International Journal of Ecology and Environmental Sciences
www.ecologyjournal.in
Online ISSN: 2664-7133; Print ISSN: 2664-7125; Impact Factor: RJIF 5.28
Received: 15-07-2020; Accepted: 30-07-2020; ; Published: 16-08-2020
Volume 2; Issue 3; 2020; Page No. 162-165

Predicting with Spatio-temporal analysis of air quality data using regression analysis

Sowmya N[1], Varalakshmi N[2], Yamuna Shree A[3], Pankaja K[4]
[1-3] Computer Science Department, Visvesvarya Technological University, Bengaluru, Karnataka, India
[4] Associate Professor, Computer Science Department, Cambridge Institute of Technology, Bengaluru, Karnataka, India

Abstract
Air pollution has become an extremely serious problem, with particulate matter having significantly greater impact on human health than other contaminants. The small diameter of fine particulate matter (PM2.5) allows it to penetrate deep into the alveoli as far as the bronchioles, interfering with gas exchange within the lungs. Forecasting air quality has also become important. This study aims to forecast air quality using a combination of multiple neural networks and LSTM to extract spatial-temporal relations. The proposed predictive model considers various meteorology data information related to the elevation space to extract terrain impact on air quality. The model includes trends from multiple locations, extracted from correlations between adjacent locations, and among similar locations in the temporal domain. We also predict the PM2.5 values using regression model in this project. Experiments employing Beijing datasets show that the proposed model achieves excellent performance and outperforms current state-of-the art methods.

Keywords: convolutional neural network (CNN), long-short-term memory (LSTM), spatio-temporal analysis, big data, air quality forecast

1. Introduction
Increasing attention has been given to air quality degeneration, with particulate matter (PM) having a significant egregious impact on human health. The small diameter of fine particulate matter (PM2.5) allows it to penetrate deep into the alveoli as far as the bronchioles, interfering with gas exchange within the lungs. Data mining provides new methods to analyze air quality in the absence of physical models, and may identify hidden information in the collected data. we propose a model to provide air quality index (AQI) predictions every hour at every monitoring location. We forecast predictions from the using historical data. This study proposes a general predictive model for air quality forecasts called spatial-temporal deep neural network (ST-DNN) that incorporates various information from monitoring locations, including PM2.5, PM10, temperature, wind speed, wind direction, average wind speed, average wind direction, relative humidity, and data related to the elevation space. We first found the most relevant spatial-temporal relations among locations, then combined multiple neural network architectures using a convolutional neural network (CNN) and long short-term memory (LSTM). Target and similar location spatial-temporal features were used to increase the predictive model sensitivity and explicitly consider terrain impacts for pollutant propagation. Thus, the proposed model uses (i) temporal information based on target location historical data, (ii) spatial relationships based on related locations' data, i.e., locations with high spatial or temporal similarity. We propose a framework to mine spatial-temporal data for a given location to provide a predictive model. We develop a deep learning model combining multiple neural networks to incorporate air quality correlations among similar locations and temporal dependency at a given location. Spatial and temporal predictions are combined dynamically based on the trained neural network.

2. Proposed System
The proposed ST-DNN model combines target location temporal information, and related location spatial-temporal and terrain information. The data flow includes target and related location historical data, i.e., pollutants, meteorological conditions, and target features and their trends over the previous few hours. These input to LSTM and NN. Air quality and meteorological condition data sources are input to LSTM, and terrain related data are input to NN. The models are merged via side by side concatenation, and the variables are passed to the following layer. The model is trained hourly over the subsequent 48 hours, since the current status varies with respect to its effect on future time intervals. Thus, we pair the inputs with target feature deviations in the various time intervals to train multiple models with the same structure corresponding to the different time intervals. The advantage of this structure is that the input sizes are constant, regardless of the location and time interval. We apply regression model to predict the Air quality with PM2.5 values.

3. Impact of Air Quality in China
A 2012 study stated that pollution had little effect on economic growth which in china's case was largely dependent on physical capital expansion and increased energy consumption due to the dependency on manufacturing and heavy industries. China was predicted to continue to grow using energy inefficient and polluting industries.
A 2013 study published in Processing of the National Academy of Science found that severe pollution during 1990's cut 5.5 years from the average life expectancy of people living in northern China, where toxic air has led to increased rates of stroke, heart diseases and cancer.

A 2015 study from the non-profit organization estimated that 1.6 million people in china die each year from heart, lung and stroke problems because of polluted air.

