IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2022, accepted 10 January 2023, date of publication 13 January 2023, date of current version 19 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3236794

RESEARCH ARTICLE

Two Novel SMOTE Methods for Solving Imbalanced Classification Problems

YUAN BAO[1] **AND SIBO YANG**[2]

[1]School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[2]School of Science, Dalian Maritime University, Dalian 116026, China

Corresponding author: Sibo Yang (ysibo@dlmu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61720106005.

ABSTRACT The imbalanced classification problem has always been one of the important challenges in neural network and machine learning. As an effective method to deal with imbalanced classification problems, the synthetic minority oversampling technique (SMOTE) has its disadvantage: Some noise samples may participate in the process of synthesizing new samples; As a result, the new synthetic sample lacks its rationality, which will reduce the classification performances of the network. To remedy this shortcoming, two novel improved SMOTE method are proposed in this paper: Center point SMOTE (CP-SMOTE) method and Inner and outer SMOTE (IO-SMOTE) method. The CP-SMOTE method generates new samples based on finding several center points, then linearly combining the minority samples with their corresponding center points. The IO-SMOTE method divides minority samples into inner and outer samples, and then uses inner samples as much as possible in the subsequent process of generating new samples. Numerical experiments are conducted to prove that compared with no-sampling and conventional SMOTE methods, the CP-SMOTE and IO-SMOTE methods can achieve better classification performances.

INDEX TERMS Imbalanced classification problems, IO-SMOTE method, CP-SMOTE method, machine learning.

I. INTRODUCTION

For general balanced classification problems, the conventional neural networks can achieve good classification results. However, in the real world, there are lots of imbalanced problems, such as transaction fraud, cancer diagnosis [1], [2], virus script judgment, and so on. As far as cancer diagnosis is concerned, the number of cancer patients must be small. But it is precisely that these few cancer patients are the most important research objects. At this time, the original neural networks [3] are no longer able to obtain satisfactory classification results, especially for those minority samples. The reason for this result is that too few minority samples make the networks unable to learn the dataset efficiently. Therefore, how to deal with imbalanced problems is an important issue in machine learning.

The approaches to dealing with the imbalanced problems roughly come from two directions: algorithm improvement [4], [5] and data processing. The algorithm improvement includes feature selection [6], cost-sensitive [7], and integrated learning. And one of the effective data processing is resampling method [8], which includes undersampling [9] and oversampling methods [10]. Undersampling method is to remove some samples in the majority class to make the number of positive and negative samples balanced, and then train the network. The random undersampling (RUS) method [11] is one of the simpler undersampling methods. As the name suggests, the RUS method is to randomly select some samples from the majority S_{major} to form a sample set E; And then remove the sample set E from S_{major} to obtain a new data set $S_{minor} + S_{major} - E$. The RUS method achieves the purpose of modifying the sample distribution by changing the proportion of the majority samples, so as to make the samples more balanced. However, it also has some disadvantages. Since the sample number of the new dataset is less than that of

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

the original dataset, some information will be lost. That is, deleting the majority samples might cause the classifier to lose important information about the majority class.

In order to overcome the shortcomings of the undersampling method, researchers have proposed oversampling method [12], [13]. And the basic idea of the oversampling method is to add some minority samples to make the number of positive and negative samples balanced. The simplest random oversampling (ROS) method [14] is to randomly select some samples from the minority samples S_{minor}, and generate a sample set E by copying the selected samples, then add them to S_{minor} to obtain a new minority class set $S_{minor} + E$. However, for the ROS method, the complexity increases in the process of training the networks due to the duplication of the minority samples. On the other hand, it is easy to cause over-fitting problems, because the ROS method is simply a copy process of the initial samples, which is not conducive to the generalization performance of the network.

In order to solve the over-fitting problem [15] caused by the ROS method, and simultaneously ensure the dataset is balanced, Chawla [16] proposed a synthetic minority oversampling technique (SMOTE) method. The basic idea of the SMOTE method is as follows: For each minority sample x_i, randomly choose a sample x_i' from its neighbor (x_i' is also a minority sample); Then randomly select a point on the line between x_i and x_i' as the new synthetic minority sample.

Based on the SMOTE method, many researchers have made improvements and achieved better classification results. Borderline SMOTE [17] oversampling process is to divide the minority samples into three categories: safe, danger and noise. And then, only the danger samples are employed to generate the novel samples. Radius SMOTE first selects a minority sample x_i and calculates a radius according to the k-nearest neighbor. Then take x_i as the center and randomly find several points so that their distance to x_i is less than the radius. The R-SMOTE method [18] eliminates the limitation of generating minority class instance distribution and improves the classification accuracy of minority class. ADASYN [19] was proposed to generate new minority class samples near the original samples that were misclassified based on the k-nearest neighbor classifier.

