IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLASSIFICATION OF AIR POLLUTION LEVELS USING SUPERVISED MACHINE LEARNING ALGORITHM

Ardra Nandakumar,Reesha P.U
MSc Scholar,Assistant Professor
Department of Computer Science,
St.Joseph's College (Autonomous),Irinjalakuda,Thrissur,India

Abstract: Deterioration of air quality is one of the major issues faced by various cities in our country. The meteorological factors,burning of fossil fuels,emissions from power plants etc. interrupts the flow of fresh air and results in the spread of Sulphur dioxide, Carbon monoxide, Nitrogen dioxide, Particular pollutants which is harmfull for all living organisms. So it is necessary to understand the effectiveness of air quality monitoring programs for planning air pollution control strategies by analyzing the trends in air quality regularly. In this study, the varying trends of ambient air quality were analyzed and predicted in terms of Air Quality Index (AQI) based on the database of dangerous pollutants like NO_2 and RSPM that are monitored at different monitoring stations in Thrissur District, Kerala, India. Here, we uses SARIMA (Seasonal Autoregressive Integrated Moving Average)model to predict the future month-wise concentration of pollutants and uses SVM (Support Vector Machine) algorithm to classify the pollution levels into different categories. This knowledge on air pollution levels of future months will help the public to take measures to reduce its level to a lesser than the harmful range.

Index Terms - AQI, SARIMA MODEL, SVM.

I. INTRODUCTION

Air pollution can be considered as one of the major hazards among the environmental pollution. As each living organisms needs both fresh and good quality air for every second, none of the living organisms can survive without such air. Inhaling pollutants for long time causes irreversible damages in human health such as skin and eyes infections,irritations,heart diseases,bronchitis,asthma,lung cancer etc. Air pollution causes bad effects not only on public health but also on the environment such as acid rains,photochemical smog,ozone layer deterioration and global warming. So,an air quality forecasting system is essential to prevent this problem to an extend.In this work,the proposed system is to develop the air quality forecasting system to predict the AQI and to categorise into different levels.

Air Quality Index (AQI) is a standard that is used to show how polluted the air currently is or how polluted it is forecast to become. It is basically used to provide information about the health risks to the public as well as the concerned authorities as AQI increases. The levels of health concern can be identified as good,moderate,bad etc. Our study aims at providing this health concern levels based on the presence of RSPM and No2 in the air. The prediction of pollutant concentration of future month is done using SARIMA,one of the popular Time series analysis model.Then AQI for each pollutant is calculated.Based on these,the pollution levels will classify using Support Vector Machine,a supervised machine learning algorithm.

Machine learning can be classified as supervised learning and unsupervised learning.Supervised learning provides a function that maps from input to output,based on training examples. Where training examples consists of labelled training data.Supervised learning consists of classification and regression.Random forest,Support Vector Machines etc are some of the most common algorithms of supervised learning. Unsupervised learning make use of unlabeled data to draw some patterns or inferences.It includes clustering and association. K-means algorithm is the most commonly used for clustering and Apriori algorithm for association.

In the proposed work, Support Vector Machine (SVM) algorithm is used for classification.

II. LITERATURE SURVEY

SI.NO	PAPER TITLE	AUTHOR	OBJECTIVE	METHODOLOGY	CONCLUSION	RESULT
1	MODELING PM2.5 URBAN POLLUTION USING MACHINE LEARNING AND SELECTED METEOROLOGICAL PARAMETERS	JAN KLEINE DETERS…	TO PREDICT THE PM2.5 ON THE BASIS OF SELECTED METERIOLOGICAL PARAMETERS	DECISION TREE AND SVM	PROVIDES AN INSIGHT INTO THE MAIN LIMITATIONS REGARDING PM2.5 PREDICTION FROM METEOROLOGICAL DATA AND MACHINE LEARNING	ACCURACY OF NEARLY 89% BY DECISION TREES
2	A MACHINE LEARNING APPROACH TO PREDICT AIR QUALITY IN CALIFORNIA	FABIANA MARTINS …	TO PREDICT HOURLY POLLUTANT CONCENTRATIONS LIKE CO_2,SO_2,NO_2,GROUND LEVEL OZONE AND PARTICULATE MATTER AS WELL AS THE HOURLY AQI FOR THE STATE OF CALIFORNIA	SUPPORT VECTOR REGRESSION	THE MODEL WAS ABLE TO CLASSIFY THE AQI INTO 6 CATEGORIES	ACCURACY OF 94.1% ON UNSEEN VALIDATION DATA
3	PREDICTION OF AIR QUALITY INDEX USING HYBRID MACHINE LEARNING ALGORITHM	JASLEEN KAUR SETHI..	TO PREDICT THE AIR QUALITY INDEX	A HYBRID MACHINE LEARNING ALGORITHM	THE HYBRID ALGORITHM PREDICTION PERFORMANCE IS BETTER THAN SVM ALGORITHM	OBSERVED AN ACCURACY OF 91.25%
4	INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING	A.GNANA SOUNDARI…	TO FORECAST THE AIR QUALITY OF INDIA USING MACHINE LEARNING TO PREDICT THE AQI OF A GIVEN AREA	GRADIENT BOOSTED MULTIVARIABLE REGRESSION	SINCE THE MODEL IS CAPABLE OF PREDICTING THE CURRENT DATA WITH 95% ACCURACY IT WILL SUCCESSFULLY PREDICT THE UPCOMING AQI OF ANY PARTICULAR	ACCURACY OF 95%

