A novel framework of credit risk feature selection for SMEs during industry 4.0

Yang Lu¹,² · Lian Yang¹,² · Baofeng Shi¹,² · Jiaxiang Li¹,² · Mohammad Zoynul Abedin³

Accepted: 20 June 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
With the development of industry 4.0, the credit data of SMEs are characterized by a large volume, high speed, diversity and low-value density. How to select the key features that affect the credit risk from the high-dimensional data has become the critical point to accurately measure the credit risk of SMEs and alleviate their financing constraints. In doing so, this paper proposes a credit risk feature selection approach that integrates the binary opposite whale optimization algorithm (BOWOA) and the Kolmogorov–Smirnov (KS) statistic. Furthermore, we use seven machine learning classifiers and three discriminant methods to verify the robustness of the proposed model by using three actual bank data from SMEs. The empirical results show that although no one artificial intelligence credit evaluation method is universal for different SMEs’ credit data, the performance of the BOWOA-KS model proposed in this paper is better than other methods if the number of indicators in the optimal subset of indicators and the prediction performance of the classifier are considered simultaneously. By providing a high-dimensional data feature selection method and improving the predictive performance of credit risk, it could help SMEs focus on the factors that will allow
them to improve their creditworthiness and more easily access loans from financial institutions. Moreover, it will also help government agencies and policymakers develop policies to help SMEs reduce their credit risks.

**Keywords** Credit rating · Credit risk · Feature selection · SMEs · Binary opposite whale optimization algorithm · Kolmogorov–Smirnov statistic

**JEL Classification** C53 · D81 · G21 · E51

1 **Introduction**

A default occurs when a borrower fails to repay the loan in full. Default risk identification has become one of the most important issues in the banking industry (Dia et al., 2022; Granja et al., 2022; Oreski & Oreski, 2014; Orth, 2012). Credit scoring is a common method of assessing the credit risk of loan clients and is also the basis of credit ratings, loan decisions, and risk management for financial institutions (Crone & Finlay, 2012). For example, in application scoring, lenders often use statistical or artificial-intelligence-based predictive models to estimate the probability of an applicant’s risk of default. The financial crisis of 2008 caused huge losses around the world. Since then, banks and financial institutions have continued to pay attention to the credit risk prediction (Altman et al., 2020).

Industry 4.0 is opening up a new era of information technology, bringing great convenience and challenges to risk management. The term Industry 4.0 refers to various technologies used to facilitate the development of value chains and shorter manufacturing lead times, improved product quality, and organizational performance (Kamble et al., 2018). It provides new paradigms and opportunities for the industrial management of SMEs. Supported by big data technology and artificial intelligence algorithms, it is more convenient and cheaper than traditional enterprise management systems. However, SMEs still face credit risks in Industry 4.0. In the era of Industry 4.0, credit data of SMEs comes from a wide range of sources. Governments have SMEs’ industrial and commercial registration, judicial, tax information, banks have SMEs’ credit information such as lending, repayment, guarantee and default, SMEs have their own operation, sales, employment data, and social media platforms disclose information such as public opinion related to SMEs. The existence of big data and data informatization makes information acquisition very convenient, but also often leads to noisy information of a high dimensionality (Duan et al., 2019). Credit data is often composed of multiple dimensions. The total number of indicators is large and there are many missing values. Therefore, the feature selection of high-dimensional credit data for SMEs is particularly important before establishing credit scoring models.

The feature selection approaches include the filter, wrapper, and hybrid methods (Hinton & Salakhutdinov, 2006; Sefidian & Daneshpour, 2018). Unlike the other two methods, the filter method is used separately from the classification, so that it runs quickly and efficiently but cannot be applied to measure the default-discriminating performance of the selected indicators (Freeman et al., 2015; Nakariyakul, 2018). The indicator selection of the wrapper approach is regarded as a search optimization problem. Different indicator subsets are generated using an optimization algorithm, the classification model is used to evaluate those indicator subsets, and then the best one is selected. In general, the wrapper approach is better than the filter approach (Chen et al., 2020). The hybrid technique is a compromise between the filter and wrapper techniques with low computational complexity and high model accuracy.
The above methods have played an important role in credit risk assessment, but there are few literature reports on the application of high-dimensional data reduction under the characteristics of big data.

In this paper, a novel binary opposite whale optimization algorithm (BOWOA) is proposed to screen features that have significant discrimination on SME’s credit risk. This method combines the advantages of the binary whale optimization algorithm (BWOA) with the opposition-based learning (OBL). It can quickly realize feature screening of high-dimensional credit data for SMEs.

The contributions of this paper include the development of a novel wrapper-type credit risk feature selection methodology that combines the BOWOA with the KS method. In short, we combine the advantages of these two to construct a model which can quickly identify the crucial default features for SMEs. The proposed model enables us to select the smallest subset of indicators with the most information, from the original indicator set. Another contribution of this paper is in mining the relationship between the credit characteristics and credit risk of SMEs in the context of Industry 4.0. More broadly, the empirical analysis was conducted using Chinese credit data for 2044 SMEs, with robustness tests carried out using data for 2157 SMEs and 3111 SMEs. The proposed method can greatly reduce the computational complexity of traditional artificial intelligence wrapper techniques. This study also addresses some policy implications. Bank analysts can refer to the model to quickly check the credit default features of SMEs. Managers of SMEs can use the model to conduct risk assessments of their own companies to better understand their companies’ risk status, make credit decisions, and reduce their credit risk. Moreover, as feature selection is a common problem in data mining, this model can also be used to attribute reduction of high-dimensional data in other fields.

In the remainder of the paper, related work is reviewed in Sect. 2. Then, the methodology is described in Sect. 3. Next, the empirical results are presented in Sect. 4. Finally, the discussion and conclusions are provided in Sects. 5 and 6, respectively.

2 Literature review

SMEs are the most common form of business and play an important role in economic development. Many literatures have been analyzed and reported on the study of financing status of SMEs (Bagale et al., 2021; Beck et al., 2005). Financial institutions and bank managers have always paid attention to the effective prediction of SMEs’ default risk (Ciampi et al., 2021; Edmister, 1972; Laitinen, 1993; Medina-Olivares et al., 2022; Shi et al., 2018). However, SMEs face more difficulties than larger companies in their risk management processes (Ciampi, 2015). The technology and management models used by large companies are usually not fit for SMEs because they are either too complex or too expensive (Pereira et al., 2013). Compared with larger companies, SMEs often face different challenges and dilemmas (Wadhwa, 2012). The financing needs of large enterprises have benefited from the development of information technology and artificial intelligence technology, but small enterprises have benefited less in this regard because small enterprise loans are still constrained due to the large amounts of missing credit data. Researchers have examined various opportunities and challenges faced by SMEs and found that their production, operation and scale expansion rely heavily on access to capital investment and the returns on those investments. One very important feature affecting the sustainable development of SMEs is that it is difficult to guarantee the credit funds required for operation. Moreover, SMEs are more vulnerable to risk
exposure than large companies, and should therefore be more engaged in risk management (Blanc Alquier & Lagasse Tignol, 2006). They face more challenges in accessing resources, have less diversification of economic activity, and weaker financial structures, and tend to face layoffs in the event of a crisis (Kim & Vonortas, 2014). Although, in the era of Industry 4.0 credit risk management is crucial to the development of sustainability and business value for SMEs, an effective risk management model for them is far from being adopted (Pereira et al., 2013). All these factors are encouraging more and more researchers, policymakers, and lenders to focus on ways of improving the precision of risk prediction and lowering the default rates of small enterprises (Oliveira et al., 2017). Therefore, it is necessary to build a complete credit risk management system for SMEs, and credit risk feature selection will be the cornerstone of the establishment of such a credit system.

The feature selection technique has been widely used to diminish the dimensions of data, reduce the computational complexity of models, and improve their accuracy (Mehmanchi et al., 2021). In general, feature selection involves picking an optimal subset of indicators from the original set of indicators. This optimal subset will contain fewer indicators, but still fully reflect the information of the original set (Shi et al., 2018). If the original set is assumed to have n indicators, then the computational complexity of the problem will be $2^n$, which turns into a non-deterministic polynomial (NP) problem. There is no direct method of obtaining such a subset, and most algorithms are based on approximate solutions. Scholars usually divide feature selection techniques into three categories: filter, wrapper, and hybrid (Kohavi & John, 1997; Lappas & Yannacopoulos, 2021).

The filter method is a statistical feature selection algorithm with low computational complexity. It directly filters the optimal feature subset using the statistical characteristics of the sample data. Therefore, this kind of algorithm is very suitable for data with a large sample size. Chi and Zhang (2017) built a feature selection model for SMEs using the rank-sum test, which avoids the assumption that the training data follows a normal distribution. By maximizing the Gini coefficient with nonlinear programming, Zhang et al. (2022) found the crucial credit risk indicators of Chinese SMEs. They also found that a single indicator may have a strong default discrimination ability but, when combined with other indicators, it may not necessarily have the same ability. Thus, it is important to screen the optimal combination of indicators.