For the original SMOTE method, some noise samples might participate in the process of synthesizing new samples. Thus, the new synthetic sample lacks its rationality, which will reduce the classification performances of the classifier. The purpose of this paper is to propose two novel improved SMOTE methods: Center point SMOTE (CP-SMOTE) method and Inner and outer SMOTE (IO-SMOTE) method. The novel CP-SMOTE method generates new samples according to finding several center points, and making a linear combination of the minority samples and their corresponding center points. As another alternative method to avoid noise samples, the IO-SMOTE method divides minority samples into inner and outer parts, and then uses inner samples as much as possible in the subsequent process of generating new samples. The numerical experiments are carried out to compare the CP-SMOTE and IO-SMOTE methods with the no-sampling and conventional SMOTE methods. According to comparing the classification accuracy rate, prediction rate, recall rate, F1-measure and some other indicators, the CP-SMOTE and IO-SMOTE methods have their own advantages, and on the whole, these two methods are much better than the SMOTE method.

The remaining chapters of this paper are organized as follows: The descriptions of the CP-SMOTE and IO-SMOTE methods are given in Section II. And in Section III, some numerical experiments on four datasets and corresponding analysis are carried out after we show the experiment setting. At last, the conclusion is presented in Section IV.

II. CP-SMOTE AND IO-SMOTE METHODS

A. CP-SMOTE METHOD (CENTER POINT SMOTE METHOD)

For solving the imbalanced classification problem, the conventional SMOTE method synthesize several minority points to balance the number of various samples. However, this method blurs the boundary between the majority and minority samples. As shown in Fig. 1, suppose that A is chosen to be an oversampling point, then randomly select point B among the k-nearest neighbor points of A, and randomly generate point C on the connection line between point A and B. However, it is not difficult to see that the neighbor points of C are majority points, and even point C itself might be a majority sample. Therefore, the new sample synthesized by SMOTE method is an extremely unreasonable sample point, which will cause a particularly large error in the subsequent network training and affect the performances of the classifier.

To overcome the above-mentioned shortcomings of the SMOTE method, we propose a new center point-SMOTE (CP-SMOTE) method. First, the k-clustering method [20], [21] is used to find several regions of the minority sample distribution. For each region, calculate the Euclidean center point of all the minority points in the region where they are located. If this distance is less than the distance of any majority sample point to the center point, then randomly select a new point between the minority sample point and the center point; Otherwise, the minority sample point is abandoned.

As shown in Fig. 2, find the two regions where the minority sample is located. For the right region, we calculate the distances of the center point O to all points in this region, and calculate the closest distance d of all the majority sample points to O. For each minority sample D, if the distance between D and O is less than d, then we randomly synthesize a point between D and O; otherwise, D does not participate in synthesizing new sample points. For the left region, the similar process is applied again.

The process is given in Algorithm 1:

Step 1: Divide the imbalanced dataset into majority class samples and minority class samples.

Step 2: The k-clustering method is employed to find n regions and corresponding center point $\{O_1, O_2, \ldots, O_n\}$ of the minority sample

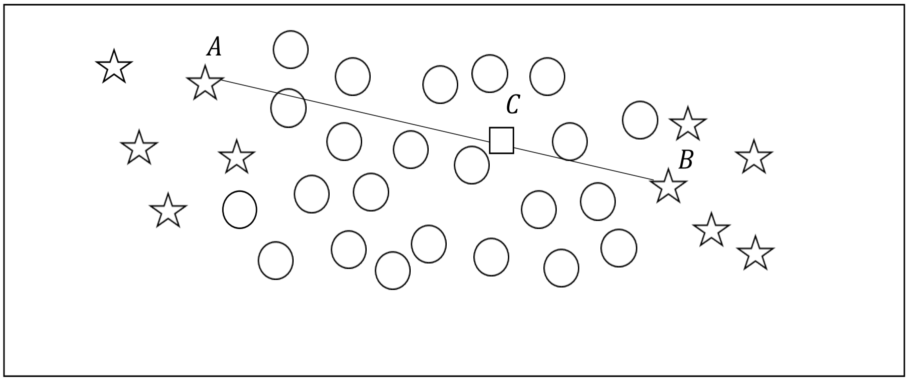


FIGURE 1. Special case of SMOTE method. The stars, circles and square denote the minority samples, majority samples and new synthetic sample, respectively.

