



PM2.5 Concentration Prediction Based on mRMR-XGBoost Model

Weijian Zhong^{1,2}, Xiaoqin Lian², Chao Gao², Xiang Chen^{1,2(✉)},
and Hongzhou Tan^{1,2}

¹ School of Electronics and Information Technology, Sun Yat-sen University,
Guangzhou 510006, China
chenxiang@mail.sysu.edu.cn

² Key Laboratory of Industrial Internet and Big Data, China National Light
Industry, Beijing Technology and Business University, Beijing 100048, China

Abstract. Air pollution is one of the main environmental pollution, in which air pollution component prediction is an important problem. At present, there have been many studies using machine learning methods to predict air pollution components. However, due to its numerous influencing factors and incomplete determination, there are still problems in accurate prediction. In this paper, the gas factors and meteorological factors collected by the self-developed integrated system are firstly used to construct the original feature set. Then, the mRMR algorithm is used to select data features from the perspective of maximum correlation and minimum redundancy. Finally, a prediction method of PM2.5 concentration in the next hour based on feature selection and XGBoost is designed by combining the data after dimension reduction with the XGBoost model. The experimental results show that mRMR algorithm can effectively select the features of air, and the prediction accuracy is improved even when only half of the features of the original data are used.

Keywords: PM2.5 · Feature selection · mRMR · XGBoost ·
Concentration prediction

1 Introduction

With the development of society, China's economic development has been accelerated, but behind it is often accompanied by environmental pollution. Air pollution is one of the main environmental pollution. Pollution gas can lead to a

The work is supported in part by the State's Key Project of Research and Development Plan under Grants (2019YFE0196400), Industry-University-Research Cooperation Project in Zhuhai (ZH22017001200072PWC), Guangdong Provincial Special Fund For Modern Agriculture Industry Technology Innovation Teams (No. 2021KJ122), and the Open Research Fund Program of Key Laboratory of Industrial Internet and Big Data, China National Light Industry, Beijing Technology and Business University.

variety of diseases, among which PM2.5 has a greater threat to human health due to its small particle size and can cause haze weather. More efficient and accurate forecasting of air pollution components can provide some reference and guiding significance for people's safe travel and related environmental protection departments' work on air pollution prevention and control.

Numerical model prediction is one of the methods for predicting atmospheric pollution components. This method needs a wide range of data and its system is complex, which is relatively immature. The other method is statistical model prediction, which is more convenient and efficient, and its prediction effect is better. Therefore, people generally use statistical models to learn the relationships between numerous relevant features and air pollution components, so as to realize the prediction of air pollution components.

A large number of researches predict air pollution components based on machine learning model. The optimization of the model is mainly in data processing, model selection and combination, and model parameter optimization [1]. In order to get the optimal combination, different combination methods should be analyzed according to the specific data set.

The quality of data features determines the upper limit of machine learning predictive performance [2]. There is a strong correlation between a number of air related features selected in this paper. In order to further improve the performance of air pollution component prediction, this paper introduces the mRMR algorithm to dimension the data features from the perspective of maximum correlation and minimum redundancy, select important features, and reduce the influence of redundancy features. Finally, XGBoost model is used to mine the information between air features and PM2.5 concentration labels, and a high-performance PM2.5 concentration prediction model for the next hour is constructed based on feature selection and XGBoost.

2 Data Set Analysis

The data used for PM2.5 concentration prediction in this paper are collected by the self-research integrated system in Foshan. The data are hourly data for four contaminated gases and four weather factors from 1 March 2020 to 2 September 2020, specifically SO₂, NO₂, PM10, PM2.5 and temperature, relative humidity, wind speed, and wind direction. Finally, combined with the hourly information on the day of the observation data, the complete data set with 9 features used in this paper is formed. The unit of gas, temperature, relative humidity, wind speed and wind direction are $\mu\text{g}/\text{m}^3$, $^{\circ}\text{C}$, $\%R.H.$, m/s and degree respectively. The characteristics of the data set are analyzed in the following part.

