir (Taiwan); the correlation coefficients between the predicted and observed DO values for the training and testing data sets were 0.75 and 0.72, respectively. Singh et al. (2009) computed the DO and BOD levels in the Gomti river (India) by using a three layer FNN with backpropagation learning; the coefficients of predicted and observed DO values for training, validating and testing data sets were 0.70, 0.74, and 0.76, respectively, and a sensitivity analysis was used to select the relevant input parameters. In another study, the FNN was applied to predict the DO in the Gruza reservoir, Serbia (Rankovic et al., 2010), but the accuracy of the results was not high enough for practical use, which may have been caused by improper training process.

In the current study, we also employed the technique of neural prediction, but we proposed new training procedures. The procedure consists of two dependent steps: (1) data clustering to form several sub-spaces and (2) manifold construction of the predicting function in each sub-space. The data are clustered by using the technique of unsupervised neural learning, and each manifold is constructed by a supervised neural network. By deploying these two steps, the error in constructing the manifold of the predicting function can be significantly reduced and the accuracy of the prediction can be clearly improved. The details of our concept are discussed in the next section.

2. Material and methods

2.1. Water quality data

The current study used the monthly water quality data obtained from the Bangkok Metropolitan Administration Department of Drainage and Sewerage for the years 2007–2011 from 276 sites covering 155 canals. A total of 13,846 data records were collected and used in the experiments, and each record consisted of 13 parameters: temperature (Temp), pH value (pH), hydrogen sulfide (H₂S), DO, BOD, chemical oxygen demand (COD), suspended solids (SS), total kjeldahl nitrogen (TKN), ammonia nitrogen (NH₃-N), nitrite nitrogen (NO₂⁻-N), nitrate nitrogen (NO₃⁻-N), total phosphorous (T-P) and total coliform (TC). The units for each surface water quality parameter are shown in Table 1 (Ministry of Natural Resource and Environment, 2009).

2.2. Studied problems

In this study, a prediction of the 1-month DO concentration in the Bangkok canals is considered. The canals of Bangkok are composed of many sites that form a complex network of water systems. The main focus is to predict the amount of DO in the current month based on the above 13 water quality parameters collected from the previous month. These parameters can be considered as a T-tuple vector and are treated as the input features in the prediction process, where T is the number of parameters used as input.

Our prediction problem is transformed into a problem of constructing a manifold of a predicting function or functional approximation in a high-dimensional space. To make the prediction accuracy as high as possible, the error associated with the constructed manifold must be minimized as much as possible. Our error minimization is based on the observation that the constructed manifold can be easily represented as a set of composite activation functions of all neurons in a supervised neural network. During the manifold construction, the amount of weight adjusted is controlled by the total error computed from the entire training data set scattering in the data space. Therefore, it is rather difficult to minimize the local error of the constructed manifold. To resolve the effect, all data in the data space must be locally clustered and used to construct a local manifold.

Let \( \mathbf{v}_i = [e_{i1}, e_{i2}, \ldots, e_{i13}]^T \) be the feature vector formed at time \( i \) and \( e_{ij} \) be the jth feature of vector \( i \). An amount of DO at time \( i \), denoted by \( o_i \), can be written as follows:

\[
o_i = f(\mathbf{v}_i) = \sum_{k=1}^{N} a_k e_{ik}, \quad 1 \leq k \leq N
\]  

where \( N \) is the number of feature vectors. The set of \( a_k \) represents the whole manifold of data space. To construct the function \( f(\cdot) \) with minimum error for every \( o_i \), the whole manifold must be partitioned into several sub-manifolds to eliminate the error effect caused by irrelevant feature vectors and their distribution. Each sub-manifold must be constructed by a function \( f(\mathbf{w}_k) \) for all \( \mathbf{v}_i \) distributed within the region of the sub-manifold. This observation leads to the following sub-problems.

2.3. Theoretical background

For this paper’s first problem, two clustering techniques were adapted. For the second problem, a feed-forward neural network was deployed to construct the sub-manifolds because of the neural network efficiency and adaptability to any data distribution. We briefly summarized the concepts of clustering and feed-forward neural network in the following sub-sections.

