iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39241

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Prediction of Air Pollution in Smart Cities Using Machine Learning Techniques

Mrs. G. Gowri[1], Mr. D. Anandhasilambarasan[2]
[1]*Department of Computer Science and Engineering,*
[2]*Department of Computer Science and Engineering,*
[1]*Dr.Mahalingam College of Engineering and Technology*
[2]*Karpagam Academy of Higher Education*

Abstract: ***Air-pollution is one of the main threats for developed societies. According to the World Health Organization (WHO), pollution is the main cause of deaths among children aged under five years. Smart cities are called to play a decisive role to increase such pollution in real-time. The increase in air pollution due to fossil fuel consumption as well as its ill effects on the climate has made air pollution forecasting an important research area in today's times. Deployment of the Internet of things (IoT) based sensors has considerably changed the dynamics of predicting air quality. prediction of spatio-temporal data has been one of the major challenges in creating a good predictive model.***
There are many different approaches which have been used to create an accurate predictive model. Primitive predictive machine learning algorithms like simple linear regression have failed to produce accurate results primarily due to lack of computing power but also due to lack of optimization techniques. A recent development in deep learning as well as improvements in computing resources has increased the accuracy of predicting time series data. However, with large spatio-temporal data sets spanning over years.
Employing regression models on the entire data can cause per date predictions to be corrupted. In this work, we look at dealing with pre-processing the times series. However, pre-processing involves a similarity measure, we explore the use of Dynamic Time Warping (DTW). K-means is then used to classify the spatio-temporal pollution data over a period of 16 years from 2000 to 2016. Here Mean Absolute error (MAE) and Root Mean Square Error (RMSE) have been used as evaluation criteria for the comparison of regression models.
Keywords: Spatio-temporal data, Primitive predictive machine learning algorithms, regression models

I. INTRODUCTION

The data set provides information of the NO2, SO2, PM10 levels for through 19 years, 2000 - 2019. Relation between the pollutants to their geographical locations translates the problem into a classification issue. The knowledge of similarity between time series is widely used for speech recognition and signature recognition. In our paper, we make use of two pieces of knowledge - factors influencing pollution and seasonality observed in every year between 2000 - 2019. With respect to these concepts to determine the similarity between time series of multiple cities and the similarity between time series of the 192 months in the years 2000 - 2019. Here worked largely with NO2 data as this has been seen to be the cause for lung diseases compared to other methods. In future the dataset will be replaced by real time data getting from sensors. SVM is particularly useful since the data involves a time series and is non-linearly related.

This method can also provide a better generalization error. Here conducted several experiments using different models and determined a low cost-complexity. There has been extensive research on developing highly accurate spatio -temporal models using different machine learning approaches. Finally the model produce each gases contribution in air pollution. This section emphasizes on some approaches are considered before choosing the appropriate model for our work.

Due to the numerous topography and extent of industrialization within the cities, predicting environment pollutant values help in foreseeing the effect and extent of pollution. Countries deploy many sensors to record different pollutant levels in urban areas also as near industrial zones, but the most index employed by governments to depict the pollution levels is that the Air Quality Index. This is often a crucial measure because it helps to work out the general quality of air which consequently is employed to work out the adverse health and climate effects which are caused to the environment.

Here, presenting a spatio temporal prediction model which might be highly effective in determining the AQI also as individual pollutant levels over a period of your time.

II. PREMODEL ANALYSIS

Air pollution prediction problem has been solved within the prevailing system by using statistical linear methods but these techniques are poor estimator for pollution prediction. They used sparse sampling, randomized matrix decompositions as a pre-processing to scale back the dimensionality of the data. They need used random forest regression technique for forecasting next 10 days.

They only took one pollutant O3 for future prediction and therefore the data subsample size is little. pollution prediction using machine learning Dynamic Neural Network (DNN) approach was carried on data generated by their low cost sensors. Prediction of Ozone Concentration in Smart City using Deep Learning is proposed. They need performed comparison with SVM, RNN machine learning algorithms and prove that deep learning neural networking perform well in accurately measuring the pollution value. They only took one pollutant and that they solved the matter linearly. They didn't mentioned how the important time data are going to be maintained.

They have used three monitoring stations for measuring ozone concentration and used Random forest and Support vector Regression for future prediction. They have found Random Forests to be more accurately estimator for predicting ozone. However, the data from three stations is of bit which they need considered only one variable for future prediction. The prevailing systems detect the air quality of a selected city selected by the user and groups it into different categories like good, satisfactory, moderate, poor, very poor, severe supported AQI (Air Quality Index). The data is displayed on a monthly, weekly or day to day. Also, once the values are forecasted, the values don't change with regard to the sudden change within the atmospheric conditions or unexpected increase in traffic. The values are detected for the entire city, and can't be verified for the accuracy of the forecasted values afterward.

