Predicting Missing Data by Using Multimodal Data Analytics

Loc Tai Tan Nguyen[1], Minh-Tam Nguyen[2], Dang-Hieu Nguyen[3],

[1,2,3]University of Information Technology, VietNam
locntt.12@grad.uit.edu.vn,tamnm.12@grad.uit.edu.vn,hieund.12@grad.uit.edu.vn

ABSTRACT

In this paper, we introduce a method using multimodal data analytics to predict missing data collected by sensors. Our approach is to find data at the near-by location and time by using the time-filtering algorithm and incrementally scanning radius to replace missing data. The method is evaluated by using MediaEval 2019 Insight for wellbeing – subtask 1 dataset and evaluation metric. The results show that the proposed method works well and predict missing data with high accuracy.

1 INTRODUCTION

Air pollution is proved to be a significant factor affect on human beings [2]. Thus, having the ability to predict air pollution is the target of many research activities [3]. Nevertheless, before being to predict air pollution, collecting air pollution data from sensors and data from objects that may impact or be impacted by air pollution may have more priority order [2]. Noise, outliers, and missing data usually happen when gathering data towards harming severely on the accuracy of a predicting stage. Thus, MediaEval 2019 Insight for wellbeing task challenges participants to recover missing data recorded by air pollution sensors (e.g., $PM2.5$) [1]. This paper reports our solution to tackle this challenge.

2 METHODOLOGY

The primary purpose of the proposed method is to define a hypothesis that can represent the associations among heterogeneous data and towards building a system that able to predict missing values in the provided dataset. The hypothesis points out that there is a strong association of heterogeneous data recording at the near-by location and time. Thus, we build the time filtering algorithm and radius-based increment scan policy to gather near-by data whose values can be used to predict missing data. The following (sub)sections describe in detail how to filter data and collect useful position information to predict missing data.

2.1 Data Processing

- **Circling Time**: This function is to collect all near-by-time data. We first cluster all given datasets into different groups so that each group has the same date and time (i.e., same day). Then only data happens within *start_time* and *end_time* are selected. It should be noted that *start_time* and *end_time* denote the time period when data missing.
- **Circling Position**: In order to collect all near-by-location data, we define the formula that calculates the distance of two coordinates. All data recorded within this distance are selected. The formula is defined as follows:

$$d = 2rsin^{-1}\left(\sqrt{sin^2\left(\frac{\alpha 2-\alpha 1}{2}\right)+cos(\alpha_1)cos(\alpha_2)sin^2\left(\frac{\beta_2-\beta_1}{2}\right)}\right)$$

where: d: is the distance between the two points; r: is the radius of the sphere; α_1, α_2: latitude of point 1 and latitude of point 2 (in radians); β_1,β_2: longitude of point 1 and longitude of point 2 (in radians). The radius is set from 1m to 100m.

2.2 Missing Data Prediction

After running the circling time and circling position, we obtain the $PM_{2.5}$ value of some nearest positions; we then calculate the Maximum, Minimum and Average of these values from a position that needs to predict. To optimizing the results, we incrementally increase the radius step by step from 1m to 20m at this time to scan all positions. According to our experience, we choose the ideal radius is 20m since within the 20m radius the predicted $PM_{2.5}$ values reach the highest accuracy.

If within 20m radius, we cannot get any point, we will take a single nearest point in [21m, 100m]. If there is no point in [0m, 100m], set value for $PM_{2.5}$ is zero and from thence we have build Algorithm 1.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

