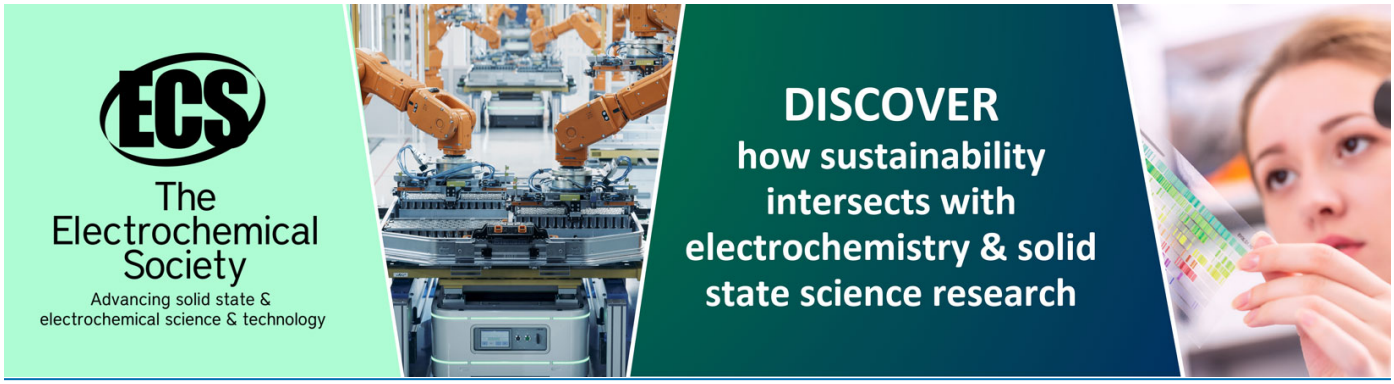
Application of XGBoost algorithm in hourly PM2.5 concentration prediction



Application of XGBoost algorithm in hourly PM2.5 concentration prediction

Bingyue Pan

College of Management, Shanghai University, Shanghai 200444, PR China
htsilu@126.com

Abstract. In view of prediction techniques of hourly PM2.5 concentration in China, this paper applied the XGBoost(Extreme Gradient Boosting) algorithm to predict hourly PM2.5 concentration. The monitoring data of air quality in Tianjin city was analyzed by using XGBoost algorithm. The prediction performance of the XGBoost method is evaluated by comparing observed and predicted PM2.5 concentration using three measures of forecast accuracy. The XGBoost method is also compared with the random forest algorithm, multiple linear regression, decision tree regression and support vector machines for regression models using computational results. The results demonstrate that the XGBoost algorithm outperforms other data mining methods.

1. Introduction

At present, China has entered the post-industrial era. However, since the management of resources at the beginning of development has not been well planned, the contradiction between China's development and environmental resources has become increasingly obvious. Air pollution problems have become the focus of attention in recent years. Since 2013, most cities in China have appeared a wide range of serious air pollution incidents. Air pollution has become one of China's major urban problems.

In general, the Air Quality Index (AQI) is used to characterize the air quality level, and its value depends on the concentration of six pollutants in the atmosphere, including PM2.5, PM10, SO_2, NO_2, CO, O_3. In these air pollutants, aerosol particles especially aerodynamic diameter of less than 2.5 μm aerosol particles (PM2.5) is extremely harmful to human health. Therefore, the monitoring and forecasting work of PM2.5 is of great scientific and practical significance.

PM2.5 concentration prediction models can be classified into two categories: chemical models and data driven models. Chemical models are based on the chemical laws to model all of the relevant chemical processes that contribute to PM2.5 formation. Data driven models use historical data to make future predictions. They are based particularly on statistical approaches.

However, the chemical transformation mechanism of air pollutants is very complex, and the current detailed list of emissions is very difficult to obtain. Therefore, there are some limitations on the prediction of PM2.5 concentration by using the chemical model. In China, various cities have established a number of air quality monitoring stations to obtain hourly urban air pollutant concentration data.

