A deep learning method for air quality prediction model using the most relevant spatial-temporal relations

A SEMINAR REPORT

Submitted by

Gowri Nair

MBT5CS047

In partial fulfilment of the requirements

for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Of the A.P. J. Abdul Kalam Technological University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MAR BASELIOS COLLEGE OF ENGINEERING AND TECHNOLOGY

MAR IVANIOS VIDYANAGAR, NALANCHIRA

THIRUVANANTHAPURAM – 695 015

November 2018

Mar Baselios College of Engineering & Technology

Mar Ivanious Vidyanagar, Nalanchira,

Thiruvananthapuram - 695 015

(Affiliated to A. P. J. Abdul Kalam Technological University)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Seminar Report titled ‘ *A deep learning method for air quality prediction model using the most relevant spatial-temporal relations* ‘ is a bonafide record of seminar presented by Gowri Nair (MBT15CS047).

Mrs. Asha S	**Dr. Vishnukumar S**	**Dr. Tessy Mathew**
Seminar Co-ordinator	Seminar Guide	Head of the Department

Place: Thiruvananthapuram

Date: 26 /11 /2018

ACKNOWLEDGEMENT

First and foremost, I thank God Almighty for His divine grace and blessings in making all these possible. May He continue to lead us in the years. It is to render my heartfelt thanks and gratitude to our principal**, Dr. T. M. George** for providing the opportunity to do this seminar during the 7th semester (2018) of my B. Tech degree course.

I am deeply thankful for our Head of the Department **, Dr.Tessy Mathew** for her support and encouragement. I would like to express my sincere gratitude to my seminar guide Asst.Professor, **Dr.Vishnukumar S**, and project coordinator, Asst.Professor, **Mrs. Asha S,** DepartmentComputer Science and Engineering for their motivation, assistance and help for the seminar. I also thank all the staff members of the Computer Science Department for providing their assistance and support. Last, but not the least, I thank all my friends and family for their valuable feedback from time to time as well their help and encouragement.

Gowri Nair

CONTENTS

List of Figures

Abstract

Particulate Matter has a great impact on human health than other contaminants making air pollution a very serious problem .The fine particulate matter (PM2.5) has very small diameter which allows it to penetrate deep into the alveoli as far as the bronchioles,. Particulate matter, on long-term exposure, interferes with gas exchange with the lungs, thereby causing cardiovascular diseases, respiratory diseases, and increase the risk of lung cancers. For this reason, forecasting air quality has become important in helping guide individuals' actions. In this model, up to 48 hours of air quality is forecasted, using a combination of multiple neural networks, including a convolutional neural network, an artificial neural network, and a long-short-term memory to extract spatial-temporal relations. Various historic meteorology data is considered in this predictive model and also, elevation space related information extract the impact of terrain on air quality. The model includes data from different locations, extracted from correlations between adjacent locations, and among similar locations in the temporal domain.

1. INTRODUCTION

Particulate Matter (PM) has a consequential impact on human health and hence more attention should be given to air quality deterioration. The fine particulate matter (PM2.5) has very small diameter which allows it to penetrate deep into the alveoli as far as the bronchioles,. Particulate matter, on long-term exposure, interferes with gas exchange with the lungs, thereby causing cardiovascular disease, respiratory disease, and increase the risk of lung cancers. Many cities have installed air quality monitoring locations with increase in public health awareness. But the problem is that all these methods only shows the current air quality and do not forecast it. Air quality prediction helps in guiding people to limit the exposure to PM2.5, for example, choosing between indoor and outdoor activities.

Accurate air quality prediction depends on a complex array of factors including emissions, traffic patterns and also meteorological factors. Meteorologists are still not able to correctly predict the wind pattern as they continuously change in both strength and direction. Moreover there are no sufficient sensors to monitor the emission from factories and vehicles. The cyclicality of particulate matter is high and it can stagnate and diffuse, polluting the surrounding environment .If PM is analysed only in the time domain, it may not consider this factor of diffusion from the surrounding environment. But if it is considered only in the spatial domain, it may not consider diffusion of PM over time. Hence temporal and spatial domain must be considered to accurately predict the air quality.

