

Air Pollution Prediction using Machine Learning

Rohit Adke¹, Suyog Bachhav², Akash Bambale³, Bhushan Wawre⁴

¹Rohit Adke, Dept. of Computer Engineering, Sinhgad College of Engineering pune, Maharashtra, India

²Suyog Bachhav, Dept. of Computer Engineering, Sinhgad College of Engineering pune, Maharashtra, India

³Akash Bambale, Dept. of Computer Engineering, Sinhgad College of Engineering pune, Maharashtra, India

⁴Bhushan Wawre, Dept. of Computer Engineering, Sinhgad College of Engineering pune, Maharashtra, India

Abstract - Urban air pollution rate has grown to alarming state across the India. Most of the cities are facing issue of poor air quality which fails to meet standards of air for good health. It is indeed necessary to develop an air pollution measurement and prediction system for smart city. Nowadays almost every city has its own prediction system which uses linear regression. Exploitation browser to visualized accurate result of prediction. For improving linear regression algorithm with maximum accuracy, we are using neural network. In which it uses Multilayer Perceptron algorithm.

Key Words: Air Pollution, Linear Regression, Multilayer Perceptron

1. INTRODUCTION

As per World Health Organization [WHO] air pollution is infectivity of the indoor or outdoor environment by any chemical and biological agent which changes characteristics of the environment. Household combustion devices, vehicles and forest fires are common origin of air pollution and noise pollution. Pollutants which are responsible for health concern include particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulphur dioxide. Air pollution cause respiratory and other diseases, which can be deadly. WHO has measured quality of air in approximately 1500 cities and Indian capital city was the one of the most polluted cities around the world. Pune is having highest concentration of particulate matter which is smaller than 2.5 micrometre.

Air pollution and lack of air quality monitoring points represent environmental and technological challenges for cities and environments around the world. To face this issue, industry has focused its efforts in finding a versatile technological alternative that allows the improvement of the air quality measuring process and provides reference values in network sites where conventional monitoring fails to cover appropriately. Unfortunately, existing products and the generated results do not represent low-cost solutions.

1.1 Different Sources of Air Pollution

There are different sources of air pollution like Vehicular emission, Industrial emission, Construction and Demolition etc. These are given below.

1. Vehicular Emission: Due to increase in vehicular traffic that includes commercial vehicles, vehicles with gears and without gears and heavy load vehicles in Pune, leads to

increase the pollution in the air and vehicles are considered to be the major cause of air pollution.

2. Industrial Emission: As the new technologies are evolving day by day so new factories are being installed. This has caused air pollution by emitting harmful smoke, gases etc.

3. Road dust, Construction and Demolition: Pavement dust and construction activities in Pune are considered to be a major contributor of increasing particulate pollution.

1.2 Our Study Area

We have taken Pune as our study area as shown in Fig. 1. Pune is one of the most polluted cities in the country. Peak levels of fine particulate matter (PM) in Pune increased by about 70%, basically due to high industrial and vehicular emissions, construction work and crop burning. The level of the airborne particulate matter- PM_{2.5} is very high in Pune. It is considered to be the most harmful pollutants to health.

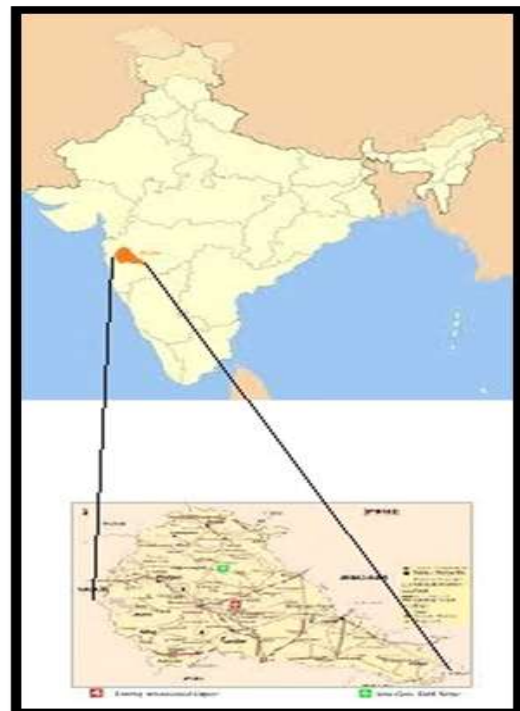


Fig -1: Study Area: Pune

2. LITERATURE REVIEW

In literature, a lot of work is done in the study and analysis of air pollution as well as predicting the future trends.

In [2] Linear regression-based air pollution prediction is done. It suggests cloud data for data analytics which can be used for taking the decision to minimize pollution. But they have used BI service and Microsoft Azure for analysis which is very expensive services. The model is not very accurate because of linear regression-based model.

In [1] machine learning based air pollution prediction is done. It suggests multilayer perceptron which results in very accurate result. But it takes large datasets and long duration for training.

In [3] Recurrent Neural Network based model for air pollution prediction is done. It suggests using machine learning algorithm and recurrent neural network for prediction which generates most accurate result but, its very expensive to implement.

3. PROPOSED APPROACH

3.1 Data set used

We have taken the data from Central Pollution Control Board (CPCB) [4]. This data set consists of six attributes that are time (in months), air pollutants like SO₂, NO₂, CO, PM10 and Ozone (O₃). Data is collected from 2000 to 2018 to predict the trends of the above pollutants in upcoming years. A snapshot of data stored in .csv file format is shown in Fig. 2.

