ABSTRACT The problem of classifying imbalanced data is one of the active research directions in machine learning and bioinformatics. The imbalance of data will greatly degrade the accuracy of classifiers. Good oversampling methods can improve the diversity and validity of new samples, which cannot only solve the imbalance problem of sample data, but also greatly improve the classification accuracy. In this study, we propose the trigonal barycenter theory and a semi-supervised oversampling method, called SOTB (Semi-supervised Oversampling method based on Trigon Barycenter theory). SOTB works to: (1) construct the non-intersecting triangles based on Mahalanobis distance; (2) combine the semi-supervised sampling method with trigonal barycenter theory to oversample the positive samples, which can cope with the data imbalance problem without affecting the quality of data. Lastly, extensive experiments were conducted to verify the effectiveness of the proposed method. The results demonstrate that SOTB can improve the validity, diversity and rationality on the distribution of the newly generated samples as well as alleviate the phenomena of over-fitting which is popular in existing oversampling approaches. In particular, when compared with the state-of-the-art oversampling methods, the results show SOTB can achieve the best classification performance.

INDEX TERMS Classification problem, trigonal barycenter theory, imbalanced data, imbalance learning, semi-supervised oversampling.

I. INTRODUCTION

Imbalanced data is pervasive in all aspects of our living life, e.g., cancer gene data from hospitals for health examination [1], software defect data generated by defect detecting software [2] and telecommunication fraud data [3], [4] formed in telecommunication systems, which can be classified into different categories, the aforementioned data are called imbalanced data [5]. Generally speaking, machine learning approaches use a large amount of sample data to train the classification model, which makes the model obtain the capability of accurately predicting the class labels of data [6], [7]. However, it will degrade the recognition rate of classification models by employing imbalanced data to train the model, resulting in a decrease in classification performance. Working as an important means to improve the performance of classification and prediction, balancing imbalanced data plays an essential role in classification research and prediction analysis [8].

Generally speaking, the number of negative samples is much greater than that of positive samples in the classification problem. However, due to people’s daily wants,
the recognition significance for positive samples is greater than that of negative samples. For example, in cancer gene detection, there may be only one cancer gene among thousands of genes. If normal genes are detected as cancer genes, it will cause unnecessary psychological burden to the patients. On the contrary, if a cancer gene is wrongly detected as a normal gene, the patients may miss the optimal treatment period and bring life-threatening to them. Another illustrative example is that, in the detection of virus programs, there are a few virus programs among a large number of software programs. If normal programs are recognized as virus programs, the computer may isolate this program, which will affect the normal operation of related programs. If the virus program is recognized to be a normal one, it may bring in serious consequences, e.g., information leaks, and computer crashes that threaten the safety of personal properties. In the research of bioinformatics, imbalanced data is pervasive in various application scenarios, e.g., identifying DNA-protein binding sites, RNA binding residues and snoRNA. The imbalanced data plays an important role in the research in bioinformatics. For example, improving the prediction accuracy of DNA-protein binding site can help us know more about the disease. Balancing imbalance data has become an important research direction in bioinformatics. So, it is of great theoretical and practical significance to cope with the problem of imbalanced classification as well as improve the recognition rate of positive samples.

However, the fundamental reason for the decline in classification performance caused by imbalanced data is that most of the classifiers train their optimal parameters by gradient descent or gradient ascent [9]. Using imbalanced data to train the classifiers, the number of training times for the negative samples is more than the number of training times for the positive samples, which implies that the parameters of the classifier are excessively trained to recognize negative samples, so that the classifier can accurately identify the negative samples instead of accurately identifying positive ones. By using a balanced dataset to train the classifier, the classifier can learn the characteristics of the positive samples thoroughly to achieve the purpose of improving the classifier results. Therefore, it is important to propose an effective and efficient method to solve the data imbalance problem in machine learning.

Contributions: In order to improve the performance of imbalanced data classification, the main contributions of this paper are given as follows:

New theory. We integrate the trigonal barycenter theory to generate new samples, which has the following advantages: 1) the new samples do not spread across the boundary of the samples, the effectiveness of oversampling can be greatly improved; 2) the new sample is uniformly distributed in the sample space, effectively describing, filling the sample space and improving the performance of machine learning; 3) the newly generated samples are rich in diversity, which can effectively alleviating the phenomenon of over-fitting.

New method. In order to effectively balance various data sets, we proposed SOTB method (Semi-supervised Oversampling method based on Trigon Barycenter theory). SOTB applies the semi-supervised theory, which can resample the imbalanced data to achieve the purpose of improving the classification accuracy.

Extensive experimental results. In order to verify the effectiveness of SOTB, we conducted extensive experiments by comparing it with the state-of-the-art oversampling methods to evaluate the performance.

II. RELATED WORK

How to effectively train the classification model has become one of the important research directions in machine learning. The researchers have paid a lot of attention and efforts in it, and consequently some approaches to solve the imbalance problem have been proposed in order to improve the performance of classification. The most straightforward methods focus on modifying the classifier, such as: Zularaert et al. [10] proposed a new activation function in the convolutional neural network, which improves the validity of the convolution layer. Zhang et al. [11] viewed the performance evaluation function as the target function and proposed a loss function in order to improve the effect of training. Zhang et al. [12] replaced the middle layers of neural networks by a new function, which saves the training time and improves the classification accuracy.