Data mining can analyse the air quality using new methods, if physical methods are not available. Data mining also considers any hidden information in the collected data. If the model is trained, it predicts the air quality more accurately and faster, compared to any physical model. Therefore, the next 48 hours of Air Quality Index (AQI) is predicted by this proposed model, and the AQI is predicted at every monitoring locations, at every hour.

This report is on a general predictive model called spatial-temporal deep neural network, *S*T-DNN. It includes various information from the monitoring locations, such as relative humidity, data related to elevation space, wind direction, average wind direction, wind speed and average wind speed. Current and previous few hours of data is used to train the model and also the meteorological conditioned data. Relevant data based on geographical and temporal correlations among monitoring locations s incorporated in this model. The most relevant spatial-temporal relations among locations are found first and multiple neural network architectures are combined using convolutional neural network.

This model thus uses,

i) temporal information based on target location's historic data,

ii) related locations' spatial relation data i.e., locations with high temporal or spatial similarity and

iii) terrain information of the surrounding locations..

The main contributions of this model are:

1. A framework to mine spatial-temporal relation data from a location to provide a predictive model.
2. A deep learning method is used to combine multiple neural networks to incorporate air quality correlation among similar locations and temporal dependency on a given location.
3. Based on trained neural network, temporal and spatial predictions are dynamically combined.

2. PROBLEM DEFINITION

Locations that have the most influential spatial-temporal relationships with the target locations are identified and then sequences for the target location based on time sequence features, which includes spatial information are predicted. Hence, according to the positions, time sequences vary, but locations are fixed. The sequences are also impacted by the spatial features, such as mountain between two regions, etc. The related spatial and temporal relationship parameters are defined first to extract the features for prediction.

Let set of locations, ***L={L1,L2,L3,..Ln}*** and set of features ***F={F1,F2,F3… Fm}***

Each location has

Definition 1**: Location coordinates*, Lc_i= (Li,xi,yi), Li*** $\in$***L*** , where xi and yi are latitude and longitude at location Li.

The distance between two location coordinates are defined since related locations could improve predictions as,

$Dsq,c = dist_{loc}\ (Lcq, Lcc)$

$= dist_{loc}\ ((Lq, xq, yq),(Lc, xc, yc)),$

Lq, Lc $\in$***L, q6=c.*** to find the most closely related locations in the spatial domain.

Definition 2: Spatial Relations Sequence Set : SRS =$\{Ds_{1,2},Ds_{1,3},...,Dsn_{-1},n\}$***,***

$Dsi, i =0, 0 < i < n+1$

The most relevant locations are SRS_cand(li,k), the set of k locations with the smallest spatial distance to li.

Definition 3 Feature Sequence Interval with Location. :

$S(Li, fj,t_{st}, ft) =\{e(Li, fj, t_{st}),e(Li,fj,t_{st+1}),..., e(Li, fj, tft)\},$

Li $\in$***L,*** ***fj*** $\in$***F, st*** $<$ ***ft.***

Distance between feature sequences of any two locations can be expressed as:

$D_{tq,c,tst,ft} = dist_{sequence}(S(Lq, ftarget, t_{st}, ft),S(Lc, ftarget, t_{st}, ft)),$

Lq, Lc $\in$***L, q6=c.***

to obtain the most related locations in the temporal domain. This report chooses PM2.5 as the feature for the target sequence prediction.

.

Definition 4: Temporal Relations Sequence Set.:

$$TRStst,\ ft = \{D_{t1,2}, t_{st}, ft, D_{t1,3}, t_{st}, ft, \ldots, D_{tn-1,n}, t_{st}, ft\}$$

Candidates TRS_cand(Li, k), the set of k locations with the smallest distance from location Li are selected.