4. Investigation of the Air Pollution Event in Beijing

Beijing-Tianjin Hebei is a super urban agglomeration on the North China Plain, with an area of 21,600 kilometers and more than100 million people. BTH has become one of the most heavily polluted areas in the world because of extensive industrialization and urbanization in the past few decades. High concentrations of fine particulates (PM2.5, with aero dynamic diameters not larger than 2.5μm) substantially reduce visibility through scattering, and they penetrate the human respiratory track. Therefore, numerous studies in recent years have focused on the formation of air pollution episodes in BTH. The air pollution is composed of multiple pollutants, which are mainly attributed to power plants, domestic pollutants, industry, and vehicle exhaust**.**

5. Methodology

Artificial Neural Networks are computing systems vaguely inspired by the biological neural network that constitute animal brain. A neural network is a network or circuit of neurons, for solving AI problems. These artificial networks are used for predictive modeling, adaptive control and applications where they can be trained via datasets.

Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. LSTM has feedforward connections. It can only process single data points, but also entire sequences of data. For example, LSTM is applicable to task such as unsegmented, connected handwriting recognition, speech recognition etc.,

LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cells remember's values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM network are well suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. Fig 1: shows the basic LSTM cell.

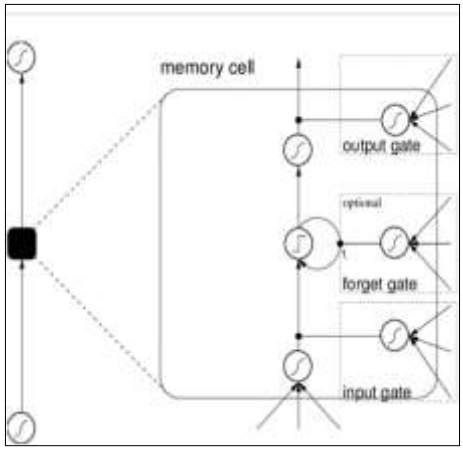

Fig 1: Basic LSTM Cell

The System architecture is shown below. System architecture is the conceptual design that defines the structure and behavior of a system. An architecture description is a formal description of a system, organized in a way that supports reasoning about the structural properties of the system. It defines the system components or building blocks and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system.

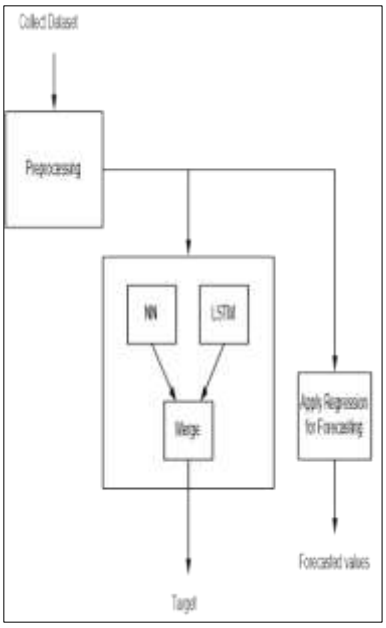

Fig 2: System architecture

5.1. Data flow diagram

The Level 1 DFD shows how the system is divided into sub-systems (processes), each of which deals with one or more of the data flows to or from an external agent, and which together provide all of the functionality of the system as a whole. It also identifies internal data stores that must be present in order for the system to do its job, and shows the flow of data between the various parts of the system.

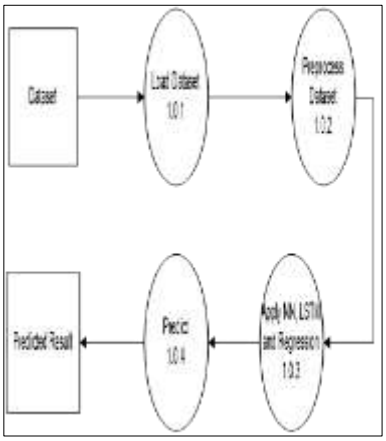

Fig 3: Data- flow daigram

5.2. Sequence Diagram

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Chart. The sequence diagrams show below.

