A2QI: An Approach for Air Pollution Estimation in MediaEval 2020

Dat Q. Duong[1,2,3,*], Quang M. Le[1,2,3,*], Dat Nguyen[1,2,3]
[1]AISIA Research Lab, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[*]Two first author have equal contribution
dat181197@gmail.com

ABSTRACT

In this paper, we present our AISIA team's contribution to the task *Insight for Wellbeing: Multimodal personal health lifelog data analysis* at *MediaEval 2020*. From the data sets provided, we extracted different types of useful attributes for the problem: the timestamp information, the geographical data, sensor data, and the semantic features from images captured by users. We proposed an approach, namely A2QI, by applying machine learning models for estimating the local AQI score and level, including Support Vector Machine and Random Forest. We evaluated the experimental data sets using Randomized Search and K-Fold cross-validation. The test sets' evaluation shows that employing a machine learning approach with appropriate features can significantly improve accuracy.

1 INTRODUCTION

In many countries worldwide, the prediction of air pollution is an increasingly undeniably significant problem. It can impact individuals and their wellbeing. In this study, we aim to use a machine learning approach using insights from the lifelog data provided by the organizer to predict the personal air pollution data as well as the individual air quality data, as given in the task description [5] of the competition MediaEval 2020. This task's primary motivation is to investigate the association between people's wellbeing and the surrounding environment's properties. The problem consists of two subtasks. In the first subtask, we explore the correlation between the air pollution data with the features we extracted from the sensor (e.g., timestamp information, the user's geographical location). In the second subtask, we utilized the features mentioned earlier, together with the semantic features extracted from cameras by users, to predict six pollutants used to calculate the AQI values.

2 OUR APPROACH

2.1 Anomaly detection

Observing the three columns of $PM_{2.5}$, NO_2, and O_3 in the training dataset for both tasks, one can see that many data points have zero value, are negative numbers, or are unreasonably large (e.g., −3000, −4900, etc.). Also, one can find a similar observation even in positive-valued data points. They are called anomalies or outliers, which have to be preprocessed before extracting features.

MediaEval'20, December 14-15 2020, Online

Now, let us consider an arbitrary column whose data needs to have a preprocessing step. One can determine these outliers in two cases: the first one includes zero and negative signed values, the other includes positive outliers (which will be defined later). For the positive outliers, we apply z-score method [1][4]. Specifically, if we consider the i^{th} qualitative data point (denoted by x_i) in the column, the formula for computing its z-score (denoted by Z_i) can be given as $Z_i = \frac{x_i - \overline{x}}{s}$, where s and $\overline{x}$ are the sum and mean value of the column, respectively.

In this work, a data point whose z-score is larger than 3.0 is called an outlier. It is worth noticing that the mean value is computing based on the positive values only, intending to avoid the influence of negative valued data points whose absolute values are large.

After detecting all the anomalies, we replace them with the average of positive values via the reason mentioned above.

2.2 Features Extraction

The problem consists of two subtasks. Each task asks for using a different data set. Nevertheless, they both include information about time, location, weather, and concentration values of contaminants related to AQI (e.g., NO_2 or O_3). Therefore, our proposed feature extraction techniques in these data types can be applied to both data sets. Also, we calculated the necessary features from the image data given in the second task.

2.2.1 Timestamp features. From the given information about time, we extract timestamp features. Specifically, we survey the correlation between the time point that the data are collected and the corresponding AQI values and ranks that need to be predicted. These features include *part of day* (*POD*) and *is rush hour* (*isRH*).

To begin with, we deduce the *POD* feature. That is, we split a day's 24-hour time into five groups, The "Early Morning" group is for the time from 5 AM to before 7 AM, the time from 7 AM to before noon is considered the "Morning" group, between noon and before 4 PM is "Afternoon" group, between 4 PM and before 8 PM is "Evening", and the remaining period between 8 PM and 5 AM is the "Night" group. From our observation, there is a noticeable increase in traffic density during the time of *Morning* and *Evening* groups, which leads to a high level of pollution caused by smoke from these means of transportation. Consequently, we expect there is a fluctuation in the data collected during these periods.

Also, we check whether a particular local measured time is a rush hour or not, which leads to extracting the second feature in the group of timestamps features, i.e., *is rush hour* (*isRH*). In

Table 1: Test results from various runs

				PM25			NO2			O3			AQI		
Task	Run	Method	Type	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
1	1	SVM	walker	**4.90**	**5.88**	**0.56**	15.31	17.92	0.50	**9.05**	**11.33**	**0.63**	**12.93**	**15.96**	**0.32**
			car	11.00	15.73	0.71	**12.22**	**14.65**	**0.35**	27.02	32.04	0.83	26.70	35.49	0.52
	2	RF	walker	**5.10**	**6.03**	**0.57**	**14.81**	**17.91**	0.50	**9.14**	**11.36**	**0.63**	**12.74**	**15.93**	**0.32**
			car	10.70	14.92	0.68	15.38	18.95	**0.41**	27.42	32.69	0.86	26.46	33.98	0.52
2	1	SVM	courses 1-4	**3.49**	**3.76**	**0.15**	**7.19**	8.68	**0.57**	**15.69**	**17.17**	**0.57**	-	-	-
	2	RF		4.57	5.43	0.20	7.70	**8.55**	0.60	15.82	17.99	0.58	-	-	-

detail, if that given point of time falls into one of these periods (7:00 AM to 9:00 AM) and (4:00 PM to 7:00 PM), it is called a rush hour. This feature is a development of the former (i.e., *POD*). We will survey the periods when the traffic density reaches the highest peak, resulting in sharp growth of AQI values and ranks.

