Multi-Source-Data-Oriented Ensemble Learning Based PM 2.5 Concentration Prediction in Shenyang

Tianfang Qi
School of Information,
Renmin University of China,
Beijing, 100872, China
qitianfang@ruc.edu.cn

Hongxun Jiang
School of Information,
Renmin University of China,
Beijing, 100872, China
jianghx@ruc.edu.cn

Xiaowen Shi
School of Information,
Renmin University of China,
Beijing, 100872, China
sxw0845@ruc.edu.cn

Abstract

Shenyang where is surrounded by smokestack industries and depends on coal heating in winter, is a classical one of cities in China northeastern which has suffered from serious air pollution, especially PM2.5. The existing research on machine learning, based on historical air-monitoring data and meteorological data, does neither forecast accurately nor identify key pollutants for PM2.5. This paper presents a multi-source-data-oriented ensemble learning for predicting PM2.5 concentration. The proposed framework incorporates not only air quality data and weather data, but also industrial emission data, especially those of winter heating enterprises, in Shenyang and nearby cities; the model also takes into account location and emission frequency of pollution sources. All these data are entered into an ensemble learning model based on Extreme Gradient Boosting (XGBoost) in order to predict PM2.5 concentration, which not only improves prediction accuracy effectively, but also provides contribution analysis of different pollutants. Experimental results show that the top two factors affecting PM2.5 concentration are: (1) air pollutant emission quantities and (2) distance from pollution sources to air-monitoring stations. According to the importance of these two factors, we refine feature selection and re-train the ensemble learning model and find that the new model performs better on 72% of evaluation indexes.

1. Introduction

With the rapid development of industrialization and the continuous urbanization in China, air pollution problems have become increasingly serious. In recent years, issues of living environment and air quality have attracted a national attention. After the State Council issued action plans for controlling air pollution in 2013, active measures were taken in many places and the overall situation has become better; however, the current situation of Shenyang, a typical of cities in northeast China, still needs constant attention. Because Shenyang is located in the region of China's traditional heavy industry base, industrial emissions have led to high levels of air pollutants, such as sulfur dioxide (SO2) and nitrogen dioxide (NO2) throughout the years, especially the large number of coal-burning heating modes after entering the winter heating period, which has led to high levels of inhalable particulate matter (PM10) and fine particulate matter (PM2.5).

The goal of improving air quality cannot be achieved without effective scientific support and systematic management decisions. Accurate prediction of the influencing factors and changing trends of air quality is the foundation. Most of the traditional prediction

URI: https://hdl.handle.net/10125/59569
ISBN: 978-0-9981331-2-6

models based on machine learning, only take air-monitoring data and meteorological data as input. These models cannot predict air pollution effectively because they are unable to make the accurate predictions and identify the formation factors. This paper proposes a multi-source-data-oriented ensemble learning model for predicting PM2.5 concentration, using air quality data and meteorological data, and in particular, distinguish the structural differences between heating and non-heating periods. The model introduces the air-monitoring data and the location of pollutant discharge and the spatial orientation of air-monitoring stations to predict PM2.5 concentration. The ensemble model, which can analyze the key features of air pollution, provides measures for the forecast and warning of air quality.

The remainder of this paper is organized as follows. Section 2 briefly describes the research work of air quality prediction, ensemble learning algorithm and feature factor selection. Section 3 introduces the algorithm of the selected machine learning. Section 4 introduces the experiment setup as well as data collection and processing. Section 5 shows the prediction results of the model, explains the comparison of different models, and performs a combination optimization of different features. Section 6 makes a summary.

2. Literature Review

2.1. Air Quality Prediction Models

According to model methods, air quality prediction models can be classified into three main categories: deterministic models, statistical models and hybrid models.

Deterministic models can be carried out without a large amount of historical data, but it requires a full understanding of the source of pollutants, the real-time emission amount and a clear description of the main chemical reactions in the process of pollutant movement, according to the research [1-3]. Deterministic models can predict the concentration of spatial resolution in places where there are no air-monitoring stations, but in some cases, they have high computational costs and require a lot of computational time to complete the prediction process.

