Air Quality Prediction System of Beijing and London

KDD CUP 2018 *

Hsiao Ting Huang
Master in NCKU
t654321ina@gmail.com

Yen Po Chien
Bachelor in NCKU
asd757817@gmail.com

Yu Cheng Chang
Bachelor in NCKU
vic85821@gmail.com

Yi Lin Wu
Bachelor in NCKU
tcfsh510alan@gmail.com

Kuang Ting Cheng
Bachelor in NCKU
timcheng74@gmail.com

Yu Hsin Chen
Master in NCKU
N26061092@mail.ncku.edu.tw

ABSTRACT

This project is to build an air quality prediction system for 2018 KDD CUP of Fresh Air. This system can predict the concentration value of several air pollutants in the next 48 hours for Beijing, China and London, UK. The architecture of our system is composed of four parts: Data crawling, Data preprocessing, Feature selection, Model construction and validation. Finally, the system shows a good performance, especially in PM2.5 and O3 concentration prediction.

Keywords

Air Quality Prediction; KDD CUP; Ensemble Model

1. INTRODUCTION

Over the past few years, the air quality in major cities such as Beijing and London has been gradually deteriorating. Among all the air pollutants, there are some that pose a threat to human health. Particulate matters(i.e., PM) are particles that are small enough to enter human body .The research[3] shows that PM with diameter smaller than 2.5 μm can easily enter lungs or blood veins and damage our respiratory and cardiovascular systems.

In this task, we seek to develop an air prediction model which can predict three indicators, PM2.5, PM10 and O3 respectively, over the coming 48 hours. To evaluate the result, we also participate in the KDD CUP, which is a competition that focus on data mining and machine learning. On each day throughout the competition, air quality data and meteorological data for both cities will be provided on the hourly basis.

*KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining, the leading professional organization of data miners.

DOI: 10.1145/1235

2. APPROACH

In this section, we will explain the architecture of our system. There are four parts : Data crawling, Data preprocessing, Feature selection, Model construction and validation. The following sections will demonstrate the details of each part.

2.1 Data Crawling

The concentration of air pollutants in the last year is provided by the KDD CUP organizer. We consider weather condition as an important factor which affects the concentration of air pollutants, so we collect the data over the past 3 years from http://beijingair.sinaapp.com/ which is an API for history air quality and weather data.

The traffic flow also has an impact on air quality. From the competition website's discussion board, we get the traffic charge form of the highway in Beijing and holiday information.

Besides, we take the elevation of each station into account, because the terrain and height influence the pollution concentration and spreading out direction. We collect the information by Google Maps Elevation API[1].

2.2 Data Preprocessing

The data need to be reorganized before we can feed in prediction models. The following are some defects :

2.2.1 Null or invalid value

The data properties are contiguous on the time domain, so it's not permitted to contain null values in any features. Therefore, we try to replace the null and invalid value with the average / svr / linear regression methods.

2.2.2 Different time format

To satisfy the competition rules, we unify the time zone to UTC ±00:00. In other side, it's convenient for the prediction system to be applied to other cities.

2.2.3 Interpolation to the terrain information

As to consider the terrain data into account, the air pollution data of every region is needed. But for a place that doesn't have accurate data, squared interpolation is used to obtain estimate values of the air pollution data. The interpolated data will than be trained with weather and terrain data in the CNN1D model described below.

2.3 Feature Selection

There are lots of features in our data-set such as concentration of several pollutants, weather condition, traffic charge, festival information and elevation of each station. It's necessary to figure out what kinds of the features should be used.

We extract a subset of relevant feature by recursive feature elimination.

2.4 Model Construction and Validation

We use some famous neural network models including ANN(Artificial Neural Network), LSTM(Long Short Term Memory), CNN1D (Convolution Neural Network), and traditional Epsilon-Support Vector Regression(SVR). Then, we implement **ensemble model** with feeding predictions of above models into an ANN model to weight each prediction and get the final result.

