applied sciences

MDPI

Article

Research on PM2.5 Concentration Prediction Based on the CE-AGA-LSTM Model

Xiaoxuan Wu [1,2,*], Chen Zhang [1,2], Jun Zhu [1] and Xin Zhang [1,2]

[1] School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China; zhangchen@hfuu.edu.cn (C.Z.); 18852617705@163.com (J.Z.); zhangxin@hfuu.edu.cn (X.Z.)
[2] Anhui Province Urban Infrastructure Big Data Technology Application Engineering Laboratory, Hefei 230601, China
* Correspondence: wuxx@hfuu.edu.cn; Tel.: +86-138-667-43776

Abstract: The PM2.5 index is an important basis for measuring the degree of air pollution. The accurate prediction of PM2.5 concentration has an important guiding role in air pollution prevention and control. The Pearson Correlation Coefficient (PCC) is a common index used to mine the correlation between meteorological factors and other air pollutants. However, this index cannot be used to mine non-linear correlations, nor can it quantitatively analyze the weight of each related attribute. In order to accurately explore the correlation between meteorological factors and other air pollutants and to achieve an accurate prediction of PM2.5 concentration, this paper proposes a short- and long-time memory (LSTM) network prediction model based on Copula entropy (CE) and the adaptive genetic algorithm (AGA). By calculating CE, the correlation between multiple meteorological factors and various atmospheric pollutants and PM2.5 was analyzed. The correlation of influencing factors was sorted according to the size of the correlation coefficients. The contribution rate of meteorological factors and atmospheric pollutants to PM2.5 concentration was determined, used as the weight of each influencing factor and predicted as the input data of the prediction model. In this paper, a long- and short-term memory network (LSTM) suitable for time series data was selected as the prediction model, while the selection of model parameters was taken into account, and the relevant parameters were sought by an adaptive genetic algorithm (AGA). The air pollutant data and meteorological data of Beijing from 1 January 2016 to 31 December 2016 were selected, and MAE and RMSE were used as evaluation indexes. By comparing the experimental results of the CE-AGA-LSTM with those of other eight prediction models (LR, SVM, RF, ARMA, ST-LSTM, LSTM, CE-LSTM and CE-RNN), we found that among the models, the CE-AGA-LSTM model provided the lowest MAE and RMSE values, i.e., 14.5 and 21.88, respectively. At the same time, the loss rate and accuracy of the CE-AGA-LSTM model were evaluated, and the experimental results verified the validity of the model.

Keywords: correlation analysis; Copula entropy; PM2.5 concentration; forecasting model; LSTM

Citation: Wu, X.; Zhang, C.; Zhu, J.; Zhang, X. Research on PM2.5 Concentration Prediction Based on the CE-AGA-LSTM Model. *Appl. Sci.* **2022**, *12*, 7009. https://doi.org/10.3390/app12147009

Academic Editors: Jingsha He and Shengzong Zhou

Received: 13 June 2022
Accepted: 4 July 2022
Published: 11 July 2022

1. Introduction

With the deepening of urbanization, the improvement of productivity and the transformation of the production mode in China, the types of air pollutants have changed considerably. As particulate matter with a diameter less than or equal to 2.5 μm in the air environment, PM2.5 causes not only serious harm to the human body, but also huge economic loss to society. PM2.5 is harmful to the human body mainly in the respiratory and cardiovascular systems; therefore, reducing PM2.5 concentrations can effectively reduce health risks [1]. Although China has improved remarkably in PM2.5 control in recent years, the task of emission reduction is still very difficult to carry out. Therefore, it is especially important to simulate and forecast the PM2.5 concentration, which is a positive guidance and reference for scientific decisions of industrial structures and the formulation of environmental pollution prevention and control measures.

Statistical methods and numerical simulation methods can be used for pollutant concentration prediction. Potential forecasts are mainly based on meteorological conditions of atmospheric dilution and dispersion capacity [2]. Warnings are issued when weather conditions are expected to meet the criteria for potentially severe pollution. Concentration forecasts are direct predictions of pollutant concentrations in an area, and the predictions are quantitative. These air pollution prediction models can be classified as parametric and non-parametric or deterministic and non-deterministic. The prediction models used in this paper were parametric models, whose outputs are uncertain when the parameters of the equations in the model have to be determined.

Previous research is of great help for the prediction of PM2.5 concentration. however, the methods used are not sufficient as they do not consider the characteristics of different periods and the future dynamic effects of the concentrations of PM2.5. Especially, if dramatic changes of meteorological factors occur, such as heavy rain or strong wind, most methods fail in the mining of the dynamic effect of meteorological factors on the concentrations of PM2.5. At the same time, prediction models fail to effectively simulate the PM2.5 concentration dependence of space and time. In view of the above limitations, this study proposes an LSTM network based on Copula entropy to accurately predict PM2.5 concentration.

The contributions of this study mainly include three aspects:

(1) Copula entropy (CE) was used to measure the statistical independence between PM2.5 and meteorological factors and air pollutants. The non-linear correlation could be mined, and it will possible to better explore the dynamic influence of characteristic states at different times on PM2.5 concentration in the future to improve the accuracy of the prediction model.
(2) This paper converted the correlation analysis results into the weights of each influencing factor, which were fed into the prediction model as input data for the prediction. This reflects the degree of influence of attribute characteristics on PM2.5 concentration change, thus improving the accuracy of the prediction model.
(3) A genetic algorithm—a heuristic search algorithm—was adopted in this paper to determine the model parameters by optimization, which avoided repeated parameter tuning and further improved the accuracy of the model prediction.