However, the aforementioned methods of using improved classifier have the following drawbacks: 1) the improvement is not obvious because the existing methods of improving the classifier works well at the current stage; 2) there is upper bound of classification accuracy, because the number of positive samples in the imbalanced dataset is insufficient, so the classifier cannot fully learn its features, thus the recognition rate of positive samples is difficult to be improved. To improve the classifier cannot fundamentally cope with the imbalance classification problem. Then, some researchers integrated the results of multiple classifiers to improve the accuracy of classification, e.g., Zhang et al. [13] proposed an ensemble method by voting the classification results of multiple convolutional neural networks in order to obtain more comprehensive results. Siers and Islam [14] combined the classification results of multiple decision trees. However, the aforementioned methods still exist the following disadvantages: 1) the time cost is high, because the assembling operation of classification results requires the training of multiple classifiers; 2) the improvement of classification accuracy is still limited due to the lack of positive samples.

In order to improve the classification performance in an effective fashion, the methods for balancing the data are proposed. Some approaches discard a part of the negative samples which makes the positive and negative samples balanced in number, which was called undersampling method [15]; Yu et al. [16] proposed to use random undersampling to predict protein residue binding sites by randomly abandoning a portion of negative samples, which can increase the accuracy
of predicting the protein residue binding position. Yen and Lee [17] found that random undersampling method will lose a lot of useful samples due to random selection. In order to avoid the problem of randomness, the negative samples are clustered, selecting representative samples in clusters as training set to improve the effectiveness of undersampling. Babar and Ade [18] believes that we should pay more consideration on the distribution characteristic of the samples and the undersampling method based on the distribution of samples can reduce the blindness of undersampling so as to guarantee the classification accuracy. In summary, undersampling can improve the classification performance of positive samples to a certain extent, but it will inevitably neglect negative samples, which makes the classifier not fully learn the features of negative samples and reduces the classification accuracy.

In order to reduce the loss of data features, several oversampling methods are proposed. The basic idea of oversampling is to generate new samples with certain validity and rich diversity, which simulates to obtain positive samples in reality in order to balance the number of positive and negative samples. The oversampling methods can solve the imbalance problem to some extent [19], [20]. Currently, the research of oversampling has become an active research direction to solve the imbalance classification problem. The most classical oversampling method is the SMOTE oversampling method proposed by Chawla et al. [21]. Its basic idea is to randomly generate samples between positive sample and its k nearest neighbors. Since the newly generated samples by SMOTE is a random value between these two samples, which increases the diversity. Han et al. [22] thought that the samples in the boundary between positive and negative samples can be easily misclassified. The proposed method works to find positive samples at the boundary of the positive sample space. Then, it performs SMOTE oversampling on these positive samples to improve the learning of boundary samples. Ramentol et al. [23] combined the SMOTE method with the fuzzy rough set theory to achieve good classification effect on the decision tree. Nakamura et al. [24] proposed the LVQ-SMOTE method to oversample the positive samples by using the learning vector quantization method. The newly generated positive samples effectively fill the space of the positive samples. Dong and Wang [25] proposed a method called Random-SMOTE, which randomly performs SMOTE oversampling among positive samples to increase the diversity of SMOTE oversampling.

The oversampling algorithm based on SMOTE can improve the classification accuracy, but there are some disadvantages in current oversampling algorithms, because most of the oversampling algorithms are based on k nearest neighbor theory for sample generation. The basic idea is that randomly generating a new sample between the currently selected sample and a k nearest neighbor. The new sample generated based on k nearest neighbors has three drawbacks:

1. The newly generated sample will distribute across the boundary of positive and negative samples.

The newly generated positive sample spans across the boundary of samples which refers to the edge of the sample space in which a certain type of samples is located.

Fig. 1 shows the distribution of two kinds of samples, black for positive samples, white for negative ones, and the central black five-pointed star is viewed as the currently selected positive sample at the initial stage. In the phase of sample generation, since the currently selected sample is located in the central region of the positive samples, the k nearest neighbors are positive sample, so the generated sample does not cross the sample boundary, as is shown in Fig. 2(a).

As shown in Fig. 2(b), the currently selected sample is located at the sample boundary, the k nearest neighbors include two types of samples, a part of k nearest neighbors are positive samples and the others are negative ones. Thus, some newly generated samples cross the sample boundary, which are called noise sample. Since some samples generated are not in the sample space, using these samples to train the classifier will greatly degrade the classification accuracy.

2. The distribution of synthetic samples is unstable.

The oversampling method based on k nearest neighbor theory mostly adopt Equation 1 to generate samples, where \( \alpha \in (0, 1) \), that is, choosing a random point between two samples as the new sample.

\[
X_{new} = \alpha \ast X_m + (1 - \alpha) \ast X_n
\] (1)

Although this method increases the diversity of samples, the distribution of the new samples is unstable. This is because the new sample is a random one between two existing samples, which will make the generated new samples concentrate in a certain area or sparse in a certain area, resulting in an unstable distribution after oversampling.
D. Liu et al.: SOTB: Semi-Supervised Oversampling Approach Based on Trigonal Barycenter Theory

Fig. 2 illustrates this problem of the sample generation method based on Equation 1.

