Prediction Of Air Quality Index Based On Support Vector Machine And Improved Butterfly Optimization Algorithm

Abstract: Air pollution is an increasingly serious problem. Air pollution has been become a main cause of environmental degradation and health effects. Accurate prediction of air quality can help improve environmental quality and human living conditions. The traditional methods for predicting the air quality index have the problem of low accuracy and efficiency. To solve this problem, this paper proposes a novel support vector machine prediction model based on improved butterfly optimization algorithm, which is called IBOA-SVM model. In the improved butterfly optimization algorithm, the sigmoid function is used to optimize the update of the parameter *c*, which increases the search diversity and improves the convergence speed. The performance of the improved butterfly optimization algorithm is verified using eight benchmark functions. Compared with performances of the traditional butterfly optimization algorithm and the particle swarm optimization algorithm, the improved butterfly optimization algorithm has strong competitiveness in accuracy and stability. We establish the IBOA-SVM prediction model for forecasting air quality based on the improved butterfly optimization algorithm. The performance of our proposed model is compared with other predicting models. The experimental results show that our proposed model has higher accuracy and efficiency in predicting the air quality index of four cities in southern China.

Keywords: Air pollution, Butterfly optimization algorithm, Support vector machine, Sigmoid function, Air quality index prediction

1 Introduction

With the continuous progress of modernization and the rapid development of industry and agriculture in China, our living standard is improving, but along with it comes the frequent occurrence of urban air pollution problems, which have many adverse effects on our health and quality of life. The air quality index (AQI) is a dimensionless indicator. It describes the degree of air pollution quantitatively. Ahigher value of AQI indicates more severe air pollution. The monitoring targets of air pollutants in China at this stage are mainly $PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3, etc., and the AQI can quantitatively describe the degree of air pollution by comprehensively evaluating these six pollutants. The AQI is currently an important indicator for confirming the state of air quality. According to the data in the 2019 China Ecological Environment State Bulletin released by the Ministry of Ecology and Environment, more than half of urban air quality does not meet the standards in China.So predicting and analyzing the AQI accurately to prevent and control air pollution in cities with substandard air quality has become

one of the important tasks of the relevant departments.

Due to the problems of low accuracy and efficiency of traditional air quality index prediction methods, traditional prediction models have been gradually losing competitiveness. In the era of artificial intelligence (AI) and big data popularity, many researchersand related technicians have developed some AQI prediction methods based on machinelearning algorithms. Ravindiran et al. used several machine learning models to predict the AQI. The experimental results indicate that machine learning models exhibit superior performance andthe catboost model is the best-performing model for AQI prediction [1]. Xiang et al. predicted the air quality index with SLR, SVR, RF, and probabilistic voting ensemble methods. The results suggest that the probabilistic voting ensemble performs the best, and the ensemble learning method composed of the above classifiers is also an efficient method to predict the AQI [2]. Aiming to predict the AQI more accurately, Wu et al. proposed an ISSA-LSTM model that combines three machine learning methods. The results indicate that the prediction accuracy can be improved by finding the variables with strong correlation with AQI [3]. SigamaniandVenkatesan proposed a multiple linear regressive (MLR) model.This model combines the correlation of two time series–dependent variables and meteorological parameters. The experimental results indicate that the model shows better performance in predicting AQI [4].Castelli et al.usedsupport vector regression (SVR) to predict hourly pollutant concentrations. The results indicate that the prediction results of the SVR model with RBF kernel are more accurate [5]. Song et al.proposed a long short-term memory (LSTM) neural network model based on theimproved jellyfish search optimizer (IJSO). Compared with other meta-heuristics algorithms, the IJSO optimizes parameters better. The results reflect that the IJSO-LSTM modelshows superior performance in the AQI prediction of Chengdu [6]. Parthiban et al.combined the EISAE-DL modelwith deep transfer learningand enhanced deep transfer learning, respectively. EISAE-DTL and EISAE-EDTL models are proposed to forecast the air quality. The experimental results show that two models have higher accuracy compared to the existing models [7]. Gao et al.proposed a difference fusion analysis model to predict AQI. Compared the other three machine learning models, the difference fusion analysis model has the smallest MSE and the highest accuracy [8].Mottahedin et al. introduced an innovative approach in environmental engineering, employing artificial intelligence techniques to forecast air quality in Semnan, Iran. Then, several machine learning (ML) models were rigorously evaluated for their performance, and a detailed analysis was conducted. By incorporating these advanced technologies, the study aims to create a reliable framework for air quality prediction. The results indicate that the adaptive neurofuzzy inference system (ANFIS) is the most effective method for predicting air quality across different seasons, showing high reliability across all datasets [9].Guo et al.proposed a new model for daily AQI prediction, termed VMD-CSA-CNN-LSTM, which employed advanced machine learning techniques, including convolutional neural networks (CNN) and long short-term memory (LSTM) networks, and leveraged the chameleon swarm algorithm (CSA) for hyperparameter optimization, integrated through a variational mode decomposition approach. This study not only advances the predictive accuracy of AQI models but also aids policymakers by providing a reliable tool for air quality management and strategic planning aimed at pollution reduction [10].

Many methods have been proposed for AQI prediction, however, there is still room for

further exploration of prediction models in accuracy. For the purpose of achieving higher accuracy, based on the relevant research of the predecessors, a novel support vector machine prediction model based on improved butterfly optimization algorithm (IBOM-SVM) is proposed in this paper. The sigmoid function is used to optimize the update of the parameter c in the butterfly optimization algorithm (BOA).Then the improved butterfly optimization algorithm combined with support vector machine (SVM) is used to predict the AQI in four cities in the South. The experimental results show the parameters of SVM obviously affect its performance. So the prediction accuracy of SVM can be improved by optimizing the parameters. The methods for optimizing parameters of SVM are grid search, gradient descent, and meta-heuristic algorithms. Meta-heuristic algorithms show superior results in optimizing the parameters of complex models. For example, genetic algorithm (GA) [11], artificial bee colony (ABC) [12], particle swarm optimization (PSO) [13], ant colony algorithm (ACA) [14], grey wolf optimizer (GWO) [15], moth-flame optimization (MFO) [16], sine cosine algorithm (SCA) [17],salp swarm algorithm (SSA) [18], grasshopper optimization algorithm (GOA) [19] and a new population-based optimization algorithm, BOA [20]. In this paper, an improved butterfly optimization algorithm (IBOA) is proposed. The performance of IBOA is tested on eight benchmark functions. The IBOA algorithm shows superior results in finding the optimal value compared with the traditional BOA algorithm. The IBOA algorithm has faster convergence speed and higher accuracy. Then, we propose IBOA-SVM model. According to the air quality index data, we use the IBOA-SVM model to predict the AQI of four cities in southern China. The comparison experiments of IBOA-SVM model, particle swarm optimization support vector machine (PSO-SVM) model and BOA-SVM model are done. The experimental results show that IBOA-SVM mode shows superior results in predicting the AQI.