3 EXPERIMENTAL RESULTS

The experimental results running on the training dataset is denoted in Table 1

Table 1. The Result of runs

Question ID, File name, Start time, End time	Number of values missing	Euclidean distances (L2 distance) Min-max [0, 1]		
		Maximum	Average	Minimum
Q1,221_2019,2019-04-06 11:00:00+09,2019-04-06 11:30:00+09	30	**0.21045918**	0.21998299	0.23630952
Q2,04_2018,2018-03-25 13:00:00+09,2018-03-25 13:30:00+09	329	**0.0330545**	0.06135831	0.31453294
Q3,213_2019,2019-04-06 12:30:00+09,2019-04-06 13:00:00+09	30	0.2923021	**0.25639416**	0.26995839
Q4,10_2018,2018-03-25 13:30:00+09,2018-03-25 14:00:00+09	351	**0.04421598**	0.04847256	0.0678512
Q5,205_2019,2019-04-06 14:30:00+09,2019-04-06 15:00:00+09	30	0.06879694	**0.05607189**	0.10070782
Q6,13_2018,2018-03-25 13:30:00+09,2018-03-25 14:00:00+09	355	0.14385279	0.29152467	**0.09042843**
Q7,118_2019,2019-03-23 13:00:00+09,2019-03-23 13:30:00+09	30	0.13827757	0.09529549	**0.0597763**
Q8,21_2018,2018-03-25 14:00:00+09,2018-03-25 14:30:00+09	292	**0.04258575**	0.13251125	0.06146689
Q9,117_2019,2019-03-23 12:00:00+09,2019-03-23 12:30:00+09	30	**0.0997428**	0.12806474	0.13343664
Q10,28_2018,2018-03-25 12:30:00+09,2018-03-25 13:00:00+09	342	**0.09445415**	0.30232973	0.16027715

Table 2 shows the results when running on the testing dataset.

Table 2. The Evaluation of run

Group_id	Method	Run_id	Score
SHT_UIT	Maximum	1	0.00483679
SHT_UIT	Average	2	0.00054178
SHT_UIT	Minimum	3	0.00046321

Experimental results are evaluated based on optimized the Maximum, Minimum and Average precision. This result shows that although our proposed method is simple but it is effective. Our best run is run with Minimum. Because the Minimum value has noise very low, the value is more accurate than the other two methods (Maximum, Average). Nevertheless, there is not a big gap among submitted runs.

4 CONCLUSIONS

We report our work at the MediaEval 2019 Insight for Wellbeing task - subtask 1. We use time-filtering algorithm and radius-based increment policy to gather near-by location and time data towards predicting missing data. The results show that our solution has high accuracy.

Algorithm 1: Recovery $PM_{2.5}$'s values from near-by location and time data

```
DataA: Merge all data in a group;
DataB: In DataA, retrieve all data in the period from
  starttime to endtime of data lost PM2.5;
DataC: A list coordinates of data lost PM2.5;
for each coordinate in DataC do
    - initialization array(PM2.5) containing values of
      PM2.5;
    - initialization array(coordinate) to store
      coordinate;
    while radius less than or equal hundred do
        for each coordinate in DataB do
            set d is distance coordinate in DataC and
              DataB;
            if d less than radius and coordinate not in
              array(coordinate) then
                - add value PM2.5 of coordinate B into
                  array(PM2.5);
                - add coordinate into
                  array(coordinate);
            else
                do nothing
            if radius greater than twenty and number
              of element in array(PM2.5) greater than
              zero then
                calculator output for PM2.5;
                - get maximun value in array(PM2.5);
                - get average all values in
                  array(PM2.5);
                - get minimun value in array(PM2.5);
                break loop on DataB and then break
                  for radius loop, go to next coordinate
                  in DataC;
            else
                do nothing
        if radius equal hundred and number of element
          in array(PM2.5) equal zero then
            set output value of PM2.5 is zero;
        else
            do nothing
```

REFERENCES

[1] Minh-Son Dao, Peijiang Zhao, Tomohiro Sato, Koji Zettsu, Duc-Tien Dang-Nguyen, Cathal Gurrin, and Ngoc-Thanh Nguyen. 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.

[2] Tomohiro Sato, Minh-Son Dao, Kota Kuribayashi, and Koji Zettsu. 2019. SEPHLA: Challenges and Opportunities Within Environment-Personal Health Archives. In *MMM*. 325–337.

[3] Peijiang Zhao and Koji Zettsu. 2018. Convolution Recurrent Neural Networks for Short-Term Prediction of Atmospheric Sensing Data. In *2018 IEEE GreenCom-CPSCom-SmartData)*. IEEE, 815–821.