Compared with deterministic models, statistical models are much easier and more efficient. Therefore, many researchers integrated deterministic methods with statistical methods to improve prediction accuracy. Some scholars designed an adaptive neuro-fuzzy model [4]. According to the data of 12-hour average air pollutants in the Yangtze river Delta Region of China, a deterministic model based on particulate matter was created. Statistical models mainly depend on historical data and trend analysis to predict the future uncertainty; they have become the basis of many areas of forecasting decision-making. At the same time, with the development of artificial intelligence (AI) technology, prediction models are gradually transitioning from traditional statistical methods to AI-based methods [5-8].

To improve prediction accuracy, hybrid models which combine advantages of different single models are widely used in the field air quality prediction. Some scholars proposed a new hybrid model for Air Quality Index (AQI) forecasting which combined a two-phase decomposition method and an extreme learning machine (ELM) optimized by differential evolution (DE) algorithm [9]; the results showed that the hybrid model based on the two-phase decomposition method had high prediction accuracy. Some scholars proposed a hybrid model based on principal component analysis (PCA) and least squares support vector machine (LSSVM), and parameters in LSSVM were optimized by cuckoo search (CS) and its generalization ability was improved [10].

Although the above models have the high accuracy, they are unable to explain the prediction results. We still have no idea about the function and degree of influence of input features, which is not conducive to us proposing relevant solutions to solve the environmental problems

based on the predicted results, and is of little practical significance. While the ensemble learning method can guarantee the accuracy of prediction, it can also output the importance of the features of the prediction model, which is conducive to our subsequent analysis and optimization. Therefore, this paper uses the ensemble learning algorithm for research.

2.2. Ensemble Learning Model

Ensemble learning has high accuracy in machine learning algorithms and is widely used for prediction. There are two major types of ensemble learning: one is based on Boosting and the other is based on Bootstrap Aggregating (Bagging). The representatives of the former algorithm are Adaboost, GBDT, XGBoost, and the representative of the latter algorithm is Random Forests (RF).

Boosting is a kind of effective integrated learning algorithm; by using Boosting, weak classifiers can be transformed to strong classifiers. Due to its efficiency and accuracy of classification, Boosting was used in face recognition [11]. When it comes to Boosting algorithms, XGBoost algorithm has a flexible and portable gradient-distributed decision-making promotion library. When dealing with large amounts of data, XGBoost can ensure high classification accuracy and low time complexity. XGBoost is used for commercial sales forecast [12], online public opinion forecast [13], e-commerce commodity recommendation [14].

From the above literatures, ensemble learning has a good prediction effect, and can output the weight of features. For this reason, this paper applies the ensemble learning algorithm to the research of urban PM2.5 concentration prediction.

2.3. Feature Engineering

The air quality prediction model mainly includes concentration prediction of PM10, PM2.5, SO2, NO2, CO and O3. When selecting relevant pollutants data, most literatures choose SO2, CO, NOx and O3 as the main influencing factors. When selecting meteorological features, temperature, humidity, wind direction, wind speed, air pressure and dew point temperature are selected as the main influencing factors. The research shows that the multi-angle consideration of the model input variables to a certain degree can improve the model's predictive performance.

There is few research taking the pollution emissions data as input. Mao et al. took Chengdu as an example [15]. According to the emission features of different pollution sources, she confirmed the identification factors for high-resolution spatial and temporal allocation as well as the estimation method for establishing the weight of spatial and temporal allocation. This method can accurately reflect the spatial and temporal distribution features of various pollutant emissions; however, to a certain extent, it relies on the accuracy and specificity of the selected features and emission inventory, and it has high requirements of data.

Therefore, this paper aims to use industrial emissions data and air-monitoring data, combined with geographical, temporal and meteorological features, through rigorous data processing and experimental dimension settings with the ensemble learning method, to predict PM2.5 concentration in Shenyang, analyze the prediction results and influencing features.

