Real Time Air Quality Evaluation Model using Machine Learning Approach

Article *in* Journal of Information Technology and Digital World · May 2022

DOI: 10.36548/jitdw.2022.1.003

Real Time Air Quality Evaluation Model using Machine Learning Approach

G. Arun[1], S. Rathi[2]

[1]PG Scholar, Dept of CSE, Government College of Technology, Anna University, Chennai, India
[2]Professor, Dept of CSE, Government College of Technology, Anna University, Chennai, India

E-mail: [1]arun.60865@gct.ac.in, [2]rathi@gct.ac.in

Abstract

In recent years, the world is being industrialized day-by-day which ultimately compels our concentration towards air quality. A gradual increase in population along with the raise in usage of vehicles and consumption of conventional energy leads to air pollution which subsequently accelerates the deterioration of air quality. And air pollution has its severe impact on human health. Many researchers have proposed various methodologies for predicting and forecasting the air quality. But it is rather important to predict the future air quality in order to reduce its impact. Therefore, this paper proposes an air quality evaluation system for future prediction. The current experiment includes three modules namely Preparation of Data, Forecasting AQI and Evaluating Air Quality. Data preparation is collecting real time data and formatting it as an input to next module. Sparse Spectrum GPR (SSGPR) is used in this study to forecast, whereas cloud model to evaluate air quality. The proposed model is capable of modelling the fuzziness and randomness. Finally, the entire model is evaluated using performance metrics like MAE, RSME and MAPE.

Keywords: Air Quality, Forecasting, SSGPR, Fuzziness, Randomness

1. Introduction

In current industrial world, there is rapid growth in population which has become a great challenge. The raise in human population leads to increase in usage of combustion fuels, vehicles and deforestation where the environment gets polluted. The ultimate impact of this will be the air pollution which has its adverse effects on the human health. It causes heart diseases, respiratory illness and lung cancer. These affects of air pollution bring the necessity of predicting the air quality that too future air quality in order to prevent its impact. So, this

DOI: https://doi.org/10.36548/jitdw.2022.1.003
Received: 10.04.2022, received in revised form: 27.04.2022, accepted: 14.05.2022, published: 25.05.2022

study proposes an air quality evaluation system to evaluate the quality of air by calculating AQI and also to forecast real time AQI.

1.1 AQI

The Air Quality Index (AQI) is a metric to express the quality of air. It's a measure that how pollution impacts human health in a short time. Its objective is to keep individuals aware of the impact of air quality. In India, EPA calculates the AQI based on the concentrations of main air pollutants in order to protect public health. Understanding the AQI is critical because it collects critical information regarding the status of air in a given point and how it may affect public health.

1.2 Calculating AQI

AQI can be calculated by averaging the concentration of major air pollutants. The concentration of all the pollutants is to be recorded for a prescribed time that to be averaged to get sub-index of each pollutant. The maximum of all the sub-indices is considered as the AQI of that location. Generally five major air pollutants PM2.5, PM10, No2, O3 and So2 are considered while calculating AQI. Among which PM2.5, PM10 and No2 are to be observed for 24 hours where as O3 and Co are to be observed for minimum 8 hours.

1.3 Necessary Conditions

It is necessary that the concentration of minimum three air pollutants to be recorded among which either pm2.5 or pm10 must exist. One more condition is that the concentration of pollutants must be recorded for minimum 16 hours to calculate AQI. The calculation of sub-index makes use of a pre-defined formula,

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} (C_p - BP_{Lo}) + I_{Lo}$$

where I = the sub-index for pollutant p

C_p = the concentration of pollutant p

BP_{Hi} = break point >= C_p

BP_{Lo} = break point <= C_p

I_{Hi} = corresponding AQI value of BP_{Hi}

I_{Lo} = corresponding AQI value of BP_{Lo}

Once the Sub-indices of all the pollutants are determined, the maximum of all will be the AQI. AQI is categorised into six classes based on its range.

