Complex & Intelligent Systems (2023) 9:3559–3579
https://doi.org/10.1007/s40747-021-00614-4

ORIGINAL ARTICLE

Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk

Mohammad Zoynul Abedin[1,2] · Chi Guotai[3] · Petr Hajek[4] · Tong Zhang[5]

Received: 25 March 2021 / Accepted: 3 December 2021 / Published online: 4 January 2022

Abstract

In small business credit risk assessment, the default and nondefault classes are highly imbalanced. To overcome this problem, this study proposes an extended ensemble approach rooted in the weighted synthetic minority oversampling technique (WSMOTE), which is called WSMOTE-ensemble. The proposed ensemble classifier hybridizes WSMOTE and Bagging with sampling composite mixtures to guarantee the robustness and variability of the generated synthetic instances and, thus, minimize the small business class-skewed constraints linked to default and nondefault instances. The original small business dataset used in this study was taken from 3111 records from a Chinese commercial bank. By implementing a thorough experimental study of extensively skewed data-modeling scenarios, a multilevel experimental setting was established for a rare event domain. Based on the proper evaluation measures, this study proposes that the random forest classifier used in the WSMOTE-ensemble model provides a good trade-off between the performance on default class and that of nondefault class. The ensemble solution improved the accuracy of the minority class by 15.16% in comparison with its competitors. This study also shows that sampling methods outperform nonsampling algorithms. With these contributions, this study fills a noteworthy knowledge gap and adds several unique insights regarding the prediction of small business credit risk.

Keywords Small business · Credit risk · Imbalanced data · Oversampling · Weighted SMOTE · Ensemble learning

✉ Mohammad Zoynul Abedin
abedinmz@yahoo.com

Chi Guotai
chigt@dlut.edu.cn

Petr Hajek
petr.hajek@upce.cz

Tong Zhang
zhangtong19900227@126.com

1 Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Tees Valley, Middlesbrough TS1 3BX, UK

2 Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur 5200, Bangladesh

3 Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China

4 Institute of System Engineering and Informatics, Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, Pardubice 532 10, Czech Republic

5 Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China

Introduction

Many researchers to date have aspired to elaborate classifiers for large corporate firms or firms that were already listed. However, in most economies, small enterprises are the principal sources of financial development and stability, stoking the engine of economic progress and growth. From a credit-approval data-modeling standpoint, small enterprises have some specific features that are unlike their larger counterparts [17,18]. Ciampi [17] stated that small enterprises are economically riskier and have a lower asset correlation with each other than do corporations. These findings underline the fact that the credit risk modeling of small enterprises should be done in a different way than that for listed companies and large firms. Recently, a credit risk appraisal has shown the economic importance of these enterprises, which is the result of the size of their personal unsecured borrowings and the rapidly growing probability of their default risk [42]. The United States and European subprime mortgage disasters are two clear examples of default scenarios. An Organization of Economic Co-operation and Development report states that the amount of default credit is increasing at

a rate of approximately 2% in China [47]. How to evaluate a lender's credit risk efficiently and accurately has been a decisive issue for all relevant practitioners, researchers, and stakeholders. Consequently, formulating a consistent model for credit risk has become a significant focus for small enterprises, so that they can ensure sustainable profits and lessen corresponding losses [10,20]. This study focuses on the credit risk assessment of small business loans in this environment, in order to advance the performance of an assessment model and improve its classification accuracy.

A feasible classification of risky and nonrisky small enterprises has been widely perceived as the primary focus of credit risk assessment modeling because an improvement of even a fraction of a percent could lead to significant future savings and profits [58]. A large body of literature has evaluated techniques for increasing the accuracy of credit risk assessment, combining statistical and artificial intelligence models. A statistical approach was used in the works of Pindado and Rodrigues [50] on logit analysis, and Duarte et al. [21] on probit analysis. A large number of studies have used artificial intelligence algorithms, such as the k-nearest neighbors (k-NN) [23], neural networks [17], support vector machines (SVMs) [3], and gradient decision tree approach [13] (see also [28] for an exhaustive review of artificial intelligence methods). However, the many experimental studies available [2,6,9,26,64] had not been able to assert the dominance of one algorithm over other competing classifiers irrespective of data traits. For example, missing values, problems with noise, outlier information with redundant and irrelevant features, and skewed class allocations had drastically affected the results of most classification algorithms. In the real world, most of the credit samples are creditworthy (majority/negative class) and the remainder are noncreditworthy or default cases (minority/positive class); that is to say, there is a class imbalance distribution. In small business loan performance data modeling, the price of misclassifying noncreditworthy applicants as creditworthy is significantly higher than the price of misclassifying creditworthy applicants as noncreditworthy [4,27]. Besides, class-skewed data distribution brings with it certain conflicts for assembling a classification algorithm; in contrast, these challenges may appear in equilibrium scenarios. Recently, modelers have optimized the forecasting performance in skewed class domains. One commonly employed sampling methodology is data-level solutions, which modify the class allocation in a given example set. Oversampling, undersampling, hybrid sampling, and ensemble learning are the major types of preprocessing approaches in credit-scoring and bankruptcy-prediction domains [62].

In a pioneering study [41] on the imbalance learning of credit risk modeling claims, it was shown that oversampling techniques performed better than any form of class imbalance learning, and, therefore, the current investigation is focused on oversampling methods. The most well known oversampling procedure, which is extensively employed in many domains, is the synthetic minority oversampling technique (SMOTE) [15]. Refinements of the SMOTE algorithm have been attempted and reported on in many studies [26,51,55,63] from the time the algorithm was launched. Using similar background material as a basis, the current study has applied a modification known as the weighted SMOTE (WSMOTE) [52]. This oversampling approach treats generated data more effectively because a specific weight is assigned to each minority data sample based on its Euclidean distance to the remaining minority data samples. This, in turn, leads to the production of more compact synthetic data in WSMOTE than in SMOTE. In addition, the current study proposes a novel ensemble approach rooted in the WSMOTE algorithm, the WSMOTE-ensemble for skewed loan performance data modeling. The proposed ensemble classifier hybridizes WSMOTE and Bagging with sampling composite mixtures (SCMs) to minimize the class-skewed constraints linked to positive and negative small business instances. It increases the multiplicity of executed algorithms as different SCMs are applied to form diverse training sets [1]. First, for a given dataset, Bagging is applied to produce nondefault majority instances. Second, the WSMOTE learner is trained to generate synthetic minority instances that are combined with the original minority data. Later, the Bags from the Bagging and the new synthetic datasets from WSMOTE are merged to obtain balanced datasets. In the study by Sun et al. [61], the C4.5 classifier, a decision tree algorithm, is trained to determine the Bag's accuracy, and the best Bags are selected for the intended experiments based on their maximum predictive power. The proposed WSMOTE-ensemble method contrasts with other existing methods, however, especially with the approach of Sun et al. [61], on several grounds, because it is more robust and prone to overfitting. In addition to the above algorithms, the current study also addresses the modification of the Chan and Stolfo [14] classifier. The original model is a hybrid sampling strategy generating random samples from the desired distribution for ensemble learning. In place of random undersampling, this study employs the WSMOTE learner to produce synthetic positive data. Accordingly, the trained algorithm is called the modified Chan undersampling (MChanUS) approach. In addition to oversampling and undersampling techniques, this study presents two hybrid data-level solutions combining (1) SMOTE with random undersampling (RUSSMOTE) and (2) undersampling and oversampling (USOS) to show the effect of different forms of class imbalance learning.

From a methodological point of view, fusion strategy (ensemble classifier approaches) is an active study area in imbalanced learning and a rare event prediction domain. It learns the new data patterns by integrating stand-alone proposals from a set of elementary algorithms. A consider-

able number of experimental studies on ensemble approaches have become available over the last decades. Notably, random forest (RF) involves the collection of unpruned forecasting, classification, or regression tree algorithms, learned on bootstrap examples of in-sample instances applying randomly chosen variables in the process of tree creation. Many empirical studies and theoretical backgrounds exhibit the strengths of RF learners [54]. However, there is a paucity of learning about small business loan performance data with RF, a rare event domain; it is a relatively new approach that must be confirmed. Inspired by the above information, the current study follows up on multiple RF-based experimental settings to assess the recitals of the different algorithms considered. In addition to the RF-based multilevel settings, this study offers data-level strategies with a C4.5 decision tree, *k*-NN, and SVM, which are the most widely used baseline classifiers in this domain. From the credit-scoring realm, the Bagging- and Boosting-based ensembles can generate better algorithmic modifications for class imbalance learning [11]. In line with these algorithmic backgrounds, the present study also hybridizes Bagging, Boosting, random committee (RC), rotation forest (RTF), and logit boost (LB) as multiple-classifier systems.

An original database used in this study comes from the 3111 records of small enterprises from a Chinese commercial bank. This paper reports on a thorough experimental study using the skewed small business credit risk data to demonstrate that the WSMOTE-ensemble-RF fusion model significantly outperforms comparative methods. Hence, the proposed method is established as a more scalable learning method for skewed credit risk data.

The remaining parts of this study are organized as follows. A literature review and issues of class imbalance problems are presented in "Related literature" and "Class imbalance problem", respectively. "WSMOTE-based methods for class imbalance problem" describes the proposed algorithms. The experimental design is presented in "Experimental design". Empirical results are highlighted in "Empirical results" and "Comparative analysis". "Conclusion" concludes the study and shows future roadmaps for the field.

Related literature

Small business credit risk modeling

Small business loan modeling has three main categories. The first is derived from information economics or capital market theory [6,56], which has a strong theoretical background, but it is not able to generate automatic credit risk predictions.

The second category is statistical classifiers. Edmister [22] was the pioneer modeler who experimented with multivariate discriminant analysis by applying 18 variables to 562 small enterprises. Following this seminal work, many researchers and practitioners moved forward to design numerous statistical models. Relevant examples include Altman and Sabato [7], Behr and Güttler [10], Mayr et al. [42], Duarte et al. [21], Arcuri and Levratto [9], Inekwe [34], Ciampi and Gordini [19], Sohn and Jeon [58], Hasumi and Hirata [31], Lin et al. [38], and Keasey et al. [35]. Most of the cited studies compare and contrast small business loan approval data by applying customary statistical algorithms, but a few take a different approach. Sohn and Jeon [58], for instance, stated that small enterprises should concentrate on setting up their predictive methodology to avoid high default rates. Moreover, credit ratings of small businesses do not necessarily reflect overall credit risk due to the mismatch between credit ratings and loss-given-default [57], and small businesses are reportedly vulnerable to local economic conditions [43]. By adopting innovative forecasting techniques, small enterprises could maintain their economic progress even in periods of credit crunch. Lin et al. [38] developed two logit regression classifiers for 429 small businesses in the United Kingdom in 2009. The researchers explored the classifiers' predictive accuracy, applying accounting-based approaches, and concluded that small business loan performance modeling has different consequences for the composition of trained algorithms. These classifiers, however, suffered from the same identical problems that plagued statistical learners (e.g., linearity, normality, and independence among predictors) and are ubiquitous in small business datasets. Applications of traditional statistical classifiers underrate the default probability of the positive class because the logit link function is symmetric and the minor instances are rare events [12]. In reality, regarding the skewed data modeling scenarios, the probability of a rare event is biased towards the less important negative class.

Lastly, a few small business studies concentrated on building ensemble classifiers for credit risk modeling to overcome the flaw of a single algorithm and show that ensemble classifiers are better executors than their baseline equivalents [5]. Recently, Zhu et al. [67] applied random subspace-based MultiBoosting to small Chinese enterprises and claimed that their proposed ensemble classifiers showed an enhanced predictive accuracy for a small-sized instance. The credit risk data evaluation derived from an ensemble classifier can fully utilize the dissimilar knowledge learned by different baseline algorithms, and it typically has a better performance than single-classifier approaches.

