HCMUS at Insight for Wellbeing Task 2019: Multimodal Personal Health Lifelog Data Analysis with Inference from Multiple Sources and Attributes

Hoang-Anh Le, Thang-Long Nguyen Ho , Minh-Triet Tran*
Faculty of Information Technology, University of Science, VNU-HCM, Vietnam
1612013@student.hcmus.edu.vn,nhtlong@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

When collecting and processing data recorded by sensors for any applications, noisy and missing data is an important problem that need to be address. This paper presents two approaches we use to predict missing air quality data in MediaEval Insight for Wellbeing Task. The first approach based on other data attributes like temperature and humidity, and the second based on data recorded from other sources. Evaluating the experimental results using the average L2 distance, we got the score of 0.9013 for the first approach and 0.0155 for the second approach.

1 INTRODUCTION

Environmental data can be used to analyse different aspects for the deveopment of the society, including the quality of personal health [2] or depressive symptoms [3]. The data can be of various sources and formats, such as spatialtemporal raster images [1] or a combination of weather, air polution, lifelog images, etc [2].

In the MediaEval Life Well Being 2019 task[2], we are given 14 categories of pollution data recorded by people who wear sensors, use smartphones and walk along pre-defined routes inside a city, and asked to develop methods that process the data to obtain insights about personal wellbeing. In subtask 1, our goal is developing a hypothesis about the associations within the heterogeneous data and build a system that is able to correctly replace segments of data of the $PM_{2.5}$ index that have been removed.

Based on the organization of the data, we found there are 2 main approaches to predict the $PM_{2.5}$ in the queries. In the first approach, we want to explore if it would be possible to find relationship between $PM_{2.5}$ values and other obtained attributes (Section 2.1). In the second approach, because in each question, there are a number of people walking in the same region in roughly the same time interval, we propose to combine values from multiple people, to infer the missing data segment of $PM_{2.5}$ (Section 2.2).

2 METHOD

2.1 Inference from other attributes

2.1.1 Using test-set only. After observing over the chart of a dataset, we omitted some features have a mean value close to zero like NO_2,3_3, some category features. We found the feature about the location of all users at any time are not so different, so we concluded that data about location and $PM_{2.5}$ are not had close relationships. Because we don't want unexplainable-relationship between coordinates and $PM_{2.5}$, we need a solution clear and stable as much as we can. We do not have adequate data about location and $PM_{2.5}$, also any pretrain model to mapping from location information to what we need. We assume coordinates value have a relationship with other attributions so coordinates' meaning can be implicitly represented through temperature and humidity feature [Figure 1], so that if we found the right function to mapping from temperature and humidity to $PM_{2.5}$, we also have the coordinates information in the result, also simplify the data. As a result, we push normalized temperature and humidity data through a multi-layer perceptron model to approach the problem.

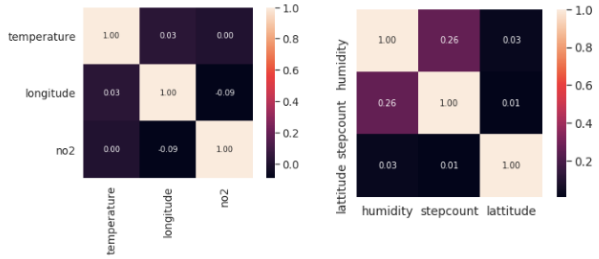

Figure 1: Top 3 in correlation heatmap on temperature and humidity

2.1.2 Using test-set as validation. We do the same as first run at this run, however, the motivation of this run is that we do not try to overfit the testing dataset of organizing, in this approach we try to generalize the method. We use the dataset development (unrated) in the contest to train set, the official dataset to the validation set. This task is preprocessed data most clean-able and optimize the loss on validation (official dataset).

2.2 Inference from other people

First, we examine and compare the coordinates and trajectories of people within the same group (same question) through time, and find that in most cases, people in the same group walked in roughly the same route, and they were at the same location together at every moment along the way (the start and end times of each person may vary) [Figure 2].

Therefore, we can conclude that given a specific time, the $PM_{2.5}$ values recorded by people within the same group are highly related because they recorded the $PM_{2.5}$ value of the same location at the same time, and we could guess the missing $PM_{2.5}$ values by the corresponding $PM_{2.5}$ values of other people in the same group.

*The first two authors contributed equally to the paper

MediaEval'19, 29-31 October 2019, Sophia Antipolis, France

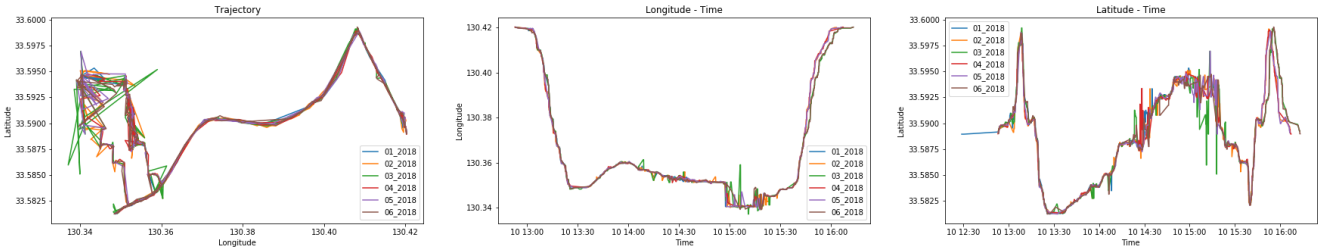

Figure 2: Trajectories and coordinates through time of people in Query Q1

However, comparing $PM_{2.5}$ values of all people in the same group, we find that these values vary considerably. Thus, we implement some statistical method to predict missing $PM_{2.5}$ values from corresponding $PM_{2.5}$ values of other people.

