Prediction of PM_{2.5} concentration based on multi-source data and self-organizing fuzzy neural network

Junfei Qiao^{1,2} · Zengzeng He^{1,2} · Shengli Du^{1,2}

Received: 25 December 2019 / Accepted: 29 February 2020 / Published online: 11 March 2020
© Springer Nature Switzerland AG 2020

Abstract

The fine particulate matter (PM_{2.5}) problem has become one of the biggest topics of concern in many cities, such as Beijing in China. This study attempts to employ a self-organizing fuzzy neural network (SOFNN) to predict the hourly concentration of PM_{2.5} using multi-source data. The multi-source data including meteorological data, pollutant concentration data and image data are first obtained through different methods. Then, a novel SOFNN, which is optimized by an improved gradient descent algorithm, is proposed based on the mutual information (MI) and sensitivity analysis (SA). More importantly, MI and SA are introduced to effectively adjust the structure of fuzzy neural network without changing the performance of the original network. Finally, the MI- and SA-based SOFNN is deployed to infer PM_{2.5} concentration at the study area. The experimental results show that our proposed model can achieve a more compact network structure, higher prediction performance and lower computation time as compared with state-of-the-art or popular methods.

Keywords PM_{2.5} · Self-organizing fuzzy neural network (SOFNN) · Multi-source data · Mutual information (MI) · Sensitivity analysis (SA) · Relevance analysis

1 Introduction

In recent years, the rapid urbanization and industrialization in many cities in China cause the deterioration of air quality owing to the improper treatment of large amounts of household waste, vehicles emissions, industrial emissions, etc. Moreover, the serious air pollution has a considerable impact on urban residents, especially sensitive people such as children and people with heart and respiratory insufficiency [23]. A large number of people suffer from various diseases or even die every year throughout the world due to the polluted air. At the same time, poor air quality will affect people's travel and production activities [32]. Besides, the heavy pollution will not only affect the image of a city, but also cause unnecessary economic losses. With the occurrence of heavy air pollution

incidents, regional atmospheric environmental problems have gradually attracted widespread attention [38]. Therefore, the improvement in air quality is of great significance to individuals and society [33]. Among all of the typical air pollutants, PM_{2.5} with the aerodynamic diameters of 2.5 μm or less has attracted a lot of attentions. Furthermore, PM_{2.5} is able to easily transmit hazardous chemicals into the human body and thus causes serious diseases by damaging cerebrovascular, cardiovascular and respiratory functions [12]. Hence, the majority of residents gradually realize the significance of predicting PM_{2.5} concentration and taking prophylactic measures before the arrival of severe smoggy conditions. To validly deal with the PM_{2.5} problem, proposing an effective method to predict hourly concentration of PM_{2.5} is desperately needed and would be beneficial to both individuals and government.

✉ Junfei Qiao, junfeiq@bjut.edu.cn; Zengzeng He, zengzenghe@gmail.com; Shengli Du, shenglidu@bjut.edu.cn | ¹Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China. ²Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China.

The modeling and prediction of $PM_{2.5}$ concentration have proven to be difficult due to the chaotic and nonlinear characters [34, 39]. Generally speaking, the methods of $PM_{2.5}$ prediction mainly consist of mechanism analysis method and statistical method [25]. The mechanism analysis method aims to analyze the causes of $PM_{2.5}$ formation and describe its diffusion process based on the complex physical and chemical changes. Saide et al. [36] employed the chemical transport model to predict PM concentration according to the linear relationship between CO and PM concentration, but the accuracy of the inventory list required for the model has a critical impact on the predicted results. Riccio et al. [35] used the Euler and Lagrangian model to represent the $PM_{2.5}$ concentration changes at a site in southern Italy. The results showed that the local $PM_{2.5}$ concentration was not only caused by local sources, but was also affected by the transportation of the surrounding areas. However, the dynamic process of the atmospheric system is so nonlinear that its theoretical analysis is complex. In addition, several parameters, such as meteorological boundary conditions and emission source lists required for mechanism analysis modeling, are difficult to obtain. Thus, the mechanism analysis method is not suitable for predicting the hourly $PM_{2.5}$ concentration.

Compared to the mechanism analysis method, it is easier to establish a statistical prediction model using the data-driven method. Chen et al. [3] predicted the PM_{10} concentration using the combination of wavelet decomposition and stepwise regression and achieved good prediction results. Elbayoumi et al. [7] adopted the concentrations of PM_{10} , $PM_{2.5}$, CO, CO_2 , wind speed, pressure and relative humidity as the inputs of the multiple linear regression (MLR) to predict PM_{10} and $PM_{2.5}$ concentrations. However, the variation of $PM_{2.5}$ concentration is a nonlinear process, and the simple linear model above cannot capture the nonlinear relationships among different variables due to the complexity and diversity of sample data.

On the other hand, artificial neural networks (ANNs) are widely used in the modeling of nonlinear systems due to their powerful nonlinear mapping ability and self-learning ability [19]. Ordieres et al. [27] used multilayer perceptron and radial basis function (RBF) neural network to predict $PM_{2.5}$ concentration and found that the prediction results of RBF neural network are more accurate. Xu et al. [41] used the echo state network (ESN) to predict the daily average $PM_{2.5}$ concentration in Shanghai and achieved better prediction results than the RBF neural network. However, the prediction accuracy of ESN still needs to be further improved.

The underlying causal mechanism including the emission and dispersion pattern of air pollutants is difficult to be ascertained by neural networks because the mapping rules in neural networks are invisible and not easy to

understand. And the fuzzy system usually handles issues such as reasoning on a high level and deals with imprecise information through linguistic expressions [1, 24]. However, the fuzzy rules are difficult to be tuned from the training data through manual operation due to the limited learning capability of fuzzy systems. Fortunately, the FNN outperforms single neural networks or fuzzy systems in prediction tasks because it combines the learning ability of neural network and the interpretability of fuzzy rule-based fuzzy system together [30, 46]. Mishra et al. [25] employed the Takagi–Sugeno FNN to predict the current $PM_{2.5}$ concentration by CO, NO_2 , O_3 , SO_2 , $PM_{2.5}$, temperature, wind speed, relative humidity and dew point temperature, and the experimental results showed that FNNs have better prediction performance than MLR and ANNs.

The parameter learning and structure adjustment are two key parts of training FNN [22]. The learning algorithms of parameters usually adopt supervised or unsupervised methods. However, compared with training methods of the parameters, it is difficult for designers to adaptively adjust the network structure to effectively solve prediction problems. Many existing results often forecast hourly $PM_{2.5}$ concentration using FNNs with an empirically fixed structure. In order to address this problem, Qiao et al. [32] used the SOFNN to establish a prediction model of hourly $PM_{2.5}$ concentration and the second-order gradient algorithm was utilized to train the network parameters. The results of experiments demonstrated that a SOFNN was more advantageous than FNNs with fixed structure in terms of $PM_{2.5}$ prediction. To the best of our knowledge, although the above method can adaptively adjust the network structure, the prediction performance of the network is deteriorated due to structural adjustment. Besides, the adjusted network architecture is not condensed because the associations of information among the rule neurons are neglected, which may result in neuron redundancy.

The MI and SA are used to evaluate the amount of internal information of rule neurons and the contribution rate of each rule neuron, respectively, which makes network structure more condensed as compared with other methods. Therefore, the MI- and SA-based SOFNN (MISA-SOFNN) is used for forecasting hourly $PM_{2.5}$ concentration by means of multi-source data. To be more concrete, the multi-source data that consist of meteorological data, pollutant concentration data and image data are first collected from different environmental authorities. Furthermore, the former two pieces of data are used for feature selection through relevance analysis and the last one is employed to extract image features. Second, the parameters of SOFNN proposed in this paper are optimized via an improved gradient descent algorithm, which is more time-saving and efficient than the gradient descent algorithm and second-order gradient algorithm. Third, the

MISA-SOFNN is employed to infer $PM_{2.5}$ concentration in the next hour by using the selected meteorological and pollutant concentration variables and the extracted image feature variables at the current moment. As shown in Fig. 1, the work flow of predicting $PM_{2.5}$ concentration mainly consists of the following steps: data collection, experimental data acquisition, model training and model examining.

In this paper, the major contributions are summarized as follows: (1) Compared to other methods, MI and SA are used to adjust the structure of FNN, which is beneficial to obtaining a prediction model with concise network structure. (2) A novel MISA-SOFNN is first designed to accurately predict $PM_{2.5}$ concentration. The self-organizing method can ensure the network output unchanged by updating weights when the structure of MISA-SOFNN is adjusted. Besides, an improved gradient descent algorithm is used to accelerate the convergence speed of MISA-SOFNN. (3) The proposed method is more comprehensive than previous methods using only meteorological or pollutant concentration data. The relevance analysis method is used for feature selection from the quantitative and qualitative perspectives, which reduces

data dimensionality, simplifies prediction model and improves prediction accuracy.