In this paper, CE was used to analyze the correlation between PM2.5 and meteorological factors and air pollutants and to select attributes. Meanwhile, the correlation analysis results were used as attribute weights and input data to predict PM2.5 pollutant concentration through the LSTM model. The organizational structure of this paper is as follows: the Section 2 introduces a background review of the theoretical basis, experimental theory and preliminary work; the Section 3 introduces the methodology applied in this paper; the Section 4 presents the experimental results of our work. Finally, we draw conclusions from the work of this paper in Section 5.

2. Related Works

2.1. Correlation Analysis

In the process of actual production, life and scientific research, it is often found that multiple factors have an impact on a certain object or phenomenon at the same time. Since there is generally a certain correlation, either strong or weak, between multiple variables, information overlaps to some extent, hindering the in-depth analysis of objects or phenomena. Correlation analysis is a statistical method to analyze whether two or more objects or phenomena are correlated and the strength of their correlation [3].

PM2.5 in the atmosphere is produced by both primary and secondary sources of pollution. Environmental monitoring stations for PM2.5 detect mainly secondary particulate matter, mainly contributed by ammonium, sulfate and nitrate, which are substances produced by SO_x and NO_x emissions [4]. Therefore, NO_2, SO_2, O_3, etc. are key factors influencing the atmospheric PM2.5 concentration [5]. The generation and flow of PM2.5 are significantly related to the local climate environment [6]. Different meteorological

conditions also have a large impact on the diffusion and transport of pollutants. Zhou [7] discussed the correlation between air pollutants and meteorological conditions in Zhumadian City and found that different seasons and meteorological conditions have a certain influence on air pollution concentrations, with a negative correlation between air pollutants and temperature and precipitation, and a positive correlation with wind speed. Co The main factors affecting the change of PM2.5 concentrations include not only the mutual transformation between various atmospheric pollutants caused by chemical interactions, but also the influence of different meteorological factors and geographical characteristics on the atmospheric environment. Since the regional geographical characteristics are relatively stable, this paper focuses on the correlation among atmospheric pollutants and between atmospheric pollutants and meteorological factors.

The measurement of correlation was proposed and studied in the early stage of statistics, and the most widely used measurement is the Pearson Correlation Coefficient (PCC). Liu.et al. [8] used PCC for correlation analysis to obtain the correlation matrix of six major AQI indicators in a city regarding available monitoring data. Zeng et al. [9] used PCC to analyze the correlation between PM2.5 concentration in summer and autumn in Beijing and six meteorological factors, including air temperature, relative humidity, wind speed, water vapor pressure, atmospheric pressure and wind direction. However, PCC is limited to linear Gaussian cases and is often inadequate for complex nonlinear natural phenomena; therefore, its application is very limited. If PCC is applied without considering the preconditions, the conclusions drawn are unreliable.

2.2. PM2.5 Concentration Prediction

In recent years, researchers have introduced artificial intelligence methods and hybrid three-dimensional models for measuring air pollutants and achieved some results. Hybrid methods have good robustness, low risk, and strong adaptability [10–12]. In terms of artificial intelligence methods, Güler Dincer et al. [13] developed a new fuzzy time series model based on the fuzzy K-Medoid clustering algorithm to predict SO_2 concentrations in Turkey. Wang. et al. [14] proposed a prediction method for PM2.5 concentration based on the LSTM and SVR hybrid model. By introducing relaxation variables into the SVR model, the LSTM model can correct large prediction errors, so to achieve better predictions.

With more and more attention paid to air pollution, researchers have also proposed many spatiotemporal prediction models to predict air pollution. Li et al. [15] presented a new ensemble reinforcement learning gated unit model. The key of this model is to establish a sub-series forecasting model by the SAE-GRU method. SAE was used to obtain low-latitude features of PM2.5 data, and GRU was applied to finish PM2.5 sub-series forecasting. Zhao et al. [16] proposed a new air quality spatio-temporal prediction model to predict future air quality based on a large amount of environmental data. Wen et al. [17] proposed a deep multi-output LSTM (DM-LSTM) neural network model that incorporates three deep learning algorithms to configure the model to extract the key factors in complex spatio-temporal relationships. Zhou et al. [18] proposed a hybrid model for spatio-temporal forecasting of PM2.5 based on graph convolutional neural network (GCN) and LSTM. Huang et al. [19] aimed at the long-term prediction of PM2.5 concentration, considering PM2.5 spatio-temporal correlation between multivariate data, and a TSMN prediction model was proposed. The model constructs a local memory component and a neighborhood component to explicitly model the temporal and spatial dependencies. Zhu et al. [20] proposed an attention-based parallel network (APNet), which can extract short-term and long-term temporal features simultaneously, based on the attention-based CNN-LSTM multilayer structure to predict PM2.5 concentration in the next 72 h. Li et al. [21] proposed a PM2.5 prediction model based on the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), differential symbolic entropy (DSE), variational mode decomposition improved by the butterfly optimization algorithm (BVMD) and kernel extreme learning machine optimized by the crow search algorithm (CSA-KELM), which was named CEEMDAN-DSE-BVMD-CSA-KELM. Hu et al. [22] proposed a hybrid machine

learning model (WD-SA-LSTM-BP model) based on simulated annealing (SA) optimization and wavelet decomposition.