Table 1. Break Points for AQI scale

AQI category(Range)	PM (24 hrs)	PM (24 hrs)	NO (24 hrs)	O (24 hrs)	CO(8hrs mg/m$^{3)}$
Good (0-50)	0-30	0-50	0-40	0-50	0-1.0
Satisfactory (51-100)	31-60	51-100	41-80	51-100	1.1-2.0
Moderately Polluted (101 – 200)	61-90	101-250	81-180	101-168	2.1-10
Poor (201 – 300)	91-120	251-350	181-280	169-208	10-17
Very Poor (301 – 400)	121-250	351-430	281-400	209-748	17-34
Severe (401 – 500)	250+	430+	400+	748+	34+

Table 2. Classes of AQI

Range	**Category**
0-50	Good
51-100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
301-500	Hazardous

The health impact of the air quality depends on the AQI at that location.

Table 3. Health Impact of AQI

AQI category(Range)	Health Impact
Good (0-50)	Minimal impact
Satisfactory(51-100)	Minor breathing discomfort to sensitive people
Moderately polluted (101 – 200)	Breathing discomfort to the people with lung and heart diseases
Poor (201 – 300)	Breathing discomfort to most people on prolonged exposure

VeryPoor(301– 400)	Respiratory illness on prolonged exposure
Severe (401 – 500)	Affects healthy people and seriously impacts those with existing diseases.

1.5 Machine Learning

In this study, machine learning approaches are opted as ML is emerging domain which is efficient to build various applications, pattern identification, etc., that helps more for the future prediction.

1.6 Data Preparation

Here, sensors are used for collecting the real time concentration of air pollutants. Arduino platform is used for programming the sensors. Based on the collected data samples AQI is calculated. All the data along with AQI are stored in thingspeak channel which is cloud storage.

1.7 Forecasting

In this study, forecasting refers the prediction of future AQI. A non-parametric Bayesian and probabilistic method named Sparse Spectrum GPR (SSGPR) is opted in this experiment. This model is trained to update its features when real-time observations occur which is said to be time continuity. This method is adopted as its computational complexity is lower than its original version GPR. It is also capable of dealing with uncertainties that too in real time. Therefore, SSGPR is chosen for forecasting the air quality in an efficient way. The air quality evaluation module includes the cloud model which is capable of quantifying the randomness and fuzziness effectively. Finally the entire model will be evaluated using the performance metrics MAE, RMSE and MAPE.

2. Related Work

A hybrid air quality warning system was established in this paper [1], which included forecasting along with evaluation. First, a hybrid forecasting model on the basis of theory "decomposition and ensemble" and integrated with the sophisticated data pre-processing approach, support vector machine is proposed as an important aspect of this system. Following that, fuzzy evaluation was used to further the research, which is also critical in the warning system. The fuzzy evaluation approach and the forecasting model are complementary. Experiments show that the created approach is far superior in terms of both accuracy and effectiveness for evaluating air quality. Furthermore, the use of forecasting as well as evaluation allows for the accurate and informative estimation of future air quality, which is a considerable benefit.

This research [2] developed a dynamic evaluation model on the basis of fuzzy synthetic evaluation approach with the objective of immediately mastering future air quality. The least square SVM, which forecasts six air pollutants concentrations, has been enhanced using a newly developed computational intelligence optimization approach. The fuzzy synthetic evaluation model which is based on the entropy weights is used to create knowledge about future air quality condition. The findings and analysis on air quality reveal that accurate and reliable forecasting of urban air pollution concentrations and air quality conditions may be objectively assessed. It is demonstrated that the suggested dynamic assessment model can give a viable tool for ambient air ambient through the simulation design.

This work [3] proposed a unique analysis–forecast system that incorporates complexity analysis, data pre-processing, and optimize–forecast modules. The proposed system performs a complexity analysis of the original series using a modified least squares support vector machine optimised by a multi-objective multi-verse optimization algorithm, and then forecasts hourly AQI series using a modified least squares support vector machine optimised by a multi-objective multi-verse optimization algorithm. Experiments using datasets from eight major Chinese cities revealed that the proposed system is capable of achieving high accuracy and stability at the same time, making it efficient and reliable for air quality monitoring.

This work [4] introduced a new hybrid model that combines a method for detecting and correcting outliers with a heuristic intelligent optimization strategy. First, data pre-processing algorithms are used to detect and correct outliers, as well as to uncover the main characteristics of the original time series; second, a widely used heuristic intelligent optimization algorithm is used to optimise the parameters of the extreme learning machine, resulting in more accurate forecasting results for each subseries; and finally, experimental results and analysis show that the presented hybrid model provides accurate prediction, outlier detection, and outlier eradication.

