Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 3526539, 13 pages
https://doi.org/10.1155/2019/3526539

Research Article

An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data

Wenhao Xie,[1,2] **Gongqian Liang**,[1] **Zhonghui Dong**,[3] **Baoyu Tan**,[4] **and Baosheng Zhang**[5]

[1]*School of Management, Northwestern Polytechnical University, 710129, China*
[2]*School of Science, Xi'an Shiyou University, 710065, China*
[3]*School of Economics and Management, Xi'an Shiyou University, 710065, China*
[4]*School of Computer Science, Xi'an Shiyou University, 710065, China*
[5]*Management Institute, Harbin Normal University, 150000, China*

Correspondence should be addressed to Wenhao Xie; xwhaoxwhao@163.com

Received 7 February 2019; Revised 6 April 2019; Accepted 21 April 2019; Published 6 May 2019

Academic Editor: Piotr Jędrzejowicz

The imbalance data refers to at least one of its classes which is usually outnumbered by the other classes. The imbalanced data sets exist widely in the real world, and the classification for them has become one of the hottest issues in the field of data mining. At present, the classification solutions for imbalanced data sets are mainly based on the algorithm-level and the data-level. On the data-level, both oversampling strategies and undersampling strategies are used to realize the data balance via data reconstruction. SMOTE and Random-SMOTE are two classic oversampling algorithms, but they still possess the drawbacks such as blind interpolation and fuzzy class boundaries. In this paper, an improved oversampling algorithm based on the samples' selection strategy for the imbalanced data classification is proposed. On the basis of the Random-SMOTE algorithm, the support vectors (SV) are extracted and are treated as the parent samples to synthesize the new examples for the minority class in order to realize the balance of the data. Lastly, the imbalanced data sets are classified with the SVM classification algorithm. F-measure value, G-mean value, ROC curve, and AUC value are selected as the performance evaluation indexes. Experimental results show that this improved algorithm demonstrates a good classification performance for the imbalanced data sets.

1. Introduction

1.1. Research Background. Data classification is one of the most common data mining techniques [1], whose task is to determine which target class some specific object in an unknown class belongs to. There are many classic classification algorithms of data mining [2–10]. However, each classification algorithm has its advantages and disadvantages. The effect of a classification algorithm is usually related to the characteristics of data. Among the many classification algorithms, SVM (support vector machine) classification algorithm, which is proposed by Vapnik et al. and is based on the structural risk minimization principle of the SLT, is widely used [11–16]. SVM shows many unique advantages in solving some classification problems such as small sample numbers, nonlinearity, and high-dimensional pattern recognition problems. The traditional SVM classification algorithm assumes two ideas: the size of the training examples of different classes in the same data set is balanced and the cost of classification error is almost similar [2]. However, in the real world, there are many imbalanced data classification problems. In addition, the classification accuracy of the minority class is often more valuable. As mentioned above, the classification model of the SVM is based on a minimization of structural risk. Therefore, for the imbalanced data, the examples in the majority class will have a greater influence on the classifier, causing its classification weight to be in favour of the majority class and then seriously affecting the distribution of classification hyperplane. So it is very important that the classification approaches can be improved at the algorithm-level or data-level to solve the imbalanced data classification, which is currently a trending issue in the data mining research field.

1.2. Related Research. In the field of data classification [17–19], the classification of imbalanced data has been the focus of attention. In recent years, the classification algorithms based on imbalanced data have attracted many researchers. The international conferences on the classification of imbalanced data have been held around the world [20–22].

At present, research approaches regarding the classification of imbalanced data by the SVM mainly divide into two categories: the improving approaches at the algorithm-level and the improving approaches at the data-level.

At the algorithm-level, the weighted SVM of the penalty coefficient C is used to control the different costs of misclassification error for the different classes. In general, a higher misclassification cost is imposed on the minority class, and a low misclassification cost is imposed on the majority class. In addition, AdaBoost algorithm—the integrating algorithm based on multiple single classifiers—and the improving algorithm based on kernel space are used widely.

At the data-level, there are two main approaches: oversampling strategy for the minority samples and undersampling strategy for the majority samples.

Oversampling technique uses some approaches, such as duplicating the minority examples or artificially synthesizing new examples from the minority class with some algorithms, to balance the classes' distribution. Chawla [23] put forward the SMOTE algorithm, which is used to synthesize the minority class examples by a linear interpolation method. The Borderline-SMOTE algorithm proposed by Han et al. [24] and the Random-SMOTE algorithm proposed by Dong [25] improved the SMOTE algorithm from different angles. Kubat and Matwin [26] discussed the mining approaches towards the samples in the imbalanced data sets. In addition, other sampling approaches [27, 28] are also classic approaches for mining minority samples.

Besides oversampling algorithm, undersampling algorithm is also a common method to deal with the imbalanced data sets. Undersampling technique mainly balances the distribution of data classes by deleting the majority class examples such as the Tomek Links algorithm [29] proposed by Wilson et al., neighbor compression rule proposed by Hart [30], and GSVM-RU algorithm proposed by Tang YC et al. [31].

2. Support Vector Classification and the Optimal Classification Hyperplane

A given set includes l samples: $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)\}$, in which $x_i \in R^n$, $y \in (-1, +1)$ $(i = 1, 2 \cdots l)$. If all the examples in D can be separated exactly by the hyperplane $wx + b = 0$ and the distance from the nearest sample point to the hyperplane is the maximum, we state that the data samples can be separated by the optimal hyperplane, which is also called the maximum margin hyperplane as shown in Figure 1.

The SVM method can also solve the classification problem of high-dimensional space. We introduce the mapping $\varphi : R_n \longrightarrow F$ and map the examples from the input space to the high-dimensional feature space while the optimal classification hyperplane is constructed in the high-dimensional feature space. In this way, the problem of the optimal classification hyperplane is transformed into the following optimization problem:

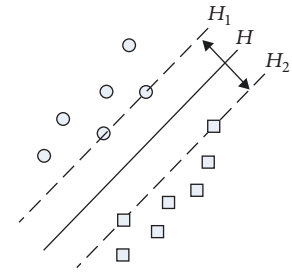

Figure 1: The maximum margin hyperplane.

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i\left(w \cdot \varphi\left(x_i\right) + b\right) - 1 + \xi_i \geq 0 \\ & \xi_i \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{1}$$

in which C is the penalty parameter, which controls the degree of penalty for misclassification samples. In addition, the greater the value of C, the greater the penalty for error.

The corresponding Lagrangian function is

$$\begin{aligned} L(w, b, \xi, \alpha) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ & - \sum_{i=1}^{l}\alpha_i\left(y_i\left(w \cdot \varphi\left(x_i\right) + b\right) - 1 + \xi_i\right) \\ & - \sum_{i=1}^{l}\beta_i\xi_i \end{aligned} \tag{2}$$

where α_i, β_i are *Lagrange* multipliers and $\alpha_i > 0$, $\beta_i > 0$. We can obtain the following dual problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j k\left(x_i, x_j\right) - \sum_{i=1}^{l}\alpha_i \\ s.t. \quad & \sum_{i=1}^{l}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, l \end{aligned} \tag{3}$$

where $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ is the kernel function.

3. Existing Algorithm Introduction

3.1. SMOTE. The oversampling technique is to increase the number of minority class samples by randomly copying the minority samples, so as to balance the examples size of the minority class and the majority class. SMOTE (synthetic minority oversampling technique) [32] is one of the most commonly used oversampling methods to solve the imbalance problems, which generates the synthetic training examples by linear interpolation for the minority class. These synthetic training examples are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and various classification techniques can be employed for the processed data. Its basic idea is as follows.

