Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(11): 19401–19415.
DOI: 10.3934/mbe.2023858
Received: 11 July 2023
Revised: 27 September 2023
Accepted: 08 October 2023
Published: 19 October 2023

Research article

Prediction of personal default risks based on a sparrow search algorithm with support vector machine model

Xu Shen[1,2,*] **and Xinyu Wang**[1]

[1] School of Economics and Management, China University of Mining and Technology, Xuzhou 221116, China
[2] School of Economics, Shandong Women's University, Jinan 250300, China

* **Correspondence:** Email: 30074@sdwu.edu.cn; Tel: 13953165173.

Abstract: Aiming at the personal credit evaluation of commercial banks, this paper constructs a classified prediction model based on machine learning methods to predict the default risk. At the same time, this paper proposes to combine the sparrow search algorithm (SSA) with the support vector machine (SVM) to explore the application of the SSA-SVM model in personal default risk prediction. Therefore, this paper takes the personal credit data as the original data, carries out statistical analysis, normalization and principal factor analysis, and substitutes the obtained variables as independent variables into the SSA-SVM model. Under the premise of the same model, the experimental results show that the evaluation indexes of the experimental data are better than the original data, which shows that it is effective for the data processing operation of the original data in this paper. On the premise of the same data, each evaluation index of the SSA-SVM model is better than the SVM model, which shows that the hybridized model established in this paper is better than the latter one in predicting personal default risk, and has certain practical value.

Keywords: default risks; SSA-SVM model; credit assessment; prediction accuracy; commercial banks

1. Introduction

To determine whether to loan to borrowers, the traditional credit business mainly relies on the subjective factors of both parties as the basis. Thomas [1] pointed out that the credit scores of UK and US residents are updated at least weekly. Zhang [2] also showed the popularity of credit scoring in the United States: Bank of America uses an internal scoring model with 97% of credit cards and 70% of its small loan business. Before the emergence of quantitative credit risk models, financial institutions mainly relied on due diligence of loan officers, such as the classic 5C evaluation method, which evaluates five aspects of credit quality of loan applicants: character, ability, capital, collateral, and condition. Due diligence largely relies on subjective judgment and the rich experience of loan officers, and there are high operational and time costs involved.

However, the personal loan business of commercial banks has begun to increase, and the traditional lending model has been unable to meet the current demand of credit business. Therefore, commercial banks hope to pursue better judgments to reduce credit defaults. At present, fintech is constantly developing, and machine learning technology is deeply involved in financial risk prevention and control, but there is still room for progress in the existing machine learning technology for this issue.

Credit scoring techniques use statistical methods and artificial intelligence models to objectively evaluate borrowing. The credit status of the applicants can save time and cost, and improve operational efficiency. Finlay believes that any slight improvement in the performance of credit scoring models can help financial institutions avoid a loss of millions of dollars [3]. The continuous improvement of the performance of credit scoring models is the goal pursued by researchers. The continuous exploration and improvement of classification techniques in credit scoring models is a hot topic in credit scoring research. The earliest credit scoring model was proposed by Durand [4], who used Fisher discriminant analysis to distinguish default loans. With the continuous innovation of technology and methods in the fields of statistics and machine learning, scholars attempt to introduce a large number of new classification methods into the field of credit scoring in order to improve the predictive performance of the model.

In real financial environments, there are still some problems, although there are plenty of studies on personal default prediction. Since the number of credit defaulters is always a small sample event and personal credit data is always an unbalanced sample, this situation needs to be taken into account. Meanwhile, when performing default prediction, appropriate data mining models need to be considered.

The rest of the paper is organized as follows: Section 2 reviews related work and elaborates the task of the paper. In Section 3, the theoretical basis is introduced. In Section 4, we construct the SSA-SVM to evaluate its index. In Section 5, we carry out empirical analysis and Section 6 concludes the paper.

2. Literature review

Researches on credit risk optimization started early. Andersson et al. [5] solved the credit risk optimization problem by building a Conditional Value at Risk model that can simultaneously adjust all positions in a portfolio of financial instruments. At the same time, Shi et al. [6] also summarized foreign research on personal loan default prediction, and reviewed the scoring models and methods used by foreign commercial banks for personal consumer credit, introducing regression analysis, neural networks and mathematical programming, and analyzed and compared the performance of various methods. Du [7] proposed to combine the problem of personal credit assessment with data mining theory, which diminishes the problem of domestic credit collection mainly depending on the subjective will of people, and provides a new research topic for personal credit assessment.

At the same time, default rate prediction models based on support vector machine models were also proposed. Vapink [8] proposed statistical learning theory, which laid the theoretical foundation of

SVM, and more and more scholars began to use SVM models in personal loan default prediction. Chong et al. [9] chose to compare the SVM model with the K-nearest neighbor model and showed that the SVM model significantly outperformed the K-nearest neighbor model in personal credit evaluation. Guo [10] found that the support vector machine based personal credit assessment model can effectively improve the prediction accuracy and has a good prospect for development. Tian [11] and Tang et al. [12] found that SVM models are superior to logistic models to a certain extent and have higher application value. Although all the above scholars believe that SVM models are superior to other models to a certain extent, there is still some room for progress in SVM models.