Fig 2: Data set Used

3.2 Flow Chart of Proposed Work

The flow chart of the proposed work is shown in Fig. 3. After collecting the data, it is preprocessed. In preprocessing data are cleaned by removing noise and filling up the missing values. We have used multilayer perception for prediction and presented it to the user on the app.

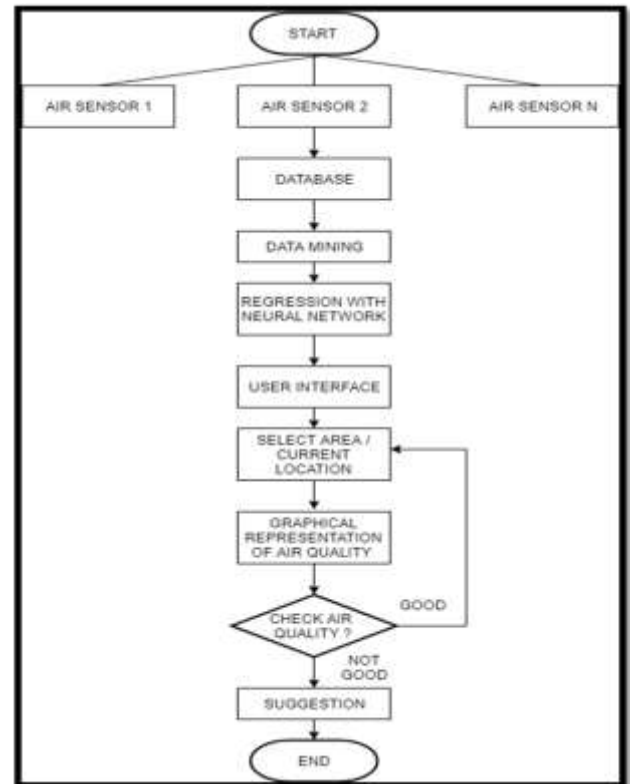


Fig 3: Flowchart of proposed work

3.3 Multilayer Perceptron

It is a class of feed forward artificial neural network. It consists of at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back- propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.

Activation Function:

A multilayer perceptron has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron, then linear algebra shows that any number of layers can be reduced to a two-layer input-output model. In MLPs some neurons use a nonlinear activation function that was developed to model the frequency of action potentials, or firing, of biological neurons.

The two common activation functions are both sigmoid, and are described by

$$y(v_i) = \tanh(v_i) \text{ and } y(v_i) = (1 + e^{-v_i})^{-1}$$

The first is a hyperbolic tangent that ranges from -1 to 1, while the other is the logistic function, which is similar in shape but ranges from 0 to 1. Here is the output of the ith

node (neuron) and is the weighted sum of the input connections.

Layers:

The MLP consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly-activating nodes. Since MLPs are fully connected, each node in one layer connects with a certain weight w_{ij} to every node in the following layer.

Learning:

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result.

This is an example of supervised learning, and is carried out through backpropagation, a generalization of the least mean squares algorithm in the linear perceptron.

We represent the error in output node j in the n th data point by $e_j(n) = d_j(n) - y_j(n)$, where d is the target value and y is the value produced by the perceptron. The node weights are adjusted based on corrections that the error in the entire output given by

$$\mathcal{E}(n) = \frac{1}{2} \sum_j e_j^2(n)$$

Using gradient descent, the change in each weight is

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

Where y_k is the output of the previous neuron and η is the learning rate, which is selected to ensure that the weights quickly converge to a response, without oscillations.

The derivative to be calculated depends on the included local field v_j , which itself varies. It is easy to prove that for an output node this derivative can be simplified to

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n) \phi'(v_j(n))$$

The analysis is more difficult for the change in weights to a hidden node, but it can be shown that the relative derivative is

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'(v_j(n)) \sum_k -\frac{\partial \mathcal{E}(n)}{\partial v_k(n)} w_{kj}(n)$$

This depends on the change in weights of the k th nodes, which represent the output layer. So, to change the hidden layer weights, the output layer weights change according to the derivative of the activation function, and so this

algorithm represents a backpropagation of the activation function

4. CONCLUSIONS

The agenda of our work is not only to bring awareness but also to minimize pollution through proper measures and ensure that the vehicles are emitting the pollutants within the range of regular pollution check. This can lead to a pollution free region in the area.

Proper pavement construction, greening of roadside space along with strict pollution laws, converting HMV from diesel to CNG, regular pollution check on two wheelers and checking industrial emission will go a long way in controlling air pollution in Pune region.

REFERENCES

- [1] Kettun Oberoi, Predicting Trends in Air Pollution in Delhi using Data Mining, IEEE-2016, 978-1-4673-6984-8/16/531.00, January 2016.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [2] Nitin Sadashiv Desai, IoT based air pollution monitoring and predictor system on Beagle Bone Black, IEEE-2017, 978-1-5090-5913-3/17/531.00, March 2017.
- [3] Yue Shan Chang, Big data platform for air quality analysis and prediction, IEEE-2018, 978-5636-4959-6/18/531.00, February 2018.
- [4] CPCB (Central pollution Control Board), 2010 Air quality monitoring, emission inventory and source appointment study for Indian cities. <http://cpcb.nic.in/Pune.pdf>