The remainder of this paper is arranged as follows. In Sect. 2, the basic information of the study area and the acquisition of the multi-source data are described in detail. In Sect. 3, several basic concepts, including feature extraction, relevance analysis, fuzzy neural network, the improved gradient descent algorithm, mutual information and sensitivity analysis, are introduced briefly. In Sect. 4, the self-organizing method is designed based on MI and SA and further used to establish a MISA-SOFNN. In Sect. 5, the meteorological and pollutant concentration data are first used for feature selection through relevance analysis and three types of contrast-sensitive features are extracted from the image data. Next, the MISA-SOFNN are used to infer $PM_{2.5}$ concentration in the next hour based on the selected meteorological and pollutant concentration variables and the extracted image feature variables at the current moment. Finally, some simulation results and discussions are also given in this section. Some concluding remarks of the whole paper are provided in Sect. 6.

Fig. 1 The work flow of predicting $PM_{2.5}$ concentration

2 Study domain and materials

2.1 Study domain

Located in the North China Plain, Beijing is surrounded by several highly industrialized cities, such as Tangshan and Tianjin. Beijing covers an area of 16,808 square kilometers and has a population of nearly 21.7 million. $PM_{2.5}$ is currently an extremely serious problem in highly populated city clusters such as the Beijing–Tianjin–Hebei region [44]. In addition, strong local emission sources and long-range transport from neighboring cities are major contributors to the $PM_{2.5}$ concentration in Beijing [17]. In this study, we focus on exploring the $PM_{2.5}$ effects on the Beijing, and the relevant experimental data of study area are readily available.

2.2 Data collection

Three types of data collected from the last 3 years are used in this study. The meteorological data and pollutant concentration data are composed of six different variables, respectively. Image data contain a lot of different photographs that are used for feature extraction. The meteorological, pollutant concentration and image data are recorded hourly, providing a total of 1200 simultaneous records.

2.2.1 Meteorological data

The hourly meteorological data are obtained from the Beijing meteorological service climate center. The meteorological data contain visibility, weather, temperature, pressure, humidity and wind speed. In order to forecast the hourly concentration of PM_{2.5} at Beijing, only meteorological data from Beijing are used in this paper.

2.2.2 Pollutant concentration data

The hourly pollutant concentration data, including PM₁₀, PM_{2.5}, NO₂, SO₂, O₃, CO, are provided by the Beijing environmental protection monitoring center. There are 12 monitoring stations of air quality in the urban area of Beijing, such as the Chaoyang Agricultural Exhibition Hall station (116.461 E, 39.937 N), Dongcheng Dongsu station (116.417 E, 39.929 N), Xicheng Wanshou Xigong station (116.352 E, 39.878 N), Haidian Wanliu station (116.287 E, 39.987 N), Fengtai Yungang station (116.146 E, 39.824 N) and Shijingshan Ancient City station (116.184 E, 39.914 N). To predict hourly PM_{2.5} concentration at Beijing, the hourly pollutant concentration data are collected from these stations. Table 1 provides the basic information about meteorological data and pollutant concentration data.

2.2.3 Image data

The photographs of Beijing are acquired from the Beijing real-time weather photographs. Their shooting interval time is at least half an hour. Those pictures taken at different locations and under various weather situations involve diversified scenes, such as square, roads, buildings, cars and parks. All photographs are of a wide range of image resolutions from 640 × 480 to 1024 × 768.

3 Preliminaries

In this section, we briefly introduce the concepts of feature extraction, relevance analysis, fuzzy neural network, improved gradient descent algorithm, mutual information and sensitivity analysis.

3.1 Feature extraction

Since the image data cannot be directly used as inputs of the prediction model, we need to extract image features from the collected images and employ these features to predict hourly PM_{2.5} concentration. Image contrast can reflect the changes in PM_{2.5} concentration [20]. Information entropy is a common and classic measurement of image contrast. The mean amount of information included in the image is represented by using information entropy. The information entropy based on phase congruence [11] is defined as

$$E_{pc} = - \sum_{i=0}^{255} P_i(s_{pc}) \cdot \log P_i(s_{pc}), \quad (1)$$

where s_{pc} comprises pixels of s and is equal to 40% maximum of the detected PC map. P is the probability value. Contrast energy is another measurement of image contrast and is used to estimate the local contrast of perceptual image [9]. Because it has the properties of simple calculation and contrast perception, contrast energy is adopted to measure image contrast [37]. Based on the existing works [4], contrast energy can be obtained as follows:

$$CE_f = \frac{\alpha \cdot Y(s_f)}{Y(s_f) + \alpha \cdot \eta} - \phi_{fr}, \quad (2)$$

Table 1 Information about the meteorological and pollutant concentration data

Data type	Variable	Unit	Interval	Data source
Meteorological data	Visibility	km	Hourly	Beijing meteorological service climate center
	Weather	–	Hourly	Beijing meteorological service climate center
	Temperature	°C	Hourly	Beijing meteorological service climate center
	Pressure	hPa	Hourly	Beijing meteorological service climate center
	Humidity	%	Hourly	Beijing meteorological service climate center
	Wind speed	m/s	Hourly	Beijing meteorological Service climate center
Pollutant concentration data	PM ₁₀	μg/m ³	Hourly	Beijing environmental protection monitoring center
	NO ₂	μg/m ³	Hourly	Beijing environmental protection monitoring center
	SO ₂	μg/m ³	Hourly	Beijing environmental protection monitoring center
	O ₃	μg/m ³	Hourly	Beijing environmental protection monitoring center
	CO	mg/m ³	Hourly	Beijing environmental protection monitoring center
	PM _{2.5}	μg/m ³	Hourly	Beijing environmental protection monitoring center

where $Y(s_f) = ((s_k \times f_h)^2 + (s_k \times f_v)^2)^{\frac{1}{2}}$, s has three channels, namely $y_b = 0.5 \times (G + R) - B$, $g_r = 0.299 \times R + 0.587 \times G + 0.114 \times B$ and $rg = R - G$ [18]. Besides, η is the contrast gain, $\alpha = \max[Y(s_f)]$, ϕ_f is used for reducing the noise, and f_v and f_h indicate vertical and horizontal second derivatives of Gaussian function, respectively. The features related to image contrast are summarized as $F_{ct} = \{E_{pc}, CE_{gr}, CE_{yb}, CE_{rg}\}$.

The lower the image brightness is, the higher the $PM_{2.5}$ concentration is [5]. Image brightness, an important factor associated with $PM_{2.5}$ concentration, dominates the effect of image enhancement. To be more specific, the appropriate image brightness may broaden the dynamic change of the image and contain more semantic information. Based on the existing works [10], the features related to image brightness are summarized as

$$F_b = \{E_{n1}, E_{n2}, E_{n3}, E_{n4}, E_{n5}, E_{n6}\}, \quad (3)$$

where $\{E_{n1}, E_{n2}, \dots, E_{n6}\}$ are six values of entropy and evaluated by $n = \{m, (1/m) | m = 3.5, 5.5, 7.5\}$.

As the $PM_{2.5}$ concentration elevates, the image colorfulness will decrease [12]. Images with high image colorfulness usually show more details and relative information. The global mean of the saturation channel can be given as follows:

$$S = \frac{1}{M} \sum_{i=1}^M T_{X \rightarrow S}[s(i)], \quad (4)$$

where $T_{X \rightarrow S}$ is a transformation function and M is the pixel number of s . Another measurement of image colorfulness is defined in Hasler and Suesstrunk [18]. Similar to Hasler and Suesstrunk [18], the four important features consist of the mean and variance of y_b and rg channel, respectively, such as $\mu_{yb}, \sigma_{yb}^2, \mu_{rg}, \sigma_{rg}^2$. Then, the measurement is defined as

$$C = \sqrt{\sigma_{yb}^2 + \sigma_{rg}^2} + k \sqrt{\mu_{yb}^2 + \mu_{rg}^2}, \quad (5)$$

where k is set to 0.3 to correct the relative importance. Therefore, the features related to image colorfulness are summarized as $F_{cl} = \{S, C\}$.

3.2 Relevance analysis

There are many ways to select input variables of prediction models [40]. However, the feature selection process of these methods is not intuitive, and the results obtained by these methods are not interpretable. In order to solve the shortcomings of these methods, the relevance analysis method is introduced to select the dominant variables related to $PM_{2.5}$ concentration from the quantitative and

qualitative perspectives. More specifically, correlation coefficients, fitting formulas and scatter diagram are used to quantitatively and qualitatively evaluate the correlation between sample variables and $PM_{2.5}$ concentration.

The Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) are used to describe the correlation strength, and they take values between -1 and 1 [26]. The correlation is stronger when the value of PCC or SCC is closer to 1 or -1 . Moreover, the two correlation coefficients represent different relationships, that is, the PCC and SCC are usually used to describe the linear correlation and nonlinear correlation, respectively.

