RESEARCH ARTICLE

Forecasting annual natural gas consumption via the application of a novel hybrid model

Feng Gao 1,2 · Xueyan Shao 1,2

Received: 27 October 2020 / Accepted: 28 December 2020 / Published online: 7 January 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract
Accurate prediction of natural gas consumption (NGC) can offer effective information for energy planning and policy-making. In this study, a novel hybrid forecasting model based on support vector machine (SVM) and improved artificial fish swarm algorithm (IAFSA) is proposed to predict annual NGC. An adaptive learning strategy based on sigmoid function is introduced to improve the performance of traditional artificial fish swarm algorithm (AFSA), which provides a dynamic adjustment for parameter moving step by step and visual scope visual. IAFSA is used to obtain the optimal parameters of SVM. In addition, the annual NGC data of China is selected as an example to evaluate the prediction performance of the proposed model. Experimental results reveal that the proposed model in this study outperforms the benchmark models such as artificial neural network (ANN) and partial least squares regression (PLS). The mean absolute percentage error (MAPE), root mean squared error (RMSE), and mean absolute error (MAE) values are as low as 0.512, 1.4958, and 1.0940. Finally, the proposed model is employed to predict NGC in China from 2020 to 2025.

Keywords Support vector machine · Improved artificial fish swarm algorithm · Natural gas consumption forecasting

Introduction
Global warming is a serious problem for countries all over the world, threatening the survival and development of human beings. Overuse of traditional fossil energy is the main cause of global warming. After the Copenhagen Conference, most countries began to explore clean and low-carbon energy transition, striving to reduce carbon emissions (Lu et al. 2020). Due to its cleanliness and high efficiency, natural gas becomes the best transitional energy in the process of clean and low-carbon energy transition. According to BP Energy Outlook 2020, it is projected that the global demand for natural gas will grow robustly over the next 15 years (BP 2020). As the demand rises, it is necessary for governments to formulate a reasonable natural gas plan in advance to ensure sufficient natural gas supplies and transmission infrastructures. Improper estimation of natural gas consumption (NGC) may lead to the imbalance of natural gas supply and demand, causing social economic losses (Shaikh and Ji 2016). Consequently, accurate prediction of NGC has great significance to the formulation of an energy policy and plan.

However, forecasting NGC is a challenging task. NGC is affected by various external factors, such as population and economic development, and the relationship between influence factors and NGC is nonlinear (Kavaklioglu 2011). Besides, the natural gas consumption time series (in particular yearly) data have the characteristics of nonlinearity, volatility, and small samples (Lu et al. 2020). In this case, it is very difficult to establish an accurate prediction model. Therefore, how to improve the prediction accuracy of NGC is a very important issue in the field of energy prediction.

Up to now, a large number of scholars have conducted studies on NGC prediction methods. Grey forecasting models and regression models are the most popular methods, widely used in NGC forecasting. For example, Shaikh et al. (2017) and Ding (2018) constructed China’s yearly NGC forecasting model by utilizing grey models. Ozdemir et al. (2016) and Melikoglu (2013) employed linear regression to forecast natural gas demand in Turkey. With the development of artificial intelligence technology, machine
learning forecasting models have been introduced to predict NGC. Szoplik (2015) used artificial neural networks (ANN) to forecast daily NGC in Szczecin (Poland). Wei et al. (2019) proposed a hybrid model combining improved singular spectrum analysis and long short-term memory to forecast daily NGC.

Despite some progress made in NGC prediction model studies, there are still some research gaps worthy of further exploration. In the current researches, the annual NGC forecasting is mainly based on grey prediction models and multiple regression models. To the best of our knowledge, there are few studies which apply machine learning models for long-term (annual) NGC forecasting. In fact, some machine learning models, such as support vector machine (SVM), have excellent fitting capabilities for nonlinear small sample data. For instance, Kavaklioglu (2011) applied SVM to predict Turkey’s electricity consumption using data from 1975 to 2006. The first 26 data are used for training and the remaining data are used for testing. Compared with random walk, SVM has a lower error. Ma et al. (2019) employed SVM to forecast building energy consumption in southern China with the test sample from 2000 to 2014. Forecasting results show that SVM can estimate building energy consumption with accuracy with MSE less than 0.001. However, the learning performance and generalization ability of SVM largely depends on the model parameters (Cherkassky and Ma 2004). To tackle this problem, some intelligent optimization algorithms are used to optimize parameters of SVM, such as genetic algorithm (GA) (Huang and Wang 2006) and particle swarm optimization algorithm (PSO) (Jiang et al. 2011; Meng et al. 2014). But these traditional intelligent optimization algorithms are prone to premature maturity and fall into local optimum (Chen et al. 2008), which may reduce the prediction accuracy.

In this paper, a novel hybrid forecasting model is proposed to predict annual NGC. The proposed model combines support vector machine (SVM) and improved artificial fish swarm algorithm (IAFSA), making full use of the ability to process small samples of SVM and the global optimization capabilities of IAFSA. IAFSA is used to optimize the parameters of SVM. The purpose of this study is to improve the prediction accuracy of NGC and provide more effective information for energy planning and policy-making.

The main contributions of this study are concluded as follows: First, an adaptive learning strategy based on sigmoid function is introduced to improve the traditional artificial fish swarm algorithm (AFSA), dealing with imbalance between local search and global search caused by fixed initial values. Three standard test functions are used to verify the strategy. Simulating results reveal that IAFSA has stronger search ability. Second, a novel hybrid model based on SVM and IAFSA is proposed to predict annual NGC. China’s NGC data is applied to validate forecasting performance of the proposed model. Experimental results show that the proposed model performs much better than benchmark models.

The rest of this study is organized as follows: “Literature review” presents the previous literature. “Methodology” introduces the foundation of support vector machine (SVM) and artificial fish swarm algorithm (AFSA), and proposes the improved artificial fish swarm algorithm (IAFSA). “Data description” describes the data in detail. “Experimental results and discussion” presents the experimental results of IAFSA-SVM and compares with benchmark models to demonstrate the forecasting performance of IAFSA-SVM. “Conclusions and future work” draws some conclusions and proposes future work.