3. Methods

3.1. Correlation Analysis of PM2.5 and Influencing Factors

CE is a mathematical concept strictly defined by Ma Jian et al. [23] in 2008 to measure the relationship of statistical independence. CE is an ideal measure of statistical independence as a correlation analysis tool for any nonlinear, non-Gaussian correlations and is a good method for causal discovery, since no assumptions need to be made. As pollutant factors affecting PM2.5 interact with each other and reflect nonlinear characteristics, the traditional analysis method suitable for linear correlations will affect the accuracy of the prediction [24]. The CE values of each factor and PM2.5 were calculated for 8 meteorological factors (air temperature, body temperature, air pressure, humidity, rainfall, wind direction, wind force, wind speed) and 5 kinds of atmospheric pollutants (PM10, SO_2, NO_2, CO, O_3), and the relevance of the influencing factors was ranked according to the size of the correlation coefficient. The top 4 factors with greater relevance to PM2.5 were taken as the main influencing factors; CE was used to analyze the correlation between the 8 main influencing factors and PM2.5 again, and the correlation coefficients were normalized as the weights of each main influencing factor and introduced into the prediction model to realize the prediction of PM2.5 pollutant concentration.

The mutual information (MI) indicator that describes the interrelation between different variables originates from the information theory and reflects the size of common information regarding different variables, i.e., the larger MI is, the stronger the correlation between the variables, and vice versa, the weaker the correlation. Assuming that there is a certain connection between the random variables X and Y, the MI between them can be calculated by Equation (1)

$$\mathrm{MI} = \frac{1}{N}\sum_{i=1}^{N} p(x_i, y_i) \ln \frac{p(x_i, y_i)}{p(x_i)p(y_j)} \tag{1}$$

where N is the sample size, x_i and y_i are the samples of the random variables X and Y, respectively.

From Equation (1), it is easy to find that when X and Y are uncorrelated, the value of MI is close to 0, while when X and Y present a functional relationship, the value of MI will be close to positive infinity. Compared with other similarity indexes, the MI index has the advantage that in addition to reflect the nonlinear correlation between variables, it does not change in size under the influence of any reversible transformation of random variables.

MI has strong information mining ability, but the joint distributions of different random variables in practical studies are often skewed and non-homogeneous, so it is difficult to find the appropriate distribution type to fit it. To solve this problem, this paper introduces the Copula theory. The Copula theory is a representation of multivariate dependencies by Copula functions [25,26].

Definition 1 (CE). *Let X be the random variable of the edge distribution u and the association density$c(u)$; the CE of X is defined as:*

$$H_c(X) = -\int_u c(u) \log c(u) du \tag{2}$$

In information theory, MI and entropy are two different concepts [27]. However, Ma and Sun proved that they are essentially the same [23], and MI is also entropy with the following relationship.

Theorem 1. *The MI of a random variable is equivalent to negative CE, i.e.,*

$$I(X) = -H_c(X) \tag{3}$$

Proof of Theorem 1.

$$I(x) = \int_x p(x) \log \frac{p(x)}{\prod_i p_i(x_i)} dx = \int_x c(u_x) \prod_i p_i(x_i) \log c(u_x) dx \tag{4}$$

$$= \int_x c(u_x) \log c(u_x) du_x = -H_c(x)$$

Based on Theorem 1, a simple nonparametric method is proposed which requires only two steps to estimate CE or MI from data.

Step 1. Estimate the empirical Copula density function (ECD).

Step 2. Estimate the CE. □

For Step 1, given the independent and identically distributed data samples $\{X_1, \cdots, X_T\}$ generated by the random variable $X = \{x_1, \cdots, x_N\}^T$, the ECD can be estimated more easily as follows.

$$F_i(x_i) = \frac{1}{T} \sum_{t=1}^{T} \chi\left(X_t^i \leq x_i\right) \tag{5}$$

where χ denotes the indicator function when $i = 1, \cdots, N$; assuming that $u = [F_1, \cdots, F_N]$, a new sample set $\{u_1, \cdots, u_T\}$ can be derived as ECD data $c(u)$.

The KNN method was suggested in [28]. Based on the Copula theory, the KNN method was used to rank the impact factors through a two-step approach for the nonparametric estimation of CE in this paper.

3.2. AGA-LSTM Prediction Model

LSTM is a subtle control of the combination of short-term memory and long-term memory through a "gate structure", which solves the problem of gradient disappearance to a certain extent and gives better results than a recurrent neural network for time series data analysis. However, in the process of constructing the LSTM model, some model parameters, such as the number of neurons in the hidden layer, the number of training times, the learning rate, etc., need to be assumed first, and the selection of these parameters also affects the prediction accuracy of the model. In this paper, a heuristic search algorithm-genetic algorithm was used to determine the model parameters by optimizing the global optimal solution and improve the accuracy of model prediction.

Genetic algorithms (GA) are a class of stochastic search algorithms that draw inspiration from natural selection and natural genetic mechanisms in biology. They are very suitable for dealing with complex and nonlinear optimization problems which are difficult to solve by traditional search methods. The crossover probability and mutation probability of genetic algorithms are the key parameters which affect their behavior and performance and directly influence the convergence of the algorithms. Adaptive genetic algorithms (AGAs) [28] enable the crossover probability and mutation probability to change automatically with fitness. In this paper, AGA and LSTM were integrated to build an AGA–LSTM prediction model so to achieve PM2.5 prediction. The basic framework of the AGA–LSTM prediction model is shown in Figure 1.

