PM$_{2.5}$ concentration prediction based on pseudo-F statistic feature selection algorithm and support vector regression

Wenbo Liu$^{1,2}$, Shengnan Liang$^{1,2,3}$ and Quan Yu$^{1,2}$

$^1$School of Mathematics and Statistics, Qiannan Normal University for Nationalities, Duyun 558000, China;
$^2$Key Laboratory of Complex Systems and Intelligent Computing, Qiannan Normal University for Nationalities, Duyun 558000, China

Abstract. Effective monitoring and prediction of PM$_{2.5}$ concentration is of great practical significance to People's Daily life and social development. In order to facilitate accurate prediction of the average daily concentration of PM$_{2.5}$, a Feature Selection Algorithm Based on Pseudo F Statistic (FSPF) was proposed in this paper, and a PM$_{2.5}$ concentration prediction model was combined with support vector regression. When constructing the model, all features are not selected into the model, and the pseudo-F statistic feature selection algorithm is used to obtain more important feature variables related to PM$_{2.5}$ concentration. On this basis, support vector regression based on different kernel functions is used to predict PM$_{2.5}$ concentration. Through real data analysis, the hyperbolic tangent kernel support vector regression model (FSPF-SVR) selected by pseudo-F statistics features has the highest prediction accuracy when predicting the average PM$_{2.5}$ concentration of Beijing in November and December 2017, with mean square errors of $1.441 \times 10^{-3}$mg$^2$/m$^6$ and $1.307 \times 10^{-3}$mg$^2$/m$^6$ respectively, which is better than the all-variable prediction model. At the same time, it is also better than random forest, weighted K nearest neighbor, Bagging regression, artificial neural network and other current mainstream machine learning regression prediction algorithms.

1. Introduction
In recent years, with the acceleration of China's industrialization and urbanization, environmental pollution, especially air pollution, has become increasingly serious, and the causes of environmental problems have become increasingly complex, which has become one of the focus issues of the public and social media. Numerous studies have shown that when people are exposed to high concentrations of air pollutants for a long time, it will cause serious damage to their health$^{[1,2]}$. PM$_{2.5}$, as the main component of air pollution, refers to atmospheric particles smaller than 2.5 microns in diameter in the air$^{[3]}$. Compared with PM$_{10}$, PM$_{2.5}$ is light in weight and small in size, and it is easy to absorb some harmful substances, such as viruses, bacteria and other pollutants, which will have a serious impact on the cardiopulmonary function of human body$^{[4]}$. Therefore, effective monitoring and prediction of PM$_{2.5}$ concentration has practical guiding significance for urban air quality assessment, arranging people to travel reasonably, and providing reasonable decision-making for government departments.

There are two main ways to predict PM$_{2.5}$ concentration, which are based on traditional statistical models and machine learning methods. In traditional statistical prediction model, the main use linear or nonlinear regression$^{[5,6]}$ and time series model$^{[7]}$, but due to the predicted variable of input and output of pollutants concentration data has very strong nonlinearity, uncertainty and instability
characteristics, leading to some statistical model is difficult to accurate prediction of pollutant concentration in the atmosphere[8,9]. In recent years, the prediction model based on machine learning algorithms compared to the traditional statistical methods showed evident advantages, and most of the research method for the integration of a variety of methods, such as wavelet analysis and artificial neural network[10], the weighted K nearest neighbour, and the BP neural network[11], empirical mode decomposition and support vector regression[12], etc.

Among many machine learning algorithms, Support Vector Regression (SVR) has the advantages of solid mathematical theory foundation, strong generalization ability, difficulty in getting into local minima, and applicable to the processing of nonlinear and non-stationary data. Therefore, this paper proposes a feature selection algorithm based on pseudo-F statistics to obtain input features related to PM$_{2.5}$ concentration. Based on this, Support Vector Regression is used to predict the average daily PM$_{2.5}$ concentration in Beijing. Compared with other literature studies, the hybrid prediction model is still adopted. The difference is that all characteristic variables related to PM$_{2.5}$ concentration are not selected into the model. Instead, importance analysis of the variables is conducted introduced during the modeling process to eliminate variables with low correlation, so as to simplify the model and improve the prediction efficiency. Through empirical data analysis, it can be verified that the FSPF-SVR model has better prediction accuracy, which is better than the full-variable prediction model in most cases. Meanwhile, compared with other mainstream machine learning, SVR has better prediction performance.

