Engineering Applications of Artificial Intelligence 32 (2014) 185–191

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Hybrid intelligent system for air quality forecasting using phase adjustment

Paulo S.G. de Mattos Neto [a,*], Francisco Madeiro [b], Tiago A.E. Ferreira [c], George D. C. Cavalcanti [d]

[a] Department of Computing, University of Pernambuco, Brazil
[b] Departament of Informatics, Catholic University of Pernambuco, Brazil
[c] Department of Statistics and Informatics, Federal Rural University of Pernambuco, Brazil
[d] Centro de Informática, Universidade Federal de Pernambuco, Brazil

ARTICLE INFO

Article history:
Received 23 October 2013
Received in revised form
27 February 2014
Accepted 19 March 2014
Available online 12 April 2014

Keywords:
Air quality forecasting
Artificial neural network
Genetic algorithm
Time series
$PM_{2.5}$
PM_{10}

ABSTRACT

The pollution caused by particulate matter (PM) concentration has a negative impact on population health, due to its relationship with several diseases. In this sense, several intelligent systems have been proposed for forecasting the PM concentration. Although it is known in the literature that PM concentration time series behave like random walk, to the authors' knowledge there is no intelligent systems developed to forecast the PM concentration that consider this characteristic. In this paper, we present an architecture developed to forecast time series guided by random walk process. The architecture, called Time-delay Added Evolutionary Forecasting (TAEF), consists of two steps: parameters optimization and phase adjustment. In the first step, a genetic optimization procedure is employed to adjust the parameters of a Multilayer Perceptron neural network that is used as the prediction model. The genetic algorithm adjusts the following parameters of the prediction model: the number of input nodes (time lags), the number of neurons in the hidden layer and the training algorithm. The second step is performed aiming to reduce the difference between the forecasting and the actual concentration value of the time series, that occur in the forecasting of the time series with random walk behavior. The approach is data-driven and only uses the past values of the pollutant concentrations to predict the next day concentration; in other words, it does not require any exogenous information. The experimental study is performed using time series of concentration levels of particulate matter ($PM_{2.5}$ and PM_{10}) from Helsinki and shows that the approach overcomes previous state-of-the-art methods by a large margin.

1. Introduction

The study of the air conditions, as well as its prediction, is an important topic due to the relationship between high concentrations of pollutants, such as particulate matter (PM), and adverse effects on human health, that is an issue of increasing public concern. Pollutants concentration forecasting (Niska et al., 2004; Pisoni et al., 2009; Siwek et al., 2011) is a relevant task, since it enables governments to warn the population regarding high levels of pollution. Several epidemiological studies have associated the concentration of those pollutants with cardiovascular and respiratory diseases (Ebelt et al., 2005; Nel, 2005; Peng et al., 2008). The Global Monitoring Report (Bank, 2008) affirms that the major urban air pollutant that affects the human health is PM.

Several types of Artificial Neural Networks (ANN) have been applied for forecasting of the PM concentration. Some works have compared the performance of different ANNs in the prediction of one step ahead. Sharma et al. (2003) used the concentration data of seven pollutants, among them $PM_{2.5}$ and PM_{10}, of the California area to evaluate the performance of four ANN models: Recurrent Network Model (RNM), Change Point Detection Model with RNM, Sequential Network Construction Model and Self Organizing Feature Model. Ordieres et al. (2005) addressed the $PM_{2.5}$ concentration in the cities of El Paso (Texas) and Ciudad Juárez (Chihuahua) to compare the performance of the Multilayer Perceptron (MLP) model, Radial Basis Function (RBF) and Square Multilayer Perceptron (SMLP). Other works proposed ANN based on MLP model. Perez and Reyes (2006) developed an integrated ANN to forecast the maximum average concentration for PM_{10} per day for city of Santiago, Chile. Kukkonen et al. (2003) combined the MLP model with homoscedastic and heteroscedastic Gaussian

* Corresponding author.
E-mail addresses: paulo.salgado@upe.br (P.S.G. de Mattos Neto), madeiro@dei.unicap.br (F. Madeiro), tiago@deinfo.ufrpe.br (T.A.E. Ferreira), gdcc@cin.ufpe.br (G.D.C. Cavalcanti).
URLS: http://www.ppgia.ufrpe.br/tiago (T.A.E. Ferreira), http://cin.ufpe.br/~viisar (G.D.C. Cavalcanti).

http://dx.doi.org/10.1016/j.engappai.2014.03.010
0952-1976/

noise (ANN-HeG) to forecast the PM_{10} concentration in Helsinki. Although most of the works have evaluated the prediction of one step ahead, Kurt and Oktay (2010) used a MLP model of three layers to forecast sulfur dioxide (SO_2), carbon monoxide (CO) and PM_{10} concentration levels for 3 days ahead for a Besiktas district.