2.3.1. K-means algorithm

K-means algorithm (MacQueen et al., 1967) is a cluster analysis technique used as a partitioning method. It uses a similarity measure, usually base on L₁-norm or L₂-norm, to cluster groups of data. The K-means algorithm is composed of the following steps (Musavi and Golabi, 2008), assuming the given data set must be clustered into K groups.

1. Generate the locations of K center points in the data space. These K points represent the initial centroids.
2. Assign each data point to the closest centroid.
3. Recalculate the locations of the K centroids with respect to the assigned data groups from step 2.
4. Repeat steps 2 and 3 until all centroids no longer change their locations.

Table 1

List of surface water quality parameters.

<table>
<thead>
<tr>
<th>Name of parameters</th>
<th>Unit of measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>Celsius</td>
</tr>
<tr>
<td>pH value</td>
<td>Standard units</td>
</tr>
<tr>
<td>Hydrogen sulfide</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Biochemical oxygen demand</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Chemical oxygen demand</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Suspended solids</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Total kjeldahl nitrogen</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Ammonia nitrogen</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Nitrite nitrogen</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Nitrate nitrogen</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Total phosphorous</td>
<td>Milligrams per liter</td>
</tr>
<tr>
<td>Total coliform</td>
<td>Most probable number per 100 milliliter</td>
</tr>
</tbody>
</table>
The aim of this algorithm is to find $K$ centroids for the $K$ clusters. The algorithm minimizes the following objective function:

$$ J = \sum_{j=1}^{K} \sum_{i=1}^{N} \| \mathbf{v}_i - \mathbf{c}_j \|^2 $$

where $\mathbf{v}_i$ is the $i$th data point, $\mathbf{c}_j$ is the $j$th cluster centroid, $\| \cdot \|$ is $L_2$-norm between a data point $\mathbf{v}_i$ and the cluster centroid $\mathbf{c}_j$, and $N$ is the number of data points.

2.3.2. Fuzzy c-mean algorithm

The fuzzy c-mean (FCM) is a popular clustering technique for which a data point can be assigned to two or more clusters based on its cluster wise membership degrees. This method is frequently used for pattern recognition (Bezdek, 1978). The algorithm minimizes the following objective function:

$$ J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \| \mathbf{v}_i - \mathbf{c}_j \|^2 $$

where $N$ is the number of data points, $C$ is the number of clusters, $m$ is any real number greater than 1, $u_{ij}^m$ is the degree of membership of $\mathbf{v}_i$ in the cluster $j$, $\mathbf{v}_i$ is the $i$th data, $\mathbf{c}_j$ is the center of the $j$th cluster and $\| \cdot \|$ is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function as shown in Eq. (3). The updates on membership degree $u_{ij}^m$ and the $\mathbf{c}_j$ cluster centroid are given by

$$ u_{ij}^m = \frac{1}{\sum_{t=1}^{C} \left( \frac{\| \mathbf{v}_i - \mathbf{c}_j \|}{\| \mathbf{v}_i - \mathbf{c}_t \|} \right)^{(2/m) - 1}} $$

and

$$ \mathbf{c}_j = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{v}_i}{\sum_{i=1}^{N} u_{ij}^m}. $$

This iteration stops when

$$ \text{max} \{ ||u_{ij}^m(t + 1) - u_{ij}^m(t)|| \} < \varepsilon, $$

where $\varepsilon$ is a terminating criterion set between 0 and 1 and $u_{ij}^m(t)$ is the membership degree at iteration step $t$. Eventually, this procedure converges to a local minimum or a saddle of point of $J$. The algorithm is composed of the following steps:

1. Initialize matrix $\mathbf{U}(0) = [u_{ij}^m]$ and set $t = 0$.
2. At $t$-step, calculate the centers vectors $\mathbf{C}(t) = \{ \mathbf{c}_1(t), \mathbf{c}_2(t), \ldots, \mathbf{c}_C(t) \}^T$ based on $\mathbf{U}(t)$.
3. Update $\mathbf{U}(t)$ and set $t = t + 1$.
4. If max $\{ ||u_{ij}^m(t + 1) - u_{ij}^m(t)|| \} < \varepsilon$ then STOP otherwise return to step 2.