A number of data driven prediction models were considered for the PM2.5 concentration prediction, among which most popular and widely used models are artificial neural networks (ANNs), random forest (RF), multiple linear regression (MLR) and support vector machines for regression (SVMreg) methods. Niska et al. used genetic algorithm and neural network method to predict the NO_2

concentration after 24h and achieved high accuracy [1]. Kurt et al. used neural networks to predict the concentrations of SO_2, PM10 and CO within three days, and researched the influence of the prediction date on the accuracy of the model as a factor [2]. In addition to the neural network, Yeganeh et al. predicted the daily CO concentration by combining the least squares method with the support vector machine method [3].

Main contributions of this paper are: (i) the application of the XGBoost algorithm for hourly PM2.5 concentration prediction in Tianjin, China; (ii) the comparison of the XGBoost model with other well-known data mining models.

2. Extreme Gradient Boosting Algorithm

In this section, XGBoost algorithm, a scalable machine learning system for tree boosting would be discussed.

Boosting is a machine learning technique that can be used for regression and classification problems. It generates a weak learner at each step and accumulates it into the total model. If the weak learner for each step is based on the gradient direction of the loss function, it can be called Gradient Boosting Machines (GBM) [4].

The main difference between Random Forest (RF) and Gradient Boosted Machines is that while in RF, trees are built independently to, each other, GBM adds a new tree to complement already built ones [5].

2.1. Objective function

Assume that a dataset is $\mathcal{D} = \{(x_i, y_i): i = 1 \cdots n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$, then we get n observations with m features each and with a corresponding variable y. Let $\hat{y}_i$ be defined as a result given by an ensemble represented by the generalised model as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

In the above formula, f_k is a regression tree, and $f_k(x_i)$ represents the score given by the k-th tree to the i-th observation in data. In order to functions f_k, the following regularised objective function should be minimized:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{2}$$

l is the loss function. In order to prevent too large complexity of the model, the penalty term Ω is included as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2 \ , \tag{3}$$

where γ and λ are parameters controlling penalty for the number of leaves T and magnitude of leaf weights w respectively. The purpose of $\Omega(f_k)$ is to prevent over-fitting and to simplify models produced by this algorithm.

2.2. Iteractive method

An iteractive method is used to minimise the objective function. The objective function that minimized in j-th iteractive we want to add f_j is:

$$\mathcal{L}^j = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(j-1)} + f_j(x_i)) + \Omega(f_j) \tag{4}$$

This function can be simplified by using the Taylor expansion. And a formula can be derived for loss reduction after the tree split from given node:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma, \tag{5}$$

where I is a subset of the available observations in the current node and I_L, I_R are subsets of the available observations in the left and right nodes after the split. The functions g_i and h_i are defined as follows:

$$g_i = \partial_{\hat{y}_i^{(j-1)}} l(y_i, \hat{y}_i^{(j-1)}) \tag{6}$$

$$h_i = \partial^2_{\hat{y}_i^{(j-1)}} l(y_i, \hat{y}_i^{(j-1)}) \tag{7}$$

The best split at any given node can be found from the formula $\mathcal{L}_{split}$. This formula depends only on the loss function and the regularisation parameter γ. It is easy to see that this algorithm can optimise any loss function that could provide the first and second-order gradients.

There are three reasons why XGBoost performs better than other tree boosting methods. They are: (i) the introduction of the regularised loss function; (ii) the weights of each new tree can be scaled down by a given constant η, which reduces an influence of a single tree on the final score. (iii) column-sampling which works in a similar way as random forests.

This algorithm is implemented in the 'xgboost' package for the 'Python' language provided by the creators of the algorithm.

3. Data

Historical PM2.5 concentration data has been taken from nineteen air-monitoring stations around Tianjin for the period December 1, 2016- December 30, 2016. The locations of the nineteen air-monitoring stations are Beichen Science and Technology Park, Binshui West road, Dali Road, Dazhigu Eight road, Fourth Street, Hanbei Road, Hangtian Road, Hexi Road, Jingu road, Nanjin Road, NanKou Road, South Road, Qianjin Road, Qinjian Road, Municipal monitoring center, Tianshan Road, Tuanpo Road, road, Yongming Road, Yuejin Road and Zhongshan North Road.

