Journal of Hydrology 534 (2016) 104–112

Contents lists available at ScienceDirect

Journal of Hydrology

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

Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution

Ozgur Kisi [a,*], Kulwinder Singh Parmar [b]

[a] Canik Basari University, Architecture and Engineering Faculty, Civil Eng. Dept., 55080 Samsun, Turkey
[b] Dr. B. R. Ambedkar National Institute of Technology, Department of Mathematics, Jalandhar, India

ARTICLE INFO

Article history:
Received 3 August 2015
Received in revised form 28 November 2015
Accepted 12 December 2015
Available online 29 December 2015
This manuscript was handled by Laurent Charlet, Editor-in-Chief, with the assistance of Antonino Cancelliere, Associate Editor

Keywords:
Chemical oxygen demand (COD)
Mathematical modeling
Estimation
Least square support vector machine
Multivariate adaptive regression splines
M5 model tree

SUMMARY

This study investigates the accuracy of least square support vector machine (LSSVM), multivariate adaptive regression splines (MARS) and M5 model tree (M5Tree) in modeling river water pollution. Various combinations of water quality parameters, Free Ammonia (AMM), Total Kjeldahl Nitrogen (TKN), Water Temperature (WT), Total Coliform (TC), Fecal Coliform (FC) and Potential of Hydrogen (pH) monitored at Nizamuddin, Delhi Yamuna River in India were used as inputs to the applied models. Results indicated that the LSSVM and MARS models had almost same accuracy and they performed better than the M5Tree model in modeling monthly chemical oxygen demand (COD). The average root mean square error (RMSE) of the LSSVM and M5Tree models was decreased by 1.47% and 19.1% using MARS model, respectively. Adding TC input to the models did not increase their accuracy in modeling COD while adding FC and pH inputs to the models generally decreased the accuracy. The overall results indicated that the MARS and LSSVM models could be successfully used in estimating monthly river water pollution level by using AMM, TKN and WT parameters as inputs.

1. Introduction

Water is an essential element for life, but we are living in a water starved world. River water and ground water play an important role to fulfill the requirement of quality water around the globe. The quality river water also affects the quality of ground water (Parmar and Bhadwaj, 2013). In old times, numerous civilizations developed on the river banks just for the availability of fresh and pure water, but unfortunately now the rivers are influenced by urbanization, industrialization and other human activities. Contamination in stream water draws consideration of government, public, NGO's and environmentalists in India and the world over. Numerous rivers have been dying at an alarming rate because of the quality of water (Parmar et al., 2009; Kisi et al., 2012). Water quality and its enhancement have a close connection with the presence of chemical oxygen demand (COD). Oxygen concentration act as an important indicator of the water quality (Hanbay et al., 2009).

There are many forecast models, which have been developed for prediction of long-term precipitation (Doyle and Barros, 2011; Dokmen and Aslan, 2013). In these methods, time series modeling on the ground of statistics has been used. Statistical modeling has many advantages over mathematical models. But the shortcomings of the statistical approach include handling nonlinear characteristics of data because the statistical models are usually based on the linear correlations of the data can be expressed with a correlation coefficient. To overcome the shortcomings of the statistical methods, least square support vector machine (LSSVM), multivariate adaptive regression splines (MARS), M5 model tree, models are developed to address the nonlinearity of data (Nayak et al., 2004; Partal and Kisi, 2007; Wieland and Mirschel, 2008; Hanbay et al., 2009; Kisi, 2009, 2013; Alves et al., 2011; Maheshwaran and Khosa, 2012, 2013; Kisi and Tombul, 2013; Soni et al., 2014; Cheng and Cao, 2014; Bhardwaj and Parmar, 2013, 2015).

LSSVM model is based on kernel methods, which have proved capable of estimating more accurate than different techniques, for example, linear models ARIMA, neural networks, ANFIS, neuro-fuzzy systems, in terms of various different assessment measures during both the validation and test stages (Hong and Pai, 2006; Xu et al., 2006; Liu et al., 2007; Wang et al., 2009). The

* Corresponding author.
E-mail addresses: okisi@basari.edu.tr (O. Kisi), kulmaths@gmail.com, kulmaths@ipu.ac.in (K.S. Parmar).

http://dx.doi.org/10.1016/j.jhydrol.2015.12.014
0022-1694/

primary inspiration for utilizing kernel techniques as a part of the field of time series prediction is their capacity to estimate time series data precisely when the fundamental model could be non-linear, non-stationary and not characterized a priori (Sapankevych and Sankar, 2009). Multivariate adaptive regression splines (MARS), is a new artificial intelligence method. It was initially presented by Friedman (1991). MARS has been observed to be a quick, flexible and precise technique for forecasting continuous and binary output variables. MARS models develop this useful functional relation from a set of coefficients and basis functions from the regression data. The fundamental favorable position of MARS is that the relationship of the MARS models is additive and interactive, which includes fewer variable interactions (Yang et al., 2003, 2004; Leathwick et al., 2006; Lee et al., 2006). Currently, M5 model trees have been utilized effectively for flood forecasting (Solomatine and Xue, 2004), water level–discharge relationship (Bhattacharya and Solomatine, 2005), rainfall–runoff modeling (Solomatine and Dulal, 2003), sedimentation modeling (Bhattacharya and Solomatine, 2006), and ET_0 modeling (Pal and Deswal, 2009). Pal and Deswal (2009) researched the capability of M5 model tree based regression approach to model daily ET_0 utilizing four inputs, involving solar radiation, average air temperature, average relative humidity, and average wind speed.