The wrapper method is proposed by John et al. (1994). Compared with the filter algorithm, the wrapper algorithm is more accurate, but its computational complexity is greater. In order to seek the best subset of features, wrapper algorithms need to train themselves using classification models. These characteristics make wrapper algorithms powerless in the big data era. Maldonado and Weber (2009) constructed a wrapper algorithm with the support vector machine (SVM) model as a kernel function to screen the optimal subset of features for identifying default status. Chen and Xiang (2017) used the group least absolute shrinkage and selection operator (LASSO) method to select a group of optimal features. To get the key indicator for oil credit, Bellotti and Crook (2009) used the SVM model and compare it to logistic regression, and found the SVM model to be competitive for feature selection.

The hybrid method combines the advantages of the filter and wrapper algorithms. The idea of this kind of technique is to filter out a pool of features using the filter algorithm, and then use the wrapper method on the subsets from this pool. Therefore, the hybrid algorithm makes a trade-off between the accuracy of the model and computational complexity. Oreski and Oreski (2014) combined the genetic algorithm (GA) and neural networks, proposing a hybrid feature selection algorithm for evaluating credit risk. Chen and Li (2010) introduced a hybrid feature selection method based on the SVM, and their study shows that the hybrid method is robust. More recently, Lappas and Yannacopoulos (2021) constructed a machine
learning framework combining expert experience with the GA for hybrid feature selection for credit risk evaluation.

The techniques used for searching the characteristic space in the wrapper and hybrid strategies are different. Because of the computational complexity of the problem, artificial intelligence optimization algorithms are commonly used, such as the GA, particle swarm optimization (PSO), and the whale optimization algorithm (WOA) (Bhattacharya et al., 2018; Kaur & Arora, 2018). The WOA is an optimization algorithm that simulates the natural behavior of humpback whales (a spiral motion and bubble-net foraging). It was proposed by Mirjalili and Lewis (2016) and is used for parameter optimization and large-scale function optimization. Compared with PSO, the WOA has higher storage efficiency because it only stores the globally optimal value during the iteration. Aljarah et al. (2016) applied the WOA to the optimization of weights and biases in artificial neural networks. Compared with PSO, the GA, and back propagation (BP), the WOA can not only avoid local optimization, but also improve convergence speed. Tharwat et al. (2017) used the WOA to find the optimal model parameters of the SVM so as to reduce classification error. To prevent the WOA from becoming trapped in a local optimum, Sayed et al. (2018) applied chaos theory, using the regularity and semi-randomness of the chaotic system to improve the WOA’s global optimization ability.

Although the existing research has achieved great success in dealing with feature selection issues for credit risk, it still has some limitations. First of all, when selecting a subset of indicators, most studies only consider the accuracy of the model, and not the total number of indicators in the subset, thus neglecting to consider the calculation cost and the data collection cost. Secondly, the computational complexity of the existing wrapper algorithms is very high, and they are inefficient at dealing with large-scale data. In the Industry 4.0 era, there is little research on the use of wrapper algorithms for large-scale credit risk samples with its high-dimension datasets. Therefore, a novel feature selection model, including more optimization algorithms and performance metrics, should be proposed. It will guide banks’ credit decision-makers and help them select the most appropriate methods for their own unique datasets.

3 Methodology

In this section, we put forward a novel credit risk feature selection model for SMEs. To begin with, the original data is standardized and sample expansion is performed on the default sample. Then, the BOWOA artificial intelligence optimization algorithms and the KS statistic are used to find the optimal subset of indicators that can accurately identify the default risk of SMEs. Finally, three measures including accuracy, area under the curve (AUC) and f1 scores are utilized to evaluate the predictive performance of the model. The reasons for these steps will be discussed in the following.

The first step is to collect SMEs’ credit data. Assuming that the sample consists of $n$ indicators and $m$ customers, the SMEs’ credit data can be described as a matrix $X$ with $m$ rows and $n$ columns. We expect to select $k$ indicators from these $n$ indicators, such that using the newly obtained data matrix $X'$ to perform default prediction is better than using the original dataset. The number $k$ of indicators should be as small as possible so that when banks collect small enterprises data, they will only need to collect $k$ indicators instead of $n$. Fewer indicators mean that the banks’ data collection costs will be lower. Also, the fewer indicators there are, the faster the computer modeling will be. In addition, since normalizing
sample data can enhance the performance of machine learning classification models, after collecting the data, necessary data cleaning and data normalization are needed. Because the credit data of SMEs is often imbalanced, this could lead to the model over-identifying SMEs that do not default. Studies have shown that sample expansion techniques have an active effect on the predictive performance of models. Therefore, sample expansion is performed on the original SMEs’ credit data.

Then, a new BOWOA method is proposed to reduce features. It is well known that the WOA is an emerging artificial intelligence optimization algorithm for high-dimensional problems (Zhang & Wen, 2021). In order to apply WOA to discrete problems, Hussien et al. (2020) propose the BWOA, which differs from the WOA as it replaces the continuous whale agent with a binary whale agent. Lappas and Yannacopoulos (2021) show that the OBL method can improve the prediction performance of unbalanced credit samples. This study combines the advantages of the OBL with BWOA, and proposes the BOWOA algorithm for feature screening of high-dimensional unbalanced credit data.

Based on the motivation of this study, those indicators that have strong default discrimination power are needed. To do so, a quantity is required to measure the default discriminatory ability of the model. This study uses the KS statistic to measure the default discriminatory ability of the model. Supposing the sample data consists of n indicators, and then these n indicators generate $2^n$ subsets of indicators. When n is relatively small, all these subsets of indicators can be traversed to find the optimal indicator group. Theoretically, if it was possible to use some kind of machine learning algorithm and do an analysis on these $2^n$ subsets of indicators, the KS statistic for each subset of indicators could be obtained. It is obvious that the subset of indicators with the highest KS statistic will be the group of indicators that would be best selected. Unfortunately, when n is not small, it is impossible to traverse all $2^n$ subsets of indicators and obtain their KS statistics, as it would involve huge computational complexity. The computational complexity here is mainly due to three aspects: the high dimensionality of the data, the diversity of the indicator groups, and the complexity of machine learning models.

The model is summarized into the following three steps. Step 1. Establish the original indicator set and standardize the original data. Perform sample expansion on the default sample. Step 2. Filter default features for SMEs using the BOWOA algorithm and the KS statistic by a linear scoring model. Step 3. Evaluate the predictive performance of the model using accuracy, AUC and f1 scores.
It is worth mentioning that the linear model is chosen to calculate the KS statistics and the weight $w_j$ of the linear function of each individual indicator. In practice, users can choose the SVM, random forest (RF), or some other methods to replace the linear function. For the linear scoring model $k \sum_{j=1}^{j} w_j x_{ij}$, we already know the value of the $j$-th indicator of the $i$-th small enterprise $x_{ij}$. To evaluate the KS statistic of the given subset of indicators, the weight coefficient $w_j$ of each indicator should be determined. If there is only one indicator ($k = 1$), the weight of the indicator is 1. For multiple indicators $\{x_1, x_2, ..., x_k\} (k > 1)$, the weight $w_j$ corresponding to the $j$-th indicator is defined as follows.

$$w_j = \frac{KS(\{x_j\})}{\sum_{j \leq k} KS(\{x_j\})}$$  \hspace{1cm} (1)

In Eq. (1), $KS(\{x_j\})$ represents the value of KS statistic of the $j$-th indicator. This formula allows quick calculation of the fitness function value of the BOWOA algorithm. Although the linear function looks simple, we found it to be effective in the empirical analysis.

### 3.1 Data standardization and SMOTE sample expansion

The original data should be normalized to eliminate the influence of different dimensions of features on the selection of indicators (Chai et al., 2019). The scoring standard of qualitative indicators is usually assessed through expert investigation and rational analysis (Bai et al., 2019). The standardization formulas for the positive, negative, and interval indicators are given below.

