Environ. Eng. Res. 2017; 22(2): 175-185
https://doi.org/10.4491/eer.2016.125

pISSN 1226-1025
eISSN 2005-968X

Influencing factors and prediction of carbon dioxide emissions using factor analysis and optimized least squares support vector machine

Siwei Wei[1†], Ting Wang[1], Yanbin Li[2]

[1]*School of Economics and Management, North China Electric Power University, Baoding 071000, China*
[2]*School of Economics and Management, North China Electric Power University, Beijing 102206, China*

ABSTRACT

As the energy and environmental problems are increasingly severe, researches about carbon dioxide emissions has aroused widespread concern. The accurate prediction of carbon dioxide emissions is essential for carbon emissions controlling. In this paper, we analyze the relationship between carbon dioxide emissions and influencing factors in a comprehensive way through correlation analysis and regression analysis, achieving the effective screening of key factors from 16 preliminary selected factors including GDP, total population, total energy consumption, power generation, steel production coal consumption, private owned automobile quantity, etc. Then fruit fly algorithm is used to optimize the parameters of least squares support vector machine. And the optimized model is used for prediction, overcoming the blindness of parameter selection in least squares support vector machine and maximizing the training speed and global searching ability accordingly. The results show that the prediction accuracy of carbon dioxide emissions is improved effectively. Besides, we conclude economic and environmental policy implications on the basis of analysis and calculation.

Keywords: Carbon dioxide emissions, Factor analysis, Fruit fly algorithm, Least squares support vector machine, Prediction

1. Introduction

With rapid economic growth, China's energy and environmental problems are very prominent and the economic development presents typical high-carbon characteristic. At present, China is in the middle stage of industrialization. The energy consumption is huge and fossil energy is dominant in energy consumption structure. The US energy information agency data show that China has surpassed the US as the world's largest carbon emitter since 2009 [1]. In addition, China's economic scale is increasing year by year and the energy consumption grows rapidly. So carbon dioxide emissions present a rapid growth trend. This shows that China's carbon emissions reduction is not only a great challenge, but also very urgent [2]. From the perspective of actual situation of resources and environment in China and focusing on the rapid and coordinated development of China's economy [3, 4], it is necessary to study the influencing factors of carbon dioxide emissions to explore the key factors of China's carbon dioxide emissions and forecast future carbon dioxide emissions accurately. Then we can put forward policy implications to promote carbon emission reduction in China, which has very important significance for successful completion of China's 2020 carbon emission reduction targets and promoting the harmonious development of China's economy, resources and environment.

Researchers at home and abroad have spared no efforts to carry out researches about carbon dioxide emissions prediction and constantly come up with advanced theoretical methods to improve forecast accuracy for years. For carbon dioxide prediction, there are two main kinds of research methods. One kind is the trend extrapolation only considering the natural laws of the subject investigated, such as regression analysis and grey prediction, among which GM(1,1) is one of the basic models and has been widely used [5-9]. This kind of methods proceed from the view of the own laws of carbon dioxide emissions. And the future changes of the subject investigated are speculated directly from historical data. Meanwhile, the forecast results are often monotonous and

Received October 5, 2016 **Accepted** December 29, 2016

† Corresponding author
Email: crazysara@126.com
Tel: +86-136-9365-2875 Fax: +86-136-9365-2875

with poor stability which cannot reflect the random variation of carbon dioxide emissions. The other kind is correlating prediction comprehensively considering the key influencing factors to obtain the forecast values of electricity consumption like factor decomposition, input-output analysis, scenario analysis, etc. [10-12]. But it is often difficult to obtain the satisfactory prediction accuracy because of the subjectivity of qualitative analysis or the own limitation of prediction models. For there exists a complex nonlinear relationship between carbon dioxide emissions and the variables that affect it, this kind of methods call for accurate mathematical models in order to obtain satisfactory prediction accuracy. At the same time, some researchers focus on the relationship between carbon emissions and the influencing factors. They quantify the relationship between carbon emissions and driving factors including economic development, energy consumption, industrial structure and so on by means of causal analysis [13, 14], input-output analysis [15], Multivariate Cointegration Analysis [16, 17], etc. And these provide scientific basis for relevant policy-making and carbon emission reduction as well as provide useful references for us.

With the popularization and application of artificial intelligence algorithms [18-20], least squares support vector machine and other intelligence algorithms represented by neural networks has been more and more applied in carbon dioxide emissions forecasting. Back propagation (BP) neural network is the most basic and important one among all kinds of neural networks algorithms. It has the advantages of self adaptation, self-organization, self-learning ability and non convexity, but it is not enough in global search, calculation speed, reliability and so on. While least squares support vector machine overcomes the defects of BP with global searching ability and faster convergence speed, which has been successfully used to solve prediction problems in many fields such as the prediction of gas concentration and wind speed [21, 22]. In addition, the predictive performance of least squares support vector machine is largely determined by the values of the two parameters. At present, some meta heuristic algorithms have been used to determine the proper values of these two parameters including particle swarm optimization, genetic algorithm, artificial bee colony algorithm, etc. However, these optimization algorithms are difficult to understand and it is slow to achieve the global optimal solution. The fruit fly optimization algorithm is a new evolutionary computation and optimization technique which is easy to understand. And it has advantages over other algorithms including simple computation, less parameters, easy adjustment, small amount of calculation, strong global searching ability and searching precision. These advantages make it easily applied to related practical problems [23]. Therefore, we select fruit fly algorithm to optimize the two necessary parameters of least squares support vector machine (LSSVM) model.

In this paper, we propose a carbon dioxide emissions prediction model based on factor analysis and LSSVM optimized by fruit fly algorithm with comprehensive consideration of the factors affecting carbon dioxide emissions. Firstly, we analyze the relationship between carbon dioxide emissions and its influencing factors in a comprehensive way by factor analysis including correlation analysis and regression analysis, achieving the effective screening of key factors. Secondly, we use fruit fly algorithm to optimize the parameters of LSSVM. Finally, the optimized least squares support vector machine is used for prediction. The combined models in this paper overcome the blindness of the parameter selection of LSSVM and maximize the training speed and the global searching ability accordingly. Then the prediction accuracy of carbon dioxide emissions is improved effectively, which is of great significance to carbon dioxide emissions controlling.