When the value of $\alpha$ is very small, the synthetic samples will distribute in the neighborhood of the currently selected sample, as shown in Fig. 2(a). If the value of $\alpha$ is large, the newly generated samples will deviate from the currently selected sample, as is shown in Fig. 2(b), and it will disrupt the stability of data distribution.

(3) The new samples do not have rich diversity.

In terms of the sample generation method according to Equation 1, the newly generated samples can only geographically located in the connecting line between two samples in space. Although the generated sample is different from the original one, the diversity of the new sample is not rich. It is impossible to generate a new sample away from the connecting line between two samples. The diversity of the new samples is insufficient, as shown in Fig. 3.

In order to overcome the aforementioned problems, in this paper, we propose a semi-supervised oversampling method based on the trigonal barycenter theory, called SOTB, which can effectively solve the following problems in oversampling methods: (1) the new sample distributes across the sample boundary, (2) the new sample is unstable in distribution, (3) the new sample is not particularly rich in diversity, and it can also alleviate the overfitting problem.

III. SEMI-SUPERVISED OVERSAMPLING METHOD BASED ON TRIGONAL BARYCENTER THEORY

In this study, the proposed semi-supervised oversampling method utilizes the characteristics of the trigonal barycenter and semi-supervised learning to generate positive samples, which can effectively solve the problems of the new sample across the sample boundary, unstable distribution of samples, and insufficient in diversity. In particular, it can reduce the phenomenon of over-fitting in the training phase, and greatly eliminate the influence of imbalance data. The semi-supervised oversampling algorithm based on the trigonal barycenter theory mainly includes the following steps: 1) distance measurement between samples, 2) non-intersecting triangle construction, 3) sample generation, and 4) semi-supervised oversampling.

A. DISTANCE MEASUREMENT BETWEEN SAMPLES

It is very scientific to measure the similarity between two samples with appropriate methods [8]. The Mahalanobis distance represents the covariance distance of two objects. Unlike the Euclidean distance, it takes into full consideration the relationship of different features, and this kind of relationship is scale-invariant. Furthermore, Mahalanobis distance is not affected by dimensions, which implies that Mahalanobis distance between two samples is independent on the unit of measurement.

Definition 1 (Covariance Matrix): Given a sample set $X = \{x_1, x_2, \ldots, x_i, \ldots, x_m\}$, there are $m$ samples, $x_i$ is a $n$-dimensional feature vector, then $X$ is a $m \times n$ matrix, and each row in $X$ is a sample. The equation of covariance matrix is given in Equation 2 and Equation 3.

$$
\Sigma = \begin{bmatrix}
\sum_{i=1}^{m} x_{i1} & \ldots & \sum_{i=1}^{m} x_{in} \\
\vdots & \ddots & \vdots \\
\sum_{i=1}^{m} x_{n1} & \ldots & \sum_{i=1}^{m} x_{nn}
\end{bmatrix} 
$$

(2)

$$
\Sigma_{ij} = \text{cov}(x_i, x_j) = E[(x_i - E(x_i)) \times (x_j - E(x_j)^T)]
$$

(3)

where $\Sigma$ represents the covariance matrix and $E(x_i)$ is the expectation of $x_i$.

Definition 2 (Mahalanobis Distance): As for the Mahalanobis distances of sample $x_i$ and $x_j$, $x_i$ and $x_j$ are column vectors, and their mahalanobis distance is calculated by Equation 4.

$$
\Gamma(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}
$$

(4)

where $\Sigma^{-1}$ represents the inverse matrix of $\Sigma$ and $\Gamma(x_i, y_j)$ represents the Mahalanobis distance between $x_i$ and $x_j$.

B. NON-INTERSECTING TRIANGLE CONSTRUCTION

The non-intersecting triangle implies that the sides of two triangles do not overlap and intersect with each other. In order to avoid the overlapping of triangles in the phase of triangle construction, we propose a non-intersecting triangle construction approach. This method regards the sample $S_{\text{border}}$ that is farthest from the mean center sample $S_{\text{mean}}$ as a benchmark, and use the Mahalanobis distance between other samples and $S_{\text{border}}$ as a measurement. By integrating this measurement, the non-intersecting triangle is constructed by an ascending order. By using this method, $\lceil \frac{m}{4} \rceil$ non-intersecting triangles are obtained, where $m$ is the number of positive samples and the symbol of $\lceil \rceil$ represents the rounding operation.

The non-intersecting triangle construction method includes the following steps. Firstly, it finds the mean center value of the positive samples by Equation 5.

$$
S_{\text{mean}} = \left( \frac{\sum_{i=1}^{m} x_{1i}}{m}, \frac{\sum_{i=1}^{m} x_{2i}}{m}, \ldots, \frac{\sum_{i=1}^{m} x_{ni}}{m} \right)
$$

(5)

where $x_{ij}$ represents the $j^{th}$ dimension of the $i^{th}$ sample, $n$ represents the number of dimensions, $1 < j < n$, $m$ represents the number of positive samples. So $S_{\text{mean}}$ is the mean center of the positive samples.