The rest of this paperis organized as follows. Section 2 introduces the BOA algorithm and proposes the IBOA algorithm and discusses the performance of the IBOA algorithm. In section 3, we propose the IBOA-SVM model and give the main framework of IBOA-SVM. We use the proposed IBOA-SVM model to predict the air quality index of four cities in southern China in section 4. We analyze the experimental data and elaborate on the details of the experiments. Our proposed model is tested on datasets and compared with other models and the relevant experimental results are discussed. In section 5 we present our conclusions.

2 Improved butterfly optimization algorithm

In this section, we first introduce the BOA algorithm. Then, in response to the existingproblems of the BOA algorithm, we propose a novel improved BOA algorithm. At the same time, we compare the performance of the improved BOA algorithm with the traditional BOA algorithm and other optimization algorithms.

2.1 Butterfly optimization algorithm

In nature, butterflies have their special foraging process, that is, in the process of searching for food, butterflies will produce and perceive the fragrance. The closer to the food source, the greater the concentration of fragrance. According to this characteristic, butterflies can find food quickly and accurately. Inspired by the above process, Arora and Singh proposeda meta-heuristic

algorithm known asthe BOA algorithm [20]. The algorithm regards the butterfly as the solution and the food as the optimal solution. The process of the butterfly finding food is the process of the algorithm finding the best solution through continuous iteration. In each iteration, the butterfly generates a fragrance based on its location, which is associated with its fitness.At the same time, the butterfly can sense the fragrance of the surrounding butterflies, and then conduct a search operation. The search behavior of butterfly is divided into two kinds. One is to move towards the butterfly with the highest fragrance concentration, known as global search.The other is random movement, known as local search.Where aswitching probability p is introduced to balance global search and local search.The parameters of the butterfly optimization algorithm mainly include the stimulus intensity I , the sensory modality c and the power exponent a . The formula for calculating the fragrance is as follows:

$$f = cI^a, \tag{1}$$

where f is the fragrance concentration, c is the sensory modality, I is the stimulus intensity and a is the power exponent.

In the global search phase, the search operator is close to the optimal solution, which can be represented by the following update equation:

$$x_i^{t+1} = x_i^t + \left(r^2 \times g^* - x_i^t\right) f_i, \tag{2}$$

where x_i^t is the solution vector for the ith butterfly in the iteration number t and g^* denotes the optimal solution in the current iteration. f_i denotes the fragrance concentration of the ith butterfly. r denotes a random number between [0,1].

The local search phase can be represented by the following update equation:

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) f_i, \tag{3}$$

where x_j^t and x_k^t are the j th and k th butterflies. r denotes the random number between [0,1] and the above equation also denotes the local random walking of the butterfly.

The process to find the optimal solution by BOA can be divided into the following three stages:

(1) Initialization phase

At this phase, the algorithm first randomly generates N initial solutions, and defines the search space and fitness function.Then set the initial value of c , a , p , and the maximum number of iterations T . Next, the fragrance concentration and the fitness value are calculated, and the optimal solution is found.

(2) Iterative phase

The solution vector updates its position according to the update equations (2) and (3). The solution vector selects global search or local search depends on the switching probability p and the random number r . When $r < p$, the global search is selected. when $r \geq p$, selecting local

search. When the solution vector moves to a new position, the algorithm recalculates the fragrance concentration and the fitness value of the solution and finds out the optimal solution for the current iteration.

(3) Final Phase

The algorithm continues to iterate. As the iteration proceeds, each solution constantly approaches the optimal solution. When the algorithm reaches the maximum number of iterations, the optimal solution is output.

2.2 Improved butterfly optimization algorithm

The BOA is widely used to solve various optimization problems as an emergingalgorithm.Ithas the characteristics of easy operation, fast convergence speed, and high optimization accuracy. However, like other optimization algorithms, butterfly optimization algorithm is also easyto fall into local optima. Soresearchers propose many improvementsto this algorithm. The improved BOA algorithm shows significant optimization performancein dealing with engineering problems and prediction problems.Li et al.proposed an enhanced version of the BOA algorithm based on the cross-entropy [21]. The co-evolution technique is also applied to this algorithm, and its remarkable performance is verified with three engineering optimization problems. With the purpose of solving the problem of low prediction accuracy of traditional models, an improved butterfly algorithm optimizing support vector machine for soil water content prediction is proposed by Wang et al.based on levy flight, gaussian mutation, and chaotic mapping, which can enhance population diversity and improve optimization ability.This method achieves good results in soil water content prediction [22].

However, there is still much room for improvement in the parameters of the BOA algorithm. In BOA, stimulus intensity I is related to fitness. When the optimization problem is finding the minimum value, I is inversely proportional to fitness. When the optimization problem is to find the maximum value, I is proportional to fitness. It is important to choose a suitable value of I. The parameter a determines the fragrance absorption of the butterfly, and the value of c affects the movement distance of the butterfly. It can be seen from the update equations (2) and (3)that the moving step size is proportional to the value of c. The c-value plays an important role in affecting the convergence speed and accuracy of the algorithm, so the ideal optimization should have a suitable value of c that makes the search factor approach the optimal solution quickly and accurately. In BOA, the value of c is fixed to 0.01, and the fixed value of c is insufficient to produce discriminable results affecting fitness. Therefore, this paper improves the adaptability of the BOA by introducing a dynamic c-value. The dynamic c-value can make the search factor approach the optimal solution quickly.

In order to reach our desired ideal state, this paper proposes a c-value growth model based on the sigmoid function. This model can effectively improve the convergence speed of the algorithm by controlling the growth change of the value of c. In the modified c-value growth model, the value of c can be expressed as:

$$c^{t+1} = \frac{1}{1+e^{-t}}. \tag{4}$$

As can be seen from the above equation, the growth rate of c-value increases faster when

the number of iterations is smaller and slower when the number of iterations is larger, which is in line with the adaptive situation of the algorithm in the search process. When the value of c changes faster, the algorithm has faster convergence speed. When the value of c changes slower, the search diversity can be increased. Based on the sigmoid function-based c-value growth model, the improved BOA algorithm can increase the effectiveness of the butterfly in searching for optimal values in both local and global search.

