Ecological Modelling 261–262 (2013) 1–7

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

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

Sirilak 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 (H_2S) content, DO, biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), total kjeldahl nitrogen (TKN), ammonia nitrogen (NH_3N), nitrite nitrogen (NO_2N), nitrate nitrogen (NO_3N), 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 predicting 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.

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. Soyupak et al. (2003) used a neural network approach to compute the pseudo steady state time- and

* Corresponding author. Tel.: +66 8 1904 5654.
E-mail address: sirilak.ar@student.chula.ac.th (S. Areerachakul).

0304-3800/$ – see front matter
http://dx.doi.org/10.1016/j.ecolmodel.2013.04.002

Table 1
List of surface water quality parameters.

Name of parameters	Unit of measurement
Temperature	Celsius
pH value	Standard units
Hydrogen sulfide	Milligrams per liter
Dissolved oxygen	Milligrams per liter
Biochemical oxygen demand	Milligrams per liter
Chemical oxygen demand	Milligrams per liter
Suspended solids	Milligrams per liter
Total kjeldahl nitrogen	Milligrams per liter
Ammonia nitrogen	Milligrams per liter
Nitrite nitrogen	Milligrams per liter
Nitrate nitrogen	Milligrams per liter
Total phosphorous	Milligrams per liter
Total coliform	Most probable number per 100 milliliter

space-dependent DO concentrations in three separate reservoirs having different characteristics with a limited number of input variables. Sengorur 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 (H_2S), DO, BOD, chemical oxygen demand (COD), suspended solids (SS), total kjeldahl nitrogen (TKN), ammonia nitrogen (NH_3N), nitrite nitrogen (NO_2N), nitrate nitrogen (NO_3N), 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 $\mathbf{v}_i = [e_{i,1}, e_{i,2}, \ldots, e_{i,13}]^T$ be the feature vector formed at time i and $e_{i,j}$ be the jth feature of vector i. An amount of DO at time i, denoted by o_i, can be written as follows:

$$o_i = f(\mathbf{v}_k); \quad 1 \le k \le N \tag{1}$$

where N is the number of feature vectors. The set of o_i 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 $f(\mathbf{v}_k)$ for all $\mathbf{v}_k$ distributed within the region of the sub-manifold. This observation leads to the following sub-problems.

1 How can we cluster the set of $\mathbf{v}_k$ for $1 \le k \le N$ into several groups so that the function approximated from each group has a minimum error with respect to each target o_i?
2 For a given group of $\mathbf{v}_i$ obtained in problem 1, how can we construct or approximate this sub-manifold by function $f(\cdot)$ with a minimum error?

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 L_1-norm or L_2-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.