2. Methods and Models

2.1. Factor Analysis

Factor analysis is proposed by Spearman in 1904 which is a multi-variable analytical method developed in the field of psychology based on the analysis of relationship between indicators and their influencing factors. Factor analysis determines the impact direction and extent of various factors on the object. Factor analysis cannot only analyze the impact of all factors on the object, but also analyze the influence of a certain factor on the object for further comparison or selection, which is widely used in financial analysis.

Statistical Product and Service Solutions (SPSS) is one of the world's leading statistical analysis software. With the expansion of product service area and the increase of service depth, the strategic direction of SPSS was under a significant adjustment. And it can be quickly applied to natural science, social science and science of technology [24]. Many of the world's influential newspapers and magazines have given a high degree of evaluation on SPSS automatic statistical drawing, in-depth analysis of data, convenient use, complete functions and so on [25-27]. In this paper, we use IBM SPSS Statistics for the whole process of factor analysis.

In practical problems, there often exists close relationships among variables. But we cannot determine the value of a variable by another one or several variables. That is to say, when the independent variable x takes a certain value, the dependent variable y may have more than one value. The non one-to-one correspondence or the uncertainty relation between variables is referred to as correlation. SPSS describes the degree of linear correlation among variables by drawing scatter diagram and calculating the correlation coefficient. Also SPSS uses appropriate statistical indicators to express correlation. The whole process is called SPSS correlation analysis. Regression analysis is the method studying on the influence on one variable by changes of other variables. Regression analysis can figure out the relation expressions between them according to the known information or data and speculate the value or scope of dependent variable from known independent variables as well. Correlation analysis aims at determining the extent of correlation between variables using correlation coefficient. While regression analysis focuses on the quantity change laws among variables and describes the relationship between variables through a mathematical expression in order to determine the influencing degree of one or several variables on another given variable.

As we all know, carbon dioxide emissions is influenced by multiple factors directly or indirectly in various fields. If the data of all factors are input into the prediction model, the amount of computation is too large and the computation complexity is too high. And too many factors will reduce the prediction accuracy.

Secondly, there may exist correlation or similar change law between the influencing factors of carbon dioxide emissions. So the factor screening and merging are necessary. In this paper, SPSS was used to analyze the correlation between carbon dioxide emissions and the initially selected factors to remove the ones with weak correlation. At the same time, the remaining factors were analyzed by regression analysis in order to identify the more significant and necessary variables without multicollinearity, after which further screening and determination of factors was done.

2.2. Least Squares Support Vector Machine

Support vector machine (SVM) theory is a new type of machine learning method based on Vapnik-Chervonenkis dimension theory and structural risk minimization. SVM shows excellent learning performance in the case of less statistical samples and overcomes the low generalization ability and over fitting of neural network, which is considered to be the alternative method of neural network. SVM is not only the most practical learning method in statistical learning theory but also the youngest part of it. It was originally developed by pattern recognition and extended to function regression problems later through which SVM is proved to show a good function approximation performance.

LSSVM is the extension of SVM which was proposed by SuyKens and Vandewalle [28]. It transforms the inequality constraints of traditional SVM into equality constraints and considers sum squares error loss function as the loss experience of training set, which transforms quadratic programming problems into linear equation problem. The training set is set as $\{(x_k, y_k)|k=1,\ 2,\ \ldots.,\ n\}$, in which $x_k \in R^n$ is the input data, and $y_k \in R^n$ is the output. $\varphi(\bullet)$ is the nonlinear mapping function which transfers the samples into a much higher dimensional feature space $\phi(x_k)$. The optimal decision function in the high-dimensional feature space is given by:

$$y(x) = \omega^T \cdot \varphi(x) + b \tag{1}$$

where $\varphi(x)$ is mapping function; ω is weight vector; b is constant.

Using the principle of structural risk minimization, the objective optimization function is as follows:

$$\min_{w,b,e}(\omega,\ e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{n} e_k^2 \tag{2}$$

The constraint condition is :

$$y_k = \omega^T\varphi(x_k) + b + e_k \quad k = 1,\ 2,\ \ldots,\ n \tag{3}$$

in which γ is the penalty coefficient, and e_k is slack variable.

Define the Lagrange function to solve the problem:

$$L(\omega, b, e, \alpha) = \phi(\omega,\ e) - \sum_{k=1}^{n}\left\{\alpha_k\left[\omega^T\varphi(x_k) + b + e_k - y_k\right]\right\} \tag{4}$$

where Lagrange multiplier $\alpha_k \in R$. According to the Karush-Kuhn-Tucker (KKT) conditions, $\omega,\ b,\ e_k,\ \alpha_k$ are taken as partial derivatives and required as zero.

$$\begin{cases} \omega = \sum_{k=1}^{n} \alpha_k\,\varphi(x_k) \\ \sum_{k=1}^{n} \alpha_k = 0 \\ \alpha_k = e_k\gamma \\ \omega^T\varphi(x_k) + b + e_k - y_k = 0 \end{cases} \tag{5}$$

According to Eq. (5), the optimization problem can be transformed into linear problem, which is shown as follow:

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1,\ x_1) + \frac{1}{\gamma} & \cdots & K(x_1,\ x_l) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_l,\ x_1) & \cdots & K(x_l,\ x_l) + \frac{1}{\gamma} \end{bmatrix} \begin{bmatrix} B \\ \alpha_1 \\ \vdots \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_1 \end{bmatrix} \tag{6}$$

Solve formula (6) to get α and b, then the LSSVM optimal linear regression function is :

$$f(x) = \sum_{k=1}^{l} \alpha_k\,K(x,\ x_k) + b \tag{7}$$

According to Mercer condition, $K(x,\ x_i) = \varphi(x)^T \cdot \varphi(x_l)$ is kernel function. In this paper, set radial basis function (RBF) as kernel function which is shown in Eq. (8):

$$K(x,\ x_k) = \exp\left(-\frac{|x - x_k|^2}{2\sigma^2}\right) \tag{8}$$

where σ^2 is the width of kernel function.