Definition 5: Spatial-Temporal Relations Set.:

$$STRS\_cand(Li, k) = SRS\_cand(Li, k) \cup TRS\_cand(Li, k)$$

$$, Li \in L.$$

The union SRS_cand(Li, k) and TRS_cand(Li, k) is used, as it provides a larger number of relationships for the model to learn.

Definition 6: Spatial-Temporal Sequence Prediction : $M(STRS\_cand(Li, k))[t_{tlb}, t_c]$

$$= S(Li, ftarget, t_{st0}, ft0),$$

$$t_{lb} < t_c < st0 \leqslant ft0.$$

To build the model M to predict a target feature sequence, where M returns a sequence set, S, of the target features for the period t_{st0} to t_{ft0}.S is generated from the most similar time series compared with t_{lb} to t_c, where t_{lb} is the look back time.

3. PREDICTION MODEL FRAMEWORK

Spatial-temporal analysis is performed to find the most relationships between locations, which are the most relevant, to find the sequence delays and interactions between locations using historical data and it also considers location features. Adjacent locations or location with similar features are considered as they have high correlation with the target locations. Training datasets from the top k related locations are created by the kNN-ED and kNN-DTWD relationship extractors when the processed data is sourced into the system.

Figure 3.1 shows the proposed prediction model framework model. It consists of 4 main parts.

1. Features of the air quality of target location's meteorological data over the previous few hours is sourced in to the temporal relationships extractor (TRE) and it uses LSTM model.
2. Related locations' air quality feature data, selected by kNN-ED or kNN-DTWD is fed into the spatial-temporal relationships extractor (SRE), which uses ANN model.
3. In the Terrain Extractor (TE), a CNN extracts interactions between terrain and air quality features from the terrain information in the vicinity of the target location.
4. In the merge layer, an STP i.e., a spatial-temporal predictor is used to combine the discrete outcomes of the components.

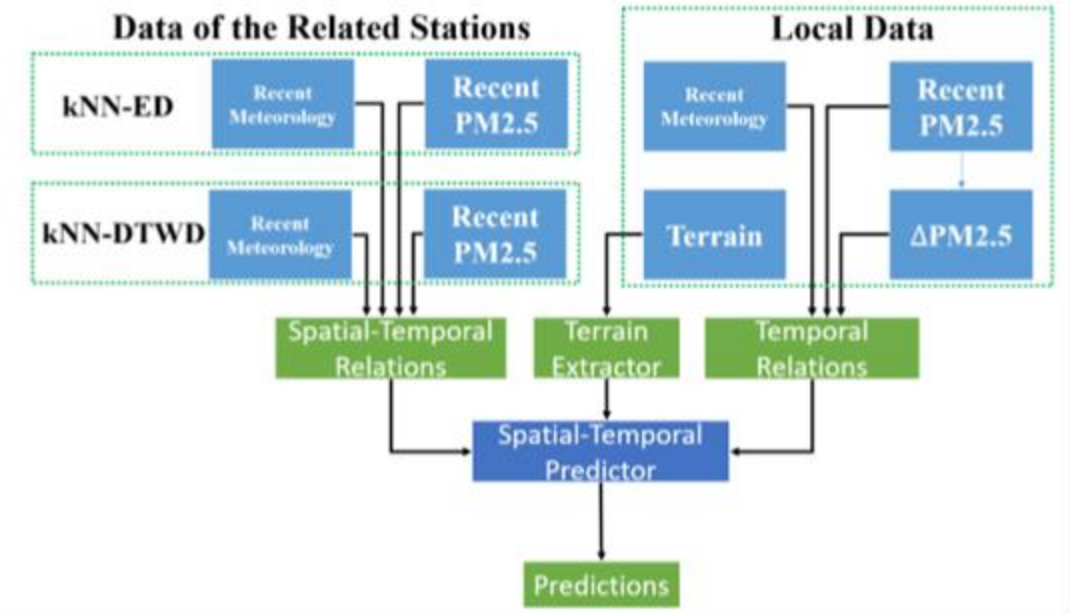

FIGURE 3.1.Prediction model framework

3.A. Mining Spatial-Temporal Relationships From Related Locations

3.A.1. k-Nearest Neighbour by Euclidean Distance (kNN-ED)