### Literature review

#### Influence factors of NGC

Analyzing and finding out the influence factors of natural gas consumption (NGC) is the key to forecasting NGC. In recent years, a large amount of literature has conducted research on the influencing factors of NGC, as summarized in Table 1. For example, Wang and Li (2020) applied grey relationship analysis to study the influence factors of natural gas demand in China’s eastern, central, and western regions. Results showed that GDP, energy consumption structure, urbanization rate, industrial structure, and population density are the main factors affecting China’s natural gas demand. Similarly, Melikoglu (2013) selected GDP and population as influence factors to estimate Turkey’s NGC. Besides GDP and population, Wadud et al. (2011) considered natural gas prices as the main external driver that affect Bangladesh’s NGC. According to his study, the relationship between NGC and GDP is positive while the relationship between NGC and natural gas price is negative. Khan (2015) studied the relationship between income, natural gas price and price of substitutes, and Pakistan’s NGC. The research results reveal that income is positively correlated with NGC in the short and long run, and natural gas price is negatively correlated with NGC. The price of substitutes has no significant impact on NGC in the short run, but has a weak positive correlation in the long run. Furthermore, through qualitative analysis, Hongxun and Hui (2019) found out the factors affecting China’s NGC including GDP, total population, urbanization rate, natural gas pipeline construction, energy consumption structure, carbon emissions, etc.

The above-mentioned literature analyzed the factors affecting NGC from various aspects, and verified the relationship between various influence factors and NGC. Due to different conditions and development stages in various countries, the influencing factors of NGC are different. To sum up, gross...
domestic products (GDP) and population are the most commonly used factors.

**NGC forecasting**

At present, there are many methods for natural gas consumption (NGC) in the literature, which can be divided into three categories: grey prediction models, regression models, and machine learning models. The grey prediction models have a good fitting ability for small sample data with poor information, and many scholars use the model for long-term NGC prediction. For example, Zheng et al. (2020) predicted the NGC of China using the optimized nonhomogeneous grey model. Ding (2018) and Zeng and Li (2016) predicted China’s natural gas demand by using a self-adapting intelligent grey model. Shaikh et al. (2017) constructed China’s NGC forecasting model by utilizing two optimized nonlinear grey models. However, the grey prediction model is based on the historical data of NGC and usually does not consider the influence of exogenous factors. When significant changes occur, historical data may be unreliable for future NGC prediction. In contrast, regression models can clearly express the relationship between influence factors and NGC, so they are also widely used in NGC forecasting. Shaikh and Ji (2016) developed a logistic model to forecast the medium-term and long-term natural gas demand in China. Melikoglu (2013) built a logistic equation for long-term natural gas demand forecasting in Turkey. Ozdemir et al. (2016) employed linear regression optimized by genetic algorithm-simulated annealing (GA-SA) algorithm to forecast NGC in Turkey. Karadede et al. (2017) and Sen et al. (2019) developed a nonlinear regression for natural gas demand (consumption) forecasting in Turkey. Nevertheless, regression models might have difficulty in capturing the complex nonlinear relationship between influence factors and NGC (Tang et al. 2012), resulting in low prediction accuracy.

With the development of artificial intelligence technology, machine learning forecasting models have been widely used in forecasting NGC. Compared with traditional forecasting models, machine learning models have stronger nonlinear fitting ability (Niu and Dai 2017) and higher forecasting accuracy. Laib et al. (2016) predicted the annual NGC in Algeria using the artificial neural network (ANN). Similarly, Szoplik (2015) used the ANN to forecast hourly natural gas demand in Szczecin (Poland). Akpinar et al. (2016) applied the ANN with weights trained by artificial bee colony algorithms to predict daily NGC. Bai and Li (2016) built a structure-calibrated support vector regression (SVR) approach to predict daily NGC. Karabiber and Xydis (2020) applied the extreme learning machine (ELM) to predict day-ahead natural gas demand in Denmark. Beyca et al. (2019) employed the SVM for forecasting monthly NGC in Istanbul, Turkey. The results revealed SVM outperforms ANN models in terms of forecasting accuracy.

According to the aforementioned literature review results, it can be found that most studies used machine learning and its hybrid models for mid- and short-term (daily, monthly) NGC prediction, and only a few studies applied ANN to predict annual NGC. However, ANN requires a lot of training data, which is not suitable for annual NGC data (small samples). Compared with ANN, SVM adopts the structural risk minimization criterion to process small sample data (Mirzaei et al. 2015) and is more suitable for annual NGC forecasting. Therefore, this article attempts to introduce SVM into NGC forecasting, and uses an improved artificial fish swarm algorithm (IAFSA) to optimize parameters. Compared with commonly used optimization algorithms such as GA and PSO, artificial fish swarm algorithm (AFSA) is a new intelligent swarm optimization algorithm with strong global search

### Table 1 Summary of factors affecting NGC in current literatures

| Country   | Time range | Influence factors                                                                 |
|-----------|------------|-------------------------------------------------------------------------------------|
| (Wang and Li 2020) | China | Annual | GDP, energy consumption structure, urbanization level, industry structure, population density |
| (Liu et al. 2018) | China | Annual | Pipeline length, urban population, natural gas price |
| (Gao and Dong 2018) | China | Annual | Pipeline length, GDP per capita, economic growth |
| (Hongxun and Hui 2019) | China | Annual | GDP, population, urbanization rate, energy consumption structure, carbon emissions |
| (Melikoglu 2013) | Turkey | Annual | GDP, population |
| (Sen et al. 2019) | Turkey | Annual | GDP, inflation |
| (Karadede et al. 2017) | Turkey | Annual | GDP, population, economic growth |
| (Ozdemir et al. 2016) | Turkey | Annual | GDP, population, economic growth |
| (Khan 2015) | Pakistan | Annual | Income, natural gas price, price of energy alternative |
| (Wadad et al. 2011) | Bengal | Annual | GDP, population, natural gas price |
| (Rafindadi and Ozturk 2015) | Malaysia | Annual | Population, capital, exports |
ability and good robustness, and is not sensitive to initial parameters (Chen et al. 2008). In order to further improve the optimization accuracy, an adaptive leaning strategy based on sigmoid function is introduced to improve AFSA.