Equation (2) shows the standardization formula for positive indicators:

$$x_{ij} = \frac{v_{ij} - \min(v_{ij})}{\max(v_{ij}) - \min(v_{ij})}$$  \hspace{1cm} (2)

where $x_{ij}$ denotes the standardized score of the $i$-th SME for the $j$-th indicator, and $v_{ij}$ denotes the raw data for the $i$-th SME and $j$-th indicator.
Equation (3) shows the standardization formula for negative indicators.

$$x_{ij} = \frac{\max(v_{ij}) - v_{ij}}{\max(v_{ij}) - \min(v_{ij})}$$

Equation (4) shows the standardization formula for interval indicators.

$$x_{ij} = \begin{cases} 
1 - \frac{q_1 - v_{ij}}{\max(q_1 - \min(v_{ij}), \max(v_{ij}) - q_2)}, & v_{ij} < q_1, \quad (a) \\
1 - \frac{v_{ij} - q_2}{\max(q_1 - \min(v_{ij}), \max(v_{ij}) - q_2)}, & v_{ij} > q_2, \quad (b) \\
1, & q_1 \leq v_{ij} \leq q_2 \quad (c)
\end{cases}$$

where \(q_1, q_2\) denote the left and right boundary of the ideal interval respectively.

The SMOTE method is a method of synthesizing new samples from a few existing samples. It has been widely used in credit risk evaluation. The main idea is to interpolate between two default sample points to obtain a new sample point without losing the original data or over-fitting.

Specifically, let \(x_1\) denote a default sample point and \(x_2\) denote another default sample point. Then, a new sample point \(x_3\) can be synthesized according to the following rules (Fernandez et al., 2018):

$$x_3 = x_1 + \text{rand}(0, 1) \times (x_2 - x_1)$$

### 3.2 Construction of indicator selection model based on BOWOA

#### 3.2.1 Initialization of population of indicator subsets based on OBL

The population of subsets of indicators needs to be initialized before using the BOWOA to find the optimal subset. The OBL algorithm proposed by Tizhoosh (2005) has often been used to optimize algorithms to improve global search performance (Ali et al., 2012; Park & Lee, 2015). Thus, OBL is used to initialize the population of subsets of indicators. By introducing OBL, we ensure that the BOWOA not only considers the subset of common indicators but also the corresponding opposite subset of indicators. This lays the foundation for the global search of the BOWOA.

To search for the optimal indicator subset in binary space, it is necessary to extend the definition of the opposite point (Wang et al., 2011) to the binary opposite point in a binary space.

**Definition 1 (Opposite Point).** Let \(X = (x_1^1, x_2^2, \ldots, x_D^D)\) denote any point in a \(D\)-dimensional space, where \(x_i^j\) is any point in the real interval \([a_i^j, b_i^j]\). Then the opposite point to \(X\) can be denoted by \(OX = (ox_1^1, ox_2^2, \ldots, ox_D^D)\), where \(ox_i^j = a_i^j + b_i^j - x_i^j, i = 1, 2, \ldots, D\).

**Definition 2 (Binary Opposite Point).** Let \(X = (x_1^1, x_2^2, \ldots, x_D^D)\) denote any point in a \(D\)-dimensional binary space, where \(x_i^j = 0\) or \(1\). Then the opposite point of \(X\) can be denoted by \(OX = (ox_1^1, ox_2^2, \ldots, ox_D^D)\), where \(ox_i^j = 1 - x_i^j, i = 1, 2, \ldots, D\).

Based on Definition 2, the basic steps for initializing the population of subsets of indicators are as follows. Firstly, the indicator subset population \(IP\) is initialized. To be specific, we randomly generate \(N\) indicator subsets \(X_i = (x_i^1, x_i^2, \ldots, x_i^D)\) as the initial indicator subset population \(IP\), where \(i = 1, 2, \ldots, N\), and \(D\) is the number of original indicators. \(x_i^D \in \{0, 1\}\),
\( x_i^D = 1 \) denotes keeping this position indicator as the indicator subset element. Secondly, the opposite indicator subset population \( OP \) is generated. According to Definition 2, we can get the opposite indicator subset \( OX_i \) of the indicator subset \( X_i \) from Step 1, and obtain the opposite indicator subset population \( OP = \{OX_1, OX_2, \ldots, OX_N\} \). Thirdly, the new initial indicator subset population \( FP \) is generated. The fitness values of the \( 2N \) indicator subsets from the above two steps are calculated and they are sorted in descending order of fitness value. The first \( N \) indicator subsets are then selected and used to construct the new initial indicator subset population, which is recorded as \( FP \).

### 3.2.2 Indicator selection based on the BOWOA

Based on the initial population of subsets of indicators, \( FP \), obtained in subsection 3.3, in this section the BOWOA is used to select indicators that are significant in identifying customer default status. The BOWOA seeks the optimal subset of indicators in two ways: exploitation and exploration (Mirjalili & Lewis, 2016). When finding the optimal subset, the BOWOA adopts the exploitation method. The process is as follows.

Let \( \Phi \) be the original set of indicators, \( \varphi \) be the optimal subset of indicators, and \( X_{gb} \) be a subset that is approximate to the optimal subset \( \varphi \).

1. **Search for the optimal subset of indicators \( \varphi \) based on the exploitation method**

   In the exploitation phase, the BOWOA uses two approaches to find the optimal subset of indicators \( \varphi \). First, “shrinking encircling” is defined, as shown in Eqs. (6–9):

   \[
   X_i(t + 1) = X_{gb}(t) - R |C \cdot X_{gb}(t) - X_i(t)| \tag{6}
   \]

   \[
   R = 2a \cdot r_1 - a \tag{7}
   \]

   \[
   C = 2 \cdot r_2 \tag{8}
   \]

   \[
   a = 2 - \frac{2t}{t_{max}} \tag{9}
   \]

   where \( t \) is the current number of iterations and \( t_{max} \) is the maximum number of iterations. \( X_i(t) \) is the \( i \)-th subset of indicators of the current iteration \( t \), \( R \) and \( C \) are coefficient vectors, and \( a \) is the convergence coefficient, which is decreased linearly from two to zero. The value of \( a \) in Eq. (9) is controlled to imitate the shrinking encircling mechanism. As the convergence factor \( a \) decreases, the fluctuation range of \( A \) decreases, and the subset of indicators \( X_i(t) \) approaches \( X_{gb}(t) \).

   Second, “spiral updating” is used, as shown in Eq. (10):

   \[
   X_i(t + 1) = |X_{gb}(t) - X_i(t)| \cdot e^{-bl} \cdot \cos(2\pi l) + X_{gb}(t) \tag{10}
   \]

   where \( b \) is a constant parameter, and \( l \) is a random real number in the interval \([-1, 1]\).

   In general, the above two ways of finding the optimal subset of indicators are implemented at the same time. To model this simultaneous behavior, it is assumed that there is a probability of 50\% of choosing the spiral updating method, and an equal probability of choosing the shrinking encircling method (Hussien et al., 2020). The mathematical model is as follows:

   \[
   X_i(t + 1) = \begin{cases} 
   X_{gb}(t) - R |C \cdot X_{gb}(t) - X_i(t)| & p < 0.5 \\
   |X_{gb}(t) - X_i(t)| \cdot e^{-bl} \cdot \cos(2\pi l) + X_{gb}(t) & p \geq 0.5 
   \end{cases} \tag{11}
   \]
When $p < 0.5$, the BOWOA searches for the optimal subset of indicators through the shrinking encircling method; otherwise, it uses the spiral method. It is worth noting that, if the updated $X_i(t + 1)$ is better than $X_{gb}$ in terms of identifying default status, we need to update $X_{gb}$ using $X_i(t + 1)$ as the approximate subset to the optimal subset of indicators for the next iteration.

(2) **Search for the optimal subset of indicators $\varphi$ based on the exploration method**

It is easy to see that the optimal subset of indicators obtained by the exploitation method may be a local optimum rather than the global optimum. To alleviate this problem, $X_{gb}$ is replaced with a random subset of indicators $X_r(t)$ to force the BOWOA to do a global search. The mathematical equation is as follows:

$$X_i(t + 1) = X_r(t) - R \cdot |C \cdot X_r(t) - X_i(t)|$$

(12)

Since this paper is searching for the optimal subset of indicators in a binary space, $X_i(t + 1)$ in Eqs. (6), and Eqs. (10, 11 and 12) to be transformed into binary form (Hussien et al., 2020). The mathematical functions are as follows:

$$S(X^{k}_{i}(t + 1)) = \frac{1}{1 + e^{-X^{k}_{i}(t+1)}}$$

(13)

$$X^{k}_{i}(t + 1) = \begin{cases} 0 & \text{rand} < S(X^{k}_{i}(t + 1)) \\ 1 & \text{rand} \geq S(X^{k}_{i}(t + 1)) \end{cases}$$

(14)

where Eq. (13) is the sigmoid function that can transform any real number into the interval $(0, 1)$, $X^{k}_{i}(t + 1)$ is the value of the $i$-th subset of indicators in the $k$-th position, and $S(X^{k}_{i}(t + 1))$ is the sigmoid value corresponding to $X^{k}_{i}(t + 1)$. Equation (14) transforms $S(X^{k}_{i}(t + 1))$ into a variable that takes either 0 and 1, achieving the purpose of the BOWOA which is to search for the optimal subset of indicators in a binary space. Algorithm 2 presents the pseudo-code of the BOWOA.
There are three reasons why the BOWOA algorithm is used for feature selection of high-dimensional credit data of SMEs. First, the BOWOA algorithm can select the optimal group of indicators with strong default identification ability. Some traditional feature selection methods, such as the RF and logistic regression, can find the importance of a single indicator, but they cannot obtain the optimal group of indicators (Zhang et al., 2022). Second, compared with other optimization algorithms, the BOWOA can quickly find the global optimal solution, and it uses stochastic optimization to avoid falling into local optimality. Third, in the age of big data, credit data often presents the characteristics of high dimensionality and noise. From the empirical analysis below, the BOWOA shows it has advantages in processing high-dimensional and imbalanced data.