3. Ensemble learning model for PM2.5 concentration prediction

Ensemble learning is a kind of robust and anti-interference model by combining the same algorithm or different algorithm. Each algorithm is a base learner, and one of the most widely used base learners is Classification and Regression Trees (CART). For a given training data $D = \{(x_1, y_1), (x_2, y_2) \ldots\ldots (x_m, y_m)\}$, it adopts a binary recursive partitioning method to deal with the binary classification problem by constructing a binary tree.

This paper focuses on the ensemble learning algorithm in the use of air quality forecast aspect, with the mainstream of Random Forests (RF), the Gradient Boosting Decision Tree (GBDT) and Extreme Gradient Boosting (XGBoost). Meanwhile, in order to compare the different performance between the ensemble algorithm and other algorithms, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) were selected. This paper will focus on RF, GBDT and XGBoost algorithms.

3.1. RF model

Bootstrap Aggregating (Bagging) is a kind of ensemble learning method with excellent and robust performance in multi-classification problems. As the representative algorithm of Bagging, RF algorithm firstly randomly selects k sample subsets from the training data $D = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2) \ldots\ldots (\mathrm{x}_m, y_m)\}$ with repeated random sampling, and correspondingly established multiple independent classification trees, which are denoted as $\{h(X, \theta_i), i = 1,2, \ldots\ldots k\}$, where θ_i is the parameter of the i[th] tree. Each tree records separately, and the final category is determined by the individual tree's classification results.

3.2. GBDT model and XGBoost model

Boosting is a kind of effective ensemble learning algorithm. By using Boosting, weak classifiers can be transformed to strong classifiers in order to get accurate classification results.

The decision tree in the GBDT algorithm belongs to the regression tree. Each iteration is to reduce the residual of the previous model and trains a new model on the gradient direction of the residual reduction.

Given the training data $T = \{(x_1, y_1), (x_2, y_2) \ldots\ldots (x_n, y_n)\}$, the fitting target is to find an estimation function $F^*(x)$, so as to make the difference between the predicted PM2.5 concentration and the real value close to 0.

As another kind of Boosting algorithm, XGBoost algorithm is based on the gradient promotion framework of a highly extensible tree structure model, and the ability of processing sparse data is outstanding. XGBoost is suitable for multi-source data for PM2.5 forecast. XGBoost can sort through the features of block processing and use multi-threading technology to ascend tree construction speed, which reduces computing time by a large proportion and breaks through the computational limitation of Boosting.

For training data $D = \{(x_i, y_i)\}(|D| = n, x_i \in R_m, y_i \in R)$, the model as following:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

x_i represents the eigenvector of the i[th] training data. The q means the leaf index corresponding to the tree and sample map. T means the number of leaves on each tree. Each tree f_k corresponds to a separate structure q and the weight w.

4. Experimental design

4.1. Experimental data collection

The collected data include meteorological monitoring stations data of Shenyang, industrial enterprises and heating enterprises pollution emissions data, coordinate information of meteorological monitoring stations and enterprises, as well as meteorological data of whole Shenyang. The raw data cover the 11 monitoring stations, 112 pollutant emission enterprises and the meteorological data of Shenyang. According to the distribution of overall data, we select 4 pieces of data with high continuity and integrity for 8 months, which are January, February, June and July of 2016 and 2017 respectively. Since Shenyang is located in the higher latitudes of China, the air pollution caused

