Applied Mathematical Modelling 39 (2015) 5907–5916

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Integrating Support Vector Regression with Particle Swarm Optimization for numerical modeling for algal blooms of freshwater ☆

Inchio Lou [a], Zhengchao Xie [a,*], Wai Kin Ung [b], Kai Meng Mok [a]

[a] *Faculty of Science and Technology, University of Macau, Macau*
[b] *Laboratory & Research Center, Macao Water Co. Ltd., Macau*

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 29 March 2015
Accepted 13 April 2015
Available online 11 May 2015

Keywords:
Algal bloom
Phytoplankton abundance
Support Vector Regression
Particle Swarm Optimization
Prediction and forecast models

ABSTRACT

Algae-releasing cyanotoxins are cancer-causing and very harmful to the human being. Therefore, it is of great significance to model how the algae population dynamically changes in freshwater reservoirs. But the practical modeling is very difficult because water variables and their internal mechanism are very complicated and non-linear. So, in order to alleviate the algal bloom problems in Macau Main Storage Reservoir (MSR), this work proposes and develops a hybrid intelligent model combining Support Vector Regression (SVR) and Particle Swarm Optimization (PSO) to yield optimal control of parameters that predict and forecast the phytoplankton dynamics. In this process, collected data for current month's variables and previous months' variables are used for model predict and forecast, respectively. In the correlation analysis of 23 water variables that monitored monthly, 15 variables such as alkalinity, Bicarbonate (HCO_3^-), dissolved oxygen (DO), total nitrogen (TN), turbidity, conductivity, nitrate, suspended solid (SS) and total organic carbon (TOC) are selected, and data from 2001 to 2008 for each of these selected variables are used for training, while data from 2009 to 2011 which are the most recent three years are used for testing. It can be seen from the numerical results that the prediction and forecast powers are respectively estimated at approximately 0.767 and 0.876, and naturally it can be concluded that the newly proposed PSO–SVR is working well and can be adopted for further studies.

1. Introduction

As a water pollution issue, freshwater algal bloom usually exists in eutrophic lakes or reservoirs due to excessive nutrients. Most species of algae (also called phytoplankton) can generate different cyanotoxins including *microcystins, cylindrospermopsis* and *nodularin*, which in turn will affect the water treatment processes and lead to negative impact on the health of public [1]. So, it is very important to understand the population dynamics of algae in the raw water storage units.

☆ This article belongs to the Special Issue: *ASPEC 2013 – 2013 International Applied Science and Precision Engineering Conference, October 2013 NanTou, Taiwan.*

* Corresponding author.
E-mail address: zxie@umac.mo (Z. Xie).

http://dx.doi.org/10.1016/j.apm.2015.04.001

For existing research, usually the major models of phytoplankton abundances simulation and prediction are based on a mechanistic approach in which the mechanism has to be relatively clear. But in practical modeling it is very difficult to model the in such a complicated system, because the system includes physical, chemical and biological processes and these processes have impact to each other as well. Because less background information and incomplete monitoring data are available, the highly nonlinear relationship between phytoplankton abundance and various water parameters is still unknown. So, it might be another option to do quantitative simulation models without considering mechanisms or simple time series models with simple structure. In existing works, these possible quantitative models have been applied for water quality simulation and prediction [2–5].

To overcome the difficulties in modeling complicated and non-linear water systems, it can be an alternative solution to adopt computational artificial intelligence as the efficient tools. In this approach it needs to consider time series effect for predicting and lack thereof for forecasting. For some existing works [6], in order to predict chlorophyll-a levels, the fundamental index of phytoplankton, principal component analysis (PCA) combined with multiple linear regressions (MLR) is applied for principle component regression (PCR). But it is not good enough to only apply PCR to solve the problem because of complex non-linearity. Then, artificial neural network (ANN) such as back propagation (BP) was applied to predict the algal bloom by assessing the eutrophication and simulating the chlorophyll-a concentration. Because ANN is self-adaptable, tolerant to errors and self-organization, it is better than PCR for non-linear simulation. ANN has been used for predicting the chlorophyll concentrations [7–9] and sludge bulking in wastewater treatment processes [10,11]. However, ANN still has some limitations such as a great amount of training data is needed, mainly depends on experience to set structure parameters, and is difficult to understand its internal working mechanism and interpret the data [6,12].

Compare the performance of the ANN and MLR, recently SVM has more advantage and developed as an efficient learning algorithm. The aim of SVR is to devise a computationally efficient way of learning well separating hyperplanes in a high dimensional feature space where good hyperplanes are ones optimizing the generalization bounds, and computationally efficient mean algorithms able to deal with sample sizes of the order of 100,000 instances [13]. Support Vector Machines (SVM) is learning machines implementing the structural risk minimization (SRM) inductive principle to obtain good generalization on a limited number of learning patterns. SRM involves simultaneous attempt to minimize the empirical risk and the VC (Vapnik–Chervonenkis) dimension. A version of a SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola [14]. This method is called Support Vector Regression (SVR). SVR is an artificial intelligent forecasting tool using a high dimensional feature space, which can model nonlinear relationships; the SVR embedded solution meaning is unique, optimal and unlikely to generate local minima; and it chooses only the necessary data points to solve the regression function, resulting in the sparseness of solution. However the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction (within a threshold ε). PSO is an emerging population-based optimization tool that simulates social behavior of biological organisms that move in groups, such as birds and fishes and developed by Dr. Kennedy and Dr. Eberhart in 1995 [15], inspired by the artificial life research results. PSO shares many similarities with evolutionary computation techniques. Each individual in PSO is assigned with a randomized velocity according to its own and its companions' flying experience. The individuals, called particles, are then flown through hyperspace. Successful applications of PSO in some optimization problems such as function minimization and NNs design, demonstrate its ability of global search [16,17]. The system is initialized with a population of random solutions and searches for optima by updating generations. However, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles.

The key contributions of this paper not only focus on the mathematical modeling itself, but also take the complete main factors that affect the algal blooms into consideration, by integrating all of those potential mechanistic blooming causative variables into both models, though only the data-driven models were applied. SVR has achieved great success in both academic and industrial platforms due to its many attractive features and promising generalization performance. When using SVR, the main problem is confronted: how to set the best kernel parameters. In application SVR, proper parameters setting can improve the SVR regression accuracy. However, inappropriate parameters in SVR lead to over-fitting or under-fitting. Different parameter settings can cause significant differences in performance. Therefore, selecting optimal hyper-parameter is an important stepin SVR design [18–20]. In order to improve the performance of SVR, PSO is applied to select the most appropriate training parameters of SVR. PSO is one of optimization techniques, which search for an optimal value of a complex objective function bird flocking and fish schooling. PSO needs only the fitness values to determine their search. However, considering the difficulty in obtaining the fitness value, due to impossibility to run a real system for each parameter combination, SVR would be an option as a fitness estimator. The aim of PSO is to shorten the time to compute and optimize the fitness value. The SVR would be used as an objective function for the PSO optimization process, to generate the outputs from the inputs. Some studies integrating PSO and SVM have been presented in the literatures [21–27]. However, there is still no research showing whether the hybrid model based on PSO and SVR can be applied in algal blooms problem that occurs in drinking water resources, which affects health of public. Furthermore, The up-to-date modeling literature only consider a limited couple of routine measured parameters, losing the key factors that are indeed important affect the algal blooms. We deliberately monitored all of relatively complete parameters over eleven years to develop a more accurate model using PSO and SVR for explaining the algal blooms mechanisms, which fills those research gaps in this study.