2.3 Performance analysis of the improved butterfly optimization algorithm

To demonstrate the superiority of IBOA, we conduct simulation experiments of the IBOA, BOA, and PSO algorithms based on eight benchmark functions. From the experimental results, it can be seen that the IBOA has higher prediction accuracy and stronger robustness.During the testing experiment, we rely on experience to choose the initial value of c is 0.01, p is 0.8, a is 0.1, the initial population size generated is 30, and T is 200. The three algorithms are run 10 times, and take the average as the optimal solution. The expression of the benchmark function used in this article is shown in Table 1.

Table 1 Standard benchmark functions used in this study

	Benchmark functions	Equation	Dim	Range	Optima
f_1	Beale	$f(x)=(1.5-x_1+x_1x_2)^2+(2.25-x_1+x_1x_2^2)^2+(2.625-x_1+x_1x_2^3)^2$	2	[-4.5,4.5]	0
f_2	Quartic	$f(x)=\sum_{i=1}^{n} ix_i^4+rand(0,1)$	30	[-1.28,1.28]	0
f_3	Schaffer	$f(x)=(x_1^2+x_2^2)^{0.25}\cdot\left[\left(50(x_1^2+x_2^2)^{0.1}+1\right)\right]$	2	[-100,100]	0
f_4	Booth	$f(x)=(x_1+2x_2-7)^2+(2x_1+x_2-5)$	2	[-10,10]	0
f_5	Matyas	$f(x)=0.26(x_1^2+x_2^2)-0.48x_1x_2$	2	[-10,10]	0
f_6	Zattl	$f(x)=(x_1^2+x_2^2-2x_1)^2+0.25x_1$	2	[-1,5]	-0.00379
f_7	Rastrigin	$f(x)=\sum_{i=1}^{n}(x_i^2-10\cos(2\pi x_i)+10)$	30	[-5.12,5.12]	0

f_8	Sphere	$f(x)=\sum_{i=1}^{n} x_i^2$	30	[-100,100]	0

To make the results more credible, we select benchmark functions with different dimensions. We solve the optimal value of these eight benchmark functions using the above three algorithms. The optimal solutions obtained by the three optimization algorithms under different benchmark functions are given in Table 2. As shown in Table 2, for the function $f_1 \sim f_8$, the optimal value obtained by the IBOA algorithm proposed in this paper is the smallest and the solution accuracy is the highest. The optimal solutions obtained by IBOA for all the benchmark functions are superior to the solutions derived by the original BOA, which indicates that the IBOA algorithm has significant superiority. To further validate the effectiveness of the algorithm, the convergence plots of the three algorithms on these benchmark functions are also given below. The convergence curves of the IBOA algorithm are compared with the BOA algorithm and PSO algorithm in Figure 1. As shown in Figure 1, the IBOA algorithm can approach the optimal solution more quickly and accurately. Thisindicates that the dynamic c -valuebased on the sigmoid function can effectively prevent the algorithm from falling into local optima. The results alsoindicate that the sigmoid function-based c -value growth model can enhance population diversity and improve optimization abilityeffectively, which achieves the expected effect. Compared with BOA, the improved algorithm shows significant superiority.

Table 2 Simulation results of different algorithms

	BOA	PSO	IBOA
f_1	6.06E-01	3.75E-02	1.33E-06
f_2	3.28E-04	1.69E-04	2.99E-06
f_3	1.19E+01	2.17E-02	1.42E-11
f_4	9.82E-02	2.18E-02	3.88E-05
f_5	9.99E-03	7.52E-18	5.94E-19
f_6	7.75E-04	-3.79E-03	-3.79E-03
f_7	1.64E-08	0.00E+00	0.00E+00
f_8	1.78E-10	1.14E-22	1.45E-23

