Air Quality Prediction with different Regression Method

YU-SHUN MAO

July 01, 2020

1. Introduction

1.1 Background

Air quality prediction is important in the current issue, along with the population growth power supplies and demand is also growing, but power supplies and production process can cause environmental pollution, the pollution at the same time can produce health risk to humans [1], a study of Germany, diesel oil and industrial emissions are the main factors that cause cancer risk (70%), in central Tehran have significant cancer risk exposure to PAHs(Polycyclic Aromatic Hydrocarbons) [2].The air we breathe is full of exhaust gases and pollutants from motor vehicles and industrial emissions and coal-fired power plants, including respirable suspended particulates (PM10, PM2.5), nitrogen dioxide (NO2), ozone and sulfur dioxide. With particle diameter of less than 2.5 micrograms of PM2.5 special attention by the medical profession, because of the fine particles into the nasal cavity, in addition to cause rhinitis, also along the respiratory tract into the lungs, affect cardiovascular health, long-term increase the chance of chronic lung disease, cardiovascular disease, increase the risk of lung cancer, even is also associated with increased risk of early death. Medical science has long confirmed that PM2.5 has a significant relationship with lung cancer risk. Particulate matter usually comes from diesel vehicles and industrial exhausts, which contain PAHs. In addition, PM2.5 can stimulate the inner wall of blood vessels and clot coagulation after it invades the body, which may lead to cardiovascular embolism and heart disease. Therefore, living in an environment with high concentration of PM2.5 for a long time will increase the risk of early death due to heart disease and chronic lung disease [3].

1.2 Problem

Data that might contribute to determining air quality might include air pollutants whole year, PM10, PM2.5, Temperature, and Weather that describe what kind of air quality is. This project aims to predict concentration of PM2.5 of next hour based hourly data.

2. Data acquisition and cleaning

2.1 Data sources

PM2.5 is the main target of our prediction because it is harmful and widely discussed. We download air quality and meteorological dataset from 2016.01.01 to 2016.12.30 in Environmental Protection Administration Executive Yuan Taiwan environmental resources database.

2.2 Data cleaning

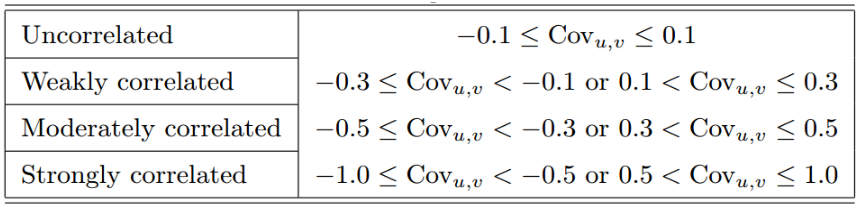
Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values or error values from sensor fault, because of lack of record keeping. I decided to replace the error values or missing values with NaN(Not a Number), and then use linear imputation to replace NaN values with substituted values, because of this dataset have fewer missing values.

2.3 Feature selection

After data cleaning, there were 8760 samples and 14 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features.

I inspected the correlation of independent variables, and found several pairs that were Strongly correlated as shown in Table 1. For example, PM2.5, and PM10 were highly correlated as shown in Table 2. This makes sense, after all, particle meters from the same source. From these strongly correlated features, only feature “PM10” was kept, others were dropped from the dataset.

