Air Quality Index Prediction in Realtime Using SVM based model in Machine Learning

Rajarshi SinhaRoy[1], *Swarnendu Sarkhel*[2]

[1] *Computer Science, St. Xavier's College, Kolkata 700016*
[2] *Computer Science, Government General Degree College Singur*

Email: rajarshi921@gmail.com, swarnendu.sarkhel04@gmail.com

Abstract

The air quality index is an index to decide the situation of the air quality. The air quality index is a measure of how air pollutants impact a persons' fitness within a time period. It is a standardized degree this is used to suggest the pollutant (so2, no2, pm 2.5, pm 10, etc.) levels. We designed a model that could estimate the air quality index based totally on ancient records of a few preceding years. The performance of this model is progressed through making use of numerous Estimation-Problem logics. Our model could be able to correctly predict the air quality index of a complete county or any nation or any bounded area supplied with the ancient records of pollutant concentration. In our model by implementing a support-vector machine, we achieved better performance than other models and for that our model gets an accuracy of 96%. With the help of support-vector machine, our model estimates the air quality to predict the air quality index of a given location primarily based totally on its ancient records of the pollution of a few preceding years. Our purpose is to increase a non-linear updatable version for real-time air quality index forecasting, to doubtlessly update the models presently being used.

Keywords: *Air quality index, Support vector machine, Prediction, Machine Learning*

1. Introduction

India is the new largest growing industrial nation, so every day we are producing record amount of pollutants like CO2, PM2.5 etc. and other harmful aerial contaminants. Air quality of a state or a country or a region is measured on the effect of pollutants on the respected regions. AQI indicates the levels of major pollutants are present on the atmosphere. Air pollution becomes very vital problem for our environment. There are various atmospheric gases which cause pollution, and each pollution has individual index and scales at different levels. The major pollutants are NO2, SO2, PM 2.5, PM 10. We collected the data from the government database, which contains pollutant concentration occurring at various places across. We calculate the individual index of the pollutant for every datapoints and find their respective AQI. Our model air quality index prediction in Realtime "AQIPR" collects all the data from a given location and then it estimates the AQI in Realtime on the basis of data points in the dataset. "AQIPR" model estimate several data by which we can obtain the regions which are more affected than other locations and it also gives us a knowledge about the data that is extracted using several techniques to know the cause and seniority of the pollutants.

2. Support Vector Machine (SVM)

There are three types of learning in machine learning: Supervised, Unsupervised, Reinforcement. Support Vector Machine is a type of supervised machine learning used for solving Classification and regression type problems. Two-group classification problems of classification algorithms are used for Support vector machine which is a supervised machine.

With respect to newer algorithms like neural networks, there are primarily two advantages expedite and better performance with sample delimited up to thousands. For text Classification where thousands of tagged samples' dataset needs to access, this algorithm is best option. The basics of Support Vector Machines and how it works are best understood with a simple example. Let us assume that we have two bags of: red ball and blue ball, and every coloured ball has features: x and y. We want a classifier that, given a pair of (x,y) coordinates, outputs if it's either red or blue. We have plotted labelled training data already on a plane (FIG 1).

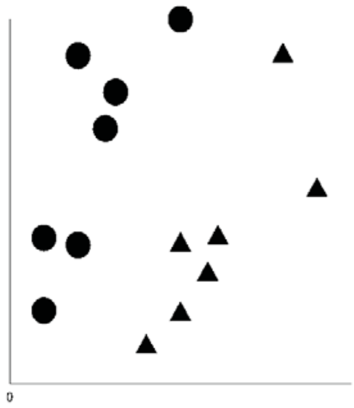


FIG 1. Data are labelled

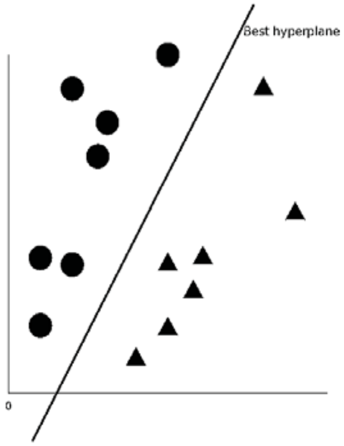


FIG 2. Hyperplane

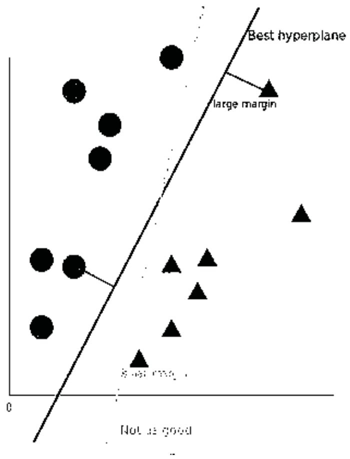


FIG 3. Hyperplane along with another plane

To produce the output on hyperplane SVM choose the data points and placed in 2D that segregates the tag best. We can consider the left-side data samples as circle and the right-side data samples as triangle with respect to the position of hyperplane and by this we take the decisions.