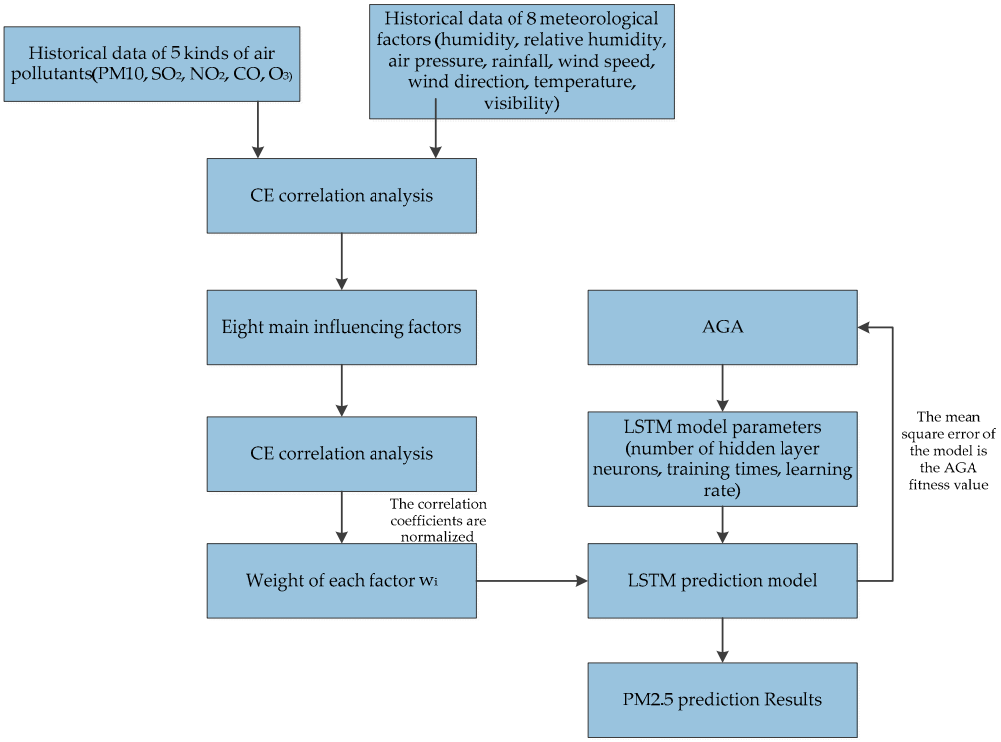

Figure 1. Basic framework of the AGA–LSTM prediction model.

Basic process of the AGA–LSTM prediction model:

Step 1. Binary encode of the number of neurons, training times and learning rates of the parameters' hidden layer in the LSTM.

Step 2. Generate the initial population N (even number).

Step 3. Establish the LSTM model, train and predict the data of training set and test set and take the predicted mean-square error as the AGA fitness value f_i.

Step 4. Select N individuals according to the roulette rule and calculate f_{avg} and $f_{\max}$.

Step 5. Each individual in the population is randomly paired into pairs, forming a total of N/2 pairs, and each pair of individuals is calculated according to the adaptive formula $P_c = \begin{cases} \frac{k_1(f_{\max}-f')}{f_{\max}-f_{avg}}, & f' > f_{avg} \\ k_2, & f' \le f_{avg} \end{cases}$ to generate the adaptive crossover probability, randomly generating R (0,1), if $R < P_c$; then, the crossover operation is performed on a pair of chromosomes.

Step 6. For all individuals N in the population, calculate the adaptive variation probability according to the adaptive variation formula $P_m = \begin{cases} \frac{k_3(f_{\max}-f)}{f_{\max}-f_{avg}}, & f > f_{avg} \\ k_4, & f \le f_{avg} \end{cases}$ and, if $R < P_m$, then perform crossover operations on that chromosome.

Step 7. Calculate the fitness of new individuals generated by crossover and mutation, which together with their parents form a new population.

Step 8. Determine whether the termination condition is satisfied; if it is satisfied, the termination returns the optimal parameters, otherwise execute step 4.

Step 9. Use the optimal parameters obtained by AGA to construct the LSTM network model; train the model and obtain the prediction results.

4. Experiments

4.1. Experimental Data

The experimental data in this paper were obtained from the hourly meteorological data (weather conditions, temperature, body temperature, barometric pressure, humidity, rainfall, wind direction, wind force, wind speed) and hourly air quality monitoring data

(PM2.5, PM10, SO_2, NO_2, CO, O_3) of Beijing from 1 January 2016 to 31 December 2016 from Nanjing Yunchuang Big Data Technology Co., Ltd. The air quality monitoring data refer to the hourly monitoring data of 12 monitoring points. In order to make the meteorological data consistent with the air quality monitoring data, the data of one monitoring point were taken uniformly. The position of the monitoring point is 116.28 longitude and 39.89 dimension; so, there were 8760 data records each, but due to some uncontrollable factors, some data were missing, and the data records finally used were 6430. The Chinese descriptions of weather conditions, wind direction, and wind power are coded as shown in Tables 1–3.

Table 1. Weather Condition Codes.

Weather Conditions	Code	Weather Conditions	Code
Clear	1	Fog	10
Haze	2	Rain and snow	11
Cloudy	3	Snow	12
Yin	4	Moderate to heavy snow	13
Light rain	5	Heavy Snow	14
Moderate to heavy rain	6	Heavy to blizzard	15
Heavy rain	7	Floating dust	16
Showers	8	Medium Rain	17
Thundershowers	9	Rainstorm	18

Table 2. Wind Code.

Wind Direction	Code
North Wind	1
Northeast wind	2
East Wind	3
Southeast Wind	4
South Wind	5
Southwest Wind	6
West Wind	7
Northwest Wind	8

Table 3. Wind Power Code.

Wind Power	Code
Breeze	1
Level 1	2
Level 2	3
Level 3	4
Level 4	5
Level 5	6

4.2. Analysis of the Experimental Results

4.2.1. Selection of Influence Factors

The hardware environment for this experiment was Intel(R) Core(TM) i7-8565U CPU 1.80 GHz, 8 GB of RAM, Windows 10 as the operating system, and Python 3.7.8 as the programming tool for this experiment.