Step 1. Setting the minority class set A, for each $x \in A$, the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and every other sample in set A.

Step 2. The sampling rate N is set according to the imbalanced proportion. For each $x \in A$, N examples $x_1, x_2 \cdots x_N$ $(N \leq k)$ are randomly selected from its k-nearest neighbors, and they construct the set A_1.

Step 3. For each example $x_k \in A_1$ $(k = 1, 2 \cdots N)$, the following formula is used to generate a new example:

$$x_{new} = x + \text{rand}(0, 1) * \left\| x - x_k \right\| \tag{4}$$

in which rand(0, 1) represents the random number between 0 and 1.

3.2. Random-SMOTE. SMOTE is a very classic oversampling approach, but it still has some deficiencies. For example, this algorithm is prone to distribution marginalization, is unable to change the sparse distribution of the samples of the minority class [33], and interpolates blindly. In view of these problems, Dong [25] proposed an improved data oversampling algorithm, which is called Random-SMOTE algorithm. Random-SMOTE algorithm can generate new minority samples in a relatively wider region, which can effectively change the sparse distribution of the minority class samples, and largely avoiding overfitting. Its basic idea is as follows.

For each sample x of the minority class set A, two examples x_1, x_2 $(x_{1,2} \neq x)$ are randomly selected from the set A, then this method will form a triangle region with x, x_1, x_2 as vertices. The process of generating new examples for the minority class is shown in Figure 2.

Step 1. Random linear interpolation is performed between x_1 and x_2 to generate N temporary examples t_i $(i = 1, 2 \cdots N)$ according to the sampling rate N.

$$t_i = x_1 + \text{rand}(0, 1) * (x_2 - x_1), \quad i = 1, 2 \cdots N \tag{5}$$

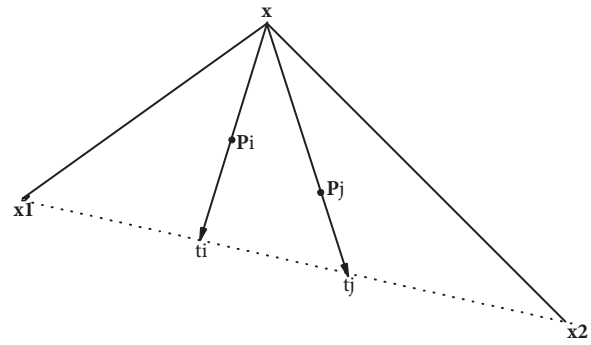

Figure 2: The synthetic examples method of the Random-SMOTE algorithm.

Step 2. Random linear interpolation is performed between t_i and x to generate N new minority class examples p_i $(i = 1, 2 \cdots N)$;

$$P_i = x + \text{rand}(0, 1) * (t_i - x) \quad i = 1, 2 \cdots N \tag{6}$$

3.3. The Contribution of Different Samples to the Classifier. Kubat et al. [26] proposed the one-side selection algorithm, which divided the samples into four types: safety samples, redundant samples, boundary samples, and noise samples by judging whether some example and its k-nearest neighbors belong to the same category. The traditional classifier has a higher recognition for the redundant samples and the safety samples. However, the boundary samples and the noise samples, which are on the junction of the two classes, tend to be confused. These examples are called "unsafe samples," and more attention should be paid to them for the classifier. Similarly, the SVM classification algorithm divides the samples into different types. Besides the noise and the redundant samples, SVM defines "safe samples" and "boundary samples" according to the values of Lagrange parameters α_i [34], in which the boundary samples are called "unsafe samples."

As is shown in Figure 3, the asterisk points represent the majority samples and the cross points represent the minority samples. The dotted circle points represent the boundary samples and the solid circle points represent the "overlapping sample points," which belong to the noise.

3.4. A Strategy of Samples' Selection for SVM Based on the Alien k-Neighbors. For the Random-SMOTE algorithm, it is not very reasonable to choose all the minority samples to be parent samples to synthesize the new examples because only the support vectors can contribute to the classification model in the SVM classification. In this way, a large number of redundant samples in the minority class will be synthesized, and this method will inevitably lead to the quality reduction of the training samples and increase training time of SVM. Hui Han et al. believed that minority samples closed to the boundary were more likely to be misclassified, so they had a greater impact on the classification results. Therefore, the Borderline-SMOTE method [24] is proposed and the synthetic examples are only generated by those samples that

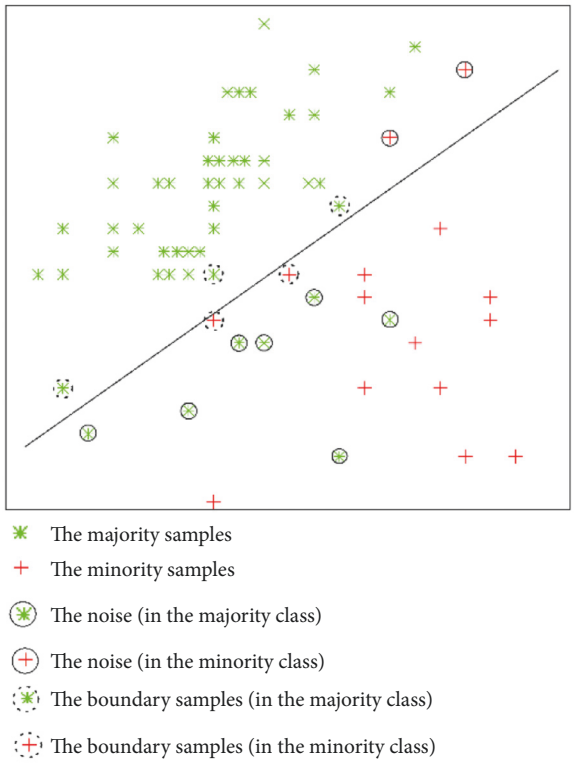

Figure 3: The different contribution of samples to the classifier.

are closed to the boundary according to this algorithm. Although this algorithm makes full use of valuable examples to synthesize the new examples, it still has some deficiencies. It judges whether some sample in the minority class belongs to the boundary samples only based on the number of most samples in the k nearest neighbors of this sample; this is not reasonable. So this method is likely to miss some valuable examples for the classification. Therefore, the key to sampling technology lies in what methods should be adopted that can retain the distribution information of the original data in order to obtain the most critical and important samples for sampling.

Chen et al. [35] proposed a samples selection method of SVM based on the alien k-neighbors. The idea of the algorithm is as follows.

The training data set is constituted via selecting the alien neighbor samples for each sample. Thus, the data set of the boundary samples is confirmed by the alien neighbors of every sample. For each sample, the algorithm can find its alien k-neighbors which are close to it. Finally, this algorithm finds the union set of all the alien neighbor samples, which contain all the boundary samples that are likely to be the support vectors.

This algorithm is different from the previous k-nearest neighbor algorithms. For each sample, this algorithm looks for the k-nearest neighbors in its alien class rather than its own class. This algorithm is not restricted by boundary conditions, but still has some shortcomings: (1) because this algorithm needs to calculate the k-nearest neighbors for all the samples belong to a certain class, it will generate a lot of redundancy and reduce the calculation speed, and (2) for the overlapping samples on the boundaries of the positive and negative classes, they are likely to be misclassified into the data set of boundary samples and be treated as the support vectors, and it will possibly reduce the generalization ability of the classifier.

