International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

An Unified Method of Exploration-Based Air Quality Prediction

R. Kavitha, Thrinath Sirna, Panini Vashisth, Rithik Shaw

Abstract:- Data mining is the application of examining large current databases in sequence to create new information. It is a classification of artificial intelligence build on the concept that systems can get from data, analyze patterns and make judgment with minimal human intervention. The forecast of air quality is done with analyzing the AQI (Air Quality Index) of the atmosphere in different areas. These predictions are done using the BP Neural network Algorithm in which the data of the gases like CO2, CO, SO2, O3, NO2, PM2.5 etc. is first classified in the system, and then the normality is checked by comparison of each gases with the normality. But the prediction cannot be fully excepted because it doesn't consider the outside weather condition of the atmosphere. This paper uses the ANN (Artificial Neural Network) technique along with BP Neural Network which analysis the weather condition of the atmosphere along with the data of the polluted gases. This paper predict more efficient air quality index of the atmosphere.

Key Word:-Data Mining, Back Propagation Neural Network (BCNN), Air Quality Index(AQI)

I. INTRODUCTION

Nowadays, people are more focusing on the importance of air quality and other aspects which are related to air but nowadays the major problem which we are facing in our daily life is the quality of air which we are living. We really need to concentrate on the air pollution because this can lead to a disaster in our life as already the pollution level has crossed the limit and we need to reduce air quality. There are different factors that affect the air quality and these factors changes from place to place. The level of air quality has crossed the upper limit due to which we have to focus on air quality. The data[1] which is being provided by metrological station are precise but not upto the mark. Our paper takes the information, data and other important values which are essential for air quality prediction, using these values in data mining techniques we are efficiently predicting new values which are more accurate. By using this we can find out the elements which are causing air pollution and there prime aspects.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Dr.R.Kavitha*, AP/Department of IT, SRMIST, Ramapram, Chennai. kavithar8@srmist.edu.in
Thrinath Sirna, UG Student, Department of IT, SRMIST, Ramapram, Chennai. thrinathsirna@gmail.com
Panini Vashisth, UG Student, Department of IT, SRMIST, Ramapram, Chennai. paninivas@gmail.com
Rithik Shaw, UG Student, Department of IT, SRMIST, Ramapram, Chennai. rithikshow@gmail.com

Taking the past values we are pre processing the data values from different source so as to remove the similar values as well as the duplicate values to get more precise results. After that we are analyzing these values by comparing the current values with historical values to train the data that are essential for the system. First we find the component which is affecting the air quality, second we will train the data which is influencing the air quality and finally we evaluate the air quality so the prediction of data is precise and efficient.

This system consists of two types of data one is existing data and the other is historical data. This paper uses both the data efficiently by combing these datasets values using specific techniques, there should be no repetition values because if it arises then the output will not be precise. Therefore we are preprocessing these values from different data sources in order to avoid co- dependency of these values.

We are taking the 144 hours upto date and past 24 hours so that we can predict the present data effectively. Air quality prediction is highly unpredictable when we are considering the same time because during the same day the PM2.5[2] concentration changes a lot. Different geographical stations have different air quality concentration values as a result it is difficult in such conditions when you have variation of values in some country of different locations. Another problem is that concentration values keeps on changing with the change in time as a result different values of output are expected when tested in different time. Another important problem is that the values from different sources are missing some or the other values which are essential for measurement and as a result these data sets cannot be combined. These past values are very important as we have to use them for prediction .Fusion of data is also very essential because this will show the variation in the values in this time.

The proposed system is as follows:

We use light GBM model to combine historical data values and future values to predict the present values. It improves our prediction and accuracy. For data missing values, we intend to combine linear estimation with data mining so that the data values problem is eliminated. On account of severe impact of PM2.5 concentration on human body, how to control PM2.5 is the major problem to be addressed. Sliding window mechanism is being used to enhance the data using deep mining[3]**.** Now the major problem is to control the PM2.5 level. There are many features such as that are majorly being associated with the PM2.5 concentration. First is Weather forecasting, timing and statistical values. This PM2.5 concentration varies from place to place.

There are three different model schemes such as time series model, regression model, and data mining model. Using these model schemes we identify the critical information using different features to discover potential vital features. On the bases of these features we can decrease random fluctuations in weather monitoring stations and make the air prediction more stable, dependent, flexible and accurate. Different weather monitoring stations produce different air quality index (AQI), these features helps us to predict the air quality with accurate results.

