JES
Journal of Engineering Sciences

Vol 13, Issue 05, MAY/2022
ISSN NO:0377-9254

A generic model to analyze and predict the location-based air quality from the sensing network data using machine learning

SOMISETTY BALA KUMARA VIGNESH, NALLAMILLI GAYATHRI SOWJANAYA, LAKAMSANI PAVANI LAKSHMI, SURLA VARAHA NAGA SATYA LOVA MAHESH
B.Tech. IV Year Students, Department of IT, PRAGATI Engineering College (Autonomous), Surampalem, A.P, India.

ABSTRACT

Air, an essential natural resource, has been compromised in terms of quality by economic activities. Considerable research has been devoted to predicting instances of poor air quality, but most studies are limited by insufficient longitudinal data, making it difficult to account for seasonal and other factors. Several prediction models have been developed using an 11-year dataset collected by Taiwan's Environmental Protection Administration (EPA). Machine learning methods, including adaptive boosting (AdaBoost), artificial neural network (ANN), random forest, stacking ensemble, and support vector machine (SVM), produce promising results for air quality index (AQI) level predictions. A series of experiments, using datasets for three different regions to obtain the best prediction performance from the stacking ensemble, AdaBoost, and random forest, found the stacking ensemble delivers consistently superior performance for R 2 and RMSE, while AdaBoost provides best results for MAE. Keywords: air quality monitoring; machine learning; air quality index

INDEX TERMS

Time-Series Prediction, Air Quality Measurement, Machine Learning

INTRODUCTION

Due to rapid urbanization and industrialization, many coun- tries around the world are facing a critical crisis of air pollution. Air pollution has become a threat to public health and a heavy influential factor on citizen's daily activity. In metropolitan cities in developing countries bothered by problems of air pollution, such as Beijing and Delhi, people usually need to wear a mask before going out [1]. Besides, outdoor activities are also constrained by the intra-day air quality.

Air pollution is caused by the presence of different air pollutants. The primary air pollutant gases are nitrogen diox- ide (NO_2), carbon monoxide (CO), ozone (O_3) andsulphur dioxide (SO_2) [2]. Another type of air pollutants is air particulate matter (PM). Among them, P M2.5 and P M10 are of particular concerns to people, which refers to atmospheric particulate matter that have a diameter of less than 2.5 µm and 10 µm. These particles can cause many respiratory or cardiovascular diseases [3]. Thus, many cities have built their own air quality monitoring stations and publish the real-time air quality information every hour. As the concern for air pollution increases, its becoming increasingly critical to measure the air quality around people [4], [5], which inform people about when is safe to perform outside activities and help them plan better routes to reach their destinations. Typ- ically, monitoring stations at fixed locations is the conven- tional approach for atmospheric factor monitoring for a large geographical district. While it is not difficult to implement such fixed sensor based monitoring system, it faces several challenges. First, huge investment is involved in building and deploying monitoring units to cover a large area. Also, it is highly dependent on neighboring environments and tends to be less accurate for farther areas. In areas close to the roads, even small distances can make a huge difference in air quality data measurement from car pollutions. Hence, new ways to collect air quality information in a cheaper and more flexible way and provide detailed air quality prediction is in demand. To address these issues,

one possible solution is to make the sensors mobile using Internet-of-Things(IoT). For exam- ple, attaching sensors on moving cars or drones proved to be a feasible method [6]. In this work, we developed the IoT devices to monitor air quality. We collected air pollution data by mounting a sensor to a car and moved around the city of Incheon, Republic of Korea. This data is then pre- processed and stored in our server. One major advantage of
using a mobile sensor is that it provides the very first hand air pollution information for an area at a particular time, when the car was moving through there. we can also cover more geographical regions and have more accurate localized information with mobile IoT sensors. While a static fixed
sensor can provide continuous feed of information about a particular area, it is not easy with a mobile sensor. However, this can be minimized by having multiple mobile sensors or assigning smaller coverage area to a mobile sensor. In this work, we propose a hybrid approach, where we de- ploy multiple static sensors as well as IoT mobile sensors to effectively monitor air quality. The static sensors can provide a holistic view by providing a continuous feed of information. On the other hand, mobile sensors can provide more accurate
data about specific areas to reduce the error from static sensors. In this paper, we build a prediction model to utilize
the collected data and provide rapid information about the air quality around people. We also developed a visualization tool to better analyze and forecast air quality and provide insights to both professional researchers and ordinary users. The main contributions of our work are summarized as follows:

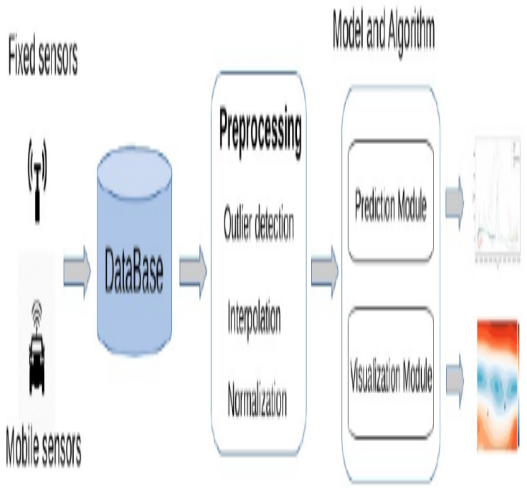
- We proposed a hybrid approach to integrate fixed and mobile IoT sensors to measure and predict air quality
data.
- We demonstrated the feasibility and effectiveness of our approach by analysing the prediction result with
different machine models.
- We developed a visualization tool to show the relative distribution of the air pollutants with a focus on P M10
and P M2.5, where it provides an intuitive understanding of the air quality around people.

The rest of our paper is organized as follows: Section 2. presents the related work on different air quality mea- surement and prediction methods. Section 3 describes the development of IoT sensors and data processing. Section 4
explains our models and algorithms. The experimental setup and results are reported in Section 5, and an analysis of the results is provided in Section 6. We summarize our work and offer conclusion in Section 7 and Section 8

RELATED WORK

To measure the air quality, several monitoring methods have been proposed and utilized. In Zheng et al.'s research [7],
they use public and private web services as well as a list of public websites to provide real-time meteorological, weather
forecasts and air quality data for their forecasting. Small unmanned aerial vehicles are used in the work of Alvarado
et al. [8] as a methodology to monitor P M10 dust particles, where they can calculate the emission rate of a source. With the development of smart city technologies, IoT devices have been shown to be an effective option to collect real time weather, road traffic, pollution and traffic information. Thus, IoT devices are also considered to enable air quality analysis [9]. In addition to the fixed sensors, public transportation in- frastructure such as buses has been used to collect air quality data [10]. Also, there is one project [11] engaged the entire community members in collecting data and developed an online air quality monitoring system based on it, which is also called crowdsourcing. Hasenfratz et al. [12] utilized sensor nodes to build a thousand models targeting at different time periods. All these aforementioned methods are either costly or time consuming. In our work, we explore the use of fixed and mobile IoT sensors together to improve the prediction performance, which has not been researched much yet.
To meet the increasing query frequency of air quality in real time and also to enable citizens to react instantly to the pollution, there has been a large body of work on building connected monitoring sensor networks to share the current air quality information with them [13]. Garzon et.al presented in [14] an air quality alert service. Their service continuously determines the areas, where the level of certain matter

concentration exceeds the preset threshold, and notify
users if they entered them. Maag et. al [15] proposed a multi-pollutant monitoring platform using wearable low-cost sensors. Compared with above methods, our system can serve the similar functions to end users practically with either fewer sensors or less demand for computation. For prediction, regression models are commonly used in the area of air quality prediction. A multivariate linear re- gression model for predicting P M 2.5 of short-period time is proposed in Zhao's work [16], which includes other gaseous pollutants such as SO2 , N O2, C O and O3. As deep learning emerged as an effective method in many applications, time series data of air pollution based on different network models have been also extensively studied and developed. Novel models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit Network (GRU) have been proved to be
a powerful sequential structures in predicting future values of air quality [9], [17] . Yi et al. [18] proposed a deep distributed fusion network to learn the characteristics of spatial disper- sion and capture all the influential factors that may have a direct or indirect effect on air quality. These aforementioned technologies fits non-linear models flexibly but usually being short of offering insight to the hidden mechanism. In addi- tion, they have not shown to necessarily outperform classical regression models in many scenarios [19]. There are also a lot of researches concentrate on approaches to model and
simulate the pollutants for prediction [20]. With a small amount of data set oriented in our project, we decided to take conventional regression models as our baseline methods because of computation efficiency, while yielding favorable results.