Algorithm 1 Geographical Relationship Set Generator (kNN-ED)

Input: Target location Li; Set of Locations coordinates Lc, where li / ∈Lc;
Number of candidates k;

Output: Set of Locations by SRS_cand(Li, k);

Let SRS_cand ←∅; **for** each Lc∈Lc **do**

Calculate distances between Li and Lc: ED(Li, 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

3.A.2. k-Nearest Neighbor By DTW Distance (knn-DTWD)

The distance between two sequences are calculated by DTW by minimizing the errors in shifting and scaling between the sequences. Although the sequences are dissimilar using Euclidean distance, DTW can restore sequence distortions by mapping the data points to corresponding intervals. Thus, DTW identified the most strongly related temporal relationships to the target location and calculated the time series feature distances between locations. The distances were sorted and the top k most similar locations chosen as candidates to predict the target location sequence. This method is refered as kNN-DTWD

Algorithm 2 Temporal Similarity Set Generator (kNN-DTWD)

Input: Target station li; Set of Locations' L, where

Li / ∈L;

Number of candidates k; Target feature ftarget;

Time Interval t_{st},ft;

Output: Set of Locations by TRS_cand(Li, k);

Let TRS_cand ←∅; **for** each l ∈L do

Calculate similarity between li and l;

$dist_{dtw}$← DTWDsim(S(Li, ftarget, t_{st}, ft), S(Lj, ftarget, t_{st}, ft), Lmin);

TRS_cand ∪{l, $dist_{dtw}$}; **end** Sort TRS_cand by $dist_{dtw}$;

if k≤Size of TRS_cand then

TRS_cand(Li, k)←first kth of TRS_cand

end

else TRS_cand(Li, k)←TRS_cand

end

When two time series intervals have non-missing values simultaneously and the common interval length exceeds lmin (i.e., la > lmin,)then the common interval the interval Is included for DTW similarity. In contrast, lb <lmin , and hence is ignored in the DTW calculation. Although the ignored cause some loss of information, but this method effectively removes most noise related errors. The average unit distance is defined as d1 =Pn 1 di Pn 1 li where di is the distance in a common interval, and li is the length of that interval; which combines multiple fragment sequences to facilitate overall similarity identification.

3.B. Prediction Model Design

The ST-DNN model combines the target location temporal information, and related location spatial-temporal and terrain information. The previous few hours data of the target and related locations is the data flow in the model, i.e., the meteorological conditions, elevation, etc. This data is input to ANN, LSTM, and Adaptive Temporal Extractor (ASE) and a matrix of 121 square sections with 11x11 coordinate lines at 500m intervals, of terrain data, where the local location is represented at the grid's center.

The minimum time interval for meteorological forecasts, lmin is set to 6 hours. Without pre-training, data consisting of meteorological conditions, pollutants, and target features of similar locations were sourced into the LSTM and ASE.

Air quality and meteorological condition data sources are input to LSTM and ASE, and terrain related data are input to CNN. At every instance status varies with respect to its effect on future time intervals and so, the model is trained at each hour over the following 48 hours. Consequently the inputs are paired with the deviations in the target feature in

the various intervals of time for training several models with the same structure corresponding to the different time intervals. Regardless of the location and time interval,

.