2.1 Data Description Statistics

The following is a descriptive statistical analysis of the data set with the help of SPSS statistical software. The statistical results are shown in Table 1. According to the average value of polluted gases, the local air quality during this period

Table 1. Data description statistics.

| Feature | Sample number | Minimum | Maximum | Mean | Standard deviation |
|-----------------|---------------|---------|---------|--------|--------------------|
| HoD | 4390 | 0.00 | 23.00 | 11.50 | 6.9210 |
| SO ₂ | 4390 | 0.10 | 9.54 | 6.32 | 1.5750 |
| NO ₂ | 4390 | 0.18 | 34.98 | 19.19 | 6.5518 |
| PM10 | 4390 | 0.48 | 190.98 | 30.73 | 28.4012 |
| PM2.5 | 4390 | 0.13 | 118.65 | 18.87 | 17.1331 |
| Temperature | 4390 | 13.9 | 40.2 | 28.31 | 5.5065 |
| Humid | 4390 | 24.74 | 100.00 | 76.40 | 14.0128 |
| Speed | 4390 | 0.00 | 37.25 | 7.47 | 4.1491 |
| Direct | 4390 | 43.60 | 322.13 | 156.27 | 58.6787 |

is good; according to the maximum value, serious air pollution exists; according to the standard value, the concentration of PM10 and PM2.5 changes relatively large.

2.2 PM2.5 Time Series

By looking at time series, we can get a general idea of the data. Time series of PM2.5 concentration values are given here, as shown in Fig. 1. As can be seen

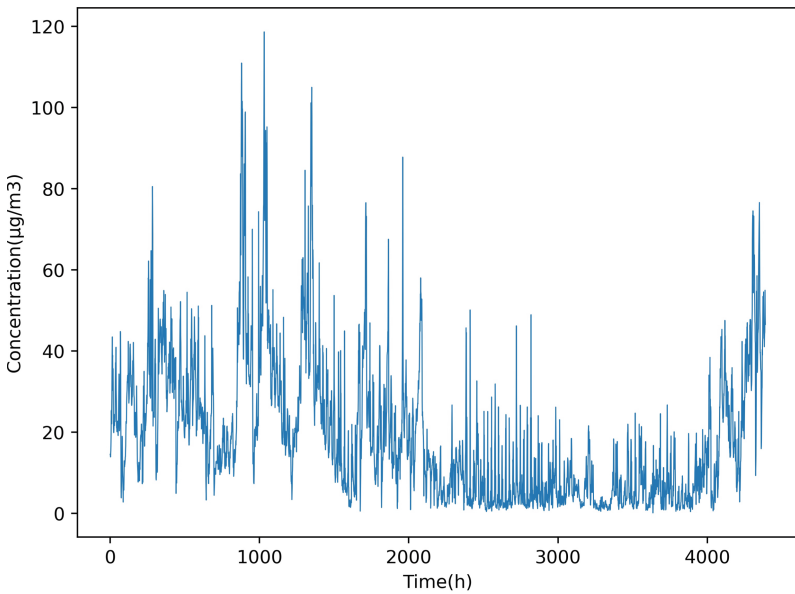


Fig. 1. Time series of PM2.5 concentration values

from the figure, the concentration of PM2.5 varies greatly, which basically stays at a low level in summer.

2.3 Normalized Mutual Information Between Features

Mutual information theory describes how much information a random variable contains another random variable, in which the value range of normalized mutual information is $[0,1]$, and the larger the value is, the larger the amount of information is. The normalized mutual information among the features in the data set is shown in Table 2, where PM2.5n represents the concentration of PM2.5 in the next hour. As can be seen from the table, there is a strong correlation between PM2.5 concentration value in the next hour and several features, and at the same time, there is a certain mutual information among other features, which will lead to a certain degree of redundancy among features.