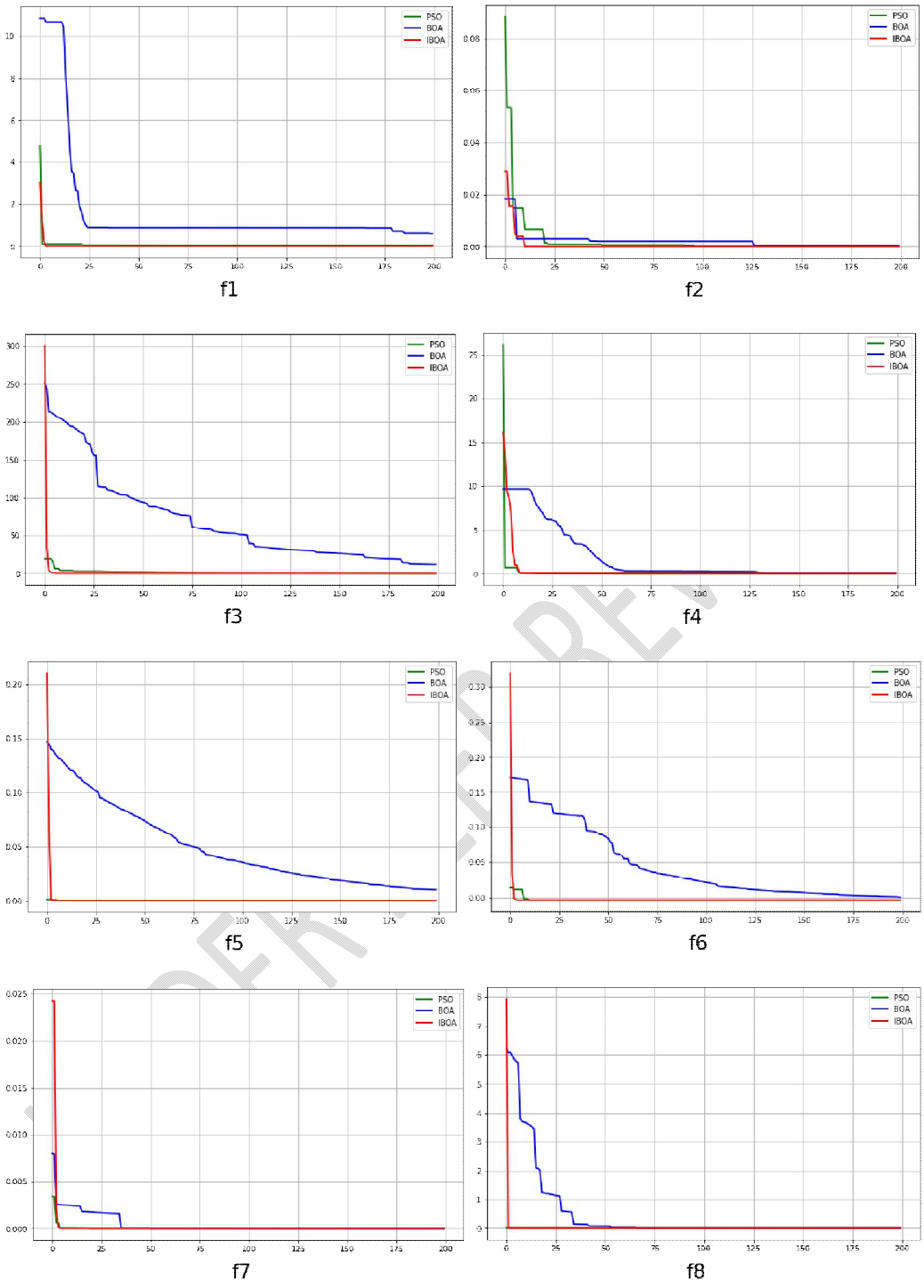


Figure 1 Convergence curves of different algorithms

3 Proposed IBOA-SVM model

3.1 Support vector machine

The support vector machine (SVM) is a machine learning algorithm first proposed by Cortes

and Vapnik for classification problems [23].Later, by introducing a ε -insensitive loss function, a regression method is obtained, namely support vector regression (SVR). SVR is a kernel-based regression method that finds a hyperplane that minimizes the structural risk. Assuming a given data set $\{(x_i, y_i)\}_1^n$, (x_i, y_i)as a sample point, SVR is to find a hyperplane that minimizes the distance from the farthest sample point to the hyperplane.The function of SVR is expressed as:

$$f(x) = w^T \phi(x) + b \ , \qquad (5)$$

where ϕ is a nonlinear mapping function, w is the weight coefficients, and b is a constant. The SVR model is transformed into solving the following optimization problems:

$$\min \frac{1}{2}\|w^2\|,$$
$$s.t.: \begin{cases} y_i - w^T \phi(x_i) - b \le \varepsilon \\ y_i - w^T \phi(x_i) - b \ge -\varepsilon \end{cases}. \qquad (6)$$

To make the prediction results of the SVR model more stable, the slack variables ξ, ξ^* are introduced to optimize the SVR model. The revised optimization problem is expressed as follows:

$$Min \frac{1}{2}\|w^2\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$
$$s.t.: \begin{cases} y_i - w^T \phi(x_i) - b \le \varepsilon + \xi_i \\ w^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1,2,...n \end{cases}, \qquad (7)$$

where C is called the penalty factor, ε is the insensitive loss parameter, and ξ, ξ^* are the slack variables.

Using the Lagrange multiplier method and optimality conditions, we find the solution to the above problem by solving its dual, and finally obtain the following expression of $f(x)$:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \, , \qquad (8)$$

where α_i, α_i^* is the Lagrange multiplier, $K(x_i, x)$ is the kernel function, the expression of $K(x_i, x)$is expressed as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \qquad (9)$$

where γ is a kernel parameter.

3.2 Basic principles of the IBOA-SVM model

The parameters C and γ have a significant impact on the performance of SVM, so the prediction accuracy of SVM can be improved by finding better parameter values [24-26]. In this paper, we use the IBOA algorithm to search optimal parameters of SVM, and propose a prediction model based on the IBOA algorithm. First, we set the parameters of IBOA, including the number of populations and the value of T. Then the algorithm generates a set of initial solutions. Each solution is represented by a two-dimensional vector (C,γ). The IBOA algorithm needs to find an optimal solution through continuous iteration, and output the optimal solution (C^*,γ^*) when the algorithm reaches the maximum number of iterations. The optimal solution (C^*,γ^*) is used in the SVM model. The output of the SVM model is used as a fitness function of the IBOA algorithm. The step-wise procedure of the IBOA-SVM algorithm is as follows:

Step 1: Set the parameters of the IBOA-SVM algorithm, including $N=10$, $T=50$, $p=0.8$, $c=0.01$, $a=0.1$, and the range of the parameters C and γ.

Step 2: Initialize the search agents according to the following equation:

$$s_p = lb_p + (ub_p - lb_p)\cdot u, \tag{10}$$

where lb_p, ub_p are the lower and upper bounds of the solution. u denotes a random number between (0,1).

Step 3: 70% of the original data is divided into a training set and 30% into a test set, and the mean squared error (MSE) is used to calculate the individual fitness.The smallest fitness value is taken as the optimal solution, and then calculate the individual scent. In this paper, the MSE and the mean absolute error (MAE) are also used to evaluate the accuracy of the model. The formula for MSE and MAE is represented as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2, \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - f_i|. \tag{12}$$

where y_i and f_i denote the true and predicted values of the ith data point, respectively, and N denotes the total number of data points.

Step 4: Update the individual position with equations (2) and (3). Update the value of c with equation (4) and continue iterating until the algorithm reaches the maximum number of iterations..

Step 5: The IBOA algorithm obtains the optimal solution.Thenwe build the SVM model using the optimal parameters.

The flowchart of the IBOA-SVM model is shown in Figure 2.