Assume A and B are two data sequences which have the same amount of data. The PCC is used to describe the linear correlation and is represented as follows:

$$PCC = \frac{\sum_{i=1}^N (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^N (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^N (B_i - \bar{B})^2}}, \quad (6)$$

where A_i and B_i are the i th data of sequence A and sequence B , respectively; $\bar{A}$ and $\bar{B}$ stand for the mean value of sequence A and sequence B , respectively; and N is the number of data points in a sequence.

The SCC is used to describe the nonlinear correlation and is defined as

$$SCC = 1 - \frac{6 \sum_{i=1}^n (m_i - n_i)}{N(N^2 - 1)}, \quad (7)$$

where N is the number of data points in a sequence. m_i and n_i are the order numbers of sequence A and sequence B , respectively.

3.3 Fuzzy neural network

The structure of FNN consists of input layer, membership function layer, rule layer, defuzzy layer and output layer and is shown in Fig. 2 [32]. The FNN model expression is described as follows:

1. The input layer: This layer has n nodes, and each represents an input variable of FNN. Furthermore, the output values of the input layer are formulated as

$$r_i = x_i \quad i = 1, 2, 3, \dots, n, \quad (8)$$

where x_i is the i th input and r_i represents the i th output.

2. The membership function layer: This layer aims to turn the input variables into fuzzy variables. Here, all neurons are membership functions, and the Gaussian function is selected as the membership function

Fig. 2 The network structure of FNN

$$a_{ij} = \exp\left[-\frac{(r_i - c_{ij})^2}{2\sigma_{ij}^2}\right] \quad j = 1, 2, 3, \dots, m, \quad (9)$$

where a_{ij} means the output of the j th membership function related to the i th input, σ_{ij} and c_{ij} are the width and center of the j th membership function related to the i th input, separately.

3. The rule layer: All neurons in this layer are fuzzy logic rules:

$$\phi_j = \prod_{i=1}^n a_{ij}, \quad (10)$$

where ϕ_j represents the output of the j th node of the rule layer.

4. Defuzzy layer: This layer has m nodes, which is the same as the number of rule layers. The outputs of the defuzzy layer are calculated as

$$\bar{\phi}_j = \phi_j / \sum_{j=1}^m \phi_j. \quad (11)$$

5. The output layer: The role of this layer is to implement a linear weighted summation of the outputs of the defuzzy layer:

$$y_o = \sum_{j=1}^m w_j \bar{\phi}_j, \quad (12)$$

where y_o is the output of the output layer and w_j represents the connection weight between the defuzzy layer and the output layer.

3.4 Improved gradient descent algorithm

Generally speaking, the convergence speed of gradient descent algorithm will be slow when it is utilized to train network parameters [15]. Therefore, the gradient descent

algorithm should be improved to achieve fast convergence and then deployed to adaptively adjust the parameters of the FNN. To better explain the improved gradient descent method, the loss function is first defined as

$$J = \frac{1}{2}(y_o - y_r)^2, \quad (13)$$

where y_r represents the real value and y_o means the output of the neural network.

Second, the update formulas for network parameters are given as

$$w'_j = w_j - \theta \frac{\partial J}{\partial w_j}, \quad (14)$$

$$c'_{ij} = c_{ij} - \theta \frac{\partial J}{\partial c_{ij}}, \quad (15)$$

$$\sigma'_{ij} = \sigma_{ij} - \theta \frac{\partial J}{\partial \sigma_{ij}}, \quad (16)$$

where w'_j , c'_{ij} and σ'_{ij} are updated parameters and θ stands for an adaptive learning rate.

Third, the adaptive learning rate proposed in this paper aims to accelerate the convergence speed of FNN. More specifically, in the early stage of training, the adaptive learning rate is so large that it is beneficial for the rapid convergence of FNN. As the number of training increases, the adaptive learning rate can be gradually reduced to avoid instability of neural networks. The adaptive learning rate is defined as follows:

$$\theta = \theta_{\max} - \frac{D}{D_{\max}}(\theta_{\max} - \theta_{\min}), \quad (17)$$

where $\theta_{\min}$ and $\theta_{\max}$ indicate the minimum and maximum learning rate, respectively; D represents the number of current training steps; $D_{\max}$ means the total number of training steps.

3.5 Mutual information

MI is commonly used to reflect the statistical dependency between two random variables, and the correlation of two neurons can also be measured by MI [28, 29, 31]. Therefore, the MI method is introduced to assess the associations of information among the rule neurons in this paper, which makes network structure more compact. The calculation of probability distribution function is significant for evaluating MI between two rule neurons. At present, there are several frequently used methods to evaluate MI, such as copula entropy method, kernel method and k -nearest neighbor method [16]. Due to the computational

simplicity and efficiency of histogram method in [13], this paper uses this method to measure MI between two rule neurons.

It is assumed that variable X and variable Y represent the outputs of two rule neurons, respectively. Firstly, we measure the entropy of discrete variable X . The histogram method divides the range of X into K_X segments, and the size of each segment is Δx . For reducing the deviation of entropy estimation, the number of segments can be defined as follows [13]:

$$K_X = \text{round} \left\{ \frac{\gamma}{6} + \frac{2}{3\gamma} + \frac{1}{3} \right\}, \quad (18)$$

$$\gamma = \sqrt[3]{8 + 324N + 12\sqrt{36N + 729N^2}}, \quad (19)$$

where N means the total number of samples. Then, the entropy of a discrete variable X can be defined as

$$H(X) = - \sum_{k=1}^{K_X} p_k(x) \log p_k(x), \quad (20)$$

where $p_k(x)$ presents the probability that a sample drops into the k th segment. It can be given as:

$$p_k(x) = \frac{n_k}{N}, \quad (21)$$

where n_k denotes the number of samples found in segment k and N denotes the total number of samples. Then, the entropy is expressed as

$$H(X) = - \sum_{k=1}^{K_X} \frac{n_k}{N} \log \frac{n_k}{N}. \quad (22)$$

Similar to the calculation process of entropy of variable X , the entropy of variable Y is expressed as

$$H(Y) = - \sum_{k=1}^{K_Y} \frac{n_k}{N} \log \frac{n_k}{N}. \quad (23)$$

Secondly, we evaluate the joint entropy of the bivariate random variable (X, Y) . The histogram method divides the xy -plane into $(K_X \times K_Y)$ cells with coordinate (i, j) , and the size of each cell is $\Delta x \times \Delta y$. The joint entropy of bivariate random variable (X, Y) can be calculated as

$$H(X, Y) = - \sum_{k_i=1}^{K_X} \sum_{k_j=1}^{K_Y} p_{k_i k_j}(x, y) \log p_{k_i k_j}(x, y), \quad (24)$$

where $p_{k_i k_j}(x, y)$ stands for the probability that a sample drops into the cell (k_i, k_j) . It can be obtained as follows:

$$p_{k_i k_j}(x, y) = \frac{n_{k_i k_j}}{N}, \quad (25)$$

where $n_{k_i k_j}$ means the number of samples found in cell (k_i, k_j) and N presents the total number of samples. Then, the joint entropy is calculated as

$$H(X, Y) = - \sum_{k_i=1}^{K_X} \sum_{k_j=1}^{K_Y} \frac{n_{k_i k_j}}{N} \log \frac{n_{k_i k_j}}{N}. \quad (26)$$

Finally, according to mathematical relationship among three entropies, MI between rule neuron and output neuron can be calculated by combining Eqs. (22), (23) and (26) together:

$$I(X; Y) = - \sum_{k_i=1}^{K_X} \frac{n_{k_i}}{N} \log \frac{n_{k_i}}{N} - \sum_{k_j=1}^{K_Y} \frac{n_{k_j}}{N} \log \frac{n_{k_j}}{N} + \sum_{k_i=1}^{K_X} \sum_{k_j=1}^{K_Y} \frac{n_{k_i k_j}}{N} \log \frac{n_{k_i k_j}}{N}. \quad (27)$$

3.6 Sensitivity analysis

SA is used to estimate the dependency of a model output to the input variables [43], which can be used to measure the contribution of each input variable to the output variance. SA adopts a sensitivity index to estimate the significance of the given inputs, and it is expressed as the fractional contribution to the output variance because of the uncertainties in inputs. Hence, we determine the contribution ratio of a rule neuron to the information transfer of the FNN output via an SA method. The main steps of the SA approach are given below [14]:

1. Defining the model and its input and output variables;
2. Assigning probability density functions for each input variable;
3. Establishing an input matrix according to sampling design;
4. Calculating the output of a given model;
5. Estimating the influence level of input variables on the output variables.