Secondly, it calculates the Mahalanobis distance between $S_{\text{mean}}$ and all other positive samples by Equation 2–Equation 4. Using the obtained Mahalanobis distance to find $S_{\text{border}}$. 

FIG. 3. Example of the new sample without diversity.
which is the farthest sample from the mean center value. Then, it computes the Mahalanobis distances between \( S_{\text{border}} \) and all positive samples besides \( S_{\text{border}} \) and sorts the distance in an ascending order and records the sequence number of samples, so that all positive samples form a sequence which is sorted by ascending order according to the Mahalanobis distance between positive samples and \( S_{\text{border}} \) where \( S_{\text{border}} \) is the first one in the sequence. Finally, by the ascending order, every three samples constitute a triangle and a total of \( \lfloor \frac{m}{3} \rfloor \) triangles are generated.

**C. SAMPLE GENERATION METHOD**

The sample generation method is the most important part of oversampling. The sample generation method based on the triangle barycenter can effectively alleviate the phenomena of new sample across the sample boundary, new sample distribution is unstable, and the diversity of new sample is not rich. The specific generation method is that calculating the barycenter of the triangle via Equation 6–Equation 7.

\[
\begin{align*}
 w_1 &= \frac{x_{i1} + x_{j1} + x_{k1}}{3} \quad (6) \\
 S_{\text{child}} &= \{w_1, w_2, \ldots, w_n\} \quad (7)
\end{align*}
\]

where \((x_{i1}, x_{j1}, x_{k1})\) represents a triangle, \(x_i\) represents the \(i^{th}\) sample, \(x_j\) represents the \(j^{th}\) eigenvalue of the \(i^{th}\) sample. And, \(S_{\text{child}}\) is the set of newly generated samples by the current triangle.

**D. SEMI-SUPERVISED OVERSAMPLING METHOD**

Semi-supervised method is widely used in machine learning for improving the accuracy of classification. In this work, we apply semi-supervised method to generate new samples. On one hand, semi-supervised method is suitable for the case with very few samples; on the other hand, semi-supervised method can iteratively generate samples to meet the requirement of a large number of samples.

SOTB can generate \( \left\lfloor \frac{m}{3} \right\rfloor \) samples per round of sampling, after one oversampling, the number of samples changes from \( m \) to \( m + \left\lfloor \frac{m}{3} \right\rfloor \). In order to cope with the data imbalance problem, we introduce a semi-supervised oversampling method, which merges the synthetic positive samples into the current positive samples to form a new positive sample set as the input of the next round of sampling. In this fashion, data balance is achieved after multiple rounds of sample generation. The phase of data generation is shown in Fig. 4.

Fig. 4 shows the phase of the semi-supervised oversampling method based on trigonal barycenter theory. The initial phase includes six positive samples, i.e., \( S_1 \sim S_6 \). The mean center \( S_{\text{mean}} \) is obtained by Equation 5. Then, it calculates the Mahalanobis distance between \( S_{\text{mean}} \) and each positive sample to find sample \( S_2 \) that is farthest from \( S_{\text{mean}} \). The next step is to calculate the Mahalanobis distance of each sample to \( S_2 \), to sort samples by ascending order to construct triangles, so \( S_1 \sim S_3 \) and \( S_4 \sim S_6 \) are paired to form triangles. It is worth to notice that if the number of current samples is not a multiple of three, the remaining samples will not participate in sample generation. The next step is sample generation, on the basis of the trigonal barycenter theory to generate new samples, the sample \( S_7 \sim S_8 \) is generated. Repeating the above steps to iteratively generate the positive sample.

Because the synthetic child sample is used as the mother sample in the step of next generation, the proposed generation method is semi-supervised. The algorithm of semi-supervised oversampling based on trigonal barycenter theory is shown in Algorithm 1.

**E. ALGORITHM COMPLEXITY ANALYSIS**

For a positive sample set consisting of \( m \) positive samples, the attribute dimension of the sample is \( n \). Firstly, SOTB constructs the paired triangles by non-intersecting triangle construction method and generates several new samples through the trigonal barycenter theory. The non-intersecting triangle construction method works to first find the mean center of all positive samples by calculating the mean value of each attribute dimension w.r.t. all positive samples, and its time complexity is \( O(n) \). Then, it calculates the Mahalanobis distance between the mean center and all positive samples, and applies the bubble sorting algorithm to find the sample \( S_{\text{border}} \) farthest from \( S_{\text{mean}} \), the time complexity of these two steps is \( O(2 \times m) \). Secondly, it calculates the Mahalanobis distance of all positive samples to the boundary sample \( S_{\text{border}} \) in order to obtain \( M_{\text{mean}} \) by using the bubble sorting method. The time complexity of these two steps is \( O(m + m^2) \). Lastly, the phase of constructing non-intersecting triangle is to generate new samples, each time a non-intersecting trigonal is constructed, a new sample is generated inside this triangle. The time complexity of non-intersecting constructing and new sample generation is \( O(\left\lfloor \frac{m}{3} \right\rfloor \times n) \). Therefore, the time complexity of this SOTB is \( O(n + 3m + m^2 + \left\lfloor \frac{m}{3} \right\rfloor) \).