2.3.3. Compactness and separation quality measures

The criteria widely accepted for partitioning a data set into a number of clusters are the degree of separation of the clusters and their compactness. The optimum case implies that all parameters lead to making the partitions as similar as possible to the real partitions of the data set (Halkidi and Vazirgiannis, 1996). Additionally, a reliable quality assessment index should consider both the compactness and the separation; one of the quality measures that can be used in clustering is described as follows (Binbib et al., 2004; Halkidi and Vazirgiannis, 1996).

The compactness of the $i$th spatial data set $\mathbf{V}_i = \{ \mathbf{v}_j \in \mathbb{R}^n | 1 \leq j \leq N_i \}$, determined by $\sigma(\mathbf{V}_i)$, is computed as follows:

$$ \sigma(\mathbf{V}_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} (\mathbf{v}_j - \mathbf{v}_i)^T (\mathbf{v}_j - \mathbf{v}_i), $$

where $\mathbf{v}_i = (1/N_i) \sum_{j=1}^{N_i} \mathbf{v}_j$.

Assume a set of clusters $\{ \mathbf{V}_i \}_{i=1}^{K}$ such that $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2 \cup \ldots \cup \mathbf{V}_K$. After clustering, the following condition $\mathbf{V}_i \cap \mathbf{V}_j = \emptyset, i \neq j$ is expected to obtain. The total compactness of spatial data set with respect to $K$ clusters, denoted by $\sigma_r$, is defined as

$$ \sigma_r = \frac{1}{K} \sum_{i=1}^{K} |\sigma(\mathbf{V}_i)|. $$

The average compactness of $C$ clusters, Comp, is:

$$ \text{Comp} = \frac{\sigma_r}{K}. $$

The average scattering of data set compactness, Scatt_Comp, is:

$$ \text{Scatt_Comp} = \frac{\text{Comp} |\sigma(\mathbf{V})|}{K} $$

The more compact the clusters, the smaller the Scatt_Comp becomes. Thus, for a given spatial data set, a smaller Scatt_Comp indicates a more precise clustering scheme. The distance between clusters is defined by the average distance between the centroids of specified clusters as follows:

$$ d = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\| \mathbf{V}_i - \mathbf{V}_j \|}{K(K - 1)}. $$

The larger $d$ implies the higher degree of separation among clusters. A quality measure for clustering is defined as follows:

$$ CD = \frac{\text{Scatt_Comp}}{d}. $$

The CD index measures the distance between the two clusters for hierarchical clustering (Halkidi et al., 2001). The definition of CD indicates that both criteria of good clustering (i.e. compactness and separation) are properly combined, enabling reliable evaluation of the clustering results (Binbib et al., 2004). A small CD value indicates that all resulting clusters from the clustering scheme are compact and separated overall.

2.3.4. Multi-layer perceptron

A multi-layer perceptron (MLP) is extensively used to solve a number of different problems including pattern recognition and interpolation (Simon, 2005; Musavi and Golabi, 2008). A network may be constructed by stacking several neural layers. Theoretically, a 3-layer structured network is efficient enough for any function approximation problem. In this study, a network was comprised of three independent layers: an input layer, hidden layer or multiple hidden layers, and an output layer. The number of neurons in the input layer is equal to the number of features and the number of neurons in the output layer is equal to the number of target outputs in each training pattern. However, the number of neurons in the hidden layer cannot be known in advance because this number strongly depends upon the way the data are distributed; therefore, it is rather difficult to define this number prior to the training process. In this study, a trial-and-error process was adopted to find the most appropriate number of hidden neurons.

The neurons in different layers are interconnected with a set of weights. The output of each neuron is computed by a non-linear
activation function. In our study, we used only sigmoid activation function (Faussett, 1996) defined in the following equation:

\[ f(s) = \frac{1}{1 + e^{-s}}, \]  

where \( s \) is the dot product between the input vector and weight vector. During the learning process, the weights and bias are iteratively adjusted by minimizing the mean square error between the target output and the predicted output. Typically, the widely used training algorithm for neural networks is the backpropagation learning algorithm. But this learning rule is rather slow to make the error converge to a desired value. Thus we adopted Levenberg–Marquardt (LM) method as the learning rule in our experiment because of its faster convergence speed (Simon, 2005).