II. LITERATURE SURVEY

A Light GBM model is based on GBDT and XG boost. It can improve the efficiency of GBDT when high data is being processed. Light GBM [4] has greater amount of efficiency, less memory footprint and support side by side learning. So processing high speed data becomes easy. During the process of training data after many processing the dimension of the data will raise to more than 1 million and is the reason we need to have faster training data and less memory cost for higher efficiency. The computational cost and segmented cost is less. Pre-processing of the data takes place at a very initial stage.

B. Multidimensional data is created by fusing the raw data for every air quality monitoring station. Initial integration takes place between grid dataset and station dataset. The blue dot represents the reference points of the grid and the red dots represents the position of air quality monitoring station. As we see the figure as the some of the reference points are close to the air quality monitoring stations, Therefore we take the meteorological station information of the reference points which are close to the monitoring station so now the reference point information will be taken for the corresponding system (Figure:1).

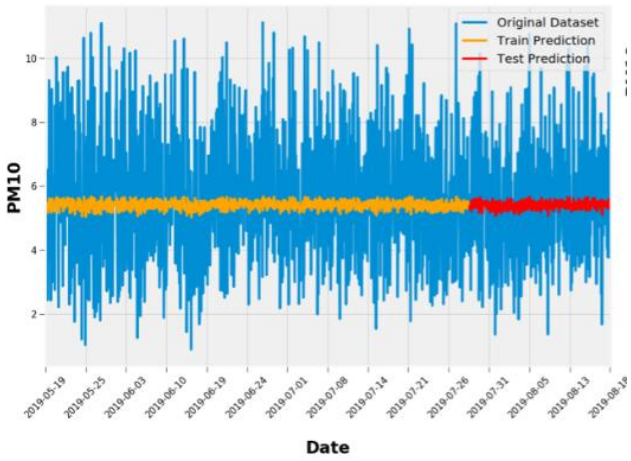

Figure:1 Dot Figure

C. There are many satellite with high resolution power that helps in predicting air quality and are being used in many environmental application but they are still not capable enough to capture high dimensional dataset for measuring air quality of each station. Now, measuring the pollutant level has become the regular part of prediction in television, newspaper etc. To measure different pollutant we need to analysis data across large timeframes and location as well, like taking 8 hours rolling air average pollution, showing, identifying similar patterns in the concentrations from multiple bases. With the advancement in the industry the quality of air is becoming worse day by day. The government is increasing the no of air quality monitoring devices so that the problem can be solved.PM2.5 is a standard level which needed to be maintained and hence we need new technology to get to know about the reason behind the worst air quality and not the measurements.

D. Getting the patterns from spatio-temporal data has become more difficult than getting the patterns from the normal numerical data types. Epidemiology[5] identifies disease patterns and variations in health risk. The problem in traditional mining is that many rules are discovered, most are very rarely used in stating objectives or questions asked. Moreover, not all rules are interesting (due to the factors of diseases), some rules might be ignored. PM2.5 concentration is important hot spot in international community. The paper is based on exploratory data analysis and visual representation.

III. SYSTEM ARCHITECTURE

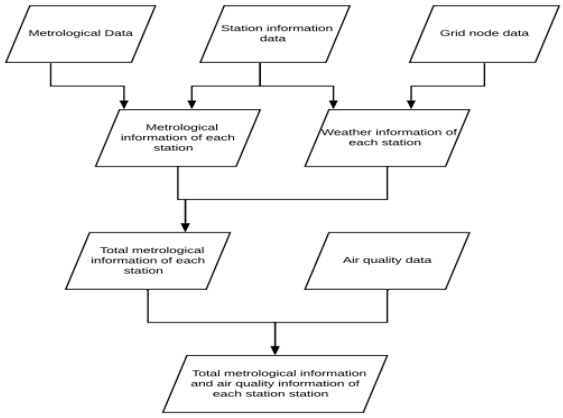

Figure:2 System Architecture

The raw data is fused across different substations to improve air quality prediction. First, the raw data enters into station information data and checks the air quality index(AQI) and sends it into the grid nodes, these grid nodes checks the air quality level ,it checks if the air quality is above the danger level of PM2.5 concentration. Next, weather information from each station is checked. Here, again the air quality index (AQI) [6] is filtered to double check the air quality index (AQI) and then it sends the data again to the weather monitoring station of each area collected .This way of predicting air quality index (AQI) for each station is more accurate and precise. Data sets of different stations across different parts in the city and the country is very different, so this model helps in improving prediction using Figure:2.