Therefore, many scholars began to shift their research focus to the optimization of SVM models, hoping to improve the prediction accuracy of the models by combining relevant algorithms. Shen et al. [13] obtained better results using the NN-SVM-KNN model. Zhong et al. [14] concluded that the traditional support vector machine model has certain problems leading to excessive computation and low accuracy, which makes it difficult to apply to large data problems, so she chose to use the LS-SVM model for experimental analysis of German credit data, and the results showed that the LS-SVM model outperformed the KNN method and the discriminant analysis method, indicating that the method is more stable and practical. Xiao et al. [15] found that the model combining principal component analysis and support vector machine significantly outperformed various credit assessment models such as neural network and K-nearest neighbor discriminant analysis in predicting personal credit assessment. Dai et al. [16] combined K-means clustering method with the SVM model to further classify the credit rating with high value of use. Liu et al. [17] chose to construct a credit evaluation model with C4.5 decision tree optimization support vector machine, which effectively solved the problem of decreasing accuracy of the traditional SVM model in predicting high-dimensional data. Wang [18] found that the predictors could not be effectively screened in the traditional SVM model, so she introduced the random forest fusion support vector machine model into the personal credit assessment problem, and proved that the method could obtain better prediction results through empirical research on experimental samples. Li [19] believed that there were many factors affecting personal credit, thus she introduced the lasso technique into personal credit assessment, and experimentally concluded that the Lasso-SVM model has high accuracy. Wang et al. [20] introduced the frog-jumping algorithm to optimize the hyper parameters based on the traditional SVM model, which solved the problem that the model parameters were difficult to determine and also had a better evaluation performance. Chen et al. [21] chose to optimize the SVM model based on the improved aspen swarm algorithm in the model parameter problem, and achieved good results.

From the above literature research results, it can be concluded that there are some differences in the impact effects brought about by the different models and parameter optimization algorithms selected in the study. How to choose a more suitable algorithm to improve the prediction accuracy has become a difficult problem in credit default prediction. Xue et al. [22] proposed the sparrow search algorithm, and it has been applied by several scholars in major fields. Wang et al. [23] applied the sparrow search algorithm to the rural road cost problem to optimize the Backpropagation neural network (BP) model to predict the rural road cost, and the experimental results showed that the SSA-BP algorithm model was better than the Radial basis function network in terms of stability and accuracy, and had better practicality. Liang et al. [24] applied the SSA-BP model to the impact ground pressure data prediction, and the experimental results showed that the optimization of BP neural network using SSA algorithm avoids the disadvantage of BP weight balance. While the sparrow search algorithm was used to optimize BP neural networks, some scholars also chose it as an optimization algorithm for support vector machines. Hu et al. [25] chose to use the sparrow search algorithm to optimize the parameters of the support vector machine as a way to achieve tool wear status recognition based on acoustic signal features and achieved a high accuracy rate. In order to improve the accuracy of short-term wind turbine power prediction, Wang et al. [26] proposed to optimize the support vector machine model with the sparrow search algorithm and compared the obtained results with GA-SVM and SVM models. The experimental results show that the SSA-SVM model can effectively improve the learning parameter selection efficiency and prediction accuracy. However, in the financial field, the sparrow search algorithm is not widely used at present.

Credit scoring models enable and support credit risk management in financial institutions. For more than half a century, Thomas et al. [27] have been part of decisions throughout the credit risk management cycle. Today, in Anderson and Ntwiga [28,29], no decisions about whom to grant a loan to, portfolio management, preventive collection actions, or even pricing are made without the support of credit scoring models. Academics and practitioners have developed different credit scoring tools to support the different decisions at each stage of the credit risk management cycle. Application scoring is used to decide whether to grant a loan to a new applicant entering the financial system. In contrast, behavioral scoring allows lenders to characterize those borrowers who have already been granted a loan, and it is used mainly for portfolio management. Finally, Paleologo et al. [30] found that collection scoring allows optimizing policies and strategies for collection and recovery.

In this paper, in order to improve the prediction accuracy, we hope to optimize the SVM model with SSA and compare it with the former in order to obtain the optimal prediction model for customer information and customer default.

3. Theoretical basis

3.1. SVM

SVM was first proposed by Vapnik in 1964. It is currently used in many fields such as pattern recognition, nonlinear regression and default prediction. SVM is based on structural risk minimization and VC theory, and the structural risk minimization is introduced to prevent overfitting of the training model. When the VC dimension is too large, the generalization ability of the model is poor, and the smaller the VC dimension is, the stronger the generalization ability of the model is. Therefore, because of the support of VC dimensional theory, the support vector machine has unique generalization performance. Here, a nonlinear support vector classifier is introduced.

Regarding the nonlinear classification problem, the kernel function is mainly used to convert it into a linear classification problem for solving. The commonly used kernel functions include linear kernel function, Gaussian kernel function, polynomial kernel function and Sigmoid kernel function. This paper mainly introduces Gaussian kernel function.

For the data sets $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$,$x_i \in R^n$, $y_i \in \{+1, -1\}$, firstly it is necessary to select the appropriate kernel function $K(x, z)$ and penalty function $C > 0$.Then, to construct Eq (1) and solve for the optimal solution $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$.