Skewed credit risk management data modeling approaches

A number of studies have addressed the skewed class scenarios of credit risk modeling. Although a range of artificial intelligence classifiers were trained for predicting the

small business credit risk, predicting skewed class instances notably confronted the customary classifiers [33]. Louzada et al. [39] investigated the performance of logistic regression (LR) and naïve logistic regression models over skewed class Brazilian credit data, and asserted that the predictive ability of trained algorithms was significantly lessened when positive high-risk instances and negative low-risk instances were extremely skewed. Antunes et al. [8] proposed Gaussian processes for real-world skewed loan approval instances and demonstrated that the proposed investigation could efficiently improve the accuracy of the trained algorithm over SVM and LR. Kim et al. [36] executed a geometric mean-based Boosting algorithm to resolve skewed class problems in bankruptcy prediction and verified that their proposed algorithm had the advantages of a high predictive power and a robust learning capability in balanced and skewed data distributions. For imbalanced credit data modeling, Sun et al. [60] combined a hybrid feature selection with SVM and multiple discriminant analysis, and they claimed that the hybrid feature selection technique was an essential tool for predicting the credit customer status. More recently, He et al. [32] proposed an extended Balance Cascade approach for six different real-world skewed credit databases and thought their methodology to be more robust in credit scoring. Evolutionary undersampling was used to favor diversity in the selected instances [48]. However, potentially valuable instances may be discarded in the undersampling approaches.

The above studies focused on skewed learning in business domains other than small enterprises. Only a few studies have concerns parallel to the ones in this study, such as the study by Gicic and Subasi [26], which applied SMOTE with RTF to a real-world microcredit dataset and determined that ensemble classifiers provided improved results. In a comparable way, Sun et al. [61] experimented with a novel decision tree ensemble algorithm for the skewed credit risk modeling of enterprises and declared that their proposed methodology could not only deal with the skewed class problem but could also augment the multiplicity of stand-alone classifiers. Therefore, the proposed methodology is intended for a rare event domain for which more empirical research is needed.

Justification of this study

Based on the surveyed studies, the use of classifier ensembles is the growing trend in credit risk prediction, and modeling skewed class small business datasets is an unfocused domain. In a study that took a different approach than the ones cited above, Haixiang et al. [28] reviewed the fundamental core learners applied to classifier ensembles and skewed data problems. They asserted that RF, neural networks, SVM, and decision trees are some widely used cardinal algorithms in the literature. In line with these findings, Sun et al. [59] condensed the obstacles of base classifiers when learning from skewed data and stated that there are dozens of learners in the existing studies that have strong points and flaws. For instance, SVM is likely to be complicated in learning for large-scale datasets and sensitive to missing values, for which, on the contrary, RF might be a feasible classifier. Following this stream of research, Rio et al. [54] claimed that the RF classifier was an eminent decision tree ensemble admired for its robustness and outstanding recitals. In addition, RF classifiers enable efficient processing of high-dimensional credit risk data without the need to perform feature selection while providing the user with feature importance [32]. Predictions of credit risk in RF are produced as an ensemble estimate from a number of simple models (decision trees) by using bootstrap samples (Bagging). Using different randomly selected training data improves the stability of the classifier and reduces its overfitting risk. Random selection of classifier variables is another beneficial quality of RF, allowing it to handle a large number of credit risk features. As a result of this embedded feature selection procedure, the number of features required for credit risk assessment can be substantially reduced, thus improving the efficiency of the RF classifier. These advantages have made RF a benchmark method in both credit scoring [37] and corporate credit rating [30]. Moreover, an ensemble-based RF learning methodology is worthwhile when one is confronted with skewed data instances [32]. Therefore, the application of RF, along with other competing ensembles, should not bias the experimental outcomes regarding positive minority instances in small business credit risk assessment. Inspired by the above findings, the current study follows up on the multiple RF-based experimental settings to assess the aspects of the different algorithms considered.

Class imbalance problem

In recent periods, many solutions have been proposed to deal with class imbalance problems, both for standard learning classifiers and ensemble algorithms. They can be categorized into three major groups:

The data-level solution is one in which the training instances are adapted to generate a more or less rebalanced class allocation that permits algorithms to perform in a way that is comparable to the standard classification. This strategy typically entails two methodologies: oversampling and undersampling. Oversampling techniques reduce class inequity by generating new minority class instances. Contrary to oversampling techniques, undersampling procedures reshape class inequity by decreasing the number of positive class instances [29]. Besides oversampling and undersampling techniques, this study presents the hybrid data-level solutions RUSSMOTE, MChanUS, and USOS to show the effect of different forms of imbalance learning.

Algorithmic modification attempts to modify the learning of forecasting classifiers concerning the negative class. Cost-sensitive learning is the most popular example of this type. A cost-sensitive classifier constructs an algorithm that usually provides a high cost for the negative class and a low cost for the positive class. Therefore, cost-sensitive learning minimizes the total cost of misclassification by changing the different types of cost ratios. The representative type of weighted learning, for instance, C4.5, k-NN, or SVM, can be applied to the class skewed problems.

Ensemble solution attempts to perk up the performance of base algorithms by introducing numerous algorithms in combination to attain new, improved performing classifiers, which are more adjusted to skewed class issues. In this study, for better classification and generalization, ensemble learning creates hybridization with data-level approaches. In particular, this study proposes a novel ensemble called the WSMOTE-ensemble algorithm. Data-level hybrid strategies based on RF, Bagging, and Boosting are reportedly able to better deal with class imbalances in loan approval datasets compared with data-level and ensemble solutions applied separately [11].

WSMOTE-based methods for class imbalance problem

SMOTE approach

SMOTE is probably the most famous algorithm for counteracting imbalanced dataset modeling [15]. It produces new instances of the negative class by operating in the "feature space" rather than the "data space". SMOTE is an oversampling technique in which each negative class instance creates $M\%$ of artificial instances comparable to the majority class instances. This augmentation in the minority instances enhances the decision accuracy of the trained algorithms.

Weighted SMOTE approach

The WSMOTE algorithm [52] is an oversampling methodology that allocates weights for generating new artificial data and employs SMOTE for a specific positive data point. Each minority instance produces an equal number of synthetic instances, and the outcome is an amendment to the SMOTE algorithm. The WSMOTE technique utilizes the Euclidean distance of each positive instance with all the other minority instances to generate a weighted scalar. This weighted scalar, together with all of the synthetic instances, creates the SMOTE generation scalar; the algorithmic pseudocode provides detailed procedures in Algorithm 1.

The minority/positive training set is taken with T instances and m variables. The Euclidean distance (ED) of each of the T minority instances is computed with regard to all positive instances. The sum of each of these distances for each i-th and j-th positive instance ($i \neq j$) provides ED_i. The distances for all of the minority instances are computed and accumulated in a matrix. Then, the ED matrix is normalized by applying the highest value, ED_max, and the lowest value, ED_min, and is named the normalized ED matrix (NED). After that, the NED scalar is customized to a revised NED matrix (RNED). The RNED matrix says that the smaller the ED of a positive instance, the higher the amount of data samples it obtains to produce the artificial sample from the total amount of artificial instances ($N\%$). The RNED matrix is computed by deducting NED_i for each minority instance from the sum of all of the NEDs. In the last stage, the weighted matrix is computed by generating each minority data portion with regard to the total sum of instances generated in the RNED matrix. Finally, this weighted matrix is utilized to obtain the SMOTE generation matrix.

Algorithm 1 Weighted SMOTE

Input: Dataset={($\mathbf{x}_i, y_i$), $\mathbf{x}_i$=($x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,m}$), i=1,2,...,T}, where T denotes the number of minority data, m denotes the number of features, and y_i denotes the class label; the amount of weighted SMOTE $N\%$; the number of nearest neighbors k.

Output: ($N\times T$)/100 synthetic minority class samples { ($\mathbf{x}_i^s$=($x_{i,1}^s$, $x_{i,2}^s, \ldots, x_{i,j}^s, \ldots, x_{i,m}^s$), i=1,2,...,($N\times T$)/100} .

Procedure:

{

Step 1: // Calculate the Euclidean distance of each of the T minority data samples

$\mathrm{ED}_i(x_i, x_l) = \sqrt{\sum_{j=1}^{m}(x_{i,j} - x_{l,j})^2}$, where $l \neq i$.

// For all the minority data, the ED are calculated and stored in the column matrix ED=[$\mathrm{ED}_1, \mathrm{ED}_2, \ldots, \mathrm{ED}_T$]

Step 2: // Normalize the ED matrix

$\mathrm{NED}_i = \frac{ED_i - ED_{min}}{ED_{max} - ED_{min}}$

Step 3: // Modify NED to a remodeled normalized matrix REND

$[\mathrm{REND}]_{T\times 1}$=sum(NED)-$[\mathrm{NED}]_{T\times 1}$

Step 4: // Calculate weight matrix W for each minority of T samples

$[W]_{T\times 1}$=$[\mathrm{REND}]_{T\times 1}$/sum(REND)

Step 5: // Calculate the SMOTE Generation Matrix G

$[G]_{T\times 1}$=$N\% \times T \times [W]_{T\times 1}$, G= [$G_1, G_2, \ldots, G_T$].

Step 6: // Generate the synthetic samples.

For i =1 to T

Compute k nearest neighbors for i, and save the indices in the $n \times n$ array.

While $G_i \neq 0$

Choose a random number between 1 and k, call it k_r. // This step chooses one of the k nearest neighbors of i.

For j= 1 to m

$x^s_{\mathrm{newindex},j} = x_{i,j} + \mathrm{rand}(0,1) \times (x_k r_{,j} - x_{i,j})$

End For

newindex++

$G_i = G_i - 1$

End While

End For

Return synthetic samples { ($\mathbf{x}_i^s$=($x_{i,1}^s, x_{i,2}^s, \ldots, x_{i,j}^s, \ldots, x_{i,m}^s$), i=1,2,...,($N\times T$)/100 } .

}

MChanUS approach

The current study applies a modification of the Chan and Stolfo [14] approach to classify the skewed small business data. The original algorithm is a hybrid sampling strategy combining random oversampling and random undersampling. In place of random sampling, this study trains the WSMOTE learner to produce synthetic positive data. Accordingly, the trained algorithm called MChanUS and the algorithmic pseudocode are given in Algorithm 2. First, the negative samples are divided into nonoverlapping subsets. Likewise, the WSMOTE algorithm is applied to generate artificial positive instances and combines them with the original minority instances, forming the new positive set of instances. Finally, the newly created minority dataset hybridizes with each of the subsets belonging to the negative class. This means that the positive minority dataset reiterates across the negative sets of instances and generates balanced datasets. The flowchart of the MChanUS algorithm illustrating the above procedures for the used small business dataset is found in Fig. 1.

Algorithm 2 MChanUS

Input: Dataset={$(\mathbf{x}_i, y_i)$, $\mathbf{x}_i$=$(x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,m})$, i=1,2,...,T}.
Output: Balanced dataset.
Procedure:
{
Step 1: // Identify data classes, i.e., positive sample and negative sample
sample_size(SS, $[Pos, Neg]_{count}$)
Step 2: // Use the WSMOTE algorithm to produce synthetic positive instances and combine them with the original minority instances
gen_synthetic_positive_samples $[SPos]_{count}$
for each $[Pos, SPos]_{count}$
get_new_positive_sample $[NPos]_{count}$
Step 3: // Create seven equal subsets of negative instances applying the random undersampling principle
make_majority_instances_set $[SNeg_1, SNeg_2, \ldots SNeg_7]_{count}$
Step 4: // Combine new positive instances with negative instances in the subsets
merge_two_newly_produced_data_sets $[NPos+SNeg_1, \ldots .. NPos+SNeg_7]_{count}$
Return the balanced dataset.
}

WSMOTE-ensemble approach

A classifier ensemble can be trained on the in-sample instances with dissimilar percentages of oversampling where the segregated SCMs are applied in the positive class dataset. It guarantees the multiplicity of baseline learners and evades overtraining to a certain level because the number of minority in-sample instances for training each baseline classifier is dissimilar. Let us assume that D learners ($d = 1, 2, ..., D$) are employed, and the segregated SCMs for training the learners fit as $\mathrm{SCM}_d = [100\times(d/D)]$, where SCM_d refers to the oversampling rate for the d-th learner. For instance, SCM_d = 20%, 40%, 60%, 80%, and 100% for D = 5.