**IV. PROPERTIES OF SOBT ALGORITHM**

The proposed semi-supervised oversampling approach based on trigonal barycenter theory has the following properties:

**Property 1:** The newly generated samples do not distribute across the sample boundary.

In terms of the real data, a certain kind of samples is distributed in a continuous sample space, so it is very difficult to find the boundary of these sample spaces. Currently, the sample boundary can only be roughly described by the existing methods. In this paper, the boundary of samples is partitioned by connecting lines between boundary samples.
Algorithm 1 Semi-Supervised Oversampling Method Based on Trigonal Barycenter Theory

Input: Imbalanced data set \( \mathcal{D} = \{x_1, x_2, \cdots, x_n\} \).

Output: Balanced data set \( D = \{x_1, x_2, \cdots, x_m\} \).

1: Partitioning \( \mathcal{D} \) into positive set \( P \) and negative set \( N \);
2: while \( P, no < N, no \) do
3: Calculating the number of positive and negative samples \( P, no \) and \( N, no \), respectively;
4: Calculating \( S_{\text{mean}} \) by Equation 5;
5: Calculating Mahalanobis distance between \( S_{\text{mean}} \) and all positive samples and storing the distance values in matrix \( M_{\text{mean}} \) by Equation 2–Equation 4;
6: Applying bubble sort method to find \( S_{\text{border}} \) that is farthest from \( S_{\text{mean}} \);
7: Calculating the Mahalanobis distance between \( S_{\text{border}} \) and each positive sample, sorting the distances by ascending order, and storing them in the matrix \( M_{\text{border}} \);
8: Initializing \( S_{\text{child}} \) and putting newly generated samples into \( S_{\text{child}} \);
9: for \( i = 0 \) to \( P, no - (P, no \% 3) \) do
10: Initializing a sample \( s \) represents the newly generated sample by the current triangle;
11: if \( i \% 3 == 0 \) then
12: \( x = M_{\text{border}}[i] \);
13: else if \( i \% 3 == 1 \) then
14: \( y = M_{\text{border}}[i] \);
15: else
16: \( z = M_{\text{border}}[i] \);
17: if \( S_{\text{child}}, no + P, no < N, no \) then
18: \( s \leftarrow \frac{x + y + z}{3} \);
19: \( S_{\text{child}} \leftarrow s/3 \);
20: end if
21: end if
22: end for
23: \( P \leftarrow P + S_{\text{child}} \);
24: end while
25: return \( D \);

Samples that are considered to be outside the boundary are regarded to be not in the same category where the samples inside the boundary area. In particular, in this study, we choose the trigonal barycenter instead of the barycenter of polygon more than three edges, because the barycenter of polygon more than three edges may not be inside of the polygon. As shown in Fig. 5, the newly generated sample is not inside the polygon, which implies that the newly generated sample spans across the sample boundary, so the newly generated sample is viewed as a noise sample. Using this kind of sample to train the classifier will greatly reduce the classification performance of the classifier.

Property 2: The new sample can uniformly distribute in the sample space.

Whether the training samples are uniformly distributed in the sample space or not has a very important influence on the training of the classifier. For example, when the sample is generated, the newly generated samples are too concentrated in a certain region, which causes the classifier will enhance the process of learning in this region, while the learning of other regions is weakened, resulting in insufficient learning w.r.t. the samples in the weakened regions. For this case, SOTB can better handle this problem, because the newly generated sample is at the barycenter of three samples. After multiple iterations, the distribution of the new samples will not be completely concentrated in a certain region, and the distribution of new samples is more uniform in the sample space to avoid the case where the sample region are too densely distrabuted. So, the classifier can be adequately and fully trained to achieve high classification accuracy.

Property 3: The newly generated sample is rich in diversity.

In the sample space, samples with rich diversity can describe the distribution of samples very well and improve the classification performance of the classifier. The reason why random oversampling methods have a limited improvement in classification accuracy is that the synthetic positive samples are similar to the original positive samples, and the new samples are not rich in diversity, and the classifier repeatedly learns the same samples, resulting in over-fitting. SMOTE, a famous oversampling method based on \( k \) nearest neighbors has certain diversity on sample generation, but the diversity is not rich enough. SOTB can generate new samples in the barycenter of three samples, which makes the new samples rich in diversity, meets the requirement of sample diversity in imbalanced classification, and effectively improves the classification accuracy. At the same time, due to the rich diversity of the newly generated samples, SOTB greatly reduces the over-fitting phenomenon.
B. EVALUATION METRICS

In experiments, we need appropriate evaluation metrics to fairly evaluate the performance of different classification methods, which can accurately reflect the classification performance. The metric of classification accuracy Acc cannot meet the evaluation requirements of the imbalance classification. For example, for a data set with an imbalance ratio equivalent to 9, if the classifier determines all samples as negative samples, the overall classification accuracy rate is 90%, but none of the positive sample was correctly classified. Due to the importance of positive samples, the key task in imbalance learning is the identification of positive samples. Therefore, when selecting evaluation metrics, we should pay more attention to identifying positive samples. In the confusion matrix of the binary classification problem, as shown in Table 2; TP represents a positive sample that is correctly classified, FN represents a positive sample that is misclassified, FP represents a negative sample that is misclassified, and TN represents a negative sample that is correctly classified.

