EasyChair Preprint

№ 5632

Prediction of Air Pollution Using Random Forest

Sahil Singh, Ayush Yadav and Akhilesh Kumar

May 27, 2021

Prediction of Air Pollution Using Random Forest

Sahi Singh
computer science and engineering
Galgotias University

Established under galgotias university act 14 of 2011
GreatorNoida,INDIA
sahil_singh.scsebtech@galgotias

Ayush Yadav
computer science and engineering
Galgotias University

Established under galgotias university act 14 of 2011

Greator Noida,INDIA
Aayushyadav15032001@gmail.com

AKHILSEH KUMAR

computer science and engineering

Galgotias University

Established under galgotias university act 14 of 2011

GreatorNoida,INDIA.

Abstract — The aim of this project is to use a heterogeneous ensemble of differential evolution with random forest method for air pollution prediction in New Delhi. This is different from existing work (independent classifier of Bayesian network and multi-label classifier used for the estimation of air pollutants) as a method is proposing to combine state-of-the-art differential evolution strategies with random forest method instead of focusing on existing single technique. When the existing approach i.e. independent and multi-label classifiers are compared with proposed approach, it shows proposed approach leads to the performance gains. Continuous ambient air quality data of two cities Delhi and Patna from Central Pollution Control Board were publically made available, from where seven pollutants (C6H6, NO2, O3, SO2, CO, PM2.5 and PM10) dataset are collected with daily average concentration. . Air pollution monitoring, is thus, becoming more and more significant. Real-time air quality information, such as concentration of PM2.5, PM10, and, NO2, is important aspect for pollution management and protecting human beings from the damages caused by air pollutants.

If we talk about air quality the PM 2.5 is Main cause for human health and city management . It also affect government policies. However, in big cites there are very few monitoring stations of air quality.So, in this Research paper we will talk about Random forest Technique to predict and Measure Pollution in big cities, the data Generated by this technique include ,real-time Traffic status, meteorology data and road information. This algorithm is also used for data training and prediction.

Keywords—air quality prediction; random forest; point of interest; traffic

I. INTRODUCTION

As we know that urbanization tends urban growth, the infrastructure of transportation depends on fossil fuels depends constantly, that's why a very huge amount of vehicle use increases traffic related pollutant emissions . Now a days Urban air pollution is a very big problem in developed and developing countries because the atmospheric pollutants had a very bad affect on human health and air pollution can also be resulted in acid rain and greenhouse effect. Disease like lung cancer are caused by these harmful pollutants. For ex, So2 and No2 are the main causes for acid rain as well as greenhouse effect. Especially in India ,Now days pollution are the very big problem in big cities like Delhi and Mumbai,where has pollutant include exhaust emissions. In Delhi there are about five million vehicles, coal burning in neighbour states. In 2019 November Delhi- NCR is like a gas chamber because of pollution Of moter vehicles and bursting of crackers in Diwali . In winters This pollution problem is at its peak in Delhi - NCR because While in summers this problem is at moderate mode. Government has also take steps measures to control this Problems by establishing air quality monitoring stations .

.

II. EASE OF USE

RELATED WORK:

Previously, there are very much studies on air quality use approach such as satellite remote sensing, wireless sensor network and dispersion model. The mathematical model such as Box model, Gaussian model are the air . How air pollution disperse in atmosphere? The main functions of meteorology , traffic volumes are the classical dispersion models. These models depends only on parameter that to simulate the pollution dispersion but it not consider some situations and conditions that are human mobility and concentrations. In meantime, model of dispersion depends on accurate data ,such as traffic emissions ,wind speed and so on Such as these factors accuracy cannot be guaranteed in certain conditions, For ex, wind speed is different in different regions this is because whether conditions and obstruction of building in determining over the structures. We can only estimate the value by determining fuel consumption and distance travelled. the other way of monitor air quality is satellite remote sensing .

Advantage of satellite data is sensing, predicting and processing air pollutants. However, many air quality managers are not yet taking the full advantage of satellite data for their applications because of challenges associated with accessing, processing, and properly interpreting observational data. That is, the certain degree of technical skill is required on the part of the data end-user, which is often problematic for the organizations with limited resources Sensor networks have also been studied extensively because of their broad applicability and, enormous application potential in areas such environmental monitoring field. A Wireless Sensor Network Air Pollution Monitor System was deployed in New Delhi for monitoring air quality ; distributed infrastructure-based wireless sensor networks and grid computing is also used for monitoring the air quality of the India. It is same reason which limits the number of stations in cities of India.