**Methodology**

**SVM**

Support vector machine (SVM) is a machine learning algorithm based on the principles of structural risk minimization (Xue and Yang 2016). It has good generalization ability for small sample, nonlinear and high-dimensional feature data (Mirzaei et al. 2015). Not only does SVM perform well in small sample, nonlinear and high-dimensional feature data estimation problems (Vapnik 1998). The support vector regression prediction, but it can also be used in regression estimation algorithm based on the principles of structural risk minimization (Zheng et al. 2019). There are three kinds of kernel functions which are commonly used: polynomial function (Eq. (9)), radial basis function (Eq. (10)), and linear function (Eq. (11)).

\[
K(X_i, X) = (X_i; X + 1)\theta
\]

(9)

\[
K(X_i, X) = e^{-\gamma \|x_i - x\|^2}
\]

(10)

\[
K(X_i, X) = X_i; X
\]

(11)

Among them, radial basis kernel function, also called Gaussian kernel function, has excellent adaptability and recognition accuracy for small sample and nonlinear problems. Therefore, this study uses the radial basis kernel function. For SVM with radial basis kernel function, the parameters include penalty factor C and kernel parameter \(\gamma\).

**IAFSA**

**Artificial fish swarm algorithm**

Artificial fish swarm algorithm (AFSA) is a new bionic optimization algorithm, proposed by Xiaolei Li in 2002 (Li et al. 2002). The core idea is to solve the optimization problem by simulating the behaviors of natural fish searching for food in water. In AFSA, the position of the artificial fish represents a feasible solution to the optimization problem. The optimization goal of AFSA is to find the position with the highest food concentration among the fish swarm, which also represents the optimal solution of the optimization problem (Zheng et al. 2019).

Suppose the search space is \(D\)-dimensional and there are \(N\) artificial fishes in total. The current position of fish \(i\) is \(X_i = (x_{i1}, x_{i2}, \ldots, x_{iD}) (i = 1, 2, \ldots, N)\), and the fitness function also called objective function, which means the food concentration of current position, is \(Y_i = f(X_i)\). The distance between fish \(i\) and fish \(j\) can be calculated by \(d_{ij} = \|X_i - X_j\|\). There are some other important parameters, including the moving step \(\text{step}\), visual scope \(\text{visual}\), crowded degree \(\delta\), and so on. In AFSA, an artificial fish performs preying behavior, swarming behavior, following behavior, and random behavior to search for the...
position with higher food concentration and record the optimal position on the bulletin board.

(1) Preying behavior

Preying behavior is a behavior in which an artificial fish swims to the position with higher food concentration. Suppose the current position of the artificial fish \(i\) is \(X_i\), we randomly select a new position \(X_j\) within its visual scope \(d_j < \text{visual} \); \(Y_i\) and \(Y_j\) are the corresponding fitness values. If, in a minimum problem, \(Y_j < Y_i\), \(X_i\) moves a step toward \(X_j\). Otherwise, continue to search within its visual scope and judge whether the new position satisfies the forward condition. If it still cannot be satisfied when the number of tries is greater than the maximum number of tries, perform random behavior. Preying behavior can be expressed as follows:

\[
X_{\text{next}} = X_i + \text{rand}() \cdot \text{step} \frac{X_j - X_i}{\|X_j - X_i\|} \tag{12}
\]

where \(\text{rand}()\) represents a random number between 0 and 1, and \(X_{\text{next}}\) is the next position of artificial fish \(i\).

(2) Random behavior

Random behavior is a behavior in which the artificial fish swims randomly within its visual scope, which is a default behavior of preying behavior. Random behavior can be expressed as follows:

\[
X_j = X_i + \text{rand}() \cdot \text{step} \tag{13}
\]

(3) Swarming behavior

In order to avoid danger, a fish swarm will spontaneously gather in groups. Within the visual scope of \(X(d_j < \text{visual})\), \(X_c\) is the central position of the fish swarm, and \(n_f\) is the number of companions. If \(\sum n_f < \delta Y_i\), it means that \(X_c\) is not too crowded with higher food concentration; \(X_i\) moves a step toward \(X_c\). Otherwise, perform preying behavior. The swarming behavior can be expressed as follows:

\[
X_{\text{next}} = X_i + \text{rand}() \cdot \text{step} \frac{X_c - X_i}{\|X_c - X_i\|} \tag{14}
\]

(4) Following behavior

When one or more artificial fish find food, fish nearby will follow. Within a visual scope of \(X(d_j < \text{visual})\), \(X_{\text{max}}\) is the position with the highest food concentration. If \(\sum n_f < \delta Y_i\), it means \(X_{\text{max}}\) is not too crowded with higher food concentration; \(X_i\) moves a step toward \(X_{\text{max}}\). Otherwise, conduct preying behavior. The following behavior can be expressed as follows:

\[
X_{\text{next}} = X_i + \text{rand}() \cdot \text{step} \frac{X_{\text{max}} - X_i}{\|X_{\text{max}} - X_i\|} \tag{15}
\]

(5) Bulletin board

The bulletin board is a place to record the status of the best individual. After an iteration, each artificial fish will compare its current fitness value with the fitness value recorded on the bulletin board. If it is less than the fitness value on the bulletin board, update the bulletin board with the current position and fitness value; otherwise, the state of the bulletin board will remain unchanged. When the iteration of the whole algorithm is finished, the output value of the bulletin board is the optimal value (Bai et al. 2017).

