TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2936 - 2940

Optimized PM2.5 Predictive model using Time Series by Genetic Algorithm Based Long Short Term Memory Networks

S. Geetha, Assistant Professor (Sr. Grade), Department of Computer Applications, MepcoSchlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu, India.
K. ThurkaiMuthuraj, Project Manager, Tata Consultancy Services, Chennai, Tamil Nadu, India.

Article Info
Volume 82
Page Number: 2936 - 2940
Publication Issue:
January-February 2020

Article History
Article Received*: 14 March 2019*
Revised*: 27 May 2019*
Accepted*: 16 October 2019*
Publication*: 18 January 2020*

Abstract:
Air pollution prediction model makes available of reliable evidence to take preventive measures for safeguarding country from severe pollutions. Air pollutants concentrations are simulated with the parameters, such as $PM_{2.5}$, CO, NO, SO_2, PM_{10}, rainfall, temperature, air pressure, humidity, and so forth. The long short term memory network is a specific type of Recurrent Neural Network majorly used in times-series data prediction. The dataset is collected from real-time air quality monitoring system maintained by Central Pollution Control Board (CPCB) through nationwide programs. Although, the LSTM model performs better always, still the hyper-parameter selection optimization is done manually. To address this concern, Genetic Algorithm based LSTM is proposed to develop the automated optimization of hyper-parameters for the prediction model to the major air pollutant concentration $PM_{2.5}$. The optimized model showed the greatest performance than the conventional models.

Keywords:*Air pollution, Long Short-term Memory Networks, Hyper Parameter, Genetic Algorithm, $PM_{2.5}$..*

I. INTRODUCTION

In recent years, many countries are suffering due to various pollutions such as water, air and land pollution, etc. Out of that, air pollution is playing a vital role in creating lot of critical health issues to the public. The primary sources of air pollution have been identified such as natural sources (forest fire, volcanic eruption, etc.), climate change (extreme weather conditions, melting glaciers, etc.), ozone hole, anthropogenic sources (burning of fuels, industry smoke, transport emission, etc.) and particulate matter pollution.Out of these, one of the primary air pollution concentration is Particulate Matter 2.5 ($PM_{2.5}$) with smog creates lot of health issues for living beings. $PM_{2.5}$ is particulate matter with 2.5 micrometre of diameter or less than. The human health is affected immenselyand lead to mortality due to long time exposure of $PM_{2.5}$[1,2]. World Health Organization states the reason as air pollution for approximately seven million mortalities on every year from stroke, lung diseases, heart disease, and respiratory problems [3,4,5]. Prediction of $PM_{2.5}$ will help the public administration to take preventive action to reduce the level of air pollution. Hence, the real-time monitoring and early prediction of PM2.5 becomes imperative. Previously, various forecasting approaches such as ARIMA [6], Multi-linear regression [7], Artificial Neural Network [8], shallow machine learning [9] and deep learning [10] were applied on the air pollution dataset and prediction of result. The deep learning becomes inevitable information technology which supports machines to learn. The special type of deep learning network under the recurrent neural network is LSTM which supports to generate model with time series data. As the air quality dataset consists of time series data, the LSTM can be used for prediction. Although, lot of models were already developed using LSTM [11] and provides better results, the

January - February 2020
ISSN: 0193 - 4120 Page No. 2936 - 2940

hyper parameter selection to be optimized manually in LSTM network. Manual selection of hyper parameters is not efficient as well as more time consuming. To overcome these problems, the hyper parameter selection to be optimized using optimization algorithms. Lot of optimization algorithms like Particle Swarm Optimization [12], Gradient Descent Algorithm [13], Genetic Algorithms [14], etc. are available and used for many optimization problems. In the proposed work, the Genetic Algorithm is chosen to optimize the hyper parameter selection in LSTM model which provides efficient model with less time consumption.

II. LSTM NETWORK

The LSTM Network with Recurrent Neural Network which supports to memorize information in current state for future use. The all LSTM neurons are memory cells. Each memory cell contains three gates. Figure 1. shows the LSTM Network.

Input Gate:

This gate decides about what new information and how much it should be remembered by the memory cell. The following equations help to the input gate to decide it (1, 2, 3).

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (1)$$

$$\widetilde{C_t} = \tanh(W_C . [h_{t-1}, x_t] + b_C) \quad (2)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (3)$$

Forget Gate:

This gate decides which information to be discarded and which information to be remembered and whether the information to be discarded completely or remembered completely. It is decided with the help of the following equation (4).