According to the particularity of the binary classification problem and the definition of the confusion matrix, the $F_1$ measure evaluation metric is the harmonic mean of two evaluation metrics of Recall and Precision, where Recall can effectively evaluate the recognition rate of positive samples and Precision can effectively evaluate the confidence of the identified positive samples. $F_1$-measure combine Recall and Precision to achieve the purpose of comprehensive evaluation. Therefore, by using the $F_1$-measure evaluation index, it is possible to comprehensively and effectively evaluate the performance of the binary classification in imbalanced classification problem.

\[
\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (8)
\]

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)
\]

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)
\]

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)
\]

C. EXPERIMENTAL SETUP

In order to verify the effectiveness of the proposed method, we used a single variable method in experiments and the experimental results were obtained by specifying the same parameter values. The comparing algorithms are used to compare the experimental data on six different imbalanced data sets. In experiments, this proposed method is compared with other four different sampling methods, including: unprocessed raw data, random oversampling, SMOTE oversampling and MAHAKIL oversampling. In order to reflect the characteristics of data balance, besides the unprocessed raw data, for other sampling methods, the positive and negative samples are equally generated, which implies IR equals 1. Furthermore, we evaluate these sampling methods by four common classifiers, including: Decision Tree, Support Vector Machine, Random Forest and Neural Networks. In order to obtain comprehensive classification results, the experimental data was randomly divided into five parts by the five-fold cross-validation method [27], and in order to eliminate the influence of random factors on experimental results, we repeated the experiment for ten times and took the average value as the final result to ensure the effectiveness of the experiment.

D. EXPERIMENTAL RESULTS ANALYSIS

The results of semi-supervised oversampling method based on trigonal barycenter theory are given in Fig. 6. Fig. 6 shows the classification results of the Decision Tree classifier in six different data sets, and SOTB performs best on the data sets of glass0, haberman, new-thyroid2 and pima, but not on vehicle1 and yeast data sets.

Fig. 7 shows the classification results of different sample methods on six different data sets by using the Support Vector Machine classifier, and the proposed SOTB sampling method performs best on the data sets of glass0, haberman, new-thyroid2 and vehicle1, but not on pima and yeast data sets.

Fig. 8 shows the classification results of in the Random Forest classifier in six different data sets, and the proposed SOTB sampling method performs best on the data sets of haberman, new-thyroid2, vehicle1 and yeast, but not on glass0 and pima data sets.
Fig. 9 shows the classification results of in the Neural Networks classifier in six different data sets, and the proposed SOTB sampling method performs best on all data sets. In the experimental results, we can see that the proposed SOTB sampling method performs the best on most of the data sets by these four commonly-used classifiers. But, for some classifiers, oversampling method does not improve the classification result. For example, in terms of the classification result on glass0 data by the random forest classifier, the unprocessed raw data obtains the best classification performance. The main reason is that different data have different distributions and it may be that the distribution of the data is suitable for one classifier, and the results of the classification result after oversampling is not as good as the unsampled raw data. Therefore, we can conclude that it is unreasonable to use a kind of classifier on some specific data set to evaluate the sampling performance.

In order to further demonstrate the effectiveness of the sampling method, we conduct statistical analysis of the experimental results by using the metric of $F_1$-score. The better the classification results are, the higher ranking value the sample method has. By statistic methods, the average ranking score is obtained. The results are given in Table 3.

According to Table 3, we can see that the unprocessed raw data has the worst classification result by these four classifiers. After the imbalanced data is balanced by random oversampling, the classification effect is slightly improved, indicating that the balancing data set can improve the classification performance which agrees with the reality. As discussed in Section IV, SMOTE is a method that randomly generate a new sample between two samples. The newly generated sample has a certain diversity, but the new sample may cross the sample boundary. The MAHAKIL method partitions the samples into two classes, pairing the $i^{th}$ sample in the first class with the $i^{th}$ sample in the second class, and each pair of samples generates a new sample at the midle-point of these two samples. The sample has a certain diversity and the new sample does not cross the sample boundary as well. Because the samples generated by the SMOTE and MAHAKIL methods have a certain diversity, the classification performance is greatly improved by applying these two sampling methods. In addition, the MAHAKIL method generates samples that do not span across the boundary, so its average ranking value is higher than that SMOTE on most classifiers, but the newly generated sample by these MAHAKIL and SMOTE is not very rich in diversity. SOTB uses the theory of trigonal barycenter to generate a richer diversity of samples when compared with SMOTE and MAHAKIL and the new sample does not cross the
The problem of binary classifications in machine learning is very popular. Traditional oversampling methods have largely failed to take in full consideration the problem of the validity of newly generated samples, the rational distribution of new samples, and the diversity of new samples. In this study, we propose a semi-supervised oversampling method based on the trigonal barycenter theory, which can cope with the following problems: the lack of validity of new samples, the sample distribution is unstable after sampling and the diversity of new samples is not rich, and the over-fitting problem in machine learning is alleviated. By comparing with the four different oversampling methods, the classification results of SOTB performs best, which not only shows the superiority of the proposed method in the oversampling methods, but also shows that it is one of the effective ways to solve the problem of imbalance classification by improving the effectiveness and diversity of existing oversampling methods.