**Improved AFSA**

In artificial fish swarm algorithm (AFSA), moving step \(\text{step}\) (hereafter \(\text{step}\)) and visual scope \(\text{visual}\) (hereafter \(\text{visual}\)) are two important parameters, influencing the optimization performance of the algorithm. But they are fixed values, remaining unchanged during the iteration process, which makes it difficult to give suitable initial values. If the values of \(\text{step}\) and \(\text{visual}\) are too large, the artificial fish can move to the global optimal value rapidly in the early stage, but there will be an oscillation phenomenon in the later stage and it is difficult to reach the optimal value accurately. If the values are too small, the artificial fish can reach the optimal value accurately but converge slowly, easy to fall into local optimum.

In order to tackle the problem aforementioned, an adaptive learning strategy of \(\text{step}\) and \(\text{visual}\) is proposed in this study. Inspired by the characteristics of sigmoid function, a dynamical weight is introduced to adjust \(\text{step}\) and \(\text{visual}\) of artificial fish to balance local search and global search. In the iteration process of IAFSA, the larger \(\text{step}\) and \(\text{visual}\) make the algorithm quickly converge and jump out of the local optimum in the early stage. The smaller \(\text{step}\) and \(\text{visual}\) can make the algorithm accurately approach the global optimum and improve the accuracy in the later stage. The improved \(\text{step}\) and \(\text{visual}\) are expressed as follows:

\[
\mu = 1 - \frac{1}{1 + e^{-\frac{\text{stepiter}}{\alpha}}} \tag{16}
\]

\[
\text{stepiter} = \text{stepiter-1} \cdot \mu \tag{17}
\]

\[
\text{visualiter} = \text{visualiter-1} \cdot \mu \tag{18}
\]
where $\mu$ represents the dynamic adjustment function (Fig. 1), $\alpha$ and $\beta$ are adjustment parameters, iter is the current number of iterations, and iter$_{\text{max}}$ is the maximum number of iterations.

Figure 1 shows the relationship between the value of $\mu$ and iterations. It can be found that the value of $\mu$ will become smaller and smaller with iteration increasing. According to formula (17) and formula (18), we can conclude that values of step and visual have the same characteristics. Therefore, IAFSA can move to the optimal value quickly with larger step and visual in the early stage and reach the optimal value with smaller step and visual in the later stage.

The process of IAFSA is as follows: The flow chart of IAFSA is shown in Fig. 2.

step 1: Initialize parameters, including population size, maximum number of iterations, moving step, visual scope, and so on.

step 2: Calculate the fitness value of each artificial fish and record the optimal fitness value and its corresponding position on the bulletin board.

step 3: Perform preying behavior, swarming behavior, and following behavior.

step 4: Compare the minimum fitness value of the three behaviors with the fitness value on the bulletin board. If less, update the bulletin board.

step 5: Update moving step and visual scope using the adaptive learning strategy.

step 6: Justify whether to reach the maximum number of iterations. If so, output the optimal fitness value and optimal position. Otherwise, continue to perform step 3 to step 6.

**Performance test**

To verify the performance of IAFSA, this study selects three standard test functions to simulate. The information of selected functions is as follows:

Schaffer function: $f_1(x,y) = 0.5 + \frac{\left(\sin\sqrt{x^2 + y^2}\right)^2 - 0.5}{\left(1 + 0.001(x^2 + y^2)\right)^2} \quad -10 < x,y < 10$

$f_1$ is a complicated two-dimensional function with lots of local minimum points and gets the global minimum value $0$ at point $(0, 0)$. Moreover, it is very difficult to find the global optimal value of $f_1$, for this function has strong volatility.

Shubert function: $f_2(x,y) = \left(\sum_{i=1}^{5}\cos((i + 1)x + i)\right) + \left(\sum_{i=1}^{5}\cos((i + 1)y + i)\right) \quad -10 < x,y < 10$

$f_2$ is also a complicated two-dimensional function with hundreds of local optimal values. The global minimum value of $f_2$ is $-186.7309$.

Rastrigrin function: $f_3(x) = \sum_{i=1}^{n}(x_i^2 - 10 \cos(2\pi x_i) + 10) \quad |x| \leq 5.12$

$f_3$ is a nonlinear multimodal function, whose peak shape fluctuates and jumps, and it is difficult to find the global optimal value. $f_3$ gets the global minimum value $0$ at point $(0, 0, \ldots, 0)$.

During the process of test, IAFSA is compared with particle swarm optimization algorithm (PSO) and AFSA. All the algorithms run on Python 3.6.10, and the parameter settings are the same as those in the literature (Zhu et al. 2020), listed in Table 2.

Considering the results of the three algorithms are unstable, every test function was simulated 20 times independently. The results are shown in Table 2.

From Table 3, we know that IAFSA finds the global optimal value of $f_1$ and $f_2$. As for $f_3$, IAFSA has the highest accuracy of optimization. Therefore, it can be demonstrated that IAFSA has strong global search ability and better search accuracy, superior to PSO and AFSA.

**IAFSA-SVM forecasting model**

The parameters of SVM influence the forecasting performance directly, so it is crucial to select suitable parameters. In the proposed model, IAFSA is used to optimize the penalty factor $C$ and kernel parameter $\gamma$ of SVM, and the position of artificial fish represents the parameters to be optimized; the search space of artificial fish is the value range of $C$ and $\gamma$. Before running the model, we need to preprocess the raw data and then split it into two datasets: training sample and testing sample. Fitness function is the mean absolute percentage error (MAPE) of the testing sample. The position of final output on the bulletin board represents the optimal parameters ($C_{\text{best}}, \gamma_{\text{best}}$). The steps of SVM parameter optimization are the same as Fig. 2.
To improve the forecasting accuracy for NGC, we present the hybrid forecasting model IAFSA-SVM, as shown in Fig. 3. There are four parts in the IAFSA-SVM forecasting model: preprocessing module, optimization module, forecasting module, and evaluation module. In the first module, original data including influence factor data and NGC data are collected and normalized. In the second module, sigmoid function is introduced to improve AFSA and use IAFSA to optimize parameters of SVM. In the third module, we apply SVM with optimal parameters to predict NGC and obtain prediction results. In the last module, we select three evaluation metrics to analyze the error of prediction results and verify the prediction accuracy of IAFSA-SVM.