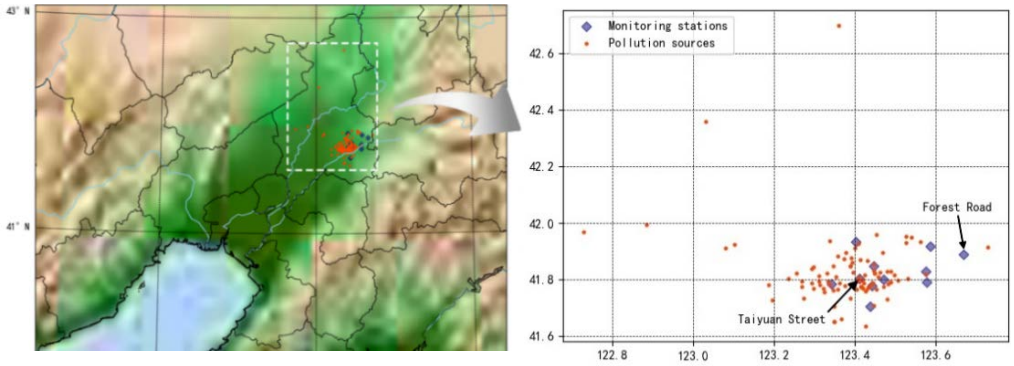

Figure 1. Location relationship between monitoring stations and pollution

by the heating period has been a huge impact, so the selection of data covers two years of the heating period and the non-heating period available to each other, which has a higher representative, as shown in table 1:

Table 1. The quantity of pollution sources in each period

Period	Range	Quantity
I-heating period	1-2 2016	42
II- non-heating period	6-7 2016	9
III- heating period	1-2 2017	67
IV- non-heating period	6-7 2017	17

The geographical position of monitoring stations and pollution emissions is shown in figure 1, where the vertical coordinate of the figure represents latitude and the horizontal coordinate represents longitude. It can be seen from figure 1 that the pollutant sources in the experiment are mostly distributed in the center area of Shenyang. Among them, most enterprises are located near in the monitoring station, and the average distance between the pollutant sources and each monitoring station is 17.8km, and the minimum distance is 0.25km. The proportion of pollutant sources to each monitoring station less than 20km is 72%. However, there are still several factories far away from each monitoring station, the furthest distance is 105km, and the proportion of distance greater than 40km is 7%.

4.2. Experimental data processing

Because the raw data is different on the scale of features in dimensions, we need to make the data standardized to eliminate the differences between features in order to avoid, to some extent, the influence of data features with larger magnitude on the effect of those with smaller magnitude; at the same time, data standardization also improves the convergence speed of the model.

According to the longitude and latitude data of pollution sources and monitoring stations, we calculate the distance (km) and azimuth angle (°) from each pollution source to each monitoring station. To get the results in a most objective way, we add the spatial division dimension features. What's more, we have divided the azimuth angle into eight directions, so as to facilitate the subsequent implementation of different experiments and the statistics of wind direction data.

We counted all the weather types and divided them into eight categories. Weather type are treated with One-Hot Encoding. The cross combination of wind direction feature and relative position of pollution sources and monitoring stations will also be presented in this paper.

Data feature dimensions and their units and standardization methods are shown in table 2:

Table 2. Feature description and standardization

Type	Symbol	Unit	Standardized
monitoring	PM2.5	μg/m3	-
	PM10	μg/m3	log(x+1)/log(max+1)
pollution	CO	mg/m3	log(x+1)/log(max+1)
	NO	μg/m3	log(x+1)/log(max+1)
	NO2	μg/m3	log(x+1)/log(max+1)
	NOx	μg/m3	log(x+1)/log(max+1)
	SO2	μg/m3	log(x+1)/log(max+1)
	O3	μg/m3	log(x+1)/log(max+1)
	SO2_t	kg/h	log(x+1)/log(max+1)
	NOx_t	kg/h	log(x+1)/log(max+1)
	pm_t	kg/h	log(x+1)/log(max+1)
meteorological	humi	%	min-max
	pressure	hpa	min-max
	temp	℃	min-max
	winddirct_x	-	sin function transformation*
	winddirct_y	-	cos function transformation *
	windspeed	m/s	min-max
	weathertype	-	One-Hot
spatial	distance	km	min-max
	azimuth	degree	min-max

