A UNIFIED PIPELINE FOR EXPLORATION-BASED AIR QUALITY PREDICTION

[1]Sowmya. S, [2]Sourabh Mishra, [3]Yash J Joysher

[1,2,3] UG Scholar, Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Abstract*: *Nowadays regular weather forecasting has started giving more and more importance to air pollution forecasts. Air quality refers to the pollutant levels in the atmosphere which tell us the extent of air pollution around us. Albeit, we must also understand that pollution as such is a phenomenon which is affected by various other factors (temperature, humidity etc) in the surrounding location at a point in time. The model proposed in this paper uses LSTM network trained with Back propagation algorithm for predicting the air quality and prior to that we use exploratory data analysis methods to determine the essential features from the dataset obtained from .The proposed forecasting model which combines with data mining techniques and neural network algorithm is based on the monitoring data of air pollution obtained from Shijiazhuang air quality monitoring stations.

Keywords: ***Air quality prediction, BPTT, EDA, LSTM*.**

I. INTRODUCTION

The country of China has seen a drastic rise in air pollution levels in recent years, particularly due to ozone and particulate matter levels [1]. Air pollutants have always been significant contributors to climate change. For example, the major greenhouse gas, ozone causes atmospheric warming [7]. Ozone influences climate change by interacting with other molecules in the air and also specifically with aerosol particles thereby making cloud adjustments [4]. Numerical air pollution predictions from meteorological stations This work is an application of numerical predictions from meteorological stations and serves as a useful system for those looking to control air pollution and emissions. We aim to create an air quality prediction system which works over three phases. The first is using exploratory data analysis to find the characteristics from our data set that are essential for predicting the air quality. This is followed by calculation of quantiles which basically helps in dividing the data we have into equal sized adjacent subgroups. The next phase is where we do the prediction. This is done using Long Short-Term Memory network [10]. It helps us to predict the next set of values in a sequence by learning from a series of observations from the past. We also observe that the model gets trained with as little error as possible since LSTM is a type of recurrent network, which means that we successively reduce the error while training the prediction system.

II. EXISTING SYSTEM

Spatio-temporal data sets were generated in few existing systems by doing a few steps of pre-processing. On this generated transactional data, the data mining algorithm was implemented. This information now available was converted into a map schema. The map schema used here was the calendar map schema [3]. The existing HBST frequent itemset mining was a method used to identify similar and repeated patterns in the spatio-temporal representation [2], [13]. Hashing was used in this proposed system to speed up the memory access process as far as possible. The data underwent pre-processing followed by application of the above-mentioned hashing algorithm [5]. This was then followed by concatenating the schema, with its patterns and the hash address. This method proved to be ineffective due to the high frequency of collisions that occurred in hashing. They generated two itemsets for every partition [6][9]. The disadvantages of these existing approaches include their high computational complexity, less prediction efficiency and the lack of dynamic behavior. These disadvantages lead us to propose a new method for air quality prediction which is explained in the following sections.

III. PROPOSED SYSTEM

In the proposed system, we utilize a set of factors like ambient pollutant levels, climatic conditions etc. of the area under consideration to apply classification algorithm and predict the expected air quality. Data mining techniques like exploratory data analysis were used to determine which features or pollution causing factors were essential for the intended prediction. Statistical components are calculated from the data available which helps in unifying the multimodal data. This eases the end user's data inquiry process from the data model. This part comprises of creating the logic and aligning the related data fields with each other. Once the information model is formed, analytics and mining techniques are used on the same.