To tag the largest we consider the distance between the closest element and the hyperplane. There are two other lines: one is near the circle types datapoint and the other is near the triangle types datapoint which is parallel to the hyperplane. If the distance between Hyperplane and two parallel lines are d- (for circle) & d+ (for triangle) then margin = (d-) + (d+) = m, will provide the maximum width which helps to predict much better. Again, SVM has two types: linear and Non-linear. If two different types of data points cannot be separated by the linear plane, then the non-linear plane is used.

3. AQIPR Model Design Approach

3.1 RAW-Data Pre-processing

Basically, we collected the raw data from different sensors. And that's leads a probability of missing and noisy data in that data set. In our AQIPR model we used several normalization methods such as Min-Max Normalization, BVA or Boundary value Analysis. Getting some faulty and NAN data is common for sensors so we used some techniques to get the average value for those input.

3.2 AQI calculation

We collected the dataset with our sensors from various places. We got the average readings of existing air quality with respect to air quality parameters, like SO2 (Sulphur dioxide), NO2 (Nitrogen dioxide), Particulate Matter 2.5 and Particulate Matter 10, etc. In several data in data set is either void or NAN because of faulty sensors. In order to remove the outliers, we have to pre-process the data. We have calculated every individual pollutants' index which gave us the docility of pollutant concentrations along with its corresponding individual indices.

3.3 Pearson correlation coefficient

From Pearson Correlation Coefficient, we can identify the efficiently co-related pollutants. NO, NO2 is co-related with NXO and PM10 is co-related with PM2.5. Now we delete those pollutant data from the data set and create a new dataset. After that, we again calculate the individual pollutant index, and using that we find the AQI.

3.4 Use DTW, Split data set and SVM

Dynamic time warping (DTW) is used as an algorithm for measuring the similarity between two temporal sequences, which may vary in speed, in time series analysis. Using this estimation, in our model AQIPR dataset is splitted into two parts of 70% and 30% of data into train and test dataset respectively to cognize the huge seasonal trends along with variations. Then we used the SVM where the kernel is linear to get the accuracy.

3.5 Accuracy calculation

False Positives (FP): When the quality of air is good and if our algorithm predicts that quality of air is poor, then it will be false positive [1].
False Negatives (FN): If the quality of air is bad and if our algorithm predicts that quality of air is good then it is also a false negative [1].

True Positives (TP): If the quality of air is bad and if our algorithm predicts correctly, that is quality of air is bad then it is true positive [1].

True Negatives (TN): If the quality of air is good and if our algorithm predicts correctly, that is quality of air is good then it is true negatives [1].

Accuracy calculation = (TP + TN) / (TP + TN + FP + FN) [1]

Accuracy is essentially an important efficiency parameter [1]. Accuracy is the ratio of properly expected commentary to total number of records [1]. We may think that if we've got excessive correctness then the mannequin is exceptional [1].

3.6 Discussion and result

In this project, we were worried about the accuracy of the algorithm model. But we tried a lot of methods to find the best accuracy to make it more reliable for the real-time project. Below we talk about the accuracy, graphs, and estimate AQI value regarding the value of pollutants.

The project's graphs are the main key to find the best results. We create the graph of the given dataset and create an AQI vs monthly graph and then reduce the data set using the Pearson correlation coefficient method. After achieving the reduced data set, we get the AQI Graph and test both. We get that both are nearly equal to each other. So, both are overlapping the graph.