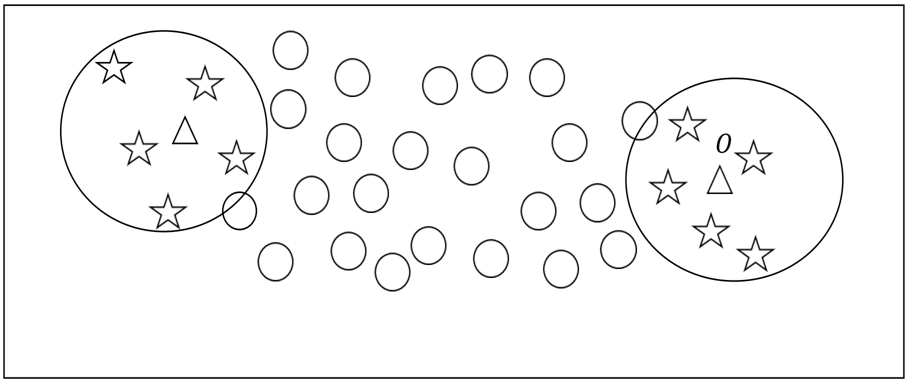


FIGURE 2. CP-SMOTE method. The stars, circles and triangle denote the minority samples, majority samples and center point, respectively.

distribution, where $O_i = \frac{1}{m}\sum_{j=1}^{m} D_{ij}$, D_{ij} is the j-th point in i-th region.

Step 3: For $i = 1$ to $i = n$, calculate the closest distance d_i of all the majority sample points to the points O_i.

Step 4: For each minority class sample P, calculate the distance dis of this sample to its corresponding center point.

Step 5: Compare dis with its corresponding d_i. If $dis < d_i$, then synthesize a point in the following criterion:

$$P_{new} = \eta P + (1 - \eta)O_i, \tag{1}$$

where $0 < \eta < 1$. Otherwise, the point P does not participate in synthesizing new sample points.

Step 6: Put the dataset obtained in steps 2-5 and the original sample set together, and then train the networks.

B. IO-SMOTE METHOD (INNER AND OUTER SMOTE METHOD)

Given an imbalanced dataset including the minority (positive) set M and the majority (negative) set N, $|M| < |N|$. Here $|M|$ and $|N|$ denote the number of M and N, respectively.

For each point $x \in M$, it can be obviously classified into positive class if most neighbor of x is positive. In this case, we call the point x inner point. On the other hand, if most neighbor of x are negative points, the class of point x is not easy to give and point x is denoted by outer point. Therefore, the minority set M is divided into two parts: inner point set and outer point set. Here, the k-nearest neighbor method is applied to find the neighbors of point x. Specifically, select two fixed positive integer c_1 and c_2, where $c_1 < c_2$. For any $x \in M$, if there exists $c \in [c_1, c_2]$ such that the number of positive points in the adjacent points of x exceeds half of $|M|$, then point x is an inner point. Otherwise, point x is an outer point. As shown in Fig. 3, for the minority sample x_1, only one of the six-nearest neighbors is the minority sample, so x_1 is an outer point; On the contrary, for x_2, five of the six-nearest neighbors are minority samples, so x_2 is an inner point.

The process is given in Algorithm 2:

Step 1: Divide the imbalanced dataset into majority set N and minority set M.

Step 2: Divide the minority set M into two parts: Inner set $inner$ and outer set $outer$.

Step 3: In the case of $inner \neq \emptyset$ and $outer \neq \emptyset$, for each point $x \in inner$ then find point $y \in outer$

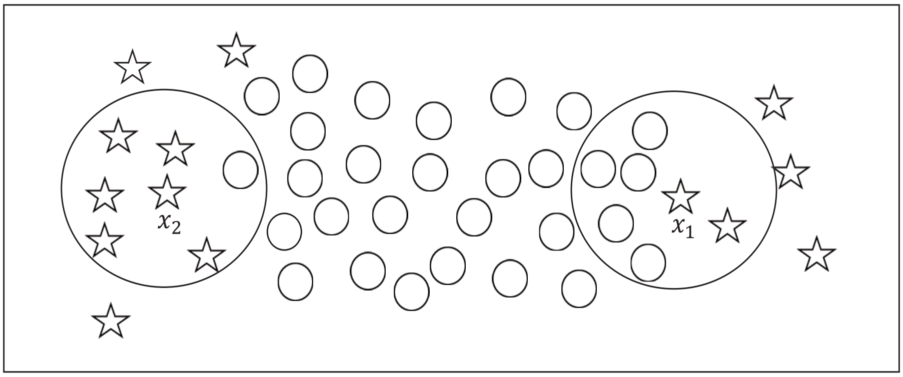


FIGURE 3. IO-SMOTE method. The stars and circles denote the minority and majority samples, respectively. x_1 is an outer point, x_2 is an inner point.

closest to the point x. IO-SMOTE method synthesizes a new point z in the following criterion:

$$z = \eta x + (1 - \eta)y, \tag{2}$$

where $0 < \eta < 1$. In this way, the point number that the IO-SMOTE method synthesizes is equal to the inner point number.

Step 4: For the case *inner* $\neq \emptyset$ or *outer* $\neq \emptyset$, randomly choose three point x_1, x_2 and x_3 from the minority set M. IO-SMOTE method synthesizes a new point z in the following criterion:

$$z = \eta_2 x_1 + (1 - \eta_2)y, \tag{3}$$

where

$$y = \eta_1 x_2 + (1 - \eta_1)x_3, \tag{4}$$

where $0 < \eta_1, \eta_2 < 1$.