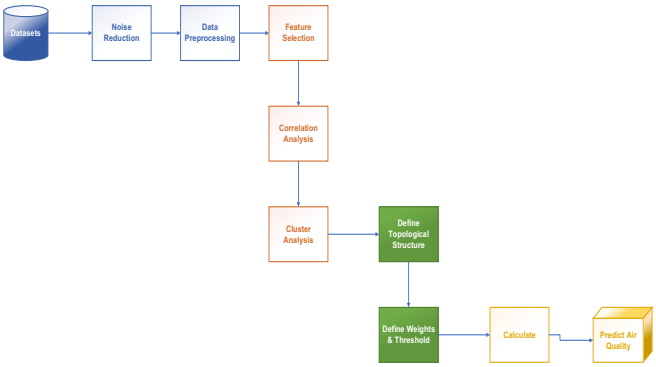

Fig. 1. System Architecture

A. Exploratory Data Analysis

We use the exploratory data analysis or the EDA method to analyze our air pollution data set and summarize the main and essential characteristics from the data. Prior to this data analysis phase, we do some pre-processing of the available data and do noise reduction in order to ensure high accuracy in the results that we obtain. In data analysis this is the first step that we implement before making use of formal statistical techniques. We use this technique to find the pollutants which significantly impact the pollution levels over certain time periods. The first phase of exploratory data analysis deals with feature selection. The features such as SO_2, NO_2, NO, CO2, and levels of various particulate matters suspended in air were found to be essential characteristics. Sulfur dioxide when inhaled causes coughing, wheezing, difficulty in breathing and in some extreme cases premature deaths [11]. In addition to this Sulphur dioxide concentrations influence the habitat suitability for plants and animals as well. Nitrogen dioxide at 25-50 ppm levels causes bronchitis and pneumonia while at higher levels i.e. above 100 ppm it causes deaths due to asphyxiation [14]. Once these essential features are identified we move to the next phase of correlation analysis. Correlation analysis is like a preliminary technique used to identify the relationships, linear or non-linear, among the features selected in the previous step. Correlation is calculated as the ratio between the covariance of two features to the product of their standard deviation. By doing so we get a correlation coefficient whose value ranges between +1 and -1. We then perform cluster analysis which gives us a clear idea about the patterns in air pollutant levels which in turn will help us in predicting the air quality at a future point in time. Using EDA methods proves to be extremely useful. Since all our data is numerical, it gives us statistical evaluation of the pollutant levels and also tells us which of those pollutants might have significant impact on the ambient air quality in the times to come.

B. Calculating Quantiles

The next phase after EDA is the calculation of quantiles. Quantiles are quantities or things that we use to split the distribution that we are working with in some uniform manner. It divides our data into equal sized adjacent groups which eases the process of representing our statistical information as well as our prediction process. The calculation of quantiles enables us to find the minimum and maximum values for all features, in turn helping us find the high and low outliers in the data. We start with the number of equal sized divisions that we want of our available data distribution. Then we decide what quantile will enable such division. Some of the quantiles that are used commonly are addressed with specific names like median is the two quantile which divides the distribution into two equal halves, quartiles are the four quantile and so on [15]. Segregating the data and grouping it together with the help of quantiles enables easier prediction and mining of the data.

C. Prediction

We use LSTM network for predicting the air quality. LSTM is expanded as Long Short-Term Memory network. Recurrent Neural Networks are considered to be better performing than normal ANN's and LSTM is one, which makes it much more efficient than previously used networks. It has the capability to learn long term dependencies [12]. This proves to be an advantage when we are predicting the air quality for a certain point in the future as the network will work based on the memory of a long span of time in the past.

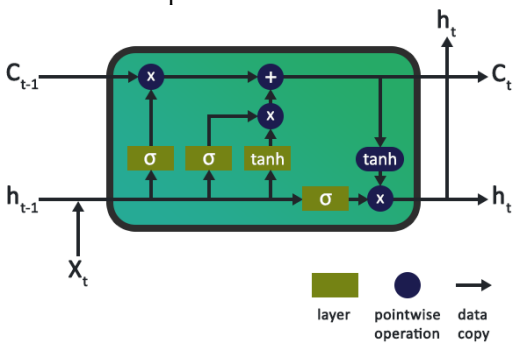

Fig. 2. LSTM Network Architecture

Backpropagation Through Time (BPTT) is the algorithm that we use to train the LSTM network. BPTT is a gradient based algorithm which is generally used for Recurrent Neural Networks [8]. The input training data for our LSTM network is ordered as a set of n input output pairs. This algorithm updates network weights to reduce the error. It's a supervised learning algorithm. In this, time steps are fed into the network to begin with. This is available as couples of input and output. The LSTM is then unfurled. This step is followed by the evaluation and gathering of errors across successive steps. The network is then rolled up followed by changing the weights based on the error accumulation in the previous step. These steps are recurrently performed, and the network gets trained. The LSTM is given one input for each timestep and it produces one output. This is done for all timesteps. BPTT unrolls all these timesteps in order to perform the training. Every timestep has a copy of the network, an input timestep and an output. For each timestep we do error calculation and accumulate it and then unrolling is done and we update the weights.