The study [5] offered a new hybrid air quality early-warning system that consists of three modules: data pre-processing, forecasting, and air quality evaluation. The purpose of a new hybrid data pre-processing technique is to extract chaotic features from raw data in order to produce a more stable series of pollution data for forecasting. A multi-objective grasshopper optimization technique is then used to increase the forecasting module's accuracy and stability. A fuzzy air quality evaluation module is also included to supply the system with thorough results. Not only does the forecasting approach achieve more accuracy and stability than previous comparison models, but the evaluation module also provides acceptable air quality data, according to the findings and comments.

The study [6] proposed a Deep Multi-output LSTM (DM-LSTM) neural network model with three deep learning algorithms (mini-batch gradient descent, dropout neuron, and L2 regularisation) for extracting the key factors of complex spatiotemporal relations and reducing error accumulation and propagation in multi-step-ahead air quality forecasting. To evaluate the proposed DM-LSTM model, three time series of PM2.5, PM10, and NOx were analysed simultaneously at five air quality monitoring stations in Taipei City, Taiwan. When integrated with three deep learning algorithms, the proposed DM-LSTM model considerably enhanced the spatio-temporal stability and accuracy of regional multi-step-ahead air quality forecasts.

The work [7] presented a new sparse Gaussian Process (GP) regression model. The major novel idea is to sparsify the spectral representation of the GP. As a result, a straightforward and practical method to regression problems has been created. It looks at the trade-offs that can be made between prediction accuracy and computing requirements, and shows that they are usually better than current state-of-the-art sparse approximations. According to this work, which examines both the weight space and function space representations, the new design implies priors over functions that are always stationary and can approximate any covariance function in this class.

3. Proposed Work

The proposed system includes three modules,

- Data Preparation
- Forecasting
- Air Quality Evaluation

3.1 Data Preparation

This module includes the collection of real time data using sensors and real time AQI calculation.

3.1.1 Collection of Data

The first and foremost task of this experiment is to organise the sensors. In this study an individual sensor is used for each air pollutant along GPRS in order to obtain location. The Arduino MKR1000 board is used here aiming to build efficient IOT projects as it comes with on on-board wi-fi feature.

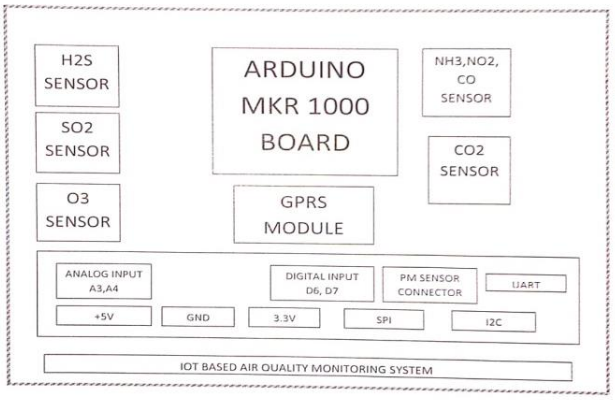

Figure 1. Block diagram of IoT Hardware

Once all the IoT hardware gets organised, it is connected to the arduino web editor in order to program the sensor. To activate the sensors the entire sketch is to be written an arduino web editor and is uploaded into the MKR1000 board. Once the uploading finishes the hardware activates at sensors start reading data. Now one more major task is to store the obtaining real time data. Therefore the current study opts thingspeak platform to store the data. Thingspeak is an user friendly and cloud platform which will be an efficient storage.

Here rather than storing all the data, the concentrations of major pollutants are only considered. Major pollutants include PM2.5, PM10, So2, O3, No2, NH3, Co. Hence, only the data of these pollutants are stored to thingspeak channel. Once the real time data is collected and it is successfully updated to thingspeak the following task will be the AQI calculation.

3.1.2 AQI Calculation

AQI Calculation includes the determination of sub-index for each pollutant and finding the maximum of all. The determination of sub-index is done using the formula that is predefined. From all the obtained sub-indices the maximum is considered to be the AQI. The sketch for this entire calculation is to be implemented and integrated with the previous one. Therefore, once the complete sketch is uploaded into the arduino board, the sensors starts reading real time data where in parallel the data is updated to channel along with the real time

AQI calculated i.e., The thingspeak channel stores the concentration of the major pollutants (PM2.5, PM10, So2, O3, NH3, Co) and also the AQI calculated at real time. Hence the initial data required for the prediction is prepared which will be formatted as an input to the forecasting module.