2.4. Proposed methodology

As previously mentioned in Section 2.2, the first problem concerns the technique of clustering the feature vectors according to their similarity. This problem may imply that the training feature space is partitioned into several connected sub-spaces. Because it is rather difficult to estimate the most appropriate number of clusters in our algorithm, we first denoted the number of clusters as a constant, \( K \). The actual value of \( K \) will be discussed in Section 3. For the second problem, the feature vectors in each sub-space are trained by using a feed-forward neural network to construct the manifold of the sub-space. The overview of our proposed algorithm to construct the DO predicting function is as follows:

**Algorithm 1 (Constructing DO predicting function).**

1. Normalize of feature vectors by using Algorithm 2.
2. Partition feature vector set \( V \) into training set \( V^{(t)} \) and testing set \( V^{(t)} \).
3. Apply \( K \)-mean clustering algorithm to group feature vectors \( V^{(t)} \) into \( K \) clusters, i.e. \( \{ V_1^{(t)}, V_2^{(t)}, \ldots, V_K^{(t)} \} \).
4. For each \( V_i^{(t)} , 1 \leq i \leq K \) do
   5. Get the distribution boundary in each dimension from each feature vector in \( V_i^{(t)} \).
   6. Train \( V_i^{(t)} \) by using a feed-forward neural network.
   7. Endfor
   8. For each \( V_i \in V^{(t)} \) do
      9. Identify the corresponding neural network of \( V_i \) with the corresponding network in step 6.
   10. Test and measure the prediction accuracy of \( V_i \) with the corresponding network in step 6.
   11. Endfor
12. Apply fuzzy \( c \)-mean clustering algorithm to group feature vectors \( V^{(t)} \) into \( C \) clusters, i.e. \( \{ V_1^{(t)}, V_2^{(t)}, \ldots, V_C^{(t)} \} \).
13. For each \( V_i^{(t)} , 1 \leq i \leq C \) do
      14. Get the distribution boundary in each dimension from each feature vector in \( V_i^{(t)} \).
      15. Train \( V_i^{(t)} \) by using a feed-forward neural network.
      16. Endfor
17. For each \( V_i \in V^{(t)} \) do
      18. Identify the corresponding neural network of \( V_i \) with the corresponding network in step 15.
      19. Test and measure the prediction accuracy of \( V_i \) with the corresponding network in step 15.
   20. Endfor

**Algorithm 2 (Normalizing feature values).**

1. \( \text{For} \ 1 \leq i \leq 13 \) do
2. Let \( a = \max (e_{ij}) \)  \( \text{min} (e_{ij}) \)
3. Let \( b = \frac{\text{min} (e_{ij})}{\max (e_{ij})} \)
4. For each \( e_{ij}, 1 \leq i \leq N \) do
   5. Compute \( e_{ij} = \frac{e_{ij} - a}{b} \)
   6. Endfor
   7. Endfor

The following three performance measures, i.e. correlation coefficient (\( R \)), mean absolute error (MAE), and mean square error (MSE) were used to evaluate the performance of our technique. The definitions of these measures are the following.

\[ R = \frac{\sum (Q_o - M_o)(Q_p - M_p)}{\sqrt{\sum (Q_o - M_o)^2 \sum (Q_p - M_p)^2}}, \]

\[ \text{MAE} = \frac{1}{N} \sum |Q_o - Q_p|, \]

\[ \text{MSE} = \frac{1}{N} \sum (Q_o - Q_p)^2, \]

where \( Q_o \) and \( Q_p \) are the observed and predicted values, \( N \) is the total number of data, \( M_o \) and \( M_p \) are the mean of the observed and predicted values.

3. Results and discussion

To evaluate the merit of our proposed technique, we conducted two separate experiments. The first experiment evaluated the accuracy of the prediction without the use of clustering so that only one neural network was deployed to construct the predicting function. The second experiment evaluated the accuracy of the prediction with a priori clustering as discussed in Section 2.3. The details and results of each experiment are summarized in the following sections.