In order to analyze the correlation between PM2.5 and meteorological factors and other atmospheric pollutants and to further select the influencing factors, this experiment ranked PM2.5 with respect to eight meteorological factors (air temperature, body temperature, air pressure, humidity, rainfall, wind direction, wind force, wind speed) and five atmospheric pollutants (PM10, NO_2, SO_2, O_3, CO) by non-parametric estimation of CE, and considered

the top four with higher correlation with the first four that displayed greater correlation with PM2.5 as the main influencing factors, thus obtaining eight influencing factors. Again, the correlation between the eight main influencing factors and PM2.5 was analyzed by the method presented in this paper, and the estimated CE values were normalized as the weights of each main influencing factor and introduced into the prediction model to achieve pollutant prediction. The experimental results are shown in Figures 2 and 3. The horizontal coordinates 1–8 of Figure 2 indicate air temperature (TMP), body temperature (FEELST), air pressure (PRES) relative humidity (HUM), precipitation (RAIN), wind direction (WDIR), wind speed (WSC) and wind speed (WSPD), respectively. The horizontal coordinates 1–5 of Figure 3 indicate PM10, NO_2, SO_2, O_3, and CO, respectively.

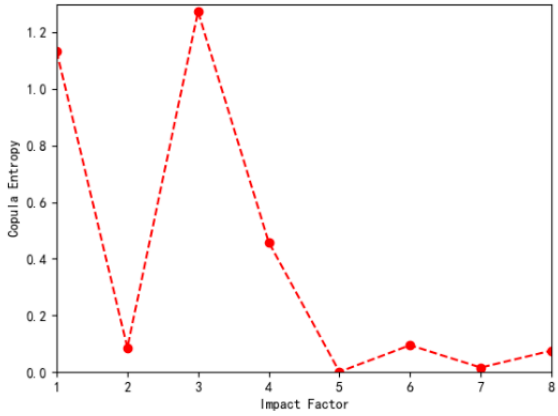

Figure 2. CE of each meteorological factor and PM2.5.

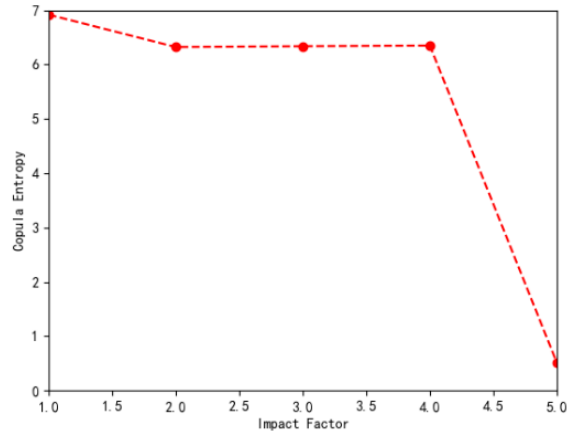

Figure 3. CE of each pollutant factor and PM2.5.

In Figure 2, the top four influencing factors with stronger correlation are PRES, TMP, HUM and WDIR according to the calculated CE results of each meteorological factor with PM2.5. Similarly, the top four influencing factors with stronger correlation are PM10, NO_2, SO_2 and O_3, according to the calculated CE results of each pollutant factor with PM2.5 in Figure 3. It was also found that PM10, NO_2, SO_2 and O_3 had stronger correlations with PM2.5 than each meteorological influence factor. Therefore, the eight influencing factors PM10, NO_2, SO_2, O_3, PRES, TMP, HUM and WDIR with stronger correlation with PM2.5 were selected; the CE results of the eight influencing factors and PM2.5 concentration were obtain to calculate the weight of each influencing factor ω_i. The attribute data of the eight impact factors x_i combined with the weights ω_i were input into the model for prediction.

4.2.2. Prediction of PM2.5 Concentration Based on the AGA-LSTM Prediction Model

Based on Section 4.2.1, eight impact factors were selected using the CE method, and the new obtained dataset had 10 dimensions, i.e., PM10, NO_2, SO_2, O_3, PRES, TMP, HUM,

WDIR, PM2.5; so, the dataset contained a total of 64,300 data records. The pattern of change of each feature at different times was further analyzed by plotting the time-series feature maps of each feature at different times, as shown in Figure 4. From Figure 4, it can be seen that each feature of the dataset had a certain periodicity.

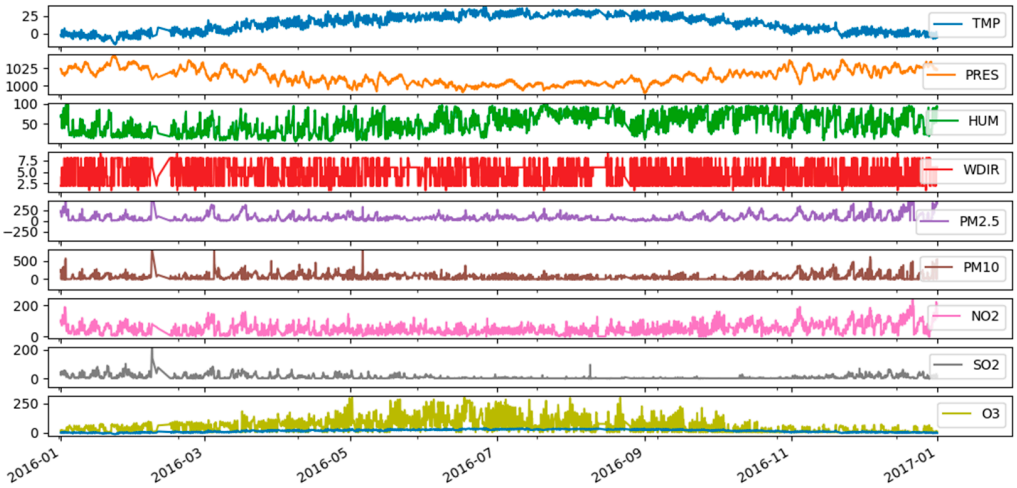

Figure 4. Time-series characteristics of PM2.5 concentrations at different times for each feature.