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha i \alpha j \, yiyjK(xi \cdot xj) - \sum_{i=1}^{N} \alpha i,$$

$$s.t. \sum_{i=1}^{N} \alpha i \, yi = 0, \quad (1)$$

$$0 \leq \alpha i \leq C, i = 1,2,\ldots,N.$$

After that, a component α_j^* is selected from the optimal solution and this component satisfies the condition $0 < \alpha_j^* < C$, using this component to calculate Eq (2).

$$b^* = y_j - \sum_{i=1}^{N} \alpha_j^* y_i K(x_i \cdot x_j) \quad (2)$$

Then the Gaussian kernel function is introduced as Eq (3), and the classification decision function in this condition is obtained as Eq (4).

$$K(x,z) = exp(-\frac{\|x-z\|^2}{2\sigma^2}), \quad (3)$$

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i^* y_i \, exp(-\frac{\|x-z\|^2}{2\sigma^2}) + b^*). \quad (4)$$

3.2. SSA

SSA was proposed by Xue et al. in 2020 and was mainly inspired by the foraging and anti-predatory behaviors of sparrows. The main rules of the SSA are described as follows:

In the model-building process, individual sparrows are divided into discoverers and joiners in a certain proportion, and the identities are not fixed. They will switch according to the superiority of the food sources found by the sparrows in their search, while the proportion remains unchanged. The individual sparrow's adaptation value represents its energy reserve, and the discoverer usually has a relatively high energy reserve to search for areas with abundant food, while the joiner also chooses to search for the best discoverer to increase its predation rate, and thus sets an alarm value to prevent the appearance of predators. When an individual sparrow detects danger and chirps, the more marginal sparrows in the group will move towards the safe area to update the optimal position, while the sparrows in the middle area will then move to approach other sparrows.

To establish the process of the SSA, first, the relevant parameters of the population and the proportion of joiners need to be initialized. Second, the fitness values are calculated and at the same time, the ranking is done. Third, the positions of predators, joiners and vigilantes are updated. Finally, the fitness value is continuously calculated to update the positions so that they meet the end conditions.

The mathematical model for the position update of the discoverer sparrow is Eq (5).

$$Xi\ (t+1) = \begin{cases} Xi\,(t) \cdot \exp(-\dfrac{i}{\alpha \cdot MaxIter}) & if \quad R_2 < ST, \\ Xi\,(t) + Q \cdot L & if \quad R_2 \geq ST. \end{cases} \quad (5)$$

where, X_i is the position vector of the first sparrow; T is the current iteration number; $MaxIter$ is the maximum iteration number; A is a random number between (0,1), R_2 ($R_2 \in [0,1]$) and ST ($ST \in [0.5, 1.0]$) for the early warning threshold and the safety threshold, respectively; Q is a random number and satisfies a positive, state distribution; L is a $1 \times d$ matrix of all 1 element (d is the dimension).

The mathematical model for updating the position of the participant Sparrow is as Eq (6).

$$X_i(t+1) = \begin{cases} Q \cdot \exp\left(\dfrac{X_{worst}(t) - X_i(t)}{i^2}\right) & if \quad i < n/2, \\ X_P(t+1) + \left|X_i(t) - X_i(t+1)\right| \bullet A^{+} \bullet L & otherwise. \end{cases} \tag{6}$$

where X_P and X_{worst} are the optimal positions of the discoverer and the worst positions globally, respectively; A represents a 1 × D matrix randomly assigned to each element 1 or –1, and A + = AT (AAT)–1.

The reconnaissance and early warning sparrow accounts for 10% to 20% of the population.

$$X_i(t+1) = \begin{cases} X_{best}(t) + \beta\left|X_i(t) - X_{best}(t)\right| & if \quad f_i < f_g, \\ X_i(t) + K\left(\dfrac{X_i(t) - X_{worst}(t)}{(f_i - f_w) + \varepsilon}\right) & if \quad f_i = f_g. \end{cases} \tag{7}$$

In the equation, f_i, f_g and f_w are the fitness values of the current sparrow individuals, and the global best and worst fitness values, respectively; X_{best} is the current global best position; K is a random number between [–1,1]; β is called the step-size control parameter; ε is the smallest constant to avoid zero in the denominator.

4. Construction of the SSA-SVM model

This section mainly introduces the construction process of the SSA-SVM model, and also discusses the process and evaluation index of this model.

4.1. Experimental procedure of the SSA-SVM model

The traditional SVM model with fixed parameters has certain universality, but may have bias for personal credit data in real financial environments. Therefore, this paper adds the SSA to the traditional SVM model to change the problem of fixed parameters of the latter.

The SSA-SVM model first needs to establish a suitable parameter range for the dataset and find the optimal parameters within this parameter range in terms of the fitness value, which is mainly applied to the dataset used in the model, i.e., each different dataset will have different optimal parameters. As a result, the parameters used in the SVM model for model training and prediction will change from the traditional SVM model parameters to the optimal parameters corresponding to the experimental data in order to obtain the optimal prediction results. In this paper, we use Pycharm for model construction, and the methodology of the SSA-SVM model is shown in Figure 1.

4.3. Evaluation matrix of the SSA-SVM model

In the binary classification problem, the confusion matrix is a standard format to represent the accuracy evaluation, which is mainly used to compare the classification results with the actual measured values, and the accuracy of the classification results can be displayed inside a confusion matrix. As shown in Table 1, TP (true non-default) denotes non-defaulted samples that are predicted by the model as non-defaulted, FP (false non-default) denotes defaulted samples that are predicted by the model as non-defaulted, FN (false default) denotes non-defaulted samples that are predicted by the model as defaulted, and TN (true default) denotes defaulted samples that are predicted by the model as defaulted.

