ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 11, November 2019

Review on Data Mining Techniques for Prediction of Air Quality

[1] Akhil Chandran N, [2] Nibin Sasidharan, [3] Aswin Kumar P M, [4] Siji Joju
[1][2][3][4] Mtech CSE Students
[1] akhilchandrann.chandran@gmail.com,[2]nibinmsasidharan@gmail.com, [3]aswinkumarpm3@gmail.com, [4]sijijoju91@gmail.com

Abstract: **- Environmental pollution has mainly been attributed to urbanization and industrial developments across the globe. Air pollution has been marked as one of the major problems of metropolitan areas around the world. especially in delhi ,the capital of India, where its administrators and residents have long been struggling with air pollution damage such as the health issues of its citizens. As far as the study area of this research is concerned, a considerable proportion of delhi air pollution is attributed to PM10 and PM2.5 pollutants. Therefore, the present study was conducted to determine the prediction models to determine air pollutions based on PM10 and PM2.5 pollution concentrations in delhi. To predict the air-pollution, the data related to day of week, month of year, topography, meteorology, and pollutant rate of two nearest neighbors as the input parameters and machine learning methods were used. These methods include a regression support vector machine, geographically weighted regression, artificial neural network and auto-regressive nonlinear neural network with an external input as the machine learning method for the air pollution prediction. A prediction model was then proposed to improve the afore-mentioned methods, by which the error percentage has been reduced and improved by 57%, 47%, 47% and 94%, respectively. The most reliable algorithm for the prediction of air pollution was autoregressive nonlinear neural network with external input using the proposed prediction model, where its one-day prediction error reached 1.79 _g/m3. Finally, using genetic algorithm, data for day of week, month of year, topography, wind direction, maximum temperature and pollutant rate of the two nearest neighbors were identified as the most effective parameters in the prediction of air pollution.Pollutants in the atmosphere is increasing day by day. The gradual increase of pollutants in the atmosphere results a severe impact on environment. So we introduce a method to predict the amount of pollutants in atmosphere in future using deep learning. The Machine learning provides different techniques to train the machine based on experience.**

***Index Terms*—Air Quality; Data mining techniques**

1. INTRODUCTION

Air pollution has been the subject of many present environmental studies due to the inclination of industrialization as well as a temporal correlation between long-term exposure to fine particulate matter and acute increases in mortality including lung cancer and cardiopulmonary [8]. . The primary air pollutants in urban areas include carbon dioxide (CO2), carbon monoxide (CO), nitrogen oxides (NO2), nitrogen monoxide (NO), and particulate matters PM2.5, PM10. However, the most concerned air pollution factor is PM2.5 or particulate matter that is up to 2.5 microns in diameter. These particles are tiny and light allowing them to stay in the atmosphere for a more extended period and also easily bypass the filters of human nose and throat due to their size property. According to C. Arden Pope [8], each 10-g/m3 elevation in long-term average PM2.5 ambient concentrations was associated with approximately 4-8 percent increased the risk of cardiopulmonary and lung cancer mortality. Air pollution is one of the most important environmental issues in both developed and developing countries. Air pollution means the existence of one or more pollutants contaminating outdoor or

indoor air in various amounts and periods which may harm human, vegetation or animal life or unexpectedly interacts with normal life or properties. The distribution of air-pollution involves a complex process depending on a number of factors. In fact, air pollution prediction, which has a non-linear dynamism, is a very difficult task and requires a close understanding of the dispersion of air pollutants in the atmosphere, which involves an immense cost . In some cases, air pollution in mega cities even exceeds the standard limit which increases the concerns. For this reason, air pollution has become a problem in many cities in the world and its investigation is considered

as a vital issue in urban management. The general sensitivity towards this problem has urged the officials to pass laws in order to prevent the air-pollution .One of urban managers' objectives is to provide the citizens with the right information to make them aware of air quality rates. The pollution information\ includes the density of daily PM2.5 and PM10 pollutants which can be announced to the concerned people by city managers as a response to the air pollution . This information may assist people to avoid the polluted areas and employ public transport facilities to reduce the level of the pollution. In addition, the concerned city managers can implement the information to control the urban traffic and the responsible pollutant industries and to increase public transport facilities in order to mitigate the level of the pollution. To achieve this goal, appropriate tools need to be used to predict air pollution

The air quality in Delhi, the capital of India, according to a WHO survey of 1600 world cities, is the worst of any major city in the world.[1][2] Air pollution in India is estimated to kill 1.5 million people every year; it is the fifth largest killer in India. India has the world's highest death rate from chronic respiratory diseases and asthma, according to the WHO. In Delhi, poor quality air irreversibly damages the lungs of 2.2 million or 50 percent of all children.