It can be seen from the whole operation process of LSSVM that kernel parameter σ^2 and penalty parameter γ are generally set based on experience, which leads to the existence of randomness and inaccuracy in the application of the LSSVM algorithm. To solve this problem, we use fruit fly optimization algorithm to optimize the parameters in order to improve the prediction accuracy of LSSVM. In addition, LSSVM shows excellent learning performance in small amount of statistical samples and overcomes low generalization ability as well as over fitting. It can effectively solve the nonlinear, high-dimensional and small sample problems [29]. So it is suitable for carbon dioxide emissions prediction in this paper. And we use MATLAB for the implement of LSSVM.

2.3. Fruit Fly Optimization Algorithm

Fruit fly optimization algorithm is a kind of intelligent optimization algorithm on the basis of fruit fly foraging behaviors proposed by Pan Wenchao in 2011. The basic concept of FOA is that fruit fly perceives food concentration according to its position, then it moves to the site of maximum or minimum concentration by comparing flavor concentration [30]. Finally the objective function extreme value can be obtained through repeated iterations of food concentration. Food finding iterative process of fruit fly swarm is shown as Fig. 1.

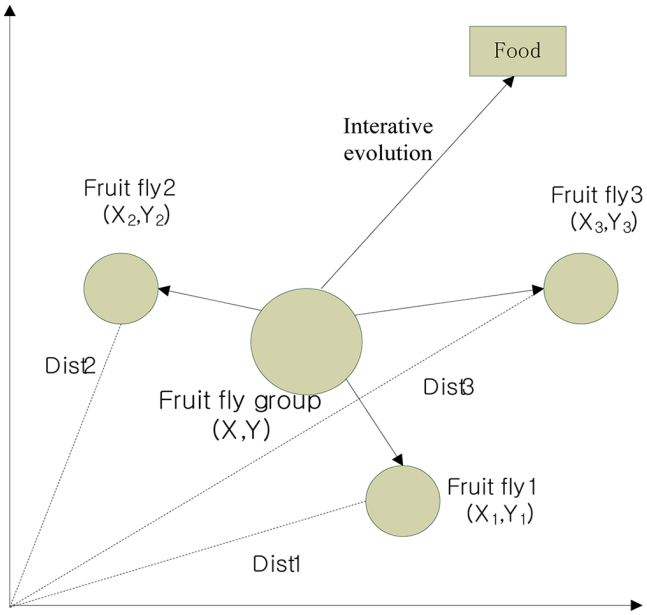

Fig. 1. Food finding iterative process of a fruit fly swarm.

According to the food finding characteristics of fruit fly swarm, the fruit fly optimization algorithm can be divided into the following steps:

(1) Randomly initialize the fruit fly swarm location (X_axis, Y_axis);

(2) Determine the random flight direction and the distance for food finding of an individual fruit fly by using olfactory:

$$X_i = X\_axis + Random\ Value \tag{9}$$

$$Y_i = Y\_axis + Random\ Value \tag{10}$$

(3) Calculate the distance between the origin and each individual fruit fly position(*Dist*), and then calculate the value of flavor concentration (*S*) which is the reciprocal of distance:

$$Dist = \sqrt{X_i^2 + Y_i^2} \tag{11}$$

$$S = \frac{1}{Dist} \tag{12}$$

(4) Put the value of flavor concentration *S* into its fitness function, then get the flavor concentration of the individual fruit fly location (*Smell*);

(5) Find out the individual fruit fly with minimal smell concentration among the fruit fly swarm:

$$[bestSmell, bestindex] = \max(Smell) \tag{13}$$

(6) Retain the best flavor concentration and its X, Y coordinates, then the fruit flies fly to the position by using vision.

$$Smellbest = bestSmell \tag{14}$$

$$X\_axis = X(bestindex) \tag{15}$$

$$Y\_axis = Y(bestindex) \tag{16}$$

(7) Repeat (2)-(5) to enter the iterative optimization. When the fitness value reaches target set or the iterative number reaches the maximal iterative number, the circulation stops and if not, go to step (6).

The two parameters in LSSVM which need optimizing mentioned in the previous are kernel parameter and penalty parameter. Calculate the distance D_i between each individual fruit fly *i* and the origin as well as taste concentration S according to formula (11), (12). Set $\gamma = 20 * S(i, 1)$, $\sigma^2 = S(i, 2)$ and substitute it to train LSSVM. Establish the fitness function by using RMSE as the fitness function value, as shown in the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_1 - \hat{y}_i)^2}{n}} \tag{17}$$

Where y_i is the actual value of the ith point, $\hat{y}_i$ is the forecast value for the ith point, n is the number of the predicted data. If the maximum number of iterations is reached or when RMSE $< 0.01\%$, end the calculation and keep the best γ and σ^2.

2.4. Hybrid Model of Factor Analysis, FOA and LSSVM

We combine SPSS factor analysis, FOA with LSSVM, which can better solve the complex nonlinear mapping problem and comprehensively reflect the characteristics of the research object. Also the hybrid model reveals the driving factors of carbon dioxide emissions and the relationship among them. At the same time, the proposed model has a higher prediction precision for the future development path of carbon dioxide emissions.