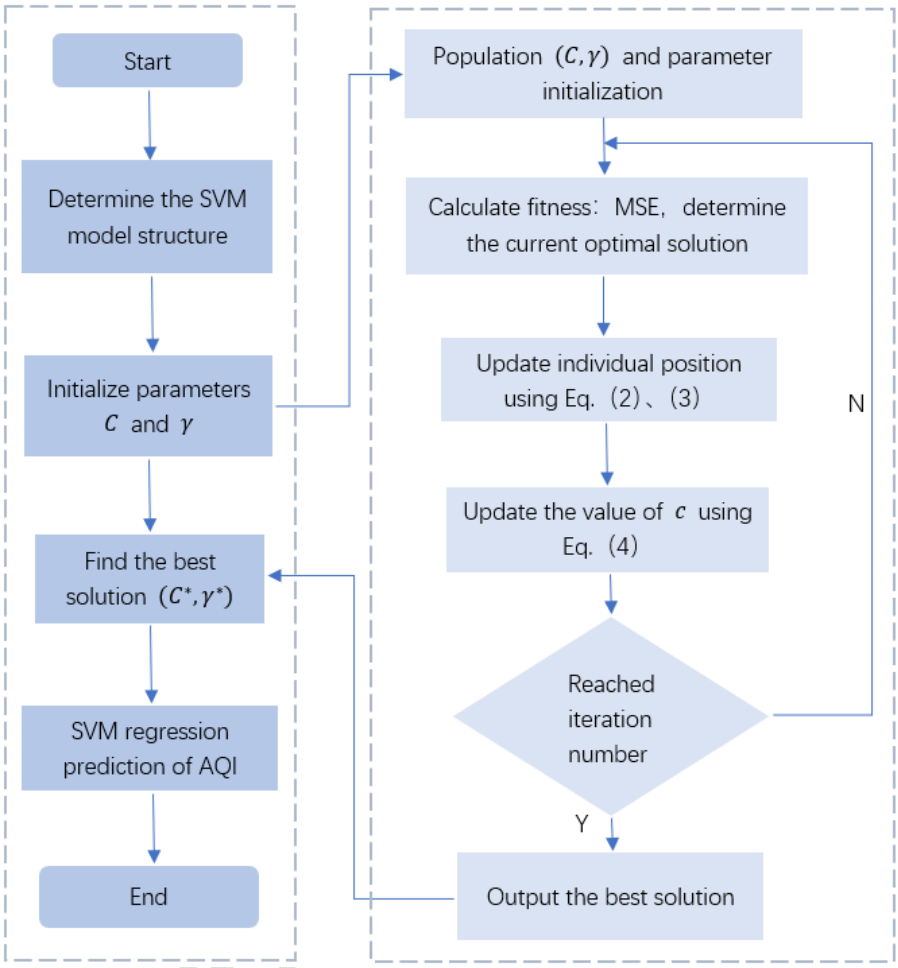


Figure 2 Flow-process diagram of IBOA-SVM

4 Experiment and discussion

4.1 Data analysis

In this paper, we select the monthly AQI and monthly average mass concentration data of six pollutant gases in four citiesin southern China from December 2013 to June 2023 for the experiments. The data contain the monthly average values of $PM_{2.5}$ ($\mu g/m^3$), PM_{10} ($\mu g/m^3$), SO_2 ($\mu g/m^3$), NO_2 ($\mu g/m^3$), CO (mg/m^3), O_3 ($\mu g/m^3$), and AQI. The data are obtained from the China Air Quality Online Monitoring and Analyzing Platform. The monthly trend of the AQI in four cities from December 2013 to June 2023 is shown in Figure 3. The darker the color in Figure 3, the more severe the pollution. $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 are chosen as the input vectors of the prediction model.