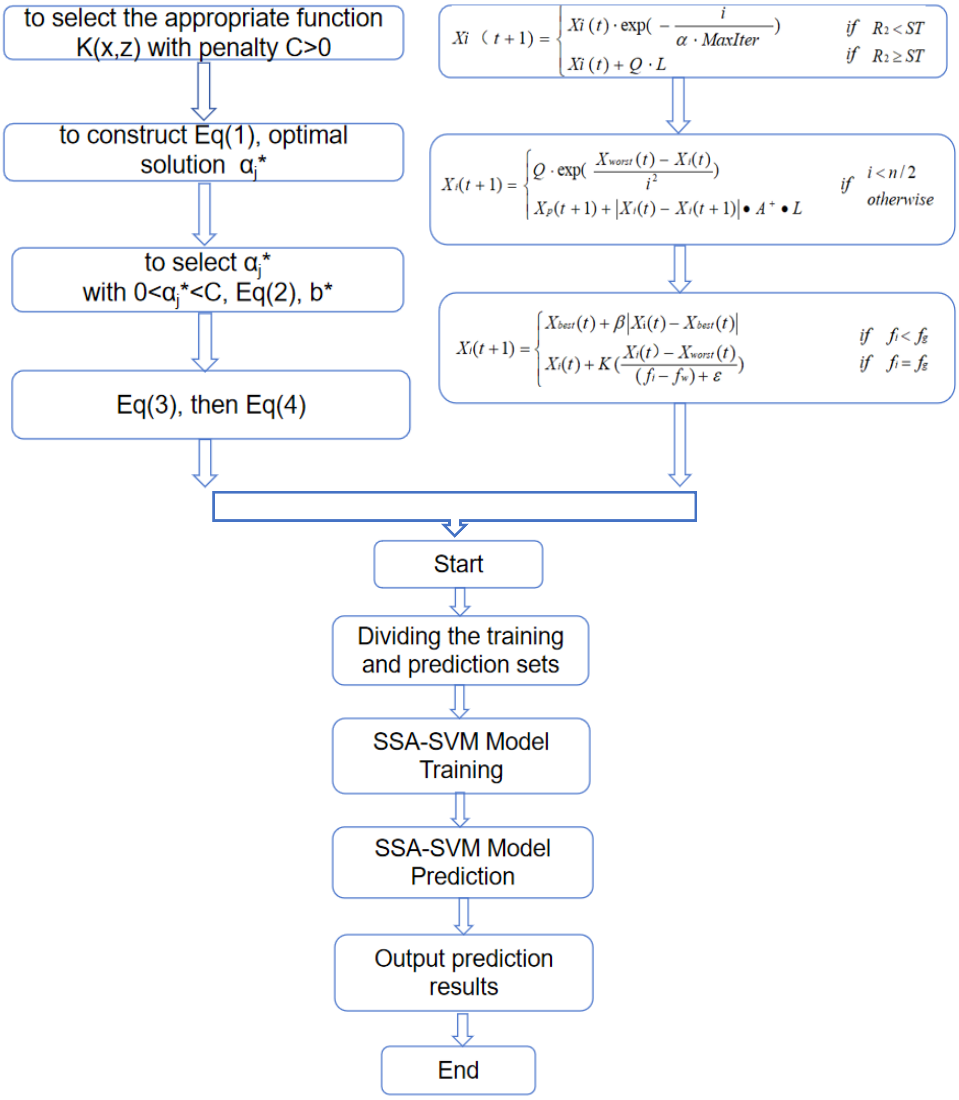
to select the appropriate function K(x,z) with penalty C>0
to construct Eq(1), optimal solution α_j^*
to select α_j^* with $0<\alpha_j^*<C$, Eq(2), b*
Eq(3), then Eq(4)
$Xi\ (t+1)=\begin{cases} Xi\,(t)\cdot\exp(-\frac{i}{\alpha\cdot MaxIter}) & if\ \ R_2<ST \\ Xi\,(t)+Q\cdot L & if\ \ R_2\ge ST \end{cases}$
$X_i(t+1)=\begin{cases} Q\cdot\exp(\frac{X_{worst}(t)-X_i(t)}{i^2}) & if\ \ i<n/2 \\ X_p(t+1)+\lvert X_i(t)-X_i(t+1)\rvert\bullet A^+\bullet L & otherwise \end{cases}$
$X_i(t+1)=\begin{cases} X_{best}(t)+\beta\lvert X_i(t)-X_{best}(t)\rvert & if\ \ f_i<f_g \\ X_i(t)+K(\frac{X_i(t)-X_{worst}(t)}{(f_i-f_w)+\varepsilon}) & if\ \ f_i=f_g \end{cases}$
Start
Dividing the training and prediction sets
SSA-SVM Model Training
SSA-SVM Model Prediction
Output prediction results
End

Figure 1. Methodology of the SSA-SVM model.

Table 1. Confusion matrix.

		True value	
Mixing matrix		Non-default sample	Sample default
Predicted value	Non-default sample	TP	FP
	Sample default	FN	TN

The model evaluation matrix in this paper mainly includes accuracy, recall, precision, ROC curve and AUC value.

(1) Accuracy

Accuracy (Accuracy) indicates the proportion of samples correctly predicted by the model to the overall sample, and is calculated as shown in Eq (8).