Firstly, SPSS correlation analysis and regression analysis were used to quantify the correlation between each influencing factor and carbon dioxide emissions. And the key factors are selected

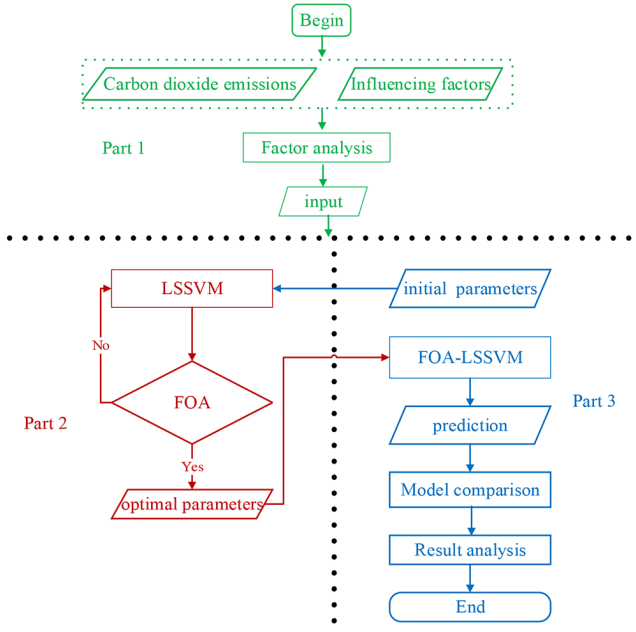

Fig. 2. Flowchart of proposed model.

to eliminate the interference of irrelevant factors. At the same time, the similarity and mutual influence of the influencing factors are excluded. Secondly, we optimize the parameters of LSSVM by fruit algorithm iterative to further improve the prediction effect. Finally, we use LSSVM optimized by FOA to construct the nonlinear regression prediction model with the identified key influencing factors as input factors and the corresponding carbon dioxide emissions as output. The specific process is shown in Fig. 2.

3. Case Analysis

3.1. The Analysis and Selection of Influencing Factors

The carbon dioxide emissions are influenced by many factors. In this paper, we analyze the influence of various factors on future development trend of carbon dioxide emissions from aspects of economic and social development. We select GDP, the primary industry GDP, the secondary industry GDP, the tertiary industry GDP, total population, per capita GDP, total energy consumption, power generation, steel production, regional final consumption, urban per capita disposable income, rural per capita net income, per capita GDP, coal consumption, private owned automobile quantity and total investment in fixed assets as 16 preliminary selected factors. The annual data of 16 influencing factors in 1980-2014 are shown in Table 1. Corresponding annual carbon dioxide emissions in 1980-2014 of China are shown in Table 2 (Data from the China Statistical Yearbook).

According to above data, we carry out the pairwise correlation

Table 1(a). The Data of Influencing Factors from 1980 to 2014

Year	GDP (100 million yuan)	Primary industry GDP (100 million yuan)	Secondary industry GDP (100 million yuan)	Tertiary industry GDP (100 million yuan)	Total population (10,000)
1980	4,551.6	1,359.4	2,180.5	1,011.6	98,705
1981	4,898.1	1,545.6	2,243.7	1,108.8	100,072
1982	5,333.1	1,761.6	2,370.6	1,200.9	101,654
1983	5,975.6	1,960.8	2,632.6	1,382.2	103,008
1984	7,226.3	2,295.5	3,089.7	1,841.1	104,357
1985	9,040	2,541.6	3,846.8	2,651.6	105,851
1986	10,308.7	2,763.9	4,469.9	3,074.9	107,507
1987	12,102.2	3,204.3	5,225.3	3,672.6	109,300
1988	15,101	3,831	6,554	4,716	111,026
1989	17,089.6	4,228	7,240	5,621.6	112,704
1990	18,774	5,017	7,678	6,079	114,333
1991	21,835.6	5,228.6	9,055.8	7,551.2	115,823
1992	27,068.3	5,800	11,640.4	9,627.9	117,171
1993	35,524.4	6,887.3	16,373	12,264.1	118,517
1994	48,459.6	9,471.4	22,333.5	16,654.7	119,850
1995	61,129.8	12,020	28,536.2	20,573.6	121,121
1996	71,572.3	13,877.8	33,665.8	24,028.7	122,389
1997	79,429.5	14,264.6	37,353.9	27,810.9	123,626
1998	84,883.7	14,618	38,808.8	31,456.8	124,761
1999	90,187.7	14,548.1	40,827.6	34,812	125,786
2000	99,776.3	14,716.2	45,326	39,734.1	126,743
2001	110,270.4	15,501.2	49,262	45,507.2	127,627
2002	121,002	16,188.6	53,624.4	51,189	128,453
2003	136,564.6	16,968.3	62,120.8	57,475.6	129,227
2004	160,714.4	20,901.8	73,529.8	66,282.8	129,988
2005	185,895.8	21,803.5	87,127.3	76,964.9	130,756
2006	217,656.6	23,313	103,163.5	91,180.1	131,448
2007	268,019.4	27,783	125,145.4	115,090.9	132,129
2008	316,751.7	32,747	148,097.9	135,906.9	132,802
2009	345,629.2	34,154	157,850.1	153,625.1	133,450
2010	408,903	39,354.6	188,804.9	180,743.4	134,091
2011	484,123.5	46,153.3	223,390.3	214,579.9	134,735
2012	534,123	50,892.7	240,200.4	243,030	135,404
2013	588,018.8	55,321.7	256,810	275,887	136,072
2014	636,138.7	58,336.1	271,764.5	306,038.2	136,782

between carbon dioxide emissions and the influencing factors with the help of IBM SPSS Statistics. We solve the Pearson correlation coefficient and conduct significant test. Operation results are seen in Table 3. From the results of correlation analysis, we can see that all the 16 influencing factors are significantly related to carbon dioxide emissions at 0.01 level. The absolute value of Pearson correlation coefficient represents the degree of correlation. Its sigh shows positive or negative correlation. Correlation test proves that the preliminary selection of factors is reasonable and practical. Specifically, we can see from the result that the correlation between total energy consumption or coal consumption and carbon dioxide emissions are the most significant. That is to say, total energy consumption and coal consumption have a positive impact on the increase of carbon dioxide emissions to a great extent. In addition, from the Pearson coefficients of GDP, the primary industry GDP, the secondary industry GDP, the tertiary industry GDP and steel production, we can draw the conclusion that the secondary industry especially heavy industry has the greatest impact on carbon dioxide emissions. And from these analyses we can get policy implications.

