Journal of Engineering Sciences
Vol 11, Issue 5,May/2020
ISSN NO:0377-9254

INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING

Sapan Bali[1]
Ms Nidhi Sengar[1] (Assistant Professor, MAIT) , Department of Information Technology, Maharaja Agrasen institute of technology, Sector-22, Delhi, India

Abstract

Air pollution is a concoction of solid particles and gases in the air. To protect the living beings I need to predict the air quality and helps to fight with the detrimental effects on flora and fauna.I aim to provide a system to predict the air quality with maximum accuracy achieved using and comparing machine learning algorithms to land with best results. The interpolation, prediction, and feature analysis of air quality are three important points in the territory of urban air computing. Air quality index of India is a standard measure used to indicate the pollutant (so2, no2, rspm, spm. etc.) levels over a period. I developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. This system can be used by central pollution control board to know air quality provided concentration of contaminated particles. My model will be capable for successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration. In my model by implementing the proposed parameter reducing formulations, I achieved better performance than the standard regression models.

Keywords : Prediction, Interpolation, feature analysis , pollutant, Gradient decent, regression models, formulations

1. INTRODUCTION

The three most important areas for urban air computing are interpolation, feature-analysis and prediction of air quality. Solutions to these topics could help us control air-pollution.

Interpolation solves the problem of less number of functional air-quality monitor stations in the city and their uneven distribution of air-quality monitor station in the city. A good feature selection technique will help us to get accurate predictive model by extracting the best feature or the most relevant feature in order to achieve prediction that is more accurate.

The problem is, that it is difficult to identify the main cause for the occurrence of air pollution hence it is difficult to identify the relevant required for interpolation and prediction of air quality of a particular region. The rest of the paper is organised as follows: Section 2 mainly focuses on analysis of different techniques and related work. Section 3 draws conclusion and then references.

2. RELATED WORK

Air pollution represents the biggest environmental risk to health. Approximately 92% of the world population live in places where air quality level exceeds WHO limits. Air pollution is one of largest causes of the top four non-communicable diseases such as stroke, lung cancer, chronic respiratory disease, and heart disease.
In 2012, one out of 9 deaths was the result of air pollution-related diseases. Over half of deaths among children less than 5 years old from acute lower respiratory infections are due to particulate matter inhaled from indoor air pollution from household solid fuels.

More than 660 million Indians breathe air that fails India's National Air Quality Standards[2]. Research suggests that meeting those standards would increase life expectancy in India by 1 year. Going further and meeting the international benchmarks of the World Health Organization is estimated to add 4.7 years to life expectancy.

According to this global estimation, in Mongolia[3] 1123 people die from air pollution-related diseases each year

Authors Kleine Deters, J., Zalakeviciute, R., Gonzalez, M. and Rybarczyk, Y. in their research paper[4] have proposed the research of outdoor pollution causing millions of premature deaths due to anthropogenic fine particulate matter or PM 2.5 in the capital city of Ecuador. A machine learning approach to based on six years of meteorological and pollution data analysis to predict concentration of PM 2.5.

3. AIR QUALITY EVALUATION

Because different pollutants have different effects, the NAAQS standards are also different. Some pollutants have standards for both long-term and short-term averaging times. The short-term standards are designed to protect against acute or short-term health effects, while the long-term standards were established to protect against chronic health effects. Because different pollutants have different effects, the NAAQS [5] standards are also different and some of them are shown in Table I. Some pollutants have standards for both long-term and short-term averaging times. The short-term standards are designed to protect against acute, or short-term, health effects, while the long-term standards were established to protect against chronic health effects.