Step 5: Put the dataset obtained in steps 3-4 and the original sample set together, and train the networks.

III. NUMERICAL EXPERIMENTS

To verify the validity of the CP-SMOTE and IO-SMOTE methods, we compare them with no-sampling and SMOTE methods on four real classification problems: ecoli1, yeast1, yeast3 and newthyroid1.

A. EXPERIMENT SETTINGS

In our experiments, *five*-fold cross validation technology will be used [22], [23], [24], [25]. For details, the dataset is equally divided into five parts, and the learning process is conducted twenty times. For each time of the training process, each part takes turns as the test set, while the rest as the training set. The above process is repeated twenty times. After adding them all together, *one hundred* classification results are achieved for each method-data pair. The contents in Tabs. 1-3 are obtained by averaging the corresponding 100 results.

We evaluate the class of a sample according to the actual output: If the actual output is less than 0.50, then we regard it as approximately equal to 0 and classify this sample into negative class; Otherwise, if the actual output is more than 0.50, then we regard it as approximately equal to 1 and classify this sample into positive class. Here, the sigmoidal function is employed as activation function:

$$g(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

The experiment process is given in Algorithm 3:

Step 1: Input the imbalanced dataset, the minority (positive) set $\boldsymbol{M} = \{\boldsymbol{m}_j | \boldsymbol{m}_j \in \mathbb{R}^n, j = 1, \ldots, M\}$ and the majority (negative) set $\boldsymbol{N} = \{\boldsymbol{n}_j | \boldsymbol{n}_j \in \mathbb{R}^n, j = 1, \ldots, N\}$.

Step 2: The above four methods are applied to generate positive samples Q to balance the number of positive samples and negative samples, respectively.

Step 3: *Five*-fold cross validation technology: $\boldsymbol{\Phi} = \boldsymbol{M} \cup \boldsymbol{N} \cup \boldsymbol{Q} = \{(\boldsymbol{x}_j, \boldsymbol{o}_j) | \boldsymbol{x}_j \in \mathbb{R}^n, \boldsymbol{o}_j = 0 \; or \; 1, j = 1, \ldots, T\}$ is equally divided into five parts: $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_5$.

Step 4: For $i = 1$ to $i = 5$, do Step 4 to Step 7. Let $\boldsymbol{\Phi}_i$ be the test samples, while $\boldsymbol{\Phi} \setminus \boldsymbol{\Phi}_i$ is the training samples.

Step 5: Train an FNN with the datasets generated by each of the above-mentioned four methods, and test the performances of these four networks.

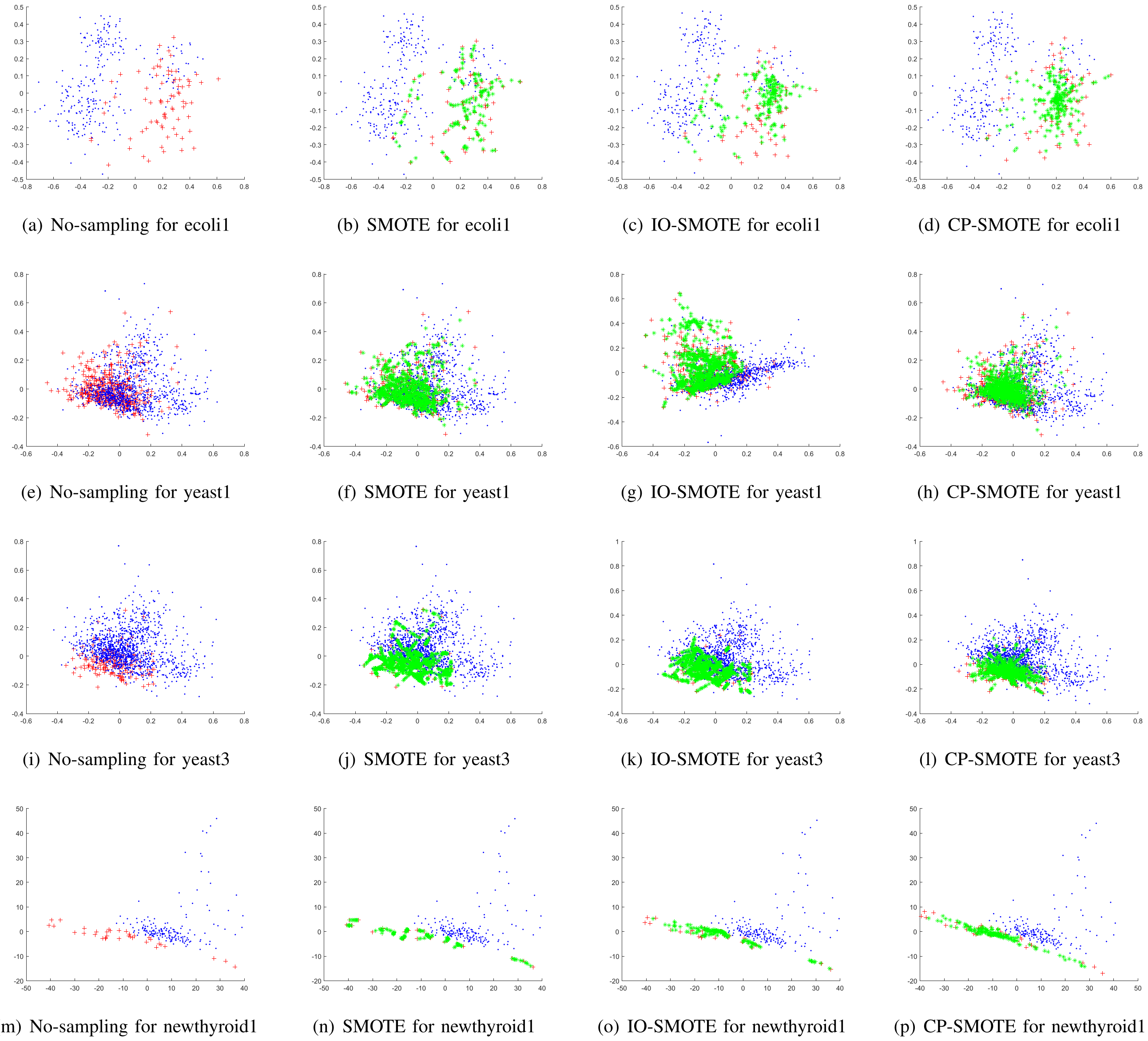
Step 6: Train an ELM with the datasets generated by each of the above-mentioned four methods, and test the performances of these four networks.