In addition to correlation analysis, we analyze the 16 factors by multiple linear regression analysis for further screening and

Table 1(b). The Data of Influencing Factors from 1980 to 2014

Year	GDP energy intensity (SCE/10,000 yuan)	Total energy consumption (10,000 tons of SCE)	Power generation (100 million kWh)	Steel production (10,000 tons)	Regional final consumption (100 million yuan)	Urban per capita disposable income (yuan)
1980	13.24	60,275	3,006	3,712	2,974.3	477.6
1981	12.53	61,364	3,093	3,560	3,282.3	500.4
1982	12.13	64,686	3,277	3,716	3,580.7	535.3
1983	11.53	68,877	3,514	4,002	4,068.6	564.6
1984	10.45	75,493	3,770	4,347	4,797.3	652.1
1985	8.48	76,682	4,107	4,679	5,931.1	739.1
1986	7.84	80,850	4,495	5,220	6,739.5	900.9
1987	7.16	86,632	4,973	5,628	7,649	1,002.1
1988	6.16	92,997	5,452	5,943	9,433	1,180.2
1989	5.67	96,934	5,848	6,159	11,043	1,373.9
1990	5.26	98,703	6,212	6,636	12,011.1	1,510.2
1991	4.75	103,783	6,775	7,100	13,628.6	1,700.6
1992	4.03	109,170	7,539	8,094	16,246.1	2,026.6
1993	3.27	115,993	8,395	8,956	20,826.9	2,577.4
1994	2.53	122,737	9,281	9,261	28,305.9	3,496.2
1995	2.15	131,176	10,070.3	9,536	36,225.7	4,283
1996	1.89	135,192	10,813.1	10,124	43,117.6	4,838.9
1997	1.71	135,909	11,355.53	10,894	47,556.7	5,160.3
1998	1.6	136,184	11,670	11,559	51,509.8	5,425.1
1999	1.56	140,569	12,393	12,426	56,681.9	5,854
2000	1.47	146,964	13,556	12,850	63,729.2	6,280
2001	1.41	155,547	14,808.02	15,163	68,617.2	659.6
2002	1.4	169,577	16,540	18,237	74,171.7	7,702.8
2003	1.44	197,083	19,105.75	22,234	79,641.5	8,472.2
2004	1.43	230,281	22,033.09	28,291	89,224.8	9,421.6
2005	1.41	261,369	25,002.6	35,324	101,604.2	10,493
2006	1.32	286,467	28,657.23	41,915	114,894.9	11,759.5
2007	1.16	311,442	32,815.53	48,929	136,438.7	13,785.8
2008	1.01	320,611	34,668.82	50,306	157,746.3	15,780.8
2009	0.97	336,126	37,146.51	57,218.23	173,093	17,174.7
2010	0.88	360,648	42,071.6	63,722.99	199,508.4	19,109.4
2011	0.8	387,043	47,130.19	68,528.31	241,579.1	21,809.8
2012	0.75	402,138	49,875.53	72,388.22	271,718.6	24,564.7
2013	0.71	416,913	54,316.35	77,904.1	301,008.4	26,955.1
2014	0.67	426,000	56,495.83	82,230.63	329,450.8	29,381

to determine the key factors of carbon dioxide emissions. The results of analysis are shown in Table 4, Table 5 and Table 6. It can be seen from the results that GDP, tertiary industry GDP, total energy consumption and per capita GDP have serious collinearity with each other or with other factors, which influences the performance and accuracy of carbon dioxide emissions prediction. At the same time, we can analyze that information contained in GDP can be fully reflected by three types of industry. And the per capita GDP is directly derived from GDP and population. The total energy consumption can be replaced by coal consumption to reflect the impacts on carbon dioxide emissions to a great extent. In short, through regression analysis we exclude GDP, tertiary industry GDP, total energy consumption and per capita GDP to optimize the influencing factor system.

3.2. Carbon Dioxide Emissions Prediction Based on FOA-LSSVM

Through the above SPSS correlation analysis and regression analysis, we identified 12 key factors of carbon dioxide emissions. The data information of key influencing factors is used as the input of FOA-LSSVM model, while the corresponding carbon dioxide

