Toxic Pollution Alert using Fog computing and Hybrid Machine learning Model(RAQ and ARIMA)

D.Sudaroli Vijayakumar[1]

[1]*Department of Information Technology, Alliance University, Bangalore, India*

sudaroli.d@alliance.edu.in

***Abstract*—**Pollution remains as a hot topic discussed globally irrespective of various measures and techniques adopted to improve the air quality. The level of pollution in many cities still exceed the limits of world health organization thus contributing various deadly diseases like cancer, stroke, heart disease and bronchitis. We find lot of compromises in the way the toxic alerts are generated as the alerts are generally generated based on the data that we obtain from the large scientific air monitoring system that are permanently installed at a specific location. The amount of data that these monitoring systems provides is relatively small and thus fatal toxic smog like occurrence prevails. Toxic smog and similar occurrences can be effectively addressed if we possess a large dataset from multiple sources. Advancements in communication technology creates an opportunity for us to integrate several sources of data through Internet of Things(IoT). This big data can be analyzed effectively with the aid of machine learning algorithms to provide a better recommendation on the toxic level in air. Along with the big data analytics, introduction of fog computing can improve the latency in data. The aim of this paper is to integrate the machine learning approaches along with the fog computing concepts to provide a toxic pollution alert.

Keywords: fog computing, cloud computing, IoT, *Machine Learning, ANN, RAQ, SVM*

.

1. INTRODUCTION

We are embraced and overwhelmed with the advent of new technologies and its contribution towards the economy. However, the other side of this growth left us with unsolvable issues that challenges the existence of human life. Depletion of ozone layer, formation of acid rain, haze, eutrophication, global climate change is all some of the buzz words that we often encounter in our day to day life. All these arises because of the urban air quality and indoor air pollution. Hence pollution in air is one of the alarming concerns for us today. This issue is being addressed in the past by creating models using the conventional approaches. According to Niharika et al., [1], most of these conventional approaches uses mathematical and statistical approaches to build a physical model and the data is coded with mathematical equations. But they failed as these models were unable to predict the extreme points of pollution, all data were treated the same irrespective of when it got generated and solving complex mathematical equations. Another important factor that is noticed in the conventional approach is the need for large computing resources. All these complexities to solve mathematical equations are ruled out with the enormous amount of data generated by the IoT devices and the intelligent machine learning approaches and the complexity relative to the computing resource can be handled with fog computing. This paper is organized as follows: Section 2 discusses the basic architecture of IoT, fog computing, air quality index value along with its relevance to our considered problem statement. Section 3 discusses the various work relative to the handling of big data along with machine learning approaches. Section 4 tries to present the hybrid approach integrating the machine learning and fog computing. Section 5 concludes the paper highlighting the future scope.

2. CONCEPTS OVERVIEW

2.1 Internet of Things(IoT)

IoT gained lot of attraction because of the impact it made across connectivity and economy. It is being anticipated by 2025, there will be 100 million IoT connected devices that will provide $100 trillion business. With such a huge importance, it becomes unavoidable to understand the working of IoT. Several definitions exist, however the easiest way to define IoT is the extendibility of network capability and computing to all the real time objects. Rajesh et al., [2] explains the concept of IoT with four important parameters: objects, data, people and process.

The concept of IoT starts with the physical devices that holds embedded technology to connect to external environment, from where the information about that device is collected so called as data. This unstructured massive information from various devices holds no meaning if it doesn’t provide any useful information. The process of converting this big data into meaningful information is carried out by human assistance which we call it as big data analytics. After the analysis procedure, the processed information is either delivered to machine or people.

The general method to enable these IoT devices is either by using RFID tags or Wireless networks. RFID tags are used to track the information about whom it is attached, and it doesn’t carry any external batteries. This simplicity gained attention in this work as the low-cost static and mobile IoT sensors can be deployed easily in bikes, people and buildings. These portable IoT sensors can travel freely across parks, pedestrian routes, major roads and side roads. They will provide more granular data as they assess air quality in real time. The different forms of data that must be considered for analysis to arrive at more appropriate prediction are the ones from satellites, emission databases, atmospheric models and the roadside granular data. This huge collection of data is made available to us with the IoT.