The set of variables used in this study are: (1) hourly PM10 concentration value; (2) hourly SO_2 concentration value; (3) hourly NO_2 concentration value; (4) hourly CO concentration value; (5) hourly O_3 concentration value.

The datasets contains 6845 samples after removing the samples which have missing values. Then after splitting the datasets, we get the train datasets for the period December 1, 2016- December 23, 2016 contains 5185 samples and the test datasets for the period December 24, 2016- December 30, 2016 contains 1663 samples. The statistics of various pollutants are given in Table 1.

Table 1. Statistics of PM2.5, PM10, CO, NO_2, SO_3, O_3.

	PM2.5	PM10	CO	NO2	SO2	O3
Maximum(μg/m³)	529	713	17.5	256	217	142
Minimum(μg/m³)	1	3	0.1	6	1	1
Average(μg/m³)	139.9	175.2	2.8	86.5	35.1	16

4. Implementation

In this section we discuss implementation of the XGBoost method as well as four other prediction methods used for comparison: the Random Forest (RF), Multiple Linear Regression (MLR), Decision Tree Regression (DTR) and Support Vector Machines for regression (SVMreg).

4.1. Statistical evaluation of model performance

Prediction performance of all models was evaluated by comparing predicted and observed PM2.5 concentration value using three measures of predict accuracy calculated from the test datasets: the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the Coefficient of Determination (also called R^2).

Assume that $y_1, \cdots, y_m, m \geq 1$ are observed values for the parameter y and $\hat{y}_1, \cdots, \hat{y}_m$ are their predict values. And $\bar{y}$ is the mean of observed values.

1. The RMSE is defined as

$$RMSE = \left(\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2\right)^{\frac{1}{2}} ; \quad (8)$$

2. The MAE is as

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| ; \quad (9)$$

3. The R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \; ; \tag{10}$$

4.2. Influence of pollutant variables on PM2.5 value

Because PM2.5 and other air pollutants may exist in the process of mutual transformation, it is necessary to analyze the correlation between PM2.5 and other air pollutants in advance. Taking all training data as samples, the influence of air pollutant factor on PM2.5 concentration was discussed. Figure 1 shows the correlations between PM2.5 and other air pollutants (PM10, CO, NO_2, O_3, SO_2).