There is an urgent need to take important steps in maintaining the quality of river water (Georgakakos et al., 2012; Shamir et al., 2015). To diagnose the problem of quality of river water firstly we should know what is the present and future pollution level of water. So forecasting of river water quality is a vital task. In this paper, long term comparative analysis of river Yamuna at Delhi, studied using LSSVM, MARS and M5 model tree prediction models. Monthly average values of last 10 years of water quality parameter COD have been considered for study. This work is noble as 40% of the population of India depends on Yamuna river water for drinking, agriculture and the daily uses. So the quality of river water is directly related to the health of the Nation. The results of these models are useful for making future government policies and also for river water management of the region. This model will provide a great contribution in the field of prediction modeling.

2. Data and methodology

Sample site, Nizamuddin of Yamuna River in Delhi was chosen according to utilization of river water. 70% of water used in Delhi depending upon Yamuna River. Yamuna River is the largest tributary river of the Ganga in northern India (Fig. 1). It originates from the Yamunotri Glacier at a height of 6387 m on south western slopes of Banderpooch peaks (38°59′N, 78°27′E) in the lower Himalayas in Uttarakhand. It travels a total length of 1376 km by crossing several states, Uttarakhand, Haryana, Himachal Pradesh, Delhi, Uttar Pradesh and has a mixing of drainage system of 366,233 km^2 before merging with Ganga at Allahabad. It contributes 40.2% of the entire Ganga Basin (CPCB, 2006). Sample site, Nizamuddin is approximately 14 km downstream from the Wazirabad barrage at Delhi, Capital of India. The distance from Hathnikund to Wazirabad is 224 km. The water quality at Nizamuddin has the impact of industrial and sewerage discharge from Haryana and Delhi. Monthly average COD data (10 years for the period 1999 Jan–2002 Apr) observed by the Central Pollution Control Board (CPCB) have been considered for the present study. The basic statistical parameters of the COD data are presented in Table 1. It is clear from the table that each data set has a different range and skews.

2.1. Least square support vector machine

Least square support vector machine (LSSVM) models are used to approximate the nonlinear relationship between input variables and output variables with certain accuracy (Suykens, 2001; Smola and Bernhard, 2004). LSSVM, originated from support vector machines (SVM) is a powerful methodology for solving problems in non-linear classification, function estimation and regression. The SVM is based on the principle of Structural Risk Minimization (SRM) and minimizes the expected error of a learning tool and so reduces the problem of over-fitting. This SVM is a supervised learning technique which is developed at AT&T Bell Laboratories by Vladimir Vapnik and his co-workers 1995 (Cortes and Vapnik, 1995). This method has been applied in pattern recognition, signal processing and non-linear regression estimation. Least squares support vector machines (LS-SVM) were proposed by Suykens and Vandewalle in 1999 and has been employed in chaotic time series prediction (Suykens and Vandewalle, 1999). LS-SVM uses a set of linear equations for training while SVM uses a quadratic optimization problem, this is the major difference between these two. LSSVM is computationally much less extensive when compared with conventional modeling methods such as multivariate linear regression (MLR), back propagation neural networks (BPNN) and partial least square regression (PLS).

Consider a given training set $\{p_k, q_k\}_{k=1}^{N}$ with input data $p_k \in R^n$ and output data $q_k \in R$ with class labels $q_k \in \{-1, +1\}$ and linear classifier

$$q(p) = sign[w^T p + b] \tag{1}$$

When the data of the two classes are separable one can say

$$\left\{ \begin{array}{ll} w^T p_k + b \geqslant +1, & \text{if } q_k = +1 \\ w^T p_k + b \leqslant -1, & \text{if } q_k = -1 \end{array} \right\} \tag{2}$$

These two sets of inequalities can be combined into one single set as follows:

$$q_k[w^T p_k + b] \geqslant 1, \quad k = 1, 2, 3, \ldots, N \tag{3}$$

SVM formulations are done within a context of convex optimization theory. The general methodology is to start formulating the problem as a constrained optimization problem, next formulate the Lagrangian and then take the conditions for optimality, finally solve the problem in the dual space of Lagrange multipliers. With resulting classifier

$$q(p) = sign\left[\sum_{k=1}^{N} \alpha_k q_k p_k^t p + b\right] \tag{4}$$

This linear SVM classifier was extended to non-separable case by Cortes and Vapnik (1995). It is done by taking an additional slack variable in the problem formulation. One modifies the set of inequalities into

$$q_k[w^T p_k + b] \geqslant 1 - \xi_k, \quad k = 1, 2, 3, \ldots, N \tag{5}$$

The LS-SVM uses equality type constraints instead of inequalities as in the classic SVM approach. This reformulation significantly simplifies a problem such that the LS-SVM solution follows directly from solving a set of linear equations rather than from a convex quadratic program. For a LS-SVM classifier, in the primal space it takes the form,

$$q(p) = sign[w^T p + b] \tag{6}$$

where b is a real constant. For nonlinear classification, the LS-SVM classifier in the dual space takes the form

$$q(p) = sign\left[\sum_{k=1}^{N} \alpha_k q_k K(p, p_k) + b\right] \tag{7}$$

where α_k is the positive real constants and b a real constant, in general, $K(p_k, p) = \langle \phi(p_k), \phi(p) \rangle, \langle \bullet, \bullet \rangle$ the inner product, and $\phi(p)$ the nonlinear map from original space to high-dimensional space. For function estimation, the LS-SVM model takes the form