IV. EXPERIMENTAL RESULTS

A dataset of 9232 pollution records in comma separated format spanning from 19/05/2019 to 18/08/2018 was used to train the entire pipeline of the air quality prediction system. From the dataset we combine the data

from the time and date fields into a single string in datetime format. We then proceed to group the same and form hourly data. This is followed by feature selection, in which we identified O3, CO, NO2, SO2, NO, CO2, PM (1, 2.5, 4, 10), TSP, Temperature, Humidity as the key features or attributes. The minimum, maximum values for each of these attributes was identified and using this the outliers were interpolated. Using the temperature and logarithmic temperature values we create dataframes and make histograms of the same.

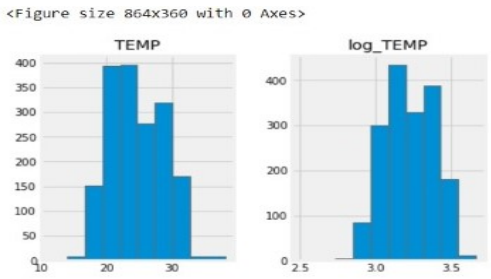

Fig. 3. Histogram for temperature dataframe

Similarly, we also generate histogram plots for humidity and particulate matter (PM).

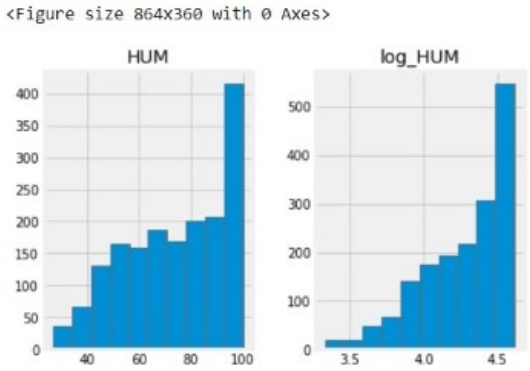

Fig. 4. Histogram for humidity dataframe

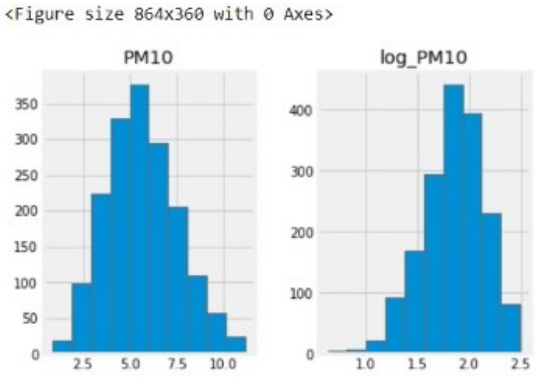

Fig. 5. Histogram for particulate matter dataframe

Utilising this data we proceed to train the LSTM using the BPTT algorithm over 500 epochs. The mean squared errors or loss is computed and accumulated over successive timesteps and the network is then rolled back to update the weights. Once the network has been trained, we use test data and generate predictions. The air quality prediction model is then evaluated using a consecutive set of experiments. The results were visualized using a fitting curve that shows the fitting effect of our model on the test data. The first curve we obtain is for temperature predictions and the next are for PM predictions, both of which are the major influential factors in air pollution predictions. These plots help us conclude that the predictions of the proposed model are consistent with the actual values.

