Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4515480, 14 pages
https://doi.org/10.1155/2020/4515480

Research Article

A VNS-EDA Algorithm-Based Feature Selection for Credit Risk Classification

Wei Chen,[1] **Zhongfei Li**,[1,2] **and Jinchao Guo**[2]

[1]*School of Business, Sun Yat-Sen University, Guangzhou 510275, China*
[2]*School of Management, Xinhua College of Sun Yat-Sen University, Guangzhou 510520, China*

Correspondence should be addressed to Zhongfei Li; lnslzf@mail.sysu.edu.cn

Received 28 December 2019; Revised 2 March 2020; Accepted 12 March 2020; Published 27 April 2020

Guest Editor: Dr. Dilbag Singh

Many quantitative credit scoring models have been developed for credit risk assessment. Irrelevant and redundant features may deteriorate the performance of credit risk classification. Feature selection with metaheuristic techniques can be applied to excavate the most significant features. However, metaheuristic techniques suffer from various issues such as being trapped in local optimum and premature convergence. Therefore, in this article, a hybrid variable neighborhood search and estimation of distribution technique with the elitist population strategy is proposed to identify the optimal feature subset. Variable neighborhood search with the elitist population strategy is used to direct its local searching in order to optimize the ergodicity, avoid premature convergence, and jump out of the local optimum in the searching process. The probabilistic model attempts to capture the probability distribution of the promising solutions which are biased towards the global optimum. The proposed technique has been tested on both publicly available credit datasets and a real-world credit dataset in China. Experimental analysis demonstrates that it outperforms existing techniques in large-scale credit datasets with high dimensionality, making it well suited for feature selection in credit risk classification.

1. Introduction

Credit risk assessment is one of the most important issues for serving small- and medium-sized enterprises (SMEs) in the commercial banking industry. Credit risk scoring model based on data mining and machine learning technique has already aroused great concern in financial credit field [1–4]. The scoring model can avoid some problems such as inaccuracy of risk classification and low precision of quantitative credit scoring, caused by the information asymmetry, to some extent. It is clear that, as an effective credit risk assessment tool, credit risk scoring model has played a more and more important role in financial institutions' credit decision-making.

Currently, data availability in credit loan is significantly enhanced by information technology. Multisource data, such as personal basic information, economy behavior, and social activity, are required for credit risk scoring purposes, in order to take preventive measures during the credit monitoring process and prioritize recovery efforts. However, these large-scale data are commonly high dimensional that leads to the curse of dimensionality. Redundant and irrelevant features in credit risk scoring model can increase the complexity of computation and produce inaccurate results in risk analysis. As we known, dimensionality reduction is used for feature extraction, abandonment, and decorrelation in machine learning. Feature selection and feature extraction are two common methods for dimensionality reduction [5–7]. Both the methods not only help in simplifying features but also can enhance stability and generalization of the classifier to improve learning ability, efficiency, and convenience. The major difference between two methods is that features obtained by feature extraction are not the original feature, while features obtained by feature selection are the part of original feature [8, 9]. From the degree of the combination with machine learning, feature selection can be divided into three categories such as filter method, wrapper method, and embedding method [9–12]. The first one is

irrespective to the subsequent machine learning process. The second one directly uses feature subset as criteria to evaluate the performance of learning model. The last technique integrates feature selection process into training a machine learning model. In practical classification problem, accuracy obtained by the wrapper and embedding method is usually higher than the filter method [13]. Besides, the optimal selection based on subspace searching can remove the redundant feature, and it makes the performance of this method is always better than the optimal selection based on sorting methods [14].

In credit risk assessment, feature selection is usually used in many different credit risk classification models and quantitative credit scoring model [15–17]. Financial institutions need to pay much attention to the most significant risk features which are from a large number of original features. Because they not only can better categorize credit risk, be further used to set interest on loan and construct prewarning management system, but also can be utilized directly for the establishment of credit risk system or evaluating index system to set up risk control mechanism. Consequently, identifying the optimal features is critical to increase the predictive accuracy and efficiency when using the credit risk scoring model. Metaheuristic techniques can be utilized in searching space reduction for feature selection [18]. However, metaheuristic techniques suffer from various issues such as being trapped in local optimum and premature convergence. Therefore, this article has proposed an efficient technique inspired by feature correlations and metaheuristic techniques to optimize feature selection quality. To evaluate the proposed method, the comparative experiments with state-of-the-art methods are carried out on three real-world credit datasets by considering classification quality measures: accuracy, F-measure, sensitivity, and specificity.

The rest of this article is organized as follows: Section 2 summarizes relevant literatures, Section 3 details the proposed technique, Section 4 describes credit datasets, Section 5 discusses experimental results, and Section 6 concludes the article and future directions.

2. Related Work

This section gives a brief review of previous works related to correlation-based feature selection and metaheuristic searching techniques and presents the aforementioned techniques in credit risk assessment.

2.1. Feature Selection Based on Correlations. Basically, the relevant features have two characteristics: the importance for the following applications such as being input variables in machine learning and low redundancy among the feature variables. A feature is considered irrelevant if it contains no information about classification. Redundancy is generally defined in terms of feature correlations. It is widely accepted that two features are redundant to each other if their values are correlated [19]. It has been shown that redundancy among the features can degrade performance. Furthermore, linear correlation analysis may not be able to detect nonlinear dependencies between features, especially for large-scale data mining problem. Therefore, feature correlations can be determined using symmetrical uncertainty (SU) [20] based on entropy which is a measure of uncertainty. It is empirically computed as follows:

$$SU(X,Y) = 2\left[\frac{H(X) - H(X \mid Y)}{H(X) + H(Y)}\right], \tag{1}$$

where $\mathrm{IG}(X \mid Y) = H(X) - H(X \mid Y)$ denotes information gain (IG) that measures the reduction in uncertainty about the value of X given the value of Y. Thus, the bigger the IG is, the more significant the correlation between X and Y is. $H(X) = -\sum_i P(x_i)\log_2(P(x_i))$ denotes $H(X)$ that represents entropy and measures the uncertainty about the values of X. Similarly, $H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j))$ denotes $H(X \mid Y)$ that is the conditional entropy and measures the uncertainty about the value of X given the value of Y.

Moreover, in order to overcome selection bias in favor of features with more values, each value should be normalized to [0, 1] for guaranteeing all values having the same scale. A value of 1 indicates that the value of either feature completely predicts the value of the other; a value of 0 indicates that X and Y are independent. Thus, SU can be used as a correlation measure between features. In feature correlation analysis, X and Y, respectively, represent the feature F_r and the classification Y. Therefore, the feature subset evaluation function of correlation-based feature selection (CFS) [21] is computed as follows:

$$M_S = k\sqrt{k + k(k-1)^{\mathrm{SU_{FF}}}}^{\mathrm{SU_{CF}}}, \tag{2}$$

where $\mathrm{SU_{CF}} = (1/k)\cdot\sum_{F_{ri}\in S}\mathrm{SU}(F_{ri}, Y)$ denotes the average correlations between feature F_r and the classification Y. $\mathrm{SU_{FF}} = (2/(k(k-1)))\cdot\sum_{F_{ri}\in S}\sum_{\substack{F_{rj}\in S \\ F_{ri}\neq Fj}}\mathrm{SU}(F_{ri}, F_{rj})$ denotes the average correlations between feature F_{ri} and the feature F_{rj}. CFS is based on feature subspace searching and derived from the definition of the evaluation function. The optimal feature subset with correlation measurement can be successfully obtained, and the predominant features are highly relevant to classification but are of small or no significant correlation to eachother.

2.2. Feature Selection Combining with Metaheuristic Techniques. Although there exist numerous feature selection techniques, the challenging issues that handle a large dimensionality number of samples has gained increasing attention [22]. In credit risk assessment, the goal of feature selection is to minimize the error with respect to the given inputs for credit classification and scoring. Piramuthu [23] discussed a few means of improving credit risk performance through data preprocessing, specifically through feature selection and construction. Because of computing complexity, metaheuristic techniques are often used to raise efficiency. Hybrid feature selection with metaheuristic techniques is usually considered as a combination of filter

and wrapper technique to improve learning performance and enhance comprehensibility.