3.2 Forecasting

In this study, the forecasting module aims at both predicting the current AQI and also the future trends. This module opts the efficient real time learning model named SSGPR. SSGPR is the advanced version of GPR [8] which is probabilistic and non-parametric Bayesian SSGPR has higher computational efficiency. That is why it is selected to perform uncertainty modelling of real time data and forecasting real time AQI. The GPR model is improvised by using sparse approximation i.e., by sparsifying the power spectrum of GPR.

The main objective of SSGPR [9] is to find the number of optimal frequencies. Once the optimal frequencies are identified, the initial mean and variance of a new observation can be calculated. Based on the mean and variance the quantification of uncertainty on the new output can be performed using the interval i.e., (UB,LB). From the study it is observed that the computation cost of SSGPR to calculate updated mean and variance is $O(m)$ and $O(m^2)$ respectively. Therefore it can be noted that the SSGPR has the complexity $O(t^3)$ which is more computationality efficient over GPR.

3.3 Air Quality Evaluation

In this study, air quality evaluation includes a cloud model [10] which evaluates the air quality using forecasted values. It is capable of transforming quantitative data to qualitative data by involving decision making method. The cloud model depends on three major parameters [11] namely Ex, En, HE. All these parameters are efficient to model fuzziness [12] as well as randomness effectively.

- Ex – Expected value of qualitative concept.
- En – Represents the fuzziness in qualitative concept.
- He – Randomness in the qualitative concept.

Air quality evaluation includes the following steps,

1. Select the pollutants to be considered to evaluate air quality.

2. Apply cloud model for each pollutant under each level.

3. Using entropy based Super scale weighting method calculate weight for each pollutant

4. According to the weights obtained from step-3 obtain certainty degree of a pollutant to determine its level of Air Quality.

5. Based on the average of results obtained by performing step-4, final certainty degree can be determined.

6. Obtain the certainty degree of each level and by observing the certainty degree, consider the level having maximum certainty degree as the air quality level.

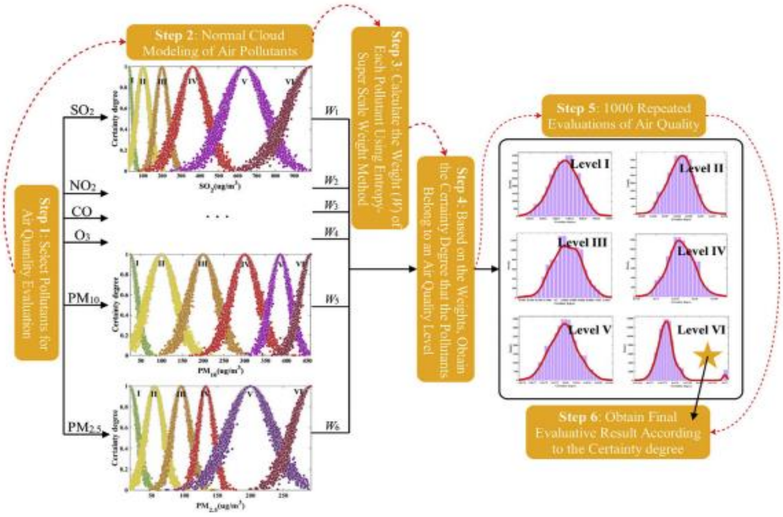
Step 1: Select Pollutants for Air Quality Evaluation
Step 2: Normal Cloud Modeling of Air Pollutants
Step 3: Calculate the Weight (W) of Each Pollutant Using Entropy-Super Scale Weight Method
Step 4: Based on the Weights, Obtain the Certainty Degree that the Pollutants Belong to an Air Quality Level
Step 5: 1000 Repeated Evaluations of Air Quality
Step 6: Obtain Final Evaluative Result According to the Certainty degree
SO2
NO2
CO
O3
PM10
PM2.5
W1
W2
W3
W4
W5
W6
I
II
III
IV
V
VI
Certainty degree
SO2(ug/m3)
PM10(ug/m3)
PM2.5(ug/m3)
Level I
Level II
Level III
Level IV
Level V
Level VI

Figure 2. Flow of Evaluating Air Quality

4. System Evaluation

The current proposed system has proposed three performance metrics namely, MAE, RMSE and MAPE to evaluate its accuracy.