Table 2. Normalized mutual information between features

| | PM2.5n | HoD | SO ₂ | NO ₂ | PM10 | PM2.5 | Temp | Humid | Speed | Direct |
|-----------------|--------|-------|-----------------|-----------------|-------|-------|-------|-------|-------|--------|
| PM2.5n | 1.000 | 0.472 | 0.781 | 0.892 | 0.932 | 0.918 | 0.703 | 0.923 | 0.870 | 0.949 |
| HoD | 0.472 | 1.000 | 0.274 | 0.429 | 0.497 | 0.472 | 0.196 | 0.483 | 0.388 | 0.528 |
| SO ₂ | 0.781 | 0.274 | 1.000 | 0.750 | 0.798 | 0.781 | 0.522 | 0.788 | 0.721 | 0.820 |
| NO ₂ | 0.892 | 0.429 | 0.750 | 1.000 | 0.906 | 0.892 | 0.670 | 0.898 | 0.842 | 0.925 |
| PM10 | 0.932 | 0.497 | 0.798 | 0.906 | 1.000 | 0.932 | 0.721 | 0.937 | 0.884 | 0.962 |
| PM2.5 | 0.918 | 0.472 | 0.781 | 0.892 | 0.932 | 1.000 | 0.702 | 0.923 | 0.870 | 0.949 |
| Temp | 0.703 | 0.196 | 0.522 | 0.670 | 0.721 | 0.702 | 1.000 | 0.711 | 0.634 | 0.745 |
| Humid | 0.923 | 0.483 | 0.788 | 0.898 | 0.937 | 0.923 | 0.711 | 1.000 | 0.876 | 0.954 |
| Speed | 0.870 | 0.388 | 0.721 | 0.842 | 0.884 | 0.870 | 0.634 | 0.876 | 1.000 | 0.903 |
| Direct | 0.949 | 0.528 | 0.820 | 0.925 | 0.962 | 0.949 | 0.745 | 0.954 | 0.903 | 1.000 |

3 mRMR Algorithm and XGBoost Prediction Model

A reliable feature set is an important part of machine learning, so we analyse air data sets from different perspectives in the previous section. All the features in the data set are collected by the actual system. However, it is not certain whether all the features are strongly correlated with PM2.5 concentration, or some redundancy may occur among the features, which reduces the prediction performance of the model. Therefore, in order to ensure the reliability of the prediction model, we need to properly process all the features before feeding the data set into the machine learning model. In addition, we also need to choose the appropriate machine learning model, according to the characteristics of the actual data set, fully mining the data relations in it, so as to achieve better prediction effect. Therefore, the following will introduce the mRMR feature selection algorithm and XGBoost prediction model used in the prediction method in this paper.

3.1 mRMR Feature Selection Algorithm

mRMR is based on mutual information, which is derived from the concept of entropy [3]. Entropy gives abstract information a certain metric, which can be used to describe the uncertainty between random things. Mutual information partly represents a common part between two random variables, and the more common information, the greater the mutual information between them, and thus the greater the interaction between them. The mutual information between random variables x and y can be calculated as follows:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \tag{1}$$

The $p(x)$ and $p(y)$ are the respective edge probability densities of the two variables, and $p(x, y)$ is their binary probability distribution. It can be seen that mutual information is the statistical mean of random variables x and y under the probability distribution of $p(x, y)$.

mRMR is a feature selection algorithm which on the one hand measures the correlation between two sets based on mutual information to identify the subset of features with the greatest correlation with the target set. On the other hand, it measures the redundancy among the features in the set on the basis of maximum correlation, so as to exclude the redundancy in the features on the basis of maximum correlation. Therefore, mRMR eliminates redundant information on the basis of retaining key features to achieve the purpose of reducing the complexity of the model and effectively preventing the over-fitting problem of machine learning models [4]. The maximum correlation in mRMR can be expressed as:

$$\max D(S, c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c). \tag{2}$$

The $I(x_i, c)$ is the mutual information described above, here specifically represents the mutual information between the variables x_i and c , and S is the feature matrix containing all x_i . $|S|$ is the characteristic matrix dimension used to represent the number of elements in S , and c is the target variable. It can be seen that the maximum correlation is expressed as the average value of mutual information. The minimum redundancy can be expressed as:

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \tag{3}$$

The $I(x_i, x_j)$ represents the mutual information between the variables x_i and x_j . Finally, the criterion for selecting the optimal feature subset is given:

$$\begin{cases} \max \psi(D, R) \\ \psi = D - R \end{cases} \tag{4}$$

3.2 XGBoost Prediction Model

Both XGBoost and GBDT algorithms belong to Boosting algorithm. Based on GBDT, XGBoost is proposed and mainly optimizes its objective function and improves the basic learning machine [5]. The XGBoost model works by providing a number of weak learners and adding up their base predictions and residuals to get the final prediction. Its expression is as follows:

$$\hat{y}_i = \sum_{i=1}^k f_i(x_i). \quad (5)$$

The k represents k weak learners, f_i represents the i th weak learner, x_i represents the i -th data set, and \hat{y}_i represents the final predicted value after accumulation.

XGBoost model is an ensemble learning framework, which contains many decision trees. Different from other ensemble learning tree models, the training process of XGBoost model is more complex. Compared with GBDT, which is also Boosting algorithm, it adds regular term to the objective function and optimizes the feature splitting process. Its objective function is expressed as follows:

$$obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{i=1}^k \Omega(f_i). \quad (6)$$

The $L(y_i, \hat{y}_i)$ represents the loss function, $\Omega(f_i)$ represents the regular term of the i -th tree, and n represents the sample size. Unlike GBDT, the loss function can be defined case by case, training a tree to make the target function as small as possible. When training the t -th tree, its objective function can be approximately expressed as follows after second-order Taylor expansion:

$$obj^t \approx \sum_{i=1}^n \left[g_i f_t(x_i) + 0.5 h_i f_t(x_i)^2 \right] + \Omega(f_t). \quad (7)$$

The g_i represents the first derivative of the loss function, h_i represents the second derivative of the loss function, and the regular term is:

$$\Omega(f_t) = \gamma T + 0.5 \lambda \sum_{j=1}^T w_j^2. \quad (8)$$

At this point, the minimum point of the objective function can be easily obtained [6], so as to finally calculate the minimum value of the objective function. The smaller the minimum of the objective function is, the better the performance of the tree is. In order to obtain the minimum of the objective function, the greedy algorithm can be used to find the tree that can get the minimum of the objective function from a series of tree structures. Then, by calculating and comparing the information before and after splitting, when a certain information gain can be obtained, the characteristic splitting is adopted. And so on, all the tree structures can be obtained eventually, until all the learning of the weak learner is finished and the training of the whole model is completed.

4 Experimental Studies

4.1 Evaluation Index of the Experimental Results

In order to objectively and reasonably judge and compare the prediction performance of the model, certain evaluation indexes need to be selected. In this paper, three evaluation indexes, $RMSE$, MAE and R^2 , are adopted, and their expressions are shown as follows:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2}, \tag{9}$$

$$MAE = \frac{1}{m} \sum_{i=1}^m |(y_i - \hat{y}_i)|, \tag{10}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{11}$$

The y_i represents the actual observed value, \hat{y}_i represents the predicted value, and m represents the number of samples.

$RMSE$ is the root mean square error, which can describe the degree of difference between the predicted value and the measured value. When the value is large, it generally means that most of the samples have large differences. MAE is the mean absolute error, which gives an average of the difference between the actual observed value and the predicted value. R^2 , the coefficient of determination, describes how much of a change in the real value is due to the predicted value.

4.2 Comparative Experiment of Prediction Model

In the selection of prediction models, SVM, KNN and RF prediction models are used to compare with XGBoost. In the experimental process, all nine features of the data set are adopted to divide the first 3900 samples in the data set into training sets, and the last 490 samples into verification sets, the same with the subsequent experiments. The experimental results are shown in Table 3. The unit of time is second.