Feature selection can be treated as a combinatorial optimization problem. The binary value of the variable x_i indicates that feature i from N features is present ($x_i = 1$) or absent ($x_i = 0$) in a reduced feature subset. The problem can be represented as follows:

$$\max_{x=(x_1,\cdots x_n)\in\{0;1\}^n} F(x), \tag{3}$$

where $F(x)$ denotes the objective function of the practical problems. In credit risk classification, $F(x)$ specifically represents classification accuracy of default risk. Selecting n appropriate features from a set of N features has been proved to be a NP-hard problem [24]. Metaheuristic techniques are used as an objective optimization method that can realize continuously adaptive optimizing searching, which is capable to balance diversity, accuracy, and efficiency. It is well suitable for solving feature selection optimization problems [25, 26]. Especially, metaheuristic techniques based on swarm intelligence and evolutionary algorithm can better complete continuously adaptive optimizing searching, which has been widely applied in many research fields [27–32]. Therefore, feature selection with metaheuristic techniques, such as genetic algorithm (GA), simulated annealing (SA), and particle swarm optimization (PSO), are commonly used to identify the optimal feature subset and enhance classification accuracy in credit risk assessment [33–38].

It can be seen more and more works incorporate metaheuristic techniques to select the relevant features, rather than enumerating all features. However, there are two main limitations to this kind of approach: trapping in local optimum and being sensitive in initial solution setting. Aiming at the two limitations, an efficient hybrid variable neighborhood search (VNS) and estimation of distribution algorithm (EDA) with the elitist population strategy is proposed, which is expected to reap the benefits of accuracy and simplicity from the traditional feature selection and keep the computational expense down from metaheuristic searching. EDA is a stochastic optimization searching algorithm based on building and sampling explicit probabilistic models of promising candidate solutions [39] and directly uses the objective function as searching message. The key aspect of EDA is that the probability distribution of the solution space is defined explicitly, whereas in most evolutionary algorithms, the distribution is defined implicitly. This explicit utilization in optimization offers some significant advantages over other metaheuristics in evolutionary computation, such as compared with the novel GA, EDA can improve a population of solutions with random sampling from the probability distribution, rather than with random and adaptive searching probabilistic method based on biosphere natural selection and genetic mechanism. GA specifically uses the designed mutation and crossover operator to generate a new adaptive population, while EDA estimates the probability distribution of the solution space and updates the probabilistic model with superior population to generate a new and better candidate population replacing the old population entirely, and the sampling procedure is repeated. Eventually, the optimal solution can be obtained until the termination of iterative constraint is met. Besides, EDA provides explicit evolutionary procedure to present how the problem is solved with a great deal of candidate solutions. Thus, EDA is successfully applied to a large number of combinatorial optimization problems, such as schedule optimization [40], multiobjective knapsack [41], and feature extraction for bioinformatics [42]. In addition, VNS is one of the incremental optimization algorithms developed from local searching techniques that expand searching space of potential solutions. It generally explores specific neighborhood based on the current incumbent solution and searches potential promising neighborhood solutions only if an improvement has been generated [43]. In general, the larger the local searching range is, the greater the chance that a high-quality optimal solution can be obtained. Specific neighborhood structure and moving operators are designed to perform a systematic searching for improving searching capability, which can reduce computational time when expanding the diversity of solutions.

3. VNS-EDA with the Elitist Population Strategy

The proposed technique achieves its objectives by using two main well-established techniques: VNS and EDA. The critical aspects are discussed in this section, including feature coding, solution of the adaptability function of objectives, and incremental searching with the elitist population strategy based on the hybridization of metaheuristic searching.

Machine learning is actually treated as the optimization process of the loss between the predicted and real data. It mainly relies on the optimization algorithms to approach the optimal solution for reaching minimum loss. However, fitting with the large-scale data is required to solve nonconvex problem, and it is always possible to fall into local optimum. Additionally, the evolutionary optimization techniques always begin with a random initial population and then evolve from one generation to another when the evolution stops. Many simulation experiments indicate that it often generates a satisfied solution for middle or small-scale applications within permissive time, not for the large-scale learning problem. Therefore, the proposed technique is based on the following two basic rules: take advantage of high-quality solutions from prior information and achieve a better tradeoff between accuracy and convergence including two aspects: one is neighborhood expansion for local searching, and another one is a probabilistic model for global searching in solution space. The proposed technique combines the good property of local heuristic searching with the elitist population and the good distribution of global heuristic searching in the evolution process.

Specifically, when problem-specific information is available, one party of initial population must be inserted a set of high-quality solutions. Inserting high-quality solutions into the initial solutions that represents the explicit significant features with low redundancy are selected to become a basic component of the population. The remainder of the

population is filled by using neighborhood operators that represent the potential significant features. The neighborhood structure is defined as the solution space to keep population diversity, which is very important to avoid premature convergence. The union of the elitist population and the new generation is defined as the predominant population, and it is indispensable to keep dominance and ensure the effectiveness of convergence. Additionally, each searching step will converge to the global optimum using a probabilistic model in the form of a probability vector of the promising solution. Therefore, the dominance and flexibility can be passed on to offspring, and the approximate optimum can be attained.

3.1. Feature Coding and Fitness Function. Feature selection is formulated as the unconstrained optimization problem. Thus, all features are transferred into bit string by binary coding. "One of K" rules is used for representing the feature mask. The bit with value 1 means one feature is selected, and the number 0 indicates this feature is not selected; that is, a list of 0 with N elements representing all features are not selected. If one feature is selected, the value corresponding to the bit variable is set to 1. Thus, the rule defines that when the ith feature of N features is selected, the ith element value is set to 1 in the list, and other unselected feature variables are set 0. Each individual in population represents a selected feature subset.

The global optimum can be reasonably found by giving suitable and comprehensive fitness evaluation function. In order to describe direct acting factors for feature selection in credit risk classification, a fitness evaluation is presented based on the classification accuracy and the number of features as depicted in equation (4). Both are two important factors that directly affect feature selection quality [44]. One predefined weight w_a is for the classification accuracy, and the other w_f is for the proportion of the selected features in all original features. The former suggests the precision and reliability of the model, and the latter reflects the complexity of the model. Generally, the higher the accuracy with the fewer features is, the higher the fitness evaluation value is:

$$\text{fitness} = w_a \times \text{accuarcy} - w_f \times \frac{m}{M}, \tag{4}$$

where accuarcy denotes the classification accuracy predicted by a fixed classifier, m is the number of the selected features, and M is the number of all original features.

3.2. Searching Space Reduction and Refining. As we known, feature selection with metaheuristic techniques typically do not require any information about the problem being solved except for the representation of solution and fitness function. Nonetheless, the drawback is that the global optimum cannot be always guaranteed because of randomicity and premature convergence. The enhancements in the proposed technique are accomplished in two stages for targeting the optimal feature subset: (1) the generation of initial solution incorporating feature correlations and (2) the generation of feature subset solution that is being more likely to the global optimum. The aim of the first stage is preliminary for restricting the solution searching space in a large-scale feature selection. Given that a good problem-specific knowledge essentially and significantly affects the performance, additional prior information can be an available measure for a good and fast selection. The second stage makes the reduced feature subset to be refined by updating probability vector of the predominant population.

In the first stage, the generation of the initial solution is performed to prevent the proposed technique from consuming time in exploring irrelevant features. The redundancy among the features can cause the degradation of classification performance. Starting point has a very important role in the searching space reduction. One technique used to bias the initial population towards good solutions is called seeding work [45, 46] by inserting high-quality solutions into the initial population. Seeding work attempts to achieve a better tradeoff between searching efficiency and convergence by incorporating prior problem-specific information. Without considering any prior information, a large solution space often results in selecting a large number of feature subsets, even can lead to the omission of relevant features. Instead, it can simplify the scale of optimization and reduce the searching scope. For example, the searching space consists of 2^d solutions in CFS, a common greedy algorithm sequential forward selection (SFS) can be used to perform an intelligent searching, and the number of solutions can be dropped to $d(d-1)/2$.