IMPLEMENTATION

In this section, we first describe the design and implemen- tation of IoT sensor device deployed in our research. Our deployment and data collection are performed in Songdo [21], South Korea, which is envisioned to be developed as a smart city. Next, we explain the preliminary processing of the acquired raw data and describe how we store and transmit the collected and cleaned data. Then, we further present the user interface to check the collected data for our analysis. Figure 1 describes the overall architecture of our proposed system

Support Vector Regression

One year after the introduction of SVM, Smola et al. [32] advanced an alternative loss function, which also allowed SVM to be applied to regression problems. Support vector regression (SVR) has been applied in the field of TS forecasting, with excellent outcomes. For instance, Drucker et al. [22], Müller et al. [23], and Cao and Tay [33] suggest that SVR is a promising method for TS forecasting, as it offers several advantages: a smaller number of free parameters, better forecast ability, and faster training.

In SVR, the idea is to map the data events X into a k-dimensional feature space *F*, through a nonlinear mapping , so that it is possible to fit a linear regression model to the data points in this space. The obtained linear learner is then used to forecast in the new feature space. Once again, the mapping from the input space into

Vol 13, Issue 05, MAY/2022
ISSN NO:0377-9254

the new feature space is defined by the kernel function.

One of the most attractive characteristics of SVR is related to the model errors; instead of minimizing the observed training error, SVR minimizes a combination of the training error and a regularization term, aimed at improving the generalization ability of the model [34]. Other attractive properties of SVR are related to the use of kernel functions, which make them applicable both to linear and nonlinear forecasting problems, and the absence of local minima in the error surface, due to the convexity of the fitness function and its constraints.

Given:

Training dataset T, represented by

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, \quad (1)$$

where $x \in X \subset \mathbb{R}^n$ are the training inputs and $y \in Y \subset \mathbb{R}$ are the training expected outputs.

A nonlinear function:

$$f(x) = w^T \Phi(x_i) + b, \quad (2)$$

where w is the weight vector, b is the bias, and $\Phi(x_1)$ is the high dimensional feature space, which is linearly mapped from the input space x; the objective is to fit the training dataset T by finding a function $f(x)$ that has the smallest possible deviation ε from the targets .

Equation (2) can be rewritten into a constrained convex optimization problem as follows:

$$\text{minimize} \quad \frac{1}{2} w^T w$$
$$\text{subject to} \begin{cases} y_i - w^T \Phi(x_i) - b \le \varepsilon \\ w^T \Phi(x_i) + b - y_i \le \varepsilon. \end{cases} \quad (3)$$

The aim of the objective function represented in equation (3) is to minimize w while satisfying the other constraints. One assumption is that $f(x)$ exists, i.e., the convex optimization problem is feasible. This assumption is not always true; therefore, one might want to trade off errors by the flatness of the estimate. Having this in mind, Vapnik reformulated equation (3) as

$$\text{minimize} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} (\xi_i^+ + \xi_i^-)$$
$$\text{subject to} \begin{cases} y_i - w^T \Phi(x_i) - b \le \varepsilon + \xi_i^+ \\ w^T \Phi(x_i) + b - y_i \le \varepsilon + \xi_i^- \\ \xi_i^+ \xi_i^- \ge 0, \end{cases} \quad (4)$$

where $C < 0$ is a prespecified constant that is responsible for regularization and represents the weight of the loss function. The first term of the objective function $w^t.w$ is the regularized term, whereas the second term $C\sum_{i=1}^{m}(\xi_i^+ + \xi_i^-)$ is called the empirical term and measures the -insensitive loss function.

To solve Equation (4), Lagrangian multipliers $(\alpha_i^+, \alpha_i^-, \eta_i^+, \eta_i^-)$ can be used to eliminate some of the primal variables. The final equation that translates the dual optimization problem of SVR is

$$\text{minimize} \quad \frac{1}{2} \sum_{i,j=1}^{m} K(x_i, x_j)(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) + \varepsilon \sum_{i=1}^{m} (\alpha_i^+ + \alpha_i^-) - \sum_{i=1}^{m} (\alpha_i^+ - \alpha_i^-)$$
$$\text{subject to} \begin{cases} \sum_{i=1}^{m} (\alpha_i^+ - \alpha_i^-) = 0 \\ \alpha_i^+, \alpha_i^- \in [0, C], \end{cases} \quad (5)$$

where k(xi,xj) is the kernel function; the above formulation allows the extension of SVR to nonlinear functions, as the kernel function allows nonlinear function approximations while maintaining the simplicity and computational efficiency of linear SVR.