### 3.2.3 Model parameter selection

1. **Fitness function**
   In the artificial intelligence optimization method, the fitness function lies at its core and directly affects the speed and effect of the optimization. The linear scoring model and the subset of indicators to predict the KS statistic are chosen as the fitness value of the
BOWOA optimization algorithm, following the definition of the KS statistic given by Rezac and Rezac (2011).

The empirical distributions of the KS statistics of the default and non-default SMEs are calculated, followed by the maximum value of the difference between the two distributions under different thresholds. Supposing that each SME has a credit score, the characteristic function of default can be defined as:

$$D(k) = \begin{cases} 1, & \text{if the } k\text{-th SME is non-default} \\ 0, & \text{if the } k\text{-th SME is default} \end{cases}$$ (15)

The empirical cumulative distribution functions of the scores of the non-default and default SMEs are as follows.

$$F_{n_1,\text{non-default}}(a) = \frac{1}{n_1} \sum_{i=1}^{n_1} I((S_i \leq q) \text{ and } (D(i) = 1))$$

$$F_{n_2,\text{default}}(a) = \frac{1}{n_2} \sum_{i=1}^{n_2} I((S_i \leq q) \text{ and } (D(i) = 0)), q \in [L, H]$$ (16)

where $S_i$ is the credit score of the $i$-th SME, $n_1$ is the number of non-default SMEs, $n_2$ is the number of default SMEs, and function $I(x)$ equals 1 if and only if statement $x$ is true. $q$ is the threshold used for judging whether a small enterprise is default. $L$ and $H$ are the lower and upper bounds of all credit scores, respectively. The KS statistic is defined as

$$KS = \max_{q \in [L, H]} |F_{n_1,\text{non-default}}(q) - F_{n_2,\text{default}}(q)|$$ (17)

As an example, Fig. 1 exhibits the estimated cumulative distribution functions for default and non-default SMEs, including a rough estimate of the KS statistic (marked by a blue line). The KS statistic measures the predictive model’s power to distinguish between

![Fig. 1 Example of KS statistic](image)
default and non-default SMEs. In general, a larger KS value implies greater discrimination power. In the credit rating of SMEs, the accuracy value and the AUC value are most often applied to judge the predictive performance of models. Few studies use the KS statistic, and we believe that the introduction of the KS statistic can give researchers a new perspective. The KS statistic measures the ability of a model to distinguish between default and non-default enterprises. This study chooses the KS statistic as the metric for three reasons. First, the credit data of SMEs is often imbalanced (Mahbobi et al., 2021). The KS statistic measures the default discrimination power of a model and can show the characteristics of imbalanced SME credit data. Second, the KS statistic is a non-parametric statistic. Its calculation depends neither on the distribution of the sample nor on the threshold used to judge default. This is more suitable for the era of Industry 4.0, with its high-dimensional and noisy data. Third, the KS statistic can reveal the default discrimination ability of a single indicator or a group of indicators (Yu et al., 2018), which makes it applicable to the feature selection problem studied in our work.

(2) Parameters settings for the BOWOA-KS model

In the BOWOA-KS model, the binary sequence needs to be initialized randomly at first to represent a subset of indicators. In order to ensure that every indicator can be randomly selected with equal probability in the screening process, following random generation rules are adopted. Every time a binary sequence is randomly generated, a positive integer N with a value from 0 to n is first generated using a uniform distribution. This represents the number of numbers with a value of 1 in the binary sequence. Then the positions in N binary sequences are randomly selected and appointed as 1. The advantage is that the total number of indicators will be obtained with the same probability, and each indicator also has equal probability of being selected.

Usually, the number of iterations of the BWOA is in the range [5, 50] (Aljarah et al., 2016), and the initial population is in the range [5,100]. If the sample has n indicators, then theoretically there should be 2n subsets. Compared to 2n, the initial population size of 100 is too small. We found that, if the initial population size was small, this led to an inability to find the optimal solution. After considering the accuracy of the model and the running speed of the program, we set the initial population of the model to 1,000 and the number of iterations to 50. The fitness function value of a population of 1,000 can be calculated quickly because the model of this paper is based on using a linear function for scoring the SMEs’ credit risk. Since the weights in the linear model can be calculated based on the fitness value of each indicator, we only need to save those fitness values in advance.

3.3 Evaluation measures

The quality of the selected subset of indicators is mainly evaluated by two aspects. The first is the total number of indicators in the subset. It is desirable to have the number of indicators to be as small as possible, because this will reduce the cost of data collection and the time required for the computer modeling. The second aspect is how well credit risk can be predicted using the subset of indicators, which is captured by metrics such as accuracy, f1 score, receiver operating characteristic curve (ROC), etc. (Rezac & Rezac, 2011). These metrics can be assessed usually based on a confusion matrix. Table 1 describes the confusion
matrix for SME credit scoring.

\[ TPR = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad FPR = \frac{FP}{FP + TN} \]  

(18)

where \( TP \) represents the number of non-default SMEs that are correctly identified, and \( FN \) represents non-default SMEs that are incorrectly identified. \( FP \) represents the number of default SMEs that are incorrectly identified, and \( TN \) represents the number of default SMEs that are correctly identified. Then the Accuracy can be given by Eq. (19) (Sun et al., 2022):

\[ \text{Accuracy} = \frac{TP + TP}{TP + FN + FP + FN} \]  

(19)

Studies usually use \( TPR \) and precision to evaluate the performance of binary classification model. However, when the results of these two measurement approaches are close, it is difficult for researchers to judge the performance of the prediction model. F1-score takes the advantages of precision and \( TPR \):

\[ \text{f1-score} = 2 \cdot \frac{TPR \cdot \text{Precision}}{TPR + \text{Precision}} \]  

(20)

The ROC curve is another measure, especially for models with imbalanced data (Abedin et al., 2021). It is an \( FPR-TPR \) diagram in which the abscissa indicates \( FPR \) and the ordinate indicates the \( TPR \). Due to the unbalanced characteristics of credit data with many non-default samples and few default samples, the prediction model is easy to misjudge default samples as non-default samples. ROC can effectively avoid this misjudgment caused by data imbalance. For convenience, researchers usually utilize the area under the ROC curve (AUC) to evaluate the prediction performance of the model for unbalanced samples. In the following, this paper uses three evaluation measures including accuracy, f1-score, and AUC to test the effectiveness of the BOWOA-KS model.

4 Empirical analysis

4.1 Sample and data sources

This paper uses credit data for 2,044 SMEs of a Chinese commercial bank, and conducts empirical analysis. The 2,044 SMEs consist of 228 default customers and 1,816 non-default customers. Thus, the sample imbalance ratio is 8:1, satisfying the characteristics of imbalanced data. Based on Standard & Poor’s, Fitch, and Moody’s (Standard & Poor’s Services, 2011; Fitch Ratings, 2013), the indicators used by the China Postal Savings Bank (2009) and the Agricultural Bank of China (2011), a total of 44 indicators of SMEs were selected. These
include five layers: “C\textsubscript{1} basic information of lenders”, “C\textsubscript{2} repayment ability”, “C\textsubscript{3} repayment willingness”, “C\textsubscript{4} guarantee”, and “C\textsubscript{5} macro environment”. It includes 44 indicators, as shown in column c of Table 2. Descriptive statistics of the indicators are shown in Table 3. The standardization for the quantitative and qualitative indicators described in Sect. 3.1 is adopted here to standardize the original data from Table 2. The standardized data are shown in Table 4, column e.