Although the sensitivity of model output can be analyzed by several tactics, all of them depend on evaluating the sensitivity index. In this paper, the input variables of the SA model are given as $[w_{11}\phi_1, w_{21}\phi_2, \dots, w_{P1}\phi_P, w_{12}\phi_1, w_{22}\phi_2, \dots, w_{P2}\phi_P, \dots, w_{1Q}\phi_1, w_{2Q}\phi_2, \dots, w_{PQ}\phi_P]$ and its dimension is $P \times Q$. The P -variable method is defined as

$$Y^k = f(w_{1k}\overline{\phi_1}, w_{2k}\overline{\phi_2}, \dots, w_{pk}\overline{\phi_p}), \quad (28)$$

where Y^k denotes the k th model output; $w_{1k}\overline{\phi_1}, w_{2k}\overline{\phi_2}, \dots, w_{pk}\overline{\phi_p}$ represent the values of input variables and the relationship between the k th output neuron and each rule neuron.

The common measurement formula of the sensitivity is defined as

$$S_j^k = \frac{\text{Var}_{w_{jk}\overline{\phi_j}}[Y^k | w_{jk}\overline{\phi_j} = \ell_j]}{\text{Var}(Y^k)}, \quad (29)$$

where S_j^k is the contribution ratio of the j th input variable to the k th network output. $w_{jk}\overline{\phi_j}$ represents the j th input variable of the k th model, $j = 1, 2, \dots, P$, $k = 1, 2, \dots, Q$. $Y^k | w_{jk}\overline{\phi_j} = \ell_j$ stands for the k th response value produced by the j th rule neuron. $\text{Var}(Y^k)$ denotes the variance of output variable of k th model. In order to enhance the readability of the formula, Eq. (29) can be further improved as follows:

$$S_j^k = \frac{\text{Var}_{w_{jk}\overline{\phi_j}}[Y^k | w_{jk}\overline{\phi_j} = \ell_j]}{\sum_{i=1}^P \text{Var}_{w_{ik}\overline{\phi_i}}[Y^k | w_{ik}\overline{\phi_i} = \ell_i]}. \quad (30)$$

Therefore, we can measure the effect of input variable on output variable using the above formula. It is evident that the contribution ratio of rule neuron will be bigger when it is more active.

4 Establishing the SOFNN based on MI and SA

This section first depicts the design process of the self-organizing method, which includes neuron combining, neuron pruning, neuron growing and threshold setting. Then, the MI- and SA-based SOFNN, namely MISA-SOFNN, is established based on the designed self-organizing method and its calculation procedure is described in detail.

4.1 Design of self-organizing method

In this paper, the following two aims will be achieved: (1) the correlation between each two rule neurons is small and all the rule neurons are informative; (2) the performance of the original network is unchanged after the structure of FNN is adjusted. Therefore, on the one hand, rule neurons are combined based on MI; on the other

hand, rule neurons with high sensitivity are split and rule neurons with low sensitivity are deleted. At the same time, the corresponding weights should be adjusted to maintain the performance of original network. The specific steps of self-organizing method are given as follows.

4.1.1 Neuron combining

The MI method is used to measure the correlation between rule neurons in this paper. The MI between the l th and the k th rule neuron, namely $I(l, k)$, can be computed by Eq. (27). The two rule neurons will be merged if the MI is larger than λ_1 (defined in Sect. 4.1.4). Assume that the l th rule neuron and the k th rule are merged to be the h th rule neuron. To ensure that the performance of the original network is unaffected by the combination of rule neurons, the parameters of network should be adjusted as follows:

$$\begin{cases} c_{ih} = c_{il} + c_{ik}, \sigma_{ih} = \sigma_{il} + \sigma_{ik}, \\ w_h = nw_l + mw_k, \end{cases} \quad (31)$$

where c_{ih} , c_{il} and c_{ik} are the centers of the h th, the l th and the k th membership function related to the i th input, respectively; σ_{ih} , σ_{il} and σ_{ik} are the widths of the h th, the l th and the k th membership function related to the i th input, separately; w_h , w_l and w_k represent the connection weights of the h th, the l th and the k th neurons in defuzzy layer related to the output layer, separately; m and n are calculated by Eqs. (35) and (36) with $m + n = 1$.

Let $[x_t, y_t]$ ($t = 1, 2, \dots, N$) denote the training samples with $x_t = (x_{t1}, x_{t2}, \dots, x_{tn})$. Based on Eqs. (8)–(12), the network outputs produced by the three rule neurons are

$$\begin{cases} y_l(x_{ti}) = \frac{w_l \exp(-\sum_{j=1}^n ((x_{ti} - c_{lj})^2 / 2\sigma_{lj}^2))}{\sum_{j=1}^m \exp(-\sum_{j=1}^n ((x_{ti} - c_{lj})^2 / 2\sigma_{lj}^2))}, \\ y_k(x_{ti}) = \frac{w_k \exp(-\sum_{j=1}^n ((x_{ti} - c_{kj})^2 / 2\sigma_{kj}^2))}{\sum_{j=1}^m \exp(-\sum_{j=1}^n ((x_{ti} - c_{kj})^2 / 2\sigma_{kj}^2))}, \\ y_h(x_{ti}) = \frac{w_h \exp(-\sum_{j=1}^n ((x_{ti} - c_{hj})^2 / 2\sigma_{hj}^2))}{\sum_{j=1}^m \exp(-\sum_{j=1}^n ((x_{ti} - c_{hj})^2 / 2\sigma_{hj}^2))}, \end{cases} \quad (32)$$

where x_{ti} is the t th input sample; $y_l(x_{ti})$ is the output produced by the l th rule neuron at t th input sample; $y_k(x_{ti})$ is the output produced by the k th rule neuron at t th input sample; $y_h(x_{ti})$ is the output produced by the h th rule neuron at t th input sample. In order to ensure that the network output is unchanged, the equation can be established as follows:

$$y_h(x_{ti}) = y_l(x_{ti}) + y_k(x_{ti}). \quad (33)$$

According to Formulas (32) and (33), Eq. (34) can be calculated as

$$w_h \exp \left(- \sum_{i=1}^n ((x_{ti} - c_{ih})^2 / 2\sigma_{ih}^2) \right) = w_l \exp \left(- \sum_{i=1}^n ((x_{ti} - c_{il})^2 / 2\sigma_{il}^2) \right) + w_k \exp \left(- \sum_{i=1}^n ((x_{ti} - c_{ik})^2 / 2\sigma_{ik}^2) \right). \quad (34)$$

Considering the condition $m + n = 1$, the value of m and n can be obtained as follows:

$$m = \frac{w_l}{w_k - w_l} \cdot (A - 1) + \frac{w_k}{w_k - w_l} \cdot B, \quad (35)$$

$$n = \frac{w_k}{w_l - w_k} \cdot (B - 1) + \frac{w_l}{w_l - w_k} \cdot A, \quad (36)$$

where

$$A = \exp \left(- \sum_{i=1}^n ((x_{ti} - c_{il})^2 / 2\sigma_{il}^2) + \sum_{i=1}^n ((x_{ti} - c_{il} - c_{ik})^2 / 2(\sigma_{il} + \sigma_{ik})^2) \right)$$

$$B = \exp \left(- \sum_{i=1}^n ((x_{ti} - c_{ik})^2 / 2\sigma_{ik}^2) + \sum_{i=1}^n ((x_{ti} - c_{il} - c_{ik})^2 / 2(\sigma_{il} + \sigma_{ik})^2) \right).$$

Then, w_h in Eq. (31) can be calculated by m and n above.

4.1.2 Neuron pruning

Rule neurons with low contribution rate evaluated by SA are useless and should be deleted. The j th rule neuron is useless if its contribution ratio S_j computed by Eq. (30) is smaller than λ_2 (defined in Sect. 4.1.4). In addition, the useless rule neuron should be deleted. To guarantee the performance of the original network unchanged, the weights of the nearest p th rule neuron that has the maximum MI to the j th rule neuron and $S_p \geq \lambda_2$ can be calculated as

$$\begin{cases} c'_{ip} = c_{ip}, \sigma'_{ip} = \sigma_{ip}, \\ w'_p = w_p + \frac{\phi_j}{\phi_p} w_j, \end{cases} \quad (37)$$

where c'_{ip} and c_{ip} are the centers of the p th membership function related to the i th input after and before pruning, respectively; σ'_{ip} and σ_{ip} are the widths of the p th membership function related to the i th input after and before pruning, respectively; w'_p and w_p are the connection weights of the p th defuzzy neuron related to the output layer after and before pruning, respectively; w_j is the

weight between the j th defuzzy neuron and output neuron; ϕ_j and ϕ_p are the outputs of the j th and the p th defuzzy neuron, respectively.