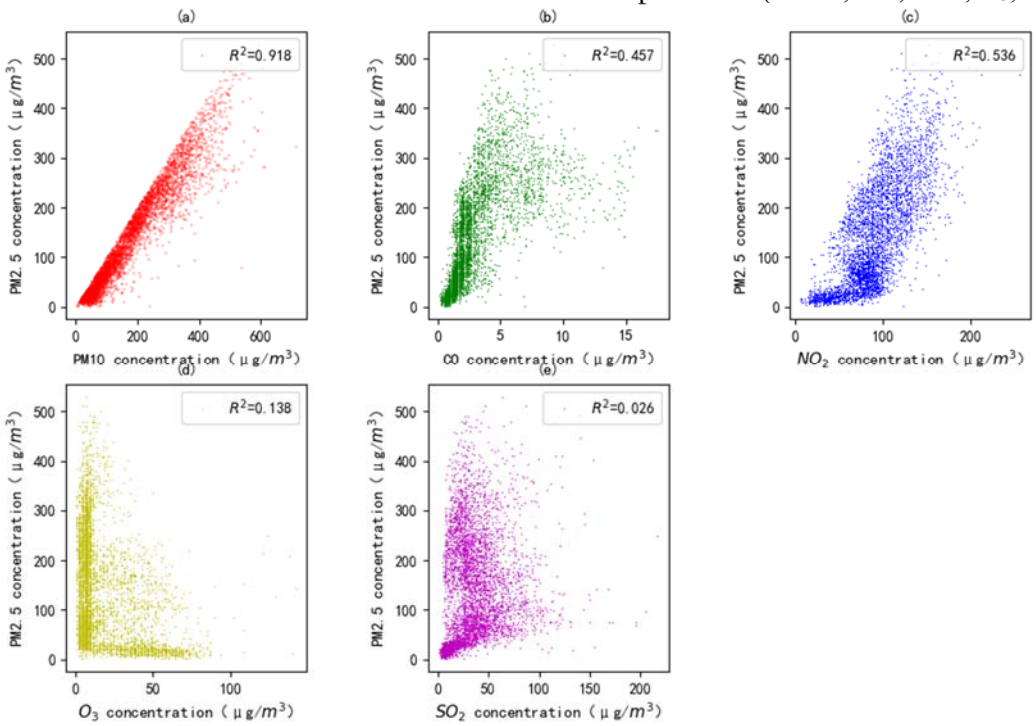


Figure 1. Correlations between PM2.5 and other air pollutants (PM10, CO, NO_2, O_3, SO_2).

From figure 1, it can be found that the concentration of PM2.5 has a certain degree of correlation with PM10, CO, NO_2 and O_3 (R^2=0.918, 0.457, 0.138, 0.536). The correlation between PM2.5 value and PM10 value was the highest. That is because there is a physical and chemical conversion process between PM2.5 and other pollutants, especially PM2.5 and PM10 would have a very high possibility of mutual conversion. In addition, part (e) shows that the correlation between PM2.5 and SO_2 is very low (R^2=0.026). Therefore, SO_2 concentration did not contribute to the prediction of PM 2.5. Through the above analysis, we decided to use the values of PM10, CO, NO_2 and O_3 as input features of the model.

4.3. Results of XGBoost and other regression model

In this section, statistical package scikit-learn in python was used to implement the XGBoost, Random Forest, Multiple Linear Regression, Decision Tree Regression and Support Vector Machines for regression. The max depths for XGBoost and DTR, respectively, are six and nine obtained by cross validation. We use the SVMreg model with the Gaussian kernel and add lasso regularization in MLR to avoid over-fitting.

Figures 2-6 display the predictions of all five models for all hours within the tests in the location Fourth Street and Table 2 summarizes the prediction performance of five models using test datasets.