4.3. Evaluation standard

In order to verify the validity of the proposed model in this paper, five criteria are adopted to evaluate the performance of proposed model. The model evaluation includes P_v, P_l, F_l, M_l and *cost*. P_v means the accuracy of PM2.5 concentration prediction; P_l means the accuracy of PM2.5 level prediction, which based on the prediction of PM2.5 concentration; false alarm rate F_l and miss alarm rate M_l means the error of PM2.5 level prediction; cost function *cost*. The definitions of each indicator are as follows:

accuracy of PM2.5 concentration prediction: $P_v = 1 - \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n} y}$

accuracy of PM2.5 level prediction: $P_l = \left(\frac{count\ of\ correct\ level\ prediction\ of\ PM2.5}{total\ prediction\ times}\right) \times 100\%$

false alarm rate of PM2.5 level prediction: $F_l = \left(\frac{count\ of\ false\ alarm\ of\ level\ prediction\ of\ PM2.5}{total\ prediction\ times}\right) \times 100\%$

miss alarm rate of PM2.5 level prediction: $M_l = \left(\frac{count\ of\ miss\ alarm\ of\ level\ prediction\ of\ PM2.5}{total\ prediction\ times}\right) \times 100\%$

cost function: $cost = \sqrt{\frac{1}{N}\sum(y_i - \hat{y}_i)^2}$

The *y* is the actual measured concentration of PM2.5, $\hat{y}$ is the predict concentration of PM2.5, concentration levels of PM2.5 in accordance with China's national standard GB3095-2012 and HJ 633-2012, as shown in the table 3:

Table 3. Definition of concentration level of PM2.5

PM2.5 concentration (μg/m³)	concentration level
0-35	Level 1
35-75	Level 2
75-115	Level 3
115-150	Level 4
150-250	Level 5
250-500	Level 6

5. Analysis of results

5.1. Algorithms comparison

At the beginning of experiments, we take the training data which contains all the features as the input into five algorithms to respectively train. Then, we use the testing data to evaluate the accuracy of models. From the perspective of the overall prediction results, the prediction performance of each model in every monitoring point is different in all periods. In order to evaluate the prediction results objectively, after analyzing 132 (11*4*3) evaluation results of P_v, P_l and *cost* indicators at 11 monitoring stations over four periods, statistical methods were used to evaluate the prediction results of each algorithm at different monitoring stations. The results of the experiments measured by P_v, P_l and *cost* for five selected algorithms are presented in Table 4:

Table 4. Evaluation of algorithm forecast result

Algorithm	optimal	P_v-mean	P_v-maximum	P_l-mean	P_l-maximum	cost-mean	cost-minimum
SVM	2	76.9	87.7	78.6	92.3	15.8	5.7
MLP	52	79.8	87.1	80.9	92.5	12.9	5.6
RF	2	77.7	86.6	78.2	90.5	14.3	6.7
GBDT	13	78.6	86.9	79.7	92.8	13.3	6.1
XGBoost	63	80.5	87.9	81.9	93.4	12.6	6.0

It can be seen from the table above that XGBoost has the best results. MLP and GBDT algorithms have lower accuracy than XGBoost in prediction results, but they still perform better than SVM and RF. Table 4 shows that the models trained by XGBoost, GBDT and MLP are better than those trained by other algorithms. However, the difference between the mean and the maximum of the prediction results of each model is generally large. Therefore, in the next section, the prediction results of the models trained by the three algorithms, i.e. XGBoost, GBDT and MLP, will be emphatically analyzed in different periods.

5.2. Periods comparison

The difference between the mean value and the maximum value of the prediction results is mainly

Table 5. Evaluation of prediction results in each period

Period	Algorithm	P_v			P_l			cost		
		mean	maximum	minimum	mean	maximum	minimum	mean	maximum	minimum
I-heating period	MLP	76.71	83.22	64.19	71.59	79.70	56.88	16.09	25.55	10.75
	GBDT	77.54	82.77	71.72	74.13	81.68	66.68	15.24	21.90	10.17
	XGBoost	80.11	85.82	76.42	77.73	85.49	73.37	14.17	20.55	9.47
II-non-heating period	MLP	82.15	87.12	74.70	86.16	91.69	79.56	9.66	13.37	5.56
	GBDT	82.64	86.89	75.82	85.68	90.62	81.04	9.25	12.73	6.05
	XGBoost	83.00	87.86	74.51	86.54	91.63	81.45	9.31	13.99	5.97
III-heating period	MLP	80.31	83.48	73.94	76.83	82.55	66.76	16.46	24.44	11.58
	GBDT	74.19	78.79	69.72	70.06	79.64	60.25	18.89	24.68	12.77
	XGBoost	77.49	82.08	72.06	74.45	82.13	62.74	17.39	24.55	11.38
IV-non-heating period	MLP	79.94	85.25	72.19	88.97	92.50	82.65	9.40	13.19	7.51
	GBDT	79.84	85.63	73.14	88.82	92.83	81.51	9.68	17.04	7.11
	XGBoost	81.50	87.95	75.54	88.95	93.37	81.85	9.46	17.11	6.41