Pollutant	Primary/ Secondary	Averaging Time	Level	Form
Carbon Monoxide (CO)	Primary	8 hours	9 ppm	Not to be exceeded more than once per year
		1 hour	35 ppm	
Lead (Pb)	Primary and secondary	Rolling 3 month average	0.15 $\mu g/m^3$	Not to be exceeded
Nitrogen Dioxide (NO2)	Primary	1 hour	100ppb	98th percentile of 1-hour daily maximum concentrations, averaged over 3 years
		1 year	53 ppb	Annual Mean
Ozone (O3)	Primary and secondary	8 hours	0.07 ppm	Annual fourth-highest daily maximum 8-hour concentration, averaged over 3 years

Table I: Naaqs Table Lists All Criteria Pollutants And Standards [5]

According to Central pollution control board the data acquired of pollutants which is for air quality degradation is show in Fig1. Eight parameters are directly contributing in the same.

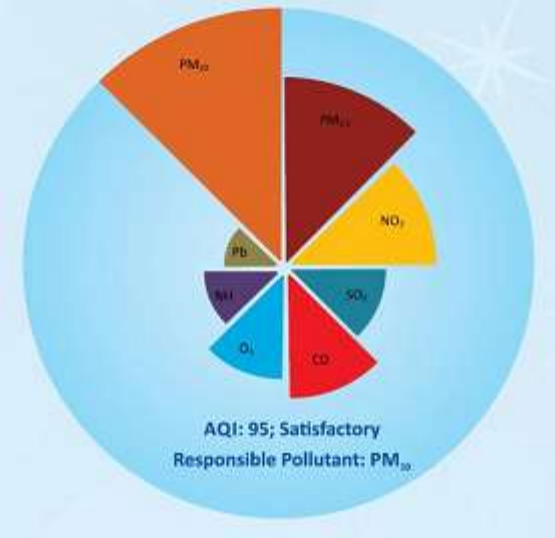

Fig1 (Source : Central pollution control board)

AQI CALCULATION CRITERIA

The air quality index of a particular data point is the aggregate of maximum indexed pollutant on that particular area. That pollutants maxsub index is taken as the air quality index of that particular location

aqi[(aqi > 0) & (aqi <= 50)] = 0
aqi[(aqi > 50) & (aqi <= 100)] = 1
aqi[(aqi > 100) & (aqi <= 200)] = 2
aqi[(aqi > 201) & (aqi <= 300)] = 3
aqi[(aqi > 301) & (aqi <= 400)] = 4
aqi[(aqi > 401)] = 5

4. MODEL FORMULATION

Semi-Supervised Learning Algorithms based on a co-training framework is proposed. The cotraining framework includes two different classifiers, the two classifiers are: Spatial Classifier and Temporal Classifier. Spatial Classifier is based on ANN and deals with spatially related features like density and length of road etc., while Temporal Classifier takes in account of temporary factors that affect the conditions like traffic and meteorology.

4.1 Prediction Using Random Forest:

Random Forest is an adaptable, simple to utilize AI algorithm that produces, even without hyperparameter tuning, an extraordinary outcome more often than not. It is additionally a standout amongst the most utilized algorithms, since its effortlessness and the way that it tends to be utilized for both classification and regression tasks. Another incredible nature of the random forest algorithm is that it is anything but difficult to gauge the general significance of each feature on the prediction. Sklearn gives an extraordinary device to this that estimates a features significance.

How it Works

Random Forest is a supervised learning model. Like you would already be able to see from it's name, it makes a forest and makes it some way or another random. The „forest" it assembles, is an outfit of Decision Trees, more often than not prepared with the "bagging" technique. The general thought of the bagging technique is that a blend of learning models expands the general outcome. Random forest forms various decision trees and consolidates them to get a progressively precise and stable predicted result.

4.2 Data Set

I acquired the dataset with various columns of sensor data from various places in India. I have the average readings of ambient air quality with respect to air quality parameters, like Sulphur dioxide (So2), Nitrogen dioxide (No2), Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM). Data acquired from the smyce has more noisy data since few of the data from the stations have been shifted or closed the period Ire marked as NAN or * or x or #.so I have to pre-process the data in order to remove the outliers. Each individual pollutant indexes, gives the relationship betIen the pollutant concentration and their corresponding individual index