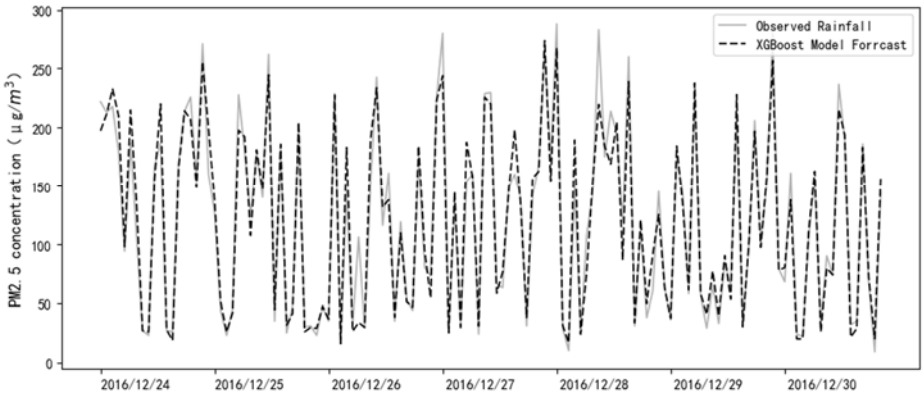


Figure 2. Prediction of PM2.5 average concentration in Fourth Street by XGBoost

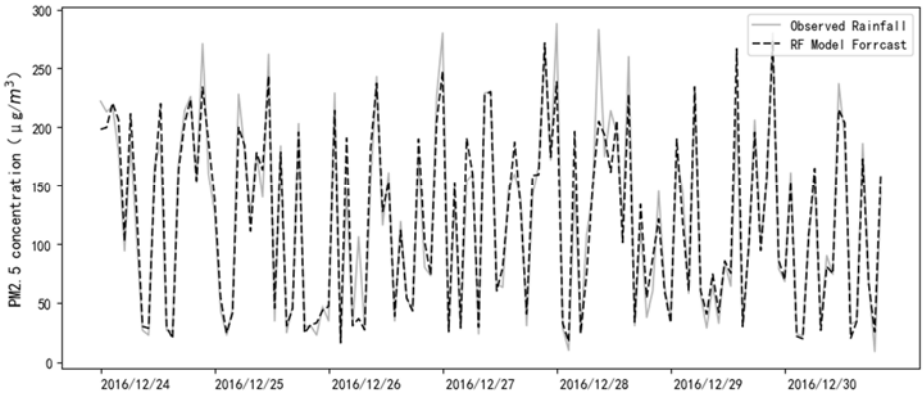


Figure 3. Prediction of PM2.5 average concentration in Fourth Street by RF

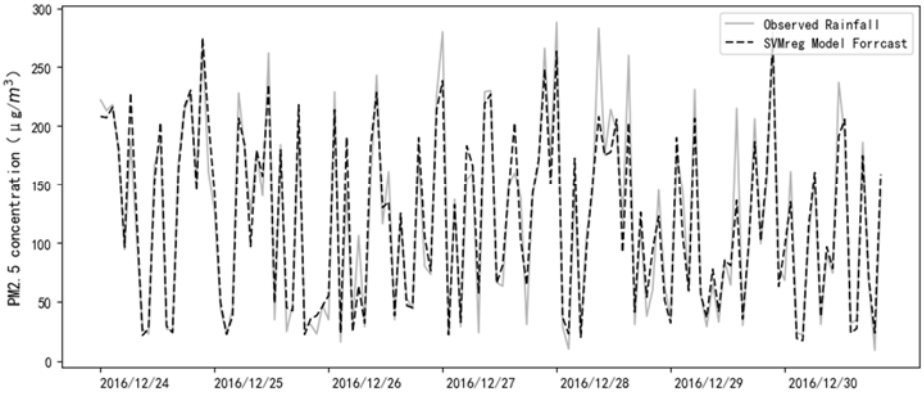


Figure 4. Prediction of PM2.5 average concentration in Fourth Street by SVMreg

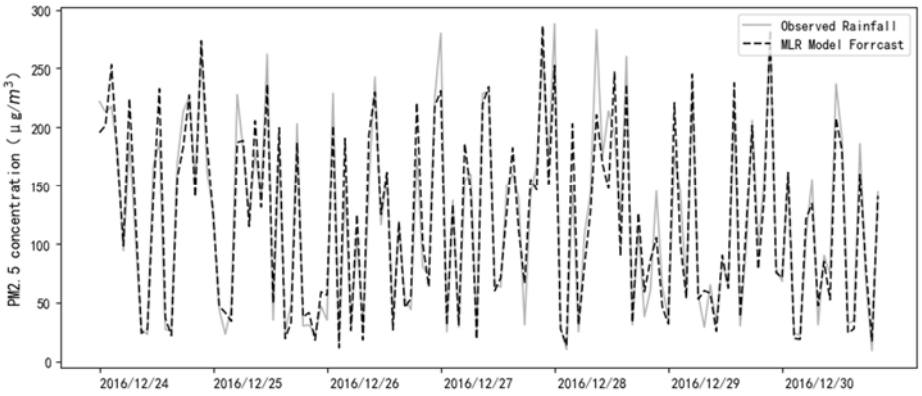


Figure 5. Prediction of PM2.5 average concentration in Fourth Street by MLR

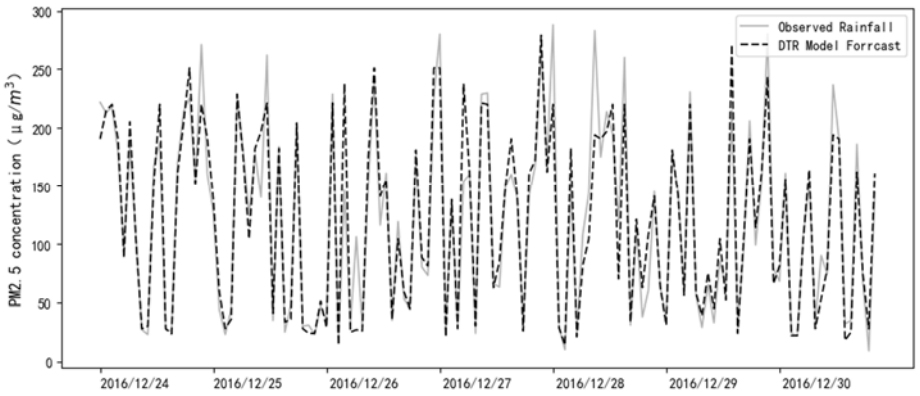


Figure 6. Prediction of PM2.5 average concentration in Fourth Street by DTR

Table 2. Comparison of five data mining models

	RMSE	MAE	R^2
XGBoost	17.298	11.774	0.9520
Random Forest (RF)	18.911	12.973	0.9426
Support Vector Machine for regressiion (SVMreg)	20.447	14.503	0.9329
Multiple Linear Regression (MLR)	21.854	16.489	0.9234
Decision Tree Regression (DTR)	25.921	15.917	0.8922

It can be seen from the Table 2 that the RMSE, MAE and R^2 values of XGBoost (RMSE=17.298, MAE=11.774, R^2=0.9520) are all better than the other four models. If we use R^2 as the evaluation criterion, the XGBoost model has the best performance, followed by Random Forest, Support Vector Regression, Multiple Linear Regression and Decision Tree Regression. However, the MAE value of Decision Tree Regression is smaller than that of Multiple Linear Regression, although its RMSE value is larger than MLR model. It is well-known that the MAE is less sensitive to outliers than the RMSE [6]. That may explains why the trends of MAE and RMSE are different in some models.

By comparing the experimental results of the XGBoost model with other four frequently used data mining models, the XGBoost method is more efficient for hourly PM2.5 concentration prediction so it is a good alternative to existing models for PM2.5 concentration prediction.

5. Conclusions and discussion