Table 1. Correlation relationship between two variables u and v

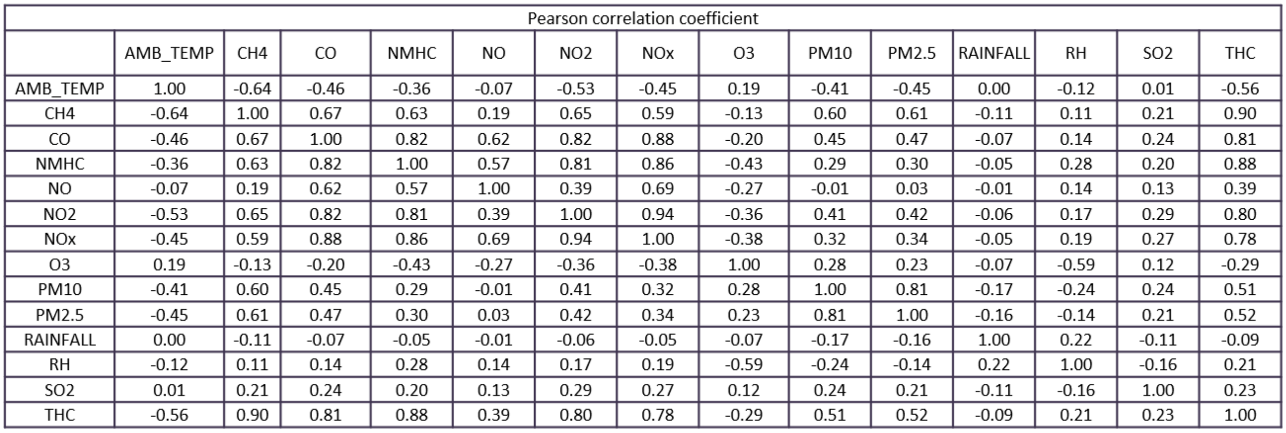


Uncorrelated	$-0.1 \leq \mathrm{Cov}_{u,v} \leq 0.1$
Weakly correlated	$-0.3 \leq \mathrm{Cov}_{u,v} < -0.1$ or $0.1 < \mathrm{Cov}_{u,v} \leq 0.3$
Moderately correlated	$-0.5 \leq \mathrm{Cov}_{u,v} < -0.3$ or $0.3 < \mathrm{Cov}_{u,v} \leq 0.5$
Strongly correlated	$-1.0 \leq \mathrm{Cov}_{u,v} < -0.5$ or $0.5 < \mathrm{Cov}_{u,v} \leq 1.0$

Table 2. Simple feature selection during data cleaning.

O3: humidity, NOX, NO2.
SO2: None.
CO: NOX, NO, NMHC, NO2, CH4, THC.
NO2: NOX, O3, CO, NMHC, THC.
PM2.5: PM10.
PM10: PM2.5.

Table 3. Pearson correlation coefficient between variables.



Pearson correlation coefficient														
	AMB_TEMP	CH4	CO	NMHC	NO	NO2	NOx	O3	PM10	PM2.5	RAINFALL	RH	SO2	THC
AMB_TEMP	1.00	-0.64	-0.46	-0.36	-0.07	-0.53	-0.45	0.19	-0.41	-0.45	0.00	-0.12	0.01	-0.56
CH4	-0.64	1.00	0.67	0.63	0.19	0.65	0.59	-0.13	0.60	0.61	-0.11	0.11	0.21	0.90
CO	-0.46	0.67	1.00	0.82	0.62	0.82	0.88	-0.20	0.45	0.47	-0.07	0.14	0.24	0.81
NMHC	-0.36	0.63	0.82	1.00	0.57	0.81	0.86	-0.43	0.29	0.30	-0.05	0.28	0.20	0.88
NO	-0.07	0.19	0.62	0.57	1.00	0.39	0.69	-0.27	-0.01	0.03	-0.01	0.14	0.13	0.39
NO2	-0.53	0.65	0.82	0.81	0.39	1.00	0.94	-0.36	0.41	0.42	-0.06	0.17	0.29	0.80
NOx	-0.45	0.59	0.88	0.86	0.69	0.94	1.00	-0.38	0.32	0.34	-0.05	0.19	0.27	0.78
O3	0.19	-0.13	-0.20	-0.43	-0.27	-0.36	-0.38	1.00	0.28	0.23	-0.07	-0.59	0.12	-0.29
PM10	-0.41	0.60	0.45	0.29	-0.01	0.41	0.32	0.28	1.00	0.81	-0.17	-0.24	0.24	0.51
PM2.5	-0.45	0.61	0.47	0.30	0.03	0.42	0.34	0.23	0.81	1.00	-0.16	-0.14	0.21	0.52
RAINFALL	0.00	-0.11	-0.07	-0.05	-0.01	-0.06	-0.05	-0.07	-0.17	-0.16	1.00	0.22	-0.11	-0.09
RH	-0.12	0.11	0.14	0.28	0.14	0.17	0.19	-0.59	-0.24	-0.14	0.22	1.00	-0.16	0.21
SO2	0.01	0.21	0.24	0.20	0.13	0.29	0.27	0.12	0.24	0.21	-0.11	-0.16	1.00	0.23
THC	-0.56	0.90	0.81	0.88	0.39	0.80	0.78	-0.29	0.51	0.52	-0.09	0.21	0.23	1.00

3. Predictive Modeling

In this study, I applied two types of time series prediction(one-step ahead prediction and multi-step ahead prediction). The multi-step ahead prediction is performed in k(step) iterations. The predicted outputs of previous iterations are fed-back as part of inputs in the current iteration, except first iteration until k iterations prediction is done.