2. Pseudo-F statistic feature selection algorithm

The purpose of feature selection is to find out the important variables that are meaningful for the classification and prediction of output variables from the numerous input variables, and eliminate the input variables that are meaningless for the classification and prediction of output variables, so as to improve the prediction performance of the classification model and reduce the complexity.

The basic idea of the feature selection algorithm for pseudo-F statistics proposed in this paper is that, considering the feature variables themselves, the more important the feature variables are, the more discrete their values are. Generally speaking, such variables are meaningless for classification prediction if the values of input variables do not change much or remain basically the same. Based on this, we construct a statistic that can reflect the discrete degree of the variable value to evaluate the importance of the input characteristic variable.

If the original data set contains $n$ observations, $p$ input features and $k$ category label variables, and the observations under each category are $n_i$, the data set can be denoted as

$$D = \{(x_i, y_i), i=1,2,...,n\}$$

Where $x_i = (x_{i1}, x_{i2}, ..., x_{ip})'$, $y_i \in \{1,2,...,k\}$, $\sum_{i=1}^{n} n_i = n$.

Based on the original data set, it is necessary to construct the statistics to evaluate the importance of each input feature. In order to facilitate the use, the sample moment is directly used to replace the total moment in the process of constructing the statistics.

Let $\bar{x}_y = \frac{1}{n} \sum_{i=1}^{n} x_{iy}$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{i1}$ respectively represent the sample mean of a characteristic variable under each category and the sample mean of a characteristic variable under all categories, which can be obtained according to the sum of squares decomposition theorem[13].

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_y)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_y - \bar{x})^2$$

(1)

Formula (1) is denoted as $SST = SSE + SSR$.

$SST$ on the left is the sum of squares of total dispersion, which reflects the difference between each characteristic value and the total mean value; $SSE$ is the sum of squares within a group or the sum of squares within a class, which reflects the difference between classes; $SSR$ is called the sum of squares.
between groups, which reflects the difference between categories. The larger the sum of squares between the groups and the smaller the sum of squares within the groups, the greater the difference of the values of the characteristic variables under different categories, and the greater the role of the input variables in the process of classification and prediction. Based on this, the pseudo-F statistics of the ratio of the sum of squares between the groups and the sum of squares within the groups are constructed and defined as follows

$$PF = \frac{SSR}{SSE} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{x}_{ij} - \bar{x})} {\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x})^2}$$ \hspace{1cm} (2)

Evaluate the importance of each input feature in the classification process according to formula (2). In practical application, p pseudo-F statistics need to be calculated, and then the importance of each variable should be arranged in descending order to find out the important features related to classification and prediction, and eliminate the features with meaningless or weak correlation.

According to the above analysis ideas, the process of feature selection algorithm (FSPF) based on pseudo-f statistics is shown in Table 1.

Table 1. FSPF Algorithm flow.

| Algorithm: FSPF Algorithm |
|---------------------------|
| **Input:** Sample set $D = \{(x_i, y_i), i = 1, 2, ..., n \}$ ; |
| **Output:** Characteristic importance $Measure = \{PF^{(1)}, PF^{(2)}, ..., PF^{(p)}\}$ . |
| 1. Divide the dataset into $k$ subsets by category label as $D_i, i = 1, 2, ..., k ;$ |
| 2. $n_i = \text{row}(D_i)$ |
| 3. for $m$ in 1:p |
| 4. for $i$ in 1:k |
| 5. $\bar{x}_{i}^{(m)} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}^{(m)}$ # The sample mean of the mth attribute in dataset $D_i$ |
| 6. end for |
| 7. Calculate the total sample mean $\bar{x}^{(m)} = \frac{1}{k} \sum_{i=1}^{k} \bar{x}_{i}^{(m)}$ |
| 8. $SSR^{(m)} = \sum_{i=1}^{k} n_i (\bar{x}_{i}^{(m)} - \bar{x}^{(m)})^2$ |
| 9. $SST^{(m)} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij}^{(m)} - \bar{x}^{(m)})^2$ |
| 10. $SSE^{(m)} = SST^{(m)} - SSR^{(m)}$ |
| 11. $PF^{(m)} = \frac{SSR^{(m)}}{SSE^{(m)}}$ |
| 12. end for |

In Table 1, $k$ represents the number of categories contained in the original dataset $D$ , $p$ represents the number of attributes, $n$ represents the sample size of dataset $D$ , and $n_i$ represents the sample size of the $i$ th category. Lines 1 to 2 of the FSPF algorithm belong to the process of dividing and marking the original data set. Lines 3 to 12 belong to the pseudo-F statistic feature selection algorithm, where lines 3 to 6 represent the sample mean of the $p$ attribute features of each category. Lines 7 through 11 represent pseudo-F statistics that calculate the attributes' characteristics.