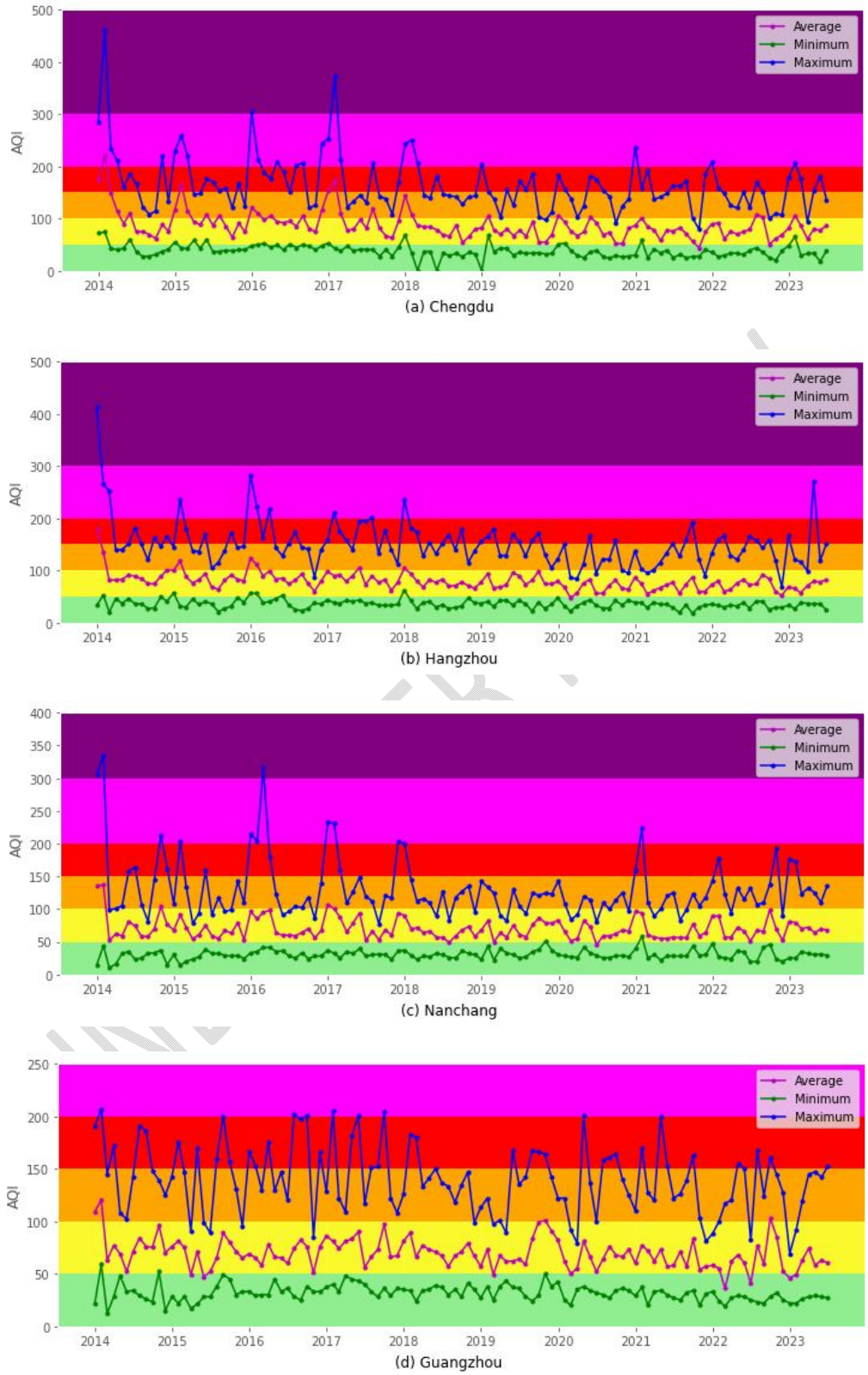


Figure 3 Monthly trend chart of AQI

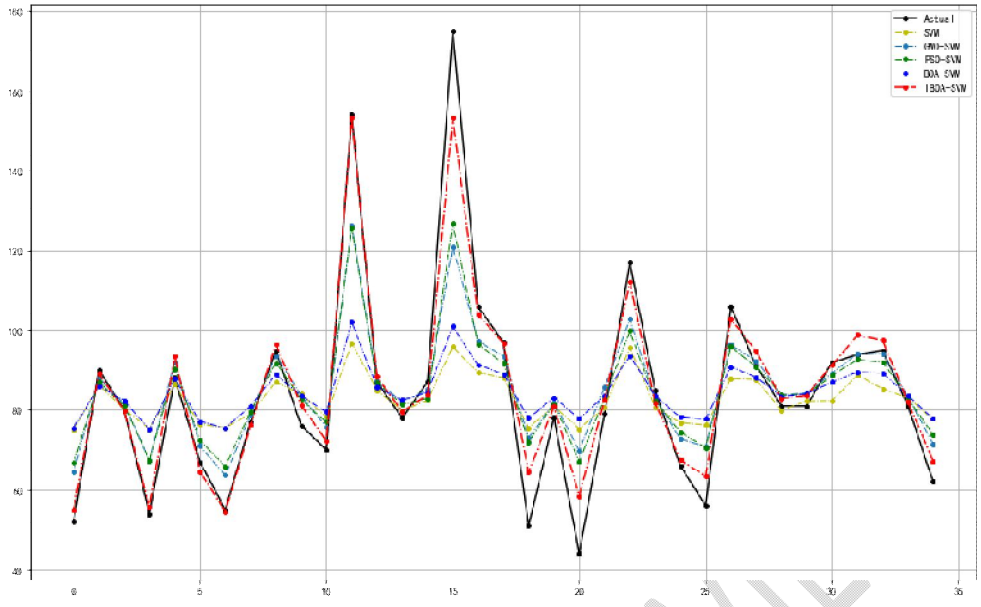
4.2 Comparison of experimental results

This paper aims to predict the AQIof four cities in southernChina more accurately, so in order to see the prediction effect of the model conveniently, we use the evaluation index MSE and MAE to quantitatively analyze the accuracy of prediction results. The smaller the values of MSE and MAE, the closer the predicted value is to the true value. This indicates that the prediction results of the model are more accurate. To verify the superiority of the IBOA-SVM model proposed in this article, we use SVM, GWO-SVM, PSO-SVM, and BOA-SVM models as comparative models.The paper uses these five models to conduct prediction experiments of the AQIof four cities in southernChina. All the experiments are carried out in the Python 3.9.12 environment. The experimental results are given below in Table 3, Figure 4, and Figure 5.

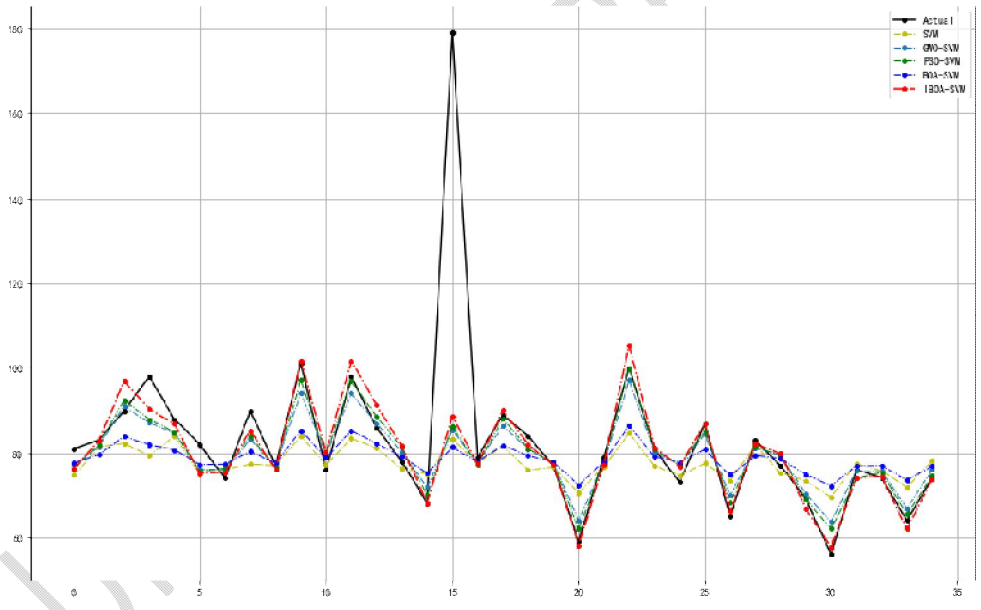
Table 3 MSE and MAE of different prediction models

		SVM	GWO-SVM	PSO-SVM	BOA-SVM	IBOA-SVM
MSE	Chengdu	428.07	294.99	288.34	423.58	**33.90**
	Hangzhou	325.29	299.40	276.68	341.14	**243.34**
	Nanchang	254.65	82.52	134.18	288.73	**27.65**
	Guangzhou	155.87	112.66	60.18	161.77	**34.50**
MAE	Chengdu	13.12	9.09	9.35	13.89	**3.83**
	Hangzhou	8.79	6.43	6.21	9.15	**4.90**
	Nanchang	10.63	4.74	6.59	11.29	**3.25**
	Guangzhou	9.39	6.98	5.15	8.73	**4.22**

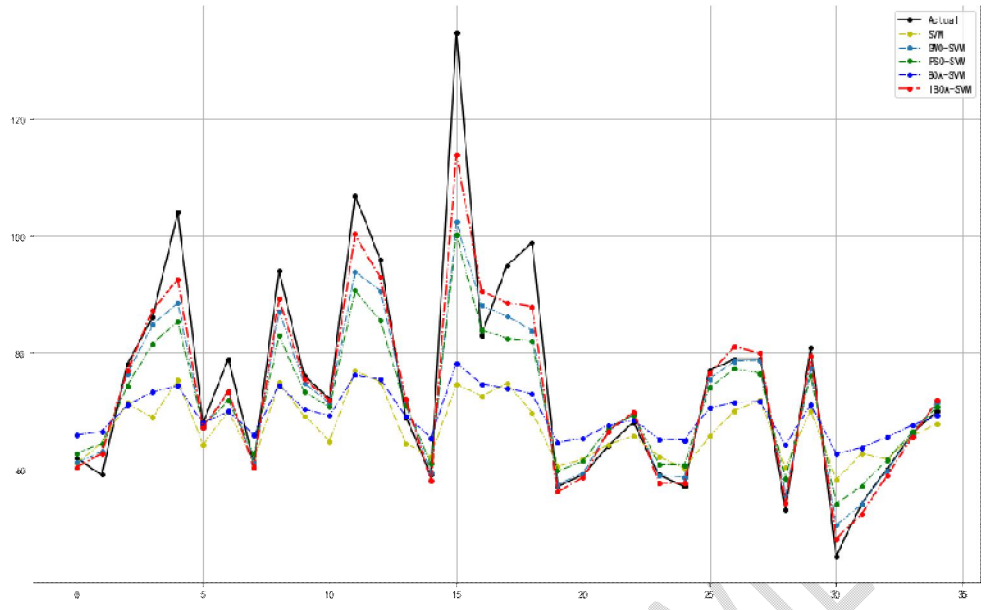
The MSE and MAE ofthe fiveprediction models are given in Table 3. The values of MSE and MAE are the average of each models run 10 times in different cities. Taking Chengdu City as an example, the MSE of IBOA-SVM model is 33.9, and the MSE of SVM, GWO-SVM, PSO-SVM, and BOA-SVM are 428.07, 294.99, 288.34, 423.58 respectively. It is easy to see that the IBOA-SVM model obtained the smallest mean squared error and had the best prediction effect. It also indicates that GWO-SVM, PSO-SVM, and BOA-SVM models are easy to fall into local optimality, while the IBOA-SVM model can effectively avoid this problem. Similarly, the IBOA-SVM model also obtained the smallest values of MSE and MAE in the other three cities. This fully demonstrates the superior performance of the new prediction model.