There are two points reference point and actual point. If the air quality is above reference point then the air quality of that area is high, if the data is below the reference point then the air quality is low.PM2.5 concentration is the perfect air quality concentration. There are several substations available in the country.

First, the PMI concentration of a particular area is collected and seen if that PMI is above danger mark. If it is above danger mark then the station signals it. Different substation collect information across different regions and air quality index (AQI) is taken at each place. Weather information is also collected across different places. If there are many pollutants in the atmosphere then Air Quality Index (AQI) [7] is depreciated.

IV. IMPLEMENTATION

A. Exploratory data analysis

It is a pre process of doing the initial stage of examination so that we can discover patterns, sporting anomalies. It is also used to check the hypothesis. One of the most essential part of this is to check the supposal value using the graphical representation. EDA and statistical are not the same. EDA[8] is based on visualization of the data as well as graphical reading of the data whereas the statistical data represents relation between two data. It is better to have a visual representation of the data so the values can be visualized easily and the output is more precise so the final outcome of the data is in form of simple summary.

B. Calculate Quantile

Quantiles focuses in a dissemination that identifies with the rank request of qualities in that conveyance. For example, you can discover any quantile by arranging the example. The center estimation of the arranged example (center quantile,50th percentile) is known as the middle. The points of confinement are the base and most extreme qualities. Some other areas between these focuses can be portrayed as far as centiles/percentiles. Centiles/percentiles are depictions of quantiles [9] comparative with 100; so the 75th percentile (upper quartile) is 75% or 75% of the route up a climbing rundown of arranged estimations of an example. The 25th percentile (lower quartile) is one fourth of the path up this rank request. Percentile rank is the extent of qualities in a conveyance that a specific worth is more noteworthy than or equivalent to.

For instance, if a student is taller than or as tall as 79% of his cohorts then the percentile rank of his stature is 79, for example he is in the 79th percentile of statures in his group. Quartiles are additionally quantiles; they separate the circulation into four equivalent parts. Percentiles are quantiles that isolate a dispersion into 100 equivalent parts and deciles are quantiles[11] that partition a circulation into 10 equivalent parts. A few creators allude to the middle as the 0.5 quantile, which implies that the extent 0.5 (half) will be underneath the middle and 0.5 will be above it. Thusly of characterizing quartiles bodes well on the off chance that you are attempting to locate a specific quantile in an informational index (for example the middle).

C. Prediction

The Long Short-Term Memory system, or LSTM organize, is a repetitive neural system that is prepared utilizing Back propagation through Time and conquers. In that capacity, it well be utilized to make enormous repetitive systems that thus can be utilized to address troublesome succession issues in AI and accomplish cutting edge results. Rather than neurons, LSTM[10] systems have memory hinders that are associated through layers. A square has segments that make it more intelligent than a traditional neuron and a memory for late groupings. A square contains doors that deal with the square's state and yield. A square works upon an info succession and each door inside a square uses the sigmoid initiation units to control whether they are activated or not, rolling out the improvement of state and expansion of data moving through the square contingent. LSTMs can be utilized to show univariate time arrangement estimating issues. These are issues involved in solitary arrangement of perceptions and a model is required to gain from the arrangement of past perceptions to foresee the following an incentive in the succession. The LSTM model will become familiar with a capacity that maps a succession of past perceptions as contribution to a yield perception. All things considered, the grouping of perceptions must be changed into numerous models from which the LSTM can learn.

V. RESULT ANALYSIS

As you can see from the figure there are three data sets blue, yellow and red. The blue data set represents the original data set which are taken from the monitoring stations that are near by monitoring station of different places .One more important thing is that we have to take the data from same place but different time so that we can also get the variation in data and if the values are same then it can be removed by the pre processing unit in the very initial stage .Now as we can see yellow fluctuation of lines in this datasets are in different time across pm 10 concentration. The original data sets are in above train prediction. This parameter is created using programming[12]. This parameter is set and if the air quality of the data set is above this line then it means the air quality is unhealthy and if it is low then it is fine. After programming this pattern with the help of original data sets and trained data sets we get the test prediction values in the same graph as in Figure:3 which shows a very less variation in the values because of pre processing of data that are being given.