A. Problem Description and Definition

Definition

A.a.Air quality Index

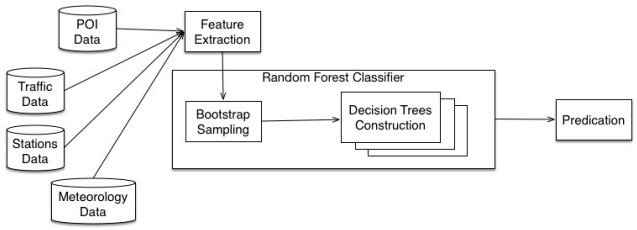
An air quality index (AQI) is a type of number which is used by many government and private agencies to tell the public of their own country that how much their air is polluted currently and in future what is the rate of increase of this AQI level. With the increase in the percentage of AQI level there will be more increase in the pollution and we all know that what happens when there is increase in level of air pollution nearly most of the people will suffer with many health problem like most of the people will have breathing problem and there will be also bad environment around us. In the whole world different countries uses types of air quality indices for the measurement of AQI level in their countries.In this paper we are going to use the standard of India for measuring AQI level. In our country measuring of AQI is mainly based on six atmospheric gases, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), suspended particulates smaller than 10 µm in the aerodynamic diameter (PM10), suspended particulates smaller than 2.5 µm in aerodynamic .The AQI value is the calculated per hour according to a formula published by India's Ministry of Environmental Protection

A. Points of interest

A point of interest, or POI, is the type of specific point location at which someone who finds their benefits our usually interested at that place which is useful for them.. e.g. restaurants and shopping malls, complex, hotels etc. surrounding us are POI.

A. RAQ algorithm

In the RAQ algorithm, all the data are gathered from that urban sensing system including air monitoring station data, meteorology data, traffic data, road information and POI data and necessary features are extracted from the heterogeneous data. These features are the most common data in the city life. Traffic-related sources like emission of air pollutants from vehicles like car, truck motorcycle ,auto etc. and POI like industries are the main sources for the air pollutant because in INDIA most of the industry are situated at Delhi and Noida and they releases their harmful gases directly to the environment without filtering the gases causing air pollution. Meteorology is main approach for dispersion of air pollutants. These data is us The training data set includes all necessary features and is divided into subsets using the bootstrap technology. Figure 5 shows the structure of the data set. A decision tree is constructed on the each subset, and the classification is done by aggregating results generated from all decision trees. Figure shows the procedure of the RAQ algorithm.



B.*Data Collection and Feature Extraction*

Meteorology Data

Meteorology data such as temperature ,humidity and barometric pressure are the most important factors that are badly affect the concentration and spread of air pollutants. Understanding the behaviour of the meteorological parameters in the planetary boundary layer is most important because atmosphere is the only medium in which air pollutants directly gets transported from the source like when there will be change in humidity it will directly get transported to air, which is governed by the meteorological parameters such as the atmospheric wind speed, wind direction, and the temperature. In this paper, we will use different types of meteorological data provided by the meteorological department of India and we use weather monitoring stations as one part of urban sensing system. Considering the accessibility of the data from meteorological Department of India, we use following meteorology data features that is: temperature (Fmt, °C), humidity (Fmh, %), barometric pressure (Fmp, mmHg), wind speed (Fmw, m/s) and visibility (Fmv, m).

B. Traffic and Road Data

As we all know that in New Delhi mainly the air pollution is caused due to emission of fossil fuels from vehicles, and this is only factor which is effecting the quality of air very severely. . In this paper, we mainly focus on two important characteristics of the traffic, which are length of road and congestion of traffic at that particular place. If the road will be very long then traffic congestion will be relatively light, and exhaust gas emissions will be at a high level because of total number of vehicles on this road. Similarly, if the road will be short and traffic congestion will be heavy. However, we do not have the method to find or observe that how much emission is being done or accurate data to quantify these two characteristics directly. Most of map service providers offer online maps and real-time traffic status through which we can get the traffic status of any place that how much traffic is there at some specific place and we will be able to predict the emission of pollutants at that particular place.As they do not publish public application interfaces for third party to get access to these data, but we can still get some useful data for determining the quality of air for example if we came to know that some website is showing the presence of more traffic and pollution at some places then it will be vbery useful for us. Essentially, these data are collected from the GPS equipment installed in cars or speed measurement sensors.

B.POI Data

POI Data The category of POIs and their density in a region indicate the land use and the function of the region as well as the traffic patterns in the region, like if there is a big shopping complex in the city then the places nearby to this complex will be more dense and crowded and this will will led to use of vehicles which will cause air pollution therefore contributing to the air quality inference of the region.