Data description

China, as the largest carbon dioxide emitter, is facing huge pressure to reduce emissions (Xu and Lin 2019). In order to reduce carbon emissions, China must change its coal-based
energy consumption structure. As a kind of clean and efficient low-carbon energy, natural gas is the best alternative to coal while renewable energy is in its infancy. Increasing the consumption of natural gas can effectively moderate the pressure of reducing emissions and improve the environment quality. In such context, China’s natural gas consumption has grown rapidly. According to the British Petroleum (BP) statistics (BP 2020), China’s NGC reached 307.33 billion m³ in 2019, an increase of 8.6% compared with 2018; since the Chinese economy entered the new normal phase in 2013, the NGC of China continues to increase and the average annual growth rate reaches 14.81%. As NGC is increasing continuously, it is very crucial for the Chinese government to formulate a reasonable natural gas plan to ensure the supply of natural gas and prevent gas shortages. Accurate prediction of NGC can avoid social economic losses caused by lack of natural gas supply and ensure energy security (Shaikh and Ji 2016). Thus, this paper selects China’s annual NGC data to validate forecasting performance of the proposed model.

NGC is usually affected by various factors such as population, GDP, industrial structure, energy consumption structure, and price. To more accurately forecast NGC, it is crucial to systematically analyze the influencing factors (Hao et al. 2020). By analyzing and summarizing the literature, this study chooses six main factors affecting China’s NGC: total population, economic development level, urbanization rate, industrial structure, energy consumption structure, and carbon dioxide emissions. The price of natural gas is also an important factor influencing NGC, but it is determined by the government in China. So, this study does not take the price into account. Influence factors are described in detail as follows:

(1) Total population (TP). Population is the basic factor of the social system, and the number of populations is one of main factors that have an impact on the NGC of a country or a region. The scale of the population directly affects the total consumption of natural gas. With the continuous improvement of economic development and consumption level, population growth will inevitably lead to an increase in NGC.

(2) Gross domestic products (GDP). GDP represents a country’s economic development level. Economic development and energy consumption influence each other; economic development cannot do without energy consumption, and energy consumption is also related to the level of economic development. Previous studies show that there is a positive relationship between NGC and GDP, and economic development level has a significant impact on NGC.

(3) Urbanization rate (UR). Urbanization rate refers to the proportion of urban population in the total population. In China, urban people consume most of residential gas. As China’s urbanization process continues to advance, the proportion of the urban population increases continuously, which means that the number of people consuming natural gas has increased.

(4) Industrial structure (IS). Industrial structure refers to the proportion of industrial added value in GDP. Due to the

Table 2 Parameter settings of different algorithms

| Algorithm       | Parameter     | Value |
|-----------------|---------------|-------|
| IAFSA & AFSA    | Population size | 50    |
|                 | Iteration number | 100   |
|                 | Max try number | 50    |
|                 | Step | 10    |
|                 | Visual | 10    |
|                 | $\delta$ | 0.618 |
| IAFSA           | $\alpha$ | 0.5   |
|                 | $\beta$ | 0.12  |
| PSO             | Population size | 50    |
|                 | Iteration number | 100   |
|                 | $w$ | 0.6   |
|                 | c1 & c2 | 1.49  |
|                 | vmin | $-1$  |
|                 | vmax | 1     |

Table 3 Results of different algorithms

| Test function | Algorithm | Best value | Worst value | Mean value | Standard value | Optimal value | $f_{optimal}$ |
|---------------|-----------|------------|-------------|------------|----------------|---------------|---------------|
| $f_1$         | IAFSA     | 0.0        | 0.009715    | 0.004857   | 0.004395       | 0.0           | 0.0           |
|               | AFSA      | 8.1851e−05| 0.006203    | 0.002314   | 0.001698       |               |               |
|               | PSO       | 1.1102e−16| 0.009715    | 0.008258   | 0.003469       |               |               |
| $f_2$         | IAFSA     | −186.7309  | −186.7309   | −186.7309  | 2.6203e−14     | −186.7309     | −186.7309     |
|               | AFSA      | −186.7176  | −183.9554   | −186.0073  | 0.6637         |               |               |
|               | PSO       | −186.7309  | −186.7309   | −186.7309  | 1.0608e−12     |               |               |
| $f_3$         | IAFSA     | 8.8817e−15| 1.9899      | 0.7135     | 0.5372         | 8.8817e−15    | 0.0           |
|               | AFSA      | 2.6138     | 5.6050      | 3.9116     | 0.8638         |               |               |
|               | PSO       | 1.0468e−10| 2.9848      | 0.7465     | 0.8250         |               |               |

$f_{optimal}$ represents the optimal value of the corresponding test function.
difference of the energy consumption index among industries, energy consumption will be affected with the adjustment of the industrial structure. Since China’s natural gas is mainly consumed by secondary industry, this study uses the proportion of secondary industry in GDP to represent the industrial structure.

(5) Energy consumption structure (ECS). Energy consumption structure refers to the proportion of NGC in primary energy consumption. Changes in the energy consumption structure will lead to changes in NGC. With the continuous optimization of China’s energy structure, the proportion of NGC in primary energy consumption increases, which will inevitably increase NGC.

(6) Carbon dioxide emissions (CE). As the largest carbon dioxide emitter, reducing carbon emissions is an urgent task for China. Natural gas is a kind of clean energy with low-carbon emissions, which gets more and more attention from government and enterprises. The Chinese government also formulated some policies to reduce emissions. These policies have accelerated the development of the natural gas market. Thus, carbon dioxide emission influences NGC to a certain degree, especially during the period energy transition.