4.1.3 Neuron growing

To avoid the drastic variation of the network output, one rule neuron with high sensitivity is split into two new rule neurons if the contribution rates of a few rule neurons are too high. One rule neuron should be split into two new rule neurons if its contribution ratio calculated by Eq. (30) is bigger than λ_3 (defined in Sect. 4.1.4). If the contribution ratio S_j of the j th rule neuron is larger than λ_3 , the j th rule neuron will split into a th and b th rule neurons. To ensure that the performance of the original network is unchanged, the weights of the a th and b th rule neurons are adjusted by [21]

$$\begin{cases} c_{ia} = c_{ij}, \sigma_{ia} = \sigma_{ij}, w_a = \alpha w_j, \\ c_{ib} = c_{ij}, \sigma_{ib} = \sigma_{ij}, w_b = (1 - \alpha) w_j, \end{cases} \quad (38)$$

where c_{ia} , c_{ib} and c_{ij} are the centers of the a th, b th and j th membership functions related to the i th input, respectively; σ_{ia} , σ_{ib} and σ_{ij} are the widths of the a th, b th and j th membership functions related to the i th input, respectively; w_a , w_b and w_j are the connection weights between defuzzy neuron and output neuron, respectively; α is the mutation parameter, and it can take either a random or fixed number ranged from 0 to 1 based on the distribution rule [21]. To make sure that the contribution rate of new rule neuron is not too high, α is set as a random number ranged from 0.2 to 0.8.

4.1.4 Threshold setting

This section will give the setting methods of the λ_1 , λ_2 and λ_3 in detail. Since the optimal thresholds are distinct for different training samples, the adaptive thresholds are set through existing experience. The MI between each two rule neurons is calculated by Eq. (27). The average of all MIs can be computed by

$$\overline{I_{ave}} = \frac{\sum_{i=1}^m \sum_{j=1}^{m-i} I(R_i : R_j)}{m(m-1)/2}, \quad (39)$$

where $I(R_i : R_j)$ means the MI between the i th rule neuron and the j th rule neuron. In addition, λ_1 can be expressed as

$$\lambda_1 = \beta_1 \overline{I_{ave}}, \quad (40)$$

where β_1 stands for a coefficient related to the training samples.

The contribution rate of each rule neuron is calculated by Eq. (30). The average of all the contribution rates can be computed by

$$\overline{S_{ave}} = \sum_{i=1}^m S_i / m, \quad (41)$$

where S_i indicates the contribution rate of the i th rule neuron. In addition, λ_2 and λ_3 can be calculated as

$$\lambda_2 = \beta_2 \overline{S_{ave}}, \quad (42)$$

$$\lambda_3 = \beta_3 \overline{S_{ave}}, \quad (43)$$

where β_2 and β_3 are coefficients related to the training samples.

4.2 Establishing the self-organizing fuzzy neural network

According to the designed self-organizing method, we proposed the MISA-SOFNN to predict the hourly concentration of $PM_{2.5}$. Furthermore, the main calculation procedure of the MISA-SOFNN is given as follows:

1. Network initialization: The initial number of rule nodes in MISA-SOFNN is set as a random number, and the inputs of the MISA-SOFNN are dominating variables including the selected meteorological and pollutant concentration variables and the extracted image features. The output layer of network has one node whose output is the $PM_{2.5}$ concentration. Several parameters, such as centers, widths and weights, are initialized as random numbers between 0 and 1.
2. Parameter learning: The MISA-SOFNN is trained by the improved gradient descent algorithm, and its parameters are adjusted by Eqs. (14)–(16) with all training samples.
3. Contribution rate calculating: When all the samples are trained, the contribution ratio of each rule neuron to the network output is calculated by Eq. (30).
4. Neuron combining: The MI between each two rule neurons is calculated using Eq. (27). The value of λ_1 is calculated based on Eq. (40). The two rule neurons whose MI is bigger than λ_1 are combined and the parameter of new rule neuron is initialized with Eq. (31).
5. Neuron pruning: The value of λ_2 is computed based on Eq. (42). The rule neurons whose contribution ratio are less than λ_2 are pruned, and the weights of the nearest neuron are adjusted by Eq. (37).
6. Neuron growing: The value of λ_3 is calculated based on Eq. (43). The rule neurons whose contribution ratios are higher than λ_3 will split into two new neurons, and the parameters of the new neurons are set according to Eq. (38).
7. Network retraining: The algorithm will go to procedure 2 to relearn the network parameters if the training process is not over. The training process will stop if the training RMSE achieves the preset RMSE (E_p) or the training step reaches the total number of training steps (D_{max}).
8. Network test: Once the MISA-SOFNN is optimized by the training samples, test samples are used to verify the generalization ability of MISA-SOFNN.

5 Experimental results and discussion

In this section, the major work can be divided into two parts. The first part is to select variables from the meteorological and pollutant concentration data by relevance analysis and extract the image feature variables from image data. The rest of the work is to evaluate and compare the performances of MISA-SOFNN and other state-of-the-art or popular methods including mechanism analysis model and neural network models.

5.1 Experimental setup

5.1.1 Parameter setting

In this paper, the MISA-SOFNN has an input layer, a membership function layer, a rule layer, a defuzzy layer and an output layer. The initial number of rule nodes is set as 2, and the framework of the MISA-SOFNN is 19-19 × 2-2-2-1. The D_{max} , E_p , maximum and minimum learning rate are determined as 800, 0.01, 0.15 and 0.01, separately. The centers, widths and weights of the MISA-SOFNN are initialized as random values between −1 and 1. The coefficients β_1 , β_2 and β_3 are set as 3, 0.5 and 3.5, respectively.

5.1.2 Experimental sample

The meteorological data, pollutant concentration data and image data are recorded every hour, and three types of data have 1200 simultaneous records, respectively. Then, a specific picture dataset for $PM_{2.5}$ concentration prediction can be established, which contains 1200 pictures and the associated meteorological and pollutant concentration variables in total. In order to ensure the reasonableness and reliability of the experimental results, 1000 pictures are retained by deleting the distorted photographs in picture dataset and their corresponding meteorological and pollutant concentration variables do not contain outliers. The experimental data adopt 1000 reserved pictures and their associated

meteorological and pollutant concentration variables, which are randomly divided into two parts by a ratio of 8 to 2, namely 800 training samples and 200 test samples. All experiments are based on the above experimental data and run with MATLAB 2014a on a PC with Windows 7 operating system Pentium IV 3.4 GHz CPU and 8 GB SDRAM memory.

5.1.3 Performance benchmarking

Seven performance indexes were introduced to evaluate the simulation results, that is, mean absolute error (MAE), Pearson linear correlation coefficient (PLCC), root-mean-square error (RMSE), index of agreement (IA), normalized mean bias (NMB), normalized mean gross error (NMGE) and mean bias (MB) [2, 6–8, 27]. In general, PLCC quantifies the overall correlation between the simulated and measured data and varies from -1 to 1 , where -1 represents the total anticorrelation, 1 is the total correlation and 0 is the absence of correlation. MAE can better reflect the average level of prediction errors. The greater the MAE value is, the worse the prediction performance of the model becomes. RMSE is used to measure the deviation between the outputs of a prediction model and the expected outputs. There is a positive correlation between the RMSE value and the prediction accuracy. IA is utilized to evaluate prediction accuracy of the model. The best IA value that can be achieved is 1 . NMB indicates over or under estimation of the model, and the ideal value is 0 . NMGE indicates mean error regardless of over or under estimation, and the ideal value is 0 . MB measures model bias and the ideal value is 0 . Moreover, seven performance indexes are defined as

$$\text{MAE} = \frac{\sum_{t=1}^N |y_p(t) - y_o(t)|}{N}, \quad (44)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^N (y_p(t) - y_o(t))^2}{N}}, \quad (45)$$

$$\text{PLCC} = \frac{\sum_{t=1}^N (y_o(t) - \bar{y}_o)(y_p(t) - \bar{y}_p)}{\sqrt{\sum_{t=1}^N (y_o(t) - \bar{y}_o)^2 \sum_{t=1}^N (y_p(t) - \bar{y}_p)^2}}, \quad (46)$$

$$\text{IA} = 1 - \frac{\sum_{t=1}^N (y_p(t) - y_o(t))^2}{\sum_{t=1}^N (|y_p(t) - \bar{y}_o| + |y_o(t) - \bar{y}_p|)^2}, \quad (47)$$

$$\text{NMB} = \frac{\sum_{t=1}^N (y_p(t) - y_o(t))}{\sum_{t=1}^N y_o(t)}, \quad (48)$$

$$\text{NMGE} = \frac{\sum_{t=1}^N |y_p(t) - y_o(t)|}{\sum_{t=1}^N y_o(t)}, \quad (49)$$

$$\text{MB} = \frac{\sum_{t=1}^N (y_p(t) - y_o(t))}{N}, \quad (50)$$

where $y_o(t)$ indicates the expected output of the t th input sample, $\bar{y}_o$ is the mean value of expected outputs of all samples, $y_p(t)$ represents the network computation output of the t th input sample, and $\bar{y}_p$ is the mean value of computation outputs of all samples. N means the total number of samples.