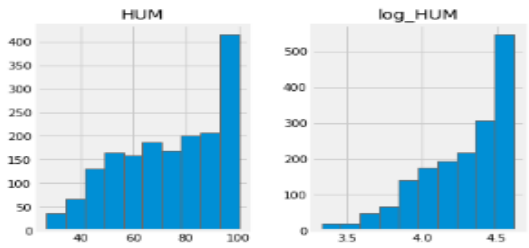

Figure:3 Air Quality Analysis

VI. CONCLUSION

Now days Air Pollution has become one of the major problem in our country and the level has even crossed the upper limit in some of the major city and one of the major reasons behind this the lack of precision and accuracy in the data of Meteorological station .

This paper is summing up the data from different meteorological stations from different places and the unique aspect is that it takes data from the same station at different time so that we can get better results and after the preprocessing of data we use the data mining technology to convert these data and once it is being done we can write the source code of the data and the graph by which we can get the output as a graphical representation .This paper will help the people to know more about the pollution level and other aspects. As we can see from the paper that is based on data mining that the output shows a graphical representation of data sets that gives the user a wide variety of dynamics of different level of pollution with there range which is done by using coding.

REFERENCES

1. R.M.Rani, Dr.M.Pushpalatha," Generation of Frequent sensor epochs using efficient Parallel Distributed mining algorithm in large IOT", Computer Communications, Volume 148, 15 December 2019, Pages 107-114
2. R.Mythili, Revathi Venkataraman, T.Sai Raj,"An attribute-based lightweight cloud data access control using hypergraph structure", The Journal of Supercomputing(JoS),Published online: 02 Jan 2020 DOI: 10.1007/s11227-019-03119-7
3. S.Sivamohan, Liza.M.K, R.Veeramani, Krishnaveni.S, Jothi.B, "Data Mining Techniques for DDOS Attack in Cloud Computing", IJCTA Interntional Scoience Press, Pg: 149-156
4. S Pandiaraj, Aishwarya, Surbhi, Alisha Minj, Priyanshu Singh, "Enabling Cloud Database Security Using Third Party Auditor", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8 Issue-4, April, 2019
5. R.Veeramani,Dr.R.Madhan Mohan, "Iot Based Speech Recognition Controlled Car using Arduino", International Journal of Engineering and Advanced Technology,Volume-9 Issue-1, October 2019
6. T.H. Feiroz khan, N.Noor Alleema, Narendra Yadav, Sameer Mishra, Anshuman Shahi "Text Document Clustering using K-Means and Dbscan by using Machine Learning",International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019
7. S.Babeetha, B. Muruganantham, S. Ganesh Kumar, A. Murugan, "An enhanced kernel weighted collaborative recommended system to alleviate sparsity", International Journal of Electrical and Computer Engineering (IJECE), Volume 10, February 2020, Page No. 447-454
8. Kavitha.R ,K.Malathi,"Recognition and Classification of Diabetic Retinopathy utilizing Digital Fundus Image with Hybrid Algorithms", October 2019,International Journal of Emgineering & Advanced Technology(IJEAT), Volume 9, Isssue 1, 109-122
9. T.Chandraleka,Jayaraj R, " Hand Gesture Robot Car using ADXL 335 ", International Journal of Engineering and Advanced Technology (IJEAT)', Volume-8 Issue-4, Nov 2019
10. H.Sangeetha,S.Abinayaa, "Smart Irrigation Systems using Sensors and GSM" in 'International Journal of Recent Technology and Engineering (IJRTE)', Volume-8 Issue-1, May 2019. Page No.:884-886
11. B.Sathya Bama,,Y.Bevish Jinila, "Attacks in Wireless sensor networks- A Research" ,International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-9S2, July 2019
12. Vellingiri, J., S. Kaliraj, S. Satheeshkumar and T. Parthiban , "A Novel Method for User Navigation Pattern Discovery and Analysis for Web Usage Mining", Journal of Computer Science 2015, vol 11 (2): Page no 372.382

AUTHORS PROFILE

Dr.R.Kavitha, Associate Professor, Dept of IT, SRMIST, Ramapuram, Chennai, have about 13yrs of experience, Published about 5 papers in WOS, 59 papers in Scopus Indexed Journal, 1 Patent. Data Mining, Cloud Computing, Machine Learinig are the area of research.

Thrinath Sirna, is currently pursuing bachelors of technology in information technology from SRM IST,Ramapuram, Chennai,Tamil Nadu,India

Panini Vashisth, is currently pursuing bachelors of technology in information technology from SRM IST,,Ramapuram ,Chennai,Tamil Nadu,India

Rithik Shaw, is currently pursuing bachelors of technology in information technology from SRM IST,,Ramapuram ,Chennai,Tamil Nadu,India