To verify whether there is correlation between the selected factors and NGC, we choose the Pearson correlation coefficient (PCC) (Eq. 19) for analysis in this study. The PCC is generally used to investigate the correlation between two variables. Because of sampling error, the correlation coefficient should be tested for significance to justify whether there is a significant correlation relationship. The significance level is 5% in this study. Under 5% significance, if $p$ is less than 5%, it means that there exists a significant correlation relationship between two variables. Calculation results are listed in Table 3.

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}$$  \hspace{1cm} (19)$$

where $r$ is the correlation coefficient, $N$ is the number of variables, $x_i$ is the independent variable, and $y_i$ is the dependent variable.

Table 4 shows that all selected factors have passed the significance test ($p < 0.05$), which means those factors are correlated with NGC. GDP has the largest correlation coefficient (0.99), that is to say, GDP has a significant impact on NGC. The second largest one is energy consumption structure, and the correlation coefficient is 0.97. Industrial structure has a weak negative correlation with NGC.

The NGC data of China were collected from 2019 British Petroleum Statistical Review of World Energy (BP 2019). The data of the influence factors were collected from the 2019 China Statistical Yearbook (NBSC 2019) and 2019 British Petroleum Statistical Review of World Energy (BP 2019). All data are annual data from 1978 to 2018, and the detailed description is shown in Table 5. In this study, we split the data into two parts: training sample (1978–2012) and testing sample (2013–2018). The training sample is used to fit the model, and the testing sample is used to verify the prediction accuracy of the model.

To prevent the different dimensions of influence factors from affecting the forecasting effect, all data are normalized. The formula is expressed as

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$  \hspace{1cm} (20)$$

where $X$ is the original data, $X_{\text{max}}$ is the maximum value, $X_{\text{min}}$ is the minimum value, and $X'$ is the normalized data, $X' \in [0, 1]$.

**Experimental results and discussion**

**Evaluation metrics**

To evaluate the forecasting accuracy of models, three error evaluation methods, widely used in previous literature, are selected as evaluation metrics, including root mean squared error (RMSE), mean absolute error (MAE), and mean
absolute percentage error (MAPE). The smaller the value of RMSE, MAE, and MAPE is, the higher the forecasting accuracy is. Those metrics can be expressed as

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

(21)

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

(22)

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

(23)

where $y_i$ is the actual value, $\hat{y}_i$ is the forecasting value, and $n$ is the number of predictions.

**Benchmark models and parameter settings**

To verify the superiority of IAFSA-SVM, this study selects support vector machine (SVM), artificial neural network (ANN), partial least squares regression (PLS), support vector machine optimized by particle swarm optimization algorithm (PSO-SVM), and support vector machine optimized by artificial fish swarm algorithm (AFSA-SVM) as benchmark models to compare with the proposed model. Among them, PLS is one of the commonly used regression models, which can eliminate the multicollinearity between influence factors and avoid spurious regression. ANN is one of the most commonly used models in machine learning models, and it is also applied to predict annual natural gas consumption in the literature (Laib et al. 2016). The reason for choosing SVM is to show the importance of parameter optimization for forecasting performance of SVM. PSO-SVM is a classic hybrid prediction model based on SVM. AFSA-SVM is used to compare the changes in prediction accuracy before and after the improvement of AFSA, highlighting the advantages of the proposed model.

All the models run on Python 3.6.10. Parameter settings of the different models are shown in Table 6.

**Results and analysis**

Figure 4 shows the real value of NGC and the prediction results of different models. Among all the models, the forecasting curve of IAFSA-SVM has the best fitting effect, the forecasting curve of ANN has a certain deviation from the real curve, and PLS has the worst fitting effect. Figure 5 is the boxplot of the relative errors of each model. Results show that the relative error of IAFSA-SVM is the smallest, and the

### Table 4 Correlation coefficients and $p$ between selected factors and NGC

| Influence factors               | $r$ | $p$ (5%) |
|---------------------------------|-----|----------|
| Total population (TP)           | 0.72| 0.000    |
| Gross domestic products (GDP)   | 0.99| 0.000    |
| Industrial structure (IS)       | −0.43| 0.005  |
| Urbanization rate (UR)          | 0.88| 0.000    |
| Energy consumption structure (ECS) | 0.97| 0.000    |
| Carbon dioxide emissions (CE)   | 0.90| 0.000    |
| Natural gas consumption (NGC)   | 1   | −        |

### Table 5 Statistical features of NGC and influence factors

|                         | Time range | Max    | Min    | Mean    | Std.    |
|-------------------------|------------|--------|--------|---------|---------|
| NGC (unit: BCM)         | 1978–2018  | 283    | 12     | 41      | 61.69   |
| TP (unit: 10 K)         | 1978–2018  | 139,538| 96,259 | 121,593.86 | 13,321.37 |
| GDP (unit: 100 M Yuan)  | 1978–2018  | 900,309.5| 3678.7 | 205,798.36 | 261,039.49 |
| IS (%)                  | 1978–2018  | 48.1   | 40.1   | 44.76   | 2.22    |
| UR (%)                  | 1978–2018  | 59.58  | 17.92  | 36.15   | 12.89   |
| ECS (%)                 | 1978–2018  | 7.43   | 1.72   | 2.94    | 1.47    |
| CE (unit: MT)           | 1978–2018  | 9428.7 | 1418.5 | 4560.44 | 2923.62 |

### Table 6 Parameter settings of different models

| Algorithm             | Parameter  | Value |
|-----------------------|------------|-------|
| IAFSA & AFSA          | Population size | 50    |
|                       | Iteration number | 30    |
|                       | Max try number   | 50    |
|                       | Step            | 10    |
|                       | Visual          | 10    |
|                       | $\delta$        | 0.618 |
| IAFSA                 | $\alpha$       | 0.5   |
|                       | $\beta$         | 0.12  |
| PSO                   | Population size | 50    |
|                       | Iteration number | 30    |
|                       | $w$             | 0.6   |
|                       | $c_1$ & $c_2$    | 1.49  |
|                       | $v_{min}$       | $-1$  |
|                       | $v_{max}$       | 1     |
Fig. 4 Forecasting results of different models

Fig. 5 Boxplot of relative error for different models
relative error of PLS is the largest. The relative error of all models is no more than 5%.