					DATA WITHIN A GIVEN REGION	
5	AIR POLLUTION PREDICTION IN SMART CITY,DEEP LEARNING APPROACH	ABDELLATIF BEKKAR..	TO PREDICT THE PM2.5 OF AIR POLLUTANTS IN THE URBAN AREA OF BEIJING	HYBRID MODEL BASED ON CNN AND LSTM	MODEL CAN EFFECTIVELY EXTRACT THE TEMPORAL AND SPATIAL FEATURES OF THE DATA THROUGH CNN AND LSTM AND IT ALSO HAS HIGH ACCURACY AND STABILITY	MEAN ABSOLUTE ERROR-6.742,ROOT MEAN SQUARE ERROR-12.921,R^2-0.989
6	FORECASTING AIR POLLUTION PARTICULATE MATTER(PM$_{2.5}$) USING MACHINE LEARNING REGRESSION MODELS	DORESWAMY…	PREDICT PM$_{2.5}$ USING MACHINE LEARNING MODELS BASED ON THE STATISTICAL CALCULATIONS OF METRICS SUCH AS MAE,MSE,RMSE AND R^2 ,TO ANALYZE THE AIR POLLUTION ON TAQMN DATA IN TAIWAN	GRADIENT BOOSTING REGRESSION	PROPOSED MODELS VALUES PERFORM BETTER COMPARES TO THE PREVIOUS MODELS , ALSO IT SHOW THAT THE ACTUAL VALUES AND PREDICTED VALUES ARE VERY CLOSE TO EACH OTHER	ROOT MEAN SQUARE ERROR-0.1302,R^2-0.9336
7	APPLICATION OF IMPROVED KNN ALGORITHM IN AIR QUALITY ASSESSMENT	YANG RUI-JUN…	TO SOLVE THE PROBLEM OF LOW ACCURACY AND LOW EFFICIENCY IN AIR QUALITY DETECTION	IMPROVED KNN ALGORITHM	THE IMPROVED KNN ALGORITHM HAS AN AIR QUALITY EVALUATION WHICH IS SUPERIOR TO THE TRADITIONAL ALGORITHM	ACCURACY OF 94.53%.

III. PROPOSED SYSTEM

The proposed system entitled as "Classification of Air Pollution levels using Supervised Machine Learning Algorithm" aimed to predict AQI of future month wise and classify them in different category. Here we collect seven-year monthly concentration of NO_2 and RSPM from Pollution Control board Thrissur. It consists of monthly minimum, average and maximum value air pollutant on that month. And calculate AQI of each pollutant and assign AQI as largest AQI value.Train these data and predict future value. Then categorise AQI to different class using Support Vector Machine algorithm. This help common people to know about air quality status.

IV. METHODOLOGY

The proposed system for classification of the air pollution levels consists of two sections. One for predicting the future monthwise values for NO_2 ,RSPM and other for the calculation of AQI and classification of pollution levels.

❖ Following are the steps involved in predicting future values:

Data Collection

Data collection is the first step of the Machine learning process. Without data,we couldn't do anything as it is the major fuel.For our proposed system,the data is collected from the directory of Pollution Control Board (PCB),Thrissur. PCB collect this data daily using a hardware device in different region. We collect monthly average of this data.Different air pollutants contribute different AQI. Highest AQI value in considered as AQI in that region. In this project we collect measure of NO_2 and RSPM monthly average of 7 years. The data contains 6 attributes like RSPM, minimum RSPM ,maximum RSPM,NO_2,minimum NO_2,maximum NO_2. We have to train and predict month wise NO_2 and RSPM based on these data and calculate Air Quality Index.

Data Preprocessing

The dataset may contain inconsistent data,missing values and repeated data. So we have to clean the dataset ,missing values have to be either delete or have to fill with mean values or some other method to get proper prediction result. Also repeated data must be removed or eliminated so as to avoid biasing of the results. Some dataset may have some extreme values which also have to be removed to get good prediction accuracy. Classification algorithms will work well only if all this preprocessing is done on the data.

Parameter Selection for the Time Series Model

Seasonal Autoregressive Integrated Moving Average,SARIMA or Seasonal ARIMA, is an extension of ARIMA that supports univariate time series data with a seasonal component. In our proposed system,SARIMA is used to forecast the future monthwise measure of air pollutants. Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements: There are 3 trend elements that require configuration. They are same as the ARIMA model;specifically:

- **P**: Trend autoregression order.
- **d:** Trend difference order.
- **q:** Trend moving average order.