In the second stage, the generation of the feature subset refining represents that the candidates without sacrificing the solution quality can be obtained by iterative searching, based on an efficient fitness evaluation through probability distribution statistics in the restricted scope. It details the union of the elitist and the potential population that may be incorporated into the probabilistic model to speed up the optimization, which can mitigate the effect of the initial choice. The local searching is often applied to a number of randomly generated initial solutions. The diversity given by the neighborhood structure are embedded into the evolutionary process so as to avoid premature convergence and not to neglect potential relevant features. Most of local searching are based on the r-flip neighborhood. In this article, we use 1 and 2-flip neighborhoods considering the computational time. After converting each feature subset to the binary coding population, 1-flip and 2-flip neighborhoods are used for a slightly varied solutions, which represents a permutation with a relatively lower computing time. New populations are attained by removing random parts and generating the replaced parts with sampling from 1-flip and 2-flip. During the processing of neighborhood searching from iteration to iteration, 1-flip neighborhood operator firstly is used to meet the requirement of slightly varied solutions. As the normalized solution space without improving their quality anymore, 2-flip neighborhood operator is performed to explore a slightly larger space. If it is possible to further improve the solution quality, 1-flip neighborhood is again used for iteration or otherwise continues to search in a larger solution space based on 2-flip neighborhood.

Although a slightly neighborhood can help to actualize local searching, the randomness may lead to the weak convergence. Thus, the high-quality solutions can be achieved by the elitist population strategy. The elite reproduction can retain good individuals and guarantee the number stability of the population in each iteration. Increasing the diversity with the elitist population strategy can mitigate risk to trap in local optimum and avoid premature convergence. In the refining process, only the union of the elitist population and the offspring which come from neighborhood structure will appear at the starting of the evolutionary process and act as the reduced feature subset. With this way, these restrictions to the searching space can reduce the irrelevant feature efficiently, and the candidates can be limited in the feasible space. Additionally, the candidates can be represented as probability vectors. The probabilistic model attempts to make the higher probability vectors sampled. The higher probability vector in the probabilistic model can be obtained from individuals with the best values of fitness evaluation, and each generation is updated by superior population, which is more strongly biased towards the global optimum. The influence of the population on the convergence rate to an optimum can be explained by the population evolution with competitive learning. Predominant population always make the probability of generating the global optimum that has increased when the iteration is continued.

More specifically, in each iteration, the candidates are ordered according to their fitness value, and the different ones which have the best fitness are selected to establish the probabilistic model. It is important to note that, sometimes, the number of selected individuals is less than or equal to the population. In either case, the population is supplemented with the best fitness individuals. The probabilistic model is constructed by the predominant variables from the best fitness individuals. The probability distribution of any individual should depend on the joint probability distribution, which defines probabilities of the corresponding variable in each generation. Due to low redundancy among the reduced features, each feature can be generated independently based on the entries in the probability vector. Suppose that the probability vector $p = (p_1, p_2, \ldots, p_n)$ which is sampled to generate new candidate is based on the probabilistic model. A value of 1 is generated in position i in the selected solution with probability p_i. In each generation, the probability vector p_i is updated according to equation (5), and each p_i is set to the proportion of 1s in the new population. Then, most best solutions are selected to give higher probabilities to the promising solutions which represent the samples of the global optimum. Equation (5) is depicted as follows:

$$p_l(x) = p\left(x|D_l^s\right) = \prod_{i=1}^{n} p_l(x_i) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{N} \delta_j\left(X_i = x_i \middle| D_l^s\right)}{N}, \tag{5}$$

where $p_i = \delta_j(X_i = x_i|D_l^s) = \begin{cases} 1 & X_i = x_i \\ 0 & \text{else} \end{cases}$ denotes the probability of generating a value of 1 in position i of solution strings. This repeated refinement of the probabilistic model can keep increasing the probability to generate predominant solutions. After a reasonable number of iteration, the global optimum with high probability would be generated and then reformulated into the feature subsets.

3.3. The General Framework of VNS-EDA Technique. In order to precisely describe the framework of VNS-EDA technique, the main steps are demonstrated, as shown in Figure 1. The detailed explanation is as follows. At first, the original features are extracted by feature redundancy elimination with correlation measurement. Then, the neighborhood operators with the elite reproduction retain the predominant population and guarantee the number stability of the population to get a good population diversity and convergence. The best candidates among the obtained locally optimal solutions are output. Furthermore, the probabilistic model is used to direct its global searching in solution space to jump out of the local optimum. Finally, new high-quality solutions are extracted and fed to a classifier, which is trained offline using labeled training data. The testing data are used to evaluate performance metrics. Thus, the proposed technique helps in reducing the searching region and has a higher effectiveness in reaching the global optimum. Algorithm 1 presents the detail of the proposed technique.

4. Credit Datasets

Two types of datasets are used for performance analysis: (a) the private ZhongAn Credit dataset which is further described in the section and (b) the publicly available German Credit dataset and LendingClub dataset, which have been frequently used as benchmarks in the literatures and in data science competitions.

4.1. ZhongAn Dataset. This private credit dataset is collected from a business loan company called ZhongAn Credit in Shenzhen, China. The company mainly serves SMEs to offer a medium-sized credit loan between CNY50000 and CNY500000. The dataset covered the period from January to December in 2017. The total number of samples is 16900, including 15419 nondefault customers who successfully fulfilled their credit obligations as positive samples and 1481 default customers who were late in performing their obligations as negative samples. It can be seen that the ratio of positive to negative samples is about 10 : 1. It is known that the classifier often breaks down when the size of the training examples per class is not balanced. Thus, the negative number 1481 is multiplied by the empirical member 5, as a reference value for the number of positive samples. Eventually, the 7405 nondefault customers are selected from 15419 nondefault customers as the training samples by the random stratified sampling method. Based on data availability and quality, the list of some typical features with their explanations can be found in Table 1. The expression and meanings of other variables are omitted due to page limit.

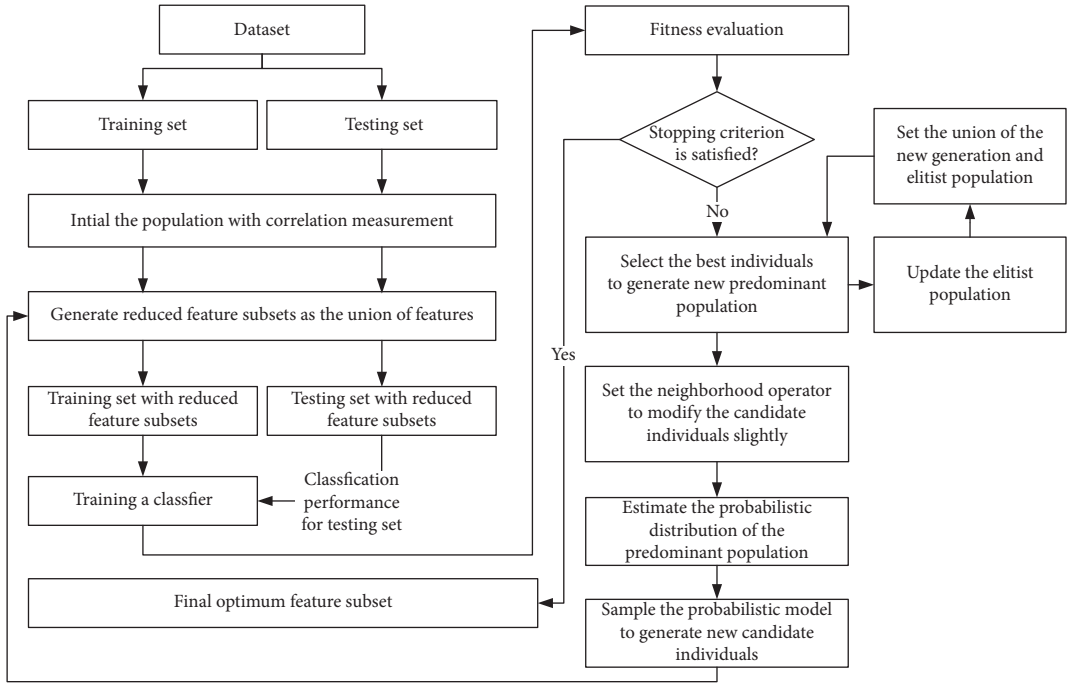

Figure 1: The general framework of the proposed VNS-EDA feature selection.