There are applications that display the real-time PM2.5 levels, while some show the forecast of a selected day. However, PM2.5 levels for dates after every week is not forecasted. This system exploits machine learning models to detect and predict PM2.5, Ground level O3 gases levels supported a knowledge set consisting of atmospheric conditions during a specific city. It finds the toxic substances from the air, but it cannot produce the simplest Normalized Root Mean Square Error (NRMSE) value and that we cannot work with the statistic data.

III.PROJECTED MODEL & ANALYSIS

The data set provides information on the town, NO2, O3, SO2 levels for through 20 years. Relation between the pollutants to their geographical locations translates the matter into a classification issue. Compared to other methods, SVM is especially useful since the data involves a timeseries and it is not suitable for non-linearly related. This method also can provide a far better generalization error. So as to predict continuous values, however, leads to the use of a variation of SVM - SVR, LSTM, ARIMA model, K-means clustering algorithms.

The proposed system finds the contribution of SO2, NO2, PM10 gases in pollution. Individual gases AQI is calculated for each of the separate pollutant concentration. Highest of all the values classify the location’s AQI at that given point in time. particulate matter, sulfur dioxide, nitrogen dioxide plays major role in polluting air. These gases are contributors for AQI calculations. Provides Information and Communication Technologies (ICT) for better health, transport and energy related facilities to the citizens and enables the govt to form efficient use of obtainable resources for the welfare of their people.

Differing types of knowledge collection sensors are deployed at various points within the town which act as a source of data for management of city resources. Better control, energy conservation, waste management, pollution control and improvement publicly safety and security are among the elemental objectives of developing a sensible city. Individual AQI is calculated for each of the separate pollutant concentration and highest of all the values classify the locations AQI at that given point in time. Particulate matter, sulfur dioxide, ground-level ozone, nitrogen dioxide and carbon monoxide gas are important contributors for AQI calculations.

AQI is calculated and reported on hourly basis at the most places to convey estimates of pollution to general public. When AQI is high, people with heart and respiratory diseases may avoid outdoor activities or may use mask to guard them. New systems are proposed which are supported the data gathered from sensors can play an important role in helping the cities manage and measure air quality. With the assistance of sensors generating data, the smart city decisions are made much faster and easier but the processing of knowledge brings its own challenges.

TABLE I: The Proposed system using following algorithms.

Problem Statement	Technique
Analyzing air quality using machine learning	Regularization and Optimization (Support Vector Machine , Support Vector Regression)
Machine Learning techniques for classifying air quality	Decision tree and K -Nearest Neighbor Algorithm
AQI Prediction	K – Means Algorithm
Low cost AQI Measuring sensors deployment and use machine learning analysis	Random Forest (Classification and Regression)
Air pollution forecast for short period of time	ARIMA Model, Dynamic Time Warpin
Low cost sensors and efficiently predicting pollution	Dynamic Neural Network (DNN)

IV. EXECUTION

When our data is comprised of attributes with varying scales, many machine learning algorithms can enjoy rescaling the attributes to all or any have an equivalent scale. This is useful for optimization algorithms is used in the core of machine learning algorithm like gradient descent. It is also useful for algorithms that weight inputs like regression and neural networks and algorithms that use distance measures like K-Nearest Neighbors. We can rescale our data using scikit-learn using the MinMaxScaler class. Here for month wise clustering, Rescaling of data applied to predict future values of gases. The following are the execution process carried to find the Integration and reduction data

1) Data Collection
2) Data Preprocessing
3) Data Classification
4) Data Clustering
5) Target Model

Table 2: Raw Data

S.NO	AREA CODE	NO2 AQI	SO2 AQI	PM10 AQI
1	01	46	34	14
2	02	46	34	14
3	03	46	34	8
4	04	34	37	8

Table 3: Preprocessed Dataset

S.NO	AREA CODE	NO2 AQI	SO2 AQI	PM10 AQI
1	01	0.634	0.256	1.0
2	02	0.821	0.347	0.965
3	03	0.342	0.167	0.289
4	04	0.628	1.0	0.876

A. Month wise Clustering of Data

After obtaining the time series data, we need to find alignment to determine the distance between two, time series to form clusters. Euclidean is one of the most common method used to determine the distance, but it is not effective for time series data. Hence, we use weighted Dynamic Time Warping to calculate the alignment between any two given time series. Dynamic Time Warping finds the optimal nonlinear alignment between two time series. To quantify this result, we calculated the alignment between the time series of 2000 and 2019.