2.2 Fog Computing

With the advent of IoT, unprecedented volume and variety of data is the privilege we can enjoy, however latency is the major concern when these IoT data is processed using cloud. Today's cloud models cannot hold such a huge volume, variety and velocity of data that the IoT generates. Current statistics shows more than two exabytes of data are generated every day and it is expected by 2020 more than 50 billion things will be connected to Internet by 2020. Moving such a huge volume for cloud analysis would require vast amount of bandwidth. Another concern is most of the devices are machines that connect to controller using industrial protocols. This must be converted to IP before sending to cloud for analysis or storage. Along with these problems, our problem statement needs the fast processing of huge volume of data. If there is a delay in analyzing all the pollutant data, then the pollution alert will be slow that can create destructions. The ideal place to analyze most IoT data is near the devices that produce and act on the data and only the historical data is sent to cloud can improve latency. This concept is called fog computing. Any device that possess storage, computing and network connectivity can be a fog node. Fog nodes are the ones that will perform the analysis. Fog computing is a suitable option [3] when data is collected at the extreme edge, large geographic area are generating data, analyze and act on the data in less than a second.

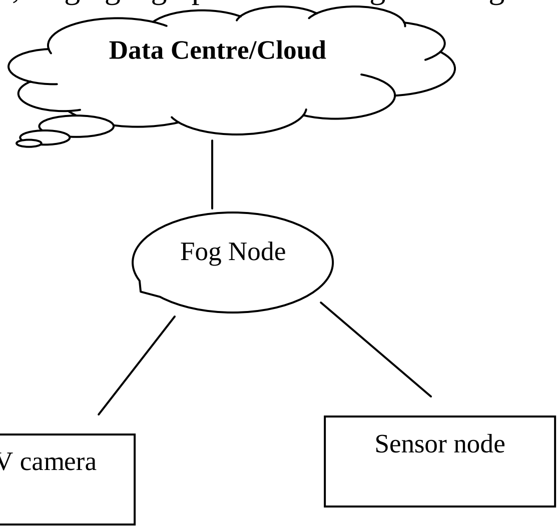


Figure 1: Architecture of Fog Computing

The most time-sensitive data are generally analyzed using the Fog node. Local and regional granular data can be aggregated and mined at fog nodes and thus timely feedback on the toxicity can be addressed rapidly.

2.3 Air Quality Evaluation:

Air quality is complex to understand and predict, as it is determined by a wide range of factors including traffic levels, industrial activity, weather conditions, temperature, wind speed, local topology and local built environment. Air quality evaluation is the only way to control air pollution. There are various types of pollutants exists however vitality is given only to the pollutants that causes immense problems to the human beings. Some of the pollutants that are identified as harmful to human beings are PM2.5, PM10, NO2, SO2, CO and O3. Along with the pollutants, Air Quality Index (AQI) is another common value used to denote the percentage of polluted air and this value is directly proportional to the health effects. Higher the AQI value, higher the adverse effects. To generate an appropriate toxic alert, a good prediction model is essential. Good prediction model can be obtained if the highly correlated input parameters are avoided and considering only the features that are highly prognostic.

3.Machine learning Air Quality Analysis:

According to the best of knowledge, studies that investigate the depth and breadth of the application of machine learning methods within air pollution epidemiology is meagre. However, S.Y. Muhammed et al. in [5] justifies that the machine learning model is best suited to solve this prediction problem. To arrive at a most appropriate solution, the idea about the existing machine learning approaches to solve this issue must be studied meticulously. Some of the works relative to the problem are as follows: E. Kalapanidas et al. in [4] addressed the effects of air pollution only from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity and classified air pollution into different levels (low, med, high, and alarm) using the feed forward multi-perceptron network and arrived at a Case Based reasoning system. This ANN based model showed promising results compared to that of the decision tree algorithms like ID3, C4.5 and its inherits.

Athanasiadis et al. [6] used a fuzzy lattice classifier to predict the categorization of O3 levels by considering the meteorological features and the other pollutants such SO2, NO2 etc. Kurt and oktay [7] predicted the daily concentrations level SO2, CO and PM10 using the neural network model. Even though they predicted the concentration three days in advance, inaccuracy prevailed as this model ignores the magnitude of numeric data. Zhao et al. [8] tried to create an improved ANN model by selecting the subset of factors from original set and the selected factors are fed into ANN for modelling which showed better results in neural network-based approaches to measure the air quality. Apart from these, other researchers tried to identify the concentration of pollutants using the neural network approach. Jiang et al. [9] compared the various models and identified that statistical models are more powerful than other models. This is an important finding to be considered in our alert mechanism. In most of the machine learning approaches that we saw so far concentrated to improve the model performance for a single task. With this understanding, highlighting some issues clarify to arrive at a more optimal solution.

i. To increase the air quality evaluation huge dataset is essential and that becomes possible only with the aid of IoT devices.
ii. Real time air quality monitor requires analysis in multiple levels so adopting a single model is not a viable option.
iii. Considering the meteorological as well as the IoT based data, hybrid machine learning models is the only option.