Seasonal Elements: There are 4 seasonal elements apart from the trend elements that are not part of ARIMA that must be configured;they are:

- **P:** Seasonal autoregressive order.
- **D:** Seasonal difference order.
- **Q:** Seasonal moving average order.
- **m:** Number of time steps for one seasonal period.

Together, the notation for an SARIMA model can be written as:

SARIMA(p,d,q)(P,D,Q)m

Inorder to train the model,define the p,d and q parameters to take any value between 0 and 2. Then generate all different combinations of p,d ,q parameters and seasonal p,d,q parameters. When evaluating and comparing models fitted with different parameters,each can be ranked against one another based on how well it fits the data or how well it can accurately predict future data points. Here we use the AIC (Akaike Information Criterion) value to measure how well a model fits the data while taking into account the overall complexity of the model. A model that fits the data very well while using lots of features will be given a larger AIC score than a model that uses fewer features to get the same goodness-of-fit. Therefore we have to find the model that gives the lowest AIC value.

Fitting the Time Series Model

We have identified the set of parameters that produces the best fitting model to our time series data. Now we have to proceed to analyze this particular model in more depth. We have to plug this optimal parameter values into a new SARIMAX model. Then fit the model using fit(). The summary attribute that results from the output of SARIMAX() will return some significant amount of information.

Validating Forecasts

The model we created can be used to produce forecasts.In this phase,the series,which will help us understand the accuracy of our forecasts. The get_prediction() and conf_int() attributes allow us to obtain the values and its confidence intervals for the forecasts of the time series.

Producing Forecasts

The get_prediction() is used to producing the forecasts of future monthwise measure of air pollutants.

❖ Following are the steps for the construction of classification model:

Data Collection

Data needed for AQI training is collected from the site of Central Pollution Control Board.This contains attributes like Average RSPM,Minimum RSPM,Maximum RSPM,Average NO_2,Minimum NO_2 ,Maximum NO_2.

Data Preprocessing

In pre-processing, the missing values must be eliminated from the dataset to ensure that the results generated are more accurate.

Model Training

A model is trained to calculate AQI (Air Quality Index) of both RSPM and NO_2. The highest AQI Value will be considered as the final AQI. This final AQI value will be trained by the SVM (Support Vector Machine) algorithm.

Support Vector Machine Algorithm: Support Vector Machine (SVM) is one of the Supervised Machine Learning algorithms which is commonly used for classification problems. The main aim of SVM algorithm is to find a hyperplane in an N-dimensional space that classifies the data points. The dimension of hyperplane depends on the number of points. If there is only two points,the hyperplane will be a line. If the number of points is three, then the hyperplane will be a two dimensional plane. When the number of points increases,it becomes difficult to imagine the hyperplane.

V. ACKNOWLEDGEMENTS

I am extremely thankful to Principal Dr.Sr. Asha Therese, Head of the Department Sr. Siji P.D for giving me constant support and guidance for preparing and presenting the paper. I express my sincere gratitude towards my guide, Ms Reesha P.U for her help, encouragement and inspiration during the preparation of this paper.

VI. CONCLUSION

Air quality prediction is one of the difficult tasks due to its changing nature, volatility, and variability in the time and space of pollutants. If it is being able to model and predict the quality of air,it can cause great impact on the health of the citizens and environment. People must know what the level of pollution in their surroundings and to plan strategies to fight against it. The proposed system will help common people as well as the concerned authorities to detect and predict pollution levels and take the necessary action against that.

REFERENCES

[1] Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, Yves Rybarczyk, “Modelling $PM_{2.5}$ Urban Pollution Using Machine Learning and Selected Meteorological Parameters “,Journal of Electrical and Computer Engineering ,vol. 2017,Article ID 5106045,14 pages,2017.

[2] Mauro Castelli, Fabiana Martins Clemente, Ales Popovic, Sara Silva, Leonardo Vanneschi, “A Machine Learning Approach to Predict Air Quality in California”, Complexity,vol 2020 , Article ID 8049504, 23 pages, 2020

[3] Jasleen Kaur Sethi,Mamta Mittal **,”** Prediction of Air Quality Index Using Hybrid Machine Learning Algorithm”,January 2021

[4] A. Gnana Soundari,J. Gnana Jeslin M.E,Akshaya A.C, “INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue) © Research India Publications.

[5] Bekkar, A., Hssina, B., Douzi, S. *et al.* Air-pollution prediction in smart city, deep learning approach. *J Big Data* **8,** 161 (2021)

[6] Doreswamy, Harishkumar, K.S.,Ibrahim Gad., Yogesh, K.M.,”Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models” Volume 2020, Pages 2057-2066.

[7] Yang Rui-jun,Ding Dan-feng,Ding Dan-feng, “Application of Improved KNN Algorithm in Air Quality Assessment”,HPCCT 2019: Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference,june 2019 Pages 108–112

[8] https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3

[9] Aditya CR, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu; "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trend and Technology (IJETT) Volume 59 Issue 4-May 2018