Intelligent models based on artificial neural network (ANN) also have been widely used to predict the PM concentration aiming best results. Some works combined Principal Component Analysis (PCA) with ANN models, as Slini et al. (2006), that proposed a hybrid system (PCA-MLP) to forecast PM_{10} concentration in Thessaloniki. Voukantsis et al. (2011) proposed a methodology that selects the input variables using PCA and the prediction is performed by combining the outputs of two methods: MLP and linear regression (LR). In that work, the time series concentration of $PM_{2.5}$ and PM_{10} in Thessaloniki and Helsinki were used. Other hybrid systems were also proposed. Niska et al. (2005) integrated a MLP model with a numerical weather prediction model HIRLAM (High Resolution Limited Area Model) to predict sequential hourly time series of concentrations of $PM_{2.5}$ in Helsinki. In that work, an input selection method based on the use of a multi-objective genetic algorithm (MOGA) was applied to reduce the number of potential meteorological input variables. For PM_{10} concentration management in Santiago, Chile, Perez (2012) combined a MLP model with a nearest neighbor method. Siwek and Osowski (2012) combined ANNs to forecast daily average concentration of PM_{10} in Warsaw, Poland.

Despite many intelligent models have been applied for forecasting of the PM concentration, to the authors' knowledge, no one considered the fact that these time series consist of a random walk process (Sitte and Sitte, 2002), typical of a Brownian motion process (Grau-Bové and Strlic, 2013; Sahu and Nicolis, 2008; Cheng et al., in press). A random walk (Sitte and Sitte, 2002) is a stochastic process that consists of successive and connected steps in which each step is chosen by a random mechanism uninfluenced by any previous step. This phenomenon also is widely seen in the financial time series representation.

In this paper, we present an evolutionary hybrid system, called TAEF method (Ferreira et al., 2008), for time series prediction of particulate matter concentrations. The system was particularly developed to forecast time series from of the phenomenon guided by random walk (Sitte and Sitte, 2002). The approach is composed of two main parts: parameters optimization and phase adjustment. The parameters optimization uses a genetic algorithm to search for the best parameters to train the predictor. A Multilayer Perceptron neural networks is used as predictor and the following parameters are adjusted: number of time lags to represent the series, number of hidden units and the algorithm to perform the training of the predictor. After, a phase adjustment procedure is performed to improve the accuracy of the predictor by automatically correcting time phase distortions. These distortions in the forecasting are common when the phenomena guided by a random walk process (Ferreira et al., 2008) are addressed, such as the PM concentration (Grau-Bové and Strlic, 2013; Sahu and Nicolis, 2008; Cheng et al., in press). Thus, in the case of PM concentration levels, the phase adjustment procedure could significantly improve the accuracy of the intelligent methods? This is a relevant aspect addressed in this paper.

This paper is organized as follows. Section 2 describes the method and the phase adjustment procedure. Section 3 presents a set of six evaluation measures used to analyze the prediction results of the architecture. Simulation results and concluding remarks are presented in Sections 4 and 5, respectively.

2. The architecture for PM concentration forecasting

Given an univariate time series database (Γ), the output of the architecture is a trained Artificial Neural Network (ANN) that is ready to predict the next day value of the time series. The first step of the architecture is the Normalization, where each time series is normalized to lie within the interval [0,1]. After normalization, the database (Γ) is divided into three disjoint parts: training (Ω), validation (Υ), and test (Δ).

Fig. 1 shows that the architecture is composed of two main modules: (i) Parameters Optimization (Section 2.1) and (ii) Phase Adjustment (Section 2.2).

2.1. Parameters optimization

The parameters optimization module is based on a modified Genetic Algorithm (GA) proposed by Leung et al. (2003), used also in other works (Moris et al., 2003; Xu et al., 2007), and is the basis for the GA used in the TAEF method. The GA searches for the best parameters of an ANN in order to improve its performance. In the search process, the GA combines exploration (global search) and exploitation (local search) strategies (Yannibelli and Amandi,

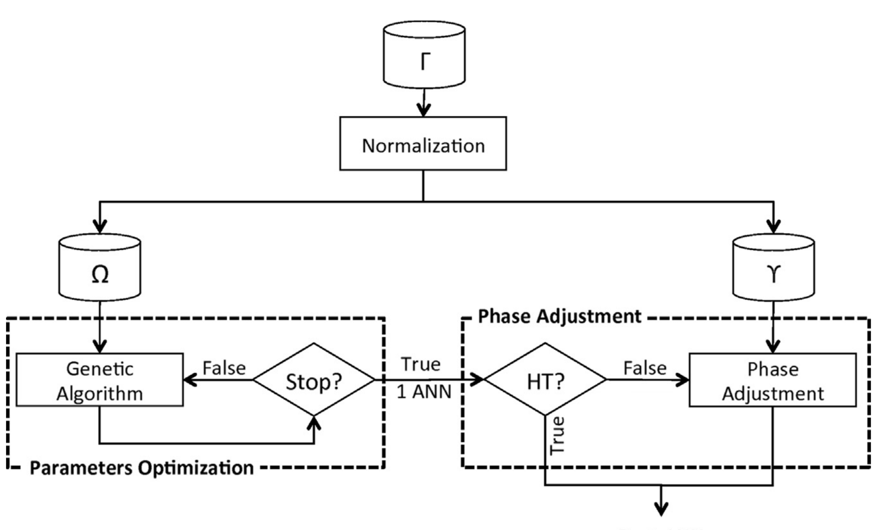

Fig. 1. Architecture of the system. "*Stop*?" corresponds to the stopping criteria of the Genetic Algorithm and "*HT*?" represents the output of a hypothesis test used to evaluate the forecasting of the Artificial Neural Network (ANN).