3.1 Regression models

3.1.1 Applying standard algorithms and their problems

I applied machine learning model (Kernel Ridge regression) and deep learning models(Convolutional Neural Network, Long Short Term Memory) to the dataset, using RMSE(Root Mean Squared Error) and MAE(Mean Absolute Error) as the tuning and evaluation metric. The results all had the same problems. The prediction errors were larger as the much larger number of time step in multi-step ahead prediction.

3.1.2 Performances of different models

I built linear regression, CNN, LSTM, MLP, and KRR using root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric. LSTM had the best performance among all models, which had lowest RMSE value and MAE value between other models as shown in Figure 1 and Figure 2.

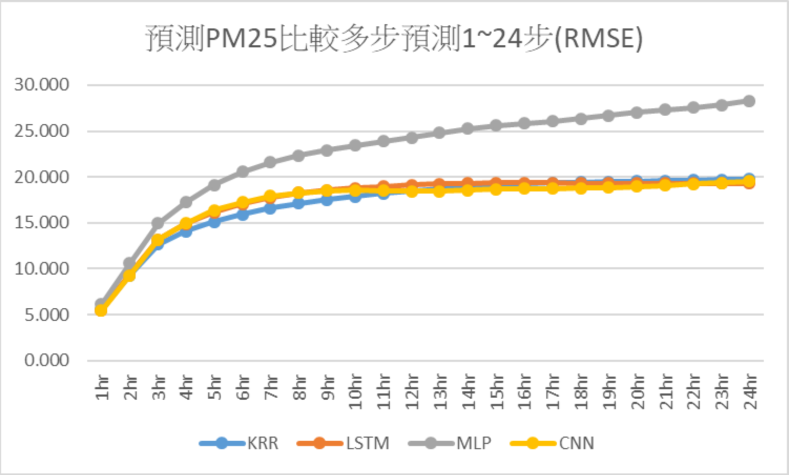


Figure 1 Applied different models for multi-step ahead prediction

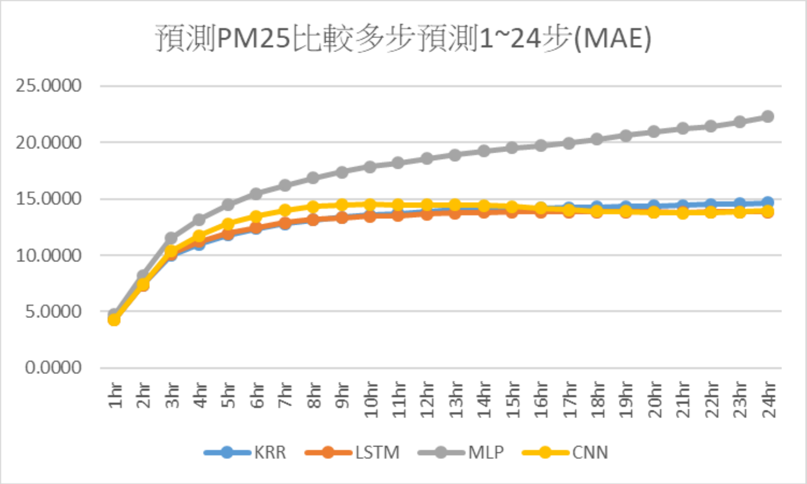


Figure 2 Applied different models for multi-step ahead prediction

4. Conclusions

In this study, I analyzed the correlation relationship between air pollutants data and meteorological data. I identified “PM2.5” and “PM10” the most correlated features that affect air quality next hour. I built regression models to predict how much concentration of “PM2.5” would change next hour or next day. These models can be very useful in helping official in a number of ways. For example, it could help estimate the time should be inspecting base on air quality prediction.

5. Future directions

In time series analysis, there are many more advanced ways to extract the characteristics of time series, such as: seasonality (seasonality), trend (trend), residual (residual), etc., through the frequency setting, you can get Weekly (7 days), monthly (30 days), annual (365 days) trends or seasonality; in the field of environmental engineering, there are many factors that affect air quality, such as climate factors such as wind speed, humidity, wind direction, and tidal changes. If these factors can be considered in air quality prediction, it is believed that air product prediction will have good results.

In the field of time series prediction, the recursive method used in multi-step prediction usually has the problem of cumulative error, which is a question worth discussing how to improve. Prediction models built by traditional machine learning methods often have upper limits of complexity, resulting in poor results in multi-step prediction. In the future, it is hoped that the complexity of the model will be increased through deep learning to help improve the effect of multi-step prediction.