4.2 Solution of feature selection problem

In this study, seven machine learning classification methods were used to test the results. These methods include the K-Nearest Neighbor (KNN), linear regression (LR), SVM, decision tree (DT), RF, the gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost), which are very commonly used in risk prediction (Abedin et al., 2021; Sun et al., 2022). These artificial intelligence classification methods were used to test whether the subset of indicators obtained using the BOWOA-KS model is reliable. Other models were also looked at for comparison. For the optimization algorithm, besides the BOWOA, the PSO algorithm was used. For the model’s fitness function, besides the KS statistic, the Gini coefficient and the AUC value were used. The accuracy or f1 score were not used because, for linear models, these two performance metrics require the default threshold to be determined. In order to avoid artificial subjectivity, models that do not require this were used as the fitness function. The RF method was also used to obtain the feature importance of each indicator in the raw data set. For a given subset of indicators, the grid search with cross validation was used to determine the hyper parameters of the machine learning classification methods (Abedin et al., 2020). Table 5 shows a list of the hyperparameters of each machine learning classification model for BOWOA-KS.

The accuracy, f1 score, AUC, and ROC curve for different subsets and machine learning methods were also obtained. The measurements of model performance are shown in Tables 6, 7 and 8, respectively. The accuracy distribution of the different machine learning methods with different numbers of indicators is shown in Fig. 2. The corresponding ROC curve of the different methods for the BOWOA-KS model is shown in Fig. 3.

In order to determine the 15 features that have the greatest impact on the 2,044 SMEs’ default risk, the RF model is adopted to screen the 44 indicators. Figure 4 shows the top 15 features.

4.3 Model robustness test

To verify the robustness of the proposed model, it was applied to two other sets of small enterprise loans in China. One of the samples contains data on 2,157 SMEs from a bank in China, with a total of 60 credit indicators, and a default ratio of about 9%. The other sample comprises 3,111 SMEs and comes from a Chinese bank, with a total of 80 credit indicators, and a default ratio of about 2%. Both samples are imbalanced. The proposed method was used to analyze these two samples and the results are shown in Tables 9, 10 and 11.

The 15 most important features from the original sets of indicators (60 and 80 indicators) are calculated in the same way as above. The results are shown in Figs. 5 and 6.
### Table 2: Original data on 2,044 SMEs

| No | (b) Criterion level | (c) Indicator | (d) Indicator type | (e) Original data for 2,044 loans $v_{ij}$ |
|----|---------------------|---------------|--------------------|------------------------------------------|
|    |                     |               |                    | Non-default SMEs | Default SMEs |
|    |                     |               |                    | (1) (1816)       | (1817) (2044) |
| 1  | C₁ Basic information of lenders | $X_1$ Age | Interval | 57...36 | 45...24 |
| ... | ... | ... | ... | ... | ... |
| 12 | C₂ Repayment ability | $X_{12}$ Skill of lenders and their families | Qualitative | 3...3 | 3...4 |
| ... | ... | ... | ... | ... | ... |
| 25 | C₃ Repayment willingness | $X_{25}$ Living conditions | Qualitative | 1...1 | 1...1 |
| ... | ... | ... | ... | ... | ... |
| 35 | C₄ Guarantee | $X_{35}$ Is there a guarantee | Qualitative | 2...1 | 0...1 |
| ... | ... | ... | ... | ... | ... |
| ... | C₅ Macro environment | ... | ... | ... | ... |
| 44 | | $X_{44}$ Engel’s coefficient | Negative | 0.373...0.364 | 0.373...0.379 |
| 45 | – | Default | – | 1...1 | 0...0 |
Table 3 Descriptive statistics of data for 2,044 SMEs

| Indicators                                      | Mean   | Var     | S.D   |
|-------------------------------------------------|--------|---------|-------|
| Age                                             | 40.932 | 70.207  | 8.379 |
| Education                                       | 3.976  | 0.976   | 0.988 |
| Marital status                                  | 1.086  | 0.240   | 0.490 |
| Gender                                          | 1.082  | 0.075   | 0.274 |
| Number of members                               | 1.634  | 0.409   | 0.640 |
| Size of labor force                             | 2.004  | 0.938   | 0.969 |
| Number of family members/Size of labor force    | 3.145  | 0.599   | 0.774 |
| Support population                              | 2.524  | 0.613   | 0.783 |
| Regional GDP growth rate                        | 18.253 | 7.195   | 2.682 |
| Consumer price index                            | 105.574| 0.599   | 0.774 |
| Resident savings deposit balance                | 5414.524| 5,157,935.048| 2271.109|
| Engel’s coefficient                             | 0.376  | 0.001   | 0.028 |

5 Discussion

To estimate the efficiency of the proposed BOWOA-KS model, the accuracy, AUC, and f1 scores were compared with those of other models, using two other credit datasets of Chinese small enterprises. By using the same initial parameters for the classifiers, a comparison of the feature selection techniques is possible. It is observed that, for different machine learning classifiers, if the number of indicators in the optimal subset of indicators and the prediction performance of the classifier are considered at the same time, the BOWOA-KS model is better (see Tables 6, 7, 8, 9, 10 and 11).

At the same time, it is evident that no artificial intelligence method is omnipotent. Specifically, the prediction performance levels of the different artificial intelligence classification methods are very close to each other. For accuracy, the prediction performances of SVM, RF, GBDT, and XGBOOST are better. If the f1 score is used for comparison, the performances of SVM and LR are worse than those of the other models. For the AUC value, the prediction performances of RF, GBDT, and XGBOOST are better. It is worth noting that, compared to the original small enterprise credit indicator set (44 indicators in total), the BOWOA-KS model established in this paper filters out fewer indicators. Although the number of indicators is reduced, we can observe that the model’s predictive power has not declined. This shows that the BOWOA-KS model established in this paper is effective. Using the PSO algorithm and the BOWOA for comparison, with the same parameter settings, we found that the indicator subset produced by the BOWOA contains fewer indicators and, except for the AUC values, has prediction performance metrics that are almost the same as for the PSO. Much to our surprise, the BOWOA-KS model finally selects only eight indicators from the original 44. That is to say, only 18% of the original set of indicators is used, but the prediction performance does not decrease. Compared with the optimal subset of indicators selected by the PSO algorithm, that selected by the BOWOA contains fewer indicators, meaning that the small enterprise default feature selection model proposed in this paper is effective and can reflect the information characteristics of the original set of indicators with fewer indicators.
| (a) No | (b) Criterion level | (c) Indicator | (d) Indicator type | (e) Standardized data for 2,044 loans $x_j$ |
|--------|-------------------|--------------|-------------------|----------------------------------|
|        |                   |              |                   | Non-default SMEs | Default SMEs |
|        |                   |              |                   | (1)    | ... | (1816) | (1817) | ... | (2044) |
| 1      | $C_1$ Basic       | $X_1$ Age    | Interval          | 0.368  | ... | 1.000  | 1.000  | ... | 0.526  |
|        | information of    |              |                   | ...    | ... | ...    | ...    | ... | ...    |
|        | lenders           |              |                   | ...    | ... | ...    | ...    | ... | ...    |
| 11     | $X_{11}$ House    |              | Positive          | 0.130  | ... | 0.285  | 0.052  | ... | 0.000  |
|        | value             |              |                   | ...    | ... | ...    | ...    | ... | ...    |
| 39     | $C_5$ Macro       | $X_{39}$ Per | Positive          | 0.089  | ... | 0.750  | 0.089  | ... | 0.205  |
|        | environment       | capita net   |                   | ...    | ... | ...    | ...    | ... | ...    |
|        |                   | income of    |                   | ...    | ... | ...    | ...    | ... | ...    |
|        |                   | rural        |                   | ...    | ... | ...    | ...    | ... | ...    |
|        |                   | households   |                   | ...    | ... | ...    | ...    | ... | ...    |
| 44     | $X_{44}$ Engel’s  |              | Negative          | 0.628  | ... | 0.701  | 0.628  | ... | 0.413  |
| 45     | Default           |              |                   | 1      | ... | 1      | 0      | ... | 0      |
### Table 5 Hyperparameters of artificial intelligence classification methods

| Methods | Parameters |
|---------|------------|
| K-NN    | $K = 5$, Euclidean distance |
| LR      | L2 penalty, 0.5 cut off |
| SVM     | Radial basis kernel function, complexity parameter $c = 1.0$, gamma = 0.1 |
| DT      | Support criterion is Gini |
| RF      | Number of trees = 200, max depth = 9, support criterion is Gini |
| GBDT    | Default booster, learning rate = 0.1, max depth = 3 |
| XGBoost | Learning rate = 0.1, max depth = 5, number of iterations = 200 |

**Fig. 2** Accuracy of different machine learning methods with different numbers of indicators

**Fig. 3** ROC curve of different methods used in the BOWOA-KS model
This provides new ideas for the selection of default characteristics of small enterprises and will reduce the cost of data collection for banks.

