Journal of Hydrology 527 (2015) 833–843

Contents lists available at ScienceDirect

Journal of Hydrology

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

Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand

Roohollah Noori [a,*], Hund-Der Yeh [b], Maryam Abbasi [c], Fatemeh Torabi Kachoosangi [d], Saber Moazami [e]

[a] Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Tehran, Iran
[b] Institute of Environmental Engineering, National Chiao Tung University, Taiwan
[c] Griffith University, Australia
[d] Department of Civil Engineering, College of Engineering, Tehran West Branch, Islamic Azad University, Tehran, Iran
[e] Department of Civil Engineering, College of Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr, Tehran, Iran

ARTICLE INFO

Article history:
Received 9 November 2014
Received in revised form 9 May 2015
Accepted 23 May 2015
Available online 1 June 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
BOD_5
Support vector machine
Uncertainty analysis
River

SUMMARY

Uncertainty is considered as one of the most important limitations for applying the results of artificial intelligence techniques (AI) in water quality management to obtain appropriate control strategies. In this research, a proper methodology was proposed to determine the uncertainty of support vector machine (SVM) for the prediction of five-day biochemical oxygen demand (BOD_5). In this regard, the SVM model was calibrated using different records for many times (here, 1000 times), to investigate model performance according to calibration pattern changes. Therefore, to implement the random selection of calibration patterns for several times, an alternative database was required. By this methodology, the parameters of SVM model will be obtained 1000 times, giving various predicted BOD_5 values each time. To evaluate the SVM model's uncertainty, the percentage of observed data bracketed by 95 percent predicted uncertainties (95PPU) and the band width of 95 percent confidence intervals (*d*-factor) were selected. Findings indicated that the SVM model was more sensitive to capacity parameter (*C*) than to kernel parameter (Gamma) and error tolerance (Epsilon). Besides, results showed that the SVM model had acceptable uncertainty in BOD_5 prediction. It is notified that the novelty of the presented methodology is beyond a mere application to water resources, and can also be used in other fields of sciences and engineering.

1. Introduction

Although there are many publications on the uncertainty analysis of deterministic water quality models (Canale and Seo, 1996; Carroll and Harms, 1999; Mailhot and Villeneuve, 2003; Wagener and Gupta, 2005; Gupta et al., 2006; Wang et al., 2006; Sin et al., 2009; Cea et al., 2011; Srivastava et al., 2014), few have focused on determining the uncertainty of artificial intelligence (AI) techniques, for the prediction of water quality parameters. AI models have been successfully applied in water quality management studies, during the past few decades (Biron and Lingireddy, 1999; Chang and Chen, 2001; Chau, 2006; Chang and Chang, 2006; Bashi-Azghadi et al., 2010a; Najah et al., 2012; Chang et al., 2013; Antanasijević et al., 2014).

The main reason of uncertainty in AI techniques is their structure and algorithm. They are usually known as data-driven models and their performance is highly dependent on the selected pattern for calibration. Therefore, these models are always sensitive to input data, and the results vary by changing the selected calibration pattern. This fact causes managers to face many problems in making appropriate decisions for monitoring water quality network. Generally, less attention has been paid to studying the uncertainty analysis of AI techniques, and few resources are available in this field. Monte-Carlo random sampling process for uncertainty analysis of AI techniques was first used in adaptive neuro-fuzzy inference system (ANFIS), to evaluate the flow from local source in a river basin (Aqil et al., 2007). Noori et al. (2010) investigated the uncertainty of artificial neural network (ANN) and ANFIS models for the prediction of carbon monoxide concentration in the atmosphere of Tehran. In other works, the uncertainty of ANFIS and ANN models was investigated for most of the water quality parameters, five-day biochemical oxygen demand (BOD_5), by Noori et al. (2013a,b), respectively.

Proper capabilities of support vector machine (SVM) in modeling nonlinear and complex phenomena have caused it to be used by many researchers, in the past decades. These efforts have been limited to the application of this model for the prediction of

* Corresponding author at: Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Enqelab Avenue, Qods Street, Azin Alley, P.O. Box 14155-6135, Tehran, Iran. Tel.: +98 9128352488; fax: +98 2166407719.
E-mail address: noor@ut.ac.ir (R. Noori).