4.Hybrid Approach for Forecasting:

The first and foremost component required to perform efficient air quality prediction is the broad set of data. In this study, data is collected from multiple sources including air monitoring station data, meteorology data, traffic data, road information or the granular data fetched regularly at intervals of one hour. Once the dataset is obtained, partitioning the data set into test and train data. In this context, the direct AQI data from the monitoring stations are the train data. Data that originated apart from the monitoring station are used as test data. The training dataset includes all the features to make the prediction, 52585 entries are considered. The features that are considered for prediction are temperature, wind speed, relative humidity, traffic index, air quality of previous day.

By studying the various machine learning approaches, we concluded that the air pollution forecasting can be accurately done only with the aid of hybrid machine learning models. To identify the better prediction model, this dataset was verified with different models based on its F-Score, Relative Absolute error(RAE), precision and Receiver operating characteristic(ROC).

Algorithm	**Precision**	**F-Score**	**ROC**	**RAE**
Naive Bayes	52.1%	0.539	0.745	85.1%
ANN	72%	0.78	0.83	61.2%
RAQ	82%	0.816	0.94	37.1%

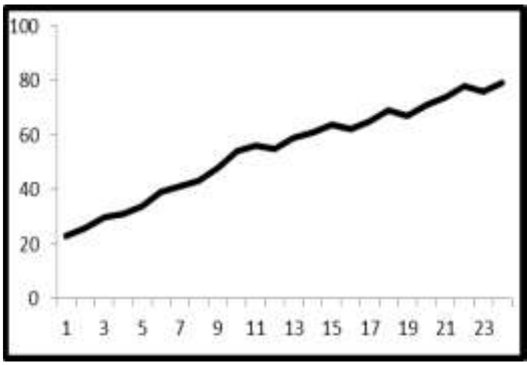
Table 1: Comparative analysis to select the best model

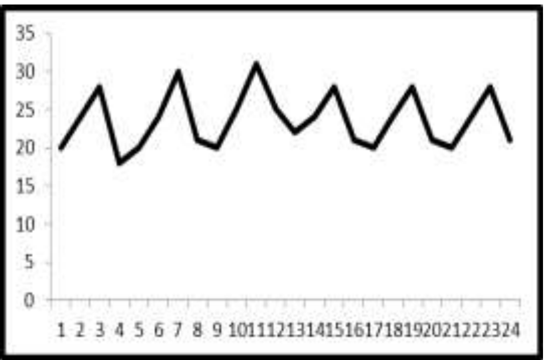
The above table clearly denotes that for the dataset chosen, the precision value didn't change irrespective of the changes in training set size variation. Lower the data points, precision didn't change so it clearly denotes that RAQ algorithm is best suitable for prediction using this dataset. The ensemble method adopted created different classifiers using different sampling strategies. Random forests are developed using the following steps:

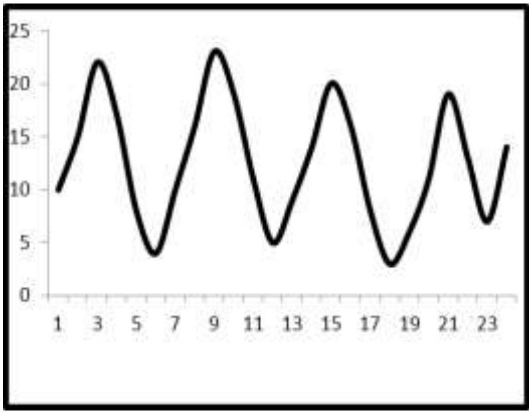
- Step 1 : Assume that the training data has N observations. One needs to generate several samples of size M ($M < N$) with replacement (called **Bagging**). Let the number of samples based on sampling of the training dataset be S_1.
- Step 2 : If the data has n predictors, sample m predictors ($m < n$).
- Step 3 : Develop trees for each of the samples generated in steps 1 using the sample of predictors from step 2 using CART.
- Step 4 : Repeat step 3 for all the samples generated in step 1.
- Step 5 : Predict the class of a new observation using majority voting based on all trees.