Input: the original features D, population size M, predominant population size N, weighting parameters w_a and w_f, neighborhood operators *one-flip* and *two-flip*, and iterations n
Output: the best feature subsets D'
(1) begin
(2) generate random M individuals of initial population $P(0)$ according to the values of the symmetrical uncertainty (correlation measurement between features)
(3) set the neighborhood operators *one-flip* and *two-flip*, designate an empty set as the elitist population $S = [\,]$
(5) while (termination criteria not met) do
(6) compute the fitness of each individuals according to equation (4) in $P(0)$
(7) $k \longleftarrow 0$
(8) select the top N promising individuals from $P(0) \cup S$ as a predominant population $P(g)$ and update the elitist population $S(g)$ with $P(g)$
(9) use the neighborhood operators to generate new individuals in $P(g)$
(10) build the probabilistic model $M(g) = P_{l+1}(X)$ from $P(g)$ according to equation (5)
(11) sample $M(g)$ to generate new candidate individuals as a new generation $P'(g)$
(12) $k \longleftarrow k + 1$
(13) end while
(14) the process ends when the termination criteria is satisfied

Algorithm 1: VNS-EDA with the elitist population strategy.

4.2. Two Public Datasets. German credit dataset is available from the UCI Repository of Machine Learning Databases [47], and it consists of 20 input features with 7 numerical and 13 categorical variables which describe the credit history, loan amounts, loan purposes, personal information, and so on. This original dataset is composed of 700 instances of creditworthy applicants and 300 instances of credit applicants who are default, and it is small scale compared with the other two datasets. The full list of features with their explanations and descriptive statistics can be found in the literature [33].

LendingClub is a US peer-to-peer (P2P) lending company which offers loan trading between \$1000 and \$40000 for high-credit worthy borrowers based on their information including borrower and loan attributes, and the standard loan period is usually three years. This original dataset can be obtained from the official website https://www.lendingclub.com/info/download-data.action. It covered the period from January to June in 2018, and loan status consisted of seven status with "Current", "Issued", "Fully Paid", "In Grace Period", "Late (31–120 days)", "Late (16–30 days)", "Charged Off", and "Default". Due to the object of credit risk classification, "Fully Paid" is considered to be the creditworthy applicants; "In Grace Period", "Late (31–120 days)", "Late (16–30 days)", "Charged Off", and "Default" are treated as

Table 1: Credit risk feature variable.

	Feature variable	Data type	
Repayment capacity	Liquidity ratio	Number	Short-term liquidity
	Quick ratio	Number	
	...	...	
	Cash ratio	Number	
	Inventory turnover	Number	Operational capability
	Accounts receivable turnover	Number	
	Total asset turnover	Number	
	Cash conversion cycle	Number	
	...	...	
	Fixed assets turnover	Number	Profit-making capability
	Asset profit ratio	Number	
	Gross profit ratio	Number	
	...	...	
	Operating profit ratio	Number	
Repayment willing	Age	Number	
	Education	Category	Junior high school, high school, bachelor degree, or above
	Housing	Category	Mortgage loan, mortgage without loan, rental, others
	Gender	Category	Male, female
	...	...	...
	Marital	Category	Married, single, divorced
	Domicile	Category	Residents, nonresidents, others

the bad credit applicants. The total number of samples is 10613. It is composed of 7692 good instances and 2921 bad instances. Because the ratio of good to bad instances is about 3 : 1, the size of the training examples per class can be considered to be balanced. Some of the important features with descriptive statistics are shown in Table 2.

5. Experiments

5.1. Implementation Details. The experiments are conducted to evaluate the proposed and existing techniques for credit risk classification. Data clearing, converting, and sampling are completed before the comparative experiments. Generally, the categorical features are transformed into the numerical ones and converted into binary string, the missing data can be filled with the median amount, and each original feature is linearly scaled to the range [0, 1]. Besides, the class label in all datasets is defined as 0 for nondefault (positive) samples and 1 for default (negative) samples. The size of the examples per class can be considered to reach a balance. Python language is used for exploratory data analysis and engineering algorithms. We build the packing algorithm component and the learning model using Scikit learn with Keras and Deap [48].

The comparisons are drawn between the proposed and three state-of-the-art techniques: population-based incremental learning (PBIL), GA, and PSO. Each generated solution represented by binary data is extracted and converted into different feature subsets. The number of features in all reduced subsets is apparently different. A simple SVM classifier is learned from large amounts of labeled data with the original feature and the reduced feature subsets. Moreover, 10-fold cross-validation has been used to verify the comparative performance metrics, which is widely used to overcome the underfitting and overfitting issue in many literatures [49]. The whole sample on one dataset are divided into three parts: training data (70%), validation data (15%), and testing data (15%).

Moreover, the common parameters are set to the same values. The parameters of the proposed technique are as the following: population size $M = 20$, predominant population size $N = 10$, and weighting parameters $w_a = 0.5$ and $w_f = 0.5$; neighborhood operators are one-flip and two-flip. The unique parameters of the other algorithms are set according to the general empirical values [50]. GA is as following: population size $M = 20$, crossover rate cxpb = 0.5, mutation rate mutp = 0.2, one-point crossover, roulette wheel selection, and elitism replacement. The parameters of PSO are set as follows: population size $N = 20$, inertia weight $w = 1$, acceleration constants $c_1 = 2$ and $c_2 = 2$, and the maximum limited velocity $V_{\max} = 6$. Specially, one incremental univariate EDA is PBIL [51] which is used to perform the probabilistic model, and the probability vector of the best solutions is shifted as follows:

$$p_{l+1}(x) = (1 - \alpha) p_l(x) + \alpha \frac{1}{N} \sum_{k=1}^{N} x_l^k, \tag{6}$$

where p_l is the probability of generating a 1 in generation l, $x_l^1, x_l^2, \ldots, x_l^N$ denotes the N best individuals, and α is the learning rate.

5.2. Experimental Results and Discussion. In order to evaluate the feasibility and effectiveness of the proposed technique, it is compared with other techniques on all experimental datasets. Table 3 shows the performance metrics of the proposed technique on three datasets, and it lists the accuracy, F-measure, sensitivity, and specificity values, which is quite consistent with training and testing samples. Typically, classification accuracy analysis of the

Table 2: LendingClub dataset with descriptive statistics.

Number	Feature	Statistics	Range
1	Loan_amnt	Avg = 15413	[1000, 40000]
2	Int_rate	Avg = 13.81	[5.31, 30.94]
3	Installment	Avg = 456.51	[30.12, 1556.8]
4	Open_acc	Avg = 11.17	[0, 57]
5	Open_acc_6 m	Avg = 1.07	[0, 12]
6	Total_bal_il	Avg = 36233.93	[0, 923836]
7	...	...	...
8	Open_rv_12 m	Avg = 1.37	[0, 17]
9	Max_bal_bc	Avg = 5255.02	[0, 104320]
10	Acc_open_past_24 mths	Avg = 4.97	[0, 31]
11	Mths_since_recent_inq	Avg = 5.90	[0, 24]
12	Num_bc_tl	Avg = 7.35	[0, 44]
13	Num_tl_op_past_12 m	Avg = 2.36	[0, 18]
14	Total_il_high_credit_limit	Avg = 45162.75	[0, 817057]
15	Annual_inc	Avg = 80672.79	[0, 93000]

Table 3: Performance metrics of VNS-EDA with the optimal feature subset.

Dataset	Accuracy (%)		F-measure (%)		Sensitivity (%)		Specificity (%)	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ZhongAn Credit	86.3	85.6	77.2	76.3	61.5	60.1	91.2	90.6
German Bank	76.2	74.5	72.9	71.4	65.6	63.8	87.7	85.1
LendingClub	74.3	75.2	70.6	71.8	60.2	61.5	87.9	89.3

Table 4: Performance results of the comparative method.

Dataset	ZhongAn Credit		German Bank		LendingClub	
	Accuracy (%)	Number of features	Accuracy (%)	Number of features	Accuracy (%)	Number of features
None	75.3	87	70.5	20	68.0	72
GA	82.7	35	74.0	8	72.8	35
PSO	82.9	36	75.5	10	72.7	32
PBIL	83.4	32	73.0	12	73.1	32
Proposed	85.6	30	74.5	10	75.2	28

proposed and other techniques is shown in Table 4 and Figure 2. Similarly, Figures 3–5 depict, respectively, F-measure, sensitivity, and specificity of the proposed and other techniques. We can draw the following observations.

