e-ISSN: 2319-8753 | p-ISSN: 2320-6710

IJIRSET

International Journal of Innovative Research in SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 6381 907 438 ijirset@gmail.com www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

Prediction of Air Pollution Using Machine Learning

CH. Muralikirshna[1], M.SivaPriya[2], V Ranjith Kumar[3], R Sai Kumar[4], T Sankeerth[5]

Assistance Professor, Department of Electronics and Communication Engineering Siddharth Institute of Engineering and Technology, Chittoor, India [1]

Department of Electronics and Communication Engineering, Siddharth Institute of Engineering and Technology , Chittoor , India[2,3,4,5,6]

ABSTRACT: The regulation of air pollutant levels is rapidly increasing and its one of the most important tasks for the governments of developing countries, especially India. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter (PM) that determine the quality of the air. When its level is high in the air, it causes serious issues on people's health. Hence, controlling it by constantly keeping a check on its level in the air is important.

KEYWORDS: Noted that the application of prediction of air pollution leads to further improvement to the world.

I. INTRODUCTION

Particulate matter can be either human-made or naturally occur. Some examples include dust, ash and sea-spray. Particulate matter (including soot) is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines. Particulate matter varies in size (i.e. the diameter or width of the particle). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers (µm). PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Different machine learning models have been applied to detect air pollution and predict PM2.5 levels based on a data set consisting of daily atmospheric conditions. Naive Bayes classification and support vector machine algorithms were applied by Dan Wei to get the minimum error with respect to prediction of the air quality in Beijing city. A fuzzy inference system was introduced by José Juan Carbajal et al. to perform parameter classification using a reasoning process and integrating them into an air quality index.

II. RELATED WORK

1.JoséJuanCarbajal-Hernándezab Luis P.Sánchez-FernándezaJesúsA.Carrasco-OchoabJoséFco.Martínez-Trinidadb: AssessmenDixian Zhu, ChangjieCai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization.

In this paper, we tackle air quality forecasting by using machine learning approaches to predict the hourly concentration of air pollutants (e.g., ozone, particle matter (PM2.5) and sulfur dioxide). Machine learning, as one of the most popular techniques, is able to efficiently train a model on big data by using large-scale optimization algorithms. Although there exist some works applying machine learning to air quality prediction, most of the prior studies are restricted to several-year data and simply train standard regression models (linear or nonlinear) to predict the hourly air pollution concentration. In this work, we propose refined models to predict the hourly air pollution concentration on the basis of meteorological data of previous days by formulating the prediction over 24 h as a multi-task learning (MTL) problem. This enables us to select a good model with different regularization techniques. We propose a useful regularization by enforcing the prediction models of consecutive hours to be close to each other and compare it with several typical regularizations for MTL, including standard Frobenius norm regularization, nuclear norm regularization, and $\ell_{2,1}$-norm regularization. Our experiments have showed that the proposed parameter-reducing formulations and consecutive-hour-related regularizations achieve better performance than existing standard regression models and existing regularizations.

2.Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai : An Empirical Study of PM2.5 Forecasting Using neural network.

In the recent years, a lot of efforts have been made to regulate air pollutant levels in most of the developed and developing countries. Fine particulate matter (PM2.5) is considered to be one of the major reasons behind deteriorating public health and a lot of efforts are being made to keep a check on PM2.5 levels. Accurately forecasting PM2.5 level is a challenging task and has been highly dependent on model based approaches. In this paper, we explore new possibilities to hourly forecast PM2.5. Choosing the right forecasting model becomes a very important aspect when it comes to improvement in prediction accuracy. We used Neural Network Autoregression (NNAR) method for the prediction task. The paper also provides a comparative analysis of prediction performance for additive version of Holt-Winters method, autoregressive integrated moving average (ARIMA) model and NNAR model. The experimentation and evaluation is done using real world measurement data from Airbox Project, which shows that our proposed method accurately does the prediction with significantly low error.

Dan wei: Predicting air pollution level in a specific city

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries, especially China. Among the pollutant index, Fine particulate matter (PM2.5) is a significant one because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. However, the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, some of these advanced techniques have been introduced into air quality research. These studies utilized selected techniques, such as Support Vector Machine (SVM) and Neural Network, to predict ambient air pollutant levels based on mostly weather and sometimes traffic variables. This project attempted to apply some machine learning techniques to predict PM2.5 levels based on a dataset consisting of daily weather and traffic parameters in Beijing, China. Due to the uncertainty of the specific number PM2.5 level, I simplified the problem to be a binary classification one, that is to classify the PM2.5 level into "High" (> 115 ug/m3) and "low" (<= 115 ug/m3). The value is chosen based on the Air Quality Level standard in China, which set 115 ug/m3 to be mild level pollution.

Pandey, Gaurav, Bin Zhang, and Le Jian. " Predictng sub-micron air pollution indicators: a machine learning approach.