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN.}} \tag{8}$$

(2) Precision

Precision (Precision) indicates the proportion of samples correctly predicted by the model to the overall samples predicted by the model for that target feature, and is calculated as shown in Eq (9).

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}. \tag{9}$$

(3) Recall

Recall (Recall) represents the proportion of samples correctly predicted by the model to the overall samples actually for that target feature, and is calculated as shown in Eq (10).

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}. \tag{10}$$

(4) ROC curve

The ROC curve is a curve with sensitivity (TPR) as the vertical coordinate and specificity (FPR) as the horizontal coordinate, and the formula for calculating TPR and FPR is shown in Eq (11).

$$\begin{gathered} TPR=\frac{TP}{TP+FN.} \\ FPR=\frac{FP}{FP+TN}. \end{gathered} \tag{11}$$

The AUC value is the area under the ROC curve, when AUC = 1, it means the model is perfect, but there is no perfect prediction model in most cases; when 0.5 < AUC < 1, the model is better than random conjecture and has some application value; when AUC = 0.5, the model is the same as random conjecture and has no application value; when AUC < 0.5 the model is weaker than random conjecture. Therefore, the larger the AUC value is, the larger the area under the ROC curve, which means that the model is more effective.

5. Empirical of the SSA-SVM model

In this section, the data used in this model are briefly described for the processing of the data, empirically analyzed to derive prediction results, and analyzed in the context of the personal credit business of commercial banks.

5.1. Selection of data, statistical analysis and normalization

The data selected in this paper come from the personal credit data of the AliCloud Tianchi public database. The original data has 700 samples (183 defaulted samples, 517 non-defaulted samples) and each sample represents the personal information of one borrower customer, 8 feature variables (6 quantitative data, 1 fixed class data and 1 target feature variable), and the model variables are shown in Table 2. In this paper, the personal information of borrowers (age, education and working years, etc.) is used as input variables, and whether the lender defaults or not is used as output variable to construct a personal credit default prediction model.

Table 2. Description of original data variables.

Feature variable type	Feature variable name	Parameter code
Quantitative data	Age	X0
	Length of service	X1
	Revenue	X2
	Debt ratio	X3
	Credit card debt	X4
	Other liabilities	X5
Fixed class data	Education level	X6
	Whether in default (0: not in default, 1: in default)	Y

For the quantitative data, the analysis was mainly conducted by drawing box plots and observing the median, upper and lower quartiles and outliers of each characteristic variable under different default conditions. It can be seen that there is a difference between the age of borrowers in the defaulted and non-defaulted samples, with borrowers in the defaulted sample being relatively younger; the box plots for working age show that borrowers in the defaulted sample have a lower working age, both of which indicate that relatively younger borrowers or borrowers with a shorter working time are more likely to default. The box plots for debt ratios, credit card debt, and other debt show that there are large differences between both the defaulted and non-defaulted samples, with the defaulted sample having significantly higher debt ratios than the non-defaulted sample, and the median and quartiles of debt ratios being higher in the defaulted sample than in the non-defaulted sample. For borrowers, default is mainly due to the inability of available funds to repay existing debts, so borrowers with low debt ratios have less debt to repay than borrowers with high debt ratios, and have higher willingness to repay, resulting in fewer defaults, which leads to a much lower debt ratio in the non-default sample than in the default sample. Therefore, the probability of default increases for relatively young borrowers in the presence of high debt ratios. Finally, the level of education does not have a more significant effect on default status, with both the defaulted and non-defaulted samples showing some degree of decline as the education level increases.

Since there are differences in the range of values for quantitative data, such as significant differences between the debt ratio and income level, the data set needs to be normalized. If the original data set is directly applied, it will not only lead to an increase in computational effort, but also cause the problem of decreasing model accuracy. There are two common methods of normalization, one is linear normalization, which is a linear transformation of the original data, and the other is the Z-score standard deviation normalization method. The first method used in this paper is to create a MinMaxScaler function in Python to scale each quantitative data to the (0, 1) range.

5.2. Principal factor analysis of the data

Principal factor analysis refers to recombining the original data set into a new set of mutually uncorrelated data sets to reflect as much information as possible with as few variables as possible. Because of the large differences between the characteristic variables in the dataset of this paper and the reduction of model computing time, this paper chooses to synthesize the original quantitative data using principal factor analysis.

Table 3. KMO and Bartlett's tests.

KMO sampling suitability quantity	0.611	
Bartlett sphericity test	Approximate cardinality	2181.502
	Degree of freedom	15
	Significance	0

The KMO and Bartlett tests were first performed on the quantitative data, as shown in Table 3, and the approximate chi-square of Bartlett's sphericity test was 2181.502, which corresponds to a significance of 0. The original hypothesis that the correlation coefficient matrix is significantly different from the unit matrix should be rejected. Also, the KMO value is 0.611, so this data is suitable for principal component analysis.

Table 4. Total variance explained.