In our future work, we focus on designing a more effective oversampling method. The implementation of such a method would require an effective model to extract the basic features of positive samples in which are used to generate new positive samples. Furthermore, we will cope with the following problems: how to evaluate the newly generated samples and how to apply the proposed method in multi-class learning research.

ACKNOWLEDGMENT

(Dingxiang Liu, Shaojie Qiao, and Nan Han contributed equally to this work.)

REFERENCES

[1] Z. Liu, D. Tang, Y. Cai, R. Wang, and F. Chen, “A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data,” Neurocomputing, vol. 266, pp. 641–650, Nov. 2017.

[2] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, “MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction,” IEEE Trans. Softw. Eng., vol. 44, no. 6, pp. 534–550, Jun. 2018.

[3] R. A. Raub, A. H. N. Hamzah, M. D. Jaafar, and K. N. Baharin, “Using subscriber usage profile risk score to improve accuracy of telecommunication fraud detection,” in Proc. Int. Conf. Comput. Intel. Cybern., Nov. 2016, pp. 127–131.

[4] R. F. Lima and A. C. M. Pereira, “A fraud detection model based on feature selection and undersampling applied to Web payment systems,” in Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WIAIT), Dec. 2015, pp. 219–222.

[5] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, vol. 6, no. 1, p. 27, Dec. 2019.

[6] S. Qiao, D. Shen, X. Wang, N. Han, and W. Zhu, “A self-adaptive parameter selection trajectory prediction approach via hidden Markov models,” IEEE Trans. Intell. Transp. Syst., vol. 16, no. 1, pp. 284–296, Feb. 2015.

[7] S. Qiao, N. Han, W. Zhu, and L. A. Gutierrez, “TraPlan: An effective Three-in-One trajectory-prediction model in transportation networks,” IEEE Trans. Intell. Transp. Syst., vol. 16, no. 3, pp. 1188–1198, Jun. 2015.

[8] S. Qiao, N. Han, J. Zhou, R.-H. Li, C. Jin, and L. A. Gutierrez, “SocialMix: A familiarity-based and preference-aware location suggestion approach,” Eng. Appl. Artif. Intell., vol. 68, pp. 192–204, Feb. 2018.

[9] A. Ramaswamy and S. Bhatnagar, “Analysis of gradient descent methods with nondiminishing bounded errors,” IEEE Trans. Autom. Control, vol. 63, no. 5, pp. 1465–1471, May 2018.

[10] J. Zullietti, F. Godin, M. Kim, A. Soete, Y. Saey, and W. De Neve, “SpliceRover: Interpretable convolutional neural networks for improved splice site prediction,” Bioinformatics, vol. 34, no. 24, pp. 4180–4188, Dec. 2018.

[11] C. Zhang, G. Wang, Y. Zhou, and J. Jiang, “A new approach for imbalanced data classification based on minimize loss learning,” in Proc. IEEE 2nd Int. Conf. Data Sci. Cyberspace (DSC), Jun. 2017, pp. 82–87.

[12] J. Zhang, Y. Zheng, and D. Qi, “Deep spatio-temporal residual networks for citywide crowd flows prediction,” in Proc. 31st AAAI Conf. Artif. Intell., San Francisco, CA, USA, Feb. 2017, pp. 1655–1661.

[13] Y. Zhang, S. Qiao, S. Ji, N. Han, D. Liu, and J. Zhou, “Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information,” Eng. Appl. Artif. Intell., vol. 79, pp. 58–66, Mar. 2019.

[14] M. J. Siens and M. Z. Islam, “Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem,” Inf. Syst., vol. 51, pp. 62–71, Jul. 2015.

[15] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, “Clustering-based undersampling in class-imbalanced data,” Inf. Sci., vol. 409–410, pp. 17–26, Oct. 2017.

[16] D.-J. Yu, J. Hu, Z.-M. Tang, H.-B. Shen, J. Yang, and J.-Y. Yang, “Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling,” Neurocomputing, vol. 104, pp. 180–190, Mar. 2013.

[17] S.-I. Yen and Y.-S. Lee, “Cluster-based under-sampling approaches for imbalanced data distributions,” Expert Syst. Appl., vol. 36, no. 3, pp. 5718–5727, Apr. 2009.

[18] V. Babar and R. Ade, “MLP-based undersampling technique for imbalanced learning,” in Proc. Int. Conf. Autom. Control Dyn. Optim. Techn. (ICACDOT), Sep. 2016, pp. 142–147.