Therefore, the number of minority class instances after SCMs is determined as $SN_{\mathrm{Min}} = S_{\mathrm{Min}} + \mathrm{round}\,(M\times\mathrm{SCM}_d)$, where M refers to the difference between instance numbers of minority (S_{Min}) and majority (S_{Maj}) instances in the initial skewed dataset. Besides, the function round (.) denotes rounding a numerical figure downward or upward based on its decimals. That is to say, when the minority sample is S_{Min} = 158 and the majority sample is S_{Maj} = 367, then their disparity is M = 209. If the oversampling composite mixtures rate SCM_d is 30%, the number of minority instances after oversampling would be SN_{Min} = 158 + round (209 × 30%) = 221. Following these procedures, fusion learners can be assembled utilizing the segregated SCM principles.

The structure of the WSMOTE-ensemble model is shown in Fig. 2. It is a blended outcome of the SCMs, WSMOTE, Bagging, and C4.5 decision tree algorithms. The WSMOTE-SCM and Bagging-SCM are trained on the minority and majority classes, respectively, to generate balanced Bags. According to the study by Sun et al. [61], the C4.5 classifier is trained to obtain the Bag's accuracy, and the best training Bags are selected for the intended experiments based on the maximum predictive power. Thus, the WSMOTE-ensemble classifier incorporates the benefits of WSMOTE and SCM to deal with the skewed class issues by generating new positive instances using the WSMOTE procedure with SCMs in each step. Unlike the oversampling mixtures from the WSMOTE algorithm, the proposed algorithm constructs several in-sample sets belonging to different numbers of minority instances. To be precise, different levels of new weighted minority instances are created to deflate the extreme credit risk of the respective class. This integration assists in enlarging the multiplicity of baseline learners and also minimizes overtraining problems. The SCM technique used in this study is an oversampling method that is applied to both minority instances (using WSMOTE) and majority instances (using Bagging with random sampling) while using differentiated sampling rates. Balanced Bags are thus obtained with different numbers of instances. Overall, the advantages of differentiated sampling rates in SCMs are as follows [61]:

- the diversity of base classifiers is increased by sampling different numbers of training instances.
- the risk of overtraining is reduced by fully exploiting the WSMOTE capacity to produce minority instances.
- when combined with Bagging, the stability of the small business credit risk classification model is improved compared with single classifier-based models.

The WSMOTE-ensemble learner also puts together the enhancements of Bagging and SCM for reducing the major-

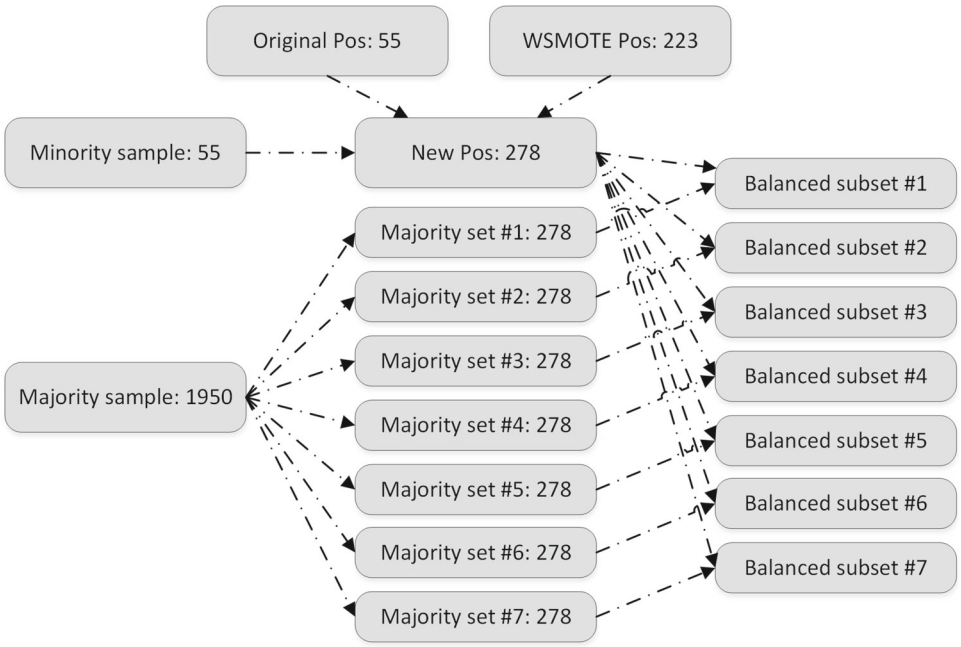

Fig. 1 Flowchart of MChanUS algorithm

ity instances with SCMs. In each training step of the baseline algorithm, the instances of the majority set from Bagging are identical to those of the minority instance set generated using WSMOTE. Accordingly, this study compiles the class-balanced Bags by mixing up WSMOTE with Bagging. Then, the baseline C4.5 classifier runs on the balanced Bags to determine each Bag's accuracy. The algorithmic pseudocode of the WSMOTE-ensemble classifier is illustrated in Algorithm 3.

The WSMOTE-ensemble classifier exploits the SCMs and generates class-balanced Bags, for which C4.5 is a baseline model. Hence, it provides us with some unique benefits for skewed data modeling compared with existing approaches. First, the WSMOTE-ensemble espouses segregated SCMs in both minority and majority sets of instances. To enlarge the set of minority instances, the WSMOTE algorithm is also trained in the data regions where no replication instances exist. Having ensured the multiplicity of baseline learners, afterward, the Bagging strategy is applied to reduce the set of majority instances. This results in the fusion of WSMOTE, SCM, Bagging, and C4.5 learner competence to address the skewed class problem. Besides the above enhancements, the WSMOTE-ensemble generates multiple class-balanced Bags from which only the best are selected based on their accuracy. This innovative class-balancing mechanism bumps up the divergence and perks up the recitals of fusion mechanisms. We believe that the proposed WSMOTE-ensemble algorithm augments a skewed data classification performance. To sum up, the proposed WSMOTE-ensemble contrasts with the existing ensemble approaches in several areas:

- The proposed WSMOTE-ensemble is an extended, enhanced edition of the approach of Sun et al. [61]. Specifically, SMOTE was replaced with its enhanced weighted-based modification, making the credit risk classifier more robust to class imbalance.
- Our explanation generalizes to a rare event domain, where instances have dissimilar traits in terms of dimension, skewness level, and size.
- It solves the overtraining dilemma of typical oversampling because it ensures the variety in the WSMOTE algorithm to generate weighted synthetic loan default instances.
- The proposed WSMOTE-ensemble classifier assembles an optimal fusion methodology for skewed data learning, which ensures both the accuracy and stability of prediction. This methodology combines the sampling method with several ensemble strategies, rather than relying on simple bootstrap aggregation of decision trees like [61].
- Compared with SMOTE, WSMOTE, random undersampling (RUS), MChanUS, USOS, and RUSSMOTE, the WSMOTE-ensemble classifier successfully evades the uncertainty of the Sun et al. [61] approach when dealing with the small number of minority instances.

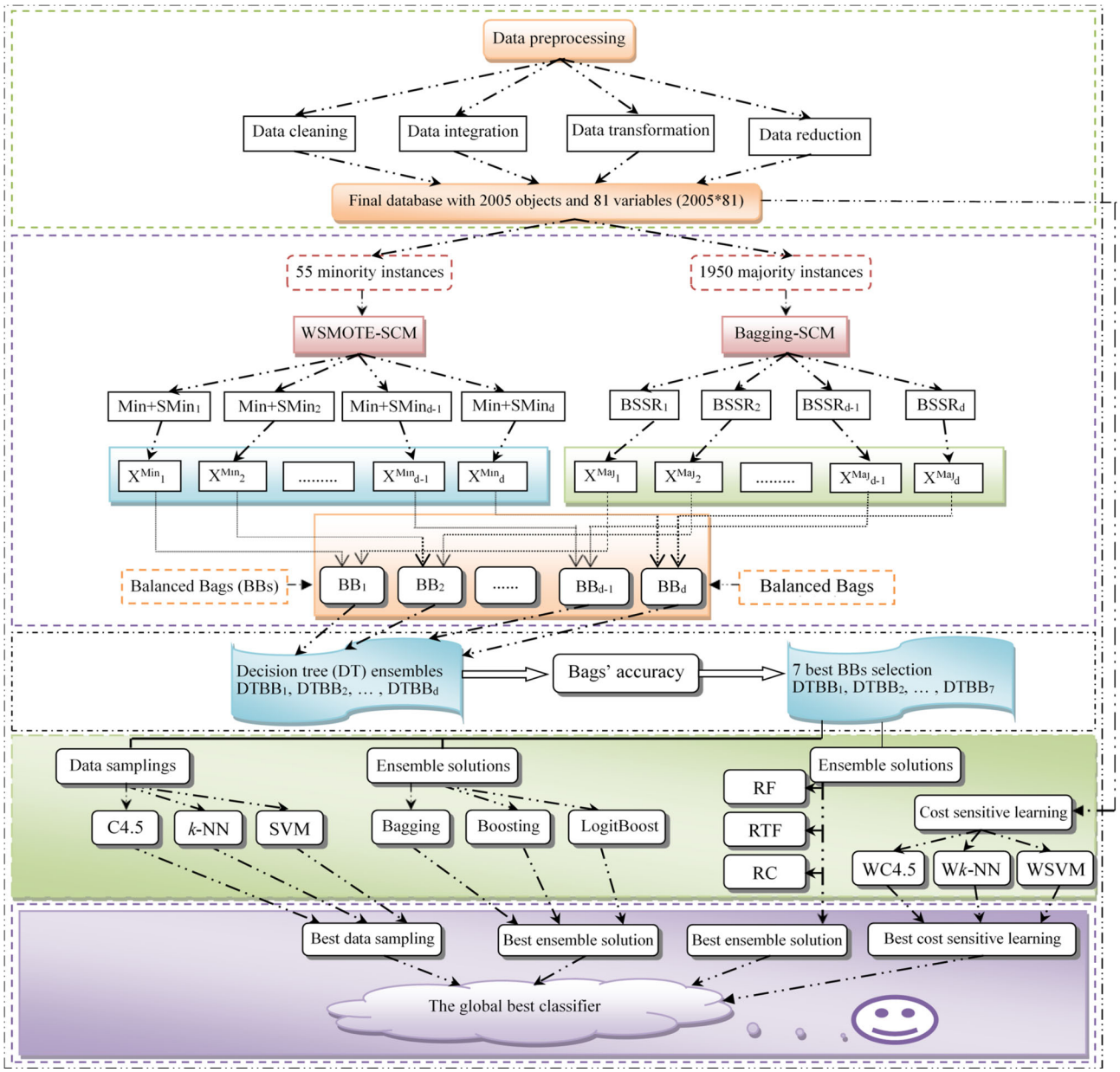

Fig. 2 Flowchart of WSMOTE-ensemble algorithm

Experimental design

Dataset and instance composition

The original commercial bank database used in this study was generated from 3111 records of firms that received small business loans in China from 1992 to 2012; of these, 3040 were creditworthy (majority class) and the remaining 71 were noncreditworthy loan customers (minority class). The current study provides the supplementary data file, which includes detailed outcomes of the trained algorithms (see Supplementary Tables S1 to S3). Details about the data source, including the bank's name, its origin, and its functionality, were excluded to maintain privacy. The experimental dataset was managed by the National Natural Science Foundation of China (NSFC), which is headed by the second contributor to this study. Due to the conditions imposed by the bank and the NSFC, it might be difficult to make the dataset public, but an interested reader may contact the project coordinator about the experimental dataset and other relevant issues. According to the Standards for Classification of Small and Medium-sized Enterprises [44], this study includes small business loan samples across multiple industries, namely, the retail, wholesale, information services, and transportation