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (4)$$

Output Gate:

This gate decides which information to be delivered from current state to next state. The neuron state is activated and multiplied which results the output information at the time t. It is decided with the help of the following equations (5, 6).

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

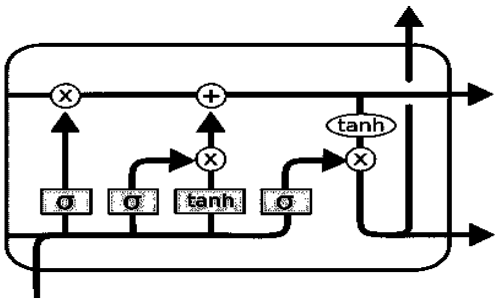


Figure 1. LSTM Network Neuron

Genetic Algorithm

Genetic algorithm is a heuristic approach which works on the basis of nature [15]. Genetic Algorithm is used widely to find near-optimal solutions to the problems with large search spaces. The GA includes mutation, crossover operators for imitating natural genetic principles. GA process includes various stages such as initialization, fitness calculation, checking termination condition, selection, mutation and crossover. Figure 2. shows the basic procedure of Genetic Algorithm.

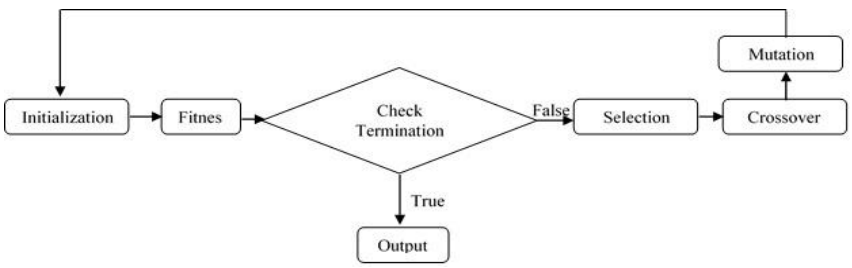


Figure 2. Process of Genetic Algorithm

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2936 - 2940

III. PRE-PROCESSING

Selection

The dataset contains various data fields like SO_2, CO, 0_3, PM_{10}, $PM_{2.5}$, NO, NO_2, Temperature, Humidity, Wind Speed, Wind Direction, etc. In this paper, only $PM_{2.5}$ is concentrated for prediction. Along with $PM_{2.5}$, the time factor also considered. Hourly $PM_{2.5}$data was taken for analysis and further prediction.

Imputation

The time series air quality data from the CPCB is collected for this work. Many pollutants such as SO_2, NOx, O_3, CO, $PM_{2.5}$, PM_{10} and NO_2 are above the limits specified by the Central Pollution Control Board (CPCB). The daily average air pollutant data is collected from 2015 – 2018. Dataset is with missing values. The missing values are attributed using average of data fields.

Normalization

The scale of $PM_{2.5}$ data is varying over the time period. It may take more time to arrive at accuracy of prediction. Hence the normalization is applied on the data to get the better prediction within specified time. Here, the Z-score normalization (7) is used to normalize the data to limit the value from 0 to 1.

$$v' = \frac{v - \bar{A}}{\sigma_A} \qquad (7)$$

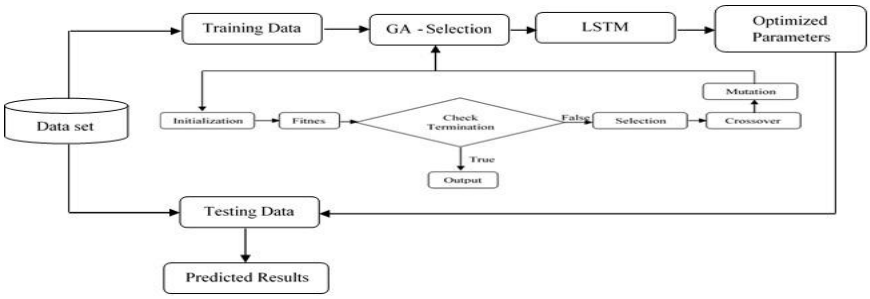


Figure 3. LSTM with GA

IV. EXPERIMENT RESULTS

The forecasting model for air quality is developed and trained with the CBCS data of Delhi. The dataset is imputed with missing values and normalized before using it for training. The next process is dividing the air quality data into train and test data. The dataset contains $PM_{2.5}$ data for 2015 - 2018. The LSTM model is developed with various hyper parameters and optimized Genetic algorithm to select the hyper parameters. The batch size is varying from 20 – 25, look back varies from 1 – 7, no. of epochs varies from 50 – 200. The model achieves prediction accuracy with selection of GA based hyper parameters. The model is developed and trained with training data and tested with testing data. The model is evaluated with various evaluation measures. The Figure 4. shows the train and test loss for $PM_{2.5}$. The Figure 5. shows the prediction or $PM_{2.5}$ by normal LSTM model and Figure.6 shows $PM_{2.5}$predictions with GA and LSTM.

