Insights for Wellbeing: Predicting Personal Air Quality Index Using Regression Approach

Amel Ksibi[1], Amina Salhi[1], Ala Alluhaidan[1], Sahar A. El_Rahman[1,2]
[1] College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[2] Electrical Engineering Department, Faculty of Engineering-Shoubra, Benha University, Cairo, Egypt
amelksibi@pnu.edu.sa , Aisalhi@pnu.edu.sa, Asalluhaidan@pnu.edu.sa , sahr_ar@yahoo.com

ABSTRACT

Providing air pollution information to individuals enables them to understand the air quality of their living environments. Thus, the association between people's wellbeing and the properties of the surrounding environment is an essential area of investigation. This paper proposes Air Quality Prediction through harvesting public/open data and leveraging them to get Personal Air Quality index. These are usually incomplete. To cope with the problem of missing data, we applied KNN imputation method. To predict Personal Air Quality Index, we apply a voting regression approach based on three base regressors which are Gradient Boosting regressor, Random Forest regressor and linear regressor. Evaluating the experimental results using the RMSE metric, we got an average score of 35.39 for Walker and 51.16 for Car.

1 INTRODUCTION

Air pollution has an intensive impact on public health and the environment[1]. Providing air pollution information to individuals enables them to understand the air quality of their living environments. Thus, the association between people's wellbeing and the properties of the surrounding environment is an essential area of investigation[2]. In fact, public atmospheric monitoring stations in urban areas provide large quantities of global air quality data (GAQD) by deploying, across the globe, expensive high-end air pollution sensors. These data including weather data (temperature, wind) and air pollution data ($PM_{2.5}$, NO_2, O_3) collected over the city, have been investigated widely for general population[3]. However, on the scale of individual people and its personal wellbeing, these research investigations are too limited, leading to a broad low accuracy and low spatio-temporal resolution, when assessing the impact of air pollution on personal health.

With the plenitude of sensing devices, developing hypotheses about the associations within the heterogenous sensors data captured from these devices, contributes towards building effective models that make it possible to understand the impact of the environment on wellbeing at the individual scale. Such models are necessary since not all cities are fully covered by standard air pollution and weather stations. The critical research question here is whether we can use only data from open sources (e.g., weather, air pollution data) to predict the personal air pollution data.

However, it is not always possible to gather plentiful amounts of such data. As a result, a key research question remains open: Can sparse or incomplete data be used to gain insight into wellbeing? Meanwhile, machine learning techniques brought more opportunities for accurate prediction of air pollution [4]. Thus, it is compulsory to find new approaches based on data analytics for personal air quality prediction challenge.

The objective of this study was to evaluate the ability of regression approaches to predict individual air pollutants values and the air quality index (AQI).

Our paper is organized as follows. In Section 2, we present state of the art on air quality prediction methods. In Section 3, we discuss proposed process for air pollutant prediction. Section 4 analyses the results while Section 5 covers discussion and conclusion.

2 RELATED WORK

City-wide air quality prediction has been of interest over the past 40 years[3]. However, all these studies focused only on determining the air pollutants values at city scale for general population. At personal scale, recent investigations are focusing on crowdsourcing computing through harvesting data from wearable sensors[5]. These sensors provide lifelog data which can be classified into two categories: numerical data (weather data, environmental variables, GPS, time, health measurements, etc).

This study focuses on personal air quality prediction using numerical lifelog data. Personal air quality is a significant indicator when evaluating the air pollution impact on personal health [6]. Predicting the personal air quality has a main challenge that is developing an effective model based on a small amount of sparse or incomplete data training dataset. To deal with this issue, Zhao et al. [7] proposed a prediction model based on CRNN (convolution recurrent neural network) for short-term $PM_{2.5}$ pollution prediction utilizing the spatial-temporal features of atmospheric sensing data. The experiments conducted using the atmospheric sensing dataset from thirty-three coastal cities in China and Fukuokas environmental monitoring dataset during 2015 to 2017.

Zhao et al. [6] designed a transfer learning model using an encoder-decoder structure using decoder transfer learning (DTL) that based on the Wasserstein distance to match the atmospheric monitoring stations data that is the source domain heterogeneous distribution and the personal air quality that is the target domain.

MediaEval'20, December 14-15 2020, Online

The aforementioned methods focus on determining personal air quality index from various factors such as whether, GPS, and environmental data. In this paper, we aim to select the most important factors that influence the prediction of the personal air quality data.

3 METHODOLOGY

Our proposed process contains two steps: data preprocessing and then training a voting regressor to predict Personal Air Quality Prediction with public/open data.