Step 7: Repeat the above procedure Steps 3-6 *twenty* times.

Step 8: Compare the *one hundred* experimental results of these four methods.

B. EXPERIMENTAL RESULTS

For these four different datasets, the SMOTE, IO-SMOTE, and CP-SMOTE methods are respectively applied to oversample the minority class samples. For newly generated samples, their characteristics are high-dimensional. To visualize these points, the PCA technique [26], [27] is employed



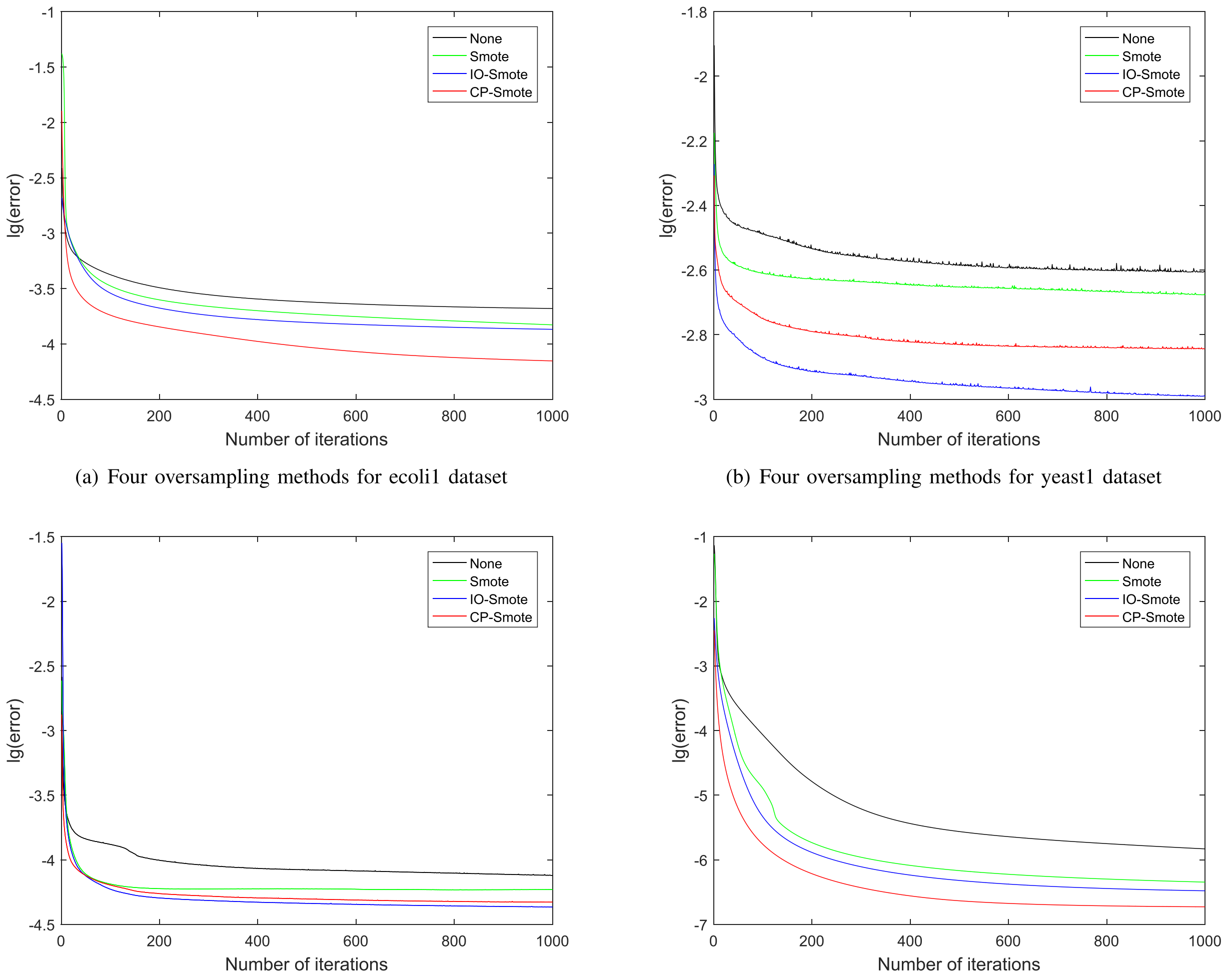
(a) No-sampling for ecoli1
(b) SMOTE for ecoli1
(c) IO-SMOTE for ecoli1
(d) CP-SMOTE for ecoli1
(e) No-sampling for yeast1
(f) SMOTE for yeast1
(g) IO-SMOTE for yeast1
(h) CP-SMOTE for yeast1
(i) No-sampling for yeast3
(j) SMOTE for yeast3
(k) IO-SMOTE for yeast3
(l) CP-SMOTE for yeast3
(m) No-sampling for newthyroid1
(n) SMOTE for newthyroid1
(o) IO-SMOTE for newthyroid1
(p) CP-SMOTE for newthyroid1