Following is about each model and its implementation:

2.4.1 Artificial Neural Network (ANN)

After feature selection, we decide to use weather condition, concentration of pollutants, time, highway charge and holiday as input features finally. In the side of activation function, **eLU** is easier to get convergence than other functions, so it's picked in our model.

The training data is the values of the past all year. And the method we take to predict is sequence to sequence. For example, the data from 12 a.m to 11 a.m. today are used to get the prediction in 12 a.m. tomorrow. The reason why we take the time interval is we usually couldn't get the real data from station in the real time. So we have to use past data to predict targets. This method is the steadiest of all, so we usually compare new method with it to measure performance.

2.4.2 Long Short Term Memory (LSTM)

LSTM

In order to find out the long-term trend of air quality and the short-term changes, we build a LSTM model. Through recursive feature elimination, we only select concentration of air pollutant as training feature. We use the past 72 hours data as input and the next 48 hours data as output to train LSTM model. To make the performance of model more intuitive, we use SMAPE as loss function when training this model. After 150 iteration, the SMAPE is about 0.6.

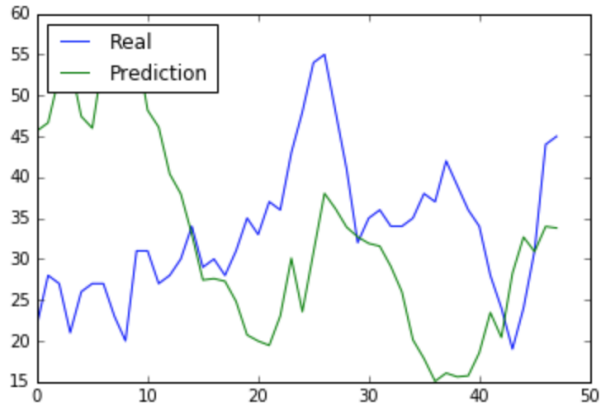
Multi-LSTM

Furthermore, since we want to take other features(not only air pollutions) into account, we attempt to use the multi-LSTM to include other features such as weather, traffic and dates.

And our few implement of multi-lstm in the KDD CUP 2018 are showing as follow:

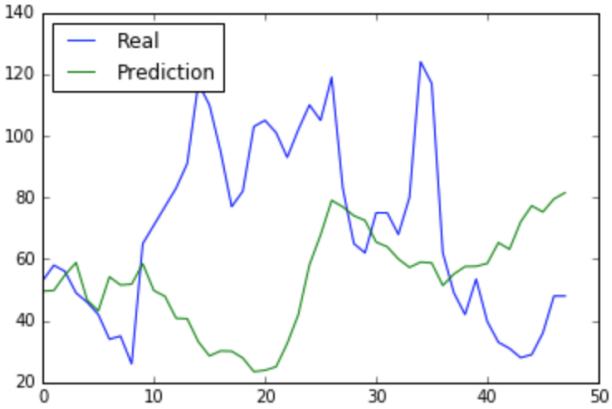
2.4.3 Convolution Neural Network 1D (CNN1D)

CNN model is good at processing coordinate data such as terrain and weather. Since regular two dimension CNN needs a lot of time and resources to train, we transform our 2D-coordinate data into a one dimensional array, each entry with multiple features. We use last 48 hours air pollution data and other features as input and the next 48 hours data as output. The result shows the SMAPE value falls around 0.5 by using SMAPE as loss function after 100 iterations.



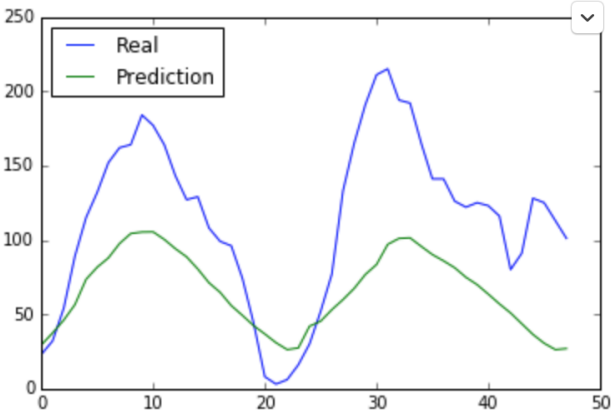

```
Predict hour:  48
SMAPE:  0.43058639714630037
```