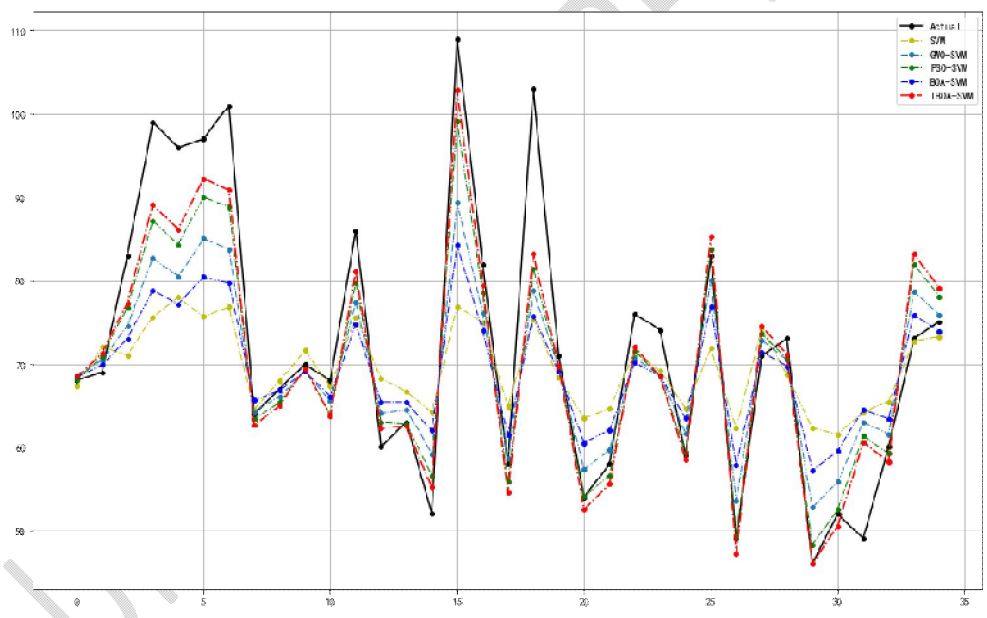
(a) Chengdu



(b) Hangzhou



(c) Nanchang



(d) Guangzhou

Figure 4 Comparison results of different prediction models

Figure 4 shows the comparison results between the real and predicted values of the AQI for four cities, the solid line represents the real value and the dashed line represents the predicted value. The predictioncurves of the five prediction models are given in Figure 4.From the experimental results, the prediction result of the IBOA-SVM model is closer to the true value, the dashed line and the solid line have the highest degree of overlap, and the predicted value and the true value of some sample points basically overlap, which indicates that the prediction accuracy of

the IBOA-SVM model proposed in this paper is the highest in the three models. The experimental results also indicate that the optimization performance of the IBOA algorithm is effectively improved. And the IBOA-SVM model is more feasible in predicting the AQI of four cities in southern China.

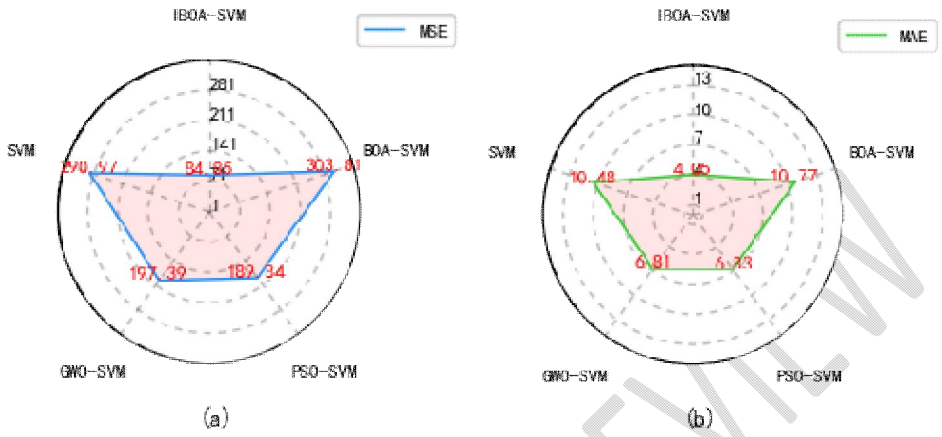


Figure 5Predicting errors of different prediction models

In order to further demonstrate its comprehensive performance, the average MSE and MAE of the IBOA-ISVM model and the other four comparison models in the four municipalities are shown in Figure 5. As can be seen from Figure 5(a), the average MSE of the IBOA-SVM model is 84.85, and the average MSE of BOA-SVM, PSO-SVM, GWO-SVM, and SVM are 303.81, 189.84, 197.39, 290.97 respectively. The IBOA-SVM model has the smallest average MSE, and the experimental results indicate that the IBOA-SVM model has higher prediction accuracy and stronger robustness thanconventional machine learning models. Figure 5(b) shows the IBOA-SVM model has the lowest value of average MAE compared to the other four models. The experimental results show that this model has stronger stability and superior performance.

5 Conclusions

In this paper, an improved butterfly optimization algorithm based on the sigmoid function to update the sensory modality c is proposed. The improved algorithm achieves an ideal state of the algorithm's optimization seeking effect by controlling convergence speed. The experimental results show that compared with the traditional BOA algorithm and PSO algorithm, the improved algorithm can significantly improve the accuracy and convergence speed of the algorithm. We proposed the IBOA-SVM prediction model based on the improved BOA algorithm and SVM. The IBOA-SVM model is used to predict the AQIoffour cities in southern China in recent years. Compared with the prediction results of the SVM model, the PSO-SVM model, the GWO-SVM model, and the BOA-SVM model, the IBOA-SVM model has better performance. Our research is of great significance for predicting air quality and controlling air pollution.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models(ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Data Availability

Data will be made available on request.