Algorithm 3 WSMOTE-ensemble

Input: Training dataset $TRDS$={ ($\mathbf{x}_i^{tr}$, y_i), i=1,2,. . . ,n_{tr}} and testing dataset $TEDS$={ ($\mathbf{x}_i^{te}$, y_i), i=1,2,. . . ,n_{te}} , n_{tr} and n_{te} denote the number of training and testing instances, respectively.
Output: The trained WSMOTE-ensemble model.
Procedure:
{
Step 1: // Produce the sampling composite mixtures
SCM= [SCM_d] (d= 1,2,. . . ,D), where D denotes the number of initial classifiers;
Step 2: // Decide on the instance set of each category in SCM rates SCM_d;
M =S_{Maj}-S_{Min}
SCM_d =[100 × (d/D)], d=1,2,. . . ,D
S^d_{Min} = S_{Min} + round (M × SCM_d) S_{Min}+ S^d_{Min}= S^N_{Min},
where M refers to the difference between the instance sets of two classes, S_{Maj} and S_{Min} are the sets of majority and minority instances, respectively, S^d_{Min} are the weighted synthetic minority instances, and S^N_{Min} is the new minority training dataset.
Step 3:
For d=1 to D
i. WSMOTE-SCM is applied to generate S^d_{Min} and merge it with the initial minority examples (S_{Min}) to produce the minority training dataset $S^N_{Min.}$
ii. Bagging-SCM is used to generate new majority instances (S_{Maj^d}). It follows random sampling with replacement. Then, S^N_{Maj} is the reduced set of negative instances.
iii. Class-balanced Bags = S^N_{Min}+ S^N_{Maj}
End For
Step 4: // Train baseline C4.5 decision classifier and determine the Bag's accuracy;
Step 5: // Pick the seven best Bags based on maximum accuracy;
Step 6: // Selected Bags are trained over the current experimental settings;
gen_imbal_data_classifier[Train{WSMOTE-ensemble}{EnsembleProposal}{ DatalevelSolution}
{AlgorithmicModification}]
Step 7: // Get the prediction results of training instances;
imbal_data_training(TRDS$_{count}$, MultilevelExperiment$_{count}$, PerformanceMeasures$_{count}$)
Step 8: // Generate the classification outputs from the testing instances;
imbal_data_testing(TEDS$_{count}$, MultilevelExperiment$_{count}$, PerformanceMeasures$_{count}$)
Step 9: // Compare the outcomes to generate a global solution.
}

sectors, as well as other sectors. For this study, the data were preprocessed and the duplicate examples were removed; following the purifications, the sample sizes were 1950 in the majority class and 55 in the minority class used as original instances (instance composition 1 (IC 1)) in the experiment. The highest imbalance ratio (IR) was 35.45. Each small business credit customer in the dataset had 48 financial variables, 27 nonfinancial variables, and 6 macroeconomic variables, for a total of 81 variables. The financial variables included the debt-to-asset ratio, net cash flow ratio, quick asset ratio, liquidity ratio, net cash flow-to-asset ratio, current asset ratio, and cash ratio. The nonfinancial variables included the audit status, patent, types of bank accounts, sales scope, and educational background of owner. The macroeconomic variables included the business cycle index, gross domestic product, and consumer price index. Table 1 illustrates the descriptive statistics of applied variables. Among the financial variables, the total of the outstanding loans to the total assets and the total of the outstanding loans to the net assets had maximum mean values of 0.99 and 0.94, respectively. Conversely, the cash flow from operating activities and the net cash flow to sales revenue had the smallest standard deviations of 0.032 and 0.0435, respectively. The results of the Student's paired t-test indicated that the values of all variables were statistically different for the default and nondefault class at the 1% level. A lack of variables for economic trends has traditionally been a primary constraint of small business data analysis. Moreover, the macroeconomic variables are critical factors that directly influence the payment behavior of any creditor [34]. The Chinese small business loan performance database, therefore, is an ideal dataset for credit risk modeling.

We focused on seven different skewed data-balancing strategies to verify the feasibility and effectiveness of the proposed WSMOTE-ensemble for the imbalanced small business loan dataset. Following the SMOTE and RUS strategies, this study generated seven different instance sets to make the dataset balanced, for example, SMOTE-1 to SMOTE-7 (instance compositions IC2 to IC8) and RUS-1 to RUS-7 (IC16 to IC22), which decreased the IR from 35.45 to 1.11 and 35.45 to 1.00, respectively. Note that the combination of IC2 to IC8 was applied to train the WSMOTE algorithm. Based on the hybrid strategies USOS and RUSSMOTE, an additional four sets of ICs were formed, namely, USOS-1 to OSUS-4 (IC30 to IC33) and RUSSMOTE-1 to RUSSMOTE-4 (IC34 to IC37), with IR ratios of 15.95 to 1.00 in each cluster, respectively. The current study also applied the MChanUS model by producing seven balanced sets of instances (IC23 to IC29). Lastly, the present study picked up the seven best WSMOTE-ensemble fusion-sampling instances (IC9 to IC15) based on the highest accuracies from the C4.5 classifier. It includes the number of selected Bags (B#37, B#29,...,B#35) and their respective samples. The seven best-performing Bags were selected to maintain uniformity with the earlier algorithms. All of these Bags were balanced. A description of sampling compositions is presented in Table 2.

To consistently evaluate the performance of the proposed methods, a five-fold cross-validation was applied. The traditional ten-fold cross-validation was not used due to the lack of minority class instances in the experimental dataset. Experiments were conducted on a personal computer with a 3.10-GHz Intel Core i5-2400 CPU and 4 GB RAM, on the Windows 7 operating system in the following program environments: MATLAB R2017b, the open-source data-mining toolkits of WEKA 3.8.0 (Waikato Environment for Knowledge Analysis), KEEL (Knowledge Extraction based

Table 1 Descriptive statistics of used variables (mean ± std)

Panel A: Financial variables		Accounts payable turnover velocity	0.8142 ± 0.3596
Debt to asset ratio	0.4642 ± 0.2490	Cash cycle	0.4240 ± 0.1916
Ratio of net CF	0.4884 ± 0.1037	Revenue growth rate	0.2138 ± 0.1820
Quick asset ratio	0.1949 ± 0.2237	Profit growth rate	0.4937 ± 0.0604
Liquidity ratio	0.1376 ± 0.1462	Total assets growth rate	0.2861 ± 0.1391
Net CF to main business income	0.1800 ± 0.1330	Rate of capital accumulation	0.4993 ± 0.0492
EBIT to current liability ratio	0.4823 ± 0.0791	Retained earnings growth rate	0.4881 ± 0.1175
(Long-term liabilities)-to-(long-term liabilities plus owners equity)	0.9000 ± 0.2760	**Panel B: Non-financial attributes**	
Net CF-to-asset ratio	0.4900 ± 0.1370	Entire period of actual operations	0.7900 ± 0.3500
Current asset ratio	0.1642 ± 0.1705	Audit status (audit or not)	0.0400 ± 0.1930
Net operating CF to net profit	0.4878 ± 0.1382	Hierarchy of new product	0.1100 ± 0.2520
Net asset to total loan	0.2100 ± 0.2590	Patent condition	0.0800 ± 0.2090
Net asset to owners' equity	0.8987 ± 0.1875	Enterprise establishment date	0.4183 ± 0.4224
Cash ratio	0.1493 ± 0.2547	Types of bank account	0.5400 ± 0.3100
Total liability to fixed assets	0.0600 ± 0.1820	Sales scope	0.4600 ± 0.2910
Outstanding loans to net assets	0.9400 ± 0.1900	Level of brand products	0.1800 ± 0.2740
Outstanding loans to total assets	0.9900 ± 0.1130	Enterprise loan ratio	0.3900 ± 0.3900
Net CF to noncurrent liability	0.0600 ± 0.1440	Educational background	0.8000 ± 0.3010
Net CF to assets ratio	0.4822 ± 0.1077	Default records	0.7400 ± 0.3800
EBITDA to liabilities	0.0441 ± 0.1009	Credit history	0.5900 ± 0.4930
Return on equity	0.1178 ± 0.1466	Marital status	0.9290 ± 0.1532
Net CF to sales revenue	0.0059 ± 0.0435	Residence status	0.6400 ± 0.4780
Net profit to sales revenue	0.0593 ± 0.0794	Residence duration	0.7900 ± 0.3860
Return on total assets	0.0997 ± 0.1219	Gender	0.9140 ± 0.1646
Operating profit margin	0.2895 ± 0.2378	Age	0.8800 ± 0.1990
Net profit to operating costs	0.0629 ± 0.1120	Automobile and real estate	0.1800 ± 0.2270
Gross profit rate	0.2600 ± 0.2680	Monthly family income	0.1100 ± 0.2220
Total profit to operating costs	0.4725 ± 0.1386	Job duration	0.4000 ± 0.3740
EBITDA	0.1673 ± 0.2376	Registered capital	0.8400 ± 0.3630
EBITDA to total revenue	0.0673 ± 0.0853	Enterprise credit in 3 years	0.8100 ± 0.3660
Net profit	0.1620 ± 0.2307	Tax records	0.8300 ± 0.3660
Net operating CF	0.5012 ± 0.0893	Legal dispute number	0.8760 ± 0.2923
Operating CF	0.0028 ± 0.0320	Business status (lawful/not)	0.4300 ± 0.2200
Receivable turnover velocity	0.0317 ± 0.1012	Number of breach of contract	0.8200 ± 0.3840
Inventory turnover velocity	0.0195 ± 0.0892	**Panel C: Macroeconomic attributes**	
Total assets turnover velocity	0.2799 ± 0.2955	The business cycle index	0.6817 ± 0.1191
Velocity of liquid assets	0.0232 ± 0.0478	Urban residents per capita savings	0.4981 ± 0.1618
Velocity of fixed assets	0.0491 ± 0.1606	GDP growth rate	0.3797 ± 0.0824
Velocity of equity	0.0980 ± 0.1648	Consumer price index	0.9848 ± 0.0438
Working capital ratio	0.2618 ± 0.1798	Citizens' per capita income	0.4681 ± 0.1309
Return on investment	0.0100 ± 0.0930	Engel coefficient	0.7294 ± 0.0699

CF cash flow, *EBIT* earnings before interest and taxes, *EBITDA* earnings before interest, taxes, depreciation and amortization

Table 2 Description of datasets with different sampling strategies used in experiment

Instance composition (IC)	# Features in all instance compositions: 81		
	# Instances in majority (negative) class (NC)	# Instances in minority (positive) class (PC)	Imbalance ratio (NC/PC)
1. Original instances	1950	55	35.45
2. SMOTE-1 (WSMOTE-1)	1950	110 (100%*55)	17.73
3. SMOTE-2 (WSMOTE-2)	1950	220 (100%*110)	8.86
4. SMOTE-3 (WSMOTE-3)	1950	440 (100%*220)	4.43
5. SMOTE-4 (WSMOTE-4)	1950	880 (100%*440)	2.22
6. SMOTE-5 (WSMOTE-5)	1950	1320 (50%*880)	1.48
7. SMOTE-6 (WSMOTE-6)	1950	1760 (100%*880)	1.11
8. SMOTE-7 (WSMOTE-7)	1950	1950 (50%*1320-30)	1.00
9. WSMOTE-ensemble-1 (B#37)	1808	Pos#55; WSmotePos#1753	1.00
10. WSMOTE-ensemble-2 (B#29)	1429	Pos#55; WSmotePos#1374	1.00
11. WSMOTE-ensemble-3 (B#40)	1950	Pos#55; WSmotePos#1895	1.00
12. WSMOTE-ensemble-4 (B#34)	1666	Pos#55; WSmotePos#1611	1.00
13. WSMOTE-ensemble-5 (B#27)	1334	Pos#55; WSmotePos#1279	1.00
14. WSMOTE-ensemble-6 (B#39)	1903	Pos#55; WSmotePos#1848	1.00
15. WSMOTE-ensemble-7 (B#35)	1713	Pos#55; WSmotePos#1658	1.00
16. RUS-1	1755 (1950*10%)	55	31.91
17. RUS-2	1455 (300)	55	26.45
18. RUS-3	1155 (300)	55	21
19. RUS-4	855 (300)	55	15.55
20. RUS-5	555 (300)	55	10.09
21. RUS-6	255 (300)	55	4.64
22. RUS-7	55 (200)	55	1.00
23. MChanUS-1	278	Pos#55; WSmotePos#223	1.00
24. MChanUS-2	278	Pos#55; WSmotePos#223	1.00
25. MChanUS-3	278	Pos#55; WSmotePos#223	1.00
26. MChanUS-4	278	Pos#55; WSmotePos#223	1.00
27. MChanUS-5	278	Pos#55; WSmotePos#223	1.00
28. MChanUS-6	278	Pos#55; WSmotePos#223	1.00
29. MChanUS-7	278	Pos#55; WSmotePos#223	1.00
30. USOS-1	1755 (1950*10%)	110 (100%*55)	15.95
31. USOS-2	1455 (300)	220 (100%*110)	6.61
32. USOS-3	1155 (300)	440 (100%*220)	2.63
33. USOS-4	880 (275)	880 (100%*440)	1.00
34. RUSSMOTE-1	1755	110	15.95
35. RUSSMOTE-2	1455	220	6.61
36. RUSSMOTE-3	1155	440	2.63
37. RUSSMOTE-4	880	880	1.00

on Evolutionary Learning) GPLv3 tool, and SPSS Modeler (SPSS 17.0).

Performance evaluation

This study employed the four most widely used evaluation criteria for imbalanced datasets. The metrics were based on a 2×2 confusion matrix, as presented in Table 3, where tp, tn, fp, and fn are the true-positive, true-negative, false-positive, and false-negative results, respectively.

Many empirical researchers with theoretical evidence claim that a highly skewed class distribution can make the overall accuracy metric almost worthless. In addition, it ignores the class efficiency from different sample clusters. Therefore, apart from the accuracy criterion, many other performance appraisals have been used to discover unseen

Table 3 The confusion matrix for credit risk assessment problem

		Predicted observations	
		Predicted positive	Predicted negative
Actual observations	Actual positive	tp	fp
	Actual negative	fn	tn

characteristics of skewed data problems. The most common are majority-class accuracy (true-positive rate, TP_{rate}) and minority-class accuracy (true-negative rate, TN_{rate}), which are typically employed to investigate the class imbalanced learning performance for each class separately. Note that TP_{rate}, Eq. (1) evaluates the proportion of nondefault loans predicted to be nondefault loans, whereas TN_{rate}, Eq. (2) evaluates the proportion of default loans predicted to be default loans.

$$\text{TP}_{\text{rate}} = tp/(tp + fn), \tag{1}$$

$$\text{TN}_{\text{rate}} = tn/(tn + fp), \tag{2}$$

$$G_m = (\text{TP}_{\text{rate}} \times \text{TN}_{\text{rate}})^{1/2}, \tag{3}$$

$$\text{AUC} = \frac{1}{2}\left(1 + \frac{tp}{tp + fn} - \frac{fp}{fp + tn}\right). \tag{4}$$

Similarly, the G-mean metric (G_m), Eq. (3), can be perceived as an indicator of the balanced performance of the classifier between the two classes. Finally, the area under the receiver operating characteristic curve (AUC) statistic, Eq. (4), is also an important measure of the discriminatory power of skewed learning tasks. An ideal skewed learning task should have an AUC of 1.0, whereas an AUC of 0.5 denotes a random classifier [45].

Statistical tests

Two categories of statistical comparisons were used for multiple and pairwise evaluations, respectively, based on the reference of García et al. [25]. The Wilcoxon signed-rank test and McNemar test were applied when only two classifiers were compared, and the Friedman test and Iman-Davenport expansion were utilized for multiple assessments based on the mean ranks of the classifiers of several ICs. Classifiers were arranged from the best to worst for each IC to obtain the mean ranks. Subsequently, the ranks were averaged for all ICs. Given M ICs, D classifiers and average ranks $R_d, d = 1,...,D$, the Iman-Davenport test worked out an F-distribution statistic with D-1 and $(D\text{-}1) \times (M\text{-}1)$ degrees of freedom to authenticate the null hypothesis as follows:

$$F = \frac{(M-1)x_F^2}{M(D-1) - x_F^2}, \tag{5}$$

$$\chi_F^2 = \frac{12M}{D(D-1)}\left[\sum R_d^2 - \frac{D(D+1)^2}{4}\right]. \tag{6}$$

A post-hoc test was employed to figure out where the differences existed if the null hypothesis that all algorithms have an equal recital was rejected. In this experiment, we used the Holm post-hoc test, because it can suitably manage the family-wise error rate and, therefore, is more precise than the Bonferroni-Dunn extension test [25]. As a control mechanism, the Holm test initially picks the classifier with the lowest average rank and then utilizes a step-down process to evaluate the control mechanism with the other classifiers. The Holm test uses the Z-statistic as follows:

$$Z = (R_{d1} - R_{d2}) \times \left\{\left(D^2 + D\right)/6M\right\}^{0.50}, \tag{7}$$

where R_{d1} and R_{d2} are the mean ranks of classifiers $d1$ and $d2$. The Holm test organizes p-values as $p_1 \leq p_2 \leq p_3 \cdots \leq p_{D-1}$, after employing the Z-statistic to point out the matching probability from the normal distribution table. The respective hypothesis is rejected if p_1 is below $\alpha/(D\text{-}1)$, and the experiments are permitted to evaluate p_2 with $\alpha/(D\text{-}2)$, and so on. All remaining hypotheses are optimized once a respective null hypothesis is not rejected. The current study follows the significance level $\alpha = 0.05$ in all statistical comparisons.

Empirical results

Algorithm selection and parameter settings

This study hybridizes class skewed learning with the base classifiers C4.5, k-NN, and SVM and the ensemble classifiers Bagging, Boosting, LB, RC, RTF, and RF. Cost-sensitive learning was also applied. The appropriate data level was used, and the hybrid sampling strategies were SMOTE, WSMOTE, WSMOTE-ensemble, RUS, MChanUS, USOS, and RUSSMOTE. Therefore, this investigation employed 69 different criteria for optimizing class-imbalanced small business loan data, as shown in Table 4. Table 4 also highlights the parameter settings and original studies of the respective algorithms, in which technical details are given to readers. The current section presents the combined experimental results. Due to space constraints, the detailed results of the four evaluation measures over multiple ICs are shown in Supplementary Tables S1 and S2. The best performance for each category is in bold; algorithms are ranked based on the AUC

Table 4 Algorithms and parameters used in experiments

Type	Algorithms and parameters
Baseline	C4.5 [53] ; pruned, confidence = 0.25, instances per leaf = 2
	k-NN [23]; k = 1, Euclidean distance
	SVM [3]; Radial basis function kernel, heuristic search, regularization term = 1.0
Data level solutions [15,52] (k = 5; # Bags = 40)	
	C4.5: SMOTE-C4.5, WSMOTE-C4.5, WSMOTE-ensemble-C4.5,RUS-C4.5, MChanUS-C4.5,USOS-C4.5, RUSSMOTE-C4.5
	k-NN: SMOTE-k-NN, WSMOTE-k-NN, WSMOTE-ensemble-k-NN, RUS-k-NN, MChanUS-k-NN, USOS-k-NN, RUSSMOTE-k-NN
	SVM: SMOTE-SVM, WSMOTE-SVM, WSMOTE-ensemble-SVM, RUS-SVM, MChanUS-SVM, USOS-SVM, RUSSMOTE-SVM
Cost-sensitive learning [65]	Weighted C4.5; $C_{\text{nondefault}}$ = 11, C_{default} = 56
	Weighted k-NN; $C_{\text{nondefault}}$ = 11, C_{default} = 56
	Weighted SVM; $C_{\text{nondefault}}$ = 11, C_{default} = 56
Ensemble solutions [36,61,67] (# Bags = 40, # iterations = 40, batch size = 100)	
Bagging	SMOTE-Bagging, WSMOTE-Bagging, WSMOTE-ensemble-Bagging, USOS-Bagging, RUS-Bagging, MChanUS-Bagging, RUSSMOTE-Bagging
Boosting	SMOTE-Boosting, WSMOTE-Boosting, WSMOTE-ensemble-Boosting, USOS-Boosting, RUS-Boosting, MChanUS-Boosting, RUSSMOTE-Boosting
LB	SMOTE-LB, WSMOTE-LB, WSMOTE-ensemble-LB, USOS-LB, RUS-LB, MChanUS-LB, RUSSMOTE-LB; shrinkage = 1.0, pool size = 1.0
RC	SMOTE-RC, WSMOTE-RC, WSMOTE-ensemble-RC, USOS-RC, RUS-RC, MChanUS-RC, RUSSMOTE-RC
RTF	SMOTE-RTF, WSMOTE-RTF, WSMOTE-ensemble-RTF, USOS-RTF, RUS-RTF, MChanUS-RTF, RUSSMOTE-RTF
RF	SMOTE-RF, WSMOTE-RF, WSMOTE-ensemble-RF, USOS-RF, RUS-RF, MChanUS-RF, RUSSMOTE-RF; # trees = 40

measure; and the statistical significance is measured by the nonparametric Wilcoxon and McNemar tests.

Study on data-level experiments

Table 5 shows the data-level evaluation results over the baseline algorithms. The values of the performance measures were calculated as averages over five-folds. Compared with the nonsampling strategy (original instances), all data-sampling methodologies had a positive impact on the classification performance of the baseline methods. Notably, the proposed WSMOTE-ensemble and WSMOTE algorithms outperformed all the base classifiers in terms of minority class accuracy (TP_{rate}), AUC, and G_m, except for the SVM for AUC; SMOTE was better in this case. Table 5 also reveals that the fusion MChanUS method performed better for the majority class (TN_{rate}) for all base classifiers. For the class imbalance learning, the results showed that WSMOTE-ensemble was more feasible for generating the reweighting of instances in favor of the minority class. The nonsampling techniques provided worse results in all evaluations. However, the degree of improvement over the nonsampling technique for the best data-level solution depended on the used base classifiers, for example, k-NN improved 114.91%, 48.10%, and 37.95% for TP_{rate}, AUC, and G_m, respectively, and C4.5 generated an improvement of 100.49%, 41.41%, and 33.14%, respectively. SVM showed the least improvement for all measures, namely, 42.47%, 17.97%, and 16.61%, respectively. In contrast to the WSMOTE-ensemble, it is worth noting that the approach of Sun et al. [61] generated 81.66% G_m and 84.00% minority class accuracy (TP_{rate}) when applied to the C4.5 classifier. Based on the parallel learning environment, our blending approach generated a 97.54% TP_{rate} and a 97.70% G_m, exhibiting perfection of 19.64% and 16.12%, respectively. Overall, the k-NN approach using the WSMOTE-ensemble data-level technique seemed to be the best solution for small business class imbalanced learning on all grounds.

Based on the AUC performance ranking and Wilcoxon signed-rank test performed for all pairs of data sampling methods over 5 folds, the improvements from the best algorithms (WSMOTE-ensemble for C4.5 and k-NN; and SMOTE for SVM) were statistically significant. RUS performed particularly poorly compared with other data sampling methods.

Table 5 Experimental performance of data sampling with baseline algorithms

Instance composition	C4.5					Sun et al. [61]				
	TP_{rate}	TN_{rate}	G_m	AUC	p(WSR)	G_m	TP_{rate}			
Original instances	0.4865	0.9812	0.6909	0.7338(8)	–	0.6863	0.5476			
SMOTE	0.8993	0.9819	0.9376	0.9406(4)	0.1160	0.7255	0.6374			
WSMOTE	0.9718	0.9251	0.9475	0.9485(3)	0.3100	–				
WSMOTE-ensemble	0.9754	0.9787	0.9770	0.9770(1)	–	0.8166	0.8400			
RUS	0.5516	0.9373	0.7145	0.7445(7)	0.0180**	–				
MChanUS	0.9185	1.0000	0.9584	0.9593(2)	0.0180**	–				
USOS	0.8806	0.9948	0.9353	0.9377(5)	0.0280**	0.7253	0.6420			
RUSSMOTE	0.7878	0.7817	0.7644	0.7847(6)	0.0180**	–				
	k-NN					SVM				
	TP_{rate}	TN_{rate}	G_m	AUC	p(WSR)	TP_{rate}	TN_{rate}	G_m	AUC	p(WSR)
Original instances	0.4546	0.9749	0.6657	0.7147(8)	–	0.6923	0.9769	0.8224	0.8346(6)	–
SMOTE	0.8452	0.9948	0.9158	0.9200(5)	0.0280**	0.8642	0.9851	0.9218	0.9846(1)	–
WSMOTE	0.9847	0.9678	0.9758	0.9762(2)	0.1760	0.9402	0.9787	0.9592	0.9594(3)	0.0280**
WSMOTE-ensemble	0.9770	0.9948	0.9859	0.9859(1)	–	0.9351	0.9774	0.9595	0.9598(2)	0.1760
RUS	0.5923	0.9117	0.7296	0.7520(7)	0.0180**	0.6330	0.9364	0.7662	0.7847(8)	0.0180**
MChanUS	0.9398	0.9989	0.9689	0.9694(3)	0.0180**	0.9001	0.9852	0.9420	0.9430(4)	0.4990
USOS	0.9183	0.9993	0.9574	0.9588(4)	0.0280**	0.8300	0.9824	0.9025	0.9062(5)	0.3100
RUSSMOTE	0.7780	0.8184	0.7807	0.7982(6)	0.0180**	0.7859	0.8272	0.6796	0.8065(7)	0.1760

Note: WSR is Wilcoxon signed-rank test, **statistically significant at $p < 0.05$

Table 6 Experimental performance of weighted classifiers

Instances composition	C4.5				k-NN			
	TP_{rate}	TN_{rate}	AUC	G_m	TP_{rate}	TN_{rate}	AUC	G_m
None	0.9815	0.3273	0.7179	0.5668	0.9908	0.3455	0.6681	0.5850
Weighted	0.9841	0.3455	0.7828	0.5831	0.9862	0.4182	0.7022	0.6422
McNemar test	$p(TP_{rate}) = 1.000$; $p(G_m) = 0.937$				$p(TP_{rate}) = 1.000$; $p(G_m) = 0.862$			
	SVM							
None	0.9821	0.3091	0.8324	0.5509				
Weighted	0.9867	0.4545	0.8982	0.6697				
McNemar test	$p(TP_{rate}) = 1.000$; $p(G_m) = 0.987$							

Study on cost-sensitive learning

Table 6 shows the cost-sensitive learning outcomes averaged over non-weighted and weighted baseline classifiers. All weighted algorithms generated progressive results compared with baseline classifiers (None), but their degree of improvement varied. In terms of the AUC, G_m, and TP_{rate}, the weighted SVM (WSVM) outperformed the other two classifiers, providing results of 0.8982, 0.6697, and 0.9867, respectively. For the TN_{rate}, WSVM was also the best amongst its equivalents. The best cost-sensitive learning (WSVM) produced improvements of 7.90%, 21.56%, 0.47%, and 47.04% over the baseline SVM in terms of the AUC, G_m, TP_{rate}, and TN_{rate} measures, respectively. The result of the McNemar test (comparing the confusion matrices obtained using the non-weighted vs. weighted baseline classifiers) suggested that cost-sensitive learning showed no significant differences in terms of the TP_{rate} and G_m for the best classifiers.

Study on the ensemble experiments

The final group of approaches for dealing with imbalanced data was based on ensemble methods. The researchers chose six classifiers that had performed the best in earlier studies [24,40,46]. To be specific, if M = [Bagging, Boosting, LB,

RC, RTF, RF] and F = [SMOTE, WSMOTE, WSMOTE-ensemble, RUS, MChanUS, USOS, RUSSMOTE], then the experimental settings followed up $M \times F$. In all ensemble cases, C4.5 was employed as the baseline algorithm, adopting the earlier ensemble methodologies, with the settings kept identical to those utilized in the baseline algorithms. The current study fixed the numbers of iterations in Boosting-based hybrids, Bags in Bagging-based hybrids, and trees in RF-based hybrids to be 40. Moreover, in LB-, RC-, and RTF-based ensembles, the batch size was 100, and the number of iterations for the latter two approaches was 40. These settings were consistent with those in earlier studies [49,66]. To authenticate whether these ensembles were advantageous in dealing with the skewed dataset, the fundamental algorithms of the corresponding ensembles were also utilized to provide comparisons.

The results for the four performance criteria are shown in Table 7. The results indicated that all of the ensemble strategies outperformed their baselines. In particular, among seven different approaches, WSMOTE-ensemble-RF, an integration of RF with WSMOTE-ensemble, performed the best in terms of the AUC (0.9910), G_m (0.9910), and TP_{rate} (0.9916), thus being more effective than SMOTE-Bagging and SMOTE-Boosting. Moreover, the effectiveness of the WSMOTE-ensemble was confirmed by the good performance of WSMOTE-ensemble-RTF, which ranked second for the AUC and the G_m.

Again, the Wilcoxon signed-rank test was conducted to compare the AUC performance (over five-folds) of the best data sampling method (WSMOTE-ensemble ranked first in all cases) with its data sampling competitors. The results show that RUS, MChanUS, USOS, and RUSSMOTE were significantly outperformed by WSMOTE-ensemble for all ensemble learning methods.

A significant finding from the results for ensemble strategies is that the WSMOTE-ensemble-RF fusion technique is a preferable choice based on the performance evaluations, and the WSMOTE-ensemble-Bagging technique is an alternative option. Both had ensured enhanced performance compared with their counterparts for skewed class data in small business credit risk.

Comparative analysis

Comparison of WSMOTE, WSMOTE-ensemble, and MChanUS

We compared the experimental results of SMOTE, WSMOTE, SMOTE-ensemble, and USOS in this separate section because WSMOTE and WSMOTE-ensemble were the outcomes of the enhanced SMOTE- and Bagging-based algorithms for imbalanced classification tasks. The fundamental procedure of the WSMOTE algorithm is to assign weights for generating new artificial data using SMOTE for a specific minority positive data point. Each of the positive instances produces an equal number of artificial instances, and the resulting outcome is an amendment to the SMOTE algorithm. Finally, the hybridized data for imbalanced learning were trained using default machine learning algorithms.

Table 8 presents the ground mean values for four evaluation measures for 129 (43×3) data samplings and 258 (43×6) ensemble learning samplings on 43 ICs (7 for SMOTE, WSMOTE, WSMOTE-ensemble, RUS, and MChanUS, and 4 for USOS and RUSSMOTE); this achieved a total of 387 experiments. As Table 8 shows, the WSMOTE-ensemble algorithm provided superior results over its counterparts in all measures except TN_{rate}. The stability of the results is illustrated in Figs. 3 and 4. The results show that the proposed WSMOTE-ensemble algorithm was effective in generating synthetic instances for the minority default class, which was achieved at the cost of a slightly deteriorated performance for the TN_{rate} by 0.2039%. Therefore, this study recommends that the WSMOTE-ensemble has distinct advantages over WSMOTE or SMOTE for balancing the skewed data classes. This is shown by its achievement of improved results across all criteria over WSMOTE and SMOTE by 7.3661% for TP_{rate}, 3.5181% for AUC, and 3.7393% for G_m. What is more, all improvements in all criteria were statistically significant at 1% for the Wilcoxon signed-rank test conducted over the 387 experiments for the four performance criteria, which strongly supports the above findings. WSMOTE-ensemble significantly outperformed all the compared sampling methods in terms of TP_{rate}, AUC, and G_m, and WSMOTE performed best for the majority class (TN_{rate}). However, WSMOTE-ensemble was not significantly outperformed for this classification measure.

Global comparisons

In this last section of the experimental analysis, an overall assessment was carried out to determine which method had performed best based on three evaluation criteria. We performed cross-family comparisons for the instance composition methods previously selected as the representatives for each case. The methods were chosen based on the AUC, G_m, and TP_{rate} because these measures are critical for class-imbalanced scenarios. In all approaches, evaluations gained from the AUC and G_m were consistent with that for the TP_{rate}; hence, an inclusive assessment was conducted for the AUC criterion. Using predominantly data-level solutions, this study selected the best sampling methods for each baseline classifier, except SVM, for which two approaches were selected. The first one was best for AUC, and the second one for G_m. Because the three cost-sensitive learning methods can stand out among the original algorithms, they stood out

Table 7 Experimental performance of ensemble learning methods

Instance composition	Bagging				Boosting				LB			
	TP_{rate}	TN_{rate}	G_m	AUC	TP_{rate}	TN_{rate}	G_m	AUC	TP_{rate}	TN_{rate}	G_m	AUC
Original instances	0.5000	0.9745	0.6980	0.7372(8)	0.5909	0.9788	0.7605	0.7845(6)	0.5600	0.9793	0.7405	0.7697(7)
SMOTE	0.9266	0.9803	0.9586	0.9535(5)**	0.8090	0.9611	0.8722	0.8850(3)	0.8416	0.9640	0.8994	0.9027(4)*
WSMOTE	0.9705	0.9577	0.9639	0.9641(2)	0.9047	0.9352	0.9197	0.9200(2)*	0.9190	0.9375	0.9281	0.9282(2)*
WSMOTE-ensemble	0.9752	0.9892	0.9822	0.9822(1)	0.9103	0.9360	0.9229	0.9231(1)	0.9191	0.9638	0.9419	0.9422(1)
RUS	0.5941	0.9189	0.7360	0.7565(7)**	0.4643	0.9208	0.6405	0.6926(8)**	0.6059	0.9221	0.7418	0.7640(8)**
MChanUS	0.9294	0.9913	0.9598	0.9603(3)**	0.8684	0.8571	0.8627	0.8627(4)**	0.8785	0.9355	0.9066	0.9070(3)**
USOS	0.9116	0.9898	0.9537	0.9544(4)**	0.7307	0.9212	0.8149	0.8260(5)**	0.7897	0.9554	0.8674	0.8726(5)**
RUSSMOTE	0.7989	0.8427	0.8082	0.8208(6)**	0.7551	0.7573	0.7296	0.7562(7)**	0.7741	0.7776	0.7563	0.7758(6)**
	RC				RTF				RF			
	TP_{rate}	TN_{rate}	G_m	AUC	TP_{rate}	TN_{rate}	G_m	AUC	TP_{rate}	TN_{rate}	G_m	AUC
Original instances	0.5882	0.9822	0.7601	0.7852(8)	0.7391	0.9808	0.8515	0.8600(6)	0.6191	0.9788	0.7784	0.7989(8)
SMOTE	0.9539	0.9884	0.9707	0.9712(4)	0.9554	0.9879	0.9717	0.9722(4)	0.9579	0.9876	0.9723	0.9728(4)
WSMOTE	0.9913	0.9360	0.9902	0.9636(5)	0.9902	0.9378	0.9627	0.9640(5)	0.9907	0.9492	0.9691	0.9698(5)
WSMOTE-ensemble	0.9893	0.9903	0.9897	0.9897(1)	0.9874	0.9919	0.9898	0.9898(1)	0.9916	0.9903	0.9910	0.9910(1)
RUS	0.6460	0.9455	0.7796	0.7957(7)**	0.6989	0.9456	0.8117	0.8222(7)**	0.6956	0.9296	0.8025	0.8126(7)**
MChanUS	0.9692	1.0000	0.9845	0.9846(2)**	0.9682	1.0000	0.9840	0.9841(2)**	0.9601	1.0000	0.9799	0.9800(2)**
USOS	0.9593	0.9987	0.9787	0.9790(3)**	0.9550	0.9961	0.9753	0.9755(3)**	0.9584	0.9987	0.9783	0.9786(3)**
RUSSMOTE	0.8177	0.8864	0.8420	0.8520(6)**	0.7355	0.8802	0.7169	0.7174(8)**	0.8074	0.8837	0.8328	0.8455(6)**

Note: **statistically significant at $p < 0.05$, *at $p < 0.10$ using the Wilcoxon signed-rank test

Table 8 Global results of WSMOTE, WSMOTE-ensemble, MChanUS, USOS, and their baseline samplings

Instance composition	Improvement for TP_{rate} (%)	Improvement for TN_{rate} (%)	Improvement for AUC (%)	Improvement for G_m (%)
WSMOTE-ensemble	–	0.51	–	–
WSMOTE	4.11***	–	1.78***	1.89***
SMOTE	7.37***	0.31	3.52***	3.74***
MChanUS	3.89***	1.03	2.21***	2.21***
USOS	9.19***	0.2040	4.18**	4.52**
RUSSMOTE	23.02 *	18.84***	22.14***	26.43***
RUS	57.97***	5.81***	26.27***	30.16***

Note: ***statistically significant at $p < 0.01$, **at $p < 0.05$, *at $p < 0.10$ using the Wilcoxon signed-rank test

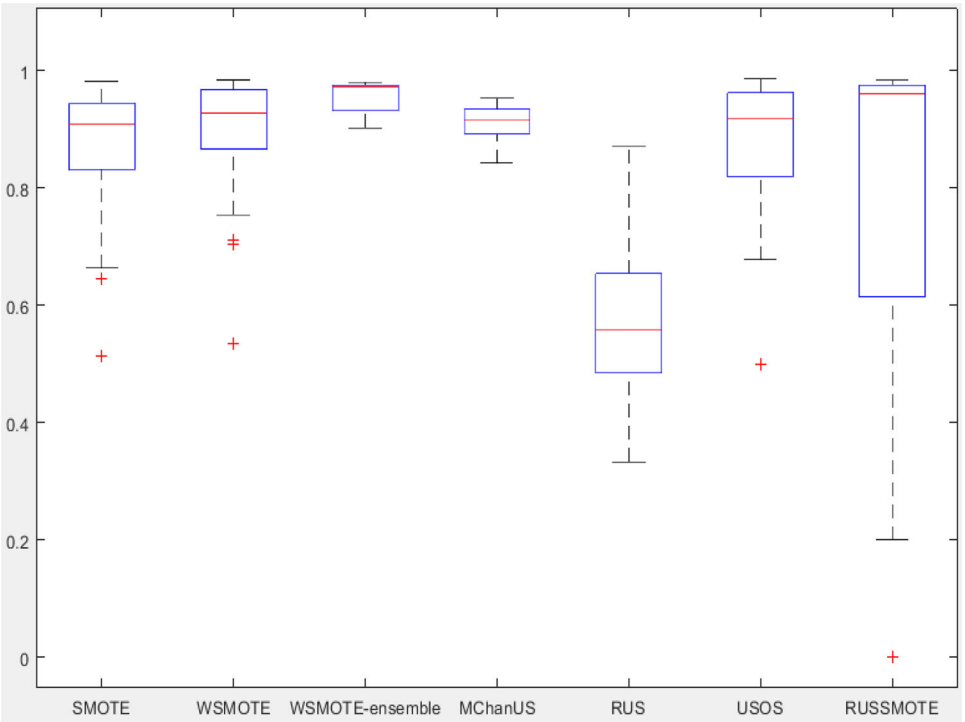

Fig. 3 Boxplot chart of global performance of different sampling strategies in minority class accuracy

from the weighted cluster. Among the blending experiments, this study retained one delegate technique of each from the Bagging-based, Boosting-based, LB-based, RC-based, RTF-based, and RF-based ensemble solutions. Altogether, 14 representative approaches were chosen for the global comparison, which provided a concise analysis. Table 9 reveals the outcome of the global assessment, in which AUC ranks were established for the various instance composition mixtures.

Table 9 demonstrates that the ensemble solutions and data-level strategies mostly dominated the results. More specifically, the RF algorithm integrated with WSMOTE-ensemble secured the top position (P#1). The blending solutions with RC learner ranked second and third (P#2, P#3); in these, WSMOTE and WSMOTE-ensemble were used as the representative algorithms. Parallel to the best algorithms, hybrid sampling, that is to say, WSMOTE-ensemble-RTF, took the fourth position (P#4). Consequently, these results indicate that blending sampling seems to be more robust than its peers. Furthermore, the blending solutions seemed to be more effective than data-level treatment because the former methods secured positions P#1, P#2, P#3, and P#4 but the data-level treatment secured positions P#5, P#7, P#8, and P#10 in the Friedman ranks. Many related studies have proven that hybrid treatment consistently attains better rankings than corresponding sampling editions [49,66]. However, the k-NN-based data-level technique was first (P#5) relative to C4.5 (P#7) and SVM (P#8 and P#10). The Bagging-based ensemble was preferable to the Boosting-based ensemble as indicated by their positions at P#6 and P#11, respectively. In contrast, cost-sensitive learners attained the lowest positions.

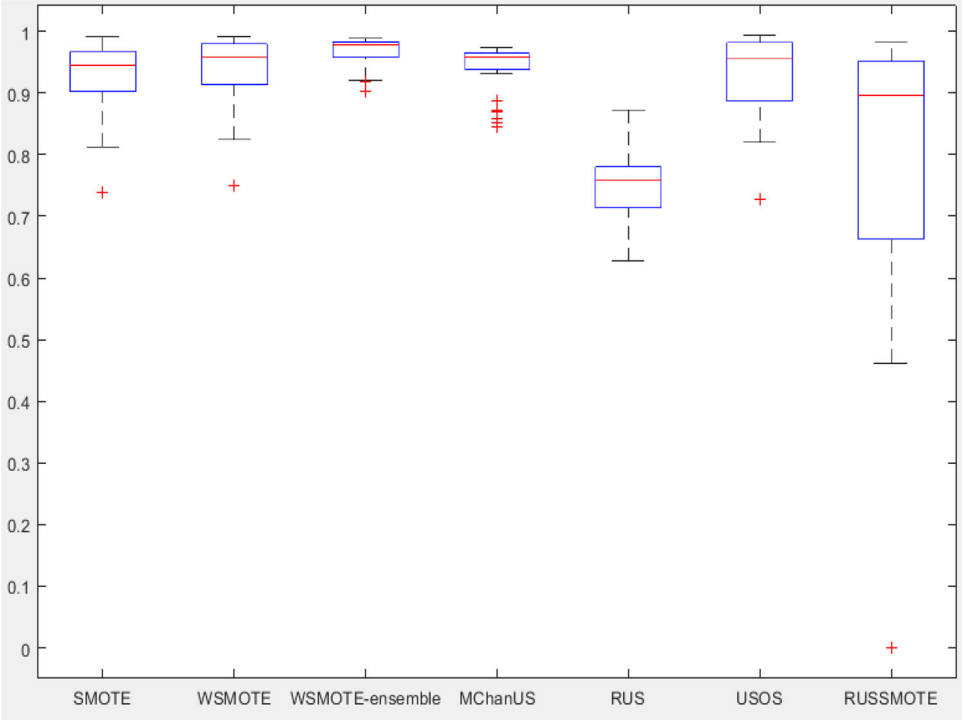

Fig. 4 Boxplot chart of AUC from the global performance of different sampling strategies

Table 9 Friedman global ranking of the top-performing classifiers for the instance composition mixtures in terms of AUC

Instance composition	Aver. rank (#Position)	Holm's p-value
WSMOTE-ensemble-RF	1.9286 (#1)	–
WSMOTERC	2.1429 (#2)	0.0500**
WSMOTE-ensemble-RC	3.0714 (#3)	0.0250**
WSMOTE-ensemble-RTF	3.1429 (#4)	0.0167**
WSMOTE-ensemble-k-NN	4.8571 (#5)	0.0125**
WSMOTE-ensemble-Bagging	5.8571 (#6)	0.0789*
WSMOTE-ensemble-C4.5	7.0000 (#7)	0.0083***
WSMOTE-ensemble-SVM	8.4286 (#8)	0.0037***
WSMOTE-ensemble-LB	9.4286 (#9)	0.0008***
SMOTE SVM	9.8571 (#10)	0.0056***
WSMOTE-ensemble-Boosting	10.4286 (#11)	0.0050***
Weighted SVM	12.7143 (#12)	0.0045***
Weighted C4.5	13.0000 (#13)	0.0042***
Weighted k-NN	13.1429 (#14)	0.0038***
Iman-Davenport p-value		0.0000***

Note: ***statistically significant at $p < 0.01$, **at $p < 0.05$, *at $p < 0.10$

Table 9 also shows the results of the Iman-Davenport test (derived from Friedman's test) and Holm post-hoc analysis. The Iman-Davenport test was performed for multiple instance composition assessments based on the mean AUC ranks. The result of this test indicates significant differences between the compared instance composition methods. To further examine the differences in AUC performance, the Holm post-hoc test was carried out. Table 9 suggests that the outstanding recital of hybrid learning (WSMOTE-ensemble-RF, P#1) was statistically significant because the Iman-Davenport test p-value was $<.01$, for which the null hypothesis was rejected. Furthermore, the Holm test showed that the performance of the top-ranked sampling strategy was statistically significant. Therefore, the significant finding was that the novel WSMOTE-ensemble sampling strategy was the best solution for credit risk prediction in class-imbalanced small business data.

Conclusion

The assessment of small business loans is important in credit risk modeling due to the irregular availability of information and the variability of time constraints. Small enterprises are subject to minimal legal requirements for data disclosure, and it is difficult for commercial banks to get detailed information about them. For this reason, commercial banks may try to base their decisions on the historical credit records of small enterprises filed with the credit bureau data manager.

From the many experimental studies available, we determined that it is not possible to assert the dominance of an algorithm over other competing classifiers irrespective of data traits. Moreover, in the real world, most of the small enterprises are creditworthy, and only a minority are noncreditworthy or default businesses; in other words, there is an imbalanced class distribution. Recently, modelers have tried to optimize the forecasting performance in skewed class scenarios. One commonly employed methodology is the oversampling of minority data points that are recognized as data-level solutions, which modify the class allocation in a given dataset. Therefore, the current investigation was mostly focused on oversampling methods. By assigning weights to new artificial data, the representative data points produced more new instances compared with their outlying counterparts, and thus, a more stable solution was obtained. The current study also proposed a novel ensemble approach rooted in the WSMOTE algorithm, or WSMOTE-ensemble for skewed data modeling. The proposed ensemble classifier hybridizes WSMOTE and Bagging with SCMs to minimize the class skewed constraints linked to default and nondefault instances.

The main empirical findings of this study can be summarized as follows:

- Experimental investigations point out that the WSMOTE-ensemble-RF blending procedure significantly outperformed the compared methods on all grounds. The WSMOTE-ensemble outstripped any other algorithmic combinations being trained. Moreover, the recitals of WSMOTE with RC also outperformed the existing SMOTE, MChanUS, and USOS sampling strategies. As a result, it is imperative to apply the WSMOTE-ensemble and WSMOTE class balancing modules in predicting the small business credit risk.
- Sampling methods outperformed nonsampling algorithms. From three base learners, k-NN with WSMOTE-ensemble was the best blending approach to a sampling strategy. Simultaneously, SVM did not seem to benefit from resampling, but it also outperformed the nonsampling strategy.
- The WSMOTE-ensemble-RF fusion approach is the preferable choice for highly imbalanced credit risk modeling. Consistent with this finding, WSMOTE-ensemble-Bagging is the second-best choice, especially in the modes of lower dimensionality and highly informative independent variables.
- Weighted algorithms can also be a feasible alternative for predicting the credit risk of small enterprises in imbalanced scenarios, with WSVM as the best choice.

With these contributions, therefore, this study fills a noteworthy knowledge gap and adds several unique insights to the literature. It has established multilevel experimental settings in a rare event domain, and produced an improved ensemble solution with $TP_{rate} = 99.16\%$, this is with 15.16% perfection in default instance classification relative to the existing decision tree ensemble solution of Sun et al. [61]. The significant improvement in performance could generate substantial savings for the financial industry and may have a huge significance in a range of financial and nonfinancial decisions. It appears extremely sensible to minimize the skewed class problem using the optimized ensemble algorithm before designing the prediction classifier. It also provides a possible clarification for the practice-oriented need of interpretability of trained algorithms without losing the accuracy of risk appraisal, which would streamline its adoption as a managerial tool by industrial organizations and users. This study will aid small business loan lenders in protecting themselves from potential borrowers with a high credit risk. The executed database can be considered a small-scale example set, although 43 sample sets have been generated by applying the multi-level data-preprocessing criteria. The current study used multi-level experimental settings applying RF, data-level samplings, and algorithmic modifications. In future research, the application of different learning techniques such as evolutionary algorithms, extreme gradient boosting, and deep neural networks may provide more diverse experimental settings. Keeping in mind its limitations, therefore, future work should focus on developing the proposed assessment methodology for small business credit risk over multiple databases in more complex experimental scenarios.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40747-021-00614-4.

Acknowledgements This work has been supported by the Key Projects of National Natural Science Foundation of China (71731003 and 71431002), the General Projects of National Natural Science Foundation of China (71471027 and 71873103), the National Social Science Foundation of China (16BTJ017), the Youth Project of National Natural Science Foundation of China (71601041), the scientific research project of the Czech Sciences Foundation Grant (19-15498S), the Aderi Intel-

ligent Technology (Xiamen) Co and Bank of Dalian as well as Postal Savings Bank of China. We thank the organizations mentioned above.

Declarations

Conflict of interest The authors have declared there is no competing interests exist.

References

1. Abedin MZ, Guotai C, Moula FE (2019c) Weighted SMOTE-ensemble algorithms: evidence from Chinese imbalance credit approval instances. In: 2019 2nd International Conference on Data Intelligence and Security (ICDIS), IEEE, pp 208–211
2. Abedin MZ, Guotai C, Colombage S, Moula FE (2018) Credit default prediction using a support vector machine and a probabilistic neural network. J Credit Risk 14(2):1–27
3. Abedin MZ, Guotai C, Moula F, Azad AS, Khan MSU (2019) Topological applications of multilayer perceptrons and support vectormachines in financial decision support systems. Int J Finance Econ 24(1):474–507
4. Abedin MZ, Guotai C, Moula FE, Zhang T, Hassan MK (2019) An optimized support vector machine intelligent technique using optimized feature selection methods: evidence from Chinese credit approval data. J Risk Model Valid 13(2):1–46
5. Abedin MZ, Chi G, Uddin MM, Satu MS, Khan MI, Hajek P (2020) Tax default prediction using feature transformation-based machine learning. IEEE Access 9:19864–19881
6. Agostino M, Gagliardi F, Trivieri F (2012) Bank competition, lending relationships and firm default risk: an investigation of Italian SMEs. Int Small Bus J 30(8):907–943
7. Altman EI, Sabato G (2007) Modelling credit risk for SMEs: evidence from the US market. Abacus 43(3):332–357
8. Antunes F, Ribeiro B, Pereira F (2017) Probabilistic modeling and visualization for bankruptcy prediction. Appl Soft Comput 60:831–843
9. Arcuri G, Levratto N (2020) Early stage SME bankruptcy: Does the local banking market matter? Small Bus Econ 54(2):421–436
10. Behr P, Güttler A (2007) Credit risk assessment and relationship lending: an empirical analysis of German small and medium-sized enterprises. J Small Bus Manage 45(2):194–213
11. Brown I, Mues C (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst Appl 39(3):3446–3453
12. Calabrese R, Marra G, Osmetti SA (2016) Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. J Oper Res Soc 67(4):604–615
13. Carmona P, Climent F, Momparler A (2019) Predicting failure in the U.S. banking sector: an extreme gradient boosting approach. Int Rev Econ Finance 61:304–323
14. Chan PK, Stolfo SJ (1998) Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proc. Fourth Int. Conf. on Knowledge Discovery and Data Mining, pp 164–168
15. Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTE-Boost: Improving prediction of the minority class in boosting. Lecture Notes Artif Intell 2838:107–119
16. Chen L, Zhou Y, Zhou D, Xue L (2017) Clustering enterprises intoeco-industrial parks: Can interfirm alliances help small and medium-sizedenterprises? J Clean Prod 168:1070–1079
17. Ciampi F (2015) Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. J Bus Res 68(5):1012–1025
18. Ciampi F (2017) The need for specific modelling of small enterprise default prediction: empirical evidence from Italian small manufacturing firms. Int J Bus Manag 12(12):251–262
19. Ciampi F, Gordini N (2013) Small enterprise default prediction modeling through artificial neural networks: an empirical analysis of Italian small enterprises. J Small Bus Manage 51(1):23–45
20. Ciampi F, Giannozzi A, Marzi G, Altman EI (2021) Rethinking SME default prediction: a systematic literature review and future perspectives. Scientometrics 1–48
21. Duarte FD, Gama APM, Gulamhussen MA (2018) Defaults in bankloans to SMEs during the financial crisis. Small Bus Econ 51(3):591–608
22. Edmister RO (1972) An empirical test of financial ratio analysis for small business failure prediction. J Financ Quant Anal 7(2):1477–1493
23. Figini S, Bonelli F, Giovannini E (2017) Solvency prediction for small and medium enterprises in banking. Decis Support Syst 102:91–97
24. Florez-Lopez R, Ramon-Jeronimo JM (2015) Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. Expert Syst Appl 42:5737–5753
25. García S, Fernández A, Luengo J, Herrera F (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf Sci 180(10):2044–2064
26. Gicic A, Subasi A (2019) Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. Expert Syst 36(2):1–22
27. Guotai C, Abedin MZ, Moula FE (2017) Modeling credit approval data with neural networks: an experimental investigation and optimization. J Bus Econ Manage 18(2):224–240
28. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
29. Hajek P, Abedin MZ (2020) A profit function-maximizing inventory backorder prediction system using big data analytics. IEEE Access 8:58982–58994
30. Hajek P, Michalak K (2013) Feature selection in corporate credit rating prediction. Knowl-Based Syst 51:72–84
31. Hasumi R, Hirata H (2014) Small business credit scoring and its pitfalls: evidence from Japan. J Small Bus Manage 52(3):555–568
32. He H, Zhang W, Zhang S (2018) A novel ensemble method for credit scoring: adaption of different imbalance ratios. Expert Syst Appl 98:105–117
33. Hernandez MA, Torero M (2014) Parametric versus nonparametric methods in risk scoring: an application to microcredit. Empir Econ 46(3):1057–1079

34. Inekwe JN (2019) Lending risk in MFIs: the extreme bounds of microeco-nomic and macroeconomic factors. J Small Bus Manage 57(2):538–558
35. Keasey K, Pindado J, Rodrigues L (2015) The determinants of the costs of financial distress in SMEs. Int Small Bus J 33(8):862–881
36. Kim MJ, Kang DK, Kim HB (2015) Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Syst Appl 42(3):1074–1082
37. Lessmann S, Baesens B, Seow HV, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur J Oper Res 247(1):124–136
38. Lin SM, Ansell J, Andreeva G (2012) Predicting default of a small business using different definitions of financial distress. J Oper Res Soc 63(4):539–548
39. Louzada F, Ferreira-Silva PH, Diniz CA (2012) On the impact of disproportional samples in credit scoring models: an application to a Brazilian bank data. Expert Syst Appl 39(9):8071–8078
40. Maldonado S, Weber R, Famili F (2014) Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Inf Sci 286:228–246
41. Marqués AI, García V, Sánchez JS (2013) On the suitability of resampling techniques for the class imbalance problem in credit scoring. J Oper Res Soc 64(7):1060–1070
42. Mayr S, Mitter C, Aichmayr A (2017) Corporate crisis and sustainable reorganization: evidence from bankrupt Austrian SMEs. J Small Bus Manage 55(1):108–127
43. Medina-Olivares V, Calabrese R, Dong Y, Shi B (2021) Spatial dependence in microfinance credit default. Int J Forecast. https://doi.org/10.1016/j.ijforecast.2021.05.009
44. Ministry of Industry and Information Technology (2011) Standard type division for middle and small-sized enterprises. National Bureau of Statistics, National Development Reform Commission. Ministry of Finance, P.R. China., Technical Report
45. Moula FE, Guotai C, Abedin MZ (2017) Credit default prediction modeling: an application of support vector machine. Risk Manage 19(2):158–187
46. Niu K, Zhang Z, Liu Y, Li R (2020) Resampling ensemble model basedon data distribution for imbalanced credit risk evaluation in P2P lending. Inf Sci 536:120–134
47. OECD (2019) Financing SMEs and entrepreneurs 2019: an OECD scoreboard. Organisation for Economic Co-operation and Development OECD, Paris
48. Papouskova M, Hajek P (2019) Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decis Support Syst 118:33–45
49. Peng L, Zhang H, Yang B, Chen Y (2014) A new approach for imbalanced data classification based on data gravitation. Inf Sci 288:347–373
50. Pindado J, Rodrigues LF (2004) Parsimonious models of financial insolvency in small companies. Small Bus Econ 22(1):51–66
51. Piri S, Delen D, Liu T (2018) A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. Decis Support Syst 106:15–29
52. Prusty MR, Jayanthi T, Velusamy K (2017) Weighted-SMOTE: a modification to SMOTE for event classification in sodium cooled fast reactors. Prog Nucl Energy 100:355-364
53. Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann
54. Rio SD, Lopez V, Benitez JM, Herrera F (2014) On the use of MapReduce for imbalanced big data using random forest. Inf Sci 285:112–137
55. Rivera WA (2017) Noise reduction a priori synthetic over-sampling for class imbalanced data sets. Inf Sci 408:146–161
56. Rostamkalaei A, Freel M (2016) The cost of growth: small firms and the pricing of bank loans. Small Bus Econ 46(2):255–272
57. Shi B, Chi G, Li W (2020) Exploring the mismatch between credit ratings and loss-given-default: a credit risk approach. Econ Model 85:420–428
58. Sohn Y, Jeon H (2010) Competing risk model for technology credit fund for small and medium-sized enterprises. J Small Bus Manage 48(3):378–394
59. Sun Y, Wong AK, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 23(4):687–719
60. Sun J, Lee YC, Li H, Huang QH (2015) Combining B&B-based hybrid feature selection and the imbalance-oriented multiple-classifier ensemble for imbalanced credit risk assessment. Technol Econ Dev Econ 21(3):351–378
61. Sun J, Lang J, Fujita H, Li H (2018) Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf Sci 425:76–91
62. Sun J, Li H, Fujita H, Fu B, Ai W (2020) Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Inf Fusion 54:128–144
63. Susan S, Kumar A (2019) SSOMaj-SMOTE-SSOMin: three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets. Appl Soft Comput 78:141–149
64. Tian S, Yu Y (2017) Financial ratios and bankruptcy predictions: an international evidence. Int Rev Econ Finance 51:510–526
65. Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng 14(3):659–665
66. Zhu B, Baesens B, Seppe KLM, Broucke V (2017) An empirical comparison of techniques for the class imbalance problem in churn prediction. Inf Sci 408:84–99
67. Zhu Y, Zhou L, Xie C, Wang GJ, Nguyen TV (2019) Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. Int J Prod Econ 211:22–33