Ingredients	Initial Eigenvalue			Extraction of the sum of squares of loads		
	Total	Percentage of variance	Cumulative %	Total	Percentage of variance	Cumulative %
1	3.087	51.443	51.443	3.087	51.443	51.443
2	1.434	23.908	75.352	1.434	23.908	75.352
3	0.602	10.033	85.385	0.602	10.033	85.385
4	0.396	6.607	91.992			
5	0.369	6.144	98.136			
6	0.112	1.864	100.000			

In the following, the SPSS24 software was chosen to conduct principal component analysis on the variable data, and the total variance explained is shown in Table 4, the first three factors have 85.385% information contribution, and the original information is lost less, which has research significance. The plot data is shown in Table 4. The information contribution of the first factor is high, and the eigenvalues of the third and later factors are small, and their contribution to the explanation of the original variables is small and can be ignored, so the first three factors are extracted in this paper. Therefore, the experimental data of this paper are shown in Table 5, and the input variables X include the variables after the principal components, i.e., the composite variables F1, F2 and F3, and one fixed class data X6, and the output variable is whether or not the lender defaults (0: no default, 1: default).

Table 5. Description of experimental data variables.

Feature Variable Type	Feature variable name	Parameter Code
Input Variables	Combined variables	F1
		F2
		F3
	Education	X6
Output Variables	Whether in default (0: not in default, 1: in default)	Y

The formulae for the composite variables F1, F2 and F3 used for the experimental data can be derived from the matrix of component score coefficients from the principal component analysis, i.e., Table 6, whose expression is Eq (12).

$$\begin{aligned} F1 &= 0.202 * X0 + 0.232 * X1 + 0.264 * X2 + 0.134 * X3 + 0.260 * X4 + 0.272 * X5 \\ F2 &= -0.304 * X0 - 0.327 * X1 - 0.228 * X2 + 0.588 * X3 + 0.212 * X4 + 0.235 * X5 \\ F2 &= 1.102 * X0 - 0.018 * X1 - 0.584 * X2 + 0.433 * X3 - 0.321 * X4 - 0.075 * X5 \end{aligned} \tag{12}$$

Table 6. Component score coefficient matrix.

		Ingredients		
		1	2	3
Age	X0	0.202	–0.304	1.012
Length of service	X1	0.232	–0.327	–0.018
Revenue	X2	0.264	–0.228	–0.584
Debt ratio	X3	0.134	0.588	0.433
Credit Card Debt	X4	0.260	0.212	–0.321
Other liabilities	X5	0.272	0.235	–0.075

5.3. Prediction results and analysis of SSA-SVM model

Based on the above analysis, the original data and the processed dataset, i.e., the experimental data, are modeled and analyzed separately. Firstly, the dataset is divided into a training set and a prediction set in the ratio of 8:2 using train_test_split in python. Secondly, the model is trained on the training set and the prediction results are output on the prediction set, and the model accuracy is compared. CPU of the computer is 11th Gen Intel(R) Core(TM) i5-1155G7 @ 2.50 GHz and the RAM is 16 GB. Execution time for experiments is about 840 seconds.

For the traditional SVM model, the parameter combinations of both the original data and the experimental data are c = 1 and gamma = 1. While, in the SSA-SVM model, after the parameter search, the optimal parameter combination of the original data becomes c = 2.79 and gamma = 0.1, and the optimal parameter combination of the experimental data becomes c = 2.41 and gamma = 0.8. After the experiment, the SVM model and the prediction results of the SSA-SVM model are shown in Tables 7 and 8, the evaluation indexes are shown in Table 9, and the ROC curves are shown in Figure 6.

Table 7. Prediction results of the original data.

Original data	Projections			
Actual	SVM model		SSA-SVM model	
	0 (not in default)	1 (breach of contract)	0 (not in default)	1 (breach of contract)
0 (not in default)	103	0	102	1
1 (breach of contract)	37	0	30	7

From Tables 7 and 9, it can be seen that the prediction accuracy of the SSA-SVM model is higher than that of the SVM model for the original data. From Tables 8 and Table 9, it can be seen that the prediction accuracy of the SSA-SVM model is higher than that of the SVM model for the experimental data. From Tables 7 and 8, it can be seen that the prediction accuracy of the SSA-SVM model is higher than that of the SVM model for the defaulted samples. From Table 9, it can be seen that for the original data, the precision of the SVM model is only 73.6%, with an accuracy and recall rate of 0, while the precision of the SSA-SVM model can reach 77.9%, with an accuracy of 87.5% and a recall rate of 18.9%. For the experimental data, the precision of the SVM model is only 75%, with an accuracy rate of 63.2% and a recall rate of 30%, while the precision of the SSA-SVM model can reach 80%, with 70% accuracy and 52.5% recall. Therefore, the precision of the experimental data is all better than the original data under the premise of the same model. The precision of the SSA-SVM model is all better than the SVM model under the premise of the same data.

Table 8. Prediction results of experimental data.

Experimental data	Projections			
Actual	SVM model		SSA-SVM model	
	0 (not in default)	1 (breach of contract)	0 (not in default)	1 (breach of contract)
0 (not in default)	93	7	91	9
1 (breach of contract)	28	12	19	21

Table 9. Accuracy, precision and recall of the model.