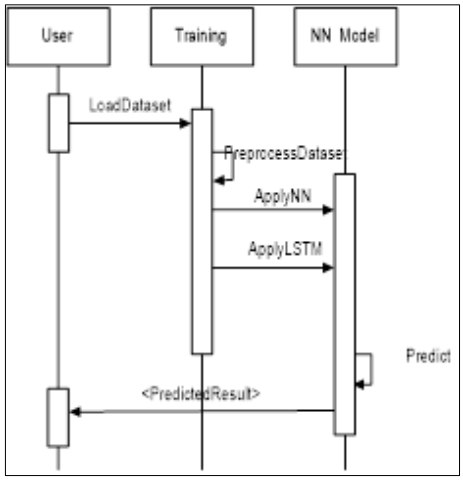

Fig 4: Sequence Diagram

6. Algorithm

Linear Regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear Regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. If the goal is prediction, or error reduction, linear regression can be used to fit a predictive model to an observed data set values of the response and explanatory variables. After developing such model, if additional values of the explanatory variable are collected without an accompanying response value, the fitted model can be used to make prediction of the response.

6.1. Regression Algorithm

1. Draw the scatterplot. Look for linear or non-linear pattern of the data and deviations from the pattern (outliers). If the pattern is non-linear, consider a transformation. If there are outliers, you may consider removing them only IF there is a non-statistical reason to do so. (Are those individuals "different" than the rest of the sampled individuals?)
2. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot (for constant standard deviation assumption) and normal probability plot (for normality assumption). If the assumptions of the model appear not to be met, a transformation may be necessary.
3. If necessary, transform the data and re-fit the least-squares regression line using the transformed data.
4. If a transformation was done, go back to step 1. Otherwise, proceed to step 5.
5. Once a "good-fitting" model is determined, write the equation of the least-squares regression line. Include the standard errors of the estimates, the estimate of σ, and R-squared.

Linear regression equation looks like this

$$Y = \theta_1 X_1 + \theta_2 X_2 + \ldots \theta_n X_n$$

Here, we have Y as our dependent variable (Sales), X's are the independent variables and all thetas are the coefficients. Coefficients are basically the weights assigned to the features, based on their importance. R-Square: It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). Mathematically, it can be written as:

$$R-Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})^2}{\sum(Y_{actual} - Y_{mean})^2}$$

The value of R-square is always between 0 and 1, where 0 means that the model does not model explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.

6.2. Geographical Relationship

Input: Target station $\not{\wedge}$i; Set of Locations' coordinate Lc, Where **li ∈ Lc**;
Output: Set of Locations by SRS_cand (li, k);
Let SRS_cand ←∅; for each lc ∈ Lc do
Calculate distances between li and lc:
SRS_cand ∪ {lc, ED (li, lc)};
end
Sort SRS_cand by ED (li, lc);
if k ≤ Size of SRS_cand then
SRS_cand(li, k) ←first kth of SRS_cand
end
else
SRS_cand(li, k) ←SRS_cand
End

7. Results

Below figures explains the graph for predicting air quality.
Figure 5. is used to indicate each and every attribute values with spatiotemporal graphs on each attribute. Spatial refers to space. Temporal refers to time. It describes a phenomenon in certain location and time

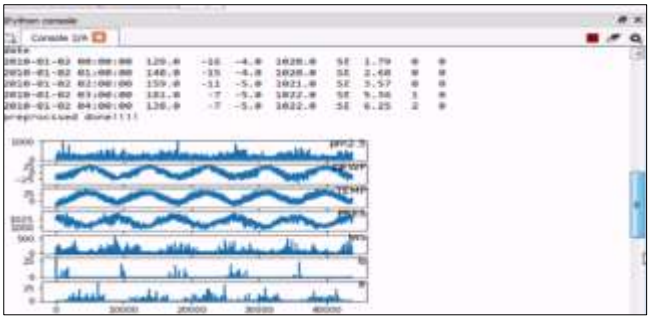

Fig 5: Displaying attribute of ST relation

Figure 6. is used to display root mean square error comparison for test and train dataset. RMSE is the standard deviation of the

residuals. Residuals are a measure of how far from the regression line data points are; In other words, it tells us how concentrated the data is around the line of best fit.