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

References

[1] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., Sonne C., 2023.Air quality prediction by machine learning models: A predictive study on the indian coastal city of Vishakhapatnam. Chemosphere. 338, 139518.https://doi.org/10.1016/j.chemosphere.2023.139518.

[2] Xiang, X., Fahad, S., Han, M.S., Naeem, M.R., Room, S., 2022. Air quality index prediction via multi-task machine learning technique: spatial analysis for human capital and intensive air quality monitoring stations. Air Quality, Atmosphere & Health. 16, 85-97.https://doi.org/10.1007/s11869-022-01255-3.

[3] Wu, H., Yang, T., Li, H., Zhou, Z., 2023. Air quality prediction model based on mRMR-RF feature selection and ISSA-LSTM. Scientific reports. 13, 12825.https://doi.org/10.1038/s41598-023-39838-4.

[4] Sigamani, S., Venkatesan, R., 2022. Air quality index prediction with influence of meteorological parameters using machine learning model for IoT application. Arabian Journal of Geosciences. 15, 340.https://doi.org/10.1007/s12517-022-09578-2.

[5] Castelli, M., Clemente, F.M., Popovic, A., Silva, S., Vanneschi, L., 2020. A Machine Learning Approach to Predict Air Quality in California. Complexity.23, 8049504. https://doi.org/10.1155/2020/8049504.

[6] Song, Q., Zou, J., Xu, M., Xi, M., Zhou, Z., 2023. Air quality prediction for Chengdu based on long short-term memory neural network with improved jellyfish search optimizer. Environmental Science and Pollution Research. 30, 64416-64442.https://doi.org/10.1007/s11356-023-26782-z.

[7] Parthiban, S., Amudha, P., Sivakumari, S.P., 2022. Exploitation of Advanced Deep Learning Methods and Feature Modeling for Air Quality Prediction. Revue d'Intelligence Artificielle. 36(6), 959-967.https://doi.org/10.18280/ria.360618.

[8] Gao, S., He, Z., Liu, Z.,Liu, J., Wang, G., Li, D., 2021. Application of difference fusion analysis based on machine learning in air quality prediction. Electronic Measurement Technology. 44(18), 85-92.10.19651/j.cnki.emt.2107055.

[9] Mottahedin, P., Chahkandi, B., Moezzi, R., Fathollahi-Fard, A.M., Ghandali, M., Gheibi, M., 2024. Air quality prediction and control systems using machine learning and adaptive

neuro-fuzzy inference system. Heliyon. 10(21), e39783-e39783. 10.1016/j.heliyon.2024.e39783

[10]Guo, Z., Jing, X., Ling, Y., Yang, Y., Jing, N., Yuan, R., Liu, Y., 2024. Optimized air quality management based on air quality index prediction and air pollutants identifcation in representative cities in China. Scientifc Reports. 14(1), 17923-17923. 10.1038/S41598-024-68972-W.

[11] Rajeev,B.S., Krishnamoorthy, C.S., 1997. Genetic algorithm-based methodologies for design optimization of trusses. Journal of Structural Engineering. 123(3), 350-358. https://doi.org/10.1061/(ASCE)0733-9445(1997)123:3(350).

[12] Karaboga, D., Basturk, B., 2007. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. Journal of Global Optimization. 39, 459-471.https://doi.org/10.1007/s10898-007-9149-x.

[13] Kennedy, J., 2010. Particle swarm optimization. Encyclopedia of machine learning,Springer-Verlag, New York. 760-766.

[14] Dorigo, M., Birattari, M., 2010. Ant colony optimization. In: Sammut C, Webb GI (Eds) Encyclopedia of machine learning. Springer US, Boston MA, pp. 36–39.

[15] Mirjalili, S., 2015. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems. 89, 228-249. https://doi.org/10.1016/j.knosys.2015.07.006.

[16] Mirjalili, S., 2016. SCA: A Sine Cosine Algorithm for solving optimization problems. Knowledge-Based Systems. 96:120-133. https://doi.org/10.1016/j.knosys.2015.12.022.

[17] Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., Saremi, S., Faris, H., Mirjalili, S.M., 2017. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software. 114, 163-191. https://doi.org/10.1016/j.advengsoft.2017.07.002.

[18] Mirjalili, S., Mirjalili, S.M., Lewis, A., 2014. Grey wolf optimizer. Advances in Engineering Software. 69, 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007.

[19] Saremi, S., Mirjalili, S., Lewis, A., 2017. Grasshopper Optimisation Algorithm: Theory and application. Advances in Engineering Software. 105, 30-47.https://doi.org/10.1016/j.advengsoft.2017.01.004.

[20] Arora, S., Singh, S., 2019. Butterfly optimization algorithm: a novel approach for global optimization. Soft Computing. 23, 715-734. https://doi.org/10.1007/s00500-018-3102-4.

[21] Li, G., Shuang, F., Zhao, P., Le, C., 2019. An Improved Butterfly Optimization Algorithm for Engineering Design Problems Using the Cross-Entropy Method. Symmetry. 11, 1-20.https://doi.org/10.3390/sym11081049.

[22] Wang, Z., Liu, Q., 2023.Soil water content prediction model based on improved butterfly algorithm optimizing support vector machine. Computer Engineering and Design. 44(2), 612-621. 10.16208/j.issn1000-7024.2023.02.040.

[23] Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Leaning. 20, 273-297. https://doi.org/10.1023/A:1022627411411.

[24] Hu, K., Jiang, H., Ji, C.-G.,Pan, Z., 2020. A modified butterfly optimization algorithm: An adaptive algorithm for global optimization and the support vector machine. Expert Systems. e12642.https://doi.org/10.1111/exsy.12642.

[25] Bao, H., Liang, G., Cai, Z., Chen, H., 2022. Random Replacement Crisscross Butterfly Optimization Algorithm for Standard Evaluation of Overseas Chinese Associations. Electronics. 11, 1080. https://doi.org/10.3390/electronics11071080.

[26] Wen, L., Cao, Y., 2020. A hybrid intelligent predicting model for exploring household CO2 emissions mitigation strategies derived from butterfly optimization algorithm. Science of the Total Environment. 727, 138572.https://doi.org/10.1016/j.scitotenv.2020.138572.