4.1 Root Mean Square Error (RMSE)

RMSE is the most commonly used methods to evaluate the validity of predictions is Root mean square error can be expressed as,

$$\text{RSME} = \sqrt{\frac{\sum_{i=1}^{N} ||y(i) - \hat{y}(i)||^2}{N}}$$

where N indicates the number of data points, y(i) is i-th measurement, and $\hat{y}(\boldsymbol{i})$ is corresponding prediction.

4.2 Mean Absolute Error (MAE)

MAE is the average of the absolute errors in a group of observations as a measure of the size of the error in the entire group.

$$\text{MAE} = \frac{|(Y_I - Y_P)|}{n}$$

where yi = actual observation

yp = predicted observation

n = number of observations or rows

4.3 Mean Absolute Percentage Error

MAPE is the mean of absolute percentage errors. Error can be defined as actual or observed value minus the forecasted value.

$$\text{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

5. Conclusion

The proposed system includes the modules of Data-Preparation, Forecasting, and Air Quality Evaluation. The real time data is collected through the sensors and real time AQI is calculated based on the data obtained. Then the Sparse Spectrum GPR model which has the ability to learn real time data with uncertainty is trained and the prediction is performed. A cloud model that deals with fuzziness and randomness is used to evaluate the quality of air. Hence, it is concluded that the proposed model is capable of predicting and forecasting the future AQI despite of the uncertainty, fuzziness and randomness of the real time data. The performance of the model is analysed using the metrics like MAE, RMSE, and MAPE. The future research may concentrate on incorporating the data normalization techniques in order to predict the AQI more accurately.

References

[1] Xu, Y., Yang, W., Wang, J., 2017a, “Air quality early-warning system for cities in China”, Atmos. Environ.148, 239-257.

[2] Li, R., Dong, Y., Zhu, Z., Li, C., Yang, H., 2019a , “A dynamic evaluation framework for ambient air pollution monitoring”, Appl. Math. Model. 65, 52-71.

[3] Li, H., Wang, J., Li, R., Lu, H., 2019b, “Novel analysis e-forecast system based on multi objective optimization for air quality index”, J. Clean. Prod. 208, 1365-1383.

[4] Wang, J., Du, P., Hao, Y., Ma, X., Niu, T., Yang, W., 2020a, “An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting”, J. Environ. Manag. 255, 109-855.

[5] Wang, Y., Wang, J. and Li, Z., 2020,”A novel hybrid air quality early-warning system based on phase-space reconstruction and multi-objective optimization: A case study in China”, Journal of Cleaner Production, 260, p.121027.

[6] Zhou, Y., Chang, F., Chang, L., Kao, I., Wang, Y., 2019,”Explore a deep learning multioutput neural network for regional multi-step-ahead air quality forecasts”, J. Clean. Prod. 209, 134-145.

[7] Lazaro-Gredilla, M., Qui nonero-Candela, J., Rasmussen, C.E., Figueiras-Vidal, A.R., ~ 2010, “Sparse spectrum Gaussian process regression”, J. Mach. Learn. Res. 11, 1865-1881.

[8] Rasmussen, C.E. and Williams, C.K., 2006, “Gaussian processes for machine learning”, vol. 1.

[9] Yang, A., Li, C., Rana, S., Gupta, S., Venkatesh, S., 2010, “Sparse spectrum gaussian process regression”, J. Mach. Learn. Res. 11, 1865-1881.

[10] Li, D., Liu, C., Gan, W., 2009,”A new cognitive model: cloud model”, Int. J. Intell. Syst. 24, 357-375.

[11] Zhang, J.G. and Singh, V.P., 2012, “Entropy-Theory and Application”, China Water & Power Press: Beijing, China, pp.79-80.

[12] Olvera-García, M.A., Carbajal-Hern andez, J.J., S anchez-Fern andez, L.P., Hern andez- Bautista, I., 2016, “Air quality assessment using a weighted Fuzzy Inference System”, Ecol. Inf. 33, 57-74.