The interpretation of data in Tables 9, 10, and 11, shows that the model established in this paper is effective and robust. They allow fewer indicators to be used to reflect the default characteristics of the original indicator set. Moreover, the sample and subsets of indicators will largely affect the prediction performance of the machine learning classification model. The performances of the different artificial intelligence classification methods show no great differences. For the dataset of 3,111 small enterprises, the prediction performance of each classifier is better than that for the 2,157 small enterprises. The reason is that the default rate among the 3,111 small enterprises is only 2%, so the model overfits the non-default samples, resulting in high-performance metrics. In addition, it was noticed that, for these two samples, the number of indicators selected by the BOWOA is less than the number
Fig. 6 15 most important features for dataset of 3,111 SMEs using RF

Table 6 Accuracy of different models for different subsets

| Total indicators | Subset      | KNN | LR | SVM | DT | RF | GBDT | XGBoost |
|------------------|-------------|-----|----|-----|----|----|------|--------|
| 44               | Original    | 0.73| 0.63| 0.89| 0.78| 0.89| 0.89 | 0.88   |
| 13               | BOWOA-AUC   | 0.77| 0.55| 0.89| 0.8 | 0.87| 0.88 | 0.86   |
| 9                | BOWOA-GINI  | 0.79| 0.62| 0.89| 0.78| 0.87| 0.86 | 0.84   |
| 8                | BOWOA-KS    | 0.78| 0.63| 0.89| 0.78| 0.87| 0.88 | 0.87   |
| 15               | PSO-AUC     | 0.74| 0.56| 0.89| 0.77| 0.88| 0.88 | 0.86   |
| 21               | PSO-GINI    | 0.75| 0.56| 0.89| 0.79| 0.88| 0.87 | 0.87   |
| 21               | PSO-KS      | 0.71| 0.62| 0.89| 0.8 | 0.88| 0.88 | 0.87   |

Table 7 f1 scores of different models for different subsets of indicators

| Total indicators | Subset      | KNN | LR | SVM | DT | RF | GBDT | XGBoost |
|------------------|-------------|-----|----|-----|----|----|------|--------|
| 44               | Original    | 0.50| 0.50| 0.47| 0.53| 0.50| 0.54 | 0.54   |
| 13               | BOWOA-AUC   | 0.53| 0.45| 0.47| 0.56| 0.51| 0.52 | 0.51   |
| 9                | BOWOA-GINI  | 0.51| 0.49| 0.47| 0.53| 0.53| 0.53 | 0.53   |
| 8                | BOWOA-KS    | 0.5 | 0.49| 0.47| 0.52| 0.55| 0.55 | 0.57   |
| 15               | PSO-AUC     | 0.49| 0.46| 0.47| 0.52| 0.53| 0.55 | 0.54   |
| 21               | PSO-GINI    | 0.53| 0.44| 0.47| 0.57| 0.5 | 0.53 | 0.55   |
| 21               | PSO-KS      | 0.47| 0.48| 0.47| 0.54| 0.48| 0.53 | 0.54   |
Table 8 AUC of different models for different subsets of indicators

| Total indicators | Subset        | KNN | LR  | SVM | DT  | RF  | GBDT | XGBoost |
|------------------|---------------|-----|-----|-----|-----|-----|------|---------|
| 44               | Original      | 0.52| 0.59| 0.54| 0.55| 0.64| 0.62 | 0.63    |
| 13               | BOWOA-AUC     | 0.58| 0.54| 0.56| 0.57| 0.63| 0.56 | 0.6    |
| 9                | BOWOA-GINI    | 0.56| 0.53| 0.49| 0.54| 0.56| 0.57 | 0.54    |
| 8                | BOWOA-KS      | 0.57| 0.57| 0.63| 0.52| 0.59| 0.65 | 0.64    |
| 15               | PSO-AUC       | 0.52| 0.56| 0.54| 0.53| 0.65| 0.59 | 0.6    |
| 21               | PSO-GINI      | 0.56| 0.53| 0.49| 0.54| 0.56| 0.57 | 0.54    |
| 21               | PSO-KS        | 0.57| 0.57| 0.63| 0.52| 0.59| 0.65 | 0.64    |

Table 9 Accuracy of different models for different subsets of indicators

| Total indicators | Subset        | KNN | LR  | SVM | DT  | RF  | GBDT | XGBoost |
|------------------|---------------|-----|-----|-----|-----|-----|------|---------|
| Dataset of 2,157 SMEs |
| 60               | Original      | 0.69| 0.62| 0.88| 0.78| 0.89| 0.88 | 0.88    |
| 13               | BOWOA-AUC     | 0.78| 0.62| 0.89| 0.77| 0.86| 0.85 | 0.85    |
| 14               | BOWOA-GINI    | 0.78| 0.64| 0.89| 0.78| 0.86| 0.87 | 0.85    |
| 18               | BOWOA-KS      | 0.77| 0.63| 0.88| 0.77| 0.88| 0.88 | 0.88    |
| 25               | PSO-AUC       | 0.75| 0.63| 0.89| 0.78| 0.88| 0.88 | 0.87    |
| 32               | PSO-GINI      | 0.73| 0.62| 0.88| 0.77| 0.88| 0.87 | 0.87    |
| 19               | PSO-KS        | 0.74| 0.62| 0.89| 0.76| 0.89| 0.87 | 0.87    |
| Dataset of 3,111 SMEs |
| 80               | Original      | 0.97| 0.93| 0.97| 0.99| 0.99| 0.98 | 0.99    |
| 16               | BOWOA-AUC     | 0.95| 0.71| 0.97| 0.97| 0.98| 0.97 | 0.98    |
| 24               | BOWOA-GINI    | 0.97| 0.88| 0.96| 0.98| 0.98| 0.98 | 0.98    |
| 48               | BOWOA-KS      | 0.96| 0.92| 0.96| 0.98| 0.98| 0.98 | 0.98    |
| 23               | PSO-AUC       | 0.97| 0.86| 0.96| 0.98| 0.99| 0.98 | 0.98    |
| 27               | PSO-GINI      | 0.96| 0.9  | 0.96| 0.98| 0.98| 0.98 | 0.98    |
| 47               | PSO-KS        | 0.97| 0.92| 0.97| 0.99| 0.98| 0.98 | 0.99    |

selected by the PSO algorithm. This also confirms that the BOWOA established in this paper is effective. Although the BOWOA-KS model has more indicators than the BOWOA-AUC and BOWOA-GINI, careful observation reveals that each artificial intelligence classifier in the BOWOA-KS model selects a subset of indicators with higher predictive performance. This means that no model can select a subset with both a small number of indicators and high predictive performance, indicating that decision-makers need to make a trade-off between the number of indicators and the predictive performance of the model. In general, the subset of indicators selected by the BOWOA-KS model is better than the subset of indicators selected by other models.

From the above discussion, we can identify those key components that affect the performance of the BOWOA-KS model. The performance is influenced by the quality of the credit
### Table 10 f1 scores of different models for different subsets of indicators

| Total indicators | Subset       | KNN | LR  | SVM | DT  | RF  | GBDT | XGBoost |
|------------------|--------------|-----|-----|-----|-----|-----|------|---------|
| **Dataset of 2,157 SMEs** |              |     |     |     |     |     |      |         |
| 60 Original      |              | 0.50| 0.49| 0.47| 0.47| 0.47| 0.49 | 0.53    |
| 13 BOWOA-AUC     |              | 0.52| 0.51| 0.48| 0.52| 0.48| 0.49 | 0.5     |
| 14 BOWOA-GINI    |              | 0.51| 0.52| 0.47| 0.51| 0.51| 0.48 | 0.49    |
| 18 BOWOA-KS      |              | 0.53| 0.51| 0.47| 0.54| 0.53| 0.52 | 0.55    |
| 25 PSO-AUC       |              | 0.53| 0.5  | 0.52| 0.54| 0.51| 0.49 | 0.52    |
| 32 PSO-GINI      |              | 0.5 | 0.49| 0.54| 0.5  | 0.47| 0.49 | 0.5     |
| 19 PSO-KS        |              | 0.47| 0.5  | 0.47| 0.53| 0.51| 0.51 | 0.53    |
| **Dataset of 3,111 SMEs** |              |     |     |     |     |     |      |         |
| 80 Original      |              | 0.78| 0.64| 0.71| 0.85| 0.80| 0.75 | 0.83    |
| 16 BOWOA-AUC     |              | 0.63| 0.44| 0.58| 0.72| 0.73| 0.67 | 0.75    |
| 24 BOWOA-GINI    |              | 0.72| 0.58| 0.68| 0.71| 0.77| 0.77 | 0.75    |
| 48 BOWOA-KS      |              | 0.74| 0.62| 0.68| 0.85| 0.79| 0.79 | 0.83    |
| 23 PSO-AUC       |              | 0.76| 0.56| 0.7  | 0.8 | 0.81| 0.77 | 0.81    |
| 27 PSO-GINI      |              | 0.7 | 0.6  | 0.69| 0.75| 0.78| 0.79 |         |
| 47 PSO-KS        |              | 0.78| 0.6  | 0.73| 0.88| 0.77| 0.77 | 0.83    |