Table 3. Evaluation indexes of prediction performance of different models

| Model | Training time | RMSE | MAE | R^2 | Forecast time |
|---------|---------------|---------|---------|---------|---------------|
| SVM | 1.033249 | 5.04196 | 3.54826 | 0.90921 | 0.140781 |
| KNN | 0.006981 | 5.06853 | 3.55584 | 0.90825 | 0.005984 |
| XGBoost | 0.228560 | 4.49273 | 3.08974 | 0.92791 | 0.000998 |
| RF | 0.966385 | 4.86782 | 3.44089 | 0.91537 | 0.007978 |

It can be found from the experimental results that the training time of XGBoost prediction model is short, and compared with other models, its prediction error is the smallest, the degree of fit is the highest, and the prediction time is the shortest. Since the XGBoost model has the best prediction effect of PM2.5 concentration in the next hour, the following experiments will be conducted on the mRMR feature selection algorithm based on this model.

4.3 Comparative Experiment of Feature Selection Algorithm

Finding out important features in the data set can reduce the feature dimension and obtain a more efficient dataset, so as to improve the performance of the prediction model. In this paper, mRMR algorithm is used for feature selection. Meanwhile, three feature selection algorithms, Pearson [7], PCA [8] and ReliefF [9], are used to compare with it. XGBoost is used as the prediction Model to study their performance in PM2.5 concentration prediction. The experimental results are shown in Table 4, which lists the evaluation indexes that each feature selection algorithm can make the prediction effect optimal and the number of features used.

Table 4. The optimal performance and the number of features of each feature selection algorithm

| | <i>RMSE</i> | <i>MAE</i> | R^2 |
|---------|----------------|----------------|----------------|
| Pearson | 4.49273[n = 9] | 3.01377[n = 8] | 0.92791[n = 9] |
| PCA | 4.53118[n = 9] | 3.19081[n = 9] | 0.92667[n = 9] |
| ReliefF | 4.49273[n = 9] | 3.08974[n = 9] | 0.92791[n = 9] |
| mRMR | 4.43299[n = 5] | 2.99519[n = 6] | 0.92982[n = 5] |

The process of feature selection is to continuously exclude features with low scores, and finally select the situation that can achieve the best prediction effect. Figure 2 and Fig. 3 are the trend graphs of predicting R^2 and *RMSE* in the process of feature selection using each algorithm, which is helpful for us to further understand the process of feature selection method. To facilitate the understanding of the important details, the longitudinal axis in the figure is limited to a certain range, where the value of the *RMSE* curve corresponding to the PCA feature selection algorithm is 16.59454 at the characteristic number of 1 and the corresponding R^2 curve is 0.01650 at the characteristic number of 1.

As can be seen from the figures, the overall trend of model prediction evaluation increases with the increase of features, because the more features there are, the more information they can provide to the model. It can also be found that the mRMR feature selection algorithm can make the model get better performance in the case of learning fewer features, while the other three algorithms can hardly get better prediction effect than using the original data set in the process