India's Ministry of Earth Sciences published a research paper in October 2018 attributing almost 41% of PM2.5 air pollution in Delhi to vehicular emissions, 21.5% to dust and 18% to industries.[3] The director of Centre for Science and Environment (CSE) alleged that the Society of Indian Automobile Manufacturers (SIAM) is lobbying "against the report" because it is "inconvenient" to the automobile industry.

Air quality index of Delhi is generally Moderate (101-200) level between January to September, and then it drastically deteriorates to Very Poor (301-400), Severe (401-500) or Hazardous (500+) levels in three months between October to December, due to various factors including stubble burning, fire crackers burning during Diwali and cold weather. In November 2017, in an event known as the Great smog of Delhi, the air pollution spiked far beyond acceptable levels. Levels of PM2.5 and PM 10 particulate matter hit 999 micrograms per cubic meter, while the safe limits for those pollutants are 60 and 100 respectively.

PM2.5 contaminants contain particles that are created by combustion or caused by the formation and compression of secondary particles. PM10 particles contain particles that are 10 micrometers in diameter and smaller and can pass through the first defensive barrier (nose and throat), damage the lungs and deposition there [8]. Studies have shown that exposure to suspended particles is associated with health effects such as cardiovascular and respiratory diseases . The World Health

Organization estimates that if the average annual concentration of PM10 is reduced from 70 _g/m3 to 20 _g/m3, then the associated deaths will be reduced by 15% . In fact, there is a relationship between the exposure to intense concentrations of suspended particles and the increase in daily and annual mortality, as well if the concentration of these pollutants is reduced while other factors are fixed then the associated deaths are reduced . These particles are very tiny and their damage to human health is high. In this study, PM2.5 and PM10 are used as pollutants to predict air pollution. Hence, air-pollution prediction is becoming one of the managerial solutions to prevent and/or mitigate its destructive implications. Therefore, it seems necessary to predict PM10 and PM2.5 pollutants using the appropriate methods. In the past few decades, two general approaches of deterministic and stochastic methods have been used to predict air-pollution. Diffusion models are among the deterministic methods developed in various regions for modeling and monitoring the air pollution. However, the output of these models relies on the input data, and in order to use them, it is necessary toaccess the data on how the pollutants disseminate and diffuse in the atmosphere

Therefore, using these models where sufficient and precise data is not accessible is problematic. Considering that the data collection needed for diffusion models is very hard and impossible at large scales, the researchers have turned to superior methods such as statistical models . Compared to the deterministic methods, statistical methods have more application in prediction of air-pollution. It is worth mentioning that factors such as air pressure, temperature, humidity, rainfall and wind affect the pollutants dissemination A study has been conducted by with the aim of predicting the density of two pollutants (CO and NOx) in industrial locations using the autoregressive model based on artificial neural network using some meteorological parameters. As a result of performance of the proposed model, Root Mean Square Error (RMSE) for CO and NOx pollutants was 0.8445 and 0.7618, and the mean absolute error (MAE) for the pollutants was 0.1451 and 0.1598, respectively. The results show the higher importance of meteorology variables in the prediction of pollutant concentration and the efficiency

of the neural network in the air pollution prediction The authors of introduced a model to improve the artificial neural network, which is a combination of air mass route analysis and wavelet transform. The rate of RMSE for the combinational model can be decreased by 40% on average. The study verified that especially on the days with a higher concentration of PM2.5 often predicted for the warned threshold of the combinational models using wavelet analysis and detection rate (DR), the RMSE can reach to the average limit of 90%. This approach indicates the potential of the proposed model in air-quality prediction system in other countries. In this research, the supervised algorithms for machine learning regression such as artificial neural network (ANN), the nonlinear autoregressive exogenous Neural Network, geographically weighted regression (GWR) and support vector regression (SVR) were used to predict PM2.5 and PM10 pollutants. To generate a dataset to be entered in the machine algorithm process, it is necessary first to interpolate the meteorological parameters for transfer of parameters from meteorology stations to air-pollution stations. In general, two types of methods including satellite imagery and ground sensors are used to collect air pollution data. Given the cost, availability and accuracy of ground sensor data for 10 years, this type of data has been used in this research

2. DATA MINING TECHNIQUES

Data Mining is the process of turning raw data into appropriate and meaningful information. Various researchers have studied and work on data mining techniques to evaluate and classify the water quality .**necessary for the proper interpretation of your figures.** There is an additional charge for color printing.