### Table 11 AUC of different models for different subsets of indicators

| Total indicators | Subset       | KNN | LR  | SVM | DT  | RF  | GBDT | XGBoost |
|------------------|--------------|-----|-----|-----|-----|-----|------|---------|
| **Dataset of 2,157 SMEs** |              |     |     |     |     |     |      |         |
| 60 Original      |              | 0.56| 0.62| 0.59| 0.53| 0.69| 0.63 | 0.62    |
| 13 BOWOA-AUC     |              | 0.55| 0.65| 0.54| 0.53| 0.58| 0.53 | 0.52    |
| 14 BOWOA-GINI    |              | 0.56| 0.66| 0.61| 0.51| 0.63| 0.64 | 0.57    |
| 18 BOWOA-KS      |              | 0.61| 0.66| 0.63| 0.55| 0.65| 0.63 | 0.63    |
| 25 PSO-AUC       |              | 0.6 | 0.61| 0.58| 0.55| 0.63| 0.61 | 0.6     |
| 32 PSO-GINI      |              | 0.56| 0.61| 0.57| 0.5 | 0.67| 0.63 | 0.6     |
| 19 PSO-KS        |              | 0.52| 0.64| 0.58| 0.55| 0.61| 0.59 | 0.59    |
| **Dataset of 3,111 SMEs** |              |     |     |     |     |     |      |         |
| 80 Original      |              | 0.90| 0.89| 0.9  | 0.86| 0.96| 0.94 | 0.95    |
| 16 BOWOA-AUC     |              | 0.73| 0.57| 0.55| 0.76| 0.89| 0.78 | 0.82    |
| 24 BOWOA-GINI    |              | 0.85| 0.85| 0.83| 0.74| 0.96| 0.95 | 0.95    |
| 48 BOWOA-KS      |              | 0.9 | 0.88| 0.9  | 0.88| 0.97| 0.94 | 0.96    |
| 23 PSO-AUC       |              | 0.85| 0.85| 0.86| 0.83| 0.96| 0.94 | 0.95    |
| 27 PSO-GINI      |              | 0.84| 0.89| 0.89| 0.77| 0.96| 0.92 | 0.9     |
| 47 PSO-KS        |              | 0.89| 0.88| 0.9  | 0.91| 0.96| 0.92 | 0.95    |
evaluation data, machine learning classification algorithms, and performance evaluation criteria. First, the most important factor affecting model performance is the quality of the credit evaluation data. The accuracy of the credit risk prediction using 3,111 SMEs is better than that of the predictions using the other two datasets, i.e. with 2,044 SMEs and 2,157 SMEs. The quality of the data generally includes three aspects: information capacity, missing values, and imbalance ratio. It can be observed that the more imbalanced the sample, the more accurate the model will be. A potential reason may be that the model overidentifies non-defaulting SMEs in an unbalanced sample. In addition, the more information the sample contains, the higher the accuracy of the model will be, which is consistent with our expectations. Second, the machine learning classification algorithms also affect the prediction performance, and no single machine learning classification algorithm is suitable for all data. Different datasets will be best served by different machine learning classification algorithms. Third, if multiple model performance evaluation criteria are considered simultaneously, for example, for the same dataset, the highest AUC and the highest accuracy rate will not necessarily be found for the same model. This indicates that the evaluation criteria used will also affect the performance of the model. The above analysis shows that the quality of credit data plays a dominant role in the prediction performance of any model. This has inspired us to work on establishing a multi-source SME credit database in the era of Industry 4.0. In addition, when analysts use a machine learning classification model to predict default risk, they should choose a classification model that suits the data.

We also calculated the 15 most important features of the other two samples using the RF method (as shown in Figs. 5 and 6). This shows that different sample sets have different features associated with defaults. For the sample of 2,044 SMEs, external environmental factors such as the Engel’s coefficient are significantly related to whether a small company defaults while, for the sample of 2,157 SMEs, the age of the company’s manager and the number of years for which the company has been operating are factors that significantly affect whether a company defaults. However, for the sample of 3,111 SMEs, indicators such as a company’s business scope and whether it has had many legal proceedings in the past three years are significantly linked to whether a company defaults. We recommend that banks consider these indicators comprehensively when making lending decisions.

6 Conclusions and future work

SMEs are the driving force behind a country’s economic development. In the Industry 4.0 era, the development of small enterprises is particularly important. The most crucial features of this era are intelligence, automation, and big data. Undeniably, the information age has brought many benefits to SMEs. At the same time, the development of SMEs still faces some challenges, especially in financing. During the COVID-19 pandemic, the cash flow of many SMEs has been limited, and some have been forced to close due to capital problems. An SME’s most common method of solving a cash flow shortage is to apply for a loan from a bank. However, the credit scoring system for SMEs is incomplete, and the method used for large enterprises cannot be applied directly to them. In this era of big data, knowing how to quickly and accurately identify indicators that are strongly related to SMEs’ credit risk from high-dimensional datasets is the key to easing SMEs’ financing constraints. Feature selection provides a possible way to solve this problem. By using an appropriate feature selection method, analysts can improve the accuracy of credit risk prediction models, reduce the computational complexity, and better understand the key default features of small enterprises.
This paper proposes a BOWOA-KS model for a wrapper algorithm that can quickly screen the default characteristics of SMEs. The BOWOA-KS model is used to conduct an empirical study on three different datasets in China. First, the methodology is verified using actual bank data for 2,044 SMEs. Then robustness tests are carried out using data for 2,157 SMEs and 3,111 SMEs. The results reveal that the proposed method can greatly reduce the computational complexity of traditional artificial intelligence wrapper techniques. We find that the optimal number of indicators and the optimal classification model differ among the datasets. If the total number of selected features and prediction performance of the model are considered comprehensively, the BOWOA-KS model is shown to be reliable and it can quickly and accurately identify the indicators that are significantly related to the default characteristics of SMEs.

In the age of Industry 4.0, the amount of information available is huge, and credit data has a very high dimensionality and contains a lot of noise. This makes it necessary to simplify the data structure in credit evaluation. The model for SMEs’ default feature selection provided in this paper can reduce the dimensionality of credit data effectively. When applying the model, operators should pay attention to the characteristics of their datasets, so as to select the most appropriate artificial intelligence optimization algorithm and model parameters. Meanwhile, financing constraints for SMEs in the post-COVID-19 era are severe. Constructing a novel credit evaluation model for SMEs could effectively reduce the information uncertainty of the market and aid the recovery of small enterprises.

Policy implications have been generated by this empirical study and methodology. In the context of Industry 4.0, artificial intelligence tools and big data acquisition technologies bring new opportunities for fast and accurate calculation of the credit risks of SMEs. Learning about the top features that are most likely to be exhibited by successful loans can help bankers not only identify creditworthy SMEs, but also guide them towards figuring out which data they should collect for their databases. Additionally, financial institutions can refer to the proposed method to reduce the dimensionality of their credit data, so as to reduce both the costs of data collection and those of modeling. This will also help SMEs to focus on those factors that will let them improve their creditworthiness and access to loans from financial institutions. At the same time, this work can benefit governmental agencies and policymakers by pinpointing the characteristics that can lead to the success of SMEs in the economy, helping them to develop policies to help SMEs guard against financial risks.

Despite the improvements it makes in terms of sample expansion, feature reduction, and the selection of performance evaluation indicators for classifier prediction, this research can be further expanded in the future. An interesting extension of this research would be to apply the BOWOA-KS models to other datasets and other types of credit applicants, which can help to evaluate the applicability of the methods. This study found the optimal subset of indicators based on the BOWOA artificial intelligence optimization algorithm. It would be interesting for scholars to use and compare other optimization algorithms in future research. Finally, for imbalanced datasets, the SMOTE algorithm was used to expand the data. Future research could compare the performance of different data expansion algorithms and find the best expansion imbalance ratio.

Acknowledgements This paper was supported by the National Natural Science Foundation of China (Nos. 72173096, 71873103, and 71731003), the Soft Science Research Project of Ministry of Agriculture and Rural Affairs of the China (No. 202122), the Humanities and Social Science Research Project of Ministry of Education of China (No. 21YJCZH107), the Social Science Foundation of Shaanxi Province (No. 2018D51), and the Credit Rating and Loan Pricing Project for Small Enterprises of the Bank of Dalian (No. 2012-01). Shi acknowledges financial support from the Tang Scholar Program of Northwest A&F University (No. 2021-04).
References

Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., & Hajek, P. (2020). Tax default prediction using feature transformation-based machine learning. *IEEE Access, 9*, 19864–19881.

