International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

A SURVEY ON AIR POLLUTANTS CONCENTRATION PREDICTION USING MACHINE LEARNING APPROACHES

POONGOTHAI R[1], DR.S.RATHI[2]

[1]*PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamilnadu, India*
[2]*Associate Professor, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamilnadu, India*

Abstract - *Air pollution is considered to be a major problem for people living in metropolitan cities. This is because air pollution significantly affects the lives of people and living in a pollution free environment signifies better quality of life. The main causes of air pollution are emissions from different transport modes, burning of fossil fuels, industrial production, forests fires, etc. Several machine learning techniques are applied to predict the concentration of air pollutants such as ozone(O_3), carbon monoxide(CO), sulphur dioxide(SO_2), total nitrogen oxide(NO_x), nitrogen dioxide(NO_2), PM2.5, PM10, NH_3 based on the meteorological data such as absolute humidity, temperature, wind speed , wind direction, etc. In this survey, we analyze the research efforts revolving on machine learning approaches for predicting the concentration of various air pollutants. This analysis will focus on evaluating the performance of the various machine learning approaches by using root mean square error(RMSE) as the evaluation metric.*

Key Words: air pollution, prediction, machine learning approaches, analysis, RMSE

1. INTRODUCTION

Air pollution is considered to be a major problem for people living in metropolitan cities. It significantly affects the lives of people and living in a pollution free environment signifies better quality of life. It causes around seven million deaths a year worldwide. This is mainly due to the growth of the population in cities, as well as the way in which we consume energy in urban areas through transport or heating systems result in the emission of huge quantities of gases that are harmful to our health. The main causes of air pollution are emissions from different transport modes, burning of fossil fuels, industrial production, forest fires, etc.

A number of meteorological conditions are critical in determining the air pollutant concentrations. Some of them are air temperature, wind speed and direction, relative humidity, incoming solar radiation, etc. Lowering the ambient temperature and incoming solar radiation slow down photochemical reactions and lead to less secondary air pollutants, such as O_3. Increasing wind speed could either increase or decrease the air pollutant concentrations. High humidity is usually associated with high concentrations of certain air pollutants such as CO and SO_2, but with low concentrations of certain air pollutants such as NO_2 and O_3.

Major outdoor air pollutants in cities include ozone(O_3), carbon monoxide(CO), sulphur dioxide(SO_2), total nitrogen oxide(NO_x), nitrogen dioxide(NO_2). It is widely believed that urban air pollution has a direct impact on human health especially in developing and industrial countries, where air quality measures are not available or minimally implemented [1]. Considering the significance of air quality on human lives, the World Health Organization(WHO) has developed guidelines for reducing the health effects of air pollution on public health by setting the limits of the concentrations of various air pollutants, some of which are ground-level ozone(O_3) sulphur dioxide(SO_2), and nitrogen dioxide(NO_2) [2].

This paper aims at analyzing the research efforts revolving on machine learning approaches for predicting the concentration of various air pollutants by using Root Mean Square Error (RMSE) value as the evaluation metric.

2. LITERATURE SURVEY

Air pollutants dataset for our survey has been collected from Central Pollution Control Board(CPCB) official website.

- *City name*: Coimbatore
- *Station name*: SIDCO Kurichi, Coimbatore- Tamilnadu
- *From date*: 01/08/2019
- *To date*: 30/10/2019

The dataset includes concentration of air pollutants, which are used in AQI(Air Quality Index) calculation, such as ozone(O_3), carbon monoxide(CO), sulphur dioxide(SO_2), total nitrogen oxide(NO_x), nitrogen dioxide(NO_2), PM2.5, PM10, NH_3 and meteorological data such as relative humidity(AH), absolute temperature(AT), wind speed(WS) , wind direction(WD).

The collected dataset must be preprocessed before applying machine learning algorithms. Air pollutant dataset has some fields, in which some value in the record are missing. Those values are replaced with the mean value using Weka 3.6.6. This process is known as data cleaning. At last, the dataset are partitioned into training and testing datasets.