The performance and good generalization of SVR depend on three training parameters: The kernel function C (the regularization parameter) ε (the insensitive zone)

Many possible kernels exist. In this work, the polynomial kernel and the radial basis function (RBF) kernel were studied. The interested reader is referred to [35] for a discussion of these and other existing kernel functions.

CONCLUSION

In this paper, we explored a new way to predict immediate air quality around people, by combining fixed and mobile sensors. Our experimental results show that our proposed hybrid distributed fixed and IoT sensor system is effective in predicting air quality around the people. In addition, our proposed system can be practically realizable by leveraging public transportation system such as buses as well as taxis to be equipped with IoT sensor devices to measure different areas. The predicted air quality data from our system can be served in various scenarios, such as planing for outdoor activities

REFERENCES

[1] "Beijing's air would be step up for smoggy delhi," http://https://www.nytimes.com/2014/01/26/world/asia/ beijings-air-would-be-step- up-for-smoggy-delhi.html, accessed January 26, 2014.
[2] M. Kampa and E. Castanas, "Human health effects of air pollution," Environmental pollution, vol. 151, no. 2, pp. 362–367, 2008.
[3] E. Boldo, S. Medina, A. Le Tertre, F. Hurley, H.-G. Mücke, F. Ballester, I. Aguilera et al., "Apheis: Health impact assessment of long-term expo- sure to pm 2.5 in 23 european cities," European journal of epidemiology, vol. 21, no. 6, pp. 449–458, 2006.
[4] J. Lin, A. Zhang, W. Chen, and M. Lin, "Estimates of daily pm2. 5 exposure in beijing using spatio-temporal kriging model," Sustainability, vol. 10, no. 8, p. 2772, 2018.
[5] Y. Jiang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang, "Maqs: a personalized mobile sensing system for indoor air quality monitoring," in Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011, pp. 271– 280.
[6] D. Zhang and S. S. Woo, "Predicting air quality using moving sensors," in Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. ACM, 2019, pp. 604–605.
[7] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 2267–2276.
[8] M. Alvarado, F. Gonzalez, P. Erskine, D. Cliff, and D. Heuff, "A methodology to monitor airborne pm10 dust particles using a small unmanned aerial vehicle," Sensors, vol. 17, no. 2, p. 343, 2017.
[9] ˙ I. Kök, M. U. ¸Sim¸sek, and S. Özdemir, "A deep learning model for air quality prediction in smart cities," in 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017, pp. 1983–1990.
[10] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas," in Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, 2013, p. 15.
[11] Y.-C. Hsu, P. Dille, J. Cross, B. Dias, R. Sargent, and I. Nourbakhsh, "Community-empowered air quality monitoring system," in Proceedingof the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 2017, pp. 1607–1619

AUTHOR PROFILE

BALA KUMARA VIGNESH SOMISETTY pursuing IVth year of B.Tech at Pragati Engineering College under the stream of Information Technology interested in learning new technologies like Salesforce, MachineLearning and worked on "Face Mask Detection using TensorFlow, Keras and OpenCV" as a part of mini project as Team member.
vigneshsommisetty@.gmail com

Vol 13, Issue 05, MAY/2022
ISSN NO:0377-9254

NALLAMILLIGAYATHRISOWJANY A pursuing IVyear of B. Tech at Pragati Engineering College under the stream of Information Technology. I'm interested in learning new technologies like Machine Learning, Block chain, IOT and I have worked on " Prediction of Chronic Kidney Disease using Random ForestAlgorithm in Machine Learning" as a part ofMini-Project as a team leader.
gayathrisowjanya226@gmail.com

PAVANI LAKSHMI LAKAMSANI pursing IV thyear of B.tech at Pragati Engineering College under the stream of Information Technology and Engineering interested in learning new technologies like Machine learning Artificial Intelligence and worked on machine learning "Voice Based Email for Visually Challenged Using Text-to-Speech and Speech-to-Text Python" as a part of mini project as a team member
pavanilakamsani@gmail.com

MAHESH SURLA pursuing IVyear of B. Tech at Pragati Engineering College under the stream of Information Technology. I'm interested in learning new technologies like Big Data, Artificial Intelligence and I have worked on " Detecting E Banking Phishing Websites Using Associative Classification in Fuzzy Logic Data Mining" as a part of Mini-Project as a team member.
mmahe4433@gmail.com