Journal of Environmental Management 196 (2017) 110–118

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead $PM_{2.5}$ concentration forecasting

Mingfei Niu [a], Kai Gan [a], Shaolong Sun [b, *], Fengying Li [c]

[a] School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China
[b] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[c] School of Mathematics and Computer Science, Ningxia Normal University, Guyuan 756000, China

ARTICLE INFO

Article history:
Received 18 July 2016
Received in revised form
22 February 2017
Accepted 26 February 2017

Keywords:
$PM_{2.5}$ concentration forecasting
Decomposition-ensemble learning paradigm
EEMD
PSR
LSSVM

ABSTRACT

$PM_{2.5}$ concentration have received considerable attention from meteorologists, who are able to notify the public and take precautionary measures to prevent negative effects on health. Therefore, establishing an efficient early warning system plays a critical role in fostering public health in heavily polluted areas. In this study, ensemble empirical mode decomposition and least square support vector machine (EEMD-LSSVM) based on Phase space reconstruction (PSR) is proposed for day-ahead $PM_{2.5}$ concentration prediction, according to the application of a decomposition-ensemble learning paradigm. The main methods of the proposed model mainly include: first, EEMD is presented to decompose the original data of $PM_{2.5}$ concentration into some intrinsic model functions (IMFs); second, PSR is applied to determine the input form of each extracted component; third, LSSVM, an effective forecasting tool, is used to predict all reconstructed components independently; finally, another LSSVM is employed to aggregate all predicted components into ensemble results for the final prediction. The empirical results show that this proposed model can outperform the comparison models and can significantly improve the prediction performance in terms of higher predictive and directional accuracy.

1. Introduction

Air-quality issues related to economic development have received wide attention around the world. One air-quality indexes, fine particulate matter ($PM_{2.5}$), is the floorboard of solid or liquid droplets and is less than or equal to 2.5 μm in the air. Higher concentration of $PM_{2.5}$ in the air, correspond to more serious air pollution. People not only discharge $PM_{2.5}$ directly, but also discharge pollutants that can be transformed into $PM_{2.5}$ in the air. Emissions mainly originate from combustion processes, such as the burning of fossil fuels and biomass (Chen et al., 2014; Li et al., 2015; Mardones and Sanhueza, 2015). $PM_{2.5}$ can exist in a large number of poison and can stay in the atmosphere for a long time. Consequently, $PM_{2.5}$ greatly influences human's health and the quality of the atmospheric environment.

China is experiencing rapid development of economy, and as a result, urban air pollution is becoming a serious environmental issue. A significant positive correlation between urbanization indicators and $urban_{PM2.5}$ shows that urban areas have considerable impact on $PM_{2.5}$ concentration (Han et al., 2014; Wu et al., 2015). Moreover, population distribution is not uniform; thus, different regions suffer from varying degrees of environmental pollution. The East China Plain and Taklimakan desert, were found to have high concentrations of fine particulate matter ($PM_{2.5}$) (Han et al., 2015). The results show that $PM_{2.5}$ concentration has large relationship with local economic development level. Since fine particulate matter enters directly into the upper respiratory tract and bronchus, and can lead to asthma, angiocarpy disease and some other ailments (Huang et al., 2012; Schweizer and Cisneros, 2014). Therefore, meteorological and medical experts believe that haze caused by fine particulate matter causes more damage to human health than sandstorms. To some extent, air pollution is inevitable as China continues to develop economically. However, this does not mean that there is no means to prevent and control pollution emissions. An integrative approach to manage and control $PM_{2.5}$ pollution in China, was proposed through a collaboration among

* Corresponding author.
E-mail addresses: gank14@lzu.edu.cn (K. Gan), sunshl13@lzu.edu.cn (S. Sun).

http://dx.doi.org/10.1016/j.jenvman.2017.02.071
0301-4797/

academics, government officials, and actors from industry (Pui et al., 2014).

$PM_{2.5}$ concentration is affected by festivals, traffic, and other factors. Several methods have been applied to forecast $PM_{2.5}$ concentration, which can be divided into two broad categories: statistical methods and machine-learning techniques (Niu et al., 2015). Dong et al. (2009) used hidden semi-Markov models (HSMMs) to forecast $PM_{2.5}$ concentration where the parameters of HSMMs were evaluated by a modified forward–backward algorithm. The HSMMs model overcomes the disadvantages of hidden Markov models (HMMs) which has limitations in the temporal structures of a forecast model. Moreover, Sun et al. (2013) presented a hidden Markov model that relies on different distribution of $PM_{2.5}$ emission. A standard HMMs model with log-normal, gamma and generalized extreme value (GEV) distributions was utilized to establish an early warning system. Form the above literature, it is not hard to find that statistical models are based on a sequence of assumptions and applied to small amount of data, and also have their own limitations on solving the nonlinear time series forecasting (Goyal et al., 2006; Li et al., 2011; Jian et al., 2012).