Our problem of predicting the concentration of various air pollutants comes under the supervised machine learning technique known as "Regression". The regression problem is a generalization of the classification problem, in which the model returns a continuous-valued output, as opposed to an output from a finite set. Machine learning algorithms that has been considered for our prediction task are as follows

- Support Vector Regression
- Decision Tree Regression
- Artificial Neural Network
- Radial Basis Function
- Linear Regression

Table -1: Description of dataset attributes

S.NO	ATTRIBUTE	DESCRIPTION
1	From Date	DD-MM-YYYY HH:MM
2	To Date	DD-MM-YYYY HH:MM
3	NO2	Concentration of nitrogen di-oxide in micrograms per cubic meter
4	NOx	Concentration of total nitrogen oxide in ppb
5	SO2	Concentration of Sulphur di-oxide in micrograms per cubic meter
6	CO	Concentration of Carbon monoxide in mg/m^3
7	NH3	Concentration of Ammonia in micrograms per cubic meter
8	Ozone	Concentration of Ozone in micrograms per cubic meter
9	PM2.5	Concentration of particulate matter(2.5) in microgram per cubic meter
10	PM10	Concentration of particulate matter(10) in microgram per cubic meter
11	AT	Absolute Temperature in degree C
12	RH	Relative Humidity in %
13	WS	Wind Speed in m/s
14	WD	Wind Direction in degree

2.1 Support Vector Regression

SVMs solve binary classification problems by formulating them as convex optimization problems. The optimization problem entails finding the maximum margin separating the hyperplane, while correctly classifying as many training points as possible. SVMs represent this optimal hyperplane with support vectors. The sparse solution and good generalization of the SVM lend themselves to adaptation to regression problems. SVM generalization to SVR is accomplished by introducing an ε-insensitive

region around the function, called the ε-tube. Although less popular than SVM, SVR has been proven to be an effective tool in real-value function estimation.

- Epsilon(ε) is set to 0.1 for optimizing the margin
- Kernel function used is Radial Basis function(RBF) which is used to transform the data into a higher dimensional feature space to make it possible to perform the linear seperation

Gaussian Radial Basis function

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

The RMSE value obtained, using SVR, for NO_2 is 7.337, NO_x is 8.911, SO_2 is 1.099, CO is 0.215, NH_3 is 10.548, Ozone is 13.064, PM2.5 is 9.611 and PM10 is 17.965

2.2 Decision Tree Regression

Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two or more branches each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**.

Parameters:

- criterion: “mse” (mean squared error) which is used to measure the quality of the split.
- splitter: “best” which is the strategy used to choose the best split.
- max_depth: None which indicates the maximum depth of the tree. If none, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples
- min_samples_split: 50 i.e. The minimum number of samples required to split an internal node:

The RMSE value obtained, using Decision Tree Regression, for NO_2 is 9.972, NO_x is 11.544, SO_2 is 1.384, CO is 0.236, NH_3 is 10.52, Ozone is 9.363, PM2.5 is 12.137 and PM10 is 19.932.

2.3 Artificial Neural Network

Artificial Neural Network(ANN) is an information processing paradigm that is inspired from the brain. Multilayer Perceptron is one of the popularly used ANNs. This layer has hidden layer which is internal to the network and has no direct contact with the external layer. The number of hidden neurons should be between the size of the input layer and size of the output layer.

Parameters:

- hidden_layer_sizes:3
- activation: ‘relu’ represents the activation function for the hidden layer

  ‘relu’, the rectified linear unit function, returns f(x) = max(0, x)
- solver: “adam” which is used for weight optimization

'adam' refers to a stochastic gradient-based optimizer . The default solver 'adam' works pretty well on relatively large datasets (with thousands of training samples or more) in terms of both training time and validation score.

- learning_rate:'constant' represents the learning rate schedule for weight updates.

The RMSE value obtained, using ANN, for NO_2 is 6.502, NO_x is 7.809, SO_2 is 1.111, CO is 0.214, NH_3 is 8.267, Ozone is 7.675, PM2.5 is 9.122 and PM10 is 16.943.

2.4 Radial Basis Function Network

RBF network is an artificial neural network with an input layer, a hidden layer, and an output layer. Hidden layer generates a signal corresponding to an input layer, and corresponding to this signal, network generates a response.

Parameters:

- hidden_layer_sizes:3
- activation: 'rbf' represents the activation function for the hidden layer

The RMSE value obtained, using RBF network, for NO_2 is 6.301, NO_x is 7.601, SO_2 is 1.001, CO is 0.154, NH_3 is 7.231, Ozone is 7.461, PM2.5 is 8.996 and PM10 is 16.543.

2.5 Linear Regression

Linear Regression is the basic and commonly used type for predictive analysis. It is a statistical approach to modeling the relationship between a dependent variable and a given set of independent variables. Multiple Linear Regression attempts to model the Relationship between two or more features and a response by fitting a linear equation to observed data.