C.Random Forest Classification

The Random Forest are an ensemble learning methods used for classification and regression. It is one of the most used algorithm because of its simplicity and more in diversity we use Random forest algorithm for multi class classification.

C.Tree Growing and Splitting

Information gain is mainly used as the criteria for classifiers According to bagging theory, random forest is strong classifier based on multiple weak classifiers.. Brieman suggests three possible values for m: 1/ 2 under root (m), under root(m), 2underroot(m). entropy is calculated as in Equation (3):

k

4 calculate information gain by
Equation ;Entropy(c)=$-\sum_{i=1} p(ci)\log 2p(ci)$ (3)

D.Prediction
For each tree, p(ci) is the estimated probability of the AQI level i. The final probability of the AQI level i p'(ci) in the random forest is defined in Equation (4), where T is the number of decision trees as mentioned before:

$p'(ci)=1/T\sum p(ci)$ (4)

The final result is determined by Equation :

$C'(i)=\max(p'(ci))$ (5)

The pseudo code of RAQ algorithm is described in Algorithm 1.

Input:
A data set S with different features: Anq, Anh, Anp, Anw, A, Ari, Aqcs, An and labeled Air Quality Index level;
unlabeled data set R; trees quantity Q; features quantity N;
Output: AQI level
1 for Q no of trees
2 we will randomly select n features from S;
3 for n features in each node
4 calculate information gain by: k
Equation ; Entropy(c)=$-\sum_{i=1} p(ci)\log 2p(ci)$
5 we will choose the maximum gain to split the data set in node.
6 we will remove used features from the features candidates;
7Now we will input unlabeled data into the trees;
5we will finally get predicted Air Quality Index level according to Equations (1) and (2);

2. Evaluation

(iii)Evaluation Method

The most useful and accurate criteria for measuring the AQI is the air quality we get from monitoring stations such as data collected from sensors .We will need two parameter one is number of tree and number of features used to build tree for the construction of Random forest. To choose the best parameters among these two we will use out of bag error to compare RAQ accuracy on the basis of different parameters pair i.e. trees and features which means the number trees required to construct a tree and number of features required to construct a random forest.The error in Random forest is calculated internally at the time of construction of trees.

Each and every tree is being constructed using a different bootstrap sample from the specified data or original data provided to us. The one-third sample is used as test cases to be input into the tree and get the classification of each test case. At the end of the run,we will take the class j that got most of the votes every time case n was out of bag . The smaller number of out of bag, the high will be the accuracy of the model. For the number of features, we will increase one by one each time from 2 to 8 in which the total number of features is 8 . For the quantity of trees, we will increase by 100 from 100 to 1000. The tree which will consume more time because more number of trees then we will ignore trees number greater then 1000 and gap will 100 to balance performance and accuracy of the tree performance and accuracy. To compare this algorithm with others, we use cross-validation method to judge the performance.

3. Results:

In this paper, we are going to predict the presence of air pollutants in air with the help of Random forest algorithm and AQI level in environment of NEW DELHI. We will use different types of data for AQI measurement like POP data, traffic data and road data also meteorology data.like humidity, temperature,speed etc.

4. Conclusions

In this paper, our model is going to predict the presence of pollution in air based on AQI level. We will determine the Air quality index at different places in New Delhi. We will use many different types of data for measuring the AQI index like we will use Meteorology Data, traffic and road data provided by some application like google map, we will also use POI data for the measurement of AQI level. We will use the random forest algorithm to predict all the regions mainly rural areas which will be not covered.
Thus;the result is yet to be determined.
.

Conflicts of Interest: We declare no conflicts of interest

Reference:

https://mausam.imd.gov.in/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4732119/

ACKNOWLEDGMENT

I have taken efforts in this project, However, it would not have been possible without the kind support and help of many individuals and organizations, We would like to extend my sincere thanks to all of them.

We would like to express our special thanks of gratitude to our teacher **Ms. Swati Sharma Mam** who gave us the golden opportunity to do this wonderful project on the topic **Prediction of Air Pollution Using Random Forest** under his guidance, which helped us in doing a lot of Research and we came across so many new things, we are really thankful to him.

Mam had guided us each and every time we came across any problem and provided us with the best available options.

She has taken pains to go through our project several times to help us in making correction to our project.

We would also like to thank our institution Galgotias University without whom this project would have been a distant reality.

My thanks and appreciations also go to my classmates in developing the project and to the people who have willingly helped me out with their abilities.

We would like to express my special gratitude and thanks to industry persons for giving us such attention and time.

Sahil Singh
Ayush Yadav
Akhilesh Kumar