4.3 Algorithm based comparison

The model is trained using random forest on the training and testing dataset in 80:20 ratio. The precision and average table is depicted in fig2. With accuracy of 0.99156 on the provided concentrations of pollutants.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10695
1	1.00	1.00	1.00	4260
2	0.99	1.00	1.00	512
3	0.00	0.00	0.00	2
4	1.00	1.00	1.00	1
5	1.00	0.56	0.71	9
accuracy			1.00	15479
macro avg	0.83	0.76	0.78	15479
weighted avg	1.00	1.00	1.00	15479

Fig 2. Precision table random forest

5. RESULTS AND CONCLUSION

Using the three algorithms as shown in fig 3. SVM, naïve bias and random forest I have predicted the air quality for a particular set of inputs and analysed the prediction accuracy in all the three algorithms out of which random has the most accuracy of near 99% in prediction.

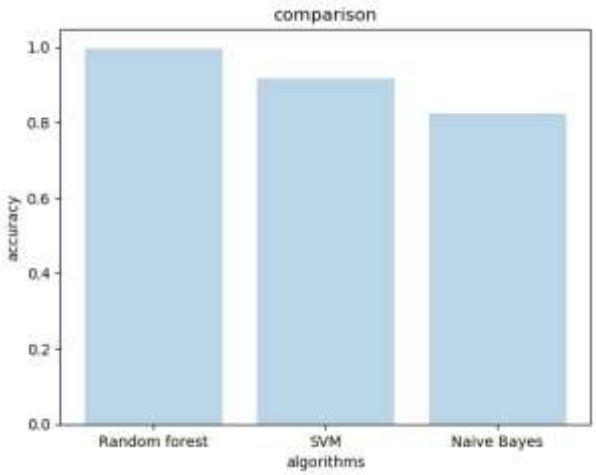

Fig3. Comparison chart

This project studies three important topics in the area of urban air computing: the interpolation, prediction, and feature analysis of fine-gained air quality. I will be performing interpolation, prediction and feature analysis on the fine grained air quality. In order to improve the performance of interpolation and prediction I use the characteristics of the spatial-temporal data and information contained in unlabeled data. I will also be adding feature selection in the input layer and removing the redundant data. By

combining the solution of interpolation, prediction and feature analysis in one I can reduce the redundant data and I will also get much better prediction result. Hence using these three techniques along with deep learning method I will predict the air quality.

ACKNOWLEDGMENT

I would like to express my very great appreciation to Ms Nidhi Sengar, Assistant professor, Maharaja agrasen institute of technology for her valuable and constructive suggestions during the planning and development of this research work. Her willingness to give her time so generously has been very much appreciated.

REFERENCES

[1]. Bellinger, C., Jabbar, M.S.M., Zaïane, O. and Osornio-Vargas, A., 2017. A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), p.907.

[2]. Michael Greenstone(University of Chicago), Santosh Harish (University of Chicago), Rohini Pande (Harvard University), Anant Sudarshan (University of Chicago), India Policy Forum July 11–12, 2017

[3]. D. Amarsaikhan, V. Battsengel, B. Nergui, M. Ganzorig, G. Bolor, February 2014, Institute of informatics , Mongolian Academy of sciences.

[4]. Kleine Deters, J., Zalakeviciute, R., Gonzalez, M. and Rybarczyk, Y., 2017. Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 2017.

[5]. NAAQS Table. (2015). [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table

[6]. Li, L., Zhang, X., Holt, J.B., Tian, J. and Piltner, R., 2011, December. Spatiotemporal interpolation methods for air pollution exposure. In Ninth Symposium of Abstraction, Reformulation, and Approximation.

[7] Liu, B.C., Binaykia, A., Chang, P.C., Tiwari, M.K. and Tsao, C.C., 2017. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjinshijiazhuang. PloS one, 12(7), p.e0179763.

[8] Saniei, R., Zangiabadi, A., Sharifikia, M. and Ghavidel, Y., 2016. Air quality classification and its temporal trend in Tehran, Iran, 2002-2012. Geospatial health.

[9] Zheng, Y., Liu, F. and Hsieh, H.P., 2013, August. U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1436-1444). ACM.