FIGURE 4. Discrete point models based on four oversampling methods in two-dimension. Red plus sign represents minority sample points, blue dots denote majority samples, green snowflakes are newly generated samples.

to reduce the dimensionality of the sample points in the n-dimensional to two-dimensional space. The distributions of these points are shown in Fig. 4, where blue represents the majority sample points, red represents the minority sample points, and green represents the synthetic sample points. Obviously, compared with the SMOTE method, the new synthetic sample points of the IO-SMOTE and CP-SMOTE are more compact, especially the CP-SMOTE method. For the CP-SMOTE method, the new synthetic sample points rarely appear near the class boundary, which will make the error smaller in the learning process.

Furthermore, the feedforward neural network (FNN) [28] and extreme learning machine (ELM) [29], [30] are employed to train the original dataset and the new datasets obtained by the above three oversampling methods (cf. Tabs. 1-2). According to these two tables, the IO-SMOTE and CP-SMOTE methods are both better than the no-sampling and SMOTE methods in terms of training and test accuracies. Moreover, the classification accuracies of the CP-SMOTE method are slightly higher than those of the IO-SMOTE method.

At the same time, we compared the error function in the neural network model (cf. Fig. 5). It can be seen that the dataset without oversampling processing has the largest error, while the dataset with SMOTE method has a significant improvement. In addition, the errors of these two novel SMOTE methods are both obviously better than that of the SMOTE method.



(a) Four oversampling methods for ecoli1 dataset

(b) Four oversampling methods for yeast1 dataset

(c) Four oversampling methods for yeast3 dataset

(d) Four oversampling methods for newthyroid1 dataset

FIGURE 5. Error functions based on four oversampling methods for four datasets.

TABLE 1. Classification accuracies for four oversampling methods in ELM.

Dataset	Ecoli1	Yeast1	Yeast3	Newthyroid1
No-sampling:				
Training	99.74%	99.90%	99.88%	99.32%
Test	99.32%	99.79%	99.76%	98.46%
SMOTE:				
Training	99.75%	99.94%	99.91%	99.60%
Test	99.40%	99.82%	99.81%	99.05%
IO-SMOTE:				
Training	99.75%	**99.95%**	**99.92%**	**99.62%**
Test	99.49%	99.83%	99.82%	99.09%
CP-SMOTE:				
Training	**99.78%**	99.94%	99.90%	**99.62%**
Test	**99.54%**	**99.90%**	**99.84%**	**99.23%**

TABLE 2. Classification accuracies for four oversampling methods in FNN.

Dataset	Ecoli1	Yeast1	Yeast3	Newthyroid1
No-sampling:				
Training	99.50%	99.82%	99.78%	99.15%
Test	99.26%	99.70%	99.71%	98.17%
SMOTE:				
Training	99.67%	99.89%	99.83%	99.48%
Test	99.47%	99.77%	99.74%	98.91%
IO-SMOTE:				
Training	99.70%	**99.91%**	99.87%	99.55%
Test	**99.57%**	99.80%	99.78%	99.02%
CP-SMOTE:				
Training	**99.71%**	**99.91%**	**99.89%**	**99.60%**
Test	99.54%	**99.83%**	**99.82%**	**99.13%**

For the purpose of evaluating the error in the learning process of these four methods, besides the classification accuracy, we also compare the following five criteria: prediction rate (PR), recalling rate (RR) [31], F1-measure [32], the standard deviation (σ) [33] and the root mean square error (RMSE) [34]. The specific calculation formulae are as follows:

$$PR := \frac{TP}{TP + FP},$$

$$RR := \frac{TP}{TP + FN},$$

TABLE 3. Five classification criteria for four datasets.