3.1. Prediction of DO without clustering

In the first experiment, 13 previously collected water quality parameters were used to form the set of input-target pairs. The 13,846 total input-target pairs were separated into the training and testing subsets at the ratio of 70:30. Instead of conducting a direct experiment based on the 13 water quality parameters, two issues were focused on to clarify how significant each water quality parameter affects the accuracy of DO prediction. In the first issue, the correlation between each parameter from the previous month and the DO concentration in the present month was computed. Those highly correlated parameters were chosen as the inputs in the DO prediction process. In the second issue, all 13 water parameters were used as the inputs in the DO prediction process. The results from both issues were compared to determine which one achieved the highest accuracy. The correlation value of each parameter with respect to DO level is summarized in Table 2.

Our experiments found that the correlations computed from parameters SS and NH3-N are the lowest. To resolve our previously discussed issues, we conducted three experiments using different parameters selected from the original 13 water quality parameters. In the first experiment, all 13 parameters were used as the inputs.

**Table 2**

<table>
<thead>
<tr>
<th>Parameters in the month ((t-1))</th>
<th>Correlation with DO in the month (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>0.191</td>
</tr>
<tr>
<td>pH</td>
<td>0.493</td>
</tr>
<tr>
<td>H2S</td>
<td>0.425</td>
</tr>
<tr>
<td>DO</td>
<td>0.169</td>
</tr>
<tr>
<td>BOD</td>
<td>0.279</td>
</tr>
<tr>
<td>COD</td>
<td>0.073</td>
</tr>
<tr>
<td>SS</td>
<td>0.004</td>
</tr>
<tr>
<td>TRN</td>
<td>0.179</td>
</tr>
<tr>
<td>NH3-N</td>
<td>0.037</td>
</tr>
<tr>
<td>NO3-N</td>
<td>0.259</td>
</tr>
<tr>
<td>NO2-N</td>
<td>0.464</td>
</tr>
<tr>
<td>T-P</td>
<td>0.145</td>
</tr>
<tr>
<td>TC</td>
<td>0.218</td>
</tr>
</tbody>
</table>
In the second experiment, all parameters except the SS parameter were used as the inputs. In the last experiment, parameters SS and NH$_3$N were excluded, and only the 11 remaining parameters were used as the inputs. The experiments were conducted 10 times by random selection of the training and testing sets, and a neural network with LM optimization and sigmoid activation function was used as the classifier. The mean square error (MSE) and number of training epochs were used as the stopping criteria and set to 0.001 and 12,000, respectively. The number of hidden neurons was varied from 13 to 80 for each training subset. The optimal structure of the MLP neural network for each type of input is shown in Table 3. The variable NumInp denotes the number of water quality parameters (e.g. in the first row, the number of considered parameters is equal to 13). It can be seen that the experiment that uses all 13 parameters gives the highest $R$ correlation coefficient, as defined in Eq. (14).

### 3.2. Prediction of DO with clustering

The experiments from the previous section determined that all of the original 13 parameters are relevant to the DO level prediction; therefore, these parameters were used again in the experiment of this section. For the cluster analysis, the average of each water quality parameter on each site was computed. We employed the $K$-mean and fuzzy $c$-mean algorithms to generate the $K$ cluster sites and then constructed the $K$ ANN models to predict the DO in $K$ clusters based on the 13 water quality parameters for each site.

#### 3.2.1. Site clustering

In this study, compactness and separation were used to measure the capability of the generated centroids to represent the corresponding feature vectors in the clusters in both the $K$-mean and fuzzy $c$-mean algorithms. The number of clusters ($K$) was arbitrarily designated so different means were obtained among clusters (Kanungo et al., 2002). The maximum number of initial clusters ($K$) was set to 6, and the CD values in Eq. (12) were used to evaluate the compactness and separation quality. The smaller the obtained CD value, the better the compactness and separation of the clusters that can be achieved. Therefore, the optimum clustering configuration and number of clusters can be established by choosing the result with the smallest CD value. The CD values of various numbers of clusters by the $K$-mean and fuzzy $c$-mean algorithms are shown in Table 4.

Table 4 shows the comparisons between the CD values of the $K$-mean and Fuzzy $c$-mean clustering results based on the number of clusters ($K$) varied from 2 to 6 clusters. It can be seen that, for each number of clusters ($K$), the $K$-mean result has a lower CD value than that of the fuzzy $c$-mean result. Therefore, the $K$-mean algorithm provides better results than the fuzzy $c$-mean algorithm in all the data sets of this experiment. In addition, the CD value from the $K$-mean algorithm is the lowest (0.385) when $K$ is equal to 5. Thus the number of clusters in the $K$-mean algorithm was set to 5 for every site.