The importance of each attribute is denoted as $PF_1 > PF_2 > ... > PF_p$ . To determine how many features are most appropriate to select into the classification model. Generally, through direct experimental analysis of the data set, the number of features is gradually increased from the most
important features. The performance index of the classification model in the initial stage will gradually increase. If the performance of the classification model decreases significantly or remains basically unchanged when the number of features increases to a certain number, the number of features will stop increasing. At this point, the number of features will be the final selected result.

3. Support vector regression

Support vector regression takes the training sample set as the data processing object. By analysing the quantitative relationship between the input variables and the numerical output variables, it constructs the "Support Vector" model, inputs the newly observed attributes into the model, and obtains the corresponding predicted value. The basic principle is as follows.

Given train set \( D = \{ (x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, ..., n \} \), we want to get the following model:

\[
f(x) = \omega^T x + b
\]

Where \( \omega = (\omega_1, ..., \omega_p)^T \), \( b \) is the parameter to be estimated. This is similar to the traditional multiple regression, but the difference is that the traditional regression model is based on the difference between the output value \( f(x_i) \) and the real value \( y_i \), and the loss is zero only when it is \( f(x_i) = y_i \). And the support vector regression is only calculated when \( |f(x_i) - y_i| > \varepsilon \).

\[
I_\varepsilon(x) = \begin{cases} 0, & |x| \leq \varepsilon \\ |x| - \varepsilon, & |x| > \varepsilon \end{cases}
\]

Formula (4) is called the \( \varepsilon \) - insensitive loss function.

SVR model can be shown

\[
\min_{\omega, b} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{n} I_\varepsilon(f(x_i) - y_i)
\]

The relaxation variable \( \xi \) is introduced and equation (5) can be expressed equivalently as

\[
\min_{\omega, b, \xi, \xi^*} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\
\text{s.t. } f(x_i) - y_i \leq \varepsilon + \xi_i^* \\
y_i - f(x_i) \leq \varepsilon + \xi_i \\
\xi_i, \xi_i^* \geq 0
\]

Where \( \xi = \max(0, |f(x) - y| - \varepsilon) \), SVR method is shown in Figure 1.

![Figure 1.](image)

**Figure 1.** The support vector regression diagram, the shaded part between the two dashed lines represents interval band, and the samples falling into it are not counted as losses.

![Figure 2.](image)

**Figure 2.** January 1, 2017-December 31, 2017 air quality index grade distribution in Beijing.
4. Experimental results and analysis

4.1. Data description and exploratory analysis

This paper studies the data objects for Beijing on January 1, 2017 - December 31, 2017, the daily average PM$_{2.5}$ concentrations, from China's air quality on-line monitoring historical data analysis platform (https://www.aqistudy.cn/historydata/Monthdata.php). The air quality data contained in the data set are PM$_{2.5}$ (μg/m$^3$), PM$_{10}$ (μg/m$^3$), SO$_2$ (μg/m$^3$), CO (μg/m$^3$), NO$_2$ (μg/m$^3$), and O$_3$ (μg/m$^3$); The meteorological data are Maxt (maximum temperature, °C), Mint (minimum temperature, °C), Weather, and Wind. The data is a multivariate dynamic time series, and its significant feature is that the sequence itself has autocorrelation. Let PM$_{2.5,1}$ represent the previous day's concentration of PM$_{2.5}$, and so on.

In order to remove the influence of dimension, the analysis data need to be normalized:

\[ x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \]

Where \( x \) is the original data, \( x_{\min} \) and \( x_{\max} \) respectively represent the minimum and maximum values of the column in which the data is located. After the prediction is completed, normalization is also required:

\[ \hat{x} = x' \times (x_{\max} - x_{\min}) + x_{\min} \]

\( \hat{x} \) is the final predictable value.