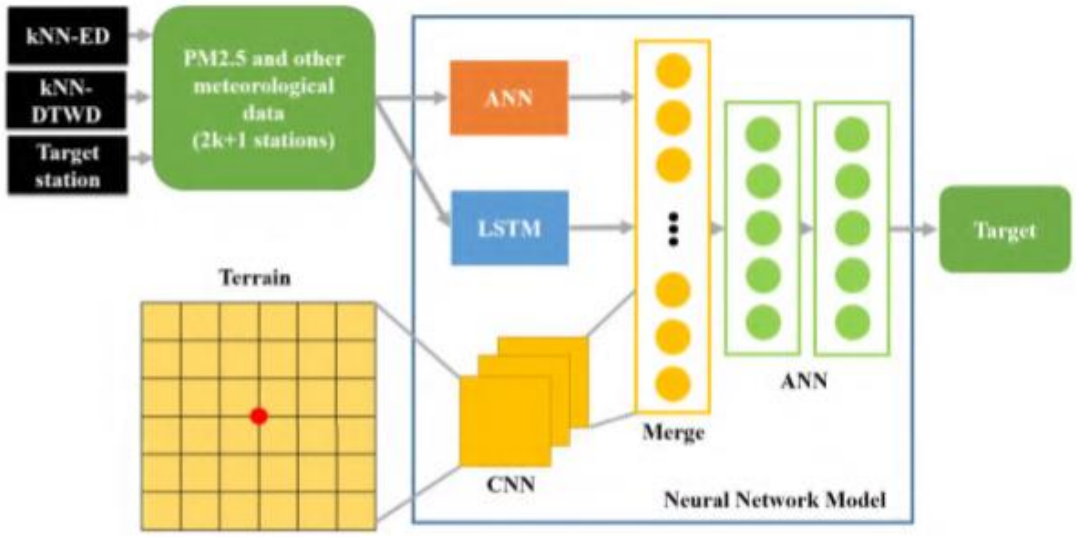

FIGURE 3.2. Prediction model design

3.B.1. Temporal Relations

Historical target location features historical target location features are input to the TRE to predict the future time series. The input time series for PM2.5 and other concentrations are continuous and coherent, and can be divided into low frequency (trends) and high frequency (rapid changes) information.

LSTM models historical time series behaviour and so, TRE LSTM is considered to obtain target location time series trends. Since ANN uses current data only, it is sensitive to rapid changes. Thus, the LSTM and ANN provide low and high frequency information, respectively, from the sequences.

3.B.2)Spatial-Temporal Relation

Pollutant dispersal means that air quality at one location can be correlated spatially with that at other locations. Since air quality at a given location is affected by local emissions as well as emissions in surrounding areas, SRE uses historical spatial-temporal neighbourhood location features inputs. Therefore, predicts target location air quality based on AQIs and meteorological data from other locations. The SRE for a location requires data mining from locations in the spatial-temporal neighbourhood using kNN-ED and kNN-DTWD, including AQIs and meteorological conditions for the previous 6 hours. The ANN SRE is included to increase model sensitivity.

3.B.3. Terrain Extractor

Due to differences in altitudes and different barriers, the relationships between locations vary. Hence, data related to terrain were included to magnify location correlations. A matrix of 121 square sections with 11x11 coordinate lines at 500m intervals, is used to capture terrain data of the nearby locations. The elevation or altitude of each point, elev, was normalized as

$$Hs = (elev - elev_{st}) / elev_{st}$$

and transformed to the relative elevation**,**

$$elev_{rel} = 1 / e^{Hs},$$

where ***Hs*** is the standardize elevation, to decrease the impact of higher altitudes. Relationships between locations could be extracted such as the impact of wind direction and wind speed for locations adjacent to mountains.

3.B.4. Merge Layer

The outcomes of TRE, SRE, and TE are concatenated and passed to the ANN layer. Local and global inputs are applied in this model. At some instances, local information is more important than global, for example, if the air circulation between to related locations is week. But sometimes, if the speed of wind is high, global information may be more important for determining air quality when wind speed is high. Thus, meteorological condition historical data at a given location is looked for, such as wind speed, wind direction, humidity, temperature, etc., to weight prediction calculations provided by the three components.