Firstly, not surprisingly, the performance metrics of the original features is the worst. The proposed technique outperforms others in terms of accuracy, F-measure, sensitivity, and specificity on ZhongAn and LendingClub datasets. Additionally, PBIL is slightly better than GP and PSO in these two datasets, and it might be because of its advantages for the searching direction of optimization process in evolutionary computing. Furthermore, the proposed technique generates the fewest reduced features among these comparison techniques, as shown in Table 4. Although the number obtained by PBIL is only a little more than the proposed technique, its performance metrics fall. It appears that PBIL can create new individuals (solutions) in each generation by evolving from fine individuals (solutions) of the previous generation. The solutions, because of the predefined solution searching space, might get into the local

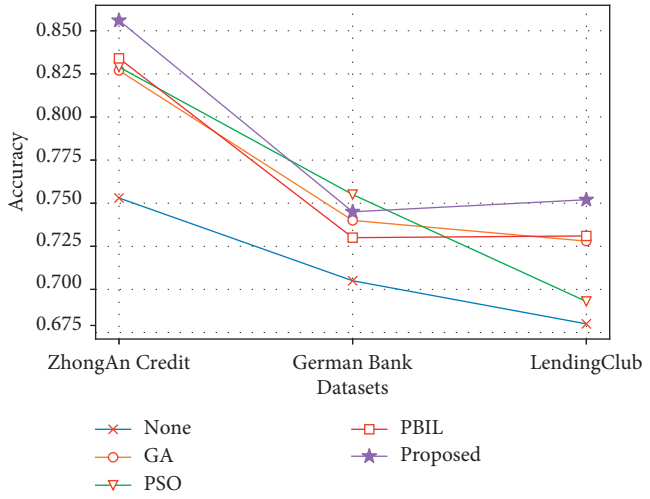

Figure 2: Comparative analysis of accuracy.

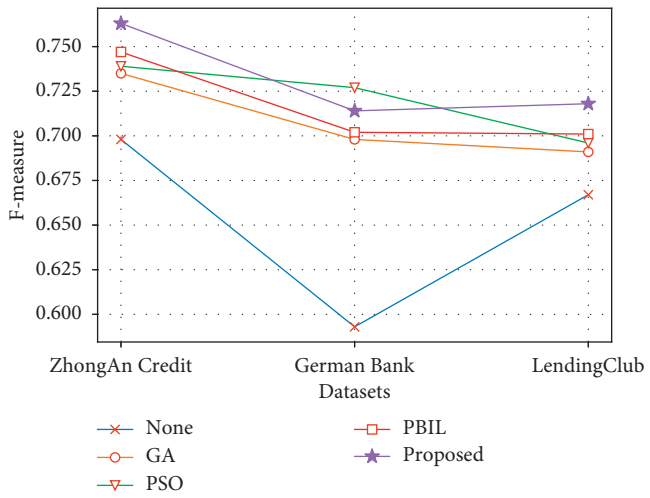

Figure 3: Comparative analysis of F-measure.

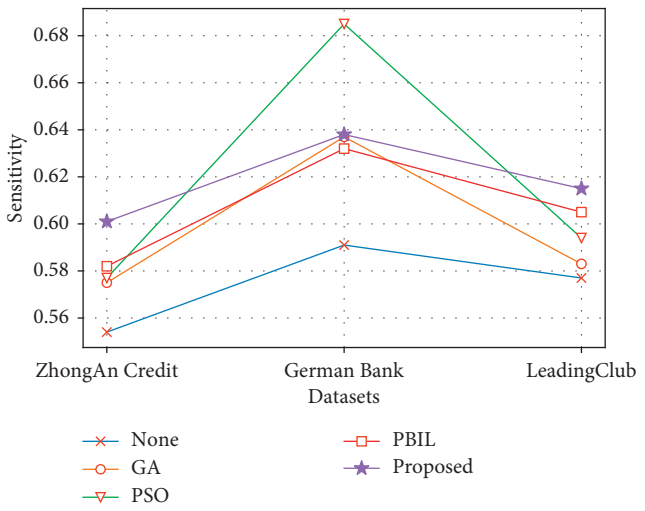

Figure 4: Comparative analysis of sensitivity.

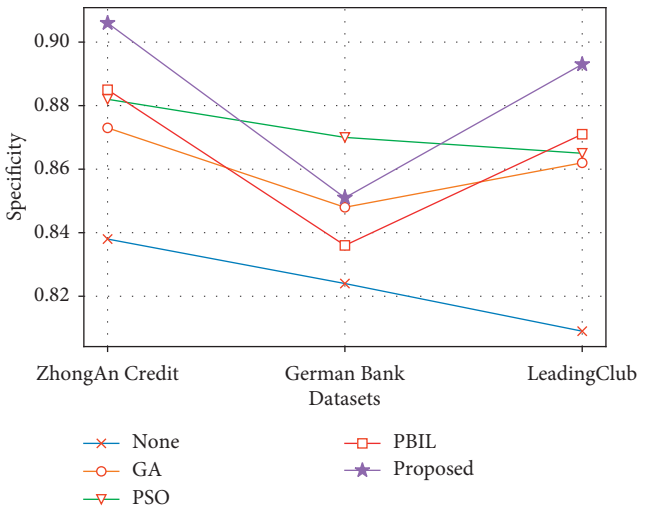

Figure 5: Comparative analysis of specificity.

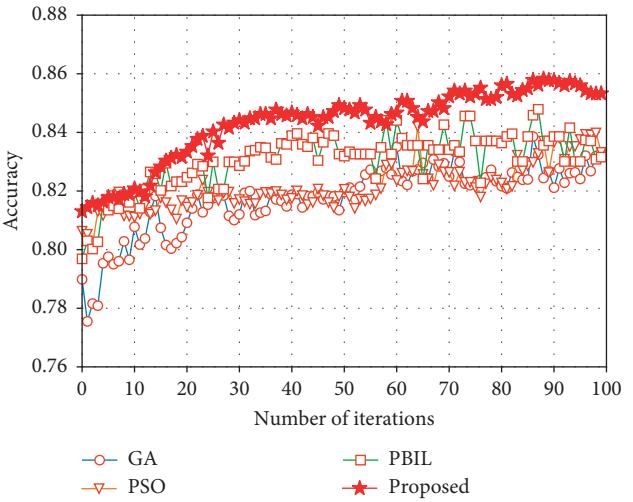

Figure 6: Comparative results of iteration on ZhongAn dataset.

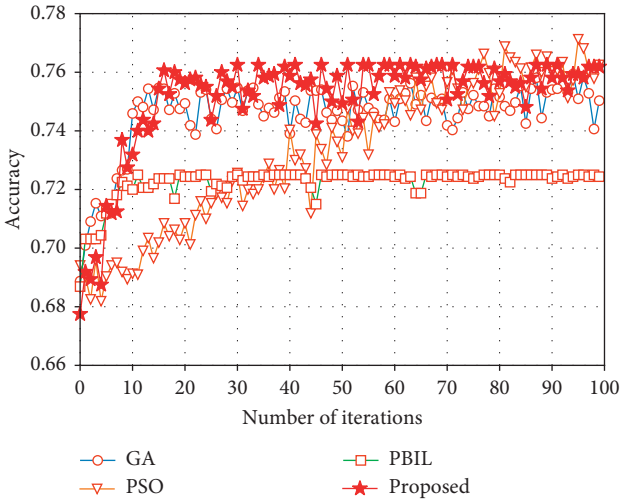

Figure 7: Comparative results of iteration on the German dataset.

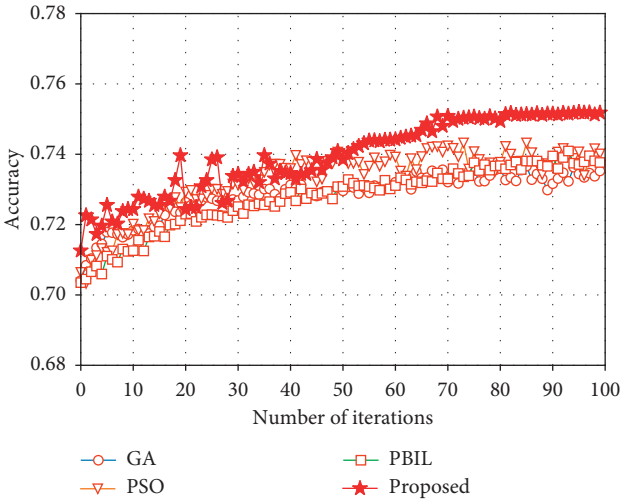

Figure 8: Comparative results of iteration on the LendingClub dataset.