Figure 1: PM2.5 in Bejing's site




```
Predict hour:  48
SMAPE:  0.5187926365074303
```

Figure 2: PM10 in Bejing's site




```
Predict hour:  48
SMAPE:  0.6036312056292006
```

Figure 3: O3 in Bejing's site

2.4.4 *Epsilon-Support Vector Regression (SVR)*

SVR, which stands for Support Vector Regressor, is a regressor. Regressors perform regression, predicting continuous ordered variables. As we know, SVR is very good at filling the missing value in a short period of continuous value. Therefore, we use the SVR to fill the missing value and also predict the air quality in the future 48 hours. Following is one of the 48 hours' prediction in Beijing's site. The red spots are predict values and the blue spots are actual values.

```
-1-ONE HOUR-----------------(0.031)
-1-NEAR HOUR----------------(0.611)
A:[21 24 23 22 23]
P:[21 19 58 68 61]
-2-FAR HOUR-----------------(0.491)
-2-48 HOUR------------------(0.477)
A:[48 54 48 43 48 45]
P:[68 62 78 78 75 72]
-3-ALL HOUR-----------------(0.634)
```

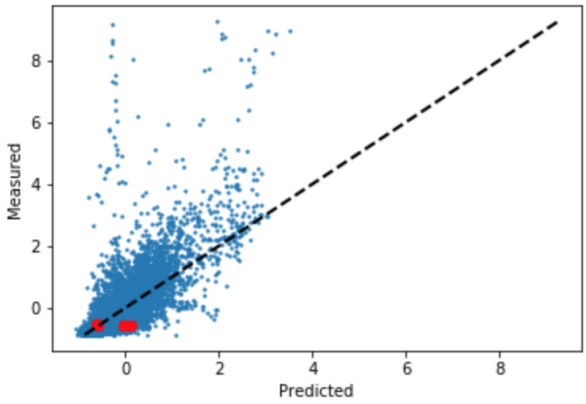


Figure 4: PM2.5 Prediction with SVR

2.4.5 *Model Validation*

In the part of validation, we select the performance of ANN model as standard. Then we try different combination of our basic 4 models and compare its performance with ANN in order to find a stronger model. Each model use data before 2017, so we can use Jan. 2018 - Apr. 2018 as validation data, and compare performance between every model and different combination of ensemble models.

3. EVALUATION

Compare our system prediction with real data, and also explain the campaign rules about evaluation.

3.1 Prediction Result

With random choice of date and station, we plot the prediction value and ground truth.

As the result shows, O3 has the best prediction result, because the concentration of O3 is severely influenced by the sunlight. The ultraviolet light will decompose oxygen to O3, so the concentration of O3 will be high in mid noon and decrease as the sun set, the pattern is relatively easy to

