IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 11 October 2023, accepted 26 October 2023, date of publication 2 November 2023, date of current version 7 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329560

RESEARCH ARTICLE

An Adaptive and Robust Method for Oriented Oversampling With Spatial Information for Imbalanced Noisy Datasets

YI DENG[1,2] AND MINGYONG LI[2,3], (Member, IEEE)
[1]School of Data Science, City University of Macau, Macau, China
[2]School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
[3]School of Computer Science and Technology, Donghua University, Shanghai 200051, China

Corresponding author: Mingyong Li (minyonglicnu@163.com)

ABSTRACT Imbalanced datasets have a large negative impact on the classifiers, biasing the classification results towards the majority class. Since imbalanced data distribution is an inevitable and significant challenge in the real world, many variants of SMOTE have been proposed. However, current oversampling methods still need improvement because they rely on hyperparameter optimization, overgeneralize due to emphasizing specific synthetic regions, randomly synthesize samples or suffer from noise performance degradation. To overcome the above problems, we propose an adaptive and robust method (OOSI) for oriented oversampling with spatial information to deal with imbalanced noisy datasets. OOSI is a rare adaptive and effective oversampling method that can fill the gaps of existing methods through dataset-specific spatial partitioning and information quantization, three-stage noise suppression, and spatially-informed generation path improvement. Firstly, a specific and adaptive clustering space is adaptively derived through the data space division of the characteristics of datasets. Then, all minority clusters are assigned a reasonable number of synthetic samples to simultaneously address intra- and inter-class imbalances by integrating the cluster samples' intra-cluster sparsity and the multi-class density information. After differentiating and identifying the noise, oriented weights are assigned based on the multi-class information level to guide the enhancement of the generation path of the synthetic samples and prevent the generation of extra noisy and overlapping samples. Extensive experiments demonstrate that the proposed algorithm outperforms 11 prominent oversampling algorithms on 11 real-world datasets with varying noise levels.

INDEX TERMS Imbalanced learning, label noise, oriented oversampling.

I. INTRODUCTION

Imbalanced noisy learning refers to the problem of training models on datasets that exhibit imbalanced class distributions and contain noisy or mislabeled samples [1]. In this case, there is a significant difference in the number of samples of different classes, which will cause the classifier to be biased towards the majority class in learning while ignoring the characteristics of the minority class, thereby impacting the classification performance [2]. Additionally, the presence of noisy or mislabeled samples further complicates the learning process as they introduce errors and mislead the model during training. The objective of this paper is to propose an adaptive and robust oversampling that adaptively divides and quantizes spatial information to inhibit the intrusion of noise and to guide reasonable sampling path improvement. Data imbalance is common and inevitable in many real-world applications, such as fraud identification [3], medical diagnosis [4], sentiment analysis [5], anomaly detection [6], and other fields. Among them, due to the characteristics of the data itself, certain classes of samples are inevitably challenging to obtain or cost highly. Meanwhile, the minority samples involve important or sensitive information. For example, in rare species recognition or cancer diagnosis, the minority class has a low occurrence rate in real life, but ignoring or misclassifying rare species and cancer will

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Ming Chen.

reduce the generalization ability and robustness of the model, resulting in severe practical consequences [7]. Therefore, how to effectively deal with data imbalance and improve the generalization ability of classifiers is a research topic with crucial theoretical significance and practical value in machine learning.

In-depth research on imbalanced learning has developed numerous algorithms, which can be broadly categorised as cost-sensitive approaches, algorithm-level approaches, and data-level approaches [8]. Cost-sensitive approaches assign higher misclassification costs to minority classes to emphasize the learning of minority classes [9]. Nevertheless, not only do the costs of misclassifying various classes depend on specific data, but it is also frequently difficult to measure precisely. Algorithm-level approaches improve learning in the minority class by enhancing or designing new algorithms to deal with imbalanced issues [10]. Data-level approaches directly manipulate datasets by resampling to equalise class number disparities [11]. As a result of their independence from particular scenarios and classifiers, data-level approaches have become the most prevalent strategies in imbalanced learning.

Data-level approaches replicate or synthesise minority samples (i.e., oversampling), remove majority samples (i.e., undersampling), or combine minority synthesis and majority removal (i.e., hybrid sampling) to balance the quantity of different classes [12]. Although eliminating samples might somewhat reduce the amount of data, undersampling can easily result in the loss of crucial information. Furthermore, by evaluating the area under the ROC curve (AUC), Batista et al. have further shown that undersampling typically performs worse than oversampling [13]. Currently, synthetic minority oversampling technology (SMOTE) is one of the most influential oversampling algorithms, which randomly synthesizes minority samples based on their k-nearest neighbors [14]. Due to its simplicity and efficiency, it has become the established sampling mechanism for subsequent oversampling algorithms.

Meanwhile, some researches have demonstrated that data imbalance is not the only factor that hinders learning. Class overlap, small separation within classes, and label noise can exacerbate the complexity of imbalanced learning, resulting in suboptimal performance [15]. In particular, the unavoidable noise itself often has a substantial influence on the learning process [16]. During model training with label noise, models may learn incorrect or misleading patterns between features and labels, which can lead to decreased accuracy. Additionally, the model could over-adapt to the noise in the training data, failing to discern between true underlying patterns and noisy labels, as a result of which the model does not generalize well beyond the training set [17]. Label noise affects the decision boundary of the classifiers. Moreover, based on the current random sampling mechanism, it is simple to introduce extra noise and overlapping samples, further increasing learning difficulty [18]. In recent years, various sampling algorithms have been proposed from different perspectives, such as noise-filtering approaches, region-emphasizing approaches, clustering-based approