Machine learning algorithms for predicting air pollutants

Jirat Boonphun, *Chalat* Kaisornsawad, and *Papis* Wongchaisuwat*

Industrial Engineering Department, Kasetsart University, 50 Ngamwongwan Rd, Ladyao Chatuchak Bangkok, Thailand

Abstract. An atmospheric particular matter, commonly recognized as PM, contains solid particles and liquid droplets suspending in an ambient air. A high concentration of PM is known to seriously cause adverse health effects to humans especially a small-sized particle, known as PM2.5. Not only health effects, environmental effects are also obviously observed. This work aims to estimate a likelihood of PM2.5 exceeding a pre-defined safety threshold. Multiple machine learning models are explored in this work. Particularly, classification models are implemented based on meteorological data and air pollutant features measured at different altitudes above a ground level. These features are shifted back to various time steps resulting in more insightful time-lagged features. Furthermore, a feature selection technique is implemented to specify a desirable set of important features. A re-sampling technique is also employed to address an unbalancing level of the response value in an original data set. The proposed models are evaluated on a case study whose data set is collected from an air monitoring station located in Bangkok, Thailand.

1 Introduction

A particular matter or PM represents a mixture of solid and liquid such as dust and ash in the air. It normally varies in a diameter or a width of the particle. In this study, we focus on PM2.5 referring to an atmospheric particular matter whose size is generally less than 2.5 micrometres. Fine particles like PM2.5 are more likely to stay longer in the air compared to other heavy particles. As a result, humans have a higher chance to inhale these fine particles into their bodies leading to adverse health effects [7]. A strong relationship between an exposure to fine particles and a mortality risk from lung or heart diseases [18] is observed. Once an amount of PM2.5 reaches a healthy threshold, it is strongly advisable to take an action to protect human health.

Most crowded cities are prone to face air pollution problems. There is no exception for Bangkok, the capital city of Thailand. According to the study by Chuersuwan et. al. [5], primary sources of PM2.5 in Bangkok are generally an automobile and a biomass burning in both traffic and residential sites. Primary problems include smog and respiratory issues whose severity has become worse during the past few years. Recently, Bangkok experienced so much smog due to PM2.5 that the government had to issue a serious health warning to residents. These motivate us to develop models to estimate an hourly concentration of an air pollutant in advanced. As a safety precaution, these predictive models are considered as a basis for applying successful pollution control processes.

Significant factors potentially impacting air pollutant concentrations include meteorological conditions such as an air temperature, a wind speed, a wind direction as well as a relative humidity [8]. High concentrations of traffic-related pollutants are strongly correlated to a low wind speed. Moreover, high concentrations of the particular matter are commonly related to high humidity in the air [10]. We further speculate that a relative humidity has a strong relationship with an amount of precipitation which in turn impacts air pollutant concentrations. In addition, we are interested in studying how PM2.5 changes with respect to other air pollutants. Due to these reasons, meteorological data and various air pollutants are utilized as initial features in forecasting PM2.5 concentrations.

In this study, we pay more attention to a situation where the amount of PM2.5 exceeds a pre-defined healthy threshold. In response, classification models are proposed in order to predict whether PM2.5 is above the safety limit. These models rely on initial features including meteorological data and multiple air pollutants which are further utilized to generate additional features. In particular, these features are shifted to different time periods resulting in various time-lagged features. Due to time-lagged features, significant number of features are retrieved as a result. In order to address this issue, a feature selection technique is implemented which controls total number of features used in the models.

We also observe that a number of instances where PM2.5 value exceeds the threshold is relatively low. In response, we employ a re-sampling technique to up-sample the minority class and down-sample the majority class prior to actually fitting the models. Classifiers including Naïve Bays, Logistic regression, Random forest, and Neural networks are implemented. To evaluate the performance of the models, traditional metrics such as a confusion matrix and F1-score are utilized.

As a case study in this work, a data set collected from an air monitoring station at Kasetsart University located in Bangkok is used to train and test the models. While

[*] Corresponding author: fengppwo@ku.ac.th

PM2.5 is measured at a fixed height above a ground level, meteorological and air pollutant features are collected at various altitudes. Instead of focusing only at ground-level information, the data set is specifically gathered along a vertical basis, i.e. at 30, 75, and 110 meters above the ground. We speculate that data collected at different heights potentially contains useful insights in forecasting the PM concentrations. Hence, we consider information measured at 30 meters as a low level while the rest are counted as a high level. To verify our assumption, features with high score due to the feature selection technique are thoroughly observed.

Features as well as an application of the models distinguish our work from others. Our main contribution is exploring other insightful features that have not been studied in prior works. Particularly, we consider meteorological and air pollutant concentrations locally measured at different heights above the ground. In addition, we introduce features at various time steps which potentially capture deeper insights. The feature selection technique is also performed to include only significant features. To address an imbalance problem observed in our data set, the re-sampling technique is employed. To the best of our knowledge, none of prior works applied similar machine learning algorithms based on a data set collected in Bangkok. Our data set is specifically gathered at the particular air monitoring station in an exceptional fashion, i.e., measuring information at 3 different heights above the ground.

We summarize the literature in Section 2. The features, the models and evaluation metrics used to test the models are thoroughly discussed in Section 3. Further descriptions regarding data collections and preparations are illustrated in Section 4. Results of our algorithms including all experiments are reported in Section 5. Discussions are provided in Section 6 while a conclusion is stated in the last section.

2 Related work

Air pollution is one of the most critical issues in many cities globally due to its adverse health effects. Bellinger et. al. [2] conducted a systemic review of applying data mining and machine learning techniques in an air pollution epidemiology. According to [2], all 400 reviewed articles are separated into 3 main research areas which include a source apportionment, a prediction of air pollution or quality or exposure, and a hypothesis generation. Our work closely aligns with the second category. An overview of various air pollution forecasting algorithms is provided in [1]. Specifically, Bai et. al. extensively reviewed theory and applications of multiple predictive models as well as further compared advantages and disadvantages among models.

The particular matter or PM especially PM2.5 is commonly counted as one of the most detrimental pollutants posing a great threat to human health. Therefore, our work primarily focuses on PM2.5 only. Substantial research has been done for developing predictive models to estimate the PM2.5 concentrations in advanced. Most of these work aim to directly forecast the value of the PM2.5 [17] [16] [9] [19] [6] [14] [12] [11] [21] [20] [15]. Instead, our study focuses on determining a likelihood of the PM2.5 exceeding the safety limit. Another area of research that is closely related to us is predicting the ground PM2.5 concentrations based on satellite data. Significant number of works had been done along this line as summarized in [4]. Comparing with these works, our study is based only on locally collected features.

A few research have been done for speculating air pollutant concentrations in Thailand. Kanabkaew [13] predicted hourly PM concentrations in Chiangmai, Thailand based on MODIS and ground-based meteorological data. While temperature and relative humidity as ground-based meteorological data and satellite data are used in [13], our work relies on additional ground-based information and other pollutant concentrations. Furthermore, [13] performed simple and multiple linear regression models but we instead use more complex classification models for predicting the likelihood of PM2.5 exceeding the threshold. In addition, Amphanthong and Busbabodhin conducted a study to predict PM10 in the northern part of Thailand based on Grey system [22].

3 Methodology

In this study, we aim to estimate an hourly fine particular matter (PM2.5) concentration in Bangkok, Thailand. Most urban populations can be adversely affected by a greatly exposure to air pollutants like PM2.5 in an ambient air. We pay more attention to the case where the pollutant concentrations are above an air quality standard. Our problem is scoped down to identify whether the PM2.5 concentration exceeds a given threshold. Several classification models are implemented.

Our proposed models are based on meteorological data measured at different height above a ground surface as well as other air pollutants as initial features. Taking into account different time period of these features results in large number of features. We hence apply a dimensionality reduction technique before applying the classification models. The re-sampling technique is also implemented to handle an imbalance data set like ours. Evaluation metrics are discussed in detailed next.

3.1 Features

Meteorological and air pollutants features used in the models are initially collected at 3 different heights which are 30, 75, and 110 meters above a ground level. At each level, we gather 4 meteorological features including Wind Speed (WS), Wind Direction (WD), Temperature (TEMP), and Relative Humidity (RH). In addition, 6 air pollutants consisting of Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Nitrogen Oxide (NOX), Sulfur Dioxide (SO_2), Carbon Monoxide (CO), and Ozone (O_3) are considered. In summary, there are 10 features for each height level resulting in 30 features in total.

We consider multiple time steps corresponding to these features. In particular, a hyperparameter n_step is defined as a number of time steps which we shift each feature back. According to a statistical test, F-value is used to measure a level of dependency between features and the response. We introduce a hyperparameter n_feature to specify a desirable number of features. In particular, we construct the final set of features by selecting top n_feature with highest F-values.

3.2 Re-sampling

The likelihood of PM2.5 exceeding the safety threshold is low among all instances in the data. Due to an imbalance of the data set, it is relatively challenging for any algorithm to provide a good performance. We employ a re-sampling technique to reduce the unbalancing level within the original data set in order to enhance the predictive power of the models. Specifically, Synthetic Minority Over-sampling Technique (SMOTE) is used to oversample the minority class and undersample the majority class [3]. For over-sampling, a random point along a line segment between the minority class sample and its k nearest neighbors is constructed. On the other hand, the majority class samples are randomly removed in order to under-sample. In our case, PM2.5 exceeding the pre-defined threshold is considered as the minority class while the rest is the majority.

3.3 Classification models and evaluation metrics

We utilize classification models to predict the likelihood of PM2.5 exceeding the safety threshold. Multiple supervised learning algorithms including Naïve Bays, Logistic regression, Random forest, and Neural networks are experimented with. The predicted probability associated with each class is achieved from the classification models. Multiple cut-off probabilities to identify the predicted class are also thoroughly investigated. We then compute traditional metrics such as the confusion matrix as well as the F1-score to evaluate the performance of these models. In order to compare among models, the F1-score is mainly considered. Different hyperparameters are calibrated to maximize the performance of the models.

4 Data preparation

All meteorological data and air pollutants are collected from the KU tower located at Kasetsart University, Bangkok, Thailand. Data collectors are installed at 30, 75 and 110 meters above the ground level in order to measure meteorological information and air pollutants in a vertical basis. These raw data are collected at every minute from 2015 to early 2019. We use an average value of each minute in the same hour as a representation of an hourly data. In summary, our data set has 36624 instances in total with 30 features at each time stamp. Among these instances, there are 2203 cases whose PM2.5 concentrations are above the threshold value which is set to 50 micrograms per cubic meter ($\mu g/m^3$). We label this case as “1” class in our binary classification model. This is accounted for approximately 6 percent.

In the original data set, numerous values are missing or irregular due to an error from data collectors. In practice, irregularities can be handled differently such as replacing with 0 or any statistical value. In contrast, our data set is in a chronical order; therefore, we propose to replace each irregularity with some default value. These default values are specifically computed based on an average of all remaining values in the last 3 hours after ignoring irregularities.

5 Results

We calibrate the hyperparmeters in order to obtain the best models’ performance. A ratio between the number of samples in the majority class over the number of samples in the minority class after resampling with respect to the SMOTE technique is set to 0.5. For the neural network classification model, 1 hidden layer with 512 hidden nodes are selected after extensively experiments. The cut-off probability of 0.5 is particularly chosen.

F-value is used to measure the significance of each feature in the models. Table 1 illustrates top 30 features based on F-value. Top n_feature with high F-value are finally selected. Different values of n_step and n_feature are experimented among all 4 models while F1-score is used as an evaluation metric. A comparison of F1-score among different models and their corresponding parameters is shown in Table 2. Additionally, the confusion matrix evaluated on 1832 instances in a test set corresponding to the best model, i.e. the Random forest with 1 time step employing top 20 features is provided in Table 3.

Table 1. Top 30 features with high F-value

PM2.5(t-1)	H30m_ NO2(t-3)	H30m_ NOX(t)	H110m_ SO2(t-3)	H75m_ SO2(t-1)
PM2.5(t-2)	H75m_ NO2(t-1)	H75m_ NO2(t-2)	H110m_ NO2(t-1)	H110m_ NO2(t-2)
PM2.5(t-3)	H110m_ SO2(t)	H110m_ SO2(t-2)	H30m_ NOX(t-3)	H75m_ NOX(t-2)
H30m_ NO2(t-1)	H75m_ NO2(t)	H30m_ NOX(t-2)	H75m_ SO2(t)	H75m_ SO2(t-2)
H30m_ NO2(t)	H110m_ SO2(t-1)	H75m_ NO2(t-3)	H75m_ NOX(t-1)	H75m_ SO2(t-3)
H30m_ NO2(t-2)	H30m_ NOX(t-1)	H110m_ NO2(t)	H75m_ NOX(t)	H110m_ NO2(t-3)

Table 2. A comparison of F1-score among models and their corresponding hyperparameters

Model	n_step	n_feature	F1-score
Naïve bays	2	10	0.6087
Logistic regression	3	25	0.7605
Random forest	1	20	0.8178
Neural network	1	20	0.7788

Table 3. The confusion matrix of the Random forest where n_step = 1 and n_feature = 20

	Predicted : No	Predicted : Yes
Actual : No	1699	29
Actual : Yes	12	92

6 Discussion

In this paper, classification models to predict the likelihood of PM2.5 exceeding the healthy threshold are proposed. We implement 4 classification models including Naïve Bays, Logistic regression, Random forest and Neural network. These models achieve reasonably good F1-score ranging from 0.6 to 0.8 with their best set of parameters. Among different models and various parameters, the Random forest model with 1 time step (n_step = 1) utilizing top 20 features provides the best performance with 0.82 F1-score.

Compared to previous works, we include novel features in the models such as meteorological and air pollutants measured at different altitudes above the ground level. To evaluate a significance of features at the high level, F-values corresponding to these features are computed. Among top 30 features with high F-value, multiple features measured at high altitudes are selected. This implies an importance of including these high-level features into the models. According to the results, various features at different time steps play a role in the models due to their high F-value. Specifically, PM2.5 itself at previous time steps, NO2, NOX and SO2 pollutants are among the top important features.

The data set used as a case study is particularly collected from the air monitoring station located in Bangkok, Thailand. From the best of our knowledge, there is no previous work which conduct a study in the same direction as our work especially for the data set locally collected in Thailand. In this data set, a severity of an unbalance level between classes is relatively high. The re-sampling technique SMOTE is implemented to address this problem. Referring to our experiments, the models with SMOTE yield better performance than the ones without SMOTE. Even though only one case study is used to evaluate the performance of the classification models in this study, the methodology including machine learning models, features and relevant techniques can be applied and generalized to other data sets.

7 Conclusion

The proposed classification models aim to predict the likelihood of PM2.5 exceeding the pre-defined threshold. Meteorological data and air pollutants features are specifically considered at different levels above a ground along a vertical axis. Adding additional features which further draw insightful information from the data potentially enhances the predictive power of the models. Higher performance of the prediction can possibly be achieved if more complete and accurate data is available. With a larger data set, more complicated models such as deep learning approaches possibly enhance an accuracy of the prediction.

We are grateful to the Environmental Faculty, Kasetsart University for providing the meteorological data and air pollutants information.

References

1. L. Bai, J. Wang, X. Ma, H. Lu. Air pollution forecasts: An overview. Int. J. Environ. Res. Public Health, **15**, 780 (2018)
2. C. Bellinger, MS. Mohomed Jabbar, O. Zaiane, A. Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health **17**, 907 (2017)
3. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. J. Artitif. Intell. Res. **16**, 321-357 (2002)
4. Y. Chu, Y. Liu, X. Li, Z. Liu, H. Lu, Y. Lu, Z. Mao, X. Chen, N. Li, M. Ren, F. Liu, L. Tian, Z. Zhu, H. Xiang. A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. Atmosphere **7**, 129 (2016)
5. N. Chuersuwan, S. Nimrat, S. Lekphet, T. Kerdkumrai. Levels and major sources of PM2.5 and PM10 in Bangkok Metropolitan Region. Environ. Int. **34**, 671-677 (2008)
6. W.G. Cobourn, An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmos. Environ. **44**, 3015-3023 (2010)
7. L. Curtis, W. Rea, P. Smith-Willis, E. Fenyves, Y. Pan. Adverse health effects of outdoor air pollutants. Environ. Int. **32**, 815-830 (2006)
8. A.T. DeGaetano, O.M. Doherty. Temporal, spatial and meteorological variations in hourly PM2.5 concentration extremes in New York City. Atmos. Environ. **38**, 1547-1558 (2004)
9. M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, D. Kenski. PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Syst. Appl. **36**, 9046-9055 (2009)
10. H.K. Elminir. Dependence of urban air pollutants on meteorology. Sci. Total Environ. **350**, 225-237 (2005)
11. X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, J. Wang. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. **107**, 118-128 (2015)
12. Z. Haiming, S. Xiaoxiao. Study on Prediction of Atmospheric PM2.5 Based on RBF Neural Network. *2013 Fourth International Conference on Digital Manufacturing & Automation.* IEEE. 1287-1289 (2013)
13. T. Kanabkaew. Prediction of hourly particulate matter concentrations in Chiangmai, Thailand using MODIS Aerosol optical depth and ground-based meteorological data. Environ. Asia **6**, 65-70 (2013)
14. I.G. McKendry. Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM10 and PM2.5) Forecasting. J. Air Waste Assoc. **52**, 1096-1101 (2002)

15. D. Mishra, P. Goyal, A. Upadhyay. Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India. Atmos. Environ. **102**, 239-248 (2015)
16. J.B. Ordieres, E.P. Vergara, R.S. Capuz, R.E. Salazar. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Model. Softw. **20**, 547-559 (2005)
17. P. Perez, A. Trier, J. Reyes. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. **34**, 1189-1196 (2000)
18. C.A. Pope, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, G.D. Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. **287**, 1132-1141 (2002)
19. W. Sun, H. Zhang, A. Palazoglu, A. Singh, W. Zhang, S. Liu. Prediction of 24-hour-average PM2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. Sci. Total Environ. **443**, 93-103 (2013)
20. D. Voukantsis, K. Karatzas, J. Kukkonen, T. Rasanen, A. Karppinen, M. Kolehmainen. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. **409**, 1266-1276 (2011)
21. Q. Zhou, H. Jiang, J. Wang, J. Zhou. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. 496, 264-274 (2014)
22. P. Amphanthong and P. Busababodhin. Forecasting PM10 in the upper northern area of Thailand with grey system theory. Burapha Science Journal. 20, 15-24 (2015)