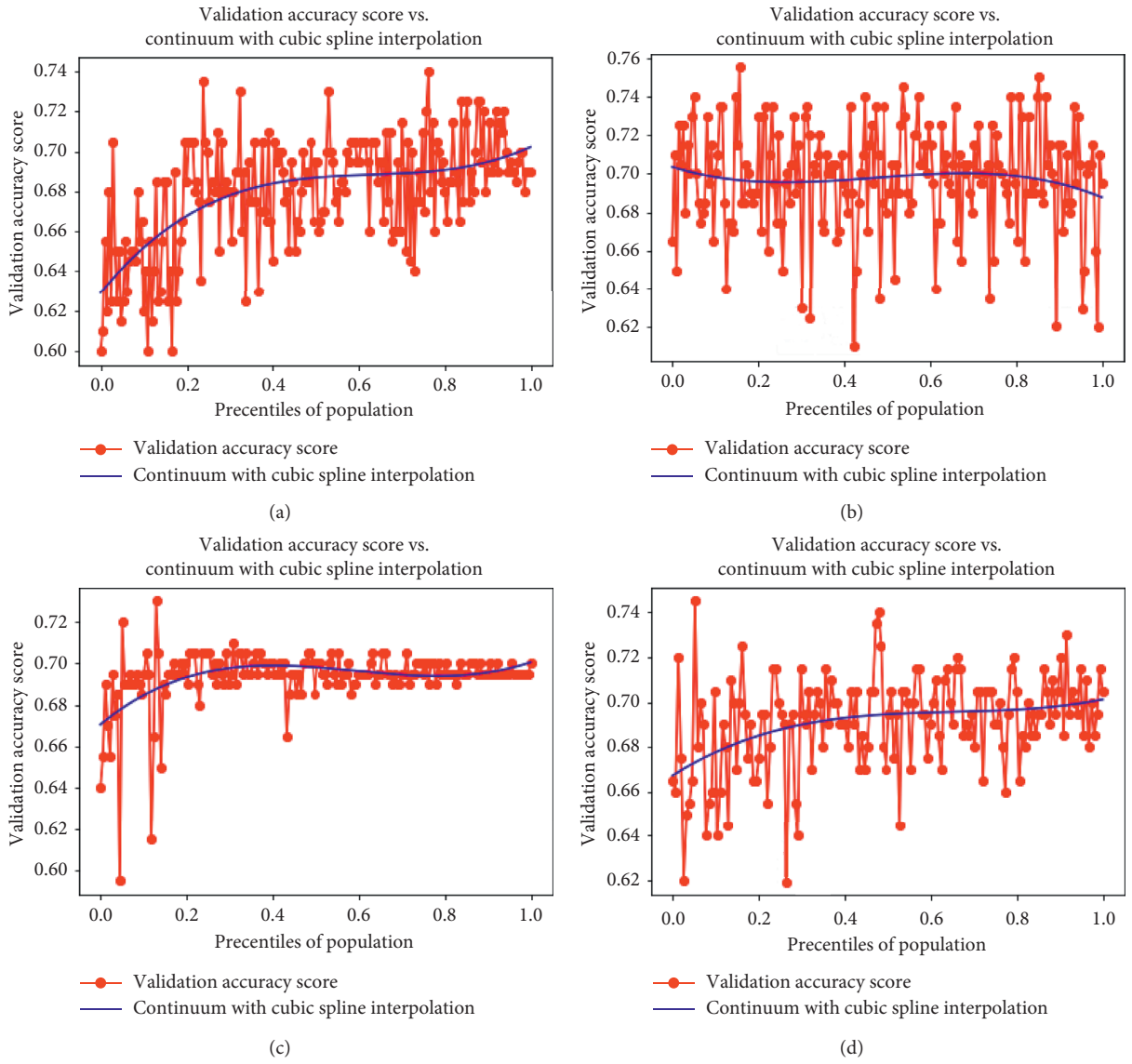

Figure 9: Comparative accuracy of all solutions on the German dataset: (a) genetic algorithm; (b) particle swarm optimization; (c) population-based incremental learning; (d) proposed.

optimal solution. However, the proposed technique uses local searching with the elitist population in noncomplete evolutionary process, so as to jump out of the local optimum and reach the global optimum. For German Credit dataset, the results suggest the proposed technique is inefficient. PSO outperforms other techniques in performance metrics although the proposed technique achieved the same 10 reduced features as the number of PSO.

Secondly, the convergent curves of ZhongAn, German, and LendingClub datasets are depicted, respectively, in Figures 6–8, which show accuracy difference in iterative searching. Taking iterative 100 times as the end, the proposed technique obviously outperforms others in searching efficiency. On the ZhongAn dataset, it can be seen the curve of the proposed technique is improved by 4.5% in the first 30 generations and by 0.3% between the 30th and the 100th generations, which indicate that the curve from 30th and 100th is a nearly straight line. However, the difference of GA and PSO remains the same between the 60th and 100th generations, and PBIL is between the 40th and 100th generations, as shown in Figure 6. Moreover, from Figure 7, it is observed that the curve of GA is similar to the proposed technique in convergent trend, and these two techniques have a higher convergence rate compared with PSO. Although they offer a slower convergence rate than PBIL, higher values are achieved. The reason behind this phenomenon may be that PBIL with poor stochastic searching results in not being able to jump out of the local optimum, as can be seen from Figure 9(c). The convergent curves on the LendingClub dataset in Figure 8 demonstrate the value of the proposed technique is entering the phase of saturation after the 60th generation. However, others are nearly identical from the 50th to 100th execution. The trends between the curves closely coincided with the presentation in