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries, especially China. Submicron particles, such as ultrafine particles (UFP, aerodynamic diameter ≤ 100 nm) and particulate matter ≤ 1.0 micrometers (PM1.0), are an unregulated emerging health threat to humans, but the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, we employed a range of machine learning techniques to predict UFP and PM1.0 levels based on a dataset consisting of observations of weather and traffic variables recorded at a busy roadside in Hangzhou, China. Based upon the thorough examination of over twenty five classifiers used for this task, we find that it is possible to predict PM1.0 and UFP levels reasonably accurately and that tree-based classification models (Alternating Decision Tree and Random Forests) perform the best for both these particles. In addition, weather variables show a stronger relationship with PM1.0 and UFP levels, and thus cannot be ignored for predicting submicron particle levels. Overall, this study has demonstrated the potential application value of systematically ct and prediction of air quality using fuzzy logic and autoregressive models:

In recent years, artificial intelligence methods have been used for the treatment of environmental problems. This work, presents two models for assessment and prediction of air quality. First, we develop a new computational model for air quality assessment in order to evaluate toxic compounds that can harm sensitive people in urban areas, affecting their normal activities. In this model we propose to use a Sigma operator to statistically asse air quality parameters using their historical data information and determining their negative impact in air quality based on toxicity limits, frequency average and deviations of toxicological tests. We also introduce a fuzzy inference system to perform parameter classification using a reasoning process and integrating them in an air quality index describing the pollution levels in five stages: *excellent, good, regular, bad* and *danger*, respectively. The second model proposed in this work predicts air quality concentrations using an autoregressive model, providing a predicted air quality index based on the fuzzy inference system previously developed. Using data from Mexico City Atmospheric Monitoring System, we perform a comparison among air quality indices developed for environmental agencies and similar models. Our results show that our models are an appropriate tool for assessing site pollution and for providing guidance to improve contingency actions in urban areas.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries, especially China. Submicron particles, such as ultrafine particles (UFP, aerodynamic diameter ≤ 100 nm) and particulate matter ≤ 1.0 micrometers (PM1.0), are an unregulated emerging health threat to humans, but the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, we employed a range of machine learning techniques to predict UFP and PM1.0 levels based on a dataset consisting of observations of weather and traffic variables recorded at a busy roadside in Hangzhou, China. Based upon the thorough examination of over twenty five classifiers used for this task, we find that it is possible to predict PM1.0 and UFP levels reasonably accurately and that tree-based classification models (Alternating Decision Tree and Random Forests) perform the best for both these particles. In addition, weather variables show a stronger relationship with PM1.0 and UFP levels, and thus cannot be ignored for predicting submicron particle levels. Overall, this study has demonstrated the potential application value of systematically ct and prediction of air quality using fuzzy logic and autoregressive models.

III. METHODOLOGY

An autoregressive (AR) model predicts future behavior based on past behavior. You only use past data to model the behavior, hence the name autoregressive.

In an AR model, the value of the outcome variable (Y) at some point t in time is like "regular" linear regression, directly related to the predictor variable (X). Where simple linear regression and AR models differ is that Y is dependent on X and previous values for Y. The AR process is an example of a stochastic process, which have degrees of uncertainty or randomness built in. The randomness means that you might be able to predict future trends pretty well with past data, but you're never going to get 100 percent accuracy. Usually, the process gets "close enough" for it to be useful in most scenarios.

AR (p) Models an AR (p) model is an autoregressive model where specific lagged values of yt are used as predictor variables. Lags are where results from one time period affect following periods. The value for "p" is called the order. For example, an AR would be a "first order autoregressive process." The outcome variable in a first order AR process at some point in time t is related only to time periods that are one period apart (i.e. the value of the variable at t – 1). A second or third order AR process would be related to data two or three periods apart.

The AR (p) model is defined by the equation:

$YT = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-1} + A_t$

Where

$Yt-1, y_{t-2} \dots y_{t-p}$ are the past series values (lags),

At is white noise (i.e. randomness),

And δ is defined by the following equation

$$\delta = \left(1 - \sum_{i=1}^{p} \phi_i\right) \mu ,$$

2) LOGISTIC REGRESSION:-

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

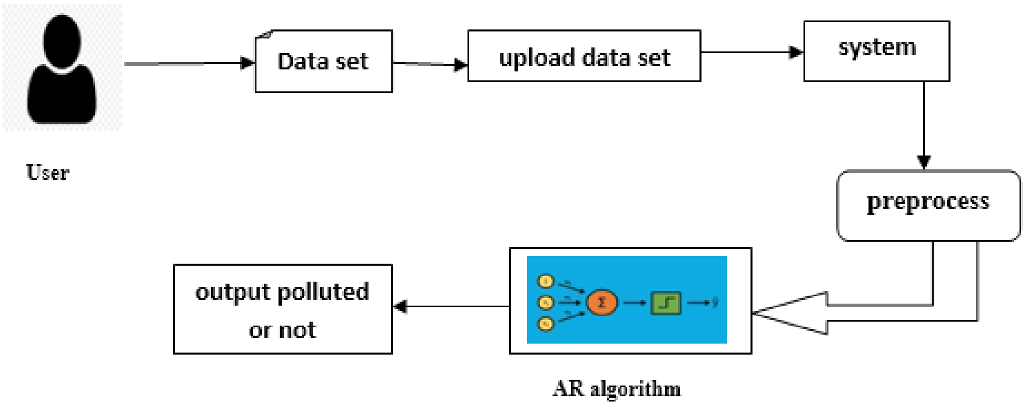
International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

