A Geo-Based Fine Granularity Air Quality Prediction Using Machine Learning and Internet-of-Things

Hang Wang[1(✉)], Yu Sun[2], and Qingquan Sun[3]

[1] University High School, Irvine, CA 92612, USA
alfredll86381762@gmail.com
[2] Department of Computer Science, California State Polytechnic University, Pomona, Pomona, CA 91768, USA
yusun@cpp.edu
[3] School of Computer Science and Engineering, California State University, San Bernardino, San Bernardino, CA 92407, USA
qsun@csusb.edu

Abstract. As the development of economy and industry, air quality decreases as one of the exchanges of our achievements. Although air pollution has already been considered as a global and critical issue over the past decades, there has not been much innovation on the way people monitor and check the quality. Most of the air quality data today is provided by government or professional sensors set up in cities, which does not provide more detailed status in smaller geo locations with finer granularity, such as specific villages, schools, and shopping malls. In this project, we use machine learning to make a mathematical model which could be used to predict the air quality for small geo locations with accuracy and fine granularity. Through series of experiments and comparisons, the most accuracy mathematical model was found, which had a difference percentage less than 20% with the real data.

Keywords: Machine learning · Air quality prediction · Internet-of-Things

1 Introduction

Air quality has received much attention in recent years due to the development of industry and environmental protection sense of people. There is a data said that from 2008 to 2013, the air pollution increased about 8% among the cities around the world. Air pollution perplexes everyone, there are about 5.5 million people died due to the air pollution. According to the report of American Lung Association, half of the American population live in an environment which has the danger of air pollution. 6 of the most polluted cities in the United States are in California. Published by UNICEF on October 31, 2026, "Clear the Air for Children" said there are about 300 million children living in extreme polluted environment.

The standards of measurement in determining the air quality are PM 2.5, PM 10, O_3, NO_2, SO_2. PM 2.5 were Fine particles which were 2.5 μm in diameter or smaller, and can only be seen with an electron microscope. PM 10 were coarse dust particles which

X. Gu et al. (Eds.): MLICOM 2017, Part I, LNICST 226, pp. 291–299, 2018.
https://doi.org/10.1007/978-3-319-73564-1_29

were 2.5 to 10 micrometers in diameter. O_3 were group of pollutants emitted during the combustion of fossil fuels. Nitrogen dioxide is an important air pollutant because it contributes to the formation of photochemical smog, which can have significant impacts on human health. SO_2 results from the burning of either sulfur or materials containing sulfur. One way to detect air quality is to put professional sensors everywhere around world in every area in every city. However, although the governments and some websites are doing this, the data are not real time and with high cost. In addition, these professional sensors only cover a limited number of big cities, leaving the air quality for most of areas unavailable. For instance, Fig. 1 shows a popular website that displays the PM 2.5 information for various locations in the world. However, the website mostly covers the major big cities, leaving the small regions and areas unprocessed.

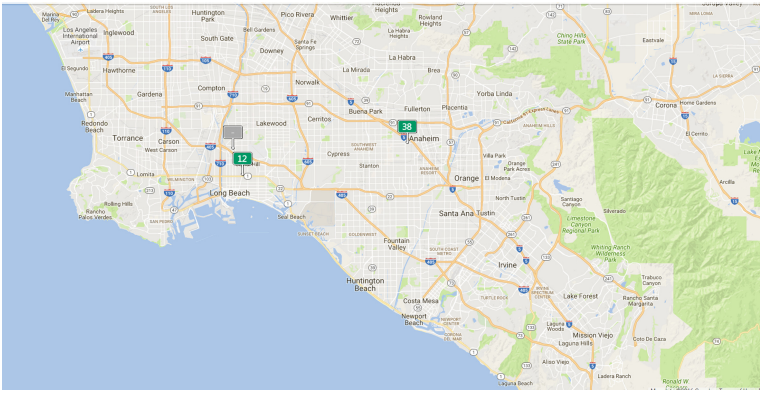


Fig. 1. A screenshot from https://waqi.info that displays city PM 2.5

In this paper, we propose to use machine learning to build a mathematical model which could be used to predict the air quality for small geo locations with accuracy and fine granularity. Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed [11]. In order to get the real air quality input data, we also built a cost-effective Internet-of-Things (IoT) solution that can monitor air quality data and send the data through the Internet at real-time. Internet of things is the internetworking of items embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data [10]. When making the air quality predictions, there are some variables that can affect the result such as the layout of the deployed monitor devices, the number of deployed monitor devices and the layout of the geo location. The goal is to use very few air quality sensors to make accurate predictions using machine learning approaches.

The rest of the paper is organized as follows: Sect. 2 gives the details on how we built the system including the architecture design and the specific components; Sect. 3 focuses on the machine learning experiments and discusses the results; Sect. 4 presents a few related work in this area, following by giving the conclusion remarks in Sect. 5, as well as pointing out the future work.

2 System Overview and Implementation

The system workflow includes building sensor monitors, collecting data and predicting air quality, as well as visualizing the data. This whole system can use the limited measured data from sensors, and through machine learning to find a suitable model to predict real time air quality. Figure 2 shows an overview of the system.

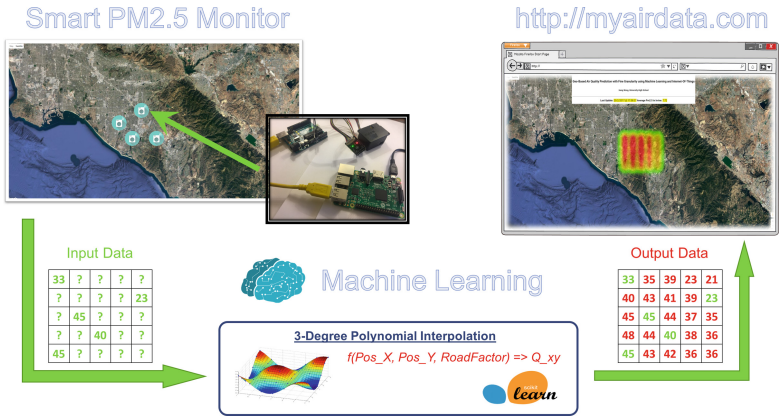


Fig. 2. The system overview

In order to get the most accurate mathematical model, an Internet-of-Things application has been built to collect air quality data, make air quality predictions and display air quality result. The application contains two modules:

Module 1. The hardware device that monitor the data and send those data to database, which allows data to be processed, followed by sending out the data to the cloud server. **Module 2.** The web-based server application that receives the data, and fits them in the mathematical model to make air quality predictions in other places using machine learning algorithms. A web-based frontend user interface has been built to visualize the air quality data and periodically refresh it (Fig. 3).

The PM 2.5 sensor used in this project is dfRobot laser dust sensor SEN0177. PM2.5 laser sensor is a digital sensor used to obtain the amount of suspended matter in air with value range from 0.3 to 10 microns. In order to read the sensor data, Arduino [2] is used. It is an electronics platform which can interact with many sensors for different purposes. Arduino is used as a single-board microcontroller programed by C or C++. One of the major limitations of the Arduino is that it is not easy to send and receive data from the Internet. Thus, Raspberry PI [3] is used to upload the data to the database.

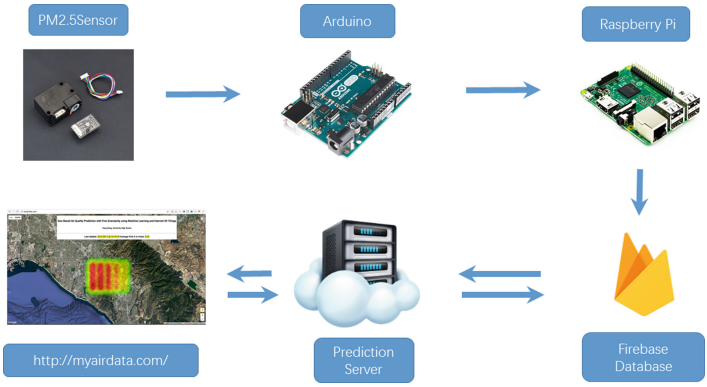


Fig. 3. The system architecture

In this project, Google Firebase is used as the database. It allows the users to upload the data from a device and get access of those data at any devices. Firebase's initial product was a real-time database, which allows developers to store and sync data across multiple clients [4].

Server here plays an important rule; it fits the data into the mathematical model and return the prediction result. The server is implemented using Python Flask [5].

In order to give out a more user-friendly experience, a web site is created to visualize the air quality data result as shown in Fig. 4. The website is based on google map. The color shown in the website represents the magnitude of pollution, red represents bad, and green represent good.

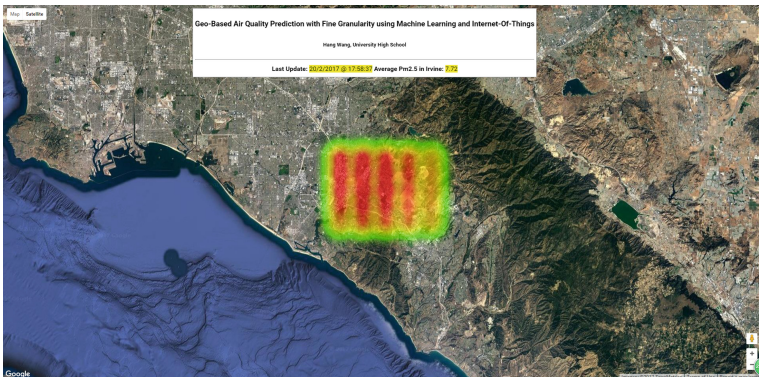


Fig. 4. The web user interface to visualize the air quality data for the city of Irvine, CA

3 Geo-Based Prediction of Air Quality

The core part of the system is to apply machine learning algorithms to predict the air quality for geo locations with fine granularity. The accuracy of the prediction model depends on the following factors:

Factor 1. The machine learning algorithm used to generate the model
Factor 2. The layout of input air quality data for known locations
Factor 3. The usage of special features (e.g. highways, population density).

Thus, to answer the question of what the best model is to predict the air quality data, a series of experiments are conducted to analyze the correlations between these factors and the accuracy.

For each experiment, we use a set of known data set as the input data, followed by running the machine learning algorithm to predict the unknown data set. Then, we compare the predicted data set values with the real actual data set, and calculate the error rate. The error rate will be based on the following formula:

$$E = (|Vexperimental - Vactual|)/((Vexperimental + Vactual)/2)$$

3.1 Experiment 1 - Machine Learning Algorithm Comparison

In this experiment, we chose 3 different common machine learning algorithms – linear regression [7], SVM [6] and polynomial interpolation [8], using the same set of air quality data we obtained from the application, we can calculate the accuracy (i.e., percentage error rate) of each machine learning algorithm.

Table 1. The input matrix with known air quality values (left) and the output matrix with air quality values being predicted (right) using linear regression

8	?	?	?	15
?	?	?	?	?
?	?	?	27	?
?	?	?	?	?
2	?	?	?	7

8	12	14.83	17.167	15
8.42	10.75	13.08	15.41	17.75
6.67	9.0	11.33	27	15.99
4.92	7.03	9.58	11.92	14.03
2	5.5	7.83	10.17	7

The Table 1 shown above is a sample experiment data used to evaluate the accuracy of linear regression. Using only 5 of the input data samples, we are able to predict the rest of the data cells using linear regression. By comparing the predicted data with the actual data values, it shows the percentage error to be: 36.53%. Using the same technique, the other two algorithms have been evaluated as shown in Fig. 5.

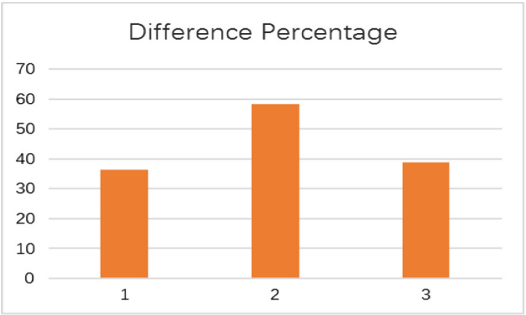


Fig. 5. Comparing the error rate of three different machine learning models

3.2 Experiment 2 - Input Data Layout Comparison

In this experiment, 4 different layouts of input air quality data are used as shown in Fig. 6. Using the best two mathematical model from Experiment 1, we can calculate the difference percentage of each layout.

Fig. 6. Different data input layout used for the known data cells

It can be seen from Fig. 7 that the second type of layout with a balanced input data points works better than the others.

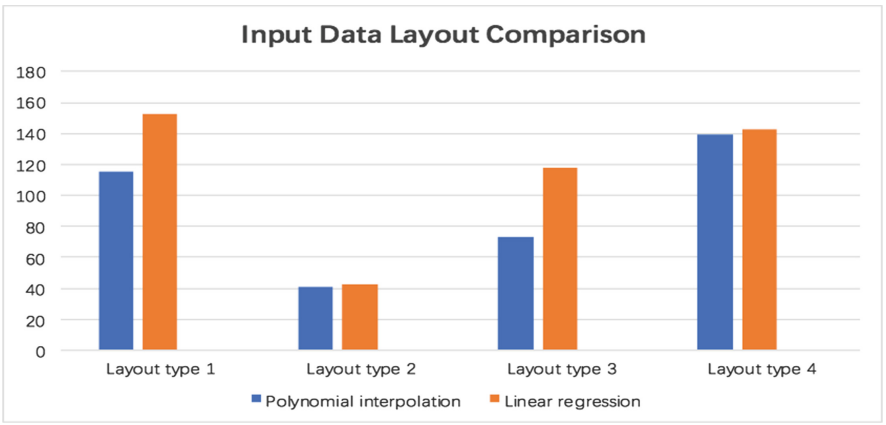
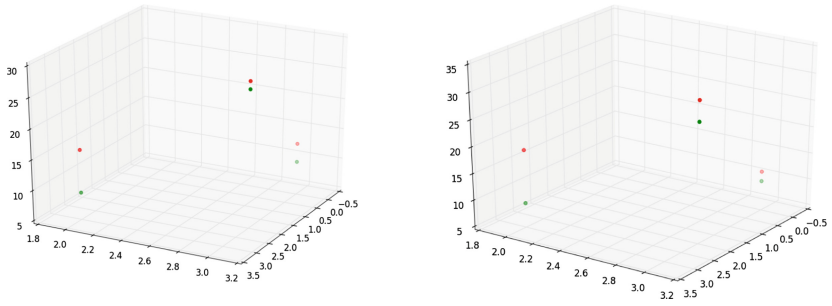


Fig. 7. Comparing the error rate of four different layouts

3.3 Experiment 3 – the Impact of Applying Special Features

As we analyze the real data set, it can be found that one factor that affects the air quality data is the highways and traffic, because the areas that are close to highways generally have slightly higher PM2.5 index than the areas without the highways. Therefore, in order to further improve the accuracy of the machine learning algorithm, we decided to add the area factor into the machine learning algorithm. Specifically, for each input data set, we also calculate the distance between the area and the highway, which will be added as the 3rd dimension to the data model. In this experiment, the locations of layouts of input air quality data are being concerned. Using the best two mathematical model from experiment 1, and best layout of air quality data from experiment 2, we can calculate the difference percentage of each model as shown in Fig. 8.