To further illustrate the prediction accuracy of IAFSA-SVM, the MAPE, RMSE, and MAE were calculated in Table 5.

Table 7 shows the comparison of the prediction accuracy of the different models. Obviously, compared with other models, IAFSA-SVM has the smallest values of MAPE, RMSE, and MAE. It means that the prediction accuracy of IAFSA-SVM is higher than that of PLS, ANN, SVM, PSO-SVM, and AFSA-SVM. Compared with AFSA-SVM, the values of three evaluation indices of IAFSA-SVM reduce by 21.99%, 14.10%, and 27.03%, which proves that the IAFSA has obvious advantages. The evaluation metrics of SVM are smaller than those of ANN, which indicates that SVM has better forecasting ability for small samples and is more suitable for annual NGC forecasting. Compared with basic SVM, hybrid models (PSO-SVM, AFSA-SVM, IAFSA-SVM) have higher prediction accuracy, which demonstrates that SVM is sensitive to parameters and it is essential to optimize its parameters.

**Future natural gas consumption trends (2020–2025)**

In this study, IAFSA-SVM is applied to predict China’s NGC from 2020 to 2025. However, forecasting the future’s NGC requires knowing the future values of influence factors. So, to accurately forecast NGC in the future, it is crucial to obtain accurate values of influence factors. Due to the impact of covid-19 in 2020, world economy will decline seriously. According to the forecasting results of the China macroeconomic forum (CMF 2020), China’s GDP growth rate will be about 3.0% in 2020. The macroeconomic research center of the Chinese Academy of Social Sciences predicted that the average growth rate of China’s economy will be 5.5% during the 14th Five-Year plan period (2021–2025) (Cass 2020). These two research institutions have certain authority and representativeness in China. Therefore, this study uses their forecasting results as the GDP growth rate of China from 2020 to 2025. Other influence factors are weakly affected by covid-19, assuming the trend remains unchanged. Since they are all small samples, GM (1,1) has a good prediction ability for small samples and poor information data. Therefore, this study uses GM (1,1) to predict the values of other influence factors from 2020 to 2025 (Table 8). Combining with values in Table 6, we apply IAFSA-SVM to predict the NGC of China from 2020 to 2025. The forecasting results are shown in Table 8.

Table 9 shows forecasting results and growth rate of NGC from 2020 to 2025. From Table 9, we can see that China’s NGC will increase continuously in the next few years, which will reach 505.19 billion m$^3$ in 2025. As for growth rate, forecasting results show that the growth rate will increase to 9.47% in the first 2 years and the $n$ decreases gradually. In 2025, the growth rate of the NGC is 7.51%, decreasing by about 2% from 2021.

**Conclusions and future work**

NGC is affected by many external factors, and the relationship between the influence factors and NGC is nonlinear.

**Table 7** Results of the evaluation index

| Model     | MAPE (%) | RMSE   | MAE   |
|-----------|----------|--------|-------|
| PLS       | 4.1114   | 13.3998| 9.6833|
| ANN       | 1.9471   | 7.5223 | 4.8206|
| SVM       | 1.3759   | 3.7720 | 3.1340|
| PSO-SVM   | 0.5981   | 1.6411 | 1.3656|
| AFSA-SVM  | 0.6564   | 1.7414 | 1.4993|
| IAFSA-SVM | 0.5120   | 1.4958 | 1.0940|

Italicized values represent the minimum values.

**Table 8** Forecasting results of influence factors from 2020 to 2025

| Year | GDP (100 M) | TP (10 K) | UR (%) | IS (%) | ECS (%) | CE (MT) |
|------|-------------|-----------|--------|--------|---------|---------|
| 2020 | 1,020,590.95| 141,059.64| 62.25  | 38.84  | 8.55    | 9424.23 |
| 2021 | 1,081,826.41| 141,781.63| 63.57  | 38.29  | 9.29    | 9472.16 |
| 2022 | 1,146,735.99| 142,507.30| 64.92  | 37.75  | 10.09   | 9520.34 |
| 2023 | 1,215,540.15| 143,236.69| 66.30  | 37.21  | 10.97   | 9568.77 |
| 2024 | 1,288,472.56| 143,969.82| 67.71  | 36.68  | 11.91   | 9617.44 |
| 2025 | 1,365,780.91| 144,706.69| 69.14  | 36.16  | 12.94   | 9666.36 |
Moreover, the amount of historical annual NGC time series data is small. These characteristics increase the difficulty of predicting NGC. In this study, a novel hybrid forecasting model (IAFSA-SVM) is proposed to improve the prediction accuracy of NGC. In this model, IAFSA is used to optimize SVM parameters. The experimental results reveal that the proposed model has higher forecasting accuracy than benchmark models. The conclusions are as follows:

(1) An adaptive learning strategy based on sigmoid function is introduced to improve the optimization ability of AFSA. The performance test results show that the improved AFSA algorithm has stronger global search ability and higher optimization accuracy.

(2) IAFSA is used to optimize SVM parameters and establish the hybrid IAFSA-SVM model to predict NGC.

(3) The annual NGC data from China is used to verify the prediction performance of the proposed model. Compared with benchmark models, the proposed model has higher prediction accuracy.

(4) Through predicting China’s NGC in 2020–2025, it can be found that the NGC in China will increase continuously in the next few years and the consumption will reach 505.19 billion m$^3$ in 2025.

Although the proposed hybrid model is superior in forecasting accuracy, there are still some problems worthy of further research. For example, many other factors have an influence on NGC, which are not fully considered in this study. In future research, other factors can be added to improve the prediction accuracy. As one of the hottest forecasting tools nowadays, deep learning can be introduced into the field of NGC prediction in the future. The model also can be applied to other energy forecasting fields, such as oil consumption forecasting and renewable energy consumption forecasting.