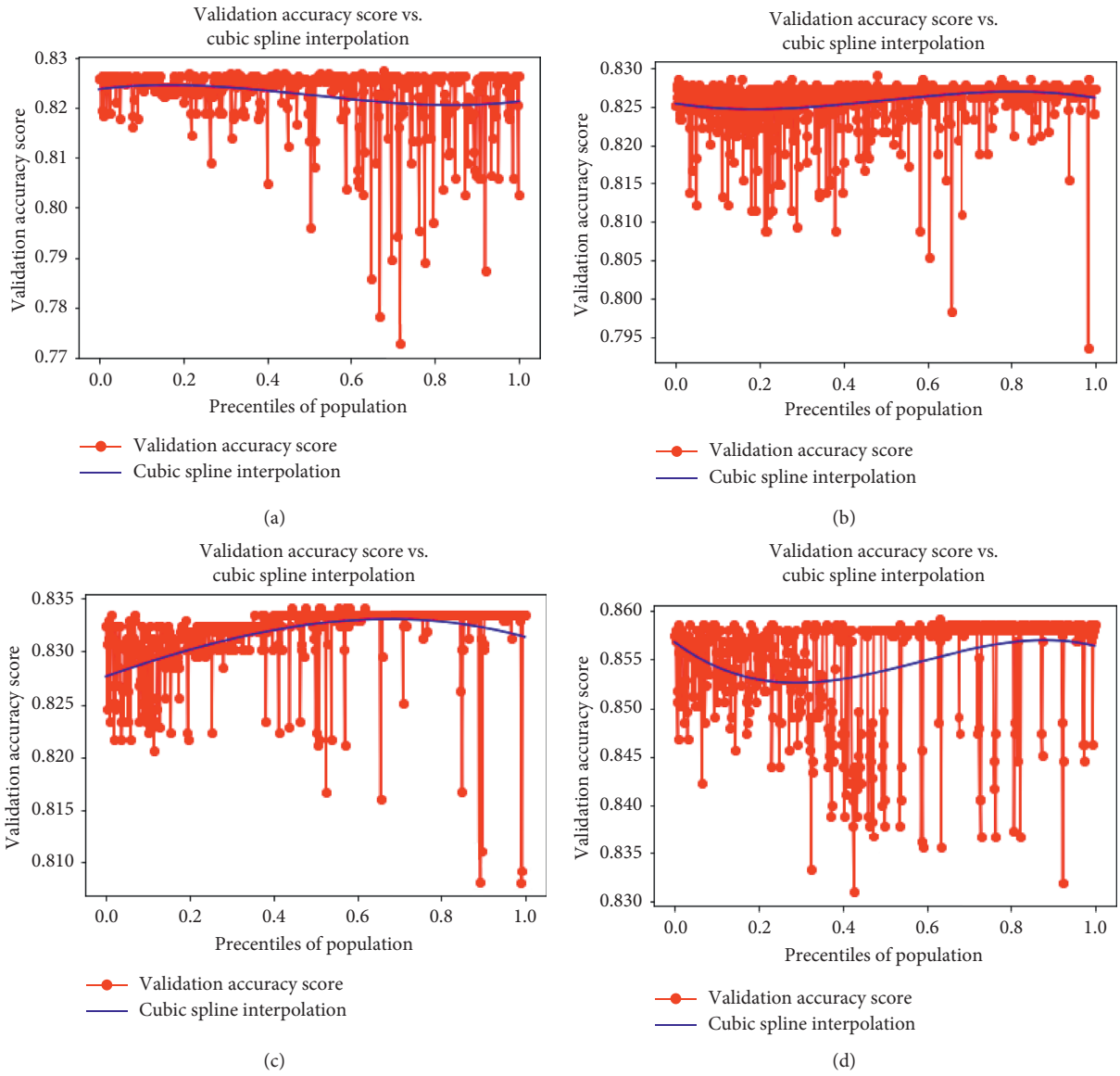

FIGURE 10: Comparative accuracy of all solutions on the ZhongAn dataset: (a) genetic algorithm; (b) particle swarm optimization; (c) population-based incremental learning; (d) proposed.

Figure 6. Besides, the interesting thing about other techniques is that the curves are almost the same, which is probably caused by the inner properties of this dataset. Based on these results, it reveals that the proposed technique can make a proper tradeoff be realized between diversity and accuracy in iterative searching.

Lastly, the proposed technique is compared with other techniques in terms of the performance generated by all reduced subsets, and the results are shown in Figures 9–11. Each value represents an accuracy obtained by each reduced feature subset. It can be seen that the measuring scales on the Y-axis are different with respect to the anticipated results, and most of values generated by the proposed technique are better than others. The figures also illustrate that the proposed technique can effectively produce more promising solutions and maintain population diversity to prevent premature convergence, except on the German dataset.

All in all, it is found that the ultimate feature number is far less than that of the original feature using the aforementioned metaheuristic techniques. The proposed technique outperforms other techniques on ZhongAn and LendingClub datasets, while it is not supported on the German dataset. It is probably that the effectiveness and reliability of the proposed technique is more significant when dealing with the large-scale data because it is required to use the metaheuristic subspace searching to optimize the ergodicity, but not necessary for small-scale data like German dataset with only 20 original features. The proposed

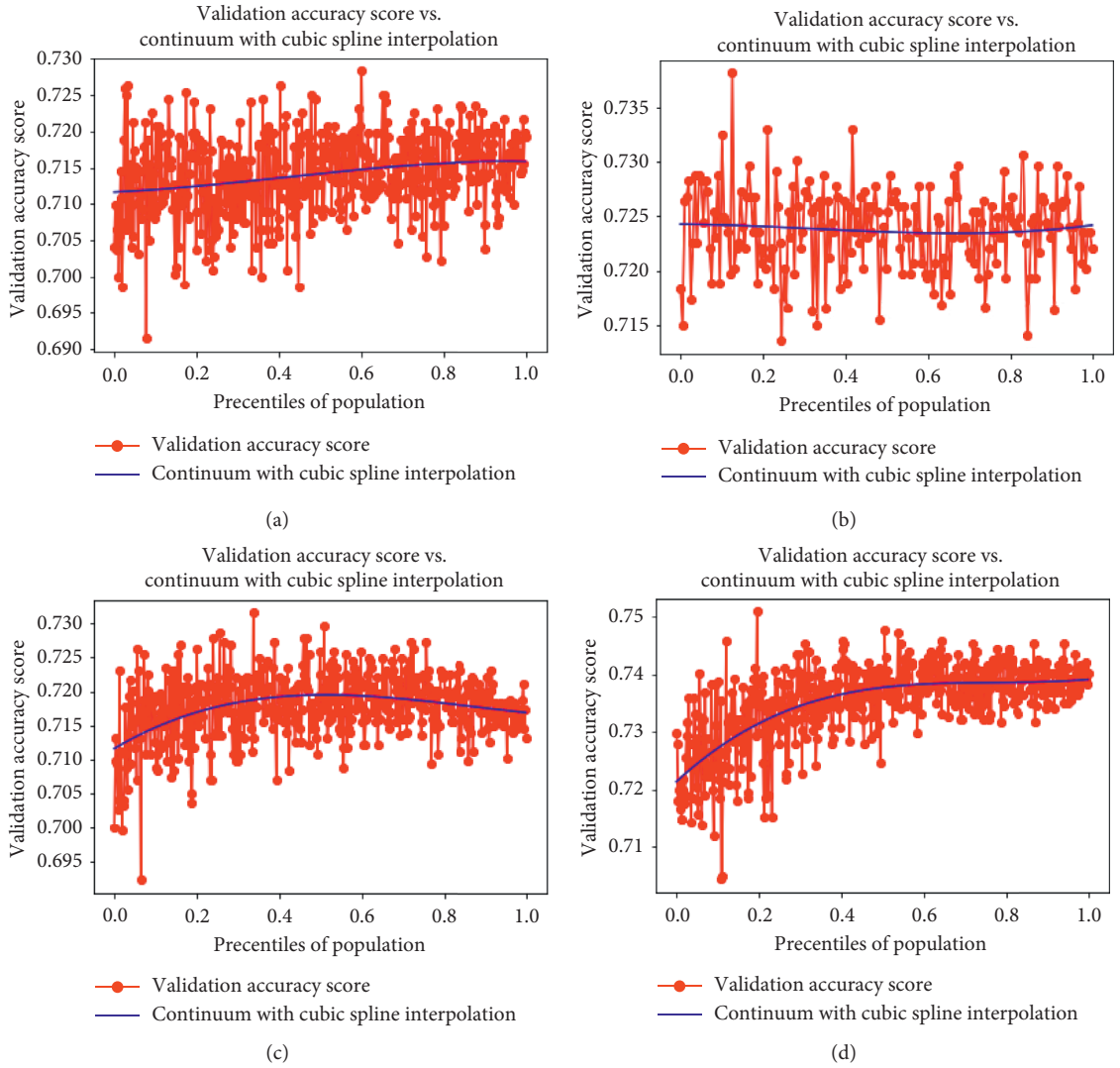

Figure 11: Comparative accuracy of all solutions on the LendingClub dataset: (a) genetic algorithm; (b) particle swarm optimization; (c) population-based incremental learning; (d) proposed.

VNS-EDA with the elitist population strategy is well suitable for large-scale credit data with high dimensionality.

6. Conclusion and Future Directions

As data availability is significantly enhanced by big data technology, as well as a significantly reduced profit margin in the credit loan industry, credit risk scoring model should be more accurate and effective. Feature selection for credit features is one of the challenging issues to deal with many irrelevant features. Previous research has shown that it still needs more sophisticated techniques to enhance the accuracy and generalization of risk classification. This article has proposed a hybrid VNS-EDA technique combining feature correlations, multiple neighborhood structures with elitist population strategy, and the probabilistic model. It includes three steps: the original feature preselection, the reduced feature refining, and the reduced feature learning. The original feature preselection can quickly extract relevant features to ameliorate the metaheuristic searching. The reduced feature refining can further generate predominant population with restricted neighborhoods to maintain population diversity and prevent premature convergence. The reduced feature learning builds an accurate classifier. The proposed technique has been tested on three credit datasets to evaluate its effectiveness. Comparisons have been drawn between the proposed and three state-of-the-art techniques by considering classification performance metrics. The mean improvement in terms of accuracy, F-measure, sensitivity, and specificity is found to be 2.8%, 2.1%, 1.9%, and 2.6% on the ZhongAn dataset, respectively. It is consistent with the results of improvement which are, respectively, 3.4%, 2.2%, 2.1%, and 2.7% obtained by the LendingClub dataset. Therefore, the proposed technique is more efficient in large-scale credit data.