5.2 Performance results

5.2.1 Input variables selection

Several studies have demonstrated that the $\text{PM}_{2.5}$ concentration is closely related to pollutants and meteorological conditions. Ni et al. [26] showed that $\text{PM}_{2.5}$ concentration was significantly affected by wind speed, weather, etc. Zheng et al. [45] pointed out that weather conditions at different times, SO_2 , NO_2 and other pollutants have a great impact on $\text{PM}_{2.5}$ concentration. In this paper, the five pollutant concentration variables, the six meteorological variables and the 12 image feature variables could be acquired from three types of data separately. Considering a large number of variables and the certain correlation existing among variables, there would be information redundancy when the neural network was established to predict hourly $\text{PM}_{2.5}$ concentration, which would make the complexity of model higher and further affect the prediction performance. Therefore, the relevance analysis method was employed to select the input variables for increasing the prediction accuracy, simplifying prediction model and reducing data dimensionality.

Note that all image features should be used as dominant variables because only image features related to $\text{PM}_{2.5}$ concentration were extracted. According to Eqs. (1)–(5), the 12 image features associated with $\text{PM}_{2.5}$ concentration could be obtained, which contained E_{pc} , CE_{gr} , CE_{yb} , CE_{rg} , E_{n1} , E_{n2} , E_{n3} , E_{n4} , E_{n5} , E_{n6} , S and C . Furthermore, the meteorological data (visibility, weather, temperature, pressure, humidity and wind speed) and pollutant concentration data (PM_{10} , NO_2 , SO_2 , O_3 , CO) were screened by relevance analysis. The $\text{PM}_{2.5}$ concentration was considered as the main factor and the other data (meteorological data, pollutant concentration data) were used as the correlation factors. Correlation scatter diagrams of main factor and correlation factors are plotted using picture dataset in Fig. 3.

From Fig. 3, we can see that there are six figures that satisfy the linear or nonlinear curve fitting relationships, which means that $PM_{2.5}$ has mathematically fitting correlation with CO , NO_2 , PM_{10} , SO_2 , visibility and wind speed. In addition, the specific meanings of the horizontal axis of the scatter plot for weather and $PM_{2.5}$ are expressed as follows: 0—sunny, 1—cloudy, 2—overcast, 3—rainy,

4—sprinkle, 5—moderate rain, 6—heavy rain, 7—rain storm, 8—thunder storm, 9—freezing rain, 10—snowy, 11—light snow, 12—moderate snow, 13—heavy snow, 14—foggy, and 15—sand storm. From the scatter plot for weather and $PM_{2.5}$, 0, 1, 2, 11 and 14 occupy almost all points, which indicates that $PM_{2.5}$ concentration at Beijing is mainly affected by weather conditions including cloudy,

Fig. 3 The scatter plots of meteorological and pollutant concentration variables with $PM_{2.5}$ concentration, respectively

overcast and foggy. The scatter plots in the remaining figures show unclear mathematical relationship between variables and $PM_{2.5}$ concentration, which demonstrates that $PM_{2.5}$ concentration has apparently unequal relationships with O_3 , temperature, humidity and air pressure. In order to further guarantee the reliability of the dominant variable selection, the PCC, SCC and fitting equations are used to quantitatively evaluate the correlation between main factor and correlation factors. The PCC is only suitable for describing the linear correlation between two factors. The nonlinear correlation can be represented by the SCC. The PCC, SCC and fitting equations of $PM_{2.5}$ and visibility, wind speed, PM_{10} , NO_2 , SO_2 , CO are calculated, where the results are shown in Table 2.

For the above results, there are some possible explanations and reasons behind it. Low visibility usually indicates a higher concentration of $PM_{2.5}$ in terms of the meteorological elements. In addition, weather conditions also have an impact on $PM_{2.5}$ concentration, for example, foggy is usually considered to increase the $PM_{2.5}$ concentration in Beijing. Although there are more severe hazy days in spring and winter in Beijing, temperature, air pressure and humidity may not be the main factors affecting $PM_{2.5}$ concentration. Furthermore, wind speed can effectively promote the atmospheric transport and help to reduce the $PM_{2.5}$ concentration value, which also meets the conclusions drawn by many years of meteorological data analysis in Beijing. As for the pollutant concentrations, they, apart from O_3 , are the dominant factors and have positive correlation with $PM_{2.5}$ concentration.

In conclusion, on the basis of three respects of considerations which are composed of scatter diagrams, correlation coefficients and fitting formulas, the meteorological variables (visibility, weather, wind speed) and the pollutant variables (PM_{10} , NO_2 , SO_2 , CO) are selected as dominant variables. In addition, the dominant variables at the current moment are used as input variables to forecast $PM_{2.5}$ concentration in the next hour.

Table 2 The calculation results of PCC and SCC between $PM_{2.5}$ concentration and dominant factors

Correlation factor	PCC	SCC	Fitting equation
Visibility	–	–0.8859	$y = -102.7\ln(x) + 396.61$
Wind Speed	–	–0.7953	$y = -111.2\ln(x) + 398.1$
PM_{10}	0.9085	–	$y = 0.7894x - 9.8103$
NO_2	–	0.8812	$y = 6.533e^{0.0267x}$
SO_2	0.8566	–	$y = 2.4233x - 9.9276$
CO	0.8406	–	$y = 49.934x - 0.8979$

–The value is not calculated in this paper; y is $PM_{2.5}$ concentration; x is dominant factor

5.2.2 Forecasting $PM_{2.5}$ concentration

In order to examine the effectiveness of the MISA-SOFNN proposed in this paper from a qualitative perspective, the MISA-SOFNN, SOG-SASOFNN and FNN with the optimal structure were employed to forecast the hourly concentration of $PM_{2.5}$. The trial-and-error method was utilized to determine the optimal structure of FNN, and its structure was identified as 19-19 × 21-21-21-1 through repeated experiments. The input variables of the SOG-SASOFNN and FNN were the same as MISA-SOFNN, and the results of experiments of MISA-SOFNN were compared with SOG-SASOFNN and FNN. In addition, the weight learning process for the MISA-SOFNN adopted the improved gradient descent algorithm, the weight learning process for FNN employed a gradient descent algorithm and the weight learning process for the SOG-SASOFNN was implemented by a second-order gradient algorithm [32]. When the MISA-SOFNN, SOG-SASOFNN and FNN were deployed to estimate the hourly concentrations of $PM_{2.5}$ at Beijing, their experimental results are shown in Figs. 4, 5 and 6.

The combining, growing and pruning processes of rule nodes of the MISA-SOFNN and SOG-SASOFNN during the training period are shown in Fig. 4 where the initial numbers of their rule neurons are both 2. From Fig. 4, the MISA-SOFNN reacts faster at the beginning of structural adjustment and acquires a more compact structure through fewer training steps when compared with the SOG-SASOFNN. Moreover, we can find that the number of rule neurons of MISA-SOFNN and SOG-SASOFNN is 21 and 25, respectively, after the structural adjustment is completed. The number of rule neurons of MISA-SOFNN is the same as that of the FNN with the optimal structure, which further indicates that the MISA-SOFNN is more efficient than SOG-SASOFNN in terms of structural adjustment.

Fig. 4 Updating process of rule nodes of MISA-SOFNN and SOG-SASOFNN

Fig. 5 The RMSE dynamic change process of MISA-SOFNN and SOG-SASOFNN in training phase

In addition, the RMSE changes in the MISA-SOFNN and SOG-SASOFNN during the training process are shown in Fig. 5. As can be seen in Fig. 5, the RMSE of MISA-SOFNN is lower than SOG-SASOFNN during the training process and the RMSE curve of MISA-SOFNN is much smoother than SOG-SASOFNN, especially at the beginning of the training. Hence, it is obvious that the MISA-SOFNN converges

with lower training RMSE and has a more compact structure under much less training steps compared with the SOG-SASOFNN.

Figure 6 shows the scatter plots between the expected and predicted hourly concentrations of $PM_{2.5}$ during the training phase and test phase for the MISA-SOFNN, FNN and SOG-SASOFNN. It is shown in Figs. 6a, b that the training and test scatter plots of MISA-SOFNN are linear and cannot contain negative predicted output values. From Fig. 6c, d, the linearities of the training and test scatter plots of FNN are worse than MISA-SOFNN and test scatter plot of FNN has several negative predicted output values. As can be seen in Fig. 6e, f, the training and test scatter plots of SOG-SASOFNN are the worst linearity among three neural networks and have many negative values that deviate significantly from expected outputs. Obviously, the overall fitting effect of the MISA-SOFNN is not only superiority to SOG-SASOFNN and FNN in training phase, but also better than them in test phase. Therefore, the MISA-SOFNN has stronger learning ability and better generalization performance than the SOG-SASOFNN and FNN.