http://dx.doi.org/10.1016/j.jhydrol.2015.05.046
0022-1694/

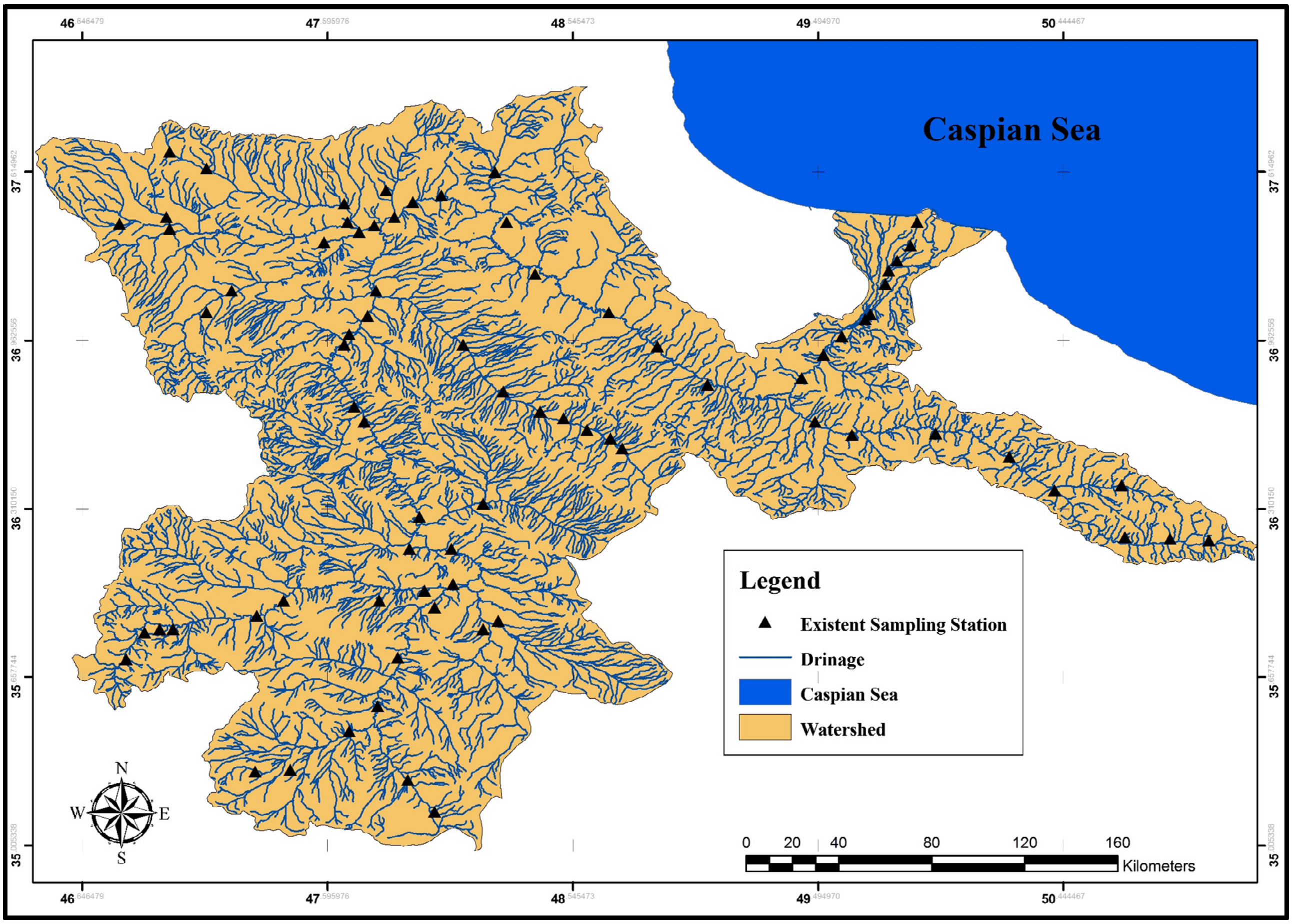

Fig. 1. The Sefidrood River basin and sampling sites.

various phenomena, in different fields of science and engineering (Najah et al., 2009; Bashi-Azghadi and Kerachian, 2010b; Singh et al., 2012, 2014). SVM technique has a different structure from ANN and ANFIS models. This fact may cause difficulties in providing an appropriate methodology for determining the uncertainty of this model. The basic idea of SVM is to use linear model to implement nonlinear class boundaries through some nonlinear mapping of the input vector into the high-dimensional feature space (Wang et al., 2009). The SVM requires no pre-determined network topology, while providing network topology for the uncertainty analysis of ANN and ANFIS models is crucial. In addition, compared to ANN and ANFIS models, SVM has a strong theoretical background which must be considered in its uncertainty analysis. Hence, determining the uncertainty of SVM is different from that of other AI techniques such as ANN and ANFIS models. Therefore, developing a method to analyze the uncertainty of SVM model has not yet been reported, not only for the prediction of water quality parameters such as BOD_5, but also in various other fields of science and engineering. Considering the mentioned facts, this study aims to investigate the uncertainty of SVM model's results, in the process of BOD_5 determination.

2. Study area and data

Data from Sefidrood River basin was used to calibrate and test the developed methodology for SVM uncertainty determination. The basin, drained by Sefidrood River, is located in the northwest of Iran (Fig. 1). The data was collected from 94 water quality monitoring stations along the river and its tributaries in spring, summer, fall and winter. Fig. 2 illustrates the variation of the used data along Sefidrood River. Generally, BOD_5 is drastically affected by other water quality parameters like DO. In Sefidrood basin, electrical conductivity (EC) is mostly affected by pollutants discharged to the river. In other words, EC is of anthropogenic source due to agricultural, industrial and urban wastewater discharges to the river. These effluents also contain a great amount of nitrates and phosphorous. Therefore, the fluctuations of EC, nitrate, and phosphorous highly affect BOD_5 variations in Sefidrood River. In addition, the locations of sampling stations, i.e. the distances from discharging points, are effective on the amount of measured BOD_5. Hence, in this study, to develop a proper SVM model for online prediction of maximum BOD_5, input vectors are selected as: the location of stations {latitude (Lat) and longitude (Lon)}, dissolved oxygen (DO) {minimum (DO_{min}), maximum (DO_{max}), and average (DO_{mean})}, EC {maximum (EC_{max}) and average (EC_{mean})}, nitrate (NO_3^-) {maximum (NO^-_{3max}) and average (NO^-_{3mean})}, and total phosphorus (TP) {maximum (TP_{max}) and average (TP_{mean})}. It must be noted that according to lack of other water quality data in the study area, in this research, only the aforementioned parameters are used for BOD_5 prediction. Although measuring nitrate and phosphorus is not as easy as measuring EC, DO and locating sampling network, the aim of this research is to eliminate time restrictions in measuring BOD_5, rather than to merely simplify the measurement. Also, maximum and minimum amounts of the mentioned parameters are used. Although this can cause a correlation between the variables, it includes useful information for BOD_5 prediction.