Table 1(c). The Data of Influencing Factors from 1980 to 2014

Year	Rural per capita net income (yuan)	per capita GDP (yuan)	Coal consumption (10,000 tons of SCE)	Private owned automobile quantity (10,000 units)	Total investment in fixed assets (100 million yuan)
1980	191.3	461.13	43,518.55	191.6	910.9
1981	223.4	489.46	44,427.54	213.62	961
1982	270.1	524.63	47,544.21	230.57	1,230.4
1983	309.8	580.11	51,037.86	248.04	1,430.1
1984	355.3	692.46	56,393.27	276.17	1,932.9
1985	397.6	854.03	58,124.96	339.59	2,543.2
1986	423.8	958.89	61,284.3	380.72	3,120.6
1987	462.6	1,107.25	66,013.58	423.47	3,791.7
1988	544.9	1,360.13	70,863.71	480.55	4,753.8
1989	601.5	1,516.33	73,669.84	529.84	4,410.4
1990	686.3	1,642.05	75,211.69	572.05	4,517
1991	708.6	1,885.26	78,978.86	628.36	5,594.5
1992	784	2,310.15	82,641.69	715.87	8,080.1
1993	921.6	2,997.41	86,646.77	848.18	13,072.3
1994	1221	4,043.35	92,052.75	973.83	17,042.1
1995	1,577.7	5,047	97,857.3	1,076.67	20,019.3
1996	1,926.1	5,847.94	99,366.12	1,137.11	22,913.5
1997	2,090.1	6,424.98	97,039.03	1,256.39	24,941.1
1998	2,162	6,803.7	96,554.46	1,355.88	28,406.2
1999	2,210.3	7,169.93	99,241.71	1,489.66	29,854.7
2000	2,253.4	7,872.33	100,670.34	1,647.77	32,917.7
2001	2,366.4	8,640.05	105,771.96	1,844.88	37,213.5
2002	2,475.6	9,419.94	116,160.25	2,091.75	43,499.9
2003	2,622.2	10,567.81	138,352.27	2,433.54	55,566.6
2004	2,936.4	12,363.79	161,657.26	2,758.51	70,477.4
2005	3,254.9	14,217	189,231.16	3,231.32	88,773.6
2006	3,587	16,558.38	207,402.11	3,788.84	109,998.2
2007	4,140.4	20,284.68	225,795.45	4,466.67	137,323.9
2008	4,760.6	23,851.43	229,236.87	5,234.23	172,828.4
2009	5,153.2	25,899.53	240,666.22	6,347.53	224,598.8
2010	5,919	30,494.44	249,568.42	7,881.97	251,683.8
2011	6,977.3	35,931.53	271,704.19	9,446.28	311,485.1
2012	7,916.6	39,446.62	275,464.53	11,028.42	374,694.7
2013	8,895.9	43,213.8	280,999.36	12,767.89	446,294.1
2014	9,892	46,507.49	281,160	14,598.11	512,020.7

Table 2. Annual Carbon Dioxide Emissions of 1980-2014

Year	Carbon dioxide emissions (10,000 tons of SCE)	Year	Carbon dioxide emissions (10,000 tons of SCE)
1980	40,235.82	1998	81,008.65
1981	39,996.16	1999	81,465.88
1982	41,857.69	2000	89,751.49
1983	44,263.54	2001	93,918.23
1984	47,909.07	2002	102,279.48
1985	51,611.69	2003	118,640.86
1986	54,744.58	2004	139,002.31
1987	58,419.65	2005	159,502.65
1988	62,232.19	2006	173,780.67
1989	63,210.71	2007	188,086.46
1990	63,025.77	2008	191,864.30
1991	65,816.06	2009	203,614.50
1992	68,025.61	2010	211,864.64
1993	71,261.93	2011	231,170.51
1994	76,501.52	2012	235,753.92
1995	79,217.35	2013	241,309.57
1996	83,495.33	2014	250,698.62
1997	81,062.92	2015	NA

Table 3. Correlation Analysis

Factor	Pearson coefficient	Significance	Factor	Pearson coefficient	Significance
GDP	0.969**	0.000	Steel production	0.990**	0.000
Primary industry GDP	0.974**	0.000	Regional final consumption	0.968**	0.000
Secondary industry GDP	0.976**	0.000	Urban per capita disposable income	0.974**	0.000
Tertiary industry GDP	0.959**	0.000	Rural per capita net income	0.964**	0.419
Total population	0.859**	0.000	Per capita GDP	0.972**	0.000
GDP energy intensity	-0.687**	0.000	Coal consumption	0.999**	0.000
Total energy consumption	0.999**	0.000	Private owned automobile quantity	0.935**	0.000
Power generation	0.991**	0.000	Total investment in fixed assets	0.923**	0.000

** indicates a significant correlation at 0.01 level (bilateral).

Table 4. Model Summary

model	R	R^2	Adjusted R^2	Error estimated of standard
1	1.000[a]	1.000	0.999	1,714.81019

[a] Predictive variables (constant): Total investment in fixed assets, GDP energy intensity, Coal consumption, Total population, Urban per capita disposable income, Primary industry GDP, Secondary industry GDP, Private owned automobile quantity, Steel production, Rural per capita net income, Regional final consumption, Power generation.

Table 5. Anova[b]

Model		Quadratic sum	df	Mean square	F	Sig.
1	Regression	1.564E11	12	1.303E10	4,432.255	.000[a]
	Residual error	64,692,627.672	22	2,940,573.985		
	Sum	1.565E11	34			

[a] Predictive variables (constant): Total investment in fixed assets, GDP energy intensity, Coal consumption, Total population, Urban per capita disposable income, Primary industry GDP, Secondary industry GDP, Private owned automobile quantity, Steel production, Rural per capita net income, Regional final consumption, Power generation

[b] Dependent variable: Carbon dioxide emissions

Table 6. Excluded Variables[b]

Model		Beta In	t	Sig.	Partial correlation	Collinearity statistics		
						Tolerance	VIF	Minimum tolerance
1	GDP	.562[a]	.547	.590	.119	1.843E-5	54264.899	1.843E-5
	Tertiary industry GDP	.262[a]	.547	.590	.119	8.478E-5	11795.523	8.478E-5
	Total energy consumption	1.282[a]	3.182	.004	.570	8.189E-5	12210.823	8.189E-5
	Per capita GDP	.739[a]	.603	.553	.130	1.288E-5	77621.713	1.288E-5

[a] Predictive variables (constant): Total investment in fixed assets, GDP energy intensity, Coal consumption, Total population, Urban per capita disposable income, Primary industry GDP, Secondary industry GDP, Private owned automobile quantity, Steel production, Rural per capita net income, Regional final consumption, Power generation
[b] Dependent variable: Carbon dioxide emissions

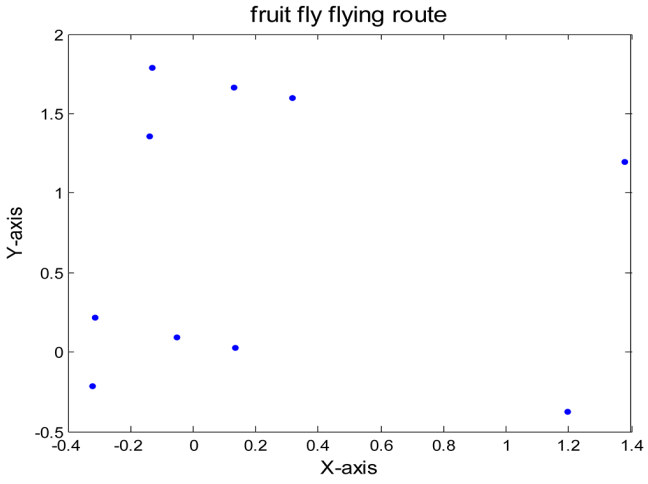

Fig. 3. Fruit fly flying route chart.