For evaluating the effectiveness of MISA-SOFNN from a quantitative perspective, four prediction models established by state-of-the-art or popular methods were used as competitors, which could be divided into two types

Fig. 6 Scatter plots between the expected and the predicted $PM_{2.5}$ concentrations for models: **a** MISA-SOFNN for training set; **b** MISA-SOFNN for test set; **c** FNN for training set; **d** FNN for test set; **e** SOG-SASOFNN for training set; and **f** SOG-SASOFNN for test set

according to their application scenarios. The first type involved three different neural networks, namely SOG-SASOFNN, FNN with the optimal structure and ESN [41]. The second type was composed of Eta-CMAQ [42]. The input variables of the SOG-SASOFNN, FNN and ESN were the same as that of the MISA-SOFNN. The weight learning process for the ESN was implemented by linear regression equation [41]. The description and evaluation of the $PM_{2.5}$ prediction in the eastern USA using the Eta-CMAQ model were detailed in Yu et al. [42]. Considering the evaluation criteria mentioned in Sect. 5.1.3 and computation time, the numerical comparison results of five prediction methods are given in Table 3.

As can be seen from Table 3, the MISA-SOFNN, generally speaking, has a better performance than FNN, SOG-SASOFNN, ESN and Eta-CMAQ. It can be found that the values of MAE, PLCC, IA, NMB, NMGE and MB for the MISA-SOFNN are better than that of the SOG-SASOFNN, FNN and ESN as well as the Eta-CMAQ in both training phase and test phase. However, the test RMSE of the MISA-SOFNN is larger than the Eta-CMAQ because the multi-source data used in this paper cover more broader scopes for $PM_{2.5}$ concentration than the Eta-CMAQ. Furthermore, the MISA-SOFNN takes less computation time and achieves better prediction performance than SOG-SASOFNN, which means that MISA-SOFNN can effectively perform the prediction tasks of $PM_{2.5}$ concentration through a limited computation time. Nevertheless, the SOG-SASOFNN spends more time than MISA-SOFNN in training phase, which may be caused by the computational complexity of the second-order algorithms. In addition, the MISA-SOFNN achieves the better prediction accuracy than FNN with the optimal

structure, indicating that the improved gradient descent algorithm proposed in this paper is superior to gradient descent algorithm. To sum up, based on the experimental analysis above, it can be concluded that the MISA-SOFNN can achieve more compact network structure, higher prediction performance and lower computation time than the SOG-SASOFNN, FNN, ESN and Eta-CMAQ in predicting the hourly concentrations of $PM_{2.5}$.

6 Conclusions

It is hard for traditional methods to validly forecast the hourly concentrations of $PM_{2.5}$ owing to high chaos and nonlinearity in the atmospheric and chemical processes. To solve these problems, a prediction method of hourly $PM_{2.5}$ concentration based on SOFNN and multi-source data has been presented in this paper. To be more concrete, multi-source data including meteorological data, pollutant concentration data and image data have first been obtained in different ways. In addition, the meteorological and pollutant concentration data have been used for feature selection using relevance analysis and the image features have been extracted from photographs in image data. Second, we have presented a novel SOFNN, which deploys the MI and SA to evaluate the amount of internal information of rule neurons and the contribution rate of each rule neuron, respectively. Besides, the parameters of the proposed model have been optimized by using an improved gradient descent algorithm. Third, the MISA-SOFNN has been used to infer $PM_{2.5}$ concentration in the next hour using the selected meteorological and pollutant

Table 3 Numerical comparison results of MISA-SOFNN, SOG-SASOFNN, FNN, ESN and Eta-CMAQ

Model	Metric							
	MAE	PLCC	RMSE	IA	NMB	NMGE	MB	Time
MISA-SOFNN								
Training	6.9047	0.9237	10.652	1.0000	−0.0010	0.0770	−0.0090	1.5881
Test	7.7981	0.9068	12.374	0.9965	−0.0169	0.0829	−1.3888	0.0163
SOG-SASOFNN								
Training	13.280	0.8227	22.705	0.9386	−0.0077	0.1771	−0.2443	3.8863
Test	15.801	0.8096	28.963	0.9299	−0.0253	0.1807	−2.1468	0.0212
FNN								
Training	10.198	0.8757	15.954	0.9770	−0.0038	0.1172	−0.1172	2.3421
Test	12.430	0.8599	19.570	0.9690	−0.0197	0.1344	−1.8191	0.0170
ESN								
Training	17.598	0.7405	32.352	0.8942	−0.0101	0.2233	−0.3132	3.4618
Test	20.964	0.7391	39.793	0.8727	−0.0322	0.2550	−2.5369	0.0196
Eta-CMAQ								
	–	0.2200 *	11.600 *	–	−0.3200 *	0.5100 *	−5.2000 *	–

–The value is not given in original paper; *the value is found in the original paper; the values of computation time are the average of training 800 steps and its unit is seconds

concentration variables and the extracted image feature variables at the current moment. The experimental results demonstrate that, compared to other existing methods, the MISA-SOFNN has a more compact network structure, higher prediction performance and lower computation time in terms of the hourly $PM_{2.5}$ prediction.

Overall, our proposed model gives desired results for predicting hourly concentration of $PM_{2.5}$, which brings the following benefits to researcher and practitioners. First, people can take precautionary measures in advance to avoid harm to their physical and mental health based on the accurate results of $PM_{2.5}$ prediction. Second, the MISA-SOFNN proposed in this paper can lay the foundation for researcher's in-depth exploration. Third, the prediction results can also provide an important reference for environmental protection departments to formulate air pollution prevention strategies. However, the quality of image data collected in this paper is not very good. We will consider using a high-pixel camera for manual acquisition in the future. Moreover, the performance of MISA-SOFNN can be further improved by considering the impact of local emissions on $PM_{2.5}$ concentration, which can guide our future work. Ultimately, the prediction method proposed in this paper can be used to predict the hourly concentration of $PM_{2.5}$ in the study area, which provides an effective method for forecasting hourly $PM_{2.5}$ concentration in other areas.

Funding This work is supported by the Key Project of National Natural Science Foundation of China (Nos. 61603009 and 61803006), the National Science and Technology Major Project (No. 2018ZX07111005) and the National Key Research and Development Project under Grants (No. 2018YFC1900800-5).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Assimakopoulos MN, Dounis A, Spanou A, Santamouris M (2013) Indoor air quality in a metropolitan area metro using fuzzy logic assessment system. *Sci Total Environ* 449:461–469. <https://doi.org/10.1016/j.scitotenv.2012.12.043>
- Carslaw DC, Ropkins K (2012) Openair—an R package for air quality data analysis. *Environ Modell Softw* 27:52–61. <https://doi.org/10.1016/j.envsoft.2011.09.008>
- Chen YY, Shi RH, Shu SJ, Gao W (2013) Ensemble and enhanced PM_{10} concentration forecast model based on stepwise regression and wavelet analysis. *Atmos Environ* 74:346–359. <https://doi.org/10.1016/j.atmosenv.2013.04.002>
- Choi LK, You J, Bovik AC (2015) Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Trans Image Process* 24(11):3888–3901. <https://doi.org/10.1109/TIP.2015.2456502>
- Davis BL, Guo JX (2000) Airborne particulate study in five cities of China. *Atmos Environ* 34(17):2703–2711. [https://doi.org/10.1016/S1352-2310\(99\)00528-2](https://doi.org/10.1016/S1352-2310(99)00528-2)
- Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S (2014) Complex time series analysis of PM_{10} and $PM_{2.5}$ for a coastal site using artificial neural network modelling and k-means clustering. *Atmos Environ* 94:106–116. <https://doi.org/10.1016/j.atmosenv.2014.04.051>
- Elbayoumi M, Ramli NA, Yusof NFFM, Yahaya ASB, Al Madhoun W, Ul-Saufie AZ (2014) Multivariate methods for indoor PM_{10} and $PM_{2.5}$ modelling in naturally ventilated schools buildings. *Atmos Environ* 94:11–21. <https://doi.org/10.1016/j.atmosenv.2014.05.007>
- Fu ML, Wang WW, Le ZC, Khorram MS (2015) Prediction of particulate matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. *Neural Comput Appl* 26(8):1789–1797. <https://doi.org/10.1007/s00521-015-1853-8>
- Groen IIA, Ghebreab S, Prins H, Lamme VAF, Scholte HS (2013) From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *J Neurosci* 33(48):18814–18824. <https://doi.org/10.1523/JNEUROSCI.3128-13.2013>
- Gu K, Wang SQ, Zhai GT, Ma SW, Yang XK, Lin WS, Zhang WJ, Gao W (2016a) Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure. *IEEE Trans Multimed* 18(3):432–443. <https://doi.org/10.1109/TMM.2016.2518868>
- Gu K, Zhai GT, Lin WS, Liu M (2016b) The analysis of image contrast: from quality assessment to automatic enhancement. *IEEE Trans Cybern* 46(1):284–297
- Gu K, Qiao JF, Li XL (2018) Highly efficient picture-based prediction of $PM_{2.5}$ concentration. *IEEE Trans Ind Electron* 66(4):3176–3184. <https://doi.org/10.1109/TIE.2018.2840515>
- Hacine-Gharbi A, Ravier P, Harba R, Mohamadi T (2012) Low bias histogram-based estimation of mutual information for feature selection. *Pattern Recognit Lett* 33(10):1302–1308. <https://doi.org/10.1016/j.patrec.2012.02.022>
- Han HG, Qiao JF (2010) A self-organizing fuzzy neural network based on a growing-and-pruning algorithm. *IEEE Trans Fuzzy Syst* 18(6):1129–1143. <https://doi.org/10.1109/TFUZZ.2010.2070841>
- Han HG, Lin ZL, Qiao JF (2017) Modeling of nonlinear systems using the self-organizing fuzzy neural network with adaptive gradient algorithm. *Neurocomputing* 266:566–578. <https://doi.org/10.1016/j.neucom.2017.05.065>
- Han M, Ren WJ (2015) Global mutual information-based feature selection approach using single-objective and multi-objective optimization. *Neurocomputing* 168:47–54. <https://doi.org/10.1016/j.neucom.2015.06.016>
- Han X, Zhang MG (2018) Assessment of the regional source contributions to $PM_{2.5}$ mass concentration in Beijing. *Atmos Ocean Sci Lett* 11(2):143–149. <https://doi.org/10.1080/16742834.2018.1412796>
- Hasler D, Suesstrunk SE (2003) Measuring colorfulness in natural images. *Proc SPIE* 5007:87–96. <https://doi.org/10.1117/12.477378>
- Herawan SG, Talib K, Putra A, Ismail AF, Shamsudin SA, Musthafah MT (2018) Prediction of generated power from steam turbine waste heat recovery mechanism system on naturally