Polynomial interpolation:18.71% Polynomial interpolation: 33.94%

Fig. 8. Comparing the error rate of applying the special feature or not

From the result in experiment 1, we can conclude that SVM is not an appropriate mathematical method for predicting air data, considering the fact that SVM is typically used to perform classification, rather than the value calculation. As the result shown in experiment 2, a series of comparison are made to get the most accurate layout for predicting air data, and we found that the second one is the most accurate. After two series of experiments, we found out that the percentage difference with actual air data is still very high. So for the third experiment, we think that environment of the predicting place is important, a geo-based perimeters are made and added into the prediction. As the result the third experiment, a mathematical mothed with percentage error of 18.71% was the best mothed that can be used for predicting air quality.

4 Related Works

There are quite a few initiatives by organizations and researchers to perform air quality related model construction and monitoring.

U-Air [11, 12] – "when urban air quality inference meets big data", uses the concept of big data to infer air quality information. They showed the usage of five sensors in Beijing and shanghai, which is a distance about 2 h by plane, and through the data they collected for both the historical and real-time to infer a real-time air quality information. However, we used a more efficient way which only takes the real-time data from the censors and other factors, through machine learning to get a very detailed air-quality data. In addition, building the cost-effective PM 2.5 sensor is another key contribution from our work.

The website waqi.info [13] displays air quality on a google map. The source of the data used by this website is unknown. However, the main limitation of this website is that the granularity of the data is not fine enough to cover most of the small cities and suburb areas. This is the main motivation for our work to predict and present the air quality data in those areas.

5 Conclusion and Future Work

In this paper, we present a machine-learning based approach to predict the air quality value with a limited number of data input points. A prediction model can be found to predict the air quality with high accuracy as hypothesis stated at the beginning of the paper. As we can see from the result shown above, the following conclusions about these different factors can be made: The polynomial interpolation model with concerning of locations and layout of contain both data at corners and center are the best model in the combinations made in the experiments. When comparing the model with the linear regression or SVM model, it turns out that the polynomial interpolation was the best model because the air quality prediction is not a simple linear problem. In addition, when the locations were being concerned, the difference percentage between prediction result and real values changed dramatically, this is because the resources of the pollution were related with these locations.

The experimental design can be improved in the model selecting. We believe that more optimized and similar mathematical model might be found to compare, particularly using deep learning, which will be one of the major works for the near future.

References

1. Kyrkilis, G., Chaloulakou, A., Kassomenos, P.A.: Development of an aggregate air quality index for an urban mediterranean agglomeration: relation to potential health effects. Environ. Int. **33**(5), 670–676 (2007)
2. Schmidt, M.: Arduino: Pragmatic bookshelf (2011)
3. Raspberry Pi 3 Model B (2017). https://www.raspberrypi.org/products/raspberry-pi-3-model-b/
4. Firebase (2017). https://en.wikipedia.org/wiki/Firebase
5. Python Flask (2017). http://flask.pocoo.org/
6. Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000 **16**(10), 906–914 (2000)

7. Seber, G., Lee, A.: Linear Regression Analysis, vol. 936. Wiley, Hoboken (2012)
8. Michie, D.: Memo functions and machine learning. Nature **218**(5136), 19–22 (1968)
9. Xia, F., Yang, L., Wang, L., Vinel, A.: Internet of things. Int. J. Commun Syst **25**(9), 1101 (2012)
10. Goldberg, D., Holland, J.: Genetic algorithms and machine learning. Mach. Learn. **3**(2), 95–99 (1988)
11. Zheng, Y., Liu, F., Hsieh, H.: U-air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013)
12. Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L., Nath, B.: Real-time air quality monitoring through mobile sensing in metropolitan areas. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, p. 15 (2013)
13. Air Pollution in the World Real-time Air Quality Index (AQI) (2017). http://waqi.info/