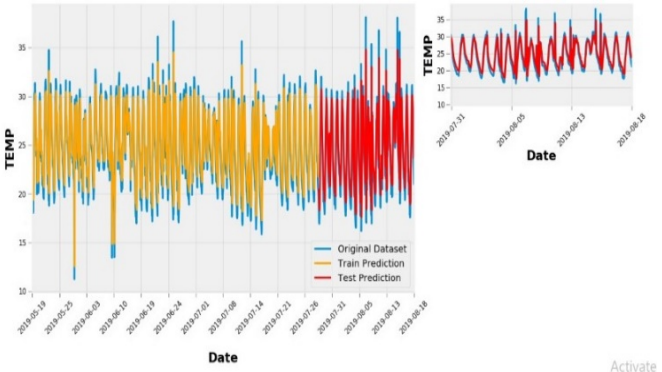

Fig. 6. Fitting curve for temperature predictions

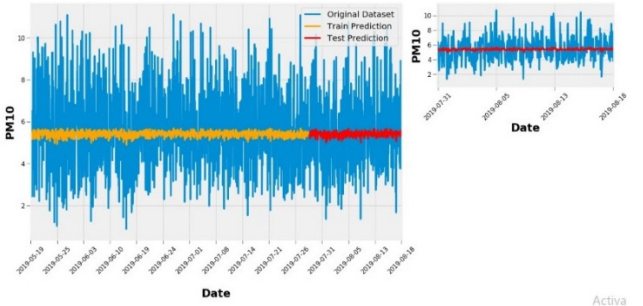

Fig. 7. Fitting curve for PM10 predictions

V. CONCLUSION

A reliable and efficient Air quality forecasting model is created by the proposed system. This model comprises of a combination of the RNN LSTM trained with the BPTT algorithm and other effective data mining methodologies. Using data mining methods, we evaluated the pollutants which had a relatively higher impact on the air quality over time and where their interrelation was comparatively lower. We concluded with a model which provides high accuracy predictions coupled with lesser number of calculations for efficiency. The prediction model proposed herein is bound to prove to be an authentic way for predictions by environmental departments.

REFERENCES

[1] W. Yu, “Spatial co-location pattern mining for location-based services in road networks,” Expert Systems with Applications, vol. 46, pp. 324–335, 2016.

[2] C. Xue,W. Song, L. Qin, Q. Dong, and X.Wen, “A spatiotemporal mining framework for abnormal association patterns in marine environments with a time series of remote sensing images,” International Journal of Applied Earth Observation and Geoinformation, vol. 38, pp. 105–114, 2015.

[3] H. Nguyen, W. Liu, and F. Chen, “Discovering congestion propagation patterns in spatio-temporal

traffic data," IEEE Transactions on Big Data, pp. 1–1, 2016.

[4] Aggarwal and D. Toshniwal, "Spatio-temporal frequent itemset mining on web data," in 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018, pp. 1160–1165.

[5] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," ACM Computing Surveys (CSUR), vol. 51, no. 4, p. 83, 2018.

[6] Y. Shao, B. Liu, S. Wang, and G. Li, "A novel software defect prediction based on atomic class-association rule mining," Expert Systems with Applications, vol. 114, pp. 237–254, 2018.

[7] S. A. Aljawarneh, R. Vangipuram, V. K. Puligadda, and J. Vinjamuri, "Gspamine: An approach to discover temporal association patterns and trends in internet of things," Future Generation Computer Systems, vol. 74, pp. 430–443, 2017.

[8] C.-H. Chee, J. Jaafar, I. A. Aziz, M. H. Hasan, and W. Yeoh, "Algorithms for frequent itemset mining: a literature review," Artiicial Intelligence Review, pp. 1–19, 2018.

[9] M. Antonelli, P. Ducange, F. Marcelloni, and A. Segatori, "A novelassociative classification model based on a fuzzy frequent pattern mining algorithm," Expert Systems with Applications, vol. 42, no. 4, pp. 208– 2097, 2015.

[10] N. Aryabarzan, B. Minaei-Bidgoli, and M. Teshnehlab, "negfin: An efficient algorithm for fast mining frequent itemsets," Expert Systems with Applications, vol. 105, pp. 129–143, 2018.

[11] U. Turdukulov, A. O. Calderon Romero, O. Huisman, and V. Retsios, "Visual mining of moving flock patterns in large spatio-temporal data sets using a frequent pattern approach," International Journal of Geographical Information Science, vol. 28, no. 10, pp. 2013–2029, 2014.

[12] L. Szathmary, "Finding frequent closed itemsets with an extended version of the eclat algorithm," in Annales Mathematicae et Informaticae, vol. 48. Ekf Liceum Kiado Eszterhazy Ter 1, Eger, 3300, Hungary, 2018, pp. 75–82.

[13] Zhang, P. Tian, X. Zhang, Q. Liao, Z. L. Jiang, and X. Wang, "Hasheclat: an efficient frequent itemset algorithm," International Journal of Machine Learning and Cybernetics, pp. 1–14, 2019.

[14] S. Qin, F. Liu, C. Wang, Y. Song, and J. Qu, "Spatial-temporal analysis and projection of extreme particulate matter (pm10 and pm2. 5) levels using association rules: A case study of the jing-jin-ji region, china," Atmospheric Environment, vol. 120, pp. 339–350, 2015.

[15] S. Skakun, N. Kussul, A. Y. Shelestov, M. Lavreniuk, and O. Kussul, "Efficiency Assessment of Multitemporal C-Band Radarsat-2 Intensity and Landsat-8 Surface Reflectance Satellite Imagery for Crop Classification in Ukraine," IEEE J. of Select. Topics in Applied Earth Obser. and Rem. Sens., vol. 9, no. 8, pp. 3712-3719, 2016.