Dataset	Ecoli1	Yeast1	Yeast3	Newthyroid1
No-sampling:				
PR	94.83%	97.67%	97.23%	92.03%
RR	95.11%	97.31%	96.74%	93.58%
F1	94.97%	97.49%	96.98%	92.80%
σ	0.0306	0.0182	0.0245	0.0364
RMSE	0.0207	0.0086	0.0117	0.0240
SMOTE:				
PR	96.14%	98.62%	98.05%	93.71%
RR	95.70%	98.04%	97.60%	93.36%
F1	95.92%	98.33%	97.82%	93.54%
σ	0.0238	0.0146	0.0173	0.0274
RMSE	0.0162	0.0051	0.0068	0.0198
IO-SMOTE:				
PR	**97.69%**	99.17%	98.60%	93.77%
RR	97.16%	98.85%	**98.34%**	94.08%
F1	97.42%	99.01%	98.47%	93.92%
σ	0.0172	0.0113	0.0148	**0.0203**
RMSE	0.0105	0.0029	0.0054	**0.0147**
CP-SMOTE:				
PR	97.45%	**99.26%**	**98.86%**	**94.52%**
RR	**97.80%**	**99.03%**	98.19%	**94.29%**
F1	**97.63%**	**99.15%**	**98.52%**	**94.41%**
σ	**0.0153**	**0.0087**	**0.0120**	0.0217
RMSE	**0.0097**	**0.0023**	**0.0038**	0.0164

$$\sigma := \sqrt{\frac{1}{S-1}\sum_{i=1}^{S-1}(y_i-\overline{y_i})^2},$$

$$RMSE := \sqrt{\frac{1}{S}\sum_{i=1}^{S}(y_i-t_i)^2},$$

$$F1-measure := \frac{2\times PR\times RR}{PR+RR}.$$

And Tab. 3 shows the prediction rate, recall rate, F1-measure, σ and RMSE of these four methods. In these four datasets, the IO-SMOTE and CP-SMOTE methods both have better performances than the no-sampling and SMOTE methods on all these five criteria. Furthermore, in the datasets of Ecoli1, Yeast1 and Yeast3, the CP-SMOTE method performs better than IO-SMOTE method; And for the rest dataset of Newthyroid1, these two methods have their own advantages under different evaluation criteria. Combined with the classification accuracies, the ranking of these four oversampling methods is: CP-SMOTE>IO-SMOTE>SMOTE >No-sampling. SMOTE and these two proposed methods are oversampling methods and do not involve network structure. Only ELM and FNN networks are enough in experiments. In fact, we will obtain similar results under other network models.

IV. CONCLUSION

This paper proposes two novel improved SMOTE methods to generate new samples: Center point SMOTE (CP-SMOTE) method and Inner and outer SMOTE (IO-SMOTE) method. The CP-SMOTE method generates new samples according to finding several center points, and then making a linear combination of the minority samples and their corresponding center points; The IO-SMOTE method divides minority samples into inner and outer samples, and then uses inner samples as much as possible in the subsequent process of generating new samples. Most of the samples generated by these two methods are far away from the classification boundary, which will make error smaller in the process of training the network.

Experiments are conducted for solving four classification problems. The experimental results reveal that the IO-SMOTE and CP-SMOTE methods both have better performances than the traditional SMOTE method.

REFERENCES

[1] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106759.

[2] Q. Li, G. Yu, J. Wang, and Y. Liu, "A deep multimodal generative and fusion framework for class-imbalanced multimodal data," *Multimedia Tools Appl.*, vol. 79, nos. 33–34, pp. 25023–25050, Sep. 2020.

[3] E. Judith and J. M. Deleo, "Artificial neural networks," *Cancer*, vol. 91, no. 8, pp. 1615–1635, 2001.

[4] J.-J. Zhang and P. Zhong, "Learning biased SVM with weighted within-class scatter for imbalanced classification," *Neural Process. Lett.*, vol. 51, no. 1, pp. 797–817, Feb. 2020.

[5] H. Zhu, H. Liu, and A. Fu, "Class-weighted neural network for monotonic imbalanced classification," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 4, pp. 1191–1201, Apr. 2021.

[6] B. Selvalakshmi and M. Subramaniam, "Intelligent ontology based semantic information retrieval using feature selection and classification," *Cluster Comput.*, vol. 22, no. 5, pp. 12871–12881, Sep. 2019.