Euclidean distance method

Euclidean distance results in no weight for phase shifted time series. For instance: if two-time series are T0: 1213110 and T1: 8121311, then with Euclidean distance, the distance between the two is calculated piecewise. (1 and 8), (2 and 1), (1 and 2) etc. However, as it is noticeable, the two-time series differ only by one position. This doesn't make the series as distant as Euclidean distance concludes it to be.

B. Dynamic Time Warping

From ARIMA model, the correlation among different levels of knowledge in context of your time series present within the data must be established. Initial experiments were focused on discovering time-series at the month level of the data, for every day of the given month. DWT clearly indicates the acute variance and lack of correlation between different values. This consequently shows a definitive lack of seasonality for a given month. To resolve this issue, we analyzed yearly time-series of various cities and established that there is noticeable correlation between time-series of various areas. Thus, to extend intra cluster correlation, the time-series of various areas which are almost like one another are merged. Within the wake of getting the time arrangement information, we've to get to make a decision the separation between two, time arrangements to shape bunches. Euclidean is one among the foremost widely known strategies wont to decide the separation, however it is not viable for time arrangement information. Consequently, we utilize weighted Dynamic Time Warping to determine the arrangement between any two given time arrangement. Dynamic time traveling finds the perfect nonlinear arrangement between two, time arrangement. To live this outcome, we determined the arrangement between the time arrangements of 2000 and 2019. Within the wake of applying both the strategies.

The outcomes are: Euclidean distance = 125

Dynamic Time Warping (window size = 10) = 73

Squared distance from one in the first, time series to every point in the other time series. With the above example, the matrix below is received. Every element is (t0 - t1)^2. With the assistance of this matrix, those distance elements are chosen such that the sum of the alignments is that the minimum sum. If that is the case, the highlighted elements would be chosen. This way, the phase difference between the 2 series doesn't contribute to the space. However, it is biased towards reducing the aforementioned effect and compares the last datum of a time series to the first of another. Hence rules like Boundary conditions - restricting the alignment derived from the matrix to start and end at the diagonal ends of the matrix, continuity conditions restricting the amount of elements compared with to seek out the shortest distance, monotonicity condition - ensuring the points compared are spaced in time from the last iteration.

Table 4: DTW Matrix

ti/t0	1	2	1	3	1	1	0
1	0	1	0	4	0	0	1
1	0	1	0	4	0	0	1
3	4	1	1	0	4	4	9
1	0	1	1	0	4	4	9
2	1	1	0	1	1	1	4
1	0	1	0	4	0	0	1
0	49	36	49	25	49	49	64