In order to better forecast PM$_{2.5}$, it is necessary to have a general understanding of the distribution of air quality levels in Beijing in 2017, as shown in Figure 2. A, B, C, D, E and F represent air quality as excellent, good, light pollution, moderately pollution, heavy pollution and severe pollution respectively. As can be seen from the annual air quality index in Beijing in Figure 2, the proportion of good grade is the highest. There are about 160 days in a year when the air quality is good, while the number of days when light pollution occurs is 75, which is slightly higher than the excellent grade. Statistics show that the number of days with good and excellent air quality accounts for about 62.46% of the year, while the number of days with light pollution or above accounts for nearly 40% of the year. Air pollution in Beijing has become a very common and serious problem, which requires effective monitoring and treatment.

In a wide range of research on PM$_{2.5}$ concentration prediction in the literature, most of the literature on the analysis to consider the associated with the concentrations of PM$_{2.5}$ air quality data (e.g., PM$_{10}$, CO, etc.) and meteorological data (temperature, wind, etc.), the common is to put all of the predictive variables to join PM$_{2.5}$ concentrations, no advance for variable selection. According to relevant literature studies [14], it is necessary to seek all relevant attribute characteristics when modeling, instead of selecting all features into the model, because some variables may be irrelevant or have little correlation with response variables, and some feature variables show redundancy. Therefore, in the prediction of PM$_{2.5}$ concentration, correlation analysis between prediction variables and response variables should be conducted to understand the correlation between variables in general.

Through the correlation analysis shown in Figure 3, we can find out which main factors affect PM$_{2.5}$ concentration. When analyzing the pollutant components in the air, it can be seen from the relevant analysis diagram that PM$_{2.5}$ and PM$_{10}$, SO$_2$, CO and NO$_2$ show a significant positive correlation, because these gas pollutants are mainly generated by coal combustion during heating and straw combustion in the vast rural areas in the northern region[15]. There is a positive correlation between PM$_{2.5}$ and PM$_{2.5,1}$, indicating that the concentration of PM$_{2.5}$ is affected by the previous day's concentration. PM$_{2.5}$ and O$_3$ show a weak negative correlation, which is because O$_3$ is mainly generated by photochemical reaction under the action of ultraviolet light. When PM$_{2.5}$ concentration is high, its extinction scatters solar radiation[16]. PM$_{2.5}$ has a negative correlation with wind and temperature, which is because when the wind is high and the temperature is high, the air flow will be accelerated, which is conducive to the diffusion of PM$_{2.5}$ particles on the surface.
Through the correlation analysis between PM$_{2.5}$ average daily concentration and various factors, it can be found that some factors do have significant correlation with PM$_{2.5}$ concentration, while some factors only have weak correlation with PM$_{2.5}$ concentration. Therefore, it is necessary to effectively screen the whole variable before the prediction of PM$_{2.5}$ concentration.

**Figure 3.** Correlation analysis between average daily PM2.5 concentration of Beijing and other variables in 2017.

**Figure 4.** Significance measurement of PM$_{2.5}$ prediction variables based on FSPF algorithm.

### 4.2. Prediction of PM$_{2.5}$ concentration based on FSPF-SVR hybrid model

FSPF algorithm is used to select the predictive variables that affect PM$_{2.5}$ concentration. Based on the value of formula (2) pseudo-F statistic, the importance of each predictive variable is measured, and the results are shown in Figure 4.

It can be seen from the importance ranking of the prediction variables given by FSPF algorithm that the degree of influence on PM$_{2.5}$ concentration is ranked from large to small as PM$_{10}$, CO, NO$_2$, O$_3$, CO$_1$, PM$_{2.5}$, PM$_{10}$, O$_{3}$, NO$_2$, SO$_2$, Maxt, Mint, PM$_{2.5}$, SO$_2$, PM$_{2.5}$, Weather, Wind, among which PM$_{10}$ and CO are the most important and the value of pseudo-F statistics is significantly higher than other input variables. NO$_2$, O$_3$, CO$_1$ and PM$_{2.5}$ also rank high in importance, but the values of their pseudo-F statistics are not different from each other. It should be noted that it is not optimal to select all the variables considered important by the algorithm into the model, and further experimental comparison is needed. According to the importance measurement of variables given by the FSPF algorithm, the variables PM$_{10}$, CO, NO$_2$, O$_3$, CO$_1$, PM$_{2.5}$ and so on, which are ranked at the top, should be considered first in the model fitting.