Name	SVM model		SSA-SVM model	
	Original data	Experimental data	Original data	Experimental data
Precision	73.6%	75%	77.9%	80%
Accuracy	0%	63.2%	87.5%	70%
Recall	0%	30%	18.9%	52.5%

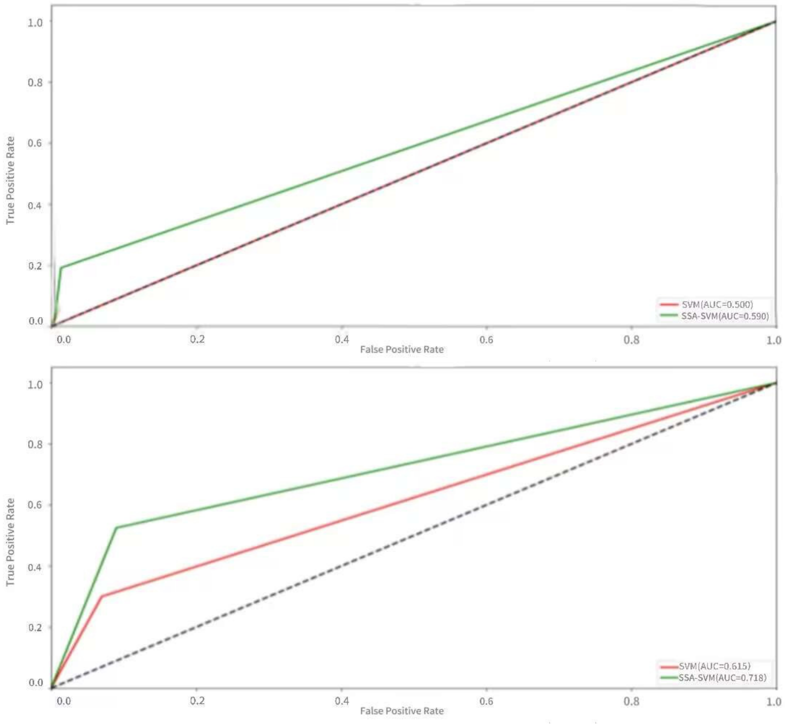

Figure 2. ROC curve.

As can be seen in Figure 2, the AUC value of the SVM model in the original data is 0.5 and the AUC value of the SSA-SVM model is 0.590, so the area occupied by the ROC curve of the SSA-SVM model is larger than the area occupied by the ROC curve of the SVM model, i.e., the SSA-SVM model outperforms the SVM model under the original data. The AUC value of the SVM model is 0.615 and the AUC value of the SSA-SVM model is 0.718, so the area occupied by the ROC curve of the SSA-SVM model is larger than that occupied by the ROC curve of the SVM model, i.e., the SSA-SVM model is superior to the SVM model under the experimental data.

The above experimental results show that, under the premise of the same model, all evaluation indexes of the experimental data are better than the original data, which indicates that the data processing operation performed on the original data in this paper is effective; under the premise of the same data, all evaluation indexes of the SSA-SVM model are better than the SVM model, which indicates that the SSA-SVM model established in this paper is better than the SVM model in personal credit default prediction and has some practical application value.

Therefore, commercial banks can use the SSA-SVM model proposed in this paper to establish a corresponding personal credit default assessment system to analyze the default status of borrowers and reduce the personal credit default phenomenon. At the same time, commercial banks can also establish corresponding default intervals based on the distribution of characteristics among borrowers, and reduce the amount of borrowing within the interval as much as possible, so as to avoid part of the default risk.

We use principal components as a preprocessing step. Since PCA are not very robust, other predictive models to predict the default probabilities can be considered, such as Figini et al. [31]. They propose a methodology for data fusion in longitudinal and survival duration models using quantitative and qualitative variables separately in the likelihood function and then combining their scores linearly by a weight, to obtain the corresponding probability of default for each SME.

6. Conclusions

In this paper, by collating the research of many scholars at home and abroad in personal credit classification prediction, we propose to establish the SSA-SVM model with sparrow search algorithm and optimized support vector machine. Prediction results show that the precision of the SVM model for the original data is only 73.6%, the accuracy and recall are 0 and the AUC value is 0.5, while SSA-SVM model for the original data prediction reaches 77.9% precision, 87.5% accuracy, 18.9% recall and 0.590 AUC value. The SVM model for experimental data has only 75% precision, 63.2% accuracy, 30% recall and 0.615 AUC value, while the SSA-SVM model for experimental data can reach 80% precision, 70% accuracy, recall 30% and AUC 0.615. Therefore, under the premise of the same model, all evaluation indexes of the experimental data are better than the original data; under the premise of the same data, all evaluation indexes of the SSA-SVM model are better than the SVM model, i.e., in terms of personal credit default prediction, the data processing operation performed on the original data in this paper is effective. The SSA-SVM model has a certain application value to the credit business of commercial banks.