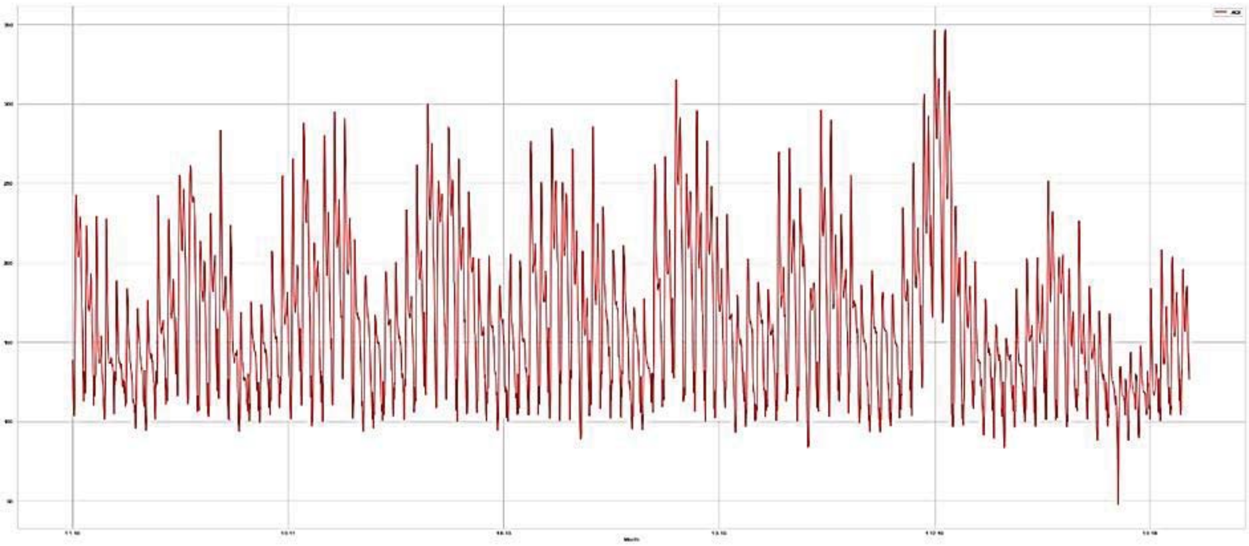
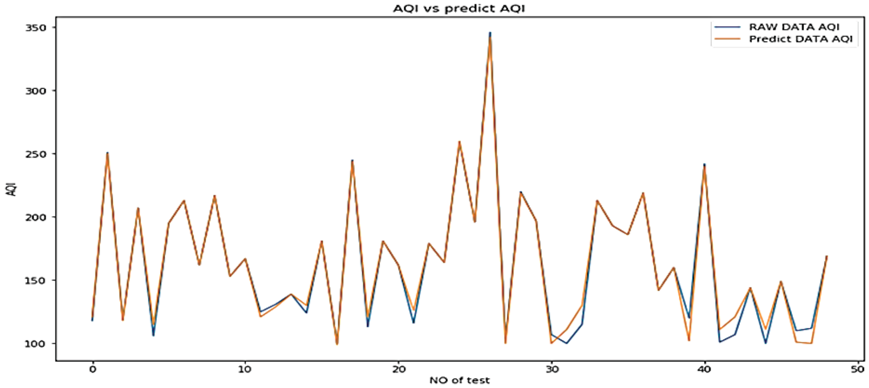


