Atmospheric Environment 80 (2013) 426–437

Contents lists available at ScienceDirect

Atmospheric Environment

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

Identifying pollution sources and predicting urban air quality using ensemble learning methods

Kunwar P. Singh [a,b,*], Shikha Gupta [a,b], Premanjali Rai [a,b]

[a] Academy of Scientific and Innovative Research, Council of Scientific & Industrial Research, New Delhi, India
[b] Environmental Chemistry Division, CSIR – Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India

HIGHLIGHTS

- Developed tree ensemble models for seasonal discrimination and air quality prediction.
- PCA used to identify air pollution sources; air quality indices used for health risk.
- Bagging and boosting algorithms enhanced predictive ability of ensemble models.
- Ensemble classification and regression models performed better than SVMs.
- Proposed models can be used as tools for air quality prediction and management.

GRAPHICAL ABSTRACT

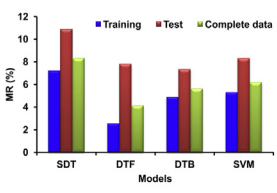
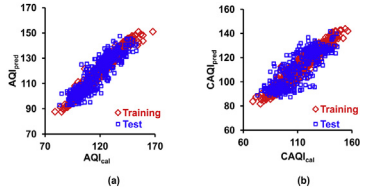


Figure shows misclassification rate in seasonal discrimination of air quality of Lucknow yielded by different models and suggest that the ensemble learning classification models (DTF and DTB) performed relatively better than SDT and SVM.



Figures show correlative distribution of calculated and model predicted values of (a) AQI, and (b) CAQI for Lucknow ambient air using DTB model.

ARTICLE INFO

Article history:
Received 23 June 2013
Received in revised form
5 August 2013
Accepted 12 August 2013

Keywords:
Ensemble learning
Decision tree forest
Decision treeboost
Air quality
Indices
Air pollution
Seasonal discrimination

ABSTRACT

In this study, principal components analysis (PCA) was performed to identify air pollution sources and tree based ensemble learning models were constructed to predict the urban air quality of Lucknow (India) using the air quality and meteorological databases pertaining to a period of five years. PCA identified vehicular emissions and fuel combustion as major air pollution sources. The air quality indices revealed the air quality unhealthy during the summer and winter. Ensemble models were constructed to discriminate between the seasonal air qualities, factors responsible for discrimination, and to predict the air quality indices. Accordingly, single decision tree (SDT), decision tree forest (DTF), and decision treeboost (DTB) were constructed and their generalization and predictive performance was evaluated in terms of several statistical parameters and compared with conventional machine learning benchmark, support vector machines (SVM). The DT and SVM models discriminated the seasonal air quality rendering misclassification rate (MR) of 8.32% (SDT); 4.12% (DTF); 5.62% (DTB), and 6.18% (SVM), respectively in complete data. The AQI and CAQI regression models yielded a correlation between measured and predicted values and root mean squared error of 0.901, 6.67 and 0.825, 9.45 (SDT); 0.951, 4.85 and 0.922, 6.56 (DTF); 0.959, 4.38 and 0.929, 6.30 (DTB); 0.890, 7.00 and 0.836, 9.16 (SVR) in complete data. The DTF and DTB models outperformed the SVM both in classification and regression which could be attributed to the incorporation of the bagging and boosting algorithms in these models. The proposed ensemble models successfully predicted the urban ambient air quality and can be used as effective tools for its management.

* Corresponding author. Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India. Tel.: +91 522 2476091; fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.08.023

1. Introduction

Air quality and its temporal and spatial variations in a region are largely determined by the nature of anthropogenic activities associated with various gaseous and particulate emissions and set of prevailing meteorological conditions there. Epidemiological studies have established the associations between the air pollutants and daily excess in mortality (Dockery and Pope, 1994) and morbidity (Kassomenos et al., 2008). Poor air quality has both acute as well as chronic health impacts (Nastos et al., 2010) and the severity of the impacts mostly depends upon two factors viz. ambient concentration of the air pollutants and its exposure time. Further, concentrations of air pollutants are subject to alter depending on the local topography, source emission, and surrounding meteorological conditions. However, among these variables, meteorological parameters are mostly responsible for causing variations in the ambient concentrations of air pollutants (Banerjee et al., 2011). Suspended particulate matter (SPM), respirable suspended particulate matter (RSPM), oxides of nitrogen (NO_x) and sulfur (SO_x), and ozone are among the major common air pollutants. RSPM refers to those SPM with nominal aerodynamic diameter of 10 μm or less. The environmental regulatory authorities (CPCB, 2009) have prescribed the guidelines for the maximum permissible levels of various air pollutants, such as SO_2 (80 μg m^{-3}), NO_2 (80 μg m^{-3}), and RSPM (100 μg m^{-3}), respectively.