Acknowledgments The authors gratefully thank Associate Professor Ming-Liang Qi and Professor Bao-Guang Xu for their suggestions and support. The authors also thank all anonymous reviewers for their help and comments. The conclusions in this paper represent the views of the authors only and do not reflect the views of any governmental or nongovernmental organization.

Authors’ contributions Feng Gao is responsible for the conceptualization, methodology, formal analysis, and writing—original draft. Xueyan Shao is responsible for conceptualization and writing—review and editing. All authors read and approved the final manuscript.

Data availability The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with ethical standards Conflict of interest The authors declare that they have no competing interests.

References

Akpinar M, Adak MF, Yumusak N, 2016. Forecasting natural gas consumption with hybrid neural networks — artificial bee colony, 2016 2nd International Conference on Intelligent Energy and Power Systems (IEPS). pp. 1-6.

Bai S, Wang L, Wang, X., 2017. Optimization of ejector geometric parameters with hybrid artificial fish swarm algorithm for PEM fuel cell, 2017 Chinese Automation Congress (CAC). pp. 3319-3322.

Bai Y, Li C (2016) Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach. Energy Build 127:571–579

Beyca OF, Ervural BC, Tatoglu E, Ozyurar PG, Zaim S (2019) Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. Energy Econ 80:937–949

BP, (2019). The British Petroleum (BP) statistical review of world energy. BP, (2020). Energy outlook 2020 edition.

Cass MRC, 2020. Research on the potential growth of the Chinese economy in the next 15 years and the main goals and indicators of economic and social development during the 14th five-year plan period. China Ind Econ (04), 5-22.

Chen X, Wang J, Sun D, Liang J, 2008. Time series forecasting based on novel support vector machine using artificial fish swarm algorithm, 2008 Fourth International Conference on Natural Computation. pp. 206-211.

Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17(1):113–126

CMF, (2020). China Macroeconomic Forum.

Ding S (2018) A novel self-adapting intelligent grey model for forecasting China’s natural-gas demand. Energy 162:393–407

Gao J, Dong X (2018) Stimulating factors of urban gas consumption in China. Nat Gas Ind 38(3):130–137

Hao J, Sun X, Feng Q (2020) A novel ensemble approach for the forecasting of energy demand based on the artificial bee colony algorithm. Energies 13(3):550

Hongxun L, Hui N (2019) Natural gas demand forecast in China based on gray-partial least square combination model. Resources Ind 21(06):9–19

Huang C-L, Wang C-J (2006) A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl 31(2):231–240

Jiang AN, Wang SY, Tang SL (2011) Feedback analysis of tunnel construction using a hybrid arithmetic based on support vector machine and particle swarm optimisation. Autom Constr 20(4):482–489

Karabiber OA, Xydís G (2020) Forecasting day-ahead natural gas demand in Denmark. J Nat Gas Sci Eng 76:103193

Karadee Y, Ozdemir G, Aydemir E (2017) Breeder hybrid algorithm approach for natural gas demand forecasting model. Energy 141:1269–1284

Kavaklioglu K (2011) Modeling and prediction of Turkey’s electricity consumption using support vector regression. Appl Energy 88(1):368–375

Khan MA (2015) Modelling and forecasting the demand for natural gas in Pakistan. Renew Sust Energ Rev 49:1145–1159

Lalib O, Kadhir MT, Chouireb L, 2016. Forecasting yearly natural gas consumption using artificial neural network for the Algerian market, 2016 4th International Conference on Control Engineering & Information Technology (CEIT). pp. 1-5.
Li X, Shao Z, Qian J, 2002. An optimizing method based on autonomous animats: fish-swarm algorithm. Syst Eng-Theory Pract (11):32-38.

Liu G, Dong X, Jiang Q, Dong C, Li J (2018) Natural gas consumption of urban households in China and corresponding influencing factors. Energy Policy 122:17–26

Lu H, Azimi M, Isley T (2019) Short-term load forecasting of urban gas using a hybrid model based on improved fruit fly optimization algorithm and support vector machine. Energy Rep 5:666–677

Lu H, Ma X, Azimi M (2020) US natural gas consumption prediction using an improved kernel-based nonlinear extension of the Arps decline model. Energy 194:116905

Ma Z, Ye C, Ma W (2019) Support vector regression for predicting building energy consumption in southern China. Energy Procedia 158:3433–3438

Melikoglu M (2013) Vision 2023: Forecasting Turkey’s natural gas demand between 2013 and 2030. Renew Sust Energ Rev 22:393–400

Meng Q, Ma X, Zhou Y (2014) Forecasting of coal seam gas content by using support vector regression based on particle swarm optimization. J Nat Gas Sci Eng 21:71–78

Mirzaei Z, Akbarpour A, Khatibinia M, Siuki AK (2015) Optimal design of homogeneous earth dams by particle swarm optimization incorporating support vector machine approach. Geomech Eng 9:6

NBSC, 2019. China statistical yearbook.

Niu D, Dai S (2017) A short-term load forecasting model with a modified particle swarm optimization algorithm and least squares support vector machine based on the denoising method of empirical mode decomposition and grey relational analysis. Energies 10(3)

Ozdemir G, Aydemir E, Olgun MO, Mulbay Z (2016) Forecasting of Turkey natural gas demand using a hybrid algorithm. Energy Sources, Part B: Econ, Plan, Policy 11(4):295–302

Rafindadi AA, Ozturk I (2015) Natural gas consumption and economic growth nexus: is the 10th Malaysian plan attainable within the limits of its resource? Renew Sust Energ Rev 49:1221–1232

Sen D, Güney ME, Tunç KMM (2019) Forecasting annual natural gas consumption using socio-economic indicators for making future policies. Energy 173:1106–1118

Shaikh F, Ji Q (2016) Forecasting natural gas demand in China: logistic modelling analysis. Int J Electr Power Energy Syst 77:25–32

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