However, this paper also has the following shortcomings, namely, the data were selected only taking into account the basic information factors of individuals and not combined with the national economic market and macroeconomic policies for the analysis. At the same time, the selected data has a vague description of the address, so it is not used in the process of empirical analysis. In the actual prediction of commercial banks, they can build a more perfect prediction model of personal credit default based on the real situation of address and other information, which may produce more accurate results.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Social Science Research Foundation of Ministry of Education of China (15YJA790051), National Social Science Fund Project of China (17BGL058) and Shandong Province Natural Science Foundation (ZR2022MG045).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. C. Thomas, A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers, *Int. J. Forecast.*, **16** (2000), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0
2. X. Q. Zhang, *Research on Credit Risk Measurement and Management of Commercial Banks in China*, Ph.D thesis, Harbin Engineering University, 2011. https://doi.org/10.7666/d.Y2236459
3. S. Finlay, Multiple classifier architectures and their application to credit risk assessment, *Eur. J. Oper. Res.*, **210** (2011), 368–378. https://doi.org/10.1016/j.ejor.2010.09.029
4. D. Durand, Risk elements in consumer instalment financing, in *NBER Books*, 1941.
5. F. Andersson, H. Mausser, D. Rosen, S. V. Uryasev, Credit risk optimization with conditional value-at-risk criterion, *Math. Prog.*, **89** (2001), 273–291. https://doi.org/10.1007/PL00011399
6. Q. Shi, Y. Jin, Consumer credit scoring model: A survey, *Stat. Res.*, **8** (2003), 36–39.
7. Z. G. Du, Overview of personal credit assessment and data mining, *South China Financ. Comput.*, **3** (2004), 8–11. https://doi.org/10.3969/j.issn.2095-0799.2004.03.004
8. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, (1995), 1–314. https://doi.org/10.1007/978-1-4757-2440-0_1
9. W. U. Chong, Y. Wang, Y. M. Guo, The model of personal credit risk assessment on support vector machine, *Oper. Res. Manage.*, **4** (2008).
10. Z. Y. Guo, Using support vector machine for the credit evaluation, *Comput. Knowl. Technol.*, **5** (2009).
11. J. W. Tian, Application of support vector machine and logistic regression model in the forecast of personal credit, *Reg. Finance Res.*, **11** (2018), 25–30.
12. H. L. Tang, B. Q. He, Z. Wei, Research on SVM-based credit evaluation model for bank personal loans, *West. Econ. Manage. Forum*, **23** (2012), 7. https://doi.org/10.3969/j.issn.2095-1124.2012.01.011

13. C. H. Shen, G. L. Liu, N. Y. Deng, An improved support vector classification and its application, *Comput. Eng.*, **8** (2005). https://doi.org/10.1038/sj.cr.7290370
14. B. Zhong, Z. Xiao, C. L. Liu, L. Chen, The credit evaluation method based on LS-SVM, *Stat. Res.*, **11** (2005).
15. Z. Xiao, W. J. Li, Personal credit scoring based on PCA and SVM, *Technol. Econ.*, **29** (2010).
16. D. P. Dai, L. P. Ni, M. Xue, Application of bank personal credit rating model based on k-means and SVM, *J. Jiangsu Univ. Sci. Technol.*, **31** (2017), 836–842.
17. X. Y. Liu, Y. M. Wang, Evaluation model for personal credit risk based on C4.5 algorithm for optimizing SVM, *Comput. Syst. Appl.*, **28** (2019), 6.
18. H. Wang, Research on personal credit assessment model based on SVM, *Sci. Technol. Entrepreneurship Mon.*, **32** (2019).
19. Q. Li, Logistic and SVM credit score models based on lasso variable selection, *J. Appl. Math. Phys.*, **7** (2019), 1131–1148. https://doi.org/10.1007/S12190-023-01877-5
20. Y. Wang, Y. F. Lu, Personal credit risk evaluation of SVM optimization based on shuffled frog leaping algorithm, *Heilongjiang Sci.*, **11** (2020), 2.
21. J. J. Chen, S. Liu, Personal credit evaluation based on SVM optimized by improved beetle swarm optimization algorithm, *Comput. Technol. Dev.*, **31** (2021), 5. https://doi.org/10.3969/j.issn.1673-629X.2021.06.024
22. J. K. Xue, B. Shen, A novel swarm intelligence optimization approach: Sparrow search algorithm, *Syst. Sci. Control Eng.*, **8** (2020), 22–34. https://doi.org/10.1080/21642583.2019.1708830
23. S. X. Wang, M. Zeng, Research on rural road cost forecast based on SSA optimized BP neural network, *Eng. Econ.*, **31** (2021), 25–29. https://doi.org/10.19298/j.cnki.1672-2442.202108025
24. Y. H. Liang, S. Y. Mao, J. F. Li, Research on rock burst data based on SSA optimized BP neural network, *Electron. Test.*, **36** (2022), 3.
25. H. Z. Hu, C. Qin, F. Guan, H. B. Zhang, S. J. An, Tool wear recognition based on sparrow search algorithm optimized support vector machine, *Sci. Technol. Eng.*, **21** (2021), 10755–10761. https://doi.org/10.3969/j.issn.1671-1815.2021.25.026
26. W. G. Wang, Y. B. Wei, X. D. Teng, Y. Huang, Short-term wind turbine generation power prediction based on sparrow search optimization support vector machine, *Intell. Comput. Appl.*, **12** (2022), 012.
27. L. Thomas, J. Crook, D. Edelman, Credit scoring and its applications, *Soc. Ind. Appl. Math.*, 2017. https://doi.org/10.1137/1.9781611974560
28. R. A. Anderson, *Credit Intelligence & Modelling: Many Paths Through the Forest of Credit Rating and Scoring*, Oxford University Press, 2022 https://doi.org/10.1093/oso/9780192844194.001.0001
29. D. B. Ntwiga, *Social Network Analysis for Credit Risk Modeling*, Ph.D. thesis, University of Nairobi, 2016.
30. G. Paleologo, A. Elisseeff, G. Antonini, Subagging for credit scoring models, *Eur. J. Oper. Res.*, **201** (2010), 490–499. https://doi.org/10.1016/j.ejor.2009.03.008
31. S. Figini, P. Giudici, Statistical merging of rating models, *J. Oper. Res. Soc.*, **62** (2011), 1067–1074. https://doi.org/10.1057/jors.2010.41