Based on the variable selection of FSPF algorithm, the support vector regression model for PM$_{2.5}$ average daily concentration prediction was established. First, the daily PM$_{2.5}$ concentration data set in Beijing from January 1, 2017 to December 31, 2017 was divided. The first 11 months were used as a training set to train the FSPF-SVR model, and the data in December were used as a test set to verify the accuracy of the model. It is important to note that the fitting precision of SVR model is dependent on the selection of kernel function, in building the model, respectively select Polynomial, gaussian radial basis and hyperbolic tangent sigmoid kernel function, the corresponding models corresponding to SVRPoly, SVRRad and SVRSig.

It can be concluded from Table 2 that the model with the best prediction effect on the average daily PM$_{2.5}$ concentration in December 2017 is the hyperbolic tangent kernel function support vector regression (FSPF-SVRSig) with three variables (PM$_{10}$, CO, NO$_2$), the corresponding MSE is 1.441·10$^{-3}$ mg$^2$/m$^6$, and the corresponding full-variable MSE is 1.769·10$^{-3}$ mg$^2$/m$^6$, indicating that the prediction accuracy of FSPF-SVRSig model is better than that of AV-SVRSig model. The model with the worst prediction effect was FSPF-SVMRad, whose MSE reached 2.744·10$^{-3}$ mg$^2$/m$^6$. From this experimental analysis, it can be concluded that the prediction accuracy of PM$_{2.5}$ average daily concentration is...
influenced by the number of variables selected and the selection of kernel function, and the prediction accuracy of the full-variable model is not necessarily the best.

Table 2. The mean square error of AV-SVR and FSPF-SVR models based on different kernel functions in December test set (Unit: mg²/m⁶).

| Model Name     | MSE       | Model Name     | MSE       |
|----------------|-----------|----------------|-----------|
| AV-SVRPoly     | 2.122 • 10⁻³ | FSPF-SVRPoly  | 2.189 • 10⁻³ |
| AV-SVRRad      | 1.546 • 10⁻³ | FSPF-SVRRad   | 2.534 • 10⁻³ |
| AV-SVRSig      | 1.769 • 10⁻³ | FSPF-SVRSig   | 1.459 • 10⁻³ |

|          | PM₁₀+CO+PM₂₅,₁ | PM₁₀+CO+NO₂ | PM₁₀+CO+NO₂+PM₂₅,₁ |
|----------|-----------------|-------------|--------------------|
| AV-SVR   | 2.189 • 10⁻³    | 2.744 • 10⁻³ | 2.420 • 10⁻³       |
| FSPF-SVR | 2.534 • 10⁻³    | 2.529 • 10⁻³ | 2.576 • 10⁻³       |

In order to further verify the difference between the full-variable model and the variable selection model and the prediction accuracy, we can adopt different experimental design schemes. The average daily PM₂₅ concentration in Beijing from January to October 2017 was selected as the training set, and the average daily PM₂₅ concentration in November was selected as the test set. The MSE of AV-SVR and FSPF-SVR models based on different kernel functions in the test set is shown in Table 3.

Table 3. The mean square error of AV-SVR and FSPF-SVR models based on different kernel functions in November test set (Unit: mg²/m⁶).

| Model Name     | MSE       | Model Name     | MSE       |
|----------------|-----------|----------------|-----------|
| AV-SVRPoly     | 2.549 • 10⁻³ | FSPF-SVRPoly  | 2.140 • 10⁻³ |
| AV-SVRRad      | 2.209 • 10⁻³ | FSPF-SVRRad   | 2.193 • 10⁻³ |
| AV-SVRSig      | 1.537 • 10⁻³ | FSPF-SVRSig   | 1.404 • 10⁻³ |

|          | PM₁₀+CO+PM₂₅,₁ | PM₁₀+CO+NO₂ | PM₁₀+CO+NO₂+PM₂₅,₁ |
|----------|-----------------|-------------|--------------------|
| AV-SVR   | 2.140 • 10⁻³    | 1.757 • 10⁻³ | 2.236 • 10⁻³       |
| FSPF-SVR | 2.193 • 10⁻³    | 1.833 • 10⁻³ | 2.214 • 10⁻³       |

Figure 5. The real average daily concentration of PM₂₅ in November 2017 was compared with the predicted value based on FSPF-SVRSig (with three input variables PM₁₀, CO and NO₂) model.