4. EVALUATION

The most difficult air pollutant to predict is PM2.5, and also it is the most widely reported metric. Hence, PM2.5 is chosen as the predictive feature in the ST-DNN model. Prediction of the concentration of PM2.5 in air can help in guiding individuals to choose between in indoor and outdoor activities, thereby limiting the exposure to it.

Index	1	2	3	4	5	6	7	8	9	10
Air Pollution Banding	Low	Low	Low	Moderate	Moderate	Moderate	High	High	High	Very High
$PM_{2.5}$ concentration ($\mu g/m^3$)	0-11	12-23	24-35	36-41	42-47	48-53	54-58	59-64	65-70	>=71
Accompanying health messages for the general population	Enjoy your usual outdoor activities			Enjoy your usual outdoor activities			Anyone experiencing discomfort such as sore eyes, cough or sore throat should consider reducing activity, particularly outdoors.			Reduce physical exertion, particularly outdoors, especially if you experience symptoms such as cough or sore throat.

FIGURE 4.1 Air quality index with corresponding message

Applying the dataset from Taiwan, all Adaptive ANN (A), LSTM (L), and CNN (C) combinations were examined to identify the best model. It was found that for the first hour prediction ST-DNN model with components A+L+C gave the best prediction and for further hours the CNN only model showed the best performance.

The kNN-DTWD based models showed superior performance to those based on kNN-ED. CNN extracted neighbourhood elevation and determined diffusion delays (direction and time) for target features.

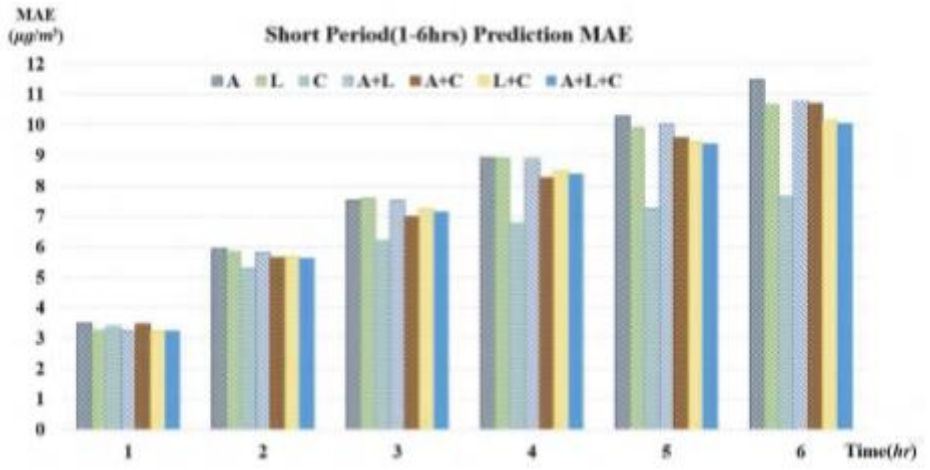

FIGURE 4.2 Model comparison

Inclusion of relative elevation is important as it reduces the local interferences and helps improve the prediction. This is because the concentration of particulate matter decreases with increase in altitude. Also, CNN provides superior performance in complex terrain.

On comparing the performance with k, for kNN-ED, the mean square error increases with increase in k and for kNN-DTWD, the mean square error decreases with increase in k. Thus kNN-DTWD outperforms the kNN-ED as it also considers the terrain.

5. CONCLUSION

The seminar was about an air quality forecasting system which uses a data driven model, ST-DNN to predict PM2.5 over 48 hours. It shows that inclusion of an LSTM helps in enhancing the first hour prediction and since a CNN can extract temporal delay factor from the spatial related data, including a CNN module helps in the enhancement of the next few hours prediction.

References *:*

[1] PING-WEI SOH1, JIA-WEI CHANG2, AND JEN-WEI HUANG 2, “Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations”, IEEE Access, Volume 6, 2018

[2]https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/