Journal of Computational and Applied Mathematics 309 (2017) 587–602

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

A hybrid wavelet kernel SVM-based method using artificial bee colony algorithm for predicting the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain)

P.J. García Nieto [a,*], E. García-Gonzalo [a], J.R. Alonso Fernández [b], C. Díaz Muñiz [b]

[a] *Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain*

[b] *Cantabrian Basin Authority, Spanish Ministry of Agriculture, Food and Environment, 33071 Oviedo, Spain*

ARTICLE INFO

Article history:
Received 21 October 2015
Received in revised form 18 January 2016

Keywords:
Support vector machines (SVMs)
Wavelet kernel
Artificial bee colony (ABC)
Microcystis aeruginosa
Woronichinia naegeliana
Regression analysis

ABSTRACT

Cyanotoxins, a kind of poisonous substances produced by cyanobacteria, are responsible for health risks in surface waters used for drinking or for recreation. Additionally, cyanobacteria blooms are often called blue–green algae. A harmful algal bloom (HAB) is an algal bloom that causes negative impacts to other organisms via production of natural toxins such as cyanotoxins. Consequently, anticipating its presence is a matter of importance to prevent risks. This study presents a novel hybrid algorithm, support vector machines with *Mexican hat wavelet kernel function* (wavelet SVMs) in combination with the *artificial bee colony* (ABC) technique, for predicting the cyanotoxin content from cyanobacterial concentrations determined experimentally in the Trasona reservoir (recreational reservoir used as a high performance training center of canoeing in the Northern Spain). This optimization technique involves kernel parameter setting in the SVM training procedure, which significantly influences the regression accuracy. Bearing this in mind, cyanotoxin contents have been predicted here by using the hybrid wavelet ABC–SVM-based model from the remaining measured water quality parameters (input variables) in the Trasona reservoir (Northern Spain) with success. In other words, the results of the present study are two-fold. In the first place, the significance of each biological and physical–chemical variable on the cyanotoxin content in the reservoir is presented through the model. Secondly, a predictive model able to predict the possible presence of cyanotoxins is obtained. The agreement of the wavelet ABC–SVM-based model with experimental data confirmed its good performance. Indeed, a coefficient of determination equal to 0.91 was obtained. Finally, conclusions of this innovative research work are exposed.

* Corresponding author. Tel.: +34 985 103417; fax: +34 985 103354.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García Nieto).

http://dx.doi.org/10.1016/j.cam.2016.01.045

1. Introduction

Algae are the simplest members of the plant kingdom, and cyanobacteria are the simplest of the algae. They have a considerable and increasing economic importance since they have both beneficial and harmful effects on human life. Cyanobacteria, also known as blue–green algae, is a phylum of bacteria that obtain their energy through photosynthesis. Indeed, cyanobacteria are aquatic and photosynthetic, that is, they live in the water, and can manufacture their own food. Because they are bacteria, they are quite small and usually unicellular, though they often grow in colonies large enough to be seen with the naked eye. Furthermore, they have a number of peculiarities that make them unique among the bacteria and algae: their evolutionary age and a modern oxygenic photosynthesis similar to other algae [1]. In order to work with sufficient accuracy with cyanobacteria, this complex problem requires the consideration of biovolume estimations because their morphological diversity ranging from small individual cells less than 1 μm in diameter to filaments greater than 20 μm in diameter and some millimeters in length [2]. Indeed, cyanobacteria and microalgae have a wide range of shapes and sizes. Therefore, cell counts are inadequate as a measure of the relative algal biomass. A standard biomass estimate is essential for comparing the relative contribution of different microalgae. As a result, depending on the equivalent geometric shapes for each microalgae different sets of equations are used to estimate their volume [2].

Cyanobacteria can be found in almost every terrestrial and aquatic habitat such as oceans, lakes, rivers, reservoirs, damp soil, temporarily moistened rocks in deserts, bare rock and soil, and even Antarctic rocks [2]. Sometimes, they become hazardous due to their uncontrolled growth giving place to the formation of extensive harmful algal blooms (HABs) [3]. In addition, some cyanobacteria produce toxins called *cyanotoxins* [4–6]. In this way, the association of toxicity with such blooms has frequently led to the closure of recreational waters when blooms are observed. Several cases of human poisoning have been documented, but a lack of knowledge prevents an accurate assessment of the risks [4–6]. At the same time, cyanotoxins are a focal environmental problem in reservoirs (e.g. the Trasona reservoir studied here) [5–7].

The main goal of this research work was to obtain the dependence relationship of cyanotoxin contents of the Trasona reservoir (output variable), expressed in micrograms per liter, as a function of the eight biological and fifteen physical–chemical input variables [8] indicated later. To fix ideas, the objective of this study is to evaluate the application of wavelet kernel support vector machines (SVMs) in combination with the Artificial Bee Colony (ABC) technique to identify cyanotoxins in the Trasona reservoir (Principality of Asturias, Northern Spain) (see Fig. 1(a) and (b)). SVM models are based on the statistical learning theory and are a new class of models that can be used for predicting values from very different fields [9–12]. SVMs are a set of related supervised learning methods used for classification and regression, and possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. The statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of SVMs [12].

In recent years, the combination of wavelet theories and SVMs has drawn considerable attention owing to its high predictive ability for a wide range of applications and better performance than other traditional learning machines. Indeed, Chen and Xie [13] used the dual-tree complex wavelet features and SVMs for pattern recognition. Yang and Wang [14] applied the wavelet technique with SVMs to DDoS (Distributed Denial of Service) intrusion detections. Widodo and Yang [15] established intelligent system for faults detection and classification of induction motor using wavelet support vector machines. Wu [16] proposed a new wavelet support vector machine to setup the nonlinear system of product sale series by combining the wavelet theory with SVMs. Tolambiya and Kalra [17] presented an effective image compression system that involved wavelet decomposition and relevance vector machine regression for compressing still images. Wei [18] presented a novel algorithm based on wavelet support vector machines (wavelet SVMs) for forecasting the hourly water levels at gauging stations.

In order to carry out the optimization mechanism corresponding to the kernel optimal hyperparameters setting in the SVM training, the artificial bee colony technique was used here with success. The artificial bee colony technique is an optimization algorithm based on the intelligent foraging behavior of honey bee swarms. Similar to other evolutionary computation algorithms such as particle swarm optimization (PSO) [19–21] or ant colony optimization [22] ABC exploits the model of social sharing of information [23–25]. According to previous researches, the SVM technique has been proved to be an effective tool to predict natural parameters, being successfully used in a wide range of environmental fields: forest modeling [26], solar radiation estimation [27,28], prediction of the air quality [29] and so on.

The Trasona reservoir, initially destined to industrial supply, is now complemented with a recreational utilization as a high performance training center of canoeing (see Fig. 1). From a biological point of view, it is a real *eutrophic ecosystem*. This ecosystem has been characterized for cyanobacterial HABs in certain periods, which sometimes have produced variable concentrations of cyanotoxins, mainly microcystins [3]. The two dominant species of the cyanobacteria community in the Trasona reservoir are *Microcystis aeruginosa* and *Woronichinia naegeliana*. It is well known that *M. aeruginosa* is potentially toxic and produces a type of toxin known as *microcystin*. Up to now, there is no evidence of the toxicity of the *W. naegeliana* in Spain and there is only a partial evidence of its toxicity outside Spain [5]. *Synergy* has been advanced as a hypothesis on how complex systems operate. Environmental systems may react in a nonlinear way to perturbations, so that the outcome may be greater than the sum of the individual component alterations. Synergistic response is a complex factor in environmental modeling [2,3].