because the concentration value and meteorological features of pollutants in the air are different in heating and non-heating periods. In addition, the number of enterprises in the heating period and the non-heating period also varies greatly, which makes the difference between each prediction model in each period more obvious. Generally, the prediction accuracy of heating period is lower than non-heating period, and its fluctuation is large, while the overall prediction effect of non-heating period is better. The results of the experiments measured by P_v, P_l and *cost* for three selected algorithms are presented in Table 5.

It can be clearly seen from the table 5 that the prediction effect of heating period is lower than that of non-heating period, and the fluctuation interval of PM2.5 concentration level prediction accuracy is the largest.

During heating period，the fluctuation of PM2.5 concentration might be the main reason which cause the different results. At the same time, the heating enterprises began to work during the heating period. The increase in the number of pollution sources caused a dramatic increase in the number of input features of the prediction model, which caused the degree of fitting of the prediction model to be affected and the fluctuation of the prediction results to be increased accordingly.

In order to improve the prediction accuracy of the model, the input features of the model are analyzed. Due to the large number of input feature dimensions, some features with low correlation will interfere the prediction accuracy of the model. Therefore, this paper extracts the features that have a great contribution to the model, and reorders the features according to the weight, then selects the part with high weight as the input of model training, so as to reduce the interference of unimportant features and improve the prediction accuracy. XGBoost algorithm can output the importance of each feature in the training data, which can be used to analyze the importance of each monitoring point in each period.

1. Ranking of the importance of monitoring data and meteorological data

Statistical methods were used to rank the importance of 25 input features, excluding industrial emission data. Among them, the features that have a large impact on PM2.5 prediction include all monitoring data, wind speed, temperature at the monitoring stations, and urban average humidity and wind direction data. See table 6 for details:

Table 6. Top 13 important features

Type	Symbol	Average ranking
monitoring	PM2.5	1
	PM10	2
	NO_2	5
	SO_2	6
	CO	7
	NO_x	8
	NO	9
	O_3	10
meteorological	humi	3
	windspeed	4
	temperature	11
	winddirct_y	12
	winddirct_x	13

2. Ranking of the importance of industrial emission data

Due to the geographical location of different monitoring stations, the density of surrounding pollution sources is also different. Therefore, two