[19] Y.-D. Zhang, Y. Zhang, F. Phillips, Z. Dong, and S. Wang, “Synthetic minority oversampling technique and fractal dimension for identifying multiple sclerosis,” Fractals, vol. 25, no. 04, Aug. 2017, Art. no. 1740010.

[20] D. Liu, S. Qiao, Y. Zhang, N. Han, J. Wei, R. Zhang, and P. Huang, “A survey on data sampling methods in imbalance classification,” J. Chongqing Univ. Technol., vol. 33, no. 7, pp. 102–112, 2019.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[22] H. Han, W. Wang, and B. Mao, “Borderline-smote: A new over-sampling method in imbalanced data sets learning,” in Proc. Int. Conf. Intell. Comput. (ICIC), Hefei, China, vol. 3644. Berlin, Germany: Springer, Aug. 2005, pp. 878–887.

[23] E. Ramoneto, Y. Caballero, R. Bello, and F. Herrera, “SMOTE-RSB: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory,” Knowl. Inf. Syst., vol. 33, no. 2, pp. 245–265, Nov. 2012.

TABLE 3. The average ranking score of different sampling method on six different data sets.

| Classifier       | Original data | Random oversampling | SMOTE | MAHAKIL | SOTB  |
|------------------|---------------|---------------------|-------|---------|-------|
| Decision Tree    | 4             | 4                   | 2.833 | 2.5     | 1.667 |
| Support Vector Machine | 3.833   | 3.167               | 3.5   | 2.667   | 1.833 |
| Random Forest    | 3.833         | 3               | 2.167 | 3.833   | 2     |
| Neural Networks  | 5             | 3.333               | 3.5   | 2.167   | 1     |
[24] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, “LVQ-SMOTE—Learning vector quantization based synthetic minority Over—sampling technique for biomedical data,” BioData Mining, vol. 6, no. 1, p. 16, Dec. 2013.

[25] Y. Dong and X. Wang, “A new over-sampling approach: Random-smote for learning from imbalanced data sets,” in Proc. 5th Int. Conf. Knowl. Sci., Eng. Manage. (KSEM), Irvine, CA, USA, vol. 7091. Berlin, Germany: Springer, Dec. 2011., pp. 343–352.

[26] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, “KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” J. Multiple-Valued Logic Soft Comput., vol. 17, nos. 2–3, pp. 255–287, 2011.

[27] S. Qiao, N. Han, Y. Gao, R.-H. Li, J. Huang, J. Guo, L. A. Gutierrez, and X. Wu, “A fast parallel community discovery model on complex networks through approximate optimization,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1638–1651, Sep. 2018.

DINGXIAO LIU is currently with the School of Software Engineering, Chengdu University of Information Technology, Chengdu, China. He has coauthored several high quality articles and participated in several projects supported by the National Natural Science Foundation of China. His research interests include artificial intelligence and bioinformatics.

SHAOJIE QIAO received the B.S. and Ph.D. degrees from Sichuan University, Chengdu, China, in 2004 and 2009, respectively. From 2007 to 2008, he was a Visiting Scholar with the School of Computing, National University of Singapore. He is currently a Professor with the School of Software Engineering, Chengdu University of Information Technology, Chengdu. He has authored more than 40 high quality articles, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE TRANSACTIONS ON SYSTEMS MAN. AND CYBERNETICS: SYSTEMS, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and so on. His research interests include artificial intelligence and bioinformatics.

NAN HAN received the M.S. and Ph.D. degrees from the Chengdu University of Traditional Chinese Medicine, Chengdu, China. She is currently an Associate Professor with the School of Management, Chengdu University of Information Technology, Chengdu. She is the author of more than 30 articles. She has participated in several projects supported by the National Natural Science Foundation of China. Her research interest includes artificial intelligence.

TAO WU received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China. He is currently an Assistant Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include complex networks, machine learning, edge computing, and privacy-preserving.

RUI MAO received the B.S. and M.S. degrees in computer science from the University of Science and Technology of China, in 1997 and 2000, respectively, and the M.S. degree in statistics and the Ph.D. degree in computer science from The University of Texas at Austin, in 2006 and 2007, respectively. He is currently the Vice Dean of the College of Computer Science and Software Engineering, Shenzhen University, and the Executive Director of the Shenzhen Institute of Computing Sciences and the Executive Director of Guangdong Province Key Laboratory of Popular High Performance Computers and Guangdong Province Engineering Center of China-made High Performance Data Computing System. His research mainly focuses on universal data management and analysis.

YONGQING ZHANG received the Ph.D. degree from Sichuan University, Chengdu, China, in 2016. He is currently an Associate Professor with the School of Computer Science, Chengdu University of Information and Technology, Chengdu. His research interests include machine learning and bioinformatics.

CHANG-AN YUAN received the Ph.D. degree from the College of Computer Science, Sichuan University, China, in 2006. He is currently a Professor with Nanning Normal University and the Guangxi College of Education. His research interests include computational intelligence and data mining.

YUEQIANG XIAO received the Ph.D. degree from the Southwestern University of Finance and Economics, China, in 2006. He is currently a Professor with the School of Management, and the Vice President of the Chengdu University of Information Technology, Chengdu, China. His research interest includes data mining.

* * *