- aspirated spark ignition engine using artificial neural network. *Soft Comput* 22(18):5955–5964. <https://doi.org/10.1007/s00500-017-2873-3>
20. Huang W, Tan JG, Kan HD, Zhao N, Song WM, Song GX, Chen GH, Jiang LL, Jiang C, Chen RJ et al (2009) Visibility, air quality and daily mortality in Shanghai, China. *Sci Total Environ* 407(10):3295–3300. <https://doi.org/10.1016/j.scitotenv.2009.02.019>
 21. Islam MM, Sattar MA, Amin MF, Yao X, Murase K (2009) A new adaptive merging and growing algorithm for designing artificial neural networks. *IEEE Trans Syst Man Cybern* 39(3):705–722. <https://doi.org/10.1109/TSMCB.2008.2008724>
 22. Kao CH, Hsu CF, Don HS (2012) Design of an adaptive self-organizing fuzzy neural network controller for uncertain nonlinear chaotic systems. *Neural Comput Appl* 21(6):1243–1253. <https://doi.org/10.1007/s00521-011-0537-2>
 23. Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. *Atmos Environ* 35(5):815–825. [https://doi.org/10.1016/S1352-2310\(00\)00385-X](https://doi.org/10.1016/S1352-2310(00)00385-X)
 24. Lin CJ (2008) An efficient immune-based symbiotic particle swarm optimization learning algorithm for TSK-type neuro-fuzzy networks design. *Fuzzy Sets Syst* 159(21):2890–2909. <https://doi.org/10.1016/j.fss.2008.01.020>
 25. Mishra D, Goyal P, Upadhyay A (2015) Artificial intelligence based approach to forecast PM_{2.5} during haze episodes: a case study of Delhi, India. *Atmos Environ* 102:239–248. <https://doi.org/10.1016/j.atmosenv.2014.11.050>
 26. Ni XY, Huang H, Du WP (2017) Relevance analysis and short-term prediction of PM_{2.5} concentrations in Beijing based on multi-source data. *Atmos Environ* 150:146–161. <https://doi.org/10.1016/j.atmosenv.2016.11.054>
 27. Ordieres JB, Vergara EP, Capuz RS, Salazar RE (2005) Neural network prediction model for fine particulate matter (PM_{2.5}) on the US-Mexico border in EL Paso (Texas) and Ciudad Juárez (Chihuahua). *Environ Modell Softw* 20(5):547–559. <https://doi.org/10.1016/j.envsoft.2004.03.010>
 28. Peng HC, Long FH, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>
 29. Peng JX, Li K, Irwin GW (2008) A new Jacobian matrix for optimal learning of single-layer neural networks. *IEEE Trans Neural Netw Learn Syst* 19(1):119–129. <https://doi.org/10.1109/TNN.2007.903150>
 30. Pratama M, Er MJ, Li X, Oentaryo RJ, Lughofer E, Arifin I (2013) Data driven modeling based on dynamic parsimonious fuzzy neural network. *Neurocomputing* 110:18–28. <https://doi.org/10.1016/j.neucom.2012.11.013>
 31. Qiao JF, Han HG (2012) Identification and modeling of nonlinear dynamical systems using a novel self-organizing RBF-based approach. *Automatica* 48(8):1729–1734. <https://doi.org/10.1016/j.automatica.2012.05.034>
 32. Qiao JF, Cai J, Han HG, Cai JX (2017) Predicting PM_{2.5} concentrations at a regional background station using second order self-organizing fuzzy neural network. *Atmosphere* 8(1):10. <https://doi.org/10.3390/atmos8010010>
 33. Qiao LP, Cai J, Wang HL, Wang WB, Zhou M, Lou SR, Chen RJ, Dai HX, Chen CH, Kan HD (2014) PM_{2.5} constituents and hospital emergency-room visits in Shanghai, China. *Environ Sci Technol* 48(17):10406–10414. <https://doi.org/10.1021/es501305k>
 34. Raga GB, Moyne LL (1996) On the nature of air pollution dynamics in Mexico City—I. Nonlinear analysis. *Atmos Environ* 30(23):3987–3993. [https://doi.org/10.1016/1352-2310\(96\)00122-7](https://doi.org/10.1016/1352-2310(96)00122-7)
 35. Riccio A, Chianese E, Agrillo G, Esposito C, Ferrara L, Tirimberio G (2014) Source apportion of atmospheric particulate matter: a joint Eulerian/Lagrangian approach. *Environ Sci Pollut Res* 21(23):13160–13168. <https://doi.org/10.1007/s11356-013-2367-5>
 36. Saide PE, Carmichael GR, Spak SN, Gallardo L, Osses AE, Mena-Carrasco MA, Pagowski M (2011) Forecasting urban PM₁₀ and PM_{2.5} pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. *Atmos Environ* 45(16):2769–2780. <https://doi.org/10.1016/j.atmosenv.2011.02.001>
 37. Scholte HS, Ghebreab S, Waldorp L, Smeulders AWM, Lamme VAF (2009) Brain responses strongly correlate with Weibull image statistics when processing natural images. *J Vis* 9(4):29–29. <https://doi.org/10.1167/9.4.29>
 38. Tao MH, Chen LF, Wang ZF, Ma PF, Tao JH, Jia SL (2014) A study of urban pollution and haze clouds over northern China during the dusty season based on satellite and surface observations. *Atmos Environ* 82:183–192. <https://doi.org/10.1016/j.atmosenv.2013.10.010>
 39. Vafa-Arani H, Jahani S, Dashti H, Heydari J, Moazen S (2014) A system dynamics modeling for urban air pollution: a case study of Tehran, Iran. *Transp Res Part D Transp Environ* 31:21–36. <https://doi.org/10.1016/j.trd.2014.05.016>
 40. Visalakshi S, Radha V (2014) A literature review of feature selection techniques and applications: review of feature selection in data mining. In: *IEEE international conference on computational intelligence and computing research*, pp 1–6. <https://doi.org/10.1109/ICCIC.2014.7238499>
 41. Xu Z, Xia XP, Liu XN, Qian ZG (2015) Combining DMSP/OLS nighttime light with echo state network for prediction of daily PM_{2.5} average concentrations in Shanghai, China. *Atmosphere* 6(10):1507–1520. <https://doi.org/10.3390/atmos6101507>
 42. Yu SC, Mathur R, Schere K, Kang DW, Pleim J, Young J, Tong D, Pouliot G, McKeen SA, Rao ST (2008) Evaluation of real-time PM_{2.5} forecasts and process analysis for PM_{2.5} formation over the eastern united states using the Eta-CMAQ forecast model during the 2004 ICARTT study. *J Geophys Res*. <https://doi.org/10.1029/2007JD009226>
 43. Yu W, Harris TJ (2009) Parameter uncertainty effects on variance-based sensitivity analysis. *Reliab Eng Syst Saf* 94(2):596–603. <https://doi.org/10.1016/j.res.2008.06.016>
 44. Zhang YL, Cao F (2015) Fine particulate matter (PM_{2.5}) in China at a city level. *Sci Rep* 5:14884. <https://doi.org/10.1038/srep14884>
 45. Zheng Y, Yi XW, Li M, Li RY, Shan ZQ, Chang E, Li TR (2015) Forecasting fine-grained air quality based on big data. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp 2267–2276. <https://doi.org/10.1145/2783258.2788573>
 46. Zhou YM, Dexter A (2013) Off-line identification of nonlinear, dynamic systems using a neuro-fuzzy modelling technique. *Fuzzy Sets Syst* 225:74–92. <https://doi.org/10.1016/j.fss.2012.11.017>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.