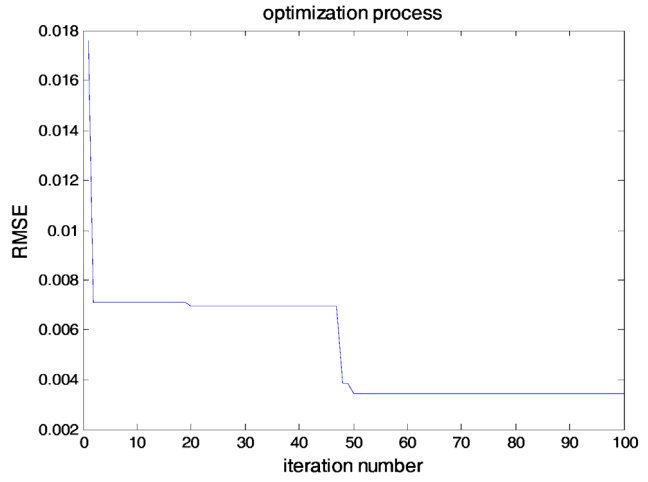

Fig. 4. Iterative curve of FOA.

emissions as the model output. Among them, the first 30 groups of sample data are taken as the training set and the last 5 groups of sample data as a test set to verify the validity and accuracy of the model. The parameters of LSSVM are optimized by fruit fly algorithm. In this paper, the initial population size is set to 40. The iterative number is set to 100. The optimized results are C = 336.6509, e = 8.1913.

3.3. Model Comparison and Result Analysis

In order to test whether the proposed model is suitable for prediction of carbon dioxide emissions and the superiority of it, we select the single LSSVM algorithm, BP neural network and GM(1,1) as contrast models for in-depth analysis. We calculated the mean absolute percentage error of four models (MAPE), percentage of the maximum absolute error (MAXAPE) and mean absolute percentage error median (MDAPE) of four models to evaluate of performance of different model in carbon dioxide emissions prediction, shown as Table 7. The fitting curves of different model test set are shown in Fig. 5, which shows the fitting effects of different models more intuitively. Both the forecast error statistical results and fitting curves of different models show that the proposed model is better than the contrast models. And the hybrid model of factor analysis, FOA and LSSVM has an incomparable fitting accuracy than other models.

Table 7. Forecast Error Statistical Results of Different Models

	GM(1,1)	BP	LSSVM	Proposed model
MAPE (%)	12.24	11.26	4.06	1.03
MaxAPE (%)	16.65	21.43	5.33	2.46
MdAPE (%)	13.03	14.03	4.37	0.88

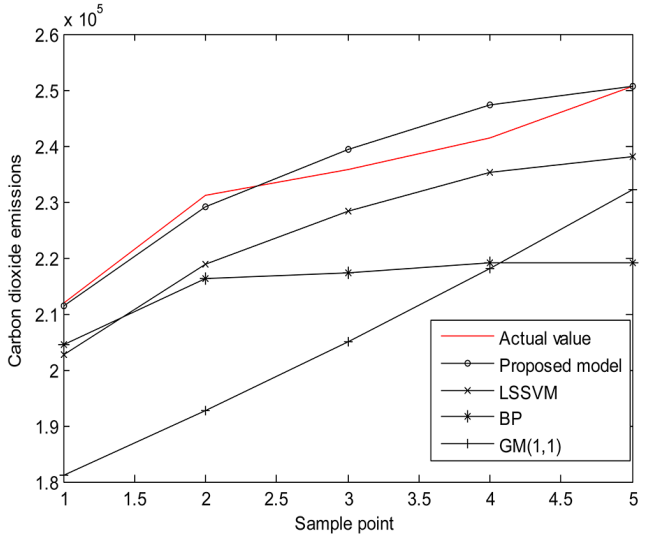

Fig. 5. Fitting curves of different models.

4. Conclusions

In this paper, we use fruit fly algorithm to select the parameters of LSSVM automatically and establish the hybrid forecasting model based on factor analysis, thus avoiding the subjectivity and blindness of man-made parameter determination. The hybrid model is used to forecast carbon dioxide emissions of China. The model comparison and result analysis show that the hybrid model in this paper has the best fitting result compared to models without parameter optimization or other conventional prediction models. Compared to similar works mentioned in the section of Introduction, we not only examine the relationship between carbon dioxide emissions and the influencing factors but also propose a new feasible and effective hybrid model involving artificial intelligence algorithms. And the prediction accuracy of proposed model has been greatly improved.

Based on analysis in the previous sections of this paper, we conclude some policy implications about carbon emission reduction. The government should actively adjust the industrial structure, eliminate backward industries as well as implement high efficiency, high technology, low pollution and low energy consumption. In addition, it is of great significance to vigorously develop clean energy such as wind power, hydro-power and solar energy, which promotes the development of green energy and the transformation of energy structure, thus achieving energy saving and emission reduction.

Facing increasingly serious problems of energy consumption and environmental pollution, the study about carbon emissions becomes more important to all countries in the world. China is a country with a large population. Its industrialization and urbanization process is far from complete. The energy consumption of production and living is still in the stage of rapid growth. We still a lot of fossil energy for a long time in the future in order to ensure the sustainable development of economy and society. The attendant problem of carbon dioxide emissions deserves particular concern. We complete the influencing factor analysis, screening and treatment of carbon dioxide emissions as well as accurate prediction for carbon dioxide emissions, which contributes to the monitoring and decision-making processes of managers. Meanwhile, it is noteworthy that carbon dioxide emissions have become a global topic and low carbon development has been approved widely by all countries. Worldwide scholars can determine the influencing factors according to particular country as well as the actual situation and carry out researches about carbon emissions using reasonable methods. And this paper can provide a reference for related researches in other continents or countries. Besides, we will focus on future scenario prediction of carbon dioxide emissions on the basis of this study for our subsequent researches.