Future directions of the proposed technique are discussed as follows: (1) the proposed technique can be also utilized to remove the irrelevant risk features when handling a large dimensionality number of samples for SMEs and individual financing. (2) The initial hyperparameters affect

the performance. The prediction results can be further optimized to improve model quality by an efficient tuning of the hyperparameters. (3) This feature selection technique is based on subspace searching. Therefore, in the near future, we can make use of advanced space reduction techniques to enhance the performance and reduce the execution time such as the combination of generative adversarial networks (GANs) with metaheuristic techniques. The population evolution can be implemented by the operator competition or cooperation through GAN with the dynamic expanding neighborhoods and be realized as a learning process with their own evolution, which is expected to obtain an approximate optimum as quickly as possible.

Data Availability

The private experimental data used to support the findings of this study have not been made available because these data belong to a private financial company, while the public experimental data are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (no. 71721001) and in part by the Natural Science Research Team of Guangdong Province of China (no. 2014A030312003).

References

[1] M. Šušteršič, D. Mramor, and J. Zupan, "Consumer credit scoring models with limited data," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4736–4744, 2009.

[2] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.

[3] F. Louzada, A. Ara, and G. B. Fernandes, "Classification methods applied to credit scoring: systematic review and overall comparison," *Surveys in Operations Research and Management Science*, vol. 21, no. 2, pp. 117–134, 2016.

[4] S. Bhatia, P. Sharma, R. Burman, S. Hazari, and R. Hande, "Credit scoring using machine learning techniques," *International Journal of Computer Applications*, vol. 161, no. 11, pp. 1–4, 2017.

[5] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature Extraction*, pp. 1–25, Springer, Berlin, Germany, 2006.

[6] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review," in *Data Classification: Algorithms and Applications*, pp. 37–64, CRC Press, Boca Raton, FL, USA, 2014.

[7] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proceedings of the IEEE Science and Information Conference*, pp. 372–378, London, UK, August 2014.

[8] R. Sandy, "Survey on dimension reduction techniques," *Journal of Computer Applications*, vol. 8, no. 5, pp. 704–710, 2006.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[10] A. I. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper," in *Proceedings of the FLAIRS Conference*, pp. 235–239, Orlando, FL, USA, May 1999.

[11] P. Somol, B. Baesens, P. Pudil, and J. Vanthienen, "Filter-versus wrapper-based feature selection for credit scoring," *International Journal of Intelligent Systems*, vol. 20, no. 10, pp. 985–999, 2005.

[12] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Feature subset selection problem using wrapper approach in supervised learning," *International Journal of Computer Applications*, vol. 1, no. 7, pp. 13–17, 2010.

[13] P. Danenas, G. Garsva, and S. Gudas, "Credit risk evaluation model development using support vector based classifiers," *Procedia Computer Science*, vol. 4, pp. 1699–1707, 2011.

[14] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[15] Y. Liu and M. Schumann, "Data mining feature selection for credit scoring models," *Journal of the Operational Research Society*, vol. 56, no. 9, pp. 1099–1108, 2005.

[16] H. Yu, X. Huang, X. Hu, and H. Cai, "A comparative study on data mining algorithms for individual credit risk evaluation," in *Proceedings of the IEEE International Conference on Management of e-Commerce and e-Government*, pp. 35–38, Chengdu, China, October 2010.

[17] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services*, vol. 27, pp. 11–23, 2015.

[18] C. Dhaenens and L. Jourdan, *Metaheuristics for Big Data*, Wiley, Hoboken, NJ, USA, 2016.

[19] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[20] M. A. Hall, *Correlation-based feature subset selection for machine learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 1999.

[21] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437–1447, 2012.

[22] D. Boughaci and A. A. K. Alkhawaldeh, "A new variable selection method applied to credit scoring," *Algorithmic Finance*, vol. 7, no. 1-2, pp. 43–52, 2018.

[23] S. Piramuthu, "On preprocessing data for financial credit risk evaluation," *Expert Systems with Applications*, vol. 30, no. 3, pp. 489–497, 2006.

[24] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. C-26, no. 9, pp. 917–922, 1977.

[25] L. Wang, H. Ni, R. Yang, V. Pappu, M. B. Fenn, and P. M. Pardalos, "Feature selection based on meta-heuristics for biomedicine," *Optimization Methods and Software*, vol. 29, no. 4, pp. 703–719, 2014.

[26] R. Diao and Q. Shen, "Nature inspired feature selection meta-heuristics," *Artificial Intelligence Review*, vol. 44, no. 3, pp. 311–340, 2015.

[27] I. S. Oh, J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1424–1437, 2004.

[28] Y. Marinakis and M. Magdalene, "A hybridized particle swarm optimization with expanding neighborhood topology for the feature selection problem," in *Hybrid Metaheuristics*, pp. 37–51, Springer, Berlin, Germany, 2013.

[29] H. Banka and S. Dara, "A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation," *Pattern Recognition Letters*, vol. 52, pp. 94–100, 2015.

[30] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[31] U. Singh and S. N. Singh, "Optimal feature selection via NSGA-II for power quality disturbances classification," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 2994–3002, 2018.

[32] H. S. Pannu, D. Singh, and A. K. Malhi, "Multi-objective particle swarm optimization-based adaptive neuro-fuzzy inference system for benzene monitoring," *Neural Computing and Applications*, vol. 31, no. 7, pp. 2195–2205, 2017.

[33] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Systems with Applications*, vol. 41, no. 4, pp. 2052–2064, 2014.

[34] J. Zhou and T. Bai, "Credit risk assessment using rough set theory and GA-based SVM," in *Proceedings of the 3rd International Conference on Grid and Pervasive Computing-Workshops*, pp. 320–325, Kunming, China, May 2008.

[35] Y. Marinakis, M. Marinaki, M. Doumpos, N. Matsatsinis, and C. Zopounidis, "Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment," *Journal of Global Optimization*, vol. 42, no. 2, pp. 279–293, 2008.

[36] C.-M. Wang and Y.-F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: a case study of credit approval data," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5900–5908, 2009.

[37] J. Wang, A.-R. Hedar, S. Wang, and J. Ma, "Rough set and scatter search metaheuristic based feature selection for credit scoring," *Expert Systems with Applications*, vol. 39, no. 6, pp. 6123–6128, 2012.

[38] H. Altinbas and G. C. Akkaya, "Improving the performance of statistical learning methods with a combined meta-heuristic for consumer credit risk assessment," *Risk Management*, vol. 19, no. 4, pp. 1–26, 2017.

[39] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 111–128, 2011.

[40] U. Aickelin and J. Li, "An estimation of distribution algorithm for nurse scheduling," *Annals of Operations Research*, vol. 155, no. 1, pp. 289–309, 2007.

[41] R. Shah and P. Reed, "Comparative analysis of multiobjective evolutionary algorithms for random and correlated instances of multiobjective d-dimensional knapsack problems," *European Journal of Operational Research*, vol. 211, no. 3, pp. 466–479, 2011.

[42] R. ArmaPñanzas, I. Inza, R. Santana et al., "A review of estimation of distribution algorithms in bioinformatics," *BioData Mining*, vol. 1, pp. 1–6, 2008.

[43] P. Hansen and N. Mladenoviú, "Variable neighborhood search: principles and applications," *European Journal of Operational Research*, vol. 130, no. 3, pp. 449–467, 2008.

[44] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pp. 90–105, Antwerp, Belgium, 2008.

[45] J. Schwarz and J. Ocenasek, "A problem knowledge-based evolutionary algorithm KBOA for hypergraph bisectioning," in *Proceedings of the 4th Joint Conference on Knowledge-Based Software Engineering*, pp. 51–58, Brno, Czech Republic, 2000.

[46] M. Pelikan and D. E. Goldberg, "Hierarchical BOA solves Ising spin glasses and MAXSAT," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1275–1286, Chicago, IL, USA, July 2003.

[47] K. Bache and M. Lichman, *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA, USA, 2013, http://archive.ics.uci.edu/ml.

[48] F. Fortin, F. Rainville, M. Gardner, M. Parizeau, and C. Gagné, "DEAP: evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, 2012.

[49] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, 2007.

[50] J. Hamon, "Combinatorial optimization for variable selection in high dimensional regression: application in animal genetic," Thesis, University of Science and Technology, Lille, France, 2013.

[51] S. Baluja, "Population-based incremental learning. a method for integrating genetic search based function optimization and competitive learning," Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1999.