3.1 Data preprocessing

The dataset used in this paper is Personal air quality dataset (PAQD) which is described in [5]. It contains weather data (e.g., temperature, humidity), atmospheric data (e.g., O_3, $PM_{2.5}$, and NO_2), GPS data, and multimedia data (e.g., images, annotation). Since the data quality and its representativeness play a crucial role in the effectiveness of prediction algorithm, we perform a process of data preprocessing to guarantee the quality of data. This process consists of missing data imputation, feature extraction and features selection.

a) Missing data imputation

Based on the hypothesis that there is a strong correlation of heterogeneous data recordings at the near-by location and time, we estimate that two recordings are close if the features that neither is missing are close. So, we can determine the values of missing features according to the mean value from the k nearest recordings. Indeed, we used sklearn.impute.KNNImputer to predict the missing values and we defined k=5.

b) Features extraction

Based on the assumption that the level of pollution may vary from one period to another on the same day and from one day to another in the same month and from one month to another in the same year, we extracted the following features from datetime component to enrich the learning model with temporal information: month number [1–12], day[1-31], hour of the day [0–23], minute[0-59].

c) Features selection

To select the most important features, we performed different combinations of features and we applied a simple regressor over the training dataset. According to the obtained results, whether data increases the RMSE. So, we decide to focus only on Time Data and GPS data to predict the values of pollutant variables O3, PM2.5, and NO2.

3.2 Personal Air Quality Prediction with public/open data

The Personal Air Quality Prediction can be represented as a regression problem where we are required to determine a continue value that is the AQI. Given the selected features, we apply a regression approach to estimate the value of each pollutant variable. For this issue, we test different regressor models over the training dataset and we obtain the best results with the voting regressor.

the voting regressor is an ensemble meta-estimator that fits several base regressors, each on the whole dataset. The algorithm then averages the individual predictions to form a final prediction. In our voting regressor, we opt for Gradient Boosting regressor , Random Forest regressor and linear regressor as base regressors. Gradient boosting regressor relies on a loss function to be optimized, a weak learner to make predictions, and an additive model to add weak learners for minimizing the loss function. This machine learning technique yields a prediction model usually by decision trees. A Random Forest Regressor is a technique that uses multiple decision trees and Bootstrap Aggregation to produce a more reliable prediction model. Linear regression, the most known regression analysis is based on a linear predictor function with unknown model parameters.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we report and discuss the experimental results achieved after submitting one run for the task1 "*Personal Air Quality Prediction with public/open data*". Table1 represents the official results for our run based on regression approach. The performance of the predictions was evaluated using root mean square error (RMSE). As can be seen in Table 1, SO_2 achieved the best results with score 12.08 using sensor data collected by walkers, while NO_2 showed the best results with score 25.02 using sensor data collected by car. Moreover, we can see that the obtained results for AQI from walker data outperforms those obtained from car data. This can be a clue that the quality of sensor data collected by walkers outperforms the quality of data collected by Car.

Table 1: Official results of the submitted run

	$PM_{2.5}$ RMSE	NO_2 RMSE	O_3 RMSE	AQI RMSE
AVG walkers	35.34	25.98	12.08	35.39
AVG car	40.93	25.02	35.98	51.16

5 CONCLUSIONS

This paper represents our first attempt to address the task "Personal Air Quality Prediction with public/open data". The proposed solution was based on data preprocessing and training voting regressor based on three base regressors. The obtained results demonstrate the quality of sensor data collected by walker. As future work, we would investigate on transfer learning over multimedia lifelog data such as egocentric photos and videos to get insights about individual wellbeing.

ACKNOWLEDGMENTS

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number PNU-DRI-RI-20-033.

REFERENCES

[1] Song, H., Lane, K. J., Kim, H., Kim, H., Byun, G., Le, M., Choi, Y., Park, C. R., & Lee, J. T. (2019). Association between Urban Greenness and Depressive Symptoms: Evaluation of Greenness Using Various Indicators, International journal of environmental research and public health, 16(2), 173.

[2] P. Vo, T. Phan, M. Dao and K. Zettsu, Association Model between Visual Feature and AQI Rank Using Lifelog Data, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 4197-4200

[3] Y.Xu, W.Yang, and J.Wang, "Air quality early-warning system for cities in china," Atmospheric Environment,vol.148,pp.239–257,2017.

[4] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, andM. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," IEEE Access, vol. 7, pp. 128 325–128 338, 2019.

[5] Dao, M. S., Zhao, P. J, Nguyen, N.T., Nguyen, T.B., Dang-Nguyen D. T., Gurrin, C., "Overview of mediaeval2020: Insights for wellbeing task - multimodal personal health lifelog data analysis," in MediaEval Benchmarking Initiative for Multimedia Evaluation, CEUR Workshop Proceedings, Dec 2020.

[6] Zhao, P. and Zettsu, K., Decoder Transfer Learning for Predicting Personal Exposure to Air Pollution, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 5620-5629.

[7] Zhao, P. and Zettsu, K., Convolution Recurrent Neural Networks for Short-Term Prediction of Atmospheric Sensing Data, The 4th IEEE International Conference on Smart Data (SmartData 2018), pp.815-821