After adopting the ensemble approach for classification, depending on the various values of the features, the air quality is categorized. After obtaining the categorization value, at this instant of time for a location the air quality is known. Now the next point that must be considered is with this value, how can we forecast the toxic smog like occurrence.

This forecasting can be done by considering the time series data on a response variable. By observing this response variable at different intervals of time, we can forecast the nature of the response variable considering the trend, seasonal, cyclical and irregular component.







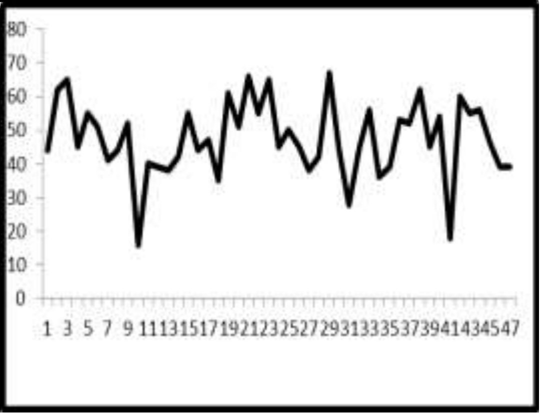


Figure 2: Trend, Seasonal, cyclical and Irregular components

By observing the trend, seasonal, cyclical and irregular patterns of the response variable corresponding accuracy values are obtained by calculating the mean absolute error (MAE) in the validation set. The mean absolute error is given by

$$\boxed{MAE = \sum_{t=1}^{n} \frac{|Y_t - F_t|}{n}}$$

Now the forecasting model after observing and calculating the mean absolute error we adopt ARIMA model as the time series is non-stationary. ARIMA has the following three components and is represented as ARIMA (p, d, q): Auto-regressive component with *p* lags AR(*p*), Integration component (*d*), Moving average with *q* lags, MA(*q*). Since the method we followed involved two models for generating the toxic alert will consume more time. So, we try to perform the forecasting computation within the device itself that is it will be embedded in the fog node.

CONCLUSION AND FUTURE SCOPE:

In this paper, with the public Beijing data I tried to identify the best model based on the ROC, precision, RME and F value. By identifying the algorithm that gave better values, I identified the bagging random forest model for performing the classification. With the classified output value, wanted to forecast the toxic level. By adopting the ARIMA model, forecasting will be done. However better result of the forecasting can be obtained by keeping the computation of ARIMA model under Fog node. Thus, the objective of generating hybrid machine learning model for toxic alert is successfully explained.

Verification of different models and adopting till random forest is successfully done using R, however checking the latency with fog is left as future scope.

REFERENCES

1. V.M. Niharika and P.S.Rao , “ A survey on air quality forecasting techniques”, International Journal of Computer Science and Information Technologies,vol.5,no.1,pp.103-107,2014.
2. Rajesh R.K and Shijimol V.R. “ Vehicular pollution monitoring and controlling using fog computing and clustering algorithm”, International journal of new innovations in Engineering and Technology, volume.4, issue 3,2016.

3. Cisco Whitepaper. “ Fog Computing and Internet of things:- Extend the cloud to where the things are”, 2015.
4. Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. Environ. Model. Softw. 2001, 16, 263–272.
5. S.Y.Muhammad,M.Makhtar,A.Rozaimee,A.Abdul and A.A.Jamal, “ Classification model for air quality using machine learning techniques,” International journal of software engineering and business applications, pp 45-52,2015.
6. Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO Symposium on Information Technologies in Environmental Engineering (ITEE’2003), Gdansk, Poland, 24–27 June 2003.
7. Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl.2010, 37, 7986–7992.
8. H.zhao, J.Zhang,K.Wang, et al., “ A GA-ANN model for air quality predicting,” IEEE,Taiwan,10 Jan. 2011.
9. Jiang, D.; Zhang, Y.; Hu, X.; Zeng, Y.; Tan, J.; Shao, D. “Progress in developing an ANN model for air pollution index forecast” 2004,38,7055-7064.