Y=b0+b1*x1+b2*x2+b3*x3+……bn*xn
Y = Dependent variable and

x1, x2, x3, …… xn = multiple independent variables

b = regression coefficient

Linear regression fits a linear model with regression coefficients to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation

Parameters:

- fit_intercept: True which is used to indicate whether to calculate the intercept for this model
- normalize: True, the regresors X will be normalized before regression by subtracting the mean and dividing by the l2-norm

The RMSE value obtained, using LR, for NO_2 is 6.607, NO_x is 8.109, SO_2 is 1.055, CO is 0.185, NH_3 is 8.396, Ozone is 8.359, PM2.5 is 9.308 and PM10 is 17.29

3. RESULT ANALYSIS

The performance evaluation metric used is Root Mean Square Error(RMSE) value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

The RMSE value obtained for air pollutants using various machine learning algorithm are as follows

Table -2: RMSE value using various Machine Learning approaches

RMSE value	SVM	DTR	ANN	RBF Network	Linear Regression
NO_2	7.337	9.972	6.502	6.301	6.607
NO_x	8.911	11.544	7.809	7.601	8.109
SO_2	1.099	1.384	1.111	1.001	1.055
CO	0.215	0.236	0.214	0.154	0.185
NH_3	10.548	10.52	8.267	7.231	8.396
Ozone	13.064	9.363	7.675	7.461	8.359
PM2.5	9.611	12.137	9.122	8.996	9.308
PM10	17.965	19.932	16.943	16.543	17.29

RMSE is an absolute measure of fit of the model to the data. Lower values of RMSE indicates better fit. RMSE has the same units as the quantity being estimated.

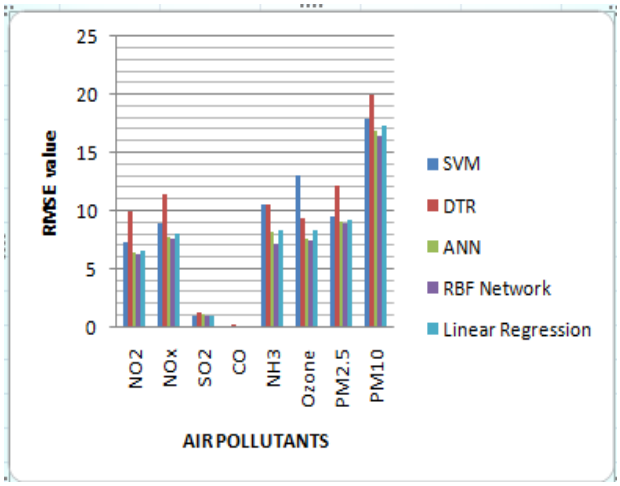

Chart -1: RMSE value obtained for various air pollutants

The above results show that Artificial Neural Network works better when compared to support vector regressor and decision tree regressor. It is found that the RBF neural network performs better than the ANN. However, all kinds of NNs produced similar and good estimates of pollutant concentrations.

4. CONCLUSION

From the survey, we conclude that among various machine learning algorithms used, neural networks are recognized as the state of the art approach for predicting the air pollutants concentration . It is found that linear regression models can in some cases be better than the other models such as decision tree regression, support vector machine, etc,. The results show that the RMSE value obtained for CO(Carbon Monoxide) is very much less when compared to all other pollutants. At the same time the

RMSE value obtained for PM10 is very high when compared to all other pollutants. Thus we can say that the models fit very well for predicting the CO gas when compared to PM10.

REFERENCES

[1] World Health Organization, "Monitoring ambient air quality for health impact assessment," WHO Regional Office Eur., Copenhagen, Denmark, Tech.

[2] World Health Organization, "WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide," World Health Org., Geneva, Switzerland, Tech.

[3] Khaled Bashir Shaban; Abdullah Kadri. Urban Air Pollution Monitoring System with forecasting models. IEEE Journal. 2016

[4] Pietro Zito; Haibo Chen; Margaret C.Bell. Predicting Real-Time Roadside CO and NO_2 concentrations using neural networks. IEEE. 2008

[5] Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 2017, 150, 146–161.

[6] Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.

[7] Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. *Adverse health effects of outdoor air pollutants*.Environ. Int. 2006, 32, 815–830.

[8] Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. Environ. Model.Softw. 2001, 16, 263–272.

[9] Jiang, D.; Zhang, Y.; Hu, X.; Zeng, Y.; Tan, J.; Shao, D. Progress in developing an ANN model for air pollution index forecast. Atmos. Environ. 2004, 38,7055–7064.