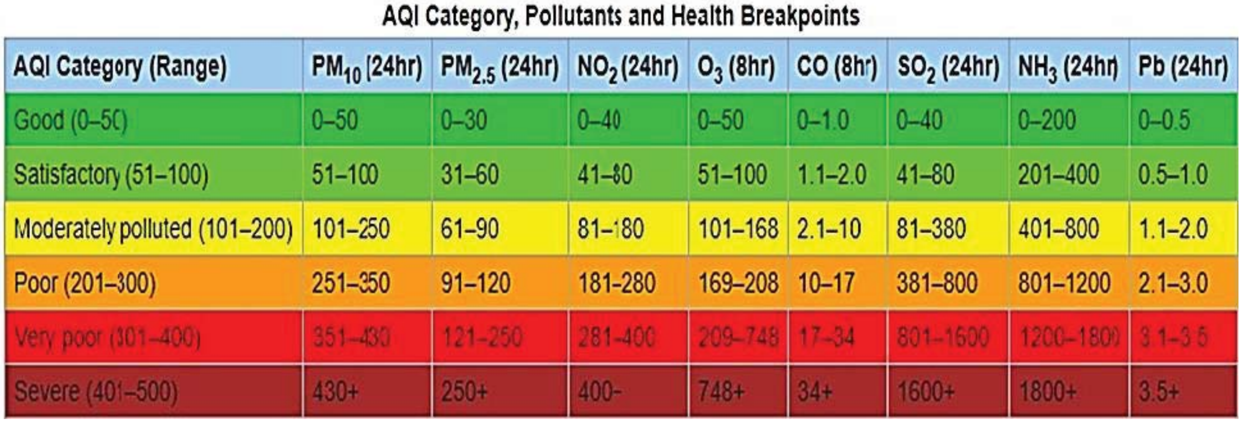
FIG 4. Month vs AQI graph

The above is graph shows AQI concerning various months after reducing the dataset. In the x-axis "Month" is considered and in the y-axis, "AQI" is considered. This is the first thing we get by calculating and plotting graphs. From here we get that the reducing data set and the original data set is quite equal. So, we can perform the rest of the project on the reducing data set. In SVM we tried several kernels to find the best accuracy and graph plot. Only in Liner kernel, we get the best accuracy of 95-98% (depends on the project data size and splitting size). Exceptions might be plotted as individual focuses. We created a box plotting graph to maintain the results as well as we create a graph. In the below graph x-axis is “Number of tests” and the y-axis is “AQI’. The orange color is for “Raw Data AQI” and the blue color is for “Predict Data AQI”. The graph tells that the line is nearly overlapping in nature.



The above graph is a close view of the previous graph (here takes only 50 data). Here we can see the nearly overlapping nature of both the dataset. This implies that our prediction is nearly close to the actual value, and our machine is 96%capable of predicting the “AQI” correctly based on a range defined from 1-6. However, it can be improved by feeding more data.

Basically, for India, there are six types of AQI which we can tag as: Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe [12]. There are some other elements besides the main pollutants are PM10, NO2, PM2.5, CO, SO2, PB, O3, NH3 [12]. Those 8 pollutants are basically nominated for the National Ambient Air Quality Standards. Impacts towards human, concentration of ambient and corresponding standards are three keys to measure the sub index of each pollutant. The worst sub-index shows the overall AQI for those pollutants. Medical experts suggested that AQI directly impact human body. The values of each pollutant’s AQI and their ambient concentrations are shown in below table with health breakpoint:



AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM$_{10}$ (24hr)	PM$_{2.5}$ (24hr)	NO$_2$ (24hr)	O$_3$ (8hr)	CO (8hr)	SO$_2$ (24hr)	NH$_3$ (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181-280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400-	748+	34+	1600+	1800+	3.5+

By this data analysis, we came to know that the best way to resampling this data is in hour format and month format. So that's the reason we got the reduced matrix, used to fit the data in ours model. But we can again optimize the model by using the gradient decent hyperparameters.

4. Conclusion

Our main aim was to predict the Air Quality Index (AQI) from its previous values. Predicting AQI will help the society and the people to take action accordingly, such as one can decide whether to go outside with a mask or without a mask or to stay inside. Calculating AQI with too many sensors is costly as well as time-consuming here in this project, we reduced the cost without any hamper to its efficiency. We had provided the data on an hourly basis so one gets information more precisely and acts accordingly.

Although our machine can predict up to 96% on a range basis, still the machine can further be improved with more data collected from sensors hourly. Another improvement can be done in the time-management field, where the data is predicted hourly, we can further improve to make it predictable in half an hour or maybe in minutes and the time-series data of every possible region needed more attention, used in our model.

Let's talk about the prediction of AQI, here our model is 96% accurate and it can predict any region's air quality index if dataset covers minimum 5 years of historical data. To alert any region our model is capable of doing that and moreover our model has an ability to track back to the destination region, by using rethinking which makes our model less dependent on historical dataset of that region.

REFERENCES

[1] K. Mahesh Babu, J. Rene Beulah, IJITEE 8, 2278 (2019)

[2] N. H. Abd Rahman, M. H. Lee, IAES 9, 33 (2020)

[3] A. Gnana Soundari, J. Gnana Jeslin and Akshaya A.C, IJAER 14, 0973 (2019)

[4] Wang Jun, Sundar A. Christopher, Geophysical research letters 30, 2095 (2003)

[5] J. He, S. Gong, Y. Yu, L. Yu, L. Wu, H. Mao, C. Song, S. Zhao, H. Liu and X. Li et al., Environmental pollution 223, 484 (2017)

[6] E. Kalapanidas, N. Avouris, Proc. ACAI 99, (2017)

[7] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, Knowledge Discovery and Data Mining 21, 2267 (2015)

[8] V. M. Niharika and P. S. Rao, International Journal of Computer Science and Information Technologies 5, 103 (2014)

[9] Liang Y-C, Maimury Y, Chen AH-L and Juarez JRC, *Applied Sciences* 10, 9151 (2020)

[10] D. J. Nowak, D. E. Crane, and J. C. Stevens, Urban Forestry &Urban Greening 4, 115 (2014)

[11] M. Caselli, L. Trizio , G. de Gennaro and P. Ielpo, Water Air Soil Pollut 201, 365 (2009)

[12] A. Kumar, International Journal of Innovative Science and Research Technology 3, 2456 (2018)