2.2.1 Average. We predict the missing $PM_{2.5}$ values by taking the average of $PM_{2.5}$ values of other people in the group in the corresponding time. [Figure 3]. However, the $PM_{2.5}$ data of these people are scatter over the time interval and not available for every second. There for we use 1D linear interpolation to predict $PM_{2.5}$ data for each person at every second before taking the average.

2.2.2 Average with bias. The average of $PM_{2.5}$ values of all people is only a reasonable prediction for the true $PM_{2.5}$ value of the environment at that moment. However, most sensors can not produce these true values, each sensor has its own inaccuracy. And since we want to predict the $PM_{2.5}$ values recorded by a specific sensor, we want to take into account this inaccuracy. Since the random noise are difficult to evaluate, we only consider the bias problem - the sensor consistently records values that lower or higher than the true values by a certain amount (the bias value).

To estimate the bias, we calculate the difference between the average $PM_{2.5}$ values and the $PM_{2.5}$ values of that sensor at each moment these values available, and take the average of these differences. After that, we add this bias to the predict values of the previous run.

2.2.3 Average (outlier removed) with bias. We observe that there are some noise in certain sensor that make some recorded $PM_{2.5}$ values become very high, having very large differences with the values of other sensor at corresponding time, making the average values become more inaccurate to estimate the true values. To remove these noise, we check at a certain time, if the difference between the value of a sensor with the average value larger than the variance by a threshold factor, than we ignore this value and recalculate the average. We also recalculate the bias value and apply it for this run.

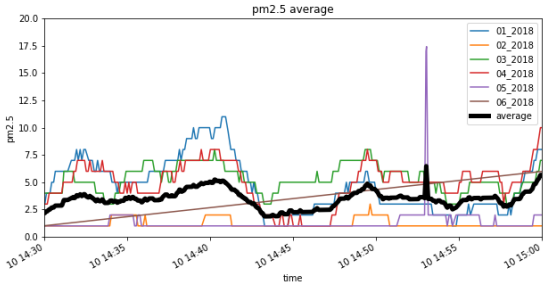

Figure 3: The average of $PM_{2.5}$ values of other people

3 EXPERIMENTS AND RESULTS

Table 1: Official evaluation result (provided by organizers)

Approach	RunID	Method	Score
1	1	MLP - Testing data	0.8141
	2	MLP - Development data	0.9013
2	3	Average	0.3384
	4	Average with bias	0.0155
	5	Average (outlier removed) with bias	0.0157

The table above shows the results of each method mentioned earlier. In this table, the scores of each run is the means of L2 distance between the predicted results and the ground truth. Our experiment results show that the second approach (predict based on other people within the group), achieve fairly good results. The result of the first approach (predict based on temperature and humidity) are not so good as the average L2 distances are still quite large. We think the reason is probably because only temperature and humidity could not give us enough information to predict the $PM_{2.5}$ values, and to have really good predictions, we should combine the information about variations of $PM_{2.5}$ values through time, the temperature and humidity values and the $PM_{2.5}$ values of other people in the same group.

4 CONCLUSION AND FUTURE WORKS

We propose two simple approaches for the Life Well Being Problem. The first approach uses a neural network to predict $PM_{2.5}$ values from other factors like temperature and humidity. The second approach using the $PM_{2.5}$ values recorded by other people at the same location and at the same time. These methods are simple but can predict the missing values quite effectively.

We think these methods could be improved further by combining them together (meaning take into account both the other attributes values and other $PM_{2.5}$ values), having a more effective noise removal method, or building a more complex regression model.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd for supporting our team with computing infrastructure.

REFERENCES

[1] Minh-Son Dao and Koji Zettsu. 2018. Complex Event Analysis of Urban Environmental Data based on Deep CNN of Spatiotemporal Raster Images. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*. 2160–2169. https://doi.org/10.1109/BigData.2018.8621916

[2] Tomohiro Sato Koji Zettsu Duc-Tien Dang-Nguyen Cathal Gurrin Ngoc-Thanh Nguyen Minh-Son Dao, Peijiang Zhao. 2019. Overview of MediaEval 2019: Insights for Wellbeing Task: Multimodal Personal Health Lifelog Data Analysis. In *MediaEval2019 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Sophia Antipolis, France.

[3] Hyeonjin Song, Kevin James Lane, Honghyok Kim, Hyomi Kim, Garam Byun, Minh Le, Yongsoo Choi, Chan Ryul Park, and Jong-Tae Lee. 2019. Association between urban greenness and depressive symptoms: Evaluation of greenness using various indicators. *International Journal of Environmental Research and Public Health* 16, 2 (2 1 2019). https://doi.org/10.3390/ijerph16020173