Atmospheric Environment 55 (2012) 357–365

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model

B. Yeganeh [a,*], M. Shafie Pour Motlagh [b], Y. Rashidi [b], H. Kamalan [c]

[a] Faculty of Civil Engineering, K. N. Toosi University of Technology, No. 1346, Vali Asr Street, Mirdamad Intersection, Tehran 1996715433, Iran
[b] Faculty of Environment, University of Tehran, Tehran, Iran
[c] Faculty of Civil Engineering, Islamic Azad University, Pardis Branch, Tehran, Iran

ARTICLE INFO

Article history:
Received 30 November 2011
Received in revised form
24 February 2012
Accepted 27 February 2012

Keywords:
CO concentration
Machine learning
Support Vector Machine
Partial Least Square
Hybrid models

ABSTRACT

Due to the health impacts caused by exposures to air pollutants in urban areas, monitoring and forecasting of air quality parameters have become popular as an important topic in atmospheric and environmental research today. The knowledge on the dynamics and complexity of air pollutants behavior has made artificial intelligence models as a useful tool for a more accurate pollutant concentration prediction. This paper focuses on an innovative method of daily air pollution prediction using combination of Support Vector Machine (SVM) as predictor and Partial Least Square (PLS) as a data selection tool based on the measured values of CO concentrations.

The CO concentrations of Rey monitoring station in the south of Tehran, from Jan. 2007 to Feb. 2011, have been used to test the effectiveness of this method. The hourly CO concentrations have been predicted using the SVM and the hybrid PLS–SVM models. Similarly, daily CO concentrations have been predicted based on the aforementioned four years measured data. Results demonstrated that both models have good prediction ability; however the hybrid PLS–SVM has better accuracy. In the analysis presented in this paper, statistic estimators including relative mean errors, root mean squared errors and the mean absolute relative error have been employed to compare performances of the models. It has been concluded that the errors decrease after size reduction and coefficients of determination increase from 56 to 81% for SVM model to 65–85% for hybrid PLS–SVM model respectively. Also it was found that the hybrid PLS–SVM model required lower computational time than SVM model as expected, hence supporting the more accurate and faster prediction ability of hybrid PLS–SVM model.

1. Introduction

In establishing ambient air quality standards, regulations have been introduced in order to set limits on the emissions of pollutants. To achieve these limits, considerations have been given to mathematical and computer modeling of air pollution. Therefore, accurate models for air pollutant prediction are needed for such models to allow forecasting and diagnosing potential compliance or non-compliance in both short- and long-term applications. Hence, it is universally agreed that air quality models are indispensable tools for assessing the impact of air pollutants on human health and the urban environment (Chan and Chan, 2000; Gokhale and Khare, 2004; Sánchez et al., 2011).

A large number of atmospheric dispersion models that aim to simulate the physical and chemical processes in the atmosphere have been used for air pollutant forecasting (Moussiopoulos et al., 1995; Yi and Prybutok, 1996). However, such models are unsuitable in many operational settings because they require significant computational efforts and a large volume of different input data. Owing to these inherent difficulties, stochastic models have been widely employed as an alternative to deterministic models to forecast air pollutant concentrations. Many linear (Robeson and Steyn, 1990) and non-linear regression models for concentration forecasting have also been reported (Chaloulakou et al., 2003; He et al., 2009).

In recent years, based on the emission and meteorological data collected from air-monitoring networks round the world, the derivation of Soft Computing Models (SCMs) by using techniques, such as Artificial Neural Network (ANN), Mixture Model and Support Vector Machine (SVM) became popular for air quality prediction (Heo and Kim, 2004; Lu and Wang, 2008).

Artificial neural network methods are regarded as cost-effective methods to achieve the prediction of air pollutants in time series that have become very popular in recent years (Gómez-Sanchis

* Corresponding author. Tel.: +98 9122583286; fax: +98 21 88261079.
E-mail address: bijan.yeganeh@yahoo.com (B. Yeganeh).

1352-2310/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2012.02.092

et al., 2006). The ANN models, in particular, the multi-layer feed-forward neural network perceptron, can be trained to approximate virtually any smooth, measurable function to produce certain promising results to date (Corcoran et al., 2003; Lu et al., 2003).

Recently, SVM paradigm (Smola and Schölkopf, 1998), has gained importance in forecasting problems related to environment (Lu and Wang, 2005; Luan et al., 2005; Osowski and Garanty, 2007; Wang et al., 2008). The SVM method, developed by Vapnik (1995), can provide an effective novel approach to improve generalization performance of neural networks and overcome the inherent drawbacks such as over-fitting training, local minima and poor generalization performance in cases with large initial data that impede the extensive application of the ANN into practice in order to achieve global solutions simultaneously (Sousa et al., 2007). Originally, the SVM model was developed for pattern recognition problems. Recently, with the introduction of ε-insensitive loss function, SVM has been extended to solve non-linear regression estimation as well as time series prediction (Broomhead and Lowe, 1998). Unlike traditional learning machines, which normally adopt the Empirical Risk Minimization Principle (ERMP) like feed-forward neural networks, SVM implements Structural Risk Minimization Principle (SRMP), which seeks to minimize an upper bound of generalization error rather than minimize training error. This process leads to better generalization than conventional methods (Lin et al., 2002).