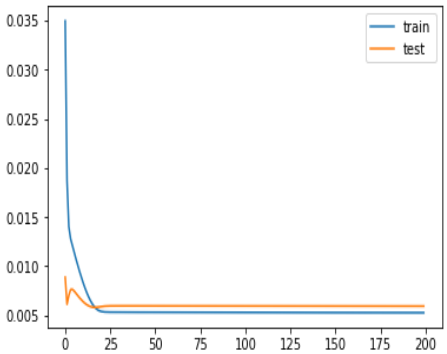
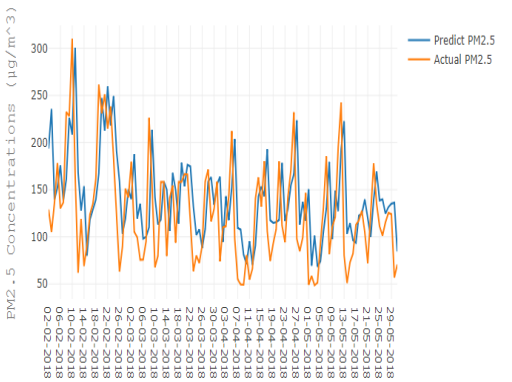


Figure 4. Train and Test Loss

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 2936 - 2940

Figure 5. Prediction using LSTM



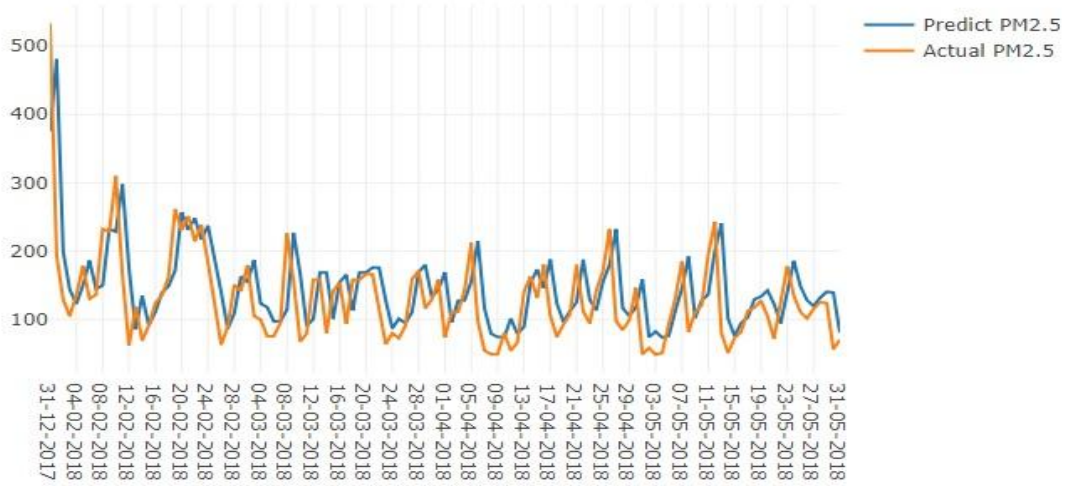


Figure 6. Prediction using LSTM with GA

V. Model Evaluation

Generally, various evaluation metrics are available to measure the performance of the statistical and deep learning models. The developed model evaluated with evaluation measures MSE, RMSE and MAE. It expands as Mean Square Error, Root Mean Square Error and Mean Absolute Error. It is used to differentiate normal LSTM model with GA based LSTM Model. Equations for MSE, RMSE and MAE are as follows.

$$MSE = \frac{1}{n}\sum_{t=0}^{n}(y_t - \tilde{y}_t)^2 \quad (7)$$

$$RMSE = SQRT(\frac{1}{n}\sum_{t=0}^{n}(y_t - \tilde{y}_t)^2) \quad (8)$$

where y_t denotes actual value, $\tilde{y}_t$ denotes predicted value and n denotes total number of observations.

$$MAE = \frac{1}{n}\sum_{t=0}^{n}|y_t - \tilde{y}_t| \quad (9)$$

where n denotes total number of predictions, y denotes predicted value.