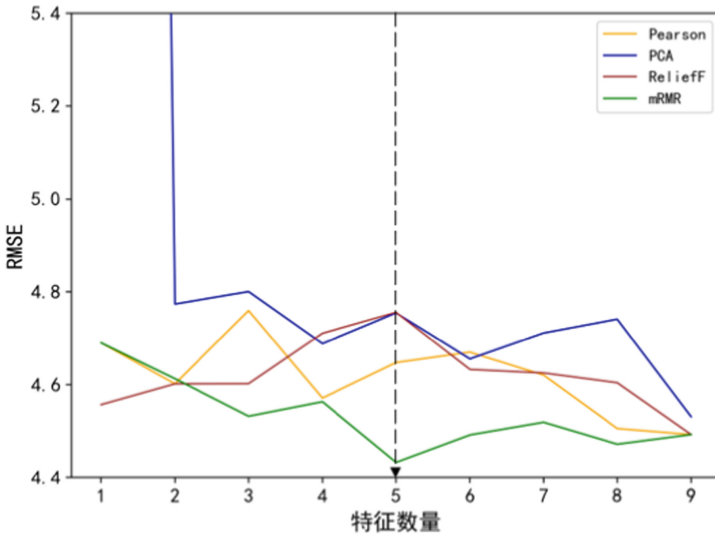


Fig. 2. RMSE curve of each feature selection method

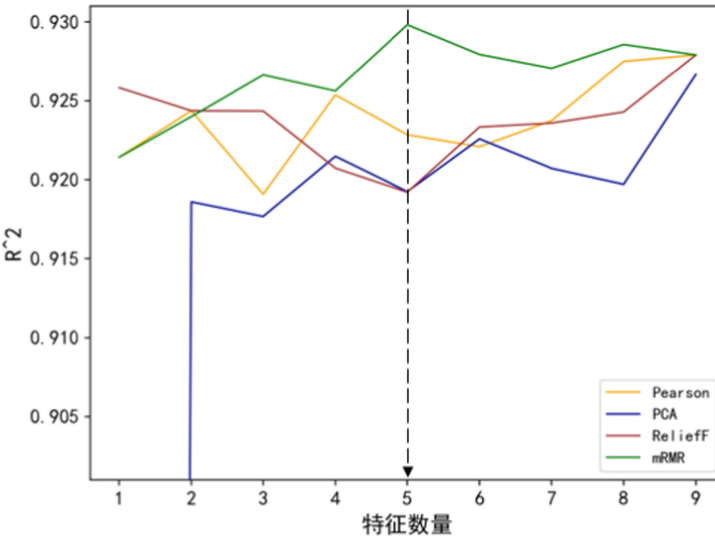


Fig. 3. R² curve of each feature selection method

of feature selection. This is because there is indeed redundancy between features, and some features help the model less than their own redundancy does. However, only when the features that contribute less to the prediction are accurately eliminated can better results be obtained. In the process of feature selection using mRMR algorithm, when the corresponding prediction performance evaluation

indexes $RMSE$ and R^2 reach the optimum, only 5 features in the data set are used, which can not only reduce the computational complexity of the model, but also improve the prediction performance of the model.

5 Conclusion

In this paper, the air related data collected by the self-research integrated system is used to form the original data set. In order to make the data set more reliable, the mRMR feature selection algorithm is used to reduce its dimension to get the feature set that can make the prediction effect optimal. Finally, the selected data set is combined with the XGBoost prediction model for learning, training and prediction, and the mRMR-XGBoost model for the next 1 h concentration of PM_{2.5} is designed. The experimental results show that the mRMR-XGBoost model has a good performance in the prediction of PM_{2.5} concentration.

References

1. Zamani Joharestani, M., Cao, C., Ni, X., et al.: PM_{2.5} prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *J. Atmosphere* **10**(7), 373 (2019)
2. Iskandaryan, D., Ramos, F., Trilles, S.: Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. *J. Appl. Sci.* **10**(7), 2401 (2020)
3. Li, J.X., Liu, X., Liu, J., Huang, J.: Prediction of PM_{2.5} concentration based on MRMR-HK-SVM model. *J. Chin. Environ. Sci.* **39**(6), 2304 (2019)
4. Xu, X., Ren, W.: Prediction of air pollution concentration based on mRMR and echo state network. *J. Appl. Sci.* **9**(9), 1811 (2019)
5. Ma, J., Yu, Z., Qu, Y., et al.: Application of the XGBoost machine learning method in PM_{2.5} prediction: a case study of Shanghai. *J. Aerosol Air Qual. Res.* **20**(1), 128–138 (2020)
6. Pan, B.: Application of XGBoost algorithm in hourly PM_{2.5} concentration prediction. In: *IOP Conference Series: Earth and Environmental Science*, vol. 113, no. 1, p. 012127. IOP Publishing (2018)
7. Benesty, J., Chen, J., Huang, Y., et al.: Pearson correlation coefficient. In: Benesty, J., Chen, J., Huang, Y. (eds.) *Noise Reduction in Speech Processing*, pp. 1–4. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00296-0_5
8. Martinez, A.M., Kak, A.C.: PCA versus LDA. *J. IEEE Trans. Pattern Anal. Mach. Intell.* **23**(2), 228–233 (2001)
9. Zhang, Y., Ding, C., Li, T.: Gene selection algorithm by combining reliefF and mRMR. *J. BMC Genom.* **9**(2), 1–10 (2008)