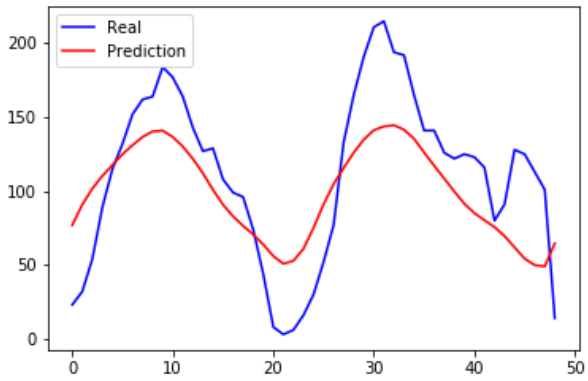


Figure 5: O3 prediction result

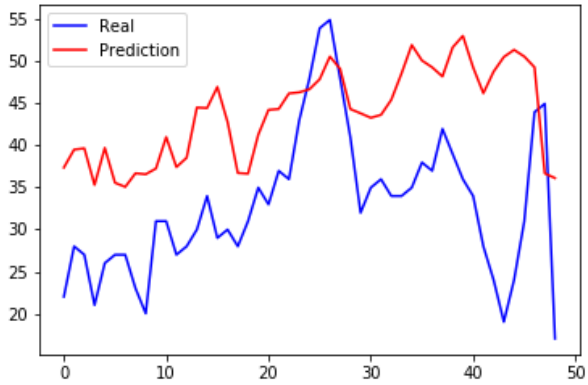


Figure 6: PM2.5 prediction result

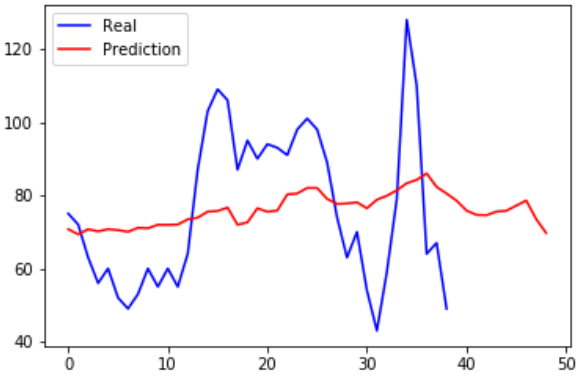


Figure 7: PM10 prediction result

predict. PM10 has the worst result due to the wide range fluctuation and the unpredictable property.

3.2 Formula

KDD CUP uses SMAPE (Symmetric Mean Absolute Percentage Error) function to evaluate the result, which is listed below.

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{(A_i + F_i)/2}$$

n = number of prediction, in this case is 48.
A_i = The prediction data.
F_i = The real air quality data.

As the formula shows, if the prediction data perfectly matches the air quality data, SMAPE would be 0. Otherwise if the prediction is null, the SAMPE would be 2. So the SMAPE indicates the model is good or bad (the lower the better).

Fig 8 show the SMAPE of the prediction from 5/1 to 5/31. The first few days due to lack of feedback and experience, our SMAPE score is relatively high. But after some modifications in various models, the SMAPE value stabilizes at around 0.45.

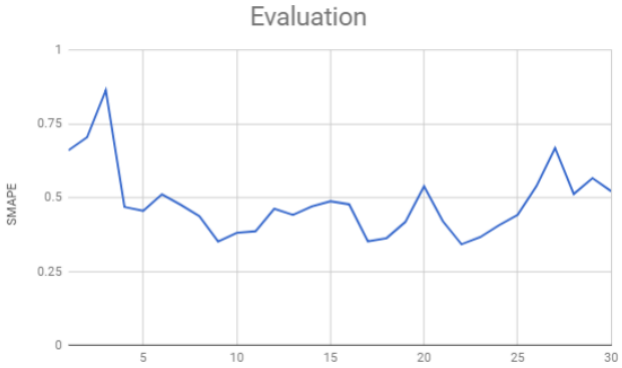


Figure 8: Performance evaluated with SMAPE

4. CONCLUSION

Finally, we get the 45'th among all participants, which is **top 6%** of the total.

Ensemble method is widely used in AI competitions. In this project, the ANN prediction is stabler than ensemble model, but sometimes it's worse in the accuracy. We find out that ensemble model is hard to detect relatively high or low value. So if the concentration is explosive high or low, the performance of this ensemble model would be bad.

The current system is specified with London and Beijing. In the future, we can expand our system to a more general use, such as predicting the air pollution globally or other kinds of air pollutants.

5. REFERENCES

[1] Google company. Google Maps Elevation API. Internet: https://developers.google.com/maps/documentation/elevation/start?hl=zh-tw

[2] Sepp Hochreiter, *LONG SHORT-TERM MEMORY*. Neural Computation 9(8):1735-1780, 1997

[3] Becker, S., Fenton, M.J., Soukup, J.M.*Involvement of microbial components and toll-like receptors 2 and 4 in cytokine responses to air pollution particles*. American journal of respiratory cell and molecular biology 27(5), 611-618 (2002)