BLOCK DIAGRAM



IV. SIMULATION RESULTS

5.1 RESULT AND DISCUSSION

Import libraries

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score,recall_score,classification_report
import matplotlib.pyplot as plt
```

Load dataset

```
df = pd.read_csv("C:/Users/shruti/Desktop/Air pollution dataset/city air pollution data.csv")
df.head()
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
0	Ahmedabad	01-01-2015	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00
1	Ahmedabad	02-01-2015	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77
2	Ahmedabad	03-01-2015	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25
3	Ahmedabad	04-01-2015	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00
4	Ahmedabad	05-01-2015	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

```
df.dropna(axis=0,inplace=True)
df.drop_duplicates(inplace=True)
df.shape
```

```
(6351, 14)
```

```
df.isnull().sum()
```

```
City       0
Date       0
PM2.5      0
PM10       0
NO         0
NO2        0
NOx        0
NH3        0
CO         0
SO2        0
O3         0
Benzene    0
Toluene    0
Xylene     0
dtype: int64
```

```
df['PM2.5'] = df['PM10'].apply(lambda x: x * 0.75 - 1.72 )
df.head()
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
2122	Amaravati	24-11-2017	85.0925	115.75	1.75	20.65	12.40	12.19	0.10	10.76	109.26	0.17	5.92	0.10
2123	Amaravati	25-11-2017	91.6550	124.50	1.44	20.50	12.08	10.72	0.12	15.24	127.09	0.20	6.50	0.06
2124	Amaravati	26-11-2017	95.0750	129.06	1.26	26.00	14.85	10.28	0.14	26.96	117.44	0.22	7.95	0.08
2125	Amaravati	27-11-2017	99.7700	135.32	6.60	30.85	21.77	12.91	0.11	33.59	111.81	0.29	7.63	0.12
2126	Amaravati	28-11-2017	76.3475	104.09	2.56	28.07	17.01	11.42	0.09	19.00	138.18	0.17	5.02	0.07

```
df['Pollution'] = df['PM2.5'].apply(lambda x: 1 if x>115 else 0)
df.to_csv('C:/Users/shruti/Desktop/Air pollution dataset/Airpolutifinaldataset.csv')
```

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

```
df.head(100)
df.to_csv('C:/Users/shruti/Desktop/Air pollution dataset/preprocessed.csv',index=False)
```

```
df['Pollution'].value_counts()
```

```
0    4876
1    1475
Name: Pollution, dtype: int64
```

```
df['City'].value_counts()
```

```
Hyderabad        1633
Delhi            1224
Visakhapatnam    1184
Amaravati         667
Amritsar          660
Kolkata           394
Chandigarh        279
Patna             191
Gurugram          119
Name: City, dtype: int64
```

Encoding variables

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['City']= le.fit_transform(df['City'])
```

```
le2 = LabelEncoder()
df['Date']= le.fit_transform(df['Date'])
```

```
X = df.drop(['Pollution','PM2.5'],axis=1)
y = df['Pollution']
```

Splitting dataset

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(random_state=0).fit(X_train, y_train)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:762: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007188|

Test accuracy

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,pred)
```

```
0.8892969569779643
```

Train accuracy

```
accuracy_score(y_train,clf.predict(X_train))
```

```
0.9077615298087739
```

Predictions

```
mypred=[5,1078,92.72,8.21,6.58,11.61,5.94,0.84,9.58,30.33,0.97,5.09,1.10]
```

```
res=clf.predict([mypred])
if res[0]==1:
    print('High Pollution')
else:
    print('Low Pollution')
```

```
Low Pollution
```

V. CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models (logistic regression and auto regression) can be efficiently used to detect the quality of air and predict the level of air pollution in the future. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities

REFERENCES

[1] Pandey, Gaurav, Bin Zhang, and Le Jian. "Predicting sub-micron air pollution indicators: a machine learning approach." Environmental Science: Processes & amp; Impacts 15.5 (2013): 996-1005.

[2] Dan Wei: Predicting air pollution level in a specific city [2014]

[3] DixianZhuc

[5] Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai: An Empirical Study of PM2.5 Forecasting Using neural network. IEEE Smart World Congress, At San Francisco, USA [2017]

[6] Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. 2003.

[7] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[8] M. Caselli& L. Trizio& G. de Gennaro& P. Ielpo. "A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model." Water Air Soil Pollut (2009) 201:7

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details

www.ijirset.com