An Improved Hybrid ARIMA and Support Vector Machine Model for Water Quality Prediction

Yishuai Guo[1,2], Guoyin Wang [2,*], Xuerui Zhang [2], and Weihui Deng [2]

[1] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
guoyishuai@cigit.ac.cn
[2] Institute of Electronic Information & Technology, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 401122, China
Wangguoyin@cigit.ac.cn

Abstract. Traditionally, the hybrid ARIMA and support vector machine model has been often used in time series forecasting. Due to the unique variability of water quality monitoring data, the hybrid model cannot easily give perfect forecasting. Therefore, this paper proposed an improved hybrid methodology that exploits the unique strength in predicting water quality time series problems. Real data sets of water quality provided by the Ministry of Environmental Protection of People's Republic of China during 2008-2014 were used to examine the forecasting accuracy of proposed model. The results of computational tests are very promising.

Keywords: ARIMA, Support vector machine, Time series forecasting, Water quality prediction.

1 Introduction

The water quality problem is a subject ongoing concern. Deterioration of water quality has initiated serious management efforts in many countries [1]. Most acceptable ecological and water related decisions are difficult to make without careful modeling, prediction and analysis of river water quality for typical development scenarios [2]. Accurate predictions of future phenomena are the lifeblood of optimal water resources management in a watershed. So far, two kind of approach have been proposed for water quality prediction [3]. One kind is the based on the mechanism of movement, physical, chemical and other factors in the water and has been widely employed in different basins [4]. But the mechanistic models usually need complete observed data and mechanism knowledge, of which are difficult to get [5]. Another kind is the models based on statistics and artificial intelligence. The rapid development of artificial intelligence provides us with more approaches for regression and better accuracy under varies situations [6-7]. For example, the support vector machine (SVM)

* Corresponding author.

D. Miao et al. (Eds.): RSKT 2014, LNAI 8818, pp. 411–422, 2014.
DOI: 10.1007/978-3-319-11740-9_38

[8-10] has been widely used for prediction and forecasting in water resources and environmental engineer.

Computer science and statistics have improved modeling approaches for discovering patterns found in water resources time series data [1]. Much effort has been devoted over the past several decades to the development and improvement of time series prediction models. One of the most important and widely used time series model is the autoregressive integrated moving average (ARIMA) model [11].

Before 2005, most of the studies reported above were simple applications of using traditional time series approaches and support vector machine [12-13]. Recently, there have been several studies suggesting hybrid models, combining the ARIMA model and support vector machine [14-17]. However, many of the real-life time series are extremely complex to be modeled using simple approaches especially when high accuracy is required. This study presents an improved hybrid model of ARIMA and SVMs to solve the water quality prediction problem.

2 Hybrid Model in Forecasting

2.1 ARIMA Model

In an autoregressive integrated moving average model (ARIMA), the future value of a variable is assumed to be a linear function of several past observations and random errors [18]. In an ARIMA model, the future value of a variable is supposed to be a linear combination of past values and past errors, expressed as follows

$$y_t = \theta_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}, \quad (1)$$

Where y_t is the value of observations and ε_t is the random error at time t, ϕ_i and θ_j are the coefficients, p and q are integers that are often referred to as autoregressive and moving average polynomials, respectively. Basically, this method has three phases: model identification, parameter estimation and diagnostic checking.
For example, the ARIMA (1, 0, 1) can be represented as follows

$$y_t = \theta_0 + \phi_1 y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}. \quad (2)$$

The residuals are modeled by the ARIMA can be represented as follows

$$\varepsilon_{1t} = y_t - y_{1t}. \quad (3)$$

Where ε_{1t} is the error of ARIMA model at time t; y_t is the value of observations at time t; y_{1t} is the value of prediction of ARIMA at time t.

The ARIMA model is basically a data-oriented approach that is adapted from the structure of the data themselves.