To overcome the shortages in the statistical studies, machine-learning techniques have been proposed to address nonlinear time series forecasting. In recent years, with an increase in data complexity, many machine-learning models have been proposed and applied to $PM_{2.5}$ concentration forecasting. Artificial neuron network (ANN) models, based on self-learning and complex mathematical framework, have been extensively used to forecast $PM_{2.5}$ concentration (Niska et al., 2004; Sousa et al., 2007; Qin et al., 2014; Babu and Reddy, 2014; Mishra et al., 2015; Wang et al., 2015a; Adams and Kanaroglou, 2016; Catalano et al., 2016). These models can capture more useful information from complex nonlinear relationships. A model combining ARIMA and ANN to forecast particulate matter in Temuco, Chile, was proposed by Díaz-Robles et al. (2008) and not only can capture 80% and 100% of pre-emergency episodes and alerts, but can also be used to forecast air quality in other districts. Voukantsis et al. (2011) used specific computational intelligence methods to forecast the concentration level of particle matter. First, principal component analysis was applied to compare patterns of air pollution; second, a multi-layer perceptron (MLP) model was used to forecast the daily concentration of PM_{10} and $PM_{2.5}$. The results prove that the performance of forecasting models did not differ significantly.

However, because the original series of $PM_{2.5}$ concentration are characterized by high volatility and irregularity, the aforementioned models cannot effectively improve precision of prediction model. Fortunately, $PM_{2.5}$ forecasting can be efficiently solved by a promising principle of "decomposition and ensemble" proposed by Tang et al. (2012) and Yu et al. (2015). The main aim of the promising principle can be expressed as follows: first, the decomposition method is used to decompose the original time series into some independent components, such as singular spectrum analysis (SSA) (Wang et al., 2014; Niu et al., 2016) and wavelet transform (WT) (Liu et al., 2014; Araghi et al., 2015); second, the forecasting method is used to forecast each decomposed component independently; third, all predicted components are aggregated as the final prediction. Some studies have confirmed that empirical mode decomposition (EMD) is an efficient decomposition method when compared with other decomposition methods (e.g., WT and SSA) (Yu et al., 2008; Zhang et al., 2009; An et al., 2012; Lin et al., 2012; Karthikeyan and Kumar, 2013). An obvious shortcoming is the frequent appearance of model mixing in the EMD model. Hence, the ensemble empirical model decomposition (EEMD) was proposed to overcome the drawback of the EMD model by adding white noise to the original time series. The EEMD technique effectively resolves the mode mixing problem and greatly improves forecasting accuracy (Jiang et al., 2014; Wang et al., 2015b).

The main contribution of the this study is proposed a decomposition-ensemble learning paradigm integrating EEMD and least square support vector machine (LSSVM) model based on phase space reconstruction (PSR) to forecast $PM_{2.5}$ concentration and compare its prediction performance with other forecasting techniques. The remaining sections of this study are organized as follows. The methodologies are introduced in Section 2. Section 3 presents the experimental analysis. Finally, the conclusions are demonstrated in Section 4.

2. Methodologies

2.1. Ensemble empirical mode decomposition (EEMD)

2.1.1. The basic theories of EMD

Empirical mode decomposition (EMD) is a novel method to process nonlinear and non-stationary data, which was initially proposed by Huang et al. (1998). The EMD model has commonly been used to decompose chaotic time series into a slice of components with different characteristic scales according to different scales of fluctuations and trends, which are transformed into intrinsic mode functions (IMFs). The IMFs components should satisfy the following two prerequisites: (1) in the whole function, the number of extrema and the zero-crossings must be equal or different at most by one; (2) the average value of the local upper and lower envelopes is zero at any point.

The original time series $y(s)$ can be decomposed through the following procedures:

(1) Identify all of the maxima and minima;
(2) Obtain the upper and lower envelopes using cubic spline interpolation;
(3) Calculate the mean envelope of the upper and lower envelopes $m(s)$ and define the differences between $y(s)$ and $m(s)$, the function is as follows:

$$d_1(s) = y(s) - m(s) \tag{1}$$

(4) Check the property of $d_1(s)$, if $d_1(s)$ satisfies the two requirements, and designate it as $IMF_1 = d_1(s)$; otherwise, set $y(s) = d_1(t)$, repeat steps 1–4 t times until the two prerequisites are satisfied and obtain $d_{1t}(s)$, set $f_1(s) = d_{1t}(s)$.
(5) Extract the $c_1(s)$ from the original time series and obtain the remaining component $c_1(s)$:

$$c_1(s) = y(s) - f_1(s) \tag{2}$$

(6) Use the component $c_1(s)$ as a new original sequence, and repeat the above steps 1–3 to obtain other $n-1$ IMFs and final residual until the stop criteria is stratified. The stop criteria can be defined by the following: the residual becomes a monotonic function or the residual can be controlled to be smaller than the predetermined value of a substantial consequence (Huang et al., 2003). The original time series $y(s)$ can be expressed as:

$$y(s) = \sum_{i=1}^{n} f_i(s) + c_n(s) \tag{3}$$

where $f_i(s)$ is the i th IMF component; $c_n(s)$ is the final residual. Moreover, the number of IMF components is limited to be less than $\log_2 n$, where n is the length of original time series.