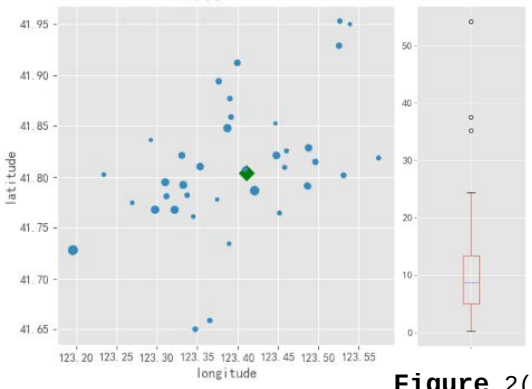

Figure 2(a). Distribution of pollution sources around Taiyuan Street

Figure 2(b). Distance of pollution sources and Taiyuan Street

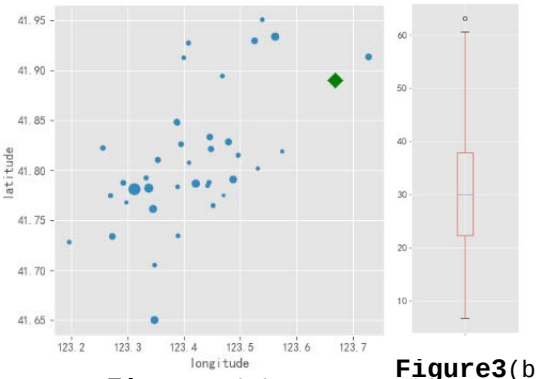

Figure 3(a). Distribution of pollution sources around Forest

Figure3(b). Distance of pollution sources and Forest Road

monitoring stations with different geographical features were selected to find out the influence of different features of pollutant emission stations on the prediction results, namely Taiyuan Street monitoring station located in the downtown area and Forest Road monitoring station with relatively remote location. Figure 1 shows the location of Taiyuan Street and Forest Road in the geographic coordinate graph, where the horizontal axis represents the longitude value and the vertical axis represents the latitude value.

For the convenience of display, the prediction results of "III- heating period" were selected for visual analysis. Figure 2(a) shows the distribution of the major pollution sources around the monitoring station of Taiyuan street. The size of the blue dot in the figure represents the importance of pollution sources. The larger the point, the more important the pollution sources are. It can be observed that compared with other monitoring stations, Taiyuan street is located in the central position of Shenyang, and there are a large number of pollution sources around and the distribution is even. Figure 2(b) is the distance between the pollution sources and the monitoring station of Taiyuan Street, whose vertical coordinate unit is km. As can be seen from figure 2, the pollution sources around Taiyuan Street monitoring points are densely distributed, and the average distance of the pollution sources that have a great influence on PM2.5 is about 10km. Most pollution sources are in a circular area with a radius of 15km centering on Taiyuan Street. At the same time, it can be found that the distance between the pollutant sources and the monitoring station has an impact on the importance of features, but there are also some cases where the pollution source is far away from the monitoring point, but it has a great impact. After analyzing the emission of these pollution sources, it is found that the emission is also a factor that influences the importance of the results.

Figure 3(a) and Figure 3(b) show the prediction results of Forest Road. Because the Forest Road monitoring station is located in the northeast corner, the most influential pollution source is located in the southwest of Forest Road. The average distance of pollution sources that have a great influence on the Forest Road monitoring point is about 30km, and most pollution sources are still within 40km around forest road. It also can be seen from the analysis results of Forest Road, there are also some cases where the pollution source is far away from the monitoring point, but it has a great impact. After analyzing the emission of these pollution sources, it is found that the emission is also a factor that influences the importance of the results.

After the above analysis, it is found that the importance of pollution sources is not entirely determined by distance, but also affected by the emission of pollution sources. Therefore, for the monitoring station of Taiyuan Street, the emission data of pollution sources within 10km and those with higher emissions within 20km are selected as important

features. For the monitoring station of forest roads, the emission data of pollution sources within the range of 30km and those with higher emissions within the range of 40km are selected as important features.

5.3. Model optimization

According to the important monitoring point data and meteorological data selected in section 5.2, as well as the important features selected according to the ranking of pollutant discharge quality and its distance from the monitoring site, the prediction model was re-trained, and the prediction results were evaluated. The prediction performance of the prediction model comparing the two trainings at two monitoring points of Taiyuan Street and Forest Road, we can see that after selecting important features, more than 72% of the indicators in the prediction result evaluation have been improved. In general, the predictive performance of the prediction model is improved after the input feature selection. Figure 4 show the comparison between the predicted values, provided by the prediction model of XGBoost algorithm training, and the actual values of PM2.5 concentration at Forest Road during IV-non-heating-period. In the figure, the red line is the actual value, the blue line is the predicted value, the vertical coordinate is PM2.5 concentration value, and the horizontal coordinate is time. It is not hard to find that the degree of fitting between the predicted value and the actual value is not satisfactory when the PM2.5 concentration is high, and it performs well in other cases.