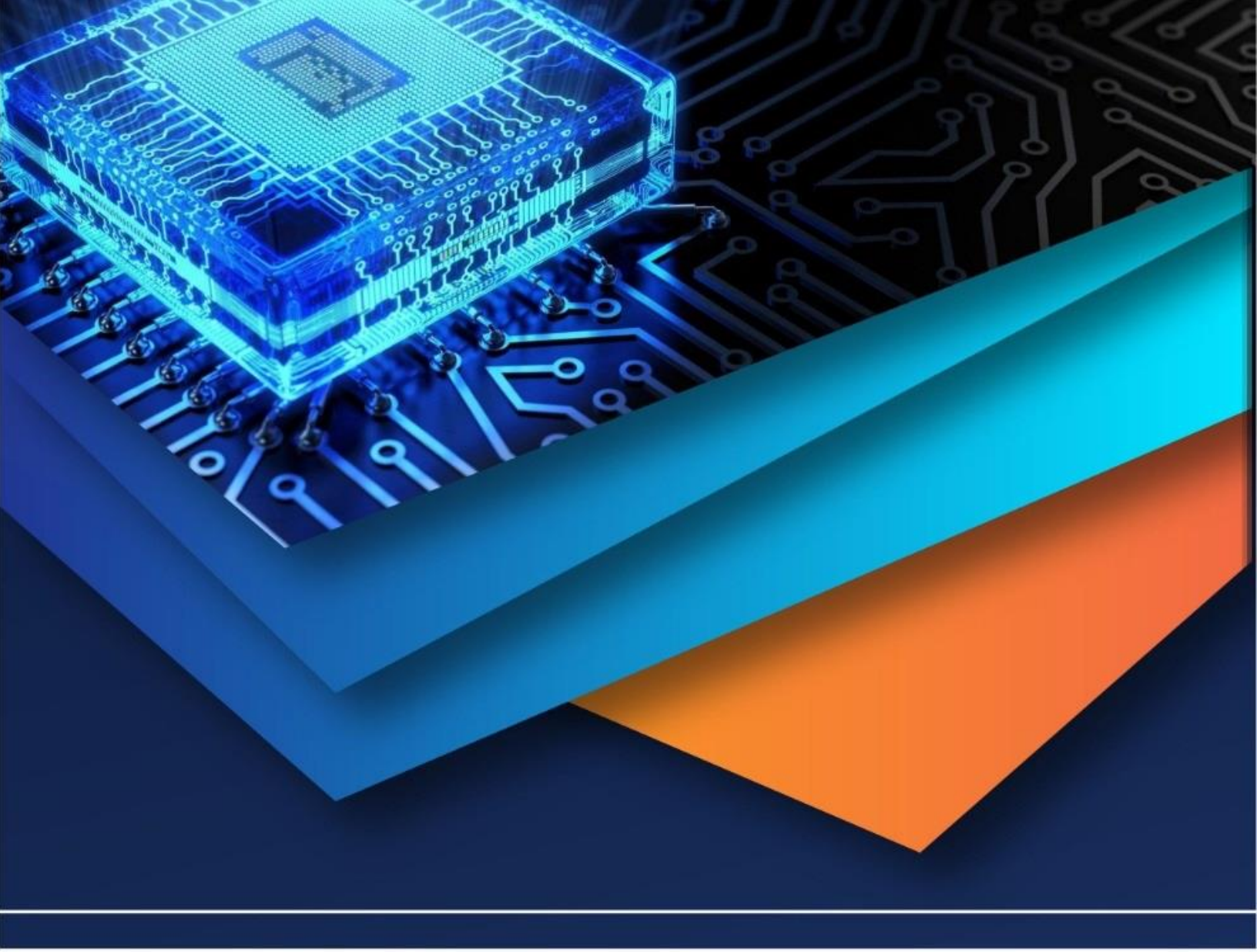
V. CONCLUSION

For predicting air pollutant level, here considered multiple models. The most suitable method as per our evaluation is to cluster months with similar behavior of pollutant levels. K-Means clustering can be used for time series prediction. We observe that pollutant levels follow seasonal behavior. Using K-Means clustering, month wise similar pollutant level behavior were clustered together. For calculating distance for clustering, we conclude that Euclidean distance is not the correct approach. Dynamic Time Warping is one of the possible measures to calculate the alignment between two time-series. Implementing DTW along with LB-Keogh (lower bound DTW) helps fasten the DTW for the given large dataset. Thus, we have k clusters, each representing a group of similar behavior patterns, with each cluster fitted to one regression line each.

City wise clustering the data by using sensors and forecast seasonal clustering. Implementing ARIMA model to create time-series regression over the clusters for prediction. This provides with one time series regression line for each cluster. Implementing Decision tree and Naïve Bayes algorithm to takes into account for spatial and temporal data in order to predict the output.

REFERENCES

[1] YING ZHANG, YANHAO WANG, QINGQING WANG1, "A Predictive Data Feature Exploration-Based Air Quality Prediction Approach", January, 2019.
[2] Richard O.Sinnott, Ziyue Guan, "Prediction of Air Pollution through Machine Learning Approaches on the Cloud," 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), no. December, pp. 1109, 2018.
[3] Asgari, Marjan, Mahdi Farnaghi, and Zeinab Ghaemi, "Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster." In Proceedings of the 2017 International Conference on Cloud and Big Data Computing, pp. 89-93. ACM, 2017.
[4] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," no. December, pp. 114, 2017.
[5] R. W. Gore, "An Approach for Classification of Health Risks Based on Air Quality Levels," pp. 5861, 2017.
[6] Bougoudis, K. Demertzis, and L. Iliadis, "EANN HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens," Neural Compute. Appl., vol. 27, no. 5, pp. 1191, 1206, 2016.
[7] A. J. Cohen et al., "Articles Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015," Lancet, vol. 6736, no. 17, pp. 112, 2016.
[8] Y. Xing, Y. Xu, M. Shi, and Y. Lian, "The impact of PM2 . 5 on the human respiratory system," vol. 8, no. I, pp. 6974, 2016.
[9] S. B. Hiregoudar , K. Manjunath, K. S.patil, "A Survey: Research Summary on Neural Networks", International Journal of Research in Engineering and Technology, ISSN: 2319 1163, Volume 03, Special Issue 03, pages 385-389, May, 2014.
[10] A. Kumar, H. Kim, and G. P. Hancke, "Environmental monitoring systems: A review," IEEE Sensors J., vol. 13, no. 4, pp. 1329–1339, Apr.2013.
[11] V. Sharma, S. Rai, A. Dev, "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN2277128X, Volume 2, Issue 10 october, 2012.
[12] U. Gehring et al., "Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life," Amer. J. Respiratory Critical Care Med., vol. 181,no. 6, pp. 596–603, 2010.
[13] O. A. Postolache, J. M. D. Pereira, and P. M. B. S. Girao, "Smart sensors network for air quality monitoring applications," IEEE Trans. Instrum. Meas., vol.58, no. 9, pp. 3253–3262, sep.2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY