A Machine Learning Based PM2.5 Forecasting Framework Using Internet of Environmental Things

Sachit Mahajan[1,2,3](✉), Hao-Min Liu[2], Ling-Jyh Chen[2], and Tzu-Chieh Tsai[3]

[1] Social Networks and Human Centered Computing Program,
Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
sachitmahajan@iis.sinica.edu.tw
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3] Department of Computer Science, National Chengchi University, Taipei, Taiwan

Abstract. Information and communication technologies have been widely used to achieve the objective of smart city development. A smart air quality sensing and forecasting system is an important part of a smart city. In this paper, we present an approach to accurately forecast hourly fine particulate matter (PM2.5). An Internet of Things (IoT) framework comprising of Airbox Devices for PM2.5 monitoring has been used to acquire the data. Our main focus is to achieve high forecasting accuracy with reduced computation time. We use a hybrid model to do the forecast and a grid based system to cluster the monitoring stations based on the geographical distance. The experimentation and evaluation is done using Airbox Devices data from 119 stations in Taichung area of Taiwan. We are able to demonstrate that a proper clustering based on geographical distance can reduce the forecasting error rate and also the computation time.

Keywords: Internet of Things (IoT) · Air quality · Smart cities

1 Introduction

Whenever we talk about a smart city, it always revolves around leveraging the advancement in the field of Information and Communication Technology (ICT). Lately, Internet of Things (IoT) has revolutionized the smart city initiative and IoT devices have become the technological backbone of smart cities [1]. IoT has revolutionized the technology evolution and can be seen as an evolution of Internet into an omnipresent network of smart interconnected objects that not only sense the information but also interact with the outside physical world [2].

In the recent years, problems related to deteriorating air quality have been a topic of concern all over the world. Rapid industrial growth and urbanization has been an important factor behind the degrading air quality. When we talk about smart cities, not only we need continuous air quality monitoring but also

Y.-B. Lin et al. (Eds.): IoTaaS 2017, LNICST 246, pp. 170–176, 2018.
https://doi.org/10.1007/978-3-030-00410-1_20

a system which accurately forecasts future air quality. There are various kind of pollutants based on human and environmental factors that get diffused in the air. One of the most important among all the pollutants is fine particulate matter whose size is 2.5 µm or less also known as PM2.5. These particles can cause serious damage to human health and can lead to respiratory problems [3]. In the past, some research has been done on forecasting air quality based on using mathematical models and air quality modelling softwares. But these methods still have some drawbacks that can be addressed. Our approach is different from the conventional approaches. We use a data centric and grid based approach which uses real-time data from Airbox Project [4] to perform the experiments and evaluation. The contribution of the paper is three-fold:

(1) We propose a neural network based Hybrid model for hourly PM2.5 prediction.
(2) We use real-time Airbox data to perform the hourly PM2.5 prediction.
(3) We evaluate our model by clustering stations into grids based on the geographical distance. We perform the evaluation using data from 119 stations in Taichung area of Taiwan and try to understand the relationship between forecast accuracy and computation time.

The rest of the paper is organized as follows. Section 2 includes some related works. That is followed by Sect. 3 which includes the methodology followed for this work. It gives the details about the data, the hybrid model and the clustering approach. Section 4 includes the results and the evaluation part based on the experiments. In Sect. 5, we conclude the paper and give some ideas about future work.

2 Related Work

There has been some previous research works which focused on air quality forecasting. In one of the related works [5], the authors performed PM2.5 prediction for the next 48 h. They used a data based approach which involved using a linear regression and neural network based prediction model. In [6], the authors proposed a Deep Hybrid Model for weather forecasting. Though it didn't forecast PM2.5 but it predicts temperature, dew point and wind. In one of the other works [7], the authors used a combination of remote sensing and meteorological data with the ground based PM2.5 observations. Some of the researchers have implemented machine learning techniques on big data to perform the computation [8]. However, there are some drawbacks as most of these techniques rely on feeding some features into the model. The features are for one particular location and the model is implemented on all the stations. It is easy to understand that different regions have different PM2.5 levels based on different sources of emission. So accurately performing forecasting using a generic models for all the stations is not really feasible. To tackle this issue, we introduce the concept of clustering the monitoring stations into grids based on the geographical distance between the stations.

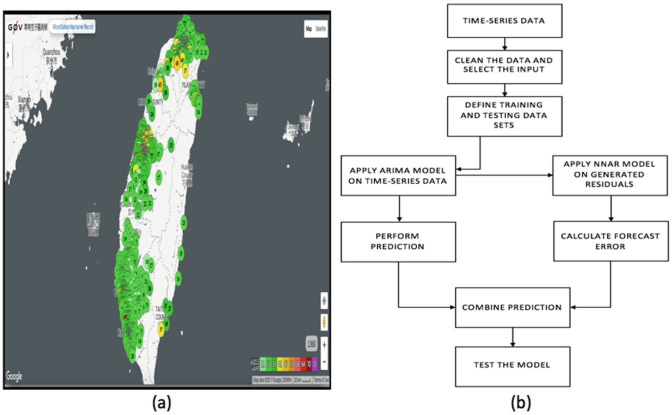
TIME-SERIES DATA
CLEAN THE DATA AND SELECT THE INPUT
DEFINE TRAINING AND TESTING DATA SETS
APPLY ARIMA MODEL ON TIME-SERIES DATA
APPLY NNAR MODEL ON GENERATED RESIDUALS
PERFORM PREDICTION
CALCULATE FORECAST ERROR
COMBINE PREDICTION
TEST THE MODEL
(a)
(b)

Fig. 1. (a)Visualization platform provided by g0v community (b) Hybrid Model flowchart

3 Methodology

In this part we will first explain about the Airbox Data used for the experiment. This section will also describe the Hybrid model used for performing the prediction and the clustering approach.

3.1 Airbox Data

The data is obtained from Airbox Project which involves deployment of IoT systems all over Taiwan. The Airbox device installation was carried out on a large scale initially in Taiwan. Initially it was initiated in Taipei City and on March 22, 2016, 150 devices were deployed around the city. The project has been widely acknowledged by people and has spread over other areas of Taiwan as well. 230 devices have been installed in Taichung area; 242 devices have been deployed in Kaohsiung area; 220 devices have been deployed in Tainan and 298 devices have been deployed in New Taipei City area of Taiwan. As of now, there are more than 1500 devices deployed in Taiwan and 24 cities all around the world. For this study, the data was collected for Airbox Devices deployed in Taichung area of Taiwan. To visualize the Airbox data, there have been some visualization systems developed. One of them is geographic information (GIS) based visualization system that combines the location with the measured data. Figure 1(a) shows the snapshot of visualization platform. The measurement data for this work was collected for the time period between January, 18 2017 and February 17, 2017.

3.2 Hybrid Model

We can divide a time-series into linear and non-linear components. The flowchart for Hybrid Model is depicted in Fig. 1(b). The Hybrid model is made by utilizing the Autoregressive Integrated Moving Average (ARIMA) model [9] and Neural Network Autoregressive (NNAR) model [10]. ARIMA model is good but it doesn't capture the non-linear components. An ARIMA (p, d, q) model has parameters p, d and q which are all integers. They must be greater than or equal to zero. The parameters respectively point to the order of the autoregressive (AR), integrated (I) and moving average (MA) components of the model. So we need a technique that can capture the non-linear components too. To solve that issue we can use Artificial Neural Networks (ANN). In NNAR [10] model, the input comprises of lagged time-series and the output is predicted time-series value. Authors in [11] described the implementation of a hybrid model. It can be represented as

$$Z_t = X_t + Y_t \tag{1}$$

In the above equation, X_t represents the linear components and Y_t represents the non linear components. In the initial step these two components have to be estimated from the data. Next step is the application of ARIMA model. ARIMA takes care of the linear components and the non- linear residuals are generated. We assume that R_t be the residuals generated at time t from the linear model. It can be represented as

$$R_t = Z_t - P_t \tag{2}$$

In the above equation, P_t is the forecast value for time t. The residuals are modelled using neural networks. If we assume that there are n input nodes, then the neural network model for residuals can be represented as

$$R_t = f(R_{t-1}, R_{t-2}, ..., R_{t-n}) + E \tag{3}$$

Neural network defines the non-linear function f and E is the randomly generated error. In the end, forecast from the neural network is generated and Eq. (3) is used to get the final output. For our hybrid model, we used an ARIMA (3, 1, 1) model where 3, 1, 1 are the values of p, d and q respectively. For neural network, we used an NNAR (9, 5) model which used 9 lagged inputs with 5 nodes in the hidden layer.

3.3 Clustering Approach

In order to reduce the computation time of the prediction on all stations, we apply clustering approach before implementing the prediction. First of all, we divided all the stations into different clusters according to their geographic locations. Then, we apply the prediction model on the average value of time series data in each cluster. According to the distribution of stations in Taichung, we have done the experiments to divide all stations into one-by-one to four-by-four clusters. One-by-one case denotes that we predict the whole region with only the average value of time series data in that region, shown in Fig. 2(a). In two-by-two case, we divided stations into four clusters according to the median value of their latitude and longitude, as it is shown in Fig. 2(b). In three-by-three case, we divided stations into nine clusters according to the 33^{th} quantile value and 67^{th} quantile value of their latitude and longitude, as it is shown in Fig. 2(c). And similarly it is done for four-by-four case as shown in Fig. 2(d).

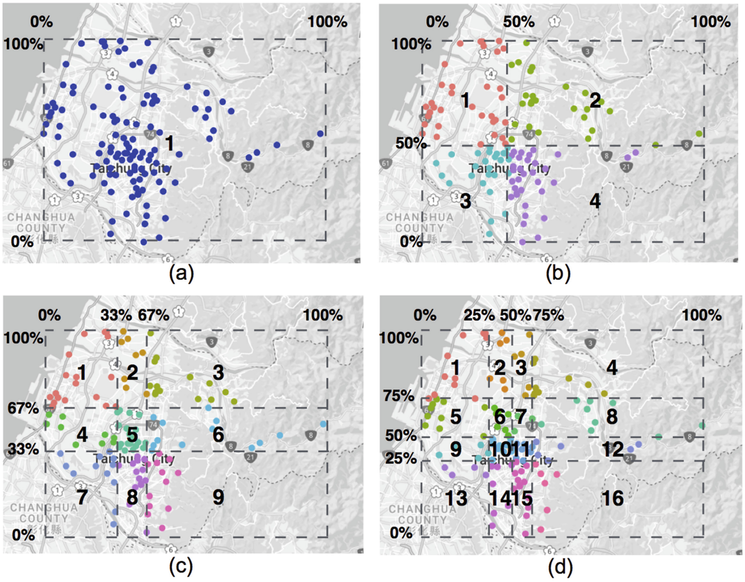

Fig. 2. (a) Without Clustering Method (b) 2 × 2 Clustering Method (c) 3 × 3 Clustering Method (d) 4 × 4 Clustering Method

Table 1. Average Mean error and RMSE of Grid 1 × 1 to Grid 4 × 4.

Grid	1 × 1	2 × 2	3 × 3	4 × 4
Ave. mean error	11.42	7.21	6.24	4.90
Ave. RMSE	13.95	8.93	8.06	6.17

4 Results and Evaluation

The experiments and evaluation were done using the real-time Airbox data. The Hybrid Model was trained using the six days hourly historical data. And the next one day hourly data was used for testing the model.

To analyse the results, we calculated the Root Mean Square Error (RMSE) and mean values. The results of average mean value and average RMSE value of Grid 1 × 1 to Grid 4 × 4 are shown in Table 1. The average of mean value and RMSE value is the weighted average, influenced by the number of stations in each cluster. The results show that both mean error and RMSE decrease with an increase in the number of grids. The mean error of Grid 4 × 4 can be reduced to less than 5 $\mu g/m^3$. This is very significant result as our aim is to perform forecast with prediction error as low as possible.

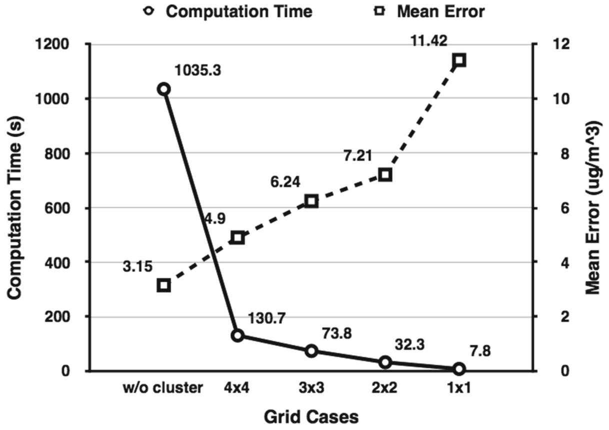

Fig. 3. Comparison of computation time and mean error among four designed cases

4.1 Evaluation

In order to evaluate our approach, we made a comparison between the computation time and mean error variation among four designed cases (4 × 4 grid, 3 × 3 grid, 2 × 2 grid, 1 × 1 grid). Before applying the grid based clustering approach, we applied the model individually on 119 stations without clustering. Although the mean error was less but still the computation time was huge. So to evaluate the system, we included the case without any cluster in the comparison result as shown in Fig. 3. The total computation time without applying clustering method went up to 1035.3 s, which is a lot when we talk about real-time computing. In the case with Grid 4 × 4 clustering, the computation time was reduced to 130.7 s, which is around 87% reduction from the design without clustering, while there was an increase of 1.74 $\mu g/m^3$ mean error as the trade off. This is an acceptable trade off considering the computation time it saves.

5 Conclusion and Future Work

In this paper, we proposed a framework that uses IoT technology and machine learning techniques to forecast PM2.5 concentration. We performed the prediction using the Hybrid model which utilizes an ARIMA model and an NNAR model. Our main aim was to reduce the computation time and at the same time make sure that we get an acceptable forecast accuracy. We followed a grid based method to efficiently group 119 monitoring stations in Taichung area of Taiwan. Grids consisted of monitoring stations clustered together according to the geographic distance. To evaluate, we tested the system for grids of different sizes and showed how the computation time can be reduced with an acceptable forecasting error.

We would like to extend this work by implementing the PM2.5 prediction framework for other regions as well. Also we would like to explore further possibilities of using different techniques to cluster monitoring stations which show similar trend over a particular duration of time. If the results are favorable, these studies can be used by environmental pollution monitoring agencies for policy making.

References

1. Jin, J., Gubbi, J., Marusic, S., Palaniswami, M.: Information framework for creating a smart city through Internet of Things. IEEE Internet Things J. **1**(2), 112–121 (2014)
2. Delic, K.A.: On resilience of IoT systems: the Internet of Things (ubiquity symposium). Ubiquity **2016**(February), article no. 1 (2016)
3. Xing, Y.-F.: The impact of PM2.5 on the human respiratory system. J. Thorac. Dis. **8**(1), E69–E74 (2016)
4. PM2.5 Open Data Portal. http://pm25.lass-net.org/en/
5. Zheng, Y., et al.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015), pp. 2267–2276. ACM, New York (2015)
6. Grover, A., Kapoor, A., Horvitz, E.: A deep hybrid model for weather forecasting. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015), pp. 379–386. ACM, New York (2015)
7. Lary, D.J., Lary, T., Sattler, B.: Using machine learning to estimate global PM2.5 for environmental health studies. Environ. Health Insights. **12;9**(Suppl. 1), 41–52 (2015)
8. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. (TIST) - Spec. Sect. Urban Comput. **5**(3), 55 (2014). Article 38
9. Christodoulos, C., Michalakelis, C., Varoutas, D.: Forecasting with limited data: combining ARIMA and diffusion models. Technol. Forecast. Soc. Change **77**(4), 558–565 (2010)
10. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts (2013)
11. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50**, 159–175 (2003)