Concern about air pollution in urban regions is receiving increasing importance worldwide (Chattopadhyay et al., 2010). The urban areas might be viewed as dense source of enormous anthropogenic emission of pollutants, which could alter the atmospheric composition, chemistry, and life cycles in its downwind regimes, extending over several hundred kilometers (Gupta et al., 2008). Petrol and diesel engines of motor vehicles were found to emit a wide variety of pollutants, principally oxides of nitrogen (NO_x) which had an increasing impact on urban air quality (Mage et al., 1996). Urban air pollution in India had increased rapidly with the population growth, number of motor vehicles, use of fuels with poor environmental performance, badly mentioned transportation system, poor land use pattern, and above all ineffective environmental regulations (Gupta et al., 2008). Developing appropriate strategies for air pollution prevention, understanding of the nature of sources and influences of meteorological conditions on their profiles are essentially required.

As for the health impact of air pollutants, air quality index (AQI) is an important indicator for general public to understand easily how bad or good the air quality is for their health and to assist in data interpretation for decision making processes related to pollution mitigation measures and environmental management. Basically, the AQI is defined as an index or rating scale for reporting the daily combined effect of ambient air pollutants recorded at the monitoring site (Kumar and Goyal, 2011). To ensure the safety of the humans and the archeological monuments, particularly in the urban areas, it is very much desired to monitor the ambient air quality on a regular basis and develop appropriate strategies for the control of the emission sources.

Accordingly, several urban air quality monitoring programs covering various major Indian cities have been initiated generating huge databases (Chattopadhyay et al., 2010) in recent past years. However, requirements of huge funds, dedicated manpower and instrumentation for such programs limit their viability. Therefore, to develop effective strategies for urban air quality management, it is essentially needed to develop appropriate methods which could be capable of predicting the air quality and enumerating the seasonal influences on ambient air in a region.

On the other hand, predictive modeling offers tools for forecasting the air quality based on past measurements. In recent years, several research efforts have been made in this direction. Atmospheric dispersion models used to predict the ground level concentration of the air pollutants around the sources (Cimorelli et al., 2005; EPA, 2005; Kesarkar et al., 2007; Bhaskar et al., 2008) require précised knowledge of several source parameters and the meteorological conditions (Collett and Oduyemi, 1997). The statistical models attempt to determine the underlying relationship between a set of input data and targets. Since, air quality data are generally very complex and exhibit nonlinear dependence, linear modeling approaches may not be suitable to model such data (Singh et al., 2012). The artificial neural networks (ANNs) are considered as standard nonlinear estimators and their predictive and generalization abilities have been well established through their successful applications in a variety of fields (Singh et al., 2012, 2013), but ANNs suffer with the problem of over-fitting in learning process.

Support vector machines (SVMs), exhibit excellent generalization abilities with non-linear systems (Singh et al., 2013), and also make use of limited data points in model building. In recent years ensemble learning methods (Snelder et al., 2009) have emerged as unbiased tools for modeling the complex relationships between set of independent and dependent variables and have been applied successfully in various research areas (Yang et al., 2010). In general, these methods are designed to overcome problems with weak predictors (Hancock et al., 2005) and have the advantage to alleviate the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for overfitting the training data (Dietterich, 2000). Decision trees (DTs) are commonly used as base predictors in building ensemble learning models (Zhang et al., 2008) and supplemented with bagging and stochastic gradient boosting techniques (Breiman, 1996; Friedman, 2002). The bagging aims minimizing of prediction variance by generating bootstrapped replica data sets, whereas, boosting creates a linear combination out of many models, where each new model is dependent on the preceding model (Friedman, 2002). Decision tree forest (DTF) and decision treeboost (DTB) implementing bagging and boosting techniques, respectively are relatively new methods for improving the accuracy of a predictive function (Yang et al., 2010). These techniques are inherently nonparametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables and can handle numerical data that are highly skewed or multi-model in nature (Mahjoobi and Etemad-Shahidi, 2008). To our knowledge, ensemble learning methods have not yet been applied to the air quality prediction.

The main objectives of this study were (i) to construct ensemble learning (EL) based classification and regression functions to enumerate the influence of seasons on the air quality; and to predict the AQI in the study region using selected air quality (RSPM, NO_2, SO_2) and meteorological parameters (air temperature, relative humidity, wind speed, sunshine hours, evaporation rate) as the estimators, (ii) to compare the predictive and generalization abilities of these modeling approaches. Accordingly, EL models were developed and the performances of these models were evaluated in terms of several statistical criteria parameters and compared with the SVM approach as benchmark. This study has shown that the application of EL methods can be useful in predicting the air quality and enumerating the seasonal influences successfully for its effective management.

2. Methods

The basic aim of this study is (a) to identify the air pollution sources, (b) to find an accurate possible classification function $\bar{f}_c$ capable of discriminating the seasonal (summer, monsoon, winter) air quality to enumerate the responsible factors, influences of