[7] H. Hu, Q. Wang, M. Cheng, and Z. Gao, "Cost-sensitive semi-supervised deep learning to assess driving risk by application of naturalistic vehicle trajectories," *Exp. Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 115041.

[8] N. M. Faber, "Comment on a recently proposed resampling method," *J. Chemometrics*, vol. 15, no. 3, pp. 169–188, Mar. 2001.

[9] H. Yu, J. Ni, and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, Feb. 2013.

[10] S. Park and H. Park, "Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic," *Computing*, vol. 103, no. 1, pp. 1–24, 2021.

[11] M. A. Tahir, J. Kittler, F. Yan, and K. Mikolajczyk, "Concept learning for image and video retrieval: The inverse random under sampling approach," in *Proc. 17th Eur. Signal Process. Conf.*, 2015, pp. 574–578.

[12] S. Kumar, M. S. Chaudhari, R. Gupta, and S. Majhi, "Multiple CFOs estimation and implementation of SC-FDMA uplink system using oversampling and iterative method," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6254–6263, Jun. 2020.

[13] Y. Yang, S. Fu, and E. T. Chung, "Online mixed multiscale finite element method with oversampling and its applications," *J. Sci. Comput.*, vol. 82, no. 2, pp. 1–20, Feb. 2020.

[14] Y. Pang, Z. Chen, L. Peng, K. Ma, C. Zhao, and K. Ji, "A signature-based assistant random oversampling method for malware detection," in *Proc. 18th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./13th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2019, pp. 256–263.

[15] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj, and S. Razia, "Reducing overfitting problem in machine learning using novel L1/4 regularization method," in *Proc. 4th Int. Conf. Trends Electron. Informat. (ICOEI)*, Jun. 2020, pp. 934–938.

[16] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. Eur. Conf. Knowl. Discovery Databases*, 2003, pp. 107–119.

[17] H. Hui, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Adv. Intell. Comput.*, 2005, pp. 878–887.

[18] M. Naseriparsa, A. Al-Shammari, M. Sheng, Y. Zhang, and R. Zhou, "RSMOTE: Improving classification performance over imbalanced medical datasets," *Health Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–13, Dec. 2020.

[19] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.

[20] R. Aishwarya and V. Nagaraju, "Automatic region of interest based medical image segmentation using spatial fuzzy K clustering method 1," *Int. J. Electron. Commun. Technol.*, vol. 3, no. 1, pp. 226–229, Mar. 2012.

[21] S. Mahak, "Image segmentation with modified K-means clustering method," *Int. J. Recent Technol. Eng.*, vol. 1, no. 2, pp. 176–179, 2012.

[22] T.-T. Wong and N.-Y. Yang, "Dependency analysis of accuracy estimates in k-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2417–2427, Nov. 2017.

[23] P. Jiang and J. Chen, "Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation," *Neurocomputing*, vol. 198, pp. 40–47, Jul. 2016.

[24] J. He and X. Fan, "Evaluating the performance of the K-fold cross-validation approach for model selection in growth mixture modeling," *Struct. Equation Model., Multidisciplinary J.*, vol. 26, no. 1, pp. 66–79, Jan. 2019.

[25] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.

[26] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Autom. Control*, vol. AC-26, no. 1, pp. 17–32, Feb. 1981.

[27] L. E. Pirogov and P. M. Zemlyanukha, "Principal component analysis for estimating parameters of the L1287 dense core by fitting model spectral maps into observed ones," *Astron. Rep.*, vol. 65, no. 2, pp. 82–94, Feb. 2021.

[28] M. Frean, "The upstart algorithm: A method for constructing and training feedforward neural networks," *Neural Comput.*, vol. 2, no. 2, pp. 198–209, Jun. 1990.

[29] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[30] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Feb. 2012.

[31] J. M. DuBois, L. S. Boylan, M. Shiyko, W. B. Barr, and O. Devinsky, "Seizure prediction and recall," *Epilepsy Behav.*, vol. 18, nos. 1–2, pp. 106–109, May 2010.

[32] R. Wang and J. Li, "Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4135–4145.

[33] H. Azami, A. Fernández, and J. Escudero, "Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis," *Med. Biol. Eng., Comput.*, vol. 55, no. 11, pp. 2037–2052, 2017.

[34] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005.

YUAN BAO received the B.S. degree in mathematics and applied mathematics from Henan University, Kaifeng, China, in 2013, and the Ph.D. degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 2020. She is currently a Postdoctoral Fellow with the School of Information Science and Technology, Dalian Maritime University. Her research interests include finite element methods and computer networks.

SIBO YANG received the B.S. and Ph.D. degrees in computational mathematics from the Dalian University of Technology, Dalian, China, in 2013 and 2020, respectively. He is currently a Lecturer with the School of Science, Dalian Maritime University, Dalian. His research interests include extreme learning machine and improvement of learning algorithms in neural networks.