However the multicollinearity, or high correlation between the independent variables in a regression equation, can make it difficult to correctly identify the most important contributors to a physical process. One method for removing such multicollinearity and redundant independent variables is to use multivariate data analysis (MDA) techniques. MDA techniques have been used for analyzing voluminous environmental data (Buhr et al., 1992). There are many data reduction techniques. Among reduction methods, PLS is a supervised feature extraction method, because it is an unsupervised dimension reduction technique when our key area of application is multivariate regression (Mitra and Yn, 2000). By principal component analysis and the synthesis of variable extraction, the most comprehensive explanatory variables that predicted the variable Y (response) were extracted. PLS can separate the information and noise of the examined system so that the appropriate models can be established to allow us to achieve more balance and provide an alternate approach to principal component analysis (PCA) technique (Mitra and Yn, 2000).

The work presented in this paper aims to examine the feasibility of applying SVM and hybrid PLS–SVM models to predict air pollutant levels in short- and long-term periods based on the measured air pollutant database in Tehran. This hybrid model provides a novel alternative for air pollutants concentration forecasting.

2. Material and methods

The hybrid PLS–SVM as a novel approach for air pollution prediction has been applied to data from an air quality monitoring station in the south of Tehran-Iran. Tehran, the capital of Islamic Republic of Iran with population of approximately 8.5 million is the most important metropolis and the largest commercial and political center of the country. Tehran is amongst a few capitals of the world, which is not located alongside a river or even close to sea. High altitude mountains surround the City from the North and East. The total area of the City is 780 km^2. There are four distinct seasons, with the annual mean rainfall at about 230 mm, and mean temperature of 17 °C. The highest temperature is 39 °C in summer and −6 °C in winter. The annual mean humidity is 40% and the highest monthly mean humidity is 65% in January and lowest being 24% in July and August. Tehran suffers from severe air pollution that sometimes makes breathing difficult. Well over 80% of the city's air pollution is due to cars.

Besides some 22 monitoring stations run by the Department of Environment, Air Quality Control Company (AQCC) a subsidiary of Tehran Municipality, also has eighteen monitoring stations in Tehran. Rey monitoring station is one of the most reliable satiation situated in the south of Tehran. To predict CO concentration in the future, the six criteria air pollutants, i.e. particulate matters (PM_{10}), total hydrocarbons (THC), nitrogen oxides (NOx), methane (CH_4), sulfur dioxide (SO_2) and ozone (O_3) in addition to six meteorological parameters, i.e. pressure (Press.), temperature (Temp.), wind direction (WD), wind speed (WS) and relative humidity (Hum.) have been used. The extended data for measurements of CO were used in this research from Jan. 2007 to Jan. 2011. The performance of hybrid PLS–SVM model has been evaluated by comparing the results with the Rey station real measured data.

2.1. Theory of Partial Least Square (PLS)

Partial Least Squares being one of the features of extraction method that was developed by Wold (Wold, 1966). PLS constructs a linear model that describes the relationship between dependent (response) variables Y and independent (predictor) variables X. This linear model tries to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. The technique of PLS is similar to the principal component analysis. It also produces linear combinations of the original surface parameters. However, PLS and PCA differ in the way they extract the principal directions. PCA ignores the information in Y when building the principal components and PLS produces the directions reflecting the relationship between Y and X. Hence, PLS results are expected to have more practical meanings.

Assume X is an $n \times p$ matrix and Y is an $n \times q$ matrix. The PLS technique works by successively extracting factors from both X and Y such that covariance between the extracted factors is maximized. PLS method can work with multivariate response variables (i.e. when Y is an $n \times q$ vector with $q > 1$). However, for this research purpose it will be assumed that a single response (target) variable i.e., Y is $n \times 1$ and X is $n \times p$, is made available as before. PLS technique attempts to find a linear decomposition of X and Y such that $X = TP^T + E$ and $Y = UQ^T + F$, where T $n \times r = X$-scores, U $n \times r = Y$-scores, P $p \times r = X$-loadings, Q $1 \times r = Y$-loadings, E $n \times p = X$-residuals, F $n \times 1 = Y$-residuals. Decomposition is finalized so as to maximize covariance between T and U. There are multiple algorithms available to solve a PLS problem. However, all algorithms follow an iterative process to extract the X-scores and Y-scores. The factors or scores for X and Y are extracted successively and the number of factors extracted (r) depends on the rank of X and Y. In this case, Y is a vector and all possible X factors will be extracted.

Extracted x-scores are linear combinations of X. For example, the first extracted x-score t of X is of the form $t = Xw$, where w is the eigenvector corresponding to the first eigenvalue of X^TYY^TX. Similarly the first y-score is $u = Yc$, where c is the eigenvector corresponding to the first eigenvalue of Y^TXX^TY. Note that X^TY denotes the covariance of X and Y. Once the first factors have been extracted, the original values of X and Y as, $X_1 = X - tt^TX$ and $Y_1 = Y - tt^TY$ would be deflated. The above process is now repeated to extract the second PLS factors (Mitra and Yn, 2000). The process continues until all possible latent factors t and u have been extracted, i.e., when X is reduced to a null matrix. The number of latent factors extracted depends on the rank of X. In PLS method, number of components (NCs) should be determined properly. Any method used for determining NCs should take into account not only the goodness of fit, but also the complexity taken to achieve that fit. In