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Petr, H. (2021). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04420-6

Agricultural Bank of China. (2011). Measures for risk classification management of credit assets of Agricultural Bank of China. Agricultural Bank of China.

Ali, M., Siarry, P., & Pant, M. (2012). An efficient differential evolution based algorithm for solving multi-objective optimization problems. *European Journal of Operational Research, 217*(2), 404–416.

Aljarah, I., Faris, H., & Mirjalili, S. (2016). Optimizing connection weights in neural networks using the whale optimization algorithm. *Soft Computing, 22*(1), 1–15.

Altman, E. I., Esentato, M., & Sabato, G. (2020). Assessing the credit worthiness of Italian SMEs and mini-bond Issuers. *Global Finance Journal, 43*, 100450.

Bagale, G. S., Vandadi, V. R., Singh, D., et al. (2021). Small and medium-sized enterprise’s contribution in digital technology. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04425-1

Bai, C. G., Shi, B. F., Liu, F., & Sarkis, J. (2019). Banking credit worthiness: Evaluating the complex relationships. *Omega, 83*, 26–38.

China Postal Savings Bank. (2009). Standard for personal credit interface of China Postal Savings Bank.

Beck, T., Demirguc-Kunt, A., & Levine, R. (2005). SMEs, growth, and poverty: cross-country evidence. *Journal of Economic Growth, 10*(3), 199–229.

Belotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), 3302–3308.

Bhattacharyya, A., Goswami, R. T., & Mukherjee, K. (2018). A feature selection technique based on rough set and improvised PSO algorithm (PSORS-FS) for permission based detection of Android malwares. *International Journal of Machine Learning and Cybernetics, 10*(7), 1893–1907.

Blanc Alquier, A. M., & Lagasse Tignol, M. H. (2006). Risk management in small and medium-sized enterprises. *Production Planning & Control, 17*(3), 273–282.

Chai, N. N., Wu, B., Yang, W. W., & Shi, B. F. (2019). A multi criteria approach for modeling small enterprise credit rating: evidence from China. *Emerging Markets Finance and Trade, 55*(11), 2523–2543.

Chen, F. L., & Li, F. C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications, 37*(7), 4902–4909.

Chen, H., & Xiang, Y. (2017). The study of credit scoring model based on group lasso. *Procedia Computer Science, 122*, 677–684.

Chen, W., Li, Z., & Guo, J. (2020). A VNS-EDA algorithm-based feature selection for credit risk classification. *Mathematical Problems in Engineering, 5*, 1–14.

Chi, G., & Zhang, Z. (2017). Multi criteria credit rating model for small enterprise using a nonparametric method. *Sustainability, 9*(10), 1834.

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises: an empirical analysis of Italian firms. *Journal of Business Research, 68*(5), 1012–1025.

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking SME default prediction: a systematic literature review and future perspectives. *Sienticmetrics, 126*(3), 2141–2188.

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: an empirical study of sample size and balancing. *International Journal of Forecasting, 28*(1), 224–238.

Dia, M., Takouda, P. M., & Golmohammadi, A. (2022). Assessing the performance of Canadian credit unions using a three-stage network bootstrap DEA. *Annals of Operations Research, 311*, 641–673.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis, 7*(2), 1477–1493.

Fernandez, A., Garcia, S., Chawla, N. V., & Herrera, F. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905.

Freeman, C., Kulić, D., & Basir, O. (2015). An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognition, 48*(5), 1812–1826.

Granja, J., Leuz, C., & Rajan, R. G. (2022). Going the extra mile: distant lending and credit cycles. *The Journal of Finance*. https://doi.org/10.1111/jofi.13114
Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Hussien, A. G., Hassanien, A. E., Houssein, E. H., & Azar, A. H. (2020). New binary whale optimization algorithm for discrete optimization problems. *Engineering Optimization, 52*(6), 945–959.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In: *The Eleventh International Conference on Machine Learning*. San Fransisco: Morgan Kaufmann, pp. 121–129.

Kamble, S. S., Angappa, G., & Gawankar, S. A. (2018). Sustainable Industry 4.0 framework: a systematic literature review identifying the current trends and future perspectives. *Process Safety and Environmental Protection, 117*, 408–425.

Kaur, G., & Arora, S. (2018). Chaotic whale optimization algorithm. *Journal of Computational Design and Engineering, 5*, 275–284.

Kim, Y., & Vonortas, N. S. (2014). Cooperation in the formative years: evidence from small enterprises in Europe. *European Journal of Operational Research, 32*(2), 795–805.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324.

Laitinen, E. K. (1993). The use of information contained in annual reports and prediction of small business failures. *International Review of Financial Analysis, 2*(3), 155–176.

Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing, 107*(29), 107391.

Mahbobi, M., Khamene, S., & Vasudevan, M. (2021). Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04114-z

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences, 179*(13), 2208–2217.

Medina-Olivares, V., Calabrese, R., Dong, Y. Z., & Shi, B. F. (2022). Spatial dependence in microfinance credit default. *International Journal of Forecasting, 38*(3), 1071–1085.

Mehmanchi, E., Gómez, A., & Prokopyev, O. A. (2021). Solving a class of feature selection problems via fractional 0–1 programming. *Annals of Operations Research, 303*, 265–295.

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software, 95*, 51–67.

Nakariyakul, S. (2018). High-dimensional hybrid feature selection using interaction information-guided search. *Knowledge-Based Systems, 145*, 59–66.

Oliveira, M. D. N. T., Ferreira, F. A. F., Pérez-Bustamante Ilander, G. O., & Jalali, M. S. (2017). Integrating cognitive mapping and MCDA for bankruptcy prediction in small-and medium-sized enterprises. *Journal of the Operational Research Society, 68*(9), 985–997.

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications, 41*(4), 2052–2064.

Orth, W. (2012). The predictive accuracy of credit ratings: Measurement and statistical inference. *International Journal of Forecasting, 28*(1), 288–296.

Park, S. Y., & Lee, J. J. (2015). Stochastic opposition-based learning using a beta distribution in differential evolution. *IEEE Transactions on Cybernetics, 46*(10), 2184–2194.

Pereira, L., Tenera, A., Bispo, J., & Wemans, J. (2013). A Risk Diagnosing Methodology Web-Based Platform for Micro, Small and Medium Businesses: Remarks and Enhancements. In: *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Springer Berlin Heidelberg.

Fitch Ratings. (2013). Global corporate finance 2012 transition and default study.

Rezac, M., & Rezac, F. (2011). How to measure the quality of credit scoring models. *Finance a Uver-Czech Journal of Economics and Finance, 61*(5), 486–507.

Sayed, G. I., Ashraf, D., & Hassanien, A. E. (2018). A new chaotic whale optimization algorithm for features selection. *Journal of Classification, 35*, 300–344.

Sefidian, A. M., & Daneshpour, N. (2018). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications, 115*, 68–94.

Shi, B. F., Meng, B., Yang, H. F., Wang, J., & Shi, W. L. (2018). A novel approach for reducing attributes and its application to small enterprise financing ability evaluation. *Complexity, 1–17*, 1032643.

Standard & Poor’s Ratings Services. (2011). S&P’s study of China’s top corporates highlights their significant financial risks. 2011. New York: Standard & Poor’s.

Sun, Y., Chai, N. N., Dong, Y. Z., & Shi, B. F. (2022). Assessing and predicting small industrial enterprise’s credit ratings: a fuzzy decision making approach. *International Journal of Forecasting, 38*(3), 1158–1172.
Tharwat, A., Moemen, Y. S., & Hassanien, A. E. (2017). Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines. Journal of Biomedical Informatics, 68, 132–149.

Tizhoosh, H. R. (2005). Opposition-based learning: a new scheme for machine intelligence. Proceedings of International Conference on Computational Intelligence for Modeling Control and Automation, 1, 695–701.

Wadhwa, R. S. (2012). Flexibility in manufacturing automation: a living lab case study of Norwegian metal-casting SMEs. Journal of Manufacturing Systems, 31(4), 444–454.

Wang, H., Wu, Z. J., Rahnamayan, S., Liu, Y., & Ventresca, M. (2011). Enhancing particles warm optimization using generalized opposition-based learning. Information Sciences, 181(20), 4699–4714.

Yu, S., Chi, G., & Jiang, X. (2018). Credit rating system for small businesses using the k-s test to select an indicator system. Management Decision, 57(1), 229–247.

Zhang, X. M., & Wen, S. C. (2021). Hybrid whale optimization algorithm with gathering strategies for high-dimensional problems. Expert Systems with Applications, 179, 115032.

Zhang, Z. P., Chi, G. T., Colombage, S., & Zhou, Y. (2022). Credit scoring model based on a novel group feature selection method: the case of Chinese small-sized manufacturing enterprises. Journal of the Operational Research Society, 73(1), 122–138.

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