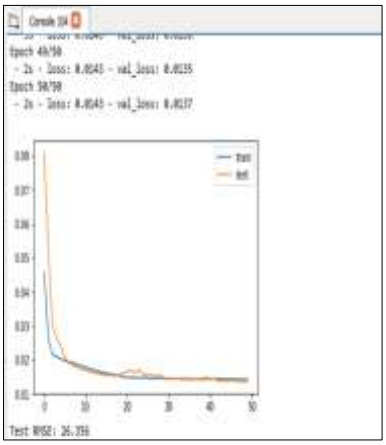

Fig 6: RMSE comparison Graph

Figure 7. shows the graph of mean absolute error comparison in LSTM model. MAE is a measure of the errors between paired observations expressing the same phenomenon. It is known as scale dependent accuracy measure and therefore cannot be used to make comparisons between series using different scales

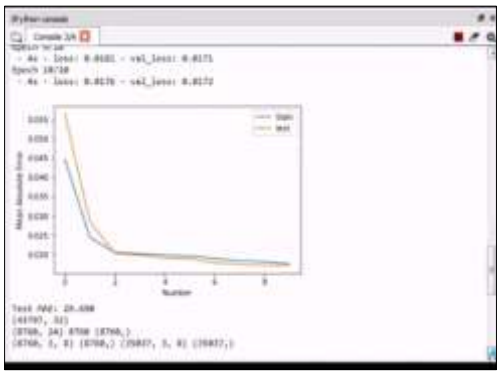

Fig 7: MEA Comparison in LSTM

Figure 8. Explains the logistics regression used to describe data and explains the relationship between one dependent variable.

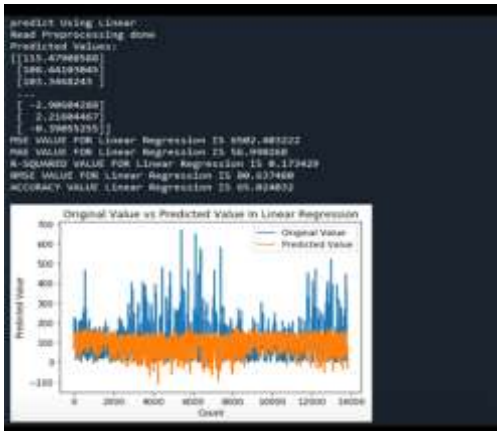

Fig 8: Logistic Regression Graph

8. Conclusion

This Paper proposed another procedure for forecasting system using data driven models, ST-DNN, to predict PM2.5 over 48 hours. The proposed method is also generally applicable to other pollutants, etc. The proposed ST-DNN shows that including an LSTM module enhanced first hour predictions, with NN module inclusion being more useful for longer time frame predictions, since NN can extract the temporal delay factor from surrounding target features by learning spatial information. We evaluated the proposed models using real-world Taiwan and Beijing datasets. Relevant location selection was verified to be important, with-in a inclusion of all locations causing increased model noise and hence poorer prediction performance. We applied Regression approach to predict the quality of air with PM2.5 values. The proposed methods outperformed all baselines and comparative models considered.

9. References

1. Ch M Kuo HJ, ‘‘Identifying spatial mixture and distributions of PM2.5 and PM10 in Taiwan during and after a dust storm,’’ Atmos. Environ, 2012; 54:728-737.
2. Kurt A, Oktay AB. ‘‘Forecasting air pollut indicator levels with geographic models 3days in advance using neural networks,’’ Expert Syst. Appl. 2010; 37(12):7986-7992.
3. Chen LJ, Ho YH, Hsieh HH, Huang ST, Lee HC, Mahajan S, *et al.* ‘‘ADF: An anomaly detection framework for large-scale PM2.5 sensing systems,’’ IEEE Internet Things J. 2018; 52(2):559-570.
4. Chang HL. ‘‘Evaluation and application of the short-range (0-6hr) pqpfs from an ensemble prediction system based on laps,’’ Ph. D dissertation, Graduate Inst. Atmos. Phys., Nat. Central Univ., Taoyuan, Taiwan, 2014.
5. Houdan WM, Barnard WR. ‘‘Evaluating the contribution of PM2.5 precursor gases and re-entrained road emissions to mobile source PM2.5 particulate matter emissions,’’ MACTEC Federal Programs, Research Triangle Park, NC, USA, 2004.
6. Qin S, Liu F, Wang C, Song Y, Qu J. ‘‘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,’’ Atmos. Environ, 2015; 120:339-350.
7. Liu CM, Young CY, Lee YC. ‘‘Influence of Asian dust storms on air quality in Taiwan,’’ Sci. Total Environ. 2006; 368(2-3):884-897.
8. Rakthanmanon T, *et al.* ‘‘Searching and mining trillions of time series subsequences underdynamictimewarping,’’ inProc.18thACMSIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, 262-270.