In this experiment, the input data were normalized using Equation (6) to map the values to decimals between 0, 1, and finally the dataset was processed into the data format needed for supervised learning and input into the prediction model.

$$x = \frac{x - \min}{\max - \min} \tag{6}$$

The format of each supervised learning data was: $(x_1\omega_1, x_2\omega_2, x_3\omega_3, \cdots, x_{10})$, where the first nine dimensions are the input training data at time t – 1 and the last dimension is the label, i.e., PM2.5 concentration at time t. The length of each input sequence was 24. To predict the PM2.5 concentration of the next day, data of the previous seven days were used as the model input data for prediction, and the PM2.5 concentration data of the first day after the seven days were used as the model output for the prediction. Therefore, the input data of the prediction model were 7*24*10, and the output data were 24*1. The data were divided into two parts for the experiment, with 70% of them as the training set, and 30% as the test set. The number of neurons in the hidden layer, the number of training times, and the learning rate in the LSTM model were determined by AGA, and the relevant parameters of the prediction model were set as shown in Table 4.

Table 4. Prediction Model-Related Parameter Settings.

Model Parameters	Set Value
Input_size	10
Output_size	1
hidden_size	32
num_layers	2
Activation function	ReLU
batch_size	60
Optimization algorithm	Adam
lr	0.001
loss function	MSE
epochs	2000
dropout	0.25

To verify the feasibility and accuracy of the proposed method in this paper, nine methods of prediction models were designed for comparison tests, namely, the Linear Regression models (LR), the Support Vector Regression models (SVM), the Random forest (RF), Autoregressive Moving Average models (ARMA), the ST-LSTM model in reference [16], the direct input LSTM prediction model with all attributes (LSTM), the CE-based LSTM prediction model (CE+LSTM), the proposed method in this paper (CE+AGA-LSTM) and the CE-based RNN prediction model (CE+RNN). To evaluate the performance of the prediction models, five samples (five 7-day time series data) were randomly selected as input data to predict PM2.5 concentrations, which were measured by three metrics: MAE, RMSE and R^2. The three indicators were obtained from Equations (7)–(9). The results of the experimental comparison are shown in Tables 5–7. The best results are marked in bold.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

$$R^2 = \mathrm{SSR}/\mathrm{SST} = 1 - \mathrm{SSE}/\mathrm{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{9}$$

Here, $\hat{y}_i$ and y_i represent the predicted and true values of PM2.5 concentration, respectively.

Tables 5–7 show MAE, RMSE and R^2 averages for the nine models over the five sample sets. Each column value indicates MAE, RMSE and R^2 values of the PM2.5 concentration of that model in 1–5 sample sets, and the results show that the proposed method in this paper (CE-AGA-LSTM) had the smallest MAE and RMSE and the largest R^2. Meanwhile, we found that the RF prediction model and the LSTM prediction model gave the same results as the CE-LSTM prediction model, while the CE-RNN prediction model had the worst performance. As can be seen from the above tables:

- Compared with the RNN model, the LSTM model is more suitable for model prediction of time series data, and its unique "gate structure" better realizes the effective combination of short-term memory and long-term memory.
- Although the LSTM model is more suitable for the prediction of time series data, for PM2.5 concentration prediction, long-term prediction in days, weeks and months is different from short-term prediction in seconds and minutes in data processing and prediction model construction; therefore, to meet the needs of the actual situation, according to the prediction of time, different prediction models should be designed. This is also a key direction of future research.

Table 5. Mean Values of MAE in five Experiments.

Sample	MAE								
	LR	**SVM**	**RF**	**ARMA**	**ST-LSTM**	**LSTM**	**CE+LSTM**	**CE+AGA-LSTM**	**CE+RNN**
1	16.826	25.516	15.747	16.735	15.521	15.9372	16.1229	**14.2270**	24.8390
2	16.782	25.782	15.032	17.352	15.602	16.2981	15.5084	**14.5400**	22.1463
3	17.855	26.792	15.248	17.395	15.272	16.2664	15.5939	**14.4312**	21.7606
4	17.364	26.455	15.289	16.897	14.583	16.1936	15.3963	**14.0790**	20.4152
5	17.522	26.374	14.553	16.823	14.890	15.5610	15.4919	**15.2141**	19.2932
Average	17.2698	26.1838	15.1738	17.0404	15.1736	16.05126	15.62268	**14.49826**	21.69086

Table 6. Mean Values of RMSE in five Experiments.

Sample	RMSE								
	LR	**SVM**	**RF**	**ARMA**	**ST-LSTM**	**LSTM**	**CE+LSTM**	**CE+AGA-LSTM**	**CE+RNN**
1	26.302	28.433	25.365	27.352	25.377	25.0623	24.4878	**21.7831**	36.0257
2	26.585	29.576	25.487	27.433	26.432	24.4848	24.0016	**21.9509**	32.4109
3	26.033	29.988	24.355	27.982	25.780	24.7315	24.3585	**21.9374**	32.2737
4	26.278	29.774	24.708	27.321	25.821	24.6910	23.8413	**21.4525**	30.8680
5	26.592	29.061	24.583	27.458	25.337	25.2206	23.5822	**22.3172**	28.7630
Average	26.358	29.3664	24.8996	27.5092	25.7494	24.83804	24.05428	**21.88822**	32.06826

Table 7. Mean Values of R^2 in five Experiments.