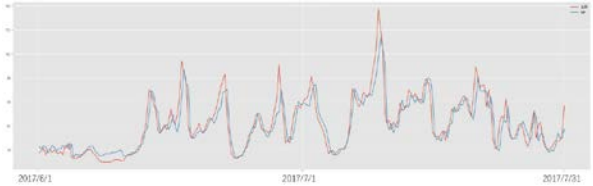

Figure 4. Comparison between the predicted value and the actual value of Forest Road

6. Conclusions

In this paper, the ensemble learning algorithm is used to predict PM2.5 concentration by inputting monitoring data, pollutant emission data, meteorological data and spatial data, and the visualization tool is used to explain the experimental process and results. The experimental results show that the results in heating period is not as good as non-heating period. Because it may be affected by the high concentration of PM2.5 data and the complex environmental situations. By using XGBoost to extract key features again and again, we find that at least 72% of the projections have improved at the selected monitoring stations. Feature selection is found to be helpful in improving prediction accuracy.

This method is also applicable to air quality prediction in other cities. When making PM2.5 concentration prediction in other cities, more pollutant emission data can be appropriately included to improve the accuracy of prediction and the interpretability of the model. The limitation of this study is that the input data only include the data of industrial production and heating emissions, while the data of traffic pollution and life pollution are not available. Therefore, it will be affected to a certain extent by intense changes of external pollution. In future studies, when the accuracy of input data is guaranteed, the spatial structure of pollution sources such as point sources, line sources, surface sources and body sources can be taken into account to improve prediction accuracy.

7. References

[1] McKeen S, Grell G, Peckham S, et al, “An evaluation of real-time air quality forecasts and their urban emissions over eastern Texas during the summer of 2006 Second Texas Air Quality Study field study”, Journal of Geophysical Research: Atmospheres, 2009, 114(D7).

[2] Chuang M T, Zhang Y, Kang D, “Application of

WRF/Chem-MADRID for real-time air quality forecasting over the Southeastern United States", Atmospheric environment, 2011, 45(34): 6241-6250.
[3] Shimadera H, Hayami H, Ohara T, et al, "Numerical simulation of extreme air pollution by fine particulate matter in China in winter 2013", Asian Journal of Atmospheric Environment, 2014, 8(1): 25-34.
[4] Song Y, Qin S, Qu J, et al, "The forecasting research of early warning systems for atmospheric pollutants: A case in Yangtze River Delta region", Atmospheric Environment, 2015, 118: 58-69.
[5] Sun W, Zhang H, Palazoglu A, et al, "Prediction of 24-hour-average PM 2.5 concentrations using a hidden Markov model with different emission distributions in Northern California", Science of the total environment, 2013, 443: 93-103.
[6] Zong X, Wu Z, "SVR Smoggy Forecast Model Based on GA Method Optimization", DEStech Transactions on Computer Science and Engineering, 2015 (cib).
[7] Mihalache, SF, Popescu M , Oprea M, " Particulate Matter 2.5 Air Pollution Forecasting Based on Artificial Intelligence, Romania", International Multidisciplinary Scientific GeoConference-SGEM, 2016, 491-498.
[8] Popescu M, Mihalache S F, Oprea M, "Air Pollutants and Meteorological Parameters Influence on PM2. 5 Forecasting and Performance Assessment of the Developed Artificial Intelligence-Based Forecasting Model", REVISTA DE CHIMIE, 2017, 68(4): 864-868.
[9] Wang D, Wei S, Luo H, et al, "A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine", Science of The Total Environment, 2017, 580: 719-733.
[10] Sun W, Sun J, "Daily PM 2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm", Journal of environmental management, 2017, 188: 144-152.
[11] Yang P, Shan S, Gao W, et al, "Face recognition using ada-boosted gabor features", Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. IEEE, 2004: 356-361.
[12] Pavlyshenko B M, "Linear, machine learning and probabilistic approaches for time series analysis", Data Stream Mining & Processing (DSMP), IEEE First International Conference on. IEEE, 2016: 377-381..
[13] Li L, Situ R, Gao J, et al, "A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity", Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017: 1912-1917.
[14] Palotti J, "Learning to Rank for Personalized E-Commerce Search at CIKM Cup 2016", Tech. rep, 2016.
[15] MAO H, ZHANG K, DI B, et al, "The high-resolution temporal and spatial allocation of emission inventory for Chengdu", Acta Scientiae Circumstantiae, 2017, 1: 00