A. ANN (Artificial Neural Network)

ANN is a classification model which is grouped by interconnected nodes. It can be viewed as a circular node which is represented as an artificial neuron that reveals the output of one neuron to the input of another. The ANN model is helpful in revealing the unexposed interrelationships in the classical information, therefore expiditing the prediction idea, envision and forecasting of water quality. Based on their performance metrics, we use various formulas which have been illustrated in .ANN model is definite and systematic enough to make important and relevant decisions regarding data usage.

B. Naïve Bayes

Naïve Bayes is a classification technique which is based on probability theories which entirely demonstrate the characteristics of water quality assessment. Bayes model is easy to use for very large datasets. In other terms, a Naive Bayes assumed that the value of a distinct feature does not related to the presence or absence of any other feature, given in the class variable. It undergoes through following steps:

1. Extract, clean and classify the water quality.
2. Remove large punctuations and split them.
3. Counting Tokens and calculating the probability. This probability is called as posterior probability which is calculated by the formula described in.
4. Adding the probabilities and then wrapping up

C. Decision Tree

Decision tree is one of the predictive modeling technique used in data mining. It aids to divide the larger dataset into smaller dataset indicating a parent-child relationship. Each internal node defined as inner node is labeled with an input feature. The inner nodes which exhibit many types of attribute test, bifurcations exhibit the test outcomes and leaf nodes particularly exhibit the category of a specific type[4]. Decision tree can handle both numerical and categorical data. It is well suited with large datasets. Higher accuracy in decision tree classification technique depicts that the technique can simulate. It is able to optimize variety of input data such as nominal, numeric and textual. It is a successful supervised learning approach which has the capability of extracting the information from vast amount of data based on decision rules.

D. FNN(Fuzzy Neural Network)

ANN is basically associated with the neurons having the capability of storage and processing for the information. FNN is chosen as a algorithm for data mining introducing the artificial neural networks. It describes the integration of fuzzy Fuzzy logic with the neural network. Fuzzy neural network algorithm which deals with the prediction process is composed of five layers named as: input layer, hidden layer, fuzzification layer, fuzzy reasoning layer and reconciliation fuzzy layer. Zhu has explained the structure of FNN .

E. Back Propogation Neural Network(BPNN)

ANN consists of interconnected processing units. Each unit is known as neuron. Each neuron will receive an input

from another neuron. Weights are assigned to each neuron. These kinds of weights regulate the nature as well as strength and power of the significance involving the interlocked neurons. The respective signals named as indicators tend to be refined from each and every input and then further processed via a weighted sum to the inputs. The BPNN algorithm criteria looks for the error with the method called as steepest descent. The united weights are modified by simply moving on the way to the negative gradient of the energy function by providing emphasis at each and every iteration for evaluating the network performance. Various performance metrics are used for calculating the network error based on specific formulas. This algorithm follows four major steps:

1. Feed forward computation.
2. Applying back propogation at the output layer.
3. Applying back propogation to the hidden layer.
4. Weights updation.

This algorithm will continue its processing until the value of error function becomes too small.

F. KNN

K-nearest neighbor is an algorithm which is used for regression and classifying the quality problems. It considers various parameters which results in the ease of calculation time and predictive power. It uses a vast amount of classes to calculate the likelihood score. When several KNNs share a class, then the weights of other neighbours to it also added together. Result of such added weights is considered to be the likelihood score. These scores are then sorted in order to find the ranked list. Therefore, KNN is a very simple and effective algorithm

CONCULSION

This paper presents an evaluation for predicting air quality by applying numerous data mining techniques and methods at many different locations. Many existing evaluation methods are studied. Various algorithms have been reviewed for predicting the water quality and hence made a comparison. As a result of analyses, Artificial neural network is used frequently.

REFERENCES

[1] Vikram Reddy, Deep Air: Forecasting Air Pollution in Beijing, China (2017)

[2] Woosuk Jung, South Korea's Air Pollution: Gasping for Solutions (2017)

[3] Daniel L. Marino, Building Energy Load Forecasting using Deep Neural Networks (2016)

[4] Xiaochen Chen, House Price Prediction Using LSTM (2017)

[5]P. Kingma, Adam – A method for stochastic optimization (2014)

[6]Wojciech Zaremba, Recurrent Neural Network Regularization (2014)

[7]Kyunghyun Cho, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (2014)

[8]C. Arden Pope, Lung Cancer, Cardiopulmonary Mortality and Long-term Exposure to Fine Particulate Air Pollution (2002) doi:10.1001/jama.287.9.1132