PM2.5 is the primary pollutant in the air today, and its concentration is also an important criterion for determining the air quality. PM2.5 not only significantly reduces atmospheric visibility and sunlight, but also results in more haze days and causes serious harm to human health. Therefore, the accurate prediction of PM2.5 concentration can help the government to understand the current air quality status and trend, so that the government can formulate effective prevention or control measures. The experimental results demonstrate that the PM2.5 hourly average concentration prediction model based on XGBoost method has the characteristics of high accuracy and low over-fitting probability, and outperforms other data mining methods.

However, in the contrast experiment of five models, the sample only takes the data from the Fourth Street monitoring station. So the results may be limited by the particularity of sample data from the same station. The future research can be extended to the following two aspects. On one hand, we could explore the impact of different geographical location data on predicting effect through the comparison of prediction performance between multi sites. On the other hand, we could research more comprehensive model input characteristics, such as synchronous meteorological data, other pollutant emission data and so on. The effective integration and utilization of urban multi-source data can improve the effectiveness of PM2.5 prediction.

References

[1] Niska H, Hiltunen T, Karppinen A, Ruuskanen J and Kolehmainen M 2004 *Engineering Application of Artificial Intelligence* **17** 159-167
[2] Kurt A, Gulbagci B, Karaca F and Alagha O 2008 *Environment International* **34** 592-598
[3] Yeganeh B, Motlagh M, Shafie P, Rashidi Y and Kamalan H 2012 *Atmospheric Environment* **55** 357-365
[4] Jerome H and Friedman 2002 *Computational Statistics&Data Analysis* **4** 367-378
[5] Marcin L, Bartosz T and Magdalena M 2017 *Lecture Notes in Computer Science* **10244** 661-671
[6] Hyndman R and Koehler A 2006 *Int. J. Forecast* **22** 679–688