Model Comparisons

The models are compared based on the evaluation metrics MSE, RMSE and MAE. The following Table 1. shows result of evaluation metrics for normal LSTM model and the model developed with genetic algorithm and LSTM. Out of the result, mean square error and mean absolute error are not showing much difference in both training and testing phase. The root means square error is reduced in the model which was developed with genetic algorithm when compared to normal LSTM model.

Table 1: Model Comparison

Phase	Evaluation	LSTM	GA with LSTM
Training	MSE	0.005	**0.003**
	RMSE	0.071	**0.059**

January - February 2020
ISSN: 0193 - 4120 Page No. 2936 - 2940

	MAE	0.005	**0.003**
Testing	MSE	0.005	**0.003**
	RMSE	0.068	**0.056**
	MAE	0.005	**0.003**

VI. CONCLUSION

The proposed model achieves accurate prediction using Long Short Term Memory Network along with Genetic Algorithm. Although, the LSTM shows better prediction, when the hyper parameters are selected with the help of genetic algorithm, still the accuracy improved. This model shows the predictions more accurately than normal LSTM model. Air quality is not only depending on $PM_{2.5}$, many other attributes also contributing. Still the major pollutant is $PM_{2.5}$. So, this study concentrates only $PM_{2.5}$ predictions. Though, the model is trained with only $PM_{2.5}$data collected from Delhi. Hence, further training can be extended to make effective prediction for all air quality data.

VII. REFERENCES

1. Beelen R., Raaschounielsen O., Stafoggia M., Andersen Z.J., Weinmayr G., Hoffmann B., Wolf K., Samoli E., Fischer P., NieuwenhuijsenM.,"Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project", The Lancet, 383, 785-795, 2014.
2. X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions", Environmental Science and Pollution Research, vol. 23, no. 22, pp. 22408–22417, 2016.
3. Lubinski W., Toczyska I., Chcialowski A., PlusaT.,"Influence of air pollution on pulmonary function in healthy young men from different regions of Poland",Annals of Agricultural and Environmental Medicine,12,1-4, 2005.
4. Chahine T., Baccarelli A., Litonjua A., Wright R.O., Suh H., Gold D.R., Sparrow D., Vokonas P., Schwartz J.,"Particulate air pollution, oxidative stress genes, and heart rate variability in an elderly cohort", Environmental Health Perspectives, 115, 1617–1622, 2007.
5. Sundeep Mishra, "Is smog innocuous? Air pollution and cardiovascular disease", Indian Heart Journal, 2017.
6. Khashei, M., Bijari, M, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting", Applied Soft Computing 11(2), 2664–2675,2011.
7. Kurt, B. Gulbagci, F. Karaca, O. Alagha, "An online air pollution forecasting system using neural networks", Environment International, 34, pp. 592-598, 2008.
8. Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., "A neural network forecast for daily average PM10 concentrations in Belgium", Atmospheric Environment, 39, 3279e3289, 2005.
9. Hsieh, W. W., "Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels", Cambridge University Press, 2009.
10. I. Kok, M. U. Simsek, and S. Ozdemir, "A deep learning model for air quality prediction in smart cities", IEEE International Conference on Big Data (Big Data), pp. 1983–1990, 2017
11. S. Geetha, L.Prasika, "Smog Prediction Model using Time Series with Long-Short Term Memory", International Journal of Mechanical Engineering and Technology (IJMET), Volume 10, Issue 01, pp. 1026–1032, 2019.
12. Zapata-Hernandez J.C., Rojas-Idarraga Y.K., Orrego D.A., Murillo-Escobar J,"Prediction of Critical Air Quality Events Using Support Vector Machines and Particle Swarm Optimization",In: Torres I., Bustamante J., Sierra D. (eds) VII Latin American Congress on Biomedical Engineering CLAIB 2016, IFMBE Proceedings, vol 60. Springer, Singapore, 2017.
13. Dixian Zhu, ChangjieCai, TianbaoYangandXun Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization", Big Data and Cognitive Computing, 2018.
14. Mahmoud Reza Delavar, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra and SmaeilHatefiAfshar, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches:A Case Study Applied to the Capital City of Tehran", ISPRS Int. J. Geo-Information, 2019.
15. Majidnezhad, V,"A novel hybrid method for vocal fold pathology diagnosis based on Russianlanguage",Journal of AI and Data Mining, Shahroud university, vol. 2, no. 2,pp.141-147, 2014.