Sample	R^2								
	LR	**SVM**	**RF**	**ARMA**	**ST-LSTM**	**LSTM**	**CE+LSTM**	**CE+AGA-LSTM**	**CE+RNN**
1	0.902	0.896	0.923	0.890	0.932	0.937	0.939	**0.947**	0.903
2	0.921	0.917	0.920	0.901	0.938	0.930	0.940	**0.949**	0.905
3	0.927	0.899	0.911	0.878	0.929	0.928	0.942	**0.950**	0.866
4	0.919	0.858	0.927	0.922	0.934	0.925	0.944	**0.952**	0.879
5	0.904	0.906	0.928	0.913	0.933	0.921	0.935	**0.956**	0.893
Average	0.9146	0.8952	0.9218	0.9008	0.9332	0.9282	0.940	**0.9508**	0.8892

The loss rate and PM2.5 concentration prediction results of the method in this paper for the test set are shown in Figures 5 and 6. As can be seen in Figure 6, the prediction accuracy of the CE-AgA-LSTM prediction model reached 92.8%.

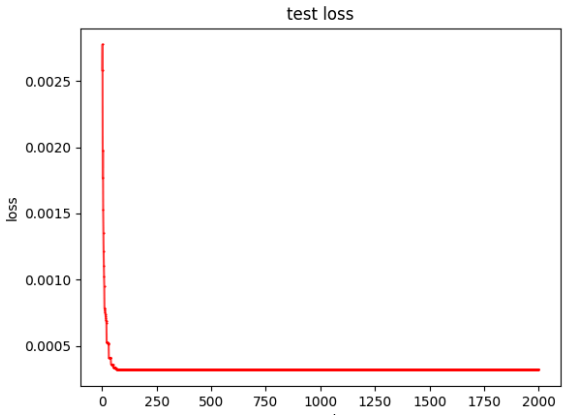

Figure 5. Loss rate of model testing.

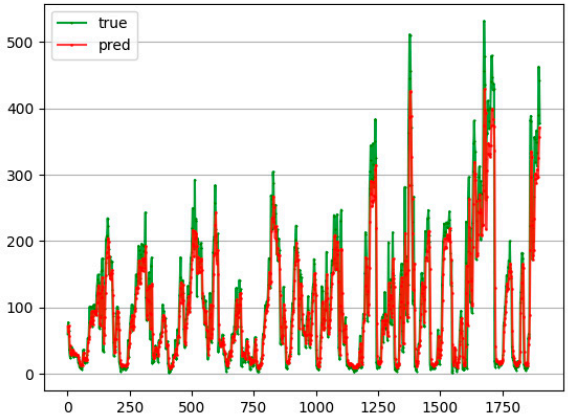

Figure 6. Predictions of model testing.

5. Conclusions

This paper proposes a prediction model for the PM2.5 concentration: the CE-AGA-LSTM model. Taking the meteorological data and air pollutant data of Beijing in 2016 as an example, the CE-AGA-LSTM prediction model and eight other models were compared based on the MAE, RMSE and R^2 indexes. The CE-AGA-LSTM prediction model used five randomly selected time series data samples as input data to predict PM2.5 concentration. Its average results in terms of MAE, RMSE and R^2 were 14.49826, 21.88822 and 0.9508, respectively. Compared with the results of other models, the results showed that the CE-AGA-LSTM prediction model is feasible. The prediction accuracy of the model on the test set reached 92.8%.

Despite its excellent performance, the model proposed in this paper also has some limitations: (1) the method in this paper only predicted the PM2.5 concentration of a single station, without considering the relationship between multiple stations in the whole city. Data of all stations in a single city can be integrated to further mine the spatial relationship between stations, so to achieve an accurate prediction of PM2.5 concentration. Data of a single site are more suitable for predicting the regional PM2.5 concentration of the site in the future. Selecting the primary site, mining the correlation between the primary site and each site and selecting the attributes would be more conducive to realizing the prediction of PM2.5 concentration of the whole city. (2) The prediction by the LSTM on stationary time series data was better than that of non-stationary data. Therefore, Empirical Mode Decomposition (EMD) was adopted to decompose and stabilize the input data, and better prediction results could be obtained. (3) In this paper, the data of the first seven days were used to predict the PM2.5 concentration of the next 24 h. We can also predict the PM2.5 concentration of the next month every day. However, for large-scale predictions, the results of the method described in this paper may not be good, so further research is necessary.