Additional information for the mean comparisons of the water quality parameters in five clusters is presented in Table 5. The mean value of water quality reflects the relationship among the parameters in each cluster. Cluster 2 shows the lowest DO and the highest BOD, indicating that the volume of oxygen contained in the water and the amount of oxygen held by the water is dependent on the water temperature, salinity, and pressure (Table 5). The amount of DO often determines the number and type of organisms living in that body of water; therefore, this cluster denotes the worst surface water quality. High COD values also indicate bad water quality, as does a high value of H$_2$S, as H$_2$S naturally occurs from decayed organic matter and has a rotten egg smell. TKN, which is the combination of NH$_3$N and organic nitrogen in the sample, is measured in milligrams per liter (mg/l). NO$_3$N originates from fertilizers, industrial sources (non-point sources), and municipal discharges (point sources) and can contaminate the surface water run-off. TC indicates the amount of microorganisms in the water, so high TC values indicate high amounts of microorganisms or bacteria and poor water quality. Table 5 shows that cluster 4 has the best water quality and cluster 2 has the worst water.

#### 3.2.2. DO prediction by $K$ ANN models

Based on the results from the previous section, five site clusters were established and the data in each site cluster were independently solved by each MLP network. The same training and testing
subsets discussed in Section 3.1 were used to create the model and to evaluate the performance. Similarly, the same type of activation function in the hidden and output layers as discussed in Section 3.1 was adapted in this experiment. The training process was terminated by using the mean square error measurement and the number of epochs, the values of which are shown in Table 6.

The selected training set used in Section 3.1 was also employed to determine the appropriate number of hidden neurons in five MLP models (Table 7). The number of hidden neurons varied from 13 to 80 neurons, and after the optimal number of hidden neurons in each MLP network was obtained, each MLP was trained by using the remaining training data with 10 times the weight initialization. Fig. 1(a)–(e) illustrate the comparison of predicted value and observed value of DO with 30 records in each cluster.

### Table 6

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Goal of minimizing mean square error</th>
<th>Number of epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.001</td>
<td>10,000</td>
</tr>
<tr>
<td>2</td>
<td>0.005</td>
<td>8000</td>
</tr>
<tr>
<td>3</td>
<td>0.003</td>
<td>8000</td>
</tr>
<tr>
<td>4</td>
<td>0.003</td>
<td>8000</td>
</tr>
<tr>
<td>5</td>
<td>0.001</td>
<td>10,000</td>
</tr>
</tbody>
</table>

3.3. Comparison between two approaches

The results of the DO predictions with and without the clustering technique on 10 data sets were compared by using the
average standard deviation of the three previously mentioned measures. We found that the DO prediction with the clustering technique provides more accurate average prediction values than those without the clustering technique in all three measures (Table 8). Furthermore, the standard deviation values of the three measures from the prediction with clustering technique are very small compared to the results from the prediction without the clustering technique. This comparison implies that the DO prediction by ANN models with the clustering technique is more consistent than is the DO prediction by ANN models without the clustering technique.

4. Conclusion

Using data collected from Bangkok canals during the years 2007–2011, this paper proposed a new procedure for predicting the amount of DO by employing a neural functional approximation based on 13 water quality parameters. The entire data set was partitioned into several connected sub-spaces using clustering techniques, and the data in each sub-space were used to construct a local manifold in the form of a neural network. Our proposed technique was applied to predict the current concentration of DO based on 13 data parameters from the previous month. The prediction error of our proposed technique is less than half of the error obtained from the prediction without clustering. Because the functionality of these results is not limited to the Bangkok area, our paper confirms that the combination of cluster analysis and neural functional approximation are a useful and efficient tool for managing natural resources and maintaining compliance with water management regulations and policy.

Acknowledgments

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

References


Streeter, B.W., Phelps, E.B., 1925. A Study of the Pollution and Natural Purification of the Ohio river III. United States Public Health Service, No. 146.