2.2.2 Location features. When surveying the factors affecting the level of pollution of a location, we consider the distance between that location and the nearest railway station, which is usually crowded with people and transports. Using the information about coordinates of a place, we extract the feature about the distance from that place to the chosen station. In this study, we use the Shibuya station (35°39′N, 139°42′E).

To compute the mentioned distance, we use the Haversine formula[2]. That is, given coordinates of two points A and B, the distance between them can be calculated as follows:

$$d(A,B) = 2 \cdot r \cdot \arcsin\left(sin^2\left(\frac{\varphi_B - \varphi_A}{2}\right) + cos(\varphi_A).cos(\varphi_B).sin^2\left(\frac{\lambda_B - \lambda_A}{2}\right)\right)^{\frac{1}{2}}, \quad (1)$$

where r is the Earth's radius, and φ_A, λ_A, φ_B, λ_B are the latitudes and longitudes of two points A and B, respectively.

2.2.3 Semantic features. In the second task, we are provided the data of images captured in different locations, which is the most challenging data type in our opinion. Our approach is to investigate if the number of cars, motorbikes, and the contrast of the images can impact the level of pollution in that captured location. We used SSD ResNet 50 (Retina Net 50)[7], a pre-trained object-detection model, to extract the mentioned features from the images of the data set.

Also, we extract features related to the contrast of the images, which can be highly correlated to the intensity of a given place's pollution. In detail, given a two-dimensional image I of size $M \times N$, we use RMS contrast formula[6] to compute its contrast. The mentioned formula can be seen in the equation (2)

$$RMS = \sqrt{\frac{1}{M.N}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left(I_{ij} - \bar{I}\right)^2} \quad (2)$$

where RMS is the contrast value that needs computing, I_{ij} is the intensity pixel of the image I at point (i, j), and $\bar{I}$ is the average intensity of all the pixels in that image.

Finally, it is worth noticing that in this study, we did not use the number of people as a feature related to image data, as the people appearing in the given images have been blurred for the sake of privacy.

3 RESULTS AND DISCUSSION

After extracting the necessary information, we evaluated two machine learning models using a Randomized Search with a 5-fold cross-validation technique to optimize the model hyper-parameters and avoid overfitting our training data. The two models we used were Support Vector Machine (SVM) [3], Random Forest (RF) [8]. It is crucial to note that we also tested other machine learning methods, e.g., Linear Regression, XGBoost, and CatBoost, and chose the two best performing models on the training data for submission. Each model is optimized and evaluated separately using different data set of each subtask. Only timestamp and geographical features were used for subtask 1, and the semantic features were combined with other feature types for subtask 2. The machine learning models were optimized based on the mean absolute error (MAE) metric.

The results on test sets are presented in Table 1. In the first subtask, we can see that using Random Forest can achieve the best result in general with data collected by a walker. For predicting the AQI value, the results of MAE, RMSE, and SMAPE, in this case, are 12.74, 15.93, and 0.32, respectively. In the second task, the best result can be achieved by using SVM. For predicting $PM_{2.5}$, the best performance in MAE, RMSE, and SMAPE are 3.49,3.76 and 0.15, respectively.

Also, if one can enhance the quality of the images captured in the data set and combine it with public weather data, the training results can be improved significantly.

4 ACKNOWLEDGEMENT

As the authors, we would like to thank AISIA Research Lab to support our team and allow us to use their computational resources for this study. Also, we would like to give our thanks to the Organization Board of MediaEval 2020 competition and Task Organizer for providing us with data sets to conduct necessary experiments.

REFERENCES

[1] 2019. Detecting Outliers in High Dimensional Data Sets using Z-Score Methodology. *International Journal of Innovative Technology and Exploring Engineering* 9, 1 (Nov. 2019), 48–53. https://doi.org/10.35940/ijitee.a3910.119119

[2] M. Basyir, M. Nasir, Suryati Suryati, and Widdha Mellyssa. 2018. Determination of Nearest Emergency Service Office using Haversine Formula Based on Android Platform. *EMITTER International Journal of Engineering Technology* 5, 2 (Jan. 2018), 270–278. https://doi.org/10.24003/emitter.v5i2.220

[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. In *Machine Learning*. 273–297.

[4] Denis Cousineau and Sylvain Chartier. 2010. Outlier detection and treatment: a review. *International Journal of Psychological Research, ISSN 2011-7922, Vol. 3, Nº. 1, 2010, pags. 58-67* 3 (01 2010).

[5] Dao, M. S., Zhao, P. J, Nguyen, N.T., Nguyen, T. Binh, Dang-Nguyen D. T., Gurrin, C. 2020. Overview of MediaEval 2020: Insights for Wellbeing Task - Multimodal Personal Health Lifelog Data Analysis. In *MediaEval Benchmarking Initiative for Multimedia Evaluation, CEUR Workshop Proceedings*.

[6] Heljä Kukkonen, Jyrki Rovamo, Kaisa Tiippana, and Risto Näsänen. 1993. Michelson contrast, RMS contrast and energy of various spatial stimuli at threshold. *Vision research* 33 (08 1993), 1431–6. https://doi.org/10.1016/0042-6989(93)90049-3

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[8] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1. 278–282 vol.1.