Author Contributions: Conceptualization, X.W.; methodology, X.W.; software, X.W. and C.Z.; validation, X.W. and J.Z.; formal analysis, X.W. and C.Z.; resources, X.W. and X.Z.; data curation, X.W. and J.Z.; writing—original draft preparation, X.W.; writing—review and editing, X.W.; project administration, C.Z.; funding acquisition, C.Z. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Universities Natural Science Research Project of Anhui Provincial, grant number KJ2021ZD0118, and the University Humanities and Social Sciences Research Project of Anhui Provincial, grant number KJ2021A0993.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brook, R.D.; Rajagopalan, S.; Pope, C.A., III; Brook, J.R.; Bhatnagar, A.; Diez-Roux, A.V.; Holguin, F.; Hong, Y.; Luepker, R.V.; Mittleman, M.A.; et al. Particulate matter air pollution and cardiovascular disease. *Circulation* **2010**, *121*, 2331–2378. [CrossRef]
2. Bai, L.; Wang, J.; Ma, X.; Lu, H. Air pollution forecasts: An overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [CrossRef]
3. Yu, Z.; Lan, D. Correlation analysis of $PM_{2.5}$ and air pollutants in Harbin City based on PLS1. *J. Ecol. Environ.* **2014**, *23*, 1953–1957.
4. Hodan, W.M.; Barnard, W.R. *Evaluating the Contribution of $PM_{2.5}$ Precursor Gases and Re-Entrained Road Emissions to Mobile Source $PM_{2.5}$ Particulate Matter Emissions*; MACTEC Federal Programs: Durham, NC, USA, 2004.
5. Kristiani, E.; Kuo, T.Y.; Yang, C.T.; Pai, K.C.; Huang, C.Y.; Nguyen, K.L.P. $PM_{2.5}$ Forecasting Model Using a Combination of Deep Learning and Statistical Feature Selection. *IEEE Access* **2021**, *9*, 68573–68582. [CrossRef]
6. Lixin, W.; Liangyu, Z.; Huan, W. Analysis and simulation study on the influence of air pollution and meteorological conditions in Baoding City. *Environ. Dev.* **2018**, *30*, 162–163.
7. Yang, Z. Analysis of Air Pollution Meteorological Correlation in Zhumadian City and Its Forecast and Early Warning System Design. Master's Thesis, Nanjing University of Information Engineering, Nanjing, China, 2018.
8. Liu, T.; Wu, M.P.; Zhang, K.D.; Liu, Y.; Zhong, J. Correlation Analysis and Control Scheme Research on $PM_{2.5}$. *Appl. Mech. Mater.* **2014**, *590*, 888–894. [CrossRef]

9. Jing, Z.; Wang, M.-e.; Zhang, H.-x. Correlation between atmospheric $PM_{2.5}$ concentration and meteorological factors during summer and autumn in Beijing, China. *Chin. J. Appl. Ecol.* **2014**, *25*, 2695–2699.
10. Yang, C.T.; Chen, S.T.; Den, W.; Wang, Y.T.; Kristiani, E. Implementation of an intelligent indoor environmental monitoring and management system in cloud. *Future Gener. Comput. Syst.* **2019**, *96*, 731–749. [CrossRef]
11. Yang, C.T.; Chen, C.J.; Tsan, Y.T.; Liu, P.Y.; Chan, Y.W.; Chan, W.C. An implementation of real-time air quality and influenza-like illness data storage and processing platform. *Comput. Hum. Behav.* **2019**, *100*, 266–274. [CrossRef]
12. Yan, R.; Liao, J.; Yang, J.; Sun, W.; Nong, M.; Li, F. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Syst. Appl.* **2021**, *169*, 114513. [CrossRef]
13. Nevin, G.D.; Özge, A. A new fuzzy time series model based on robust clustering for forecasting of air pollution. *Ecol. Inform.* **2018**, *43*, 157–164.
14. Qianying, W.; Kexin, Y. Prediction of $PM_{2.5}$ concentration based on LSTM-SVR hybrid model. *Inf. Technol. Informatiz.* **2021**, *9*, 33–36.
15. Li, Y.; Liu, Z.; Liu, H. A novel ensemble reinforcement learning gated unit model for daily $PM_{2.5}$ forecasting. *Air Qual. Atmos. Health* **2021**, *14*, 443–453. [CrossRef]
16. Zhao, F.; Liang, Z.; Zhang, Q.; Seng, D.; Chen, X. Research on $PM_{2.5}$ Spatiotemporal Forecasting Model Based on LSTM Neural Network. *Comput. Intell. Neurosci.* **2021**, *2021*, 1616806. [CrossRef]
17. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099. [CrossRef]
18. Weijian, H.; Danyang, L.; Yuan, H. Long-term prediction of $PM_{2.5}$ concentration based on deep learning. *Appl. Res. Comput.* **2021**, *38*, 1809–1814.
19. Zhou, Y.; Chang, F.J.; Chang, L.C.; Kao, I.F.; Wang, Y.S. Exploring a deep learning multi-output neural network for regional multi step-ahead air quality forecasts. *J. Clean. Prod.* **2019**, *209*, 134–145. [CrossRef]
20. Zhu, J.; Deng, F.; Zhao, J.; Zheng, H. Attention-based parallel networks (APNet) for $PM_{2.5}$ spatiotemporal prediction. *Sci. Total Environ.* **2021**, *769*, 145082. [CrossRef]
21. Li, G.; Chen, L.; Yang, H. Prediction of $PM_{2.5}$ concentration based on improved secondary decomposition and CSA-KELM. *Atmos. Pollut. Res.* **2022**, *13*, 101455. [CrossRef]
22. Hu, S.; Liu, P.; Qiao, Y.; Wang, Q.; Zhang, Y.; Yang, Y. $PM_{2.5}$ concentration prediction based on WD-SA-LSTM-BP model: A case study of Nanjing city. *Environ. Sci. Pollut. Res.* **2022**, 1–17. [CrossRef]
23. Ma, J.; Sun, Z. Mutual information is copula entropy. *Tsinghua Sci. Technol.* **2011**, *16*, 51–54. [CrossRef]
24. Jiamei, J.; Kaikai, C.; Zhexiang, W. Improved particle swarm optimization BP neural network for $PM_{2.5}$ prediction. *Comput. Eng. Des.* **2021**, *42*, 3498–3501.
25. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, NY, USA, 2007.
26. Joe, H. *Dependence Modeling with Copulas*; Chapman and Hall/CRC: London, UK, 2014.
27. Thomas, M.T.; Joy, A.T. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2012.
28. Srinivas, M.; Patnaik, L.M. Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Trans. Syst. Man Cybern.* **2002**, *24*, 656–667. [CrossRef]