4. The AKN-Random-SMOTE Algorithm

In order to avoid the deficiencies of the Random-SMOTE algorithm and the alien k-nearest mentioned above, we propose a new Random-SMOTE algorithm which is called AKN-Random-SMOTE algorithm in this paper. The support vectors are extracted by the improved alien k-neighbors algorithm, and the oversampling strategy is only performed to the boundary decision samples of the minority class rather than all the minority samples. This algorithm can reduce the computational overhead of the system and improve the classification accuracy and efficiency. Besides that, we remove the overlapping samples on the boundary of the two classes before classification, which are considered "noise samples" by us. In this way, this improved algorithm can overcome the defects of the fuzzy boundaries to some extent so that the generalization ability of the classifier can be further improved.

4.1. Deletion of the Overlapping Samples. Due to the limitations of data collection approaches, different types of data usually tend to have similar values on some attributes, which may cause the poor classification effect. These samples are located in the overlapping area of the attribute space, which are called "overlapping samples."

The classification errors of the classifier often occur in the boundary region of the two data sets. The boundary region of the two classes is such a region where the samples overlap. The situation will become more complicated especially for the imbalanced data [36]. At the same time, aiming for fixing the fuzzy boundary problem of the Random-SMOTE algorithm, we need to judge the overlapping examples before classification and then delete them so that it can reduce the probability of fuzzy boundary of the two classes and the negative influence of the noise for classification.

Setting set S, which is a set of overlapping samples; setting set A, which is the minority class; $x_i \in A$ $(i = 1, 2, \cdots n_A)$, in which n_A is the number of samples in A; setting $\overline{x}$, which is the center of the minority class A, this algorithm of deleting the overlapping samples of the minority class is as follows.

Step 1. Initialize the set S and the value of i: $S = \Phi$, $i = 1$, and calculate the center of the minority class $\overline{x} = (1/n_A) \sum_{i=1}^{n_A} x_i$.

Step 2. For each $x \in A$, judging whether all of the k-nearest neighbors of x belong to the majority class, if this condition is satisfied, x is added to the set S.

We will repeat Step 2 until all the examples in A have been searched. Then, we will judge whether S is null, if $S = \Phi$, this algorithm will end; otherwise, we will turn to Step 3.

Table 1: The information of the imbalanced data sets in the experiments.

Data set ID	Data sets	#Abb	#Attr	#Min	#Maj	IR	Data source
1	Banana	Banana	2	75	2808	0.03	KEEL
2	Haberman's Survival	Haberman	3	81	225	0.36	UCI
3	Bupa	Bupa	6	145	200	0.73	UCI
4	Appendicitis	Appendicitis	7	21	85	0.25	KEEL
5	Pima Indians Diabetes	Pima	8	268	500	0.54	KEEL
6	German Credit Data	German	20	300	700	0.43	UCI
7	Vehicle Silhouettes	Vehicle	18	199	647	0.31	KEEL
8	Led7digit	Led	7	52	448	0.12	UCI
9	Wisconsin	Wisconsin	9	241	458	0.53	UCI
10	Wine	Wine	13	48	130	0.37	UCI

Step 3. For every sample $s_i \in S \subset A$, we will do the following:

① Calculate the distance between s_i and the center $\overline{x}$ of set A, marked as d_i; calculate the distance between each y_j $(j = 1, 2 \cdots k)$ and the center $\overline{x}$ of set A, marked as l_j $(j = 1, 2, \cdots k)$, in which y_j belongs to the majority class as well as belongs to the k-nearest neighbors of s_i.

② For each d_i, if satisfying with $d_i < l_j$ $(j = 1, 2, \cdots k)$, we will delete s_i from S; otherwise, we will keep s_i in S.

We will repeat Step 3 until all the samples in S have been searched.

Step 4. S is the set of the overlapping samples of the minority class. Delete all the examples of the set S.

At last, we find the overlapping samples of the majority class in the same way and delete them.

4.2. The AKN-Random-SMOTE Algorithm. The sections above have introduced the drawback of the alien k-neighbors algorithm, which can generate a lot of redundancy. So this algorithm must be improved. In this paper, before extracting the support vectors, we first use the K-means algorithm to cluster the examples for the majority class and then extract the center of each cluster. Although the number of data sets after clustering is reduced, the spatial characteristics of the data are still reserved. Therefore, we only need to look for the alien k-neighbors for each cluster center. In this way, we can not only extract the support vectors but also avoid producing the redundant samples.

The steps of the AKN-Random-SMOTE algorithm are as follows.

Step 1. For the samples of the majority class, the K-means algorithm is used for clustering. We set the number of the clusters in the majority class as K_C. After clustering, the center of each cluster C_i $(i = 1, 2 \cdots K_C)$ of the majority class is extracted.

Step 2. For each cluster center C_i $(i = 1, 2 \cdots K_C)$, we calculate its k-nearest neighbor samples, which belong to the minority class. All the examples constitute set B, which is the set of the boundary decision samples.

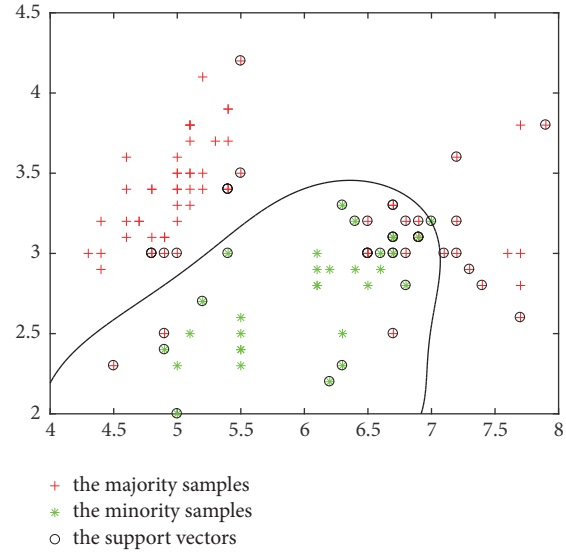

Figure 4: The original SVM classification effect without data processing.

Step 3. For each boundary decision sample of B, the Random-SMOTE algorithm is used to generate the new samples until the number of the two classes' samples is balanced.

Step 4. As mentioned in section above, we delete the overlapping samples for the two classes.

Step 5. The processed data is classified by SVM classification algorithm.

5. Experiments

5.1. Experimental Settings

5.1.1. Experimental Data Sets. In order to verify that this improved oversampling algorithm proposed in this paper has higher classification accuracy for the imbalanced data sets classification, we select the different algorithms for

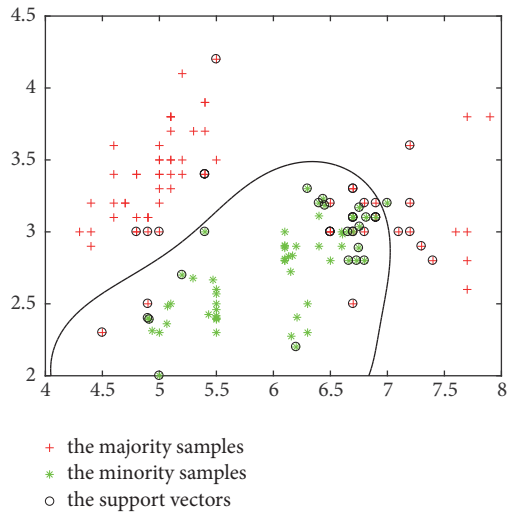

Figure 5: The classification effect processed by the SMOTE algorithm.