References

1. Worrell E, Price L, Martin N, Hendriks C, Meida LO. Carbon dioxide emissions from the global cement industry. *Annu. Rev. Energ. Environ*. 2001;26:303-329.
2. Marland G, Andres RJ, Boden TA. Global, regional, and national CO_2 emissions. Trends 93: A Compendium of Data on Global Change; 1994. p. 505-584.
3. Solomon S, Plattner GK, Knutti R, Friedlingstein P. Irreversible climate change due to carbon dioxide emissions. *Proc. Natl. Acad. Sci. USA.* 2009:106:1704-1709.
4. Du L, Wei C, Cai S. Economic development and carbon dioxide emissions in China: Provincial panel data analysis. *China Econ. Rev.* 2012;23:371-384.
5. Suganthi L, Samuel AA. Energy models for demand forecasting–A review. *Renew. Sust. Energy. Rev*. 2012;16:1223-1240.
6. Rout UK, Voβ A, Singh A, Fahl U, Blesl M, Gallachoir BPO. Energy and emissions forecast of China over a long-time horizon. *Energy* 2011;36:1-11.
7. Chang YS, Jeon S. Using the experience curve model to project carbon dioxide emissions through 2040. *Carbon Manag.* 2015;6:51-62.
8. Auffhammer M, Carson RT. Forecasting the path of China's CO_2 emissions using province-level information. *J. Environ. Econ. Manag.* 2008;55:229-247.
9. Sheta AF, Ghatasheh N, Faris H. Forecasting global carbon dioxide emission using auto-regressive with exogenous input and evolutionary product unit neural network models. Information and Communication Systems (ICICS), 2015 6th International Conference on. IEEE; 2015. p. 182-187.
10. Ramanathan R. A multi-factor efficiency perspective to the relationships among world GDP, energy consumption and carbon dioxide emissions. *Technol. Forecast. Soc.* 2006;73:483-494.
11. Lin SJ, Lu IJ, Lewis C. Grey relation performance correlations among economics, energy use and carbon dioxide emission in Taiwan. *Energ. Policy* 2007;35:1948-1955.
12. Faria AP, Fernandes GW, França MGC. Predicting the impact of increasing carbon dioxide concentration and temperature on seed germination and seedling establishment of African grasses in Brazilian Cerrado. *Austral Ecol.* 2015;40:962-973.
13. Asumadu-Sarkodie S, Owusu PA. Energy use, carbon dioxide emissions, GDP, industrialization, financial development, and population, a causal nexus in Sri Lanka: With a subsequent prediction of energy use using neural network. *Energ. Source. Part B.* 2016;11:889-899.
14. Mohiuddin O, Asumadu-Sarkodie S, Obaidullah M. The relationship between carbon dioxide emissions, energy consumption, and GDP: A recent evidence from Pakistan. *Cogent Eng.* 2016;3:1210491.
15. Sun W, Liu M. Prediction and analysis of the three major industries and residential consumption CO_2 emissions based on least squares support vector machine in China. *J. Clean. Prod.* 2016;122:144-153.
16. Asumadu-Sarkodie S, Owusu PA. Carbon dioxide emissions, GDP, energy use, and population growth: A multivariate and causality analysis for Ghana, 1971-2013. *Environ. Sci. Pollut. Res.* 2016;23:13508-13520.
17. Asumadu-Sarkodie S, Owusu PA. Multivariate co-integration analysis of the Kaya factors in Ghana. *Environ. Sci. Pollut. Res.* 2016;23:9934-9943.
18. Pauzi HM, Abdullah L. Neural network training algorithm for carbon dioxide emissions forecast: A performance comparison.

Advanced Computer and Communication Engineering Technology. Springer International Publishing; 2015. p. 717-726.
19. Chang H, Sun W, Gu X. Forecasting energy CO_2 emissions using a quantum harmony search algorithm-based DMSFE combination model. *Energies* 2013;6:1456-1477.
20. Salahuddin M, Gow J, Ozturk I. Is the long-run relationship between economic growth, electricity consumption, carbon dioxide emissions and financial development in Gulf Cooperation Council Countries robust? *Renew. Sust. Energy Rev*. 2015;51:317-326.
21. Cheng J, Qian JS, Guo YN. Least squares support vector machines for gas concentration forecasting in coal mine. *Int. J. Comput. Sci. Network Secur.* 2006;6:125-129.
22. Wang XI, Wang MW. Short-term wind speed forecasting based on wavelet decomposition and least square support vector machine. *Power Syst. Technol.* 2010;34:179-184.
23. Pan WT. A new fruit fly optimization algorithm: taking the financial distress model as an example. *Knowl-Based Syst.* 2012;26:69-74.
24. Harman HH. Modern factor analysis. Oxford: Univ. of Chicago Press; 1960. p. 10-23.
25. Levesque R. SPSS programming and data management: A guide for SPSS and SAS users. Chicago: Spss Inc.; 2005.
26. Green SB, Salkind NJ. Using SPSS for windows and macintosh: Analyzing and understanding data. 6th ed. New Jersey: Prentice Hall Press; 2010.
27. O'connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav. Res. Meth. Instrum. Comput.* 2000;32:396-402.
28. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process. Lett*. 1999;9:293-300.
29. Ismail S, Shabri A, Samsudin R. A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting. *Expert Syst. Appl*. 2011;38:10574-10578.
30. Li HZ, Guo S, Li CJ, Sun JQ. A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowl-Based Syst.* 2013;37:378-387.