It can be clearly seen from the experimental results in Table 3 that the prediction accuracy of the SVR model after variable selection is higher than that of the full-variable SVR model. The model with the best prediction effect is FSPF-SVRSig with three input variables PM₁₀, CO and NO₂, and the corresponding MSE is 1.307 • 10⁻³ (mg²/m⁶). The relatively optimal accuracy of this model is also due to the fact that PM₁₀, CO and NO₂ are the three most important variables among all the predictive variables. A comparison between the real and predicted average PM₂₅ concentration in November is given, as shown in Figure 5.

As can be seen from Figure 5, the FSPF-SVRSig model with three input variables can well predict the average daily concentration of PM₂₅ in most cases except for a few days, especially in the middle
of November, from 6 to 16 and from 22 to 30. In other words, the predicted value is close to the real value. It can be seen that sometimes it is not necessary to select all variables into the prediction model, but only to select a few more important prediction variables into the model, a relatively high precision prediction model can be obtained.

In order to further compare and verify the prediction effect of SVR model on average daily PM$_{2.5}$ concentration in many machine learning models, the mainstream machine learning algorithms at present, namely Random Forest (RF), Weighted k-nearest Neighbor (KKNN), Bagging regression, Artificial Neural Network (ANN) and Multiple Adaptive Spline Regression (MARS) and traditional mathematical statistical model Multiple Linear Regression (MLR), were selected to predict the average daily PM$_{2.5}$ concentration. The data set used was the average daily PM$_{2.5}$ concentration data of Beijing from January to November, 2017, the first 10 months were used as the training set, and the last November was used as the test set. Based on each "machine learning algorithm", the prediction mean square error (MSE) of the full variable and FSPF variable selection was obtained as shown in Table 4.

Table 4. The mean square error of PM$_{2.5}$ average daily concentration prediction of Beijing in November based on different machine learning algorithms (Unit: mg$^2$/m$^6$).

| Algorithm name | All variable MSE | FSPF variable selection MSE PM$_{10+CO+NO_2}$ |
|----------------|-----------------|-----------------------------------------------|
| SVRSig         | 1.537 $\cdot$ 10$^{-3}$ | 1.307 $\cdot$ 10$^{-3}$ |
| RF             | 2.229 $\cdot$ 10$^{-3}$ | 2.058 $\cdot$ 10$^{-3}$ |
| KKNN           | 2.893 $\cdot$ 10$^{-3}$ | 4.199 $\cdot$ 10$^{-3}$ |
| Bagging        | 1.988 $\cdot$ 10$^{-3}$ | 1.787 $\cdot$ 10$^{-3}$ |
| MARS           | 2.440 $\cdot$ 10$^{-3}$ | 2.130 $\cdot$ 10$^{-3}$ |
| ANN            | 1.714 $\cdot$ 10$^{-3}$ | 1.653 $\cdot$ 10$^{-3}$ |
| MLR            | 2.023 $\cdot$ 10$^{-3}$ | 1.795 $\cdot$ 10$^{-3}$ |

When a variety of machine learning algorithms were used to predict the average daily PM$_{2.5}$ concentration in Beijing in November 2017, the SVRSig model still had the highest prediction accuracy. The main reason is that the SVR model is supported by a more complete mathematical optimization theory. Generally speaking, it performs better than other algorithms in the regression prediction of continuous variables. On the other hand, SVR can be modeled based on different kernel functions. The existence of kernel functions makes SVR better able to deal with nonlinear problems, which is obviously better than linear regression model. It is obvious from Table 4 that after FSPF variable selection, except KKNN algorithm, the prediction accuracy of other model algorithms is better than that of the full-variable model. Through the comparative study of various model algorithms, it is further verified that it is necessary to select the importance of variables in advance in the prediction of PM$_{2.5}$ average daily concentration.

5. Conclusions

This paper proposes a hybrid prediction model based on FSPF-SVR for the prediction of PM$_{2.5}$ daily concentration in Beijing. When constructing the model, air pollutant data and meteorological parameters related to PM$_{2.5}$ concentration were selected as the prediction variables, and the daily PM$_{2.5}$ concentration of the next day was predicted based on various data indexes of the previous day. However, in the process of establishing the model, all relevant predictive variables are not selected into the model blindly.

FSPF variable selection algorithm was used to measure the importance of the prediction variables that affect the average daily PM$_{2.5}$ concentration, and analyze which variables play a greater role in the prediction of PM$_{2.5}$ concentration. The important variables were selected into the SVR model, and the non-important variables were removed, so as to simplify the model and improve the prediction accuracy. Through experimental analysis, it can be concluded that the prediction accuracy of the FSPF-SVRsig model after variable selection is the best among all algorithm models when the average daily PM$_{2.5}$ concentration in Beijing is predicted in November and December 2017.
It can be concluded that the model after variable selection is better than the full variable model in many cases. In the future work of PM$_{2.5}$ monitoring and prediction, model variables can be simplified to improve the calculation efficiency, so as to obtain more accurate prediction results. The FSPF-SVR hybrid prediction model proposed in this paper is also applicable to the prediction of PM$_{2.5}$ in other cities, and can be widely applied to other fields such as medical diagnosis, financial risk assessment and rainfall prediction.

Acknowledgements
Natural Science Foundation of Guizhou Province Education Department (NO. Qian Jiao He KYZl[2019]200); Education Quality Improvement Project of Qiannan Normal University for Nationalities(NO. 2019xjg0201)

References
[1] Khaniabadi Y O, Goudarzi G, Daryanoosh S M, et al 2016 Exposure to PM$_{10}$,NO$_2$, and O$_3$ and impacts on human health[J] Environ Sci Pollut Res Int 1-9
[2] Megido L, Suárez-Peña B, Negral L, et al. 2017 Suburban air quality: human health hazard assessment of potentially toxic elements in PM$_{10}$[J] Chemosphere 177 284-291
[3] Sun W, Sun J 2017 Daily PM2.5 Concentration Prediction Based on Principal Component Analysis and LSSVM Optimized by Cuckoo Search Algorithm[J] Journal of Environmental Management 188
[4] Peng Y, Jie L 2016 Study on spatial and temporal variation of air pollutants and optimization algorithm of artificial intelligence: a case study of Beijing [M] Beijing: Science press
[5] Li Y, Chen Q, Zhao H, et al. 2015 Variations in PM$_{10}$, PM$_{2.5}$ and PM$_{10}$ in an urban area of the Sichuan basin and their relation to meteorological factors [J] Atmosphere 6(1) 150-163
[6] Huang R, Chun L 2017 Seasonal variation characteristics and forecasting model of PM2.5 in Changsha, central city in China[J] J Environ Anal Toxicol 7(1) 429-435
[7] Hui Y, Jing Y et al 2017 PM$_{2.5}$ hourly concentration tracking prediction model Based on ARMAX [J] Journal of Tianjin University 1
[8] AOIFE D, BRUCE M, BRIAN B 2015 Real time air quality forecasting using integrated parametric and non-parametric regression techniques[J] Atmospheric Environment 103(103) 53-65
[9] PAN L, SUN B, WANG W 2011 City air quality forecasting and impact factors analysis based on grey model [J] Procedia Engineering 12 74-79
[10] Yuxiao P, Qihong L, Xiaofei C 2016 PM$_{2.5}$ prediction model based on wavelet and process neural network [J] Environmental Science and Technology 29(2) 43-46
[11] Wenli Z, Lisha X, et al 2019 Study on PM$_{2.5}$ concentration prediction model based on weighted KNN-BP Neural Network [J] Journal of Environmental Engineering and Technology 9(01) 17-21
[12] Jiangeng L, Aorong Luo 2018 Prediction of PM$_{2.5}$ mass concentration based on empirical mode decomposition of complementary sets and support vector regression [J] Journal of Beijing University of Technology 44(12) 28-36
[13] Xiaohua L 2019 Construction and application of multivariate analysis of variance model [J] Statistics and Decision 1 75-78
[14] Nilson R, Peña, José M, Bjökegren, Johan, et al 2007 Consistent Feature Selection for Pattern Recognition in Polynomial Time[J] Journal of Machine Learning Research 8(8) 589-612
[15] Yuanhua J, Xiaojie Y, Quanfeng J, et al 2018 Analysis on PM$_{2.5}$ emission factors and main components of crop straw under different combustion conditions [J] Journal of Environmental Science 1 92-100
[16] Dong L, Huifeng X, et al 2019 Prediction of PM$_{2.5}$ daily concentration based on hybrid model [J] Statistics and decision making 5 92-96