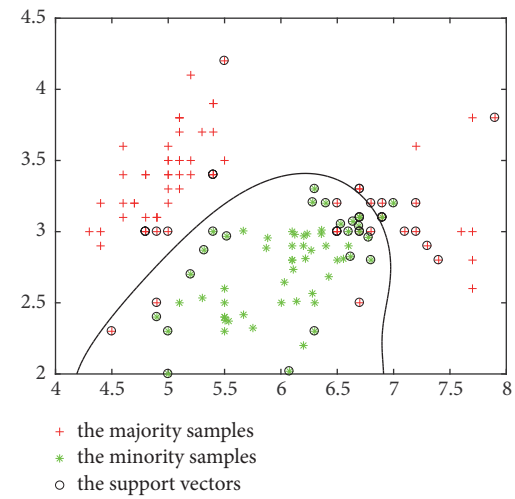

Figure 6: The classification effect processed by the Random-SMOTE algorithm.

comparison, and the experiments were carried out on 10 different imbalanced data sets. In this paper, the classification accuracy of these imbalanced data sets via different processing methods is given, and these processing methods include the original classification method without oversampling processing and the above three oversampling algorithms. The classification accuracy is valued according to the values of F-measure and G-mean.

In this study, we used 10 imbalanced data sets which are from UCI data sets and KEEL data sets; they have different sample sizes and attributes. In addition, they are also different in class imbalance ratio (IR). Table 1 summarizes the characteristics of the imbalanced data sets selected in the experiments, including the abbreviation of the data sets names (# Abb), the number of attributes of the examples (#Attr), the number of the minority examples (# Min), the number of the majority examples (# Maj), and the imbalance ratio (IR), in which IR=#Min/#Maj.

5.1.2. Parameters' Selection. In this paper, the AKN-Random-SMOTE algorithm needs to set the following parameters: K_1 and K_2, in which K_1 represents the number of clusters in the K-means algorithm, and K_2 represents the number of decision samples selected in the alien k-neighbors algorithm. In order to improve the applicability of this new algorithm, we try to set the different values of K_1 and K_2: $K_1 \in [2, 14]$, $K_2 \in [10, 200]$. In order to make the oversampling algorithms achieve their optimal classification effects and, at the same time, make the classification accuracy rate to be stable, we select the different values of the two parameters for every different data set, and they can be seen in Table 2. The value of the sampling rate is determined by the ratio of the number of the two class samples.

5.2. Performance Evaluation. The performance evaluation of the classification is an important reference for classification algorithms. The evaluation indexes, which are applied to

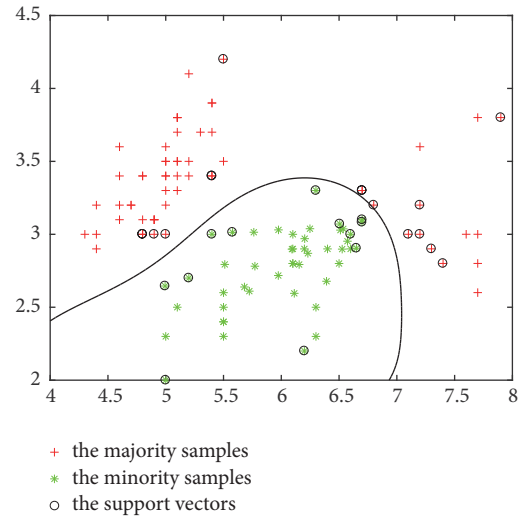

Figure 7: The classification effect processed by the AKN-Random-SMOTE algorithm.

the balanced data classification, are no longer suitable for the imbalanced data classification. The new classification performance evaluation indexes, such as precision, Recall, F-measure, and G-mean, are often used to measure the imbalanced classification performance [2].

Set the minority class as the positive class, and the majority class as the negative class. TP represents the number of the positive samples which are correctly classified by classifier; TN represents the number of the negative examples which are correctly classified by the classifier; FN represents the number of the positive examples which are wrongly classified by the classifier; FP represents the number of the negative examples which are wrongly classified by the classifier. The classification performance indexes for the minority class samples are as follows:

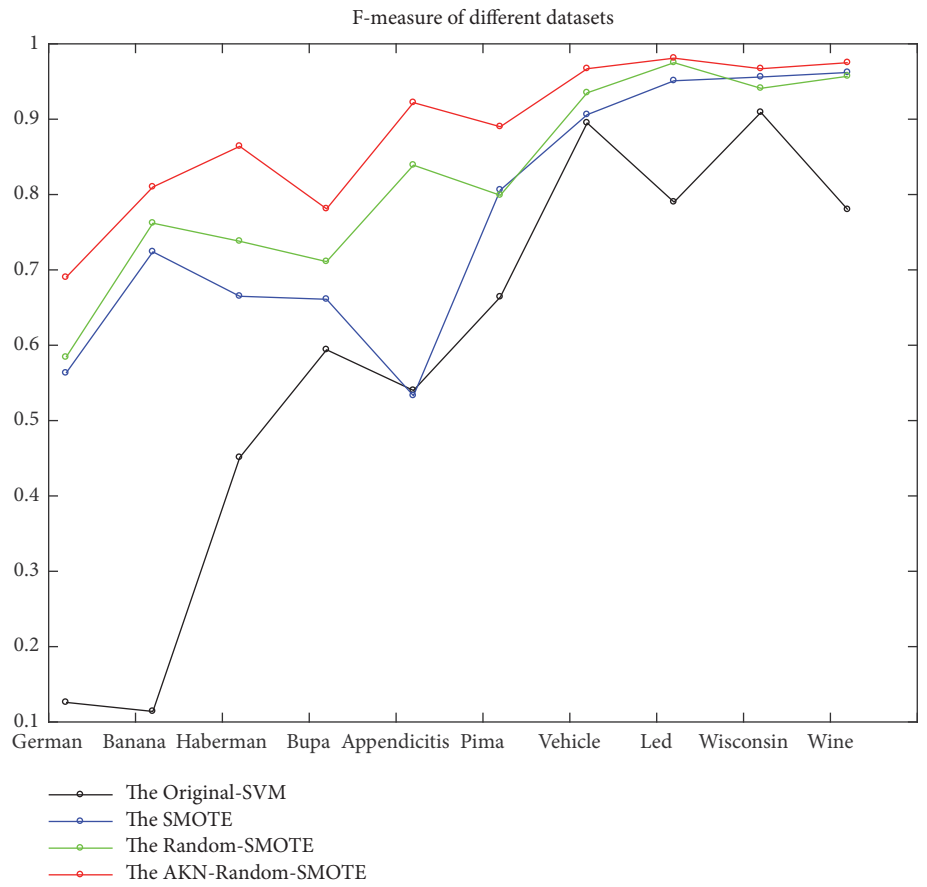

Figure 8: The polyline graph of F-measure values of the different methods.

Table 2: The values of the parameters.

Data set ID	Data sets	K_1	K_2
1	Banana	10	54
2	Haberman	14	62
3	Bupa	8	64
4	Appendicitis	5	10
5	Pima	10	109
6	German	2	200
7	Vehicle	9	86
8	Led	10	33
9	Wisconsin	5	92
10	Wine	6	30

(1) Classification precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

(2) Recall rate: the higher the value, the fewer number that the positive examples are wrongly classified by the classifier.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

(3) F-measure: this value is a classification evaluation index which considers both recall rate and precision rate.

$$F - measure = \frac{\left(1 + \beta^2\right) \times \text{precision} \times \text{Recall}}{\beta^2 \times \text{precision} + \text{Recall}} \tag{9}$$

$\beta \in [0, \infty)$, when $\beta < 1$; it emphasizes the effect of "precision rate"; when $\beta > 1$, it emphasizes the effect of "recall rate"; when $\beta = 1$, "precision rate" is as important as "recall rate"

Table 3: The comparison of classification results for the different methods.

Data set ID	Original-SVM		SMOTE		Random-SMOTE		AKN-Random-SMOTE	
	G-mean	F-measure	G-mean	F-measure	G-mean	F-measure	G-mean	F-measure
1	0.634	0.114	0.698	0.724	0.739	0.762	0.784	0.810
2	0.605	0.451	0.670	0.665	0.750	0.738	0.860	0.864
3	0.653	0.594	0.673	0.661	0.712	0.711	0.812	0.781
4	0.672	0.540	0.677	0.533	0.832	0.839	0.895	0.922
5	0.739	0.664	0.786	0.806	0.787	0.799	0.915	0.890
6	0.257	0.126	0.628	0.563	0.644	0.584	0.725	0.690
7	0.925	0.895	0.933	0.906	0.937	0.935	0.968	0.967
8	0.898	0.790	0.951	0.951	0.974	0.975	0.981	0.981
9	0.946	0.909	0.954	0.956	0.940	0.941	0.973	0.967
10	0.802	0.780	0.964	0.962	0.958	0.957	0.975	0.975

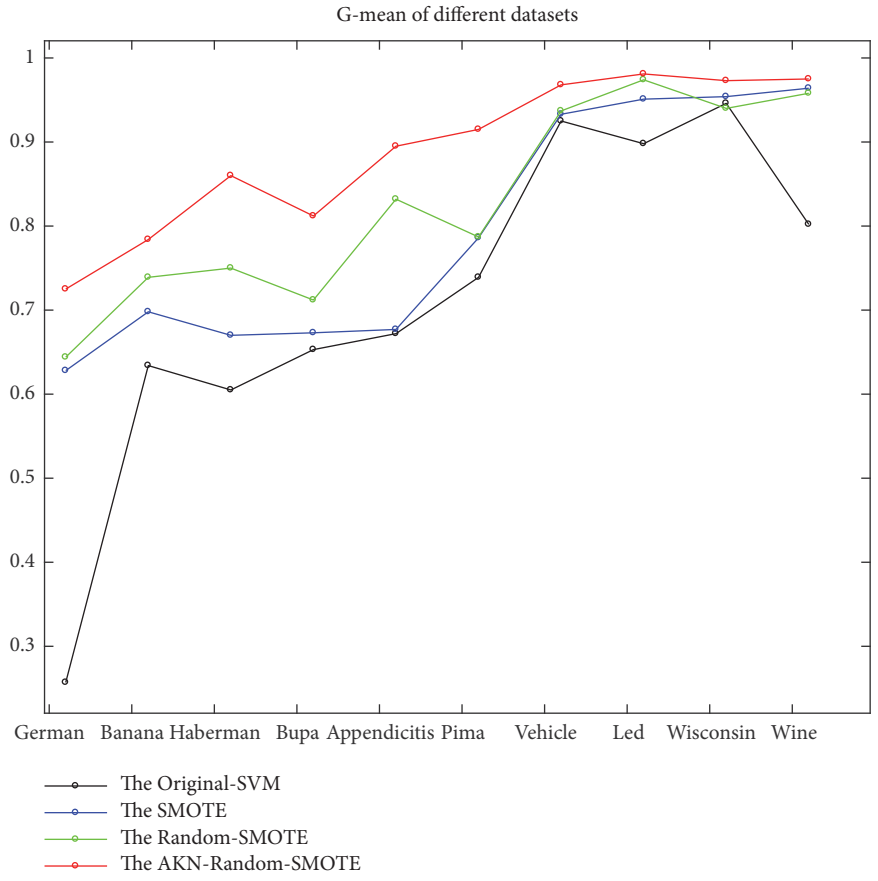

Figure 9: The polyline graph of G-mean values of the different methods.

for the classification evaluation, which is the commonly used indicator F_1:

$$F_1 = \frac{2\ \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})} \tag{10}$$

(4) G-mean: the geometric average of the classification precision of the minority class and the classification precision of the majority class.

$$G - mean = \sqrt{\frac{TP}{(\text{TP} + \text{FN})} \times \frac{TN}{(FP + TN)}} \tag{11}$$

F-measure and G-mean are two performance evaluation indexes that are usually used for the imbalanced data classification.

5.3. The Results and Analyses of the Experiments. For the imbalanced data sets mentioned above, we use the 10-fold cross validation algorithm to select the training set and the

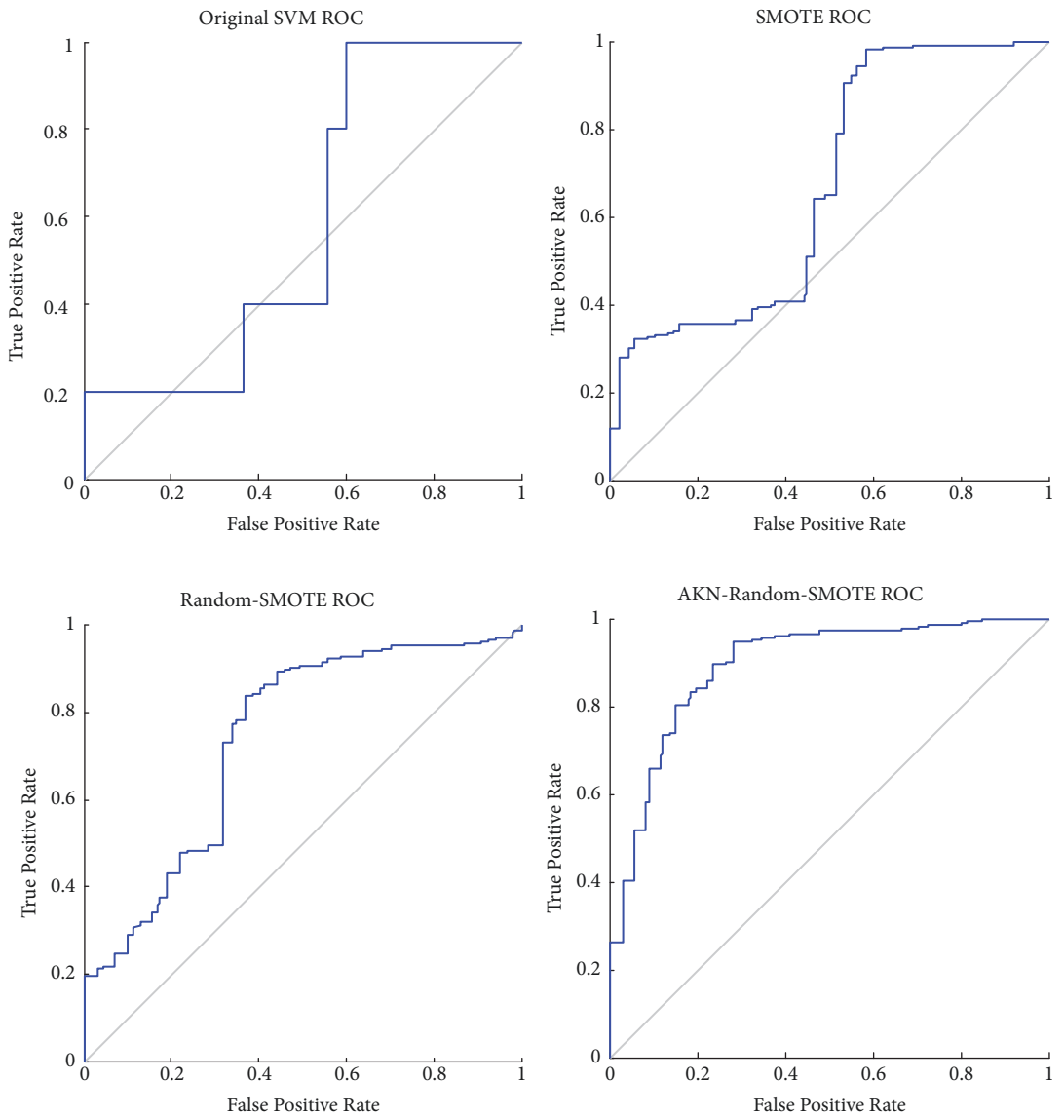

Figure 10: The ROC curves of the four methods for the Banana data set.

test set. A subset is selected as the test set each time and the average classification accuracy of the 10-times experiments is taken as a result by the cross validation algorithm.

In this paper, the SMOTE algorithm, the Random-SMOTE algorithm, and the AKN-Random-SMOTE algorithm are separately used to oversample the data in order to balance the data sets, and then the SVM classification algorithm is used to realize the classification. Finally, the classification results via the original SVM classification algorithm without data processing and the classification results via the above three oversampling algorithms processing are compared.

In order to show the classification effects lively, we first classify a two-dimensional imbalanced data set for the following situations: the data without processing and the data after the above three oversampling algorithms. The classification effects can be seen from Figures 4–7.

Secondly, we use the SMOTE algorithm, the Random-SMOTE algorithm, and the AKN-Random-SMOTE algorithm to reconstruct the data for the 10 imbalanced data sets. Finally, we use the SVM algorithm for the classification. In Table 3, "Original-SVM," "SMOTE," "Random-SMOTE," and "AKN-Random-SMOTE", respectively, represent the SVM classification results of the imbalanced data sets without data processing and the classification results processed after the three different oversampling algorithms. We take the F-measure values and the G-mean values as the classification performance evaluation indicators. We calculate the values many times for each algorithm and get the stable average value. The experimental comparison results are shown in Table 3.

From Table 3, we can see the AKN-Random-SMOTE algorithm proposed in this paper has higher G-mean values and F-measure values than other methods for the 10 data

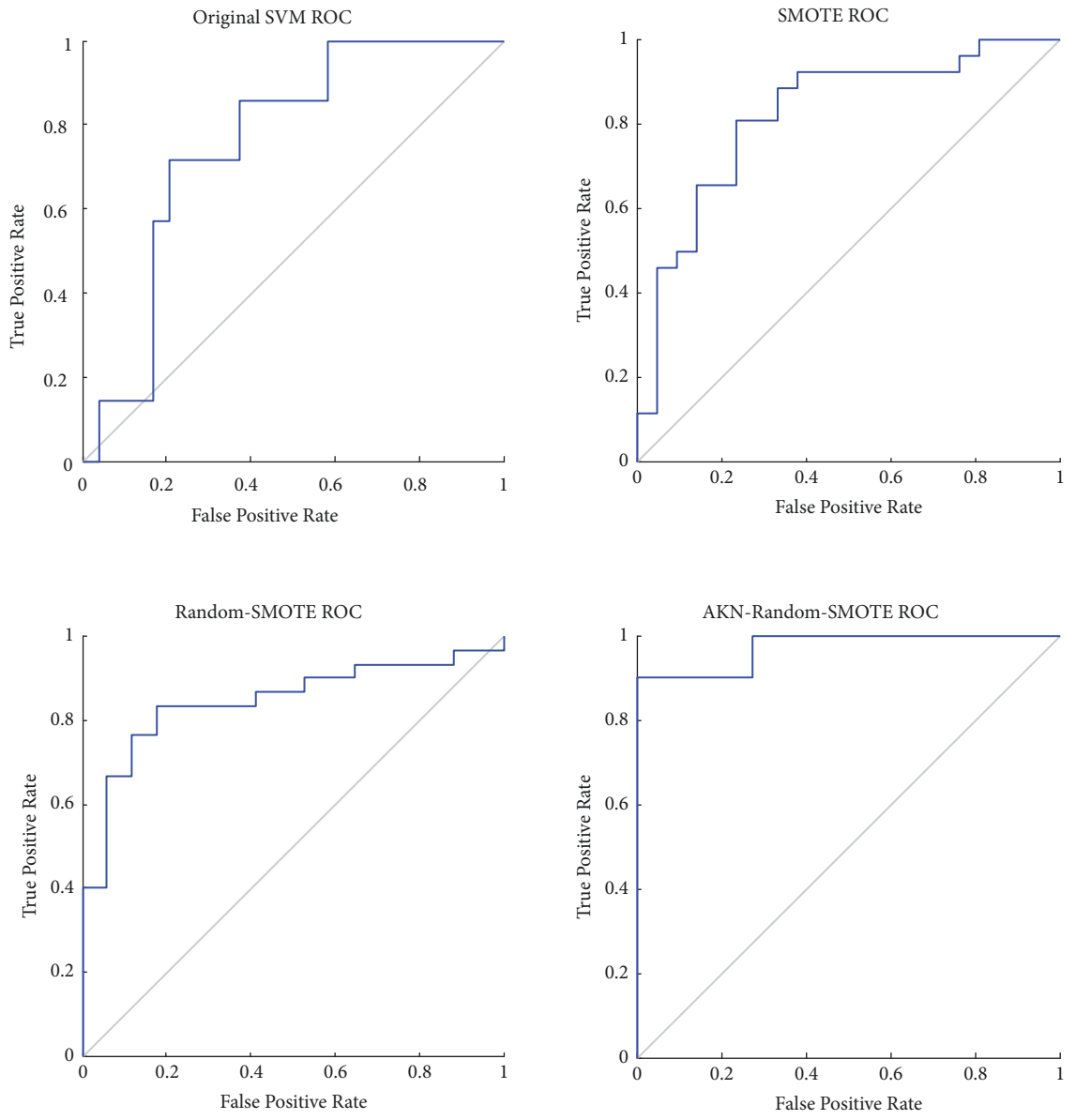

Figure 11: The ROC curves of the four methods for the Haberman data set.

sets. For the comparison purpose, the F-measure average values and the G-mean average values for the data sets were further given in Figures 8 and 9 in the form of polyline graphs.

Finally, we used the statistical tools, ROC curve (Receiver Operating Characteristic Curve) and AUC value (Area under ROC Curve), to evaluate the performance of the classification methods for the imbalanced data sets. ROC curve calculates a series of sensitivity and specificity by setting a number of different critical values for the continuous variables and then takes the sensitivity as the ordinate and the specificity as the abscissa to draw a curve. The closer the ROC curve is to the upper-left coordinate point (0, 1), i.e., the larger the value of the AUC is, and the higher the classification accuracy of the classifier will be. In general, the value of AUC is between 0.5 and 1. When AUC>0.5, the closer the value of AUC is to 1, the better the classification accuracy is. We generally believe that (1) when 0.5<AUC<0.7, the classification accuracy is low; (2) when 0.7<AUC<0.9, the classifier has a certain accuracy; (3) when AUC> 0.9, the classification accuracy is higher.

Because ROC curve has a good advantage that it does not change greatly with the change of the positive and the negative samples' distribution, ROC curve and AUC value are often used to evaluate the classification effect for the imbalanced data.

Using the data sets, Banana, Haberman, Pima, and Led as examples, we have plotted the ROC curves for the four different data processing methods (see Figures 10–13) and calculated the AUC values (see Table 4) of the ROC curves. It can be seen from the ROC curves and AUC values that the AKN-Random-SMOTE algorithm proposed in this paper has better classification effect than other algorithms.

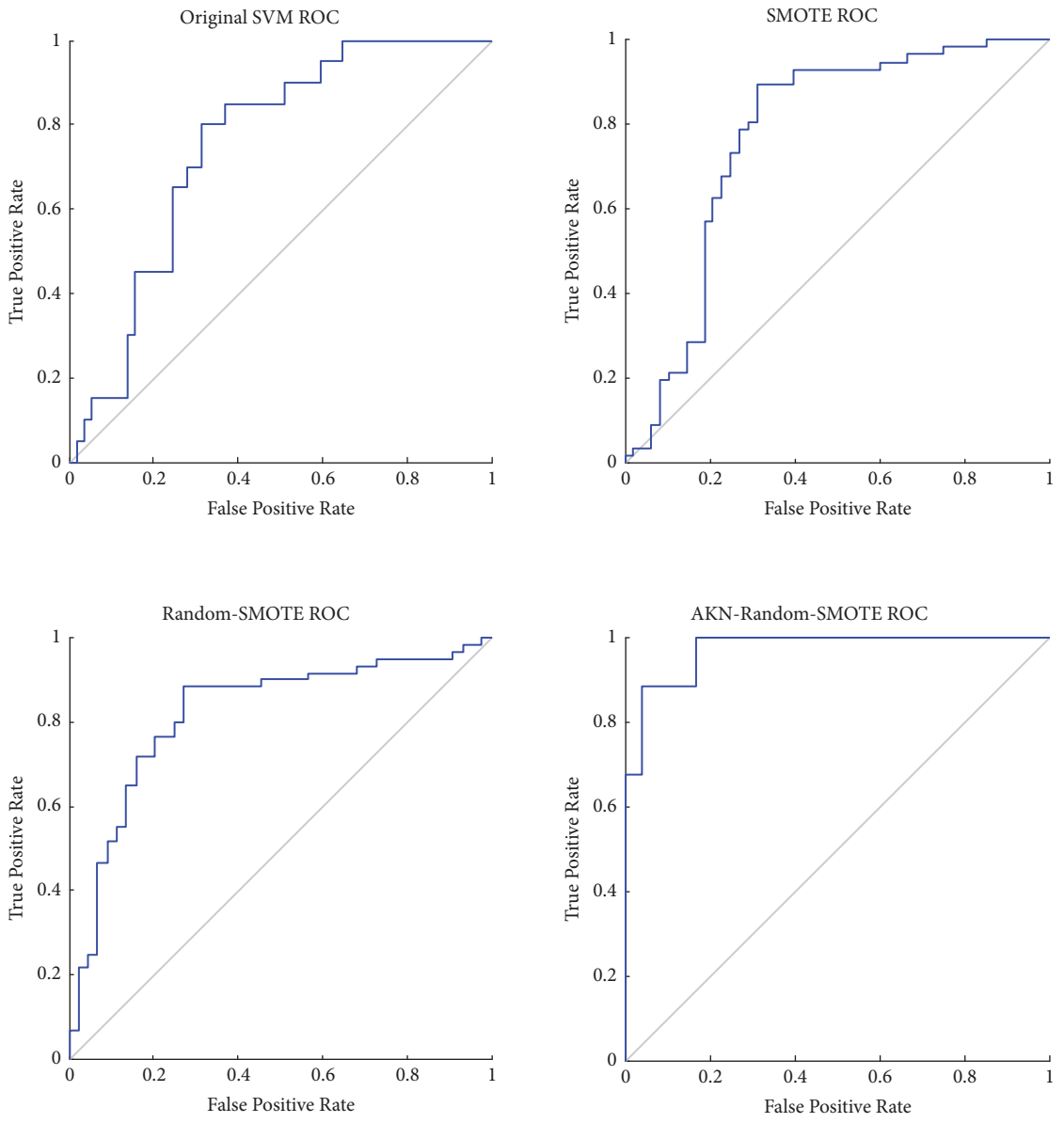

Figure 12: The ROC curves of the four methods for the Pima data set.

Table 4: The AUC values of the different methods for the four data sets.

Data sets	Original SVM	SMOTE	Random-SMOTE	AKN-Random-SMOTE
Banana	0.5845	0.6695	0.7378	0.8912
Haberman	0.7560	0.8205	0.8451	0.9727
Pima	0.7491	0.7716	0.8189	0.9718
Led	0.8804	0.9676	0.9910	0.9962

6. Conclusion

In this paper, an improved oversampling algorithm based on the samples' selection strategy is proposed. The support vectors are extracted and treated as the parent samples for oversampling to realize the balance of the data sets. Finally, the imbalanced data sets are classified with SVM classification algorithm. The experiments compare the classification effects via the different processing methods for the 10 imbalanced data sets including the classification approach without data processing and the classification approaches with data processing by SMOTE, Random-SMOTE, and AKN-Random-SMOTE. We select F-measure value, G-mean value, ROC curve, and AUC value as the evaluation indexes. The experimental results show that the AKN-Random-SMOTE

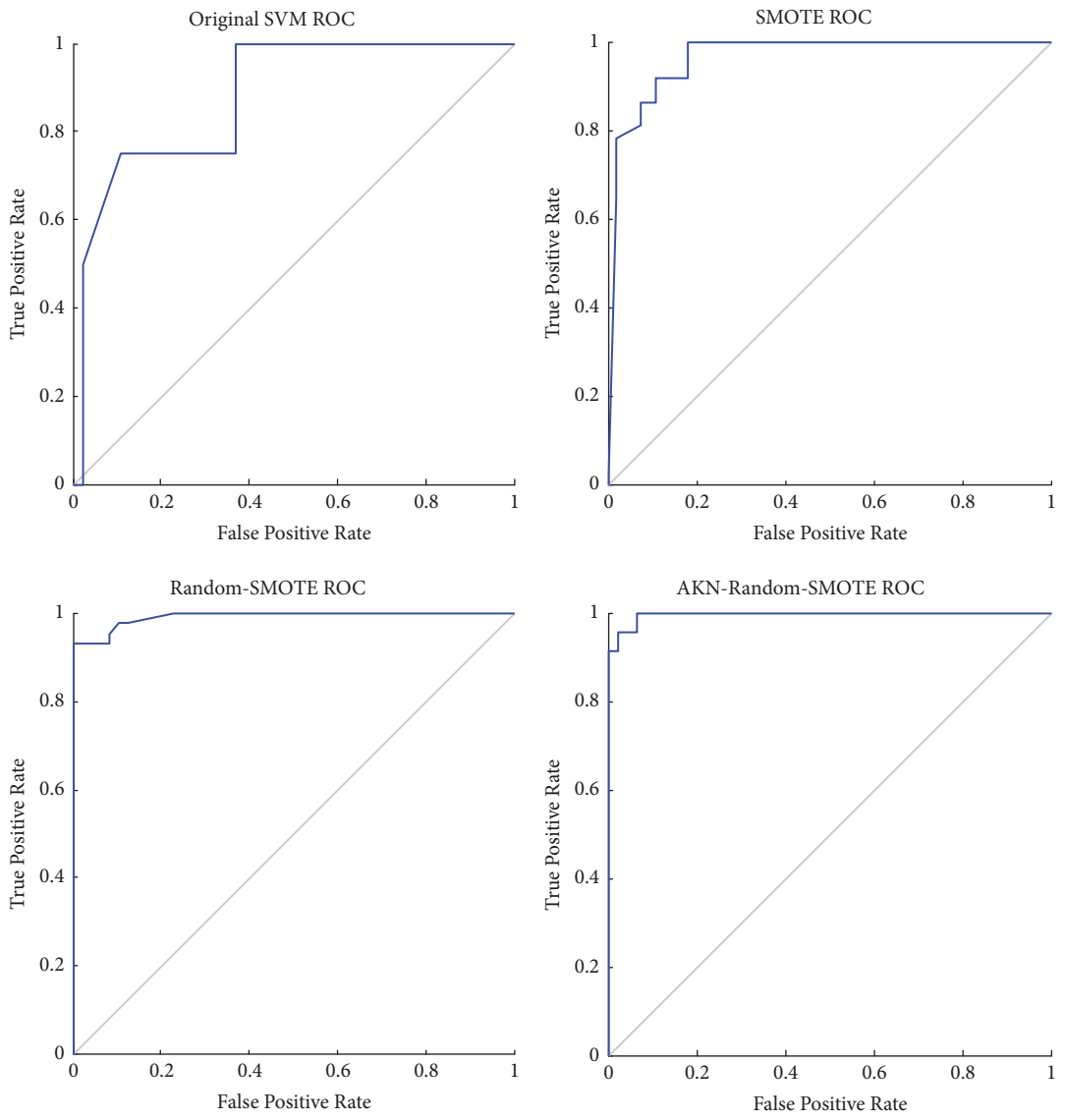

Figure 13: The ROC curves of the four methods for the Led data set.

algorithm has good performance for the imbalanced data classification.

In view of the problems exiting in the experiments, two issues can be considered to be resolved in the future work. Firstly, for the K-means clustering algorithm and the alien k-neighbors algorithm, the choices of parameters have a great impact on the experimental results. Therefore, a reasonable algorithm can be considered to optimize the parameters so that the algorithm has a stable and optimal classification effect. Secondly, the improved oversampling method can be combined with the novel or improved classification algorithm as a comprehensive solution for both the data-level and algorithm-level for further research.

Data Availability

The imbalanced data sets data used to support the findings of this study have been deposited in UCI data set and KEEL data set and they are available openly (the URL of the UCI data set is http://archive.ics.uci.edu/ml/datasets.php and the URL of the KEEL data set is https://sci2s.ugr.es/keel/dataset.php).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China 71702039.

References

[1] J. Han, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers Inc, San Francisco, Calif, USA, 2005.

[2] Q. Tang, *Research of Classification on Imbalanced Data Sets and Its Application in Student loans Credit Risk Management*, Wuhan, China, 2012.

[3] R. Andrews, J. Diederich, and A. B. Tickle, "A survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-Based Systems*, vol. 8, no. 6, pp. 373–389, 1995.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] P. LangIey, W. Iba, and K. Thompson, "An analysis of Bayesian classifers," in *Proceedings of the 10th National Conference of Artificial Intelligence*, pp. 223–228, AAAl Press, San Jose, Calif, USA, 1992.

[6] M. Ramoni and P. Sebastiani, "Robust Bayes classifiers," *Artificial Intelligence*, vol. 125, no. 1-2, pp. 209–226, 2001.

[7] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Machine Learning*, vol. 10, no. 1, pp. 57–78, 1993.

[8] J. R. Quinlan, *C4.5 Programs for Machine Leaning*, Morgan-Kaufmann Publishers Inc, San Francisco, Calif, USA, 1993.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, London, UK, 2000.

[10] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, Cambridge, Mass, USA, 2001.

[11] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[12] S. Abe, *Support Vector Machines for Pattern Classification*, Springer, New York, NY, USA, 2005.

[13] H.-P. Huang and Y.-H. Liu, "Fuzzy support vector machines for pattern recognition and data mining," *International Journal of Fuzzy Systems*, vol. 4, no. 3, pp. 826–835, 2002.

[14] Q. Wu, "Research on SVM learning algorithm based on optimization theory," pp. 2-3, 2009.

[15] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM, Pittsburgh, PA, USA, July 1992.

[16] L. Qin, *Research on Datamining Method for Imbalanced Dataset Based on Improved SMOTE*, Nanjing university of aeronautics and astronautics, Nanjing, China, 2017.

[17] Y. Li et al., "The protein classification combing the site evolution distance and SVM," *Chinese Journal Computer*, vol. 31, no. 1, pp. 43–50, 2008.

[18] P. Xu et al., "Internet traffic classification using support vector machine," *Journal of Computer Research and Development*, vol. 141, no. 3, pp. 407–414, 2009.

[19] X. Yan, *Comprehensive Oversampling and Underdamping Study of Imbalanced Data Sets*, Northeast electric power university, Jilin City, China, 2016.

[20] K. Yoon and S. Kwek, "An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics," in *Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, IEEE, Rio de Janeiro, Brazil, November 2005.

[21] X. Chen J et al., "An application of classification analysis for skewed class distribution in therapeutic drug monitoring the case of vancomycin," in *the Proceedings of 4th International Workshop on Design of Reliable Communication Networks*, pp. 35–39, IEEE, Alberta, Canada, 2003.

[22] P. Radivojac, U. Korad, K. M. Sivalingam, and Z. Obradovic, "Learning from class imbalanced data in wireless sensor networks," in *Proceedings of the 2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall*, pp. 3030–3034, IEEE, Orlando, FL, USA, October 2003.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[24] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, vol. 3644 of *Lecture Notes in Computer Science*, pp. 878–887, Springer, Berlin, Germany, 2005.

[25] Y. J. Dong, *The Study on Random-SMOTE for the Classification of Imbalanced Data Sets*, Dalian university of technology, Dalian, China, 2009.

[26] M. Kubat and S. Matwin, "Addressing the curse if imbalanced training sets: one sided selection," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 179–186, Nashville, Tennessee, 1997.

[27] R. Agarwal et al., "A new framework for learning classifier models in data mining (a case-study in network intrusion detection)," in *Proceedings of the 1st SIAM International Conference on Data Mining*, pp. 1–17, Chicago, IL, USA, 2001.

[28] M. V. Joshi and V. Kumar, "Classification using ripple down structure (a case for rare classes)," in *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 321–332, Orlando, Fla, USA, 2004.

[29] D. Randall Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257–286, 2000.

[30] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515-516, 1968.

[31] Y. Tang, Y.-Q. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 1, pp. 281–288, 2009.

[32] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: improving prediction of the minority class in boosting," in *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 107–119, Springer, Dubrovik, Croatie, 2003.

[33] M. Weiss G and H. Hirsh, "A quantitative study of small disjuncts," in *Proceedings of the In Proceedings of the 17th National Conference on Artificial Intelligence*, pp. 665–670, AAAI Press, Austin, TX, USA, 2000.

[34] Lu. J. R., *Research on classification method of high-dimensional class-imbalanced data sets base on SVM*, Harbin Institute of Technology, Harbin, China, 2016.

[35] J. N. Chen et al., "Speeding up algorithm for support vector machine based on alien neighbor," *Computer Engineering*, vol. 44, no. 5, pp. 19–24, 2018.

[36] H. T. Xiong et al., "Study on the class overlap and its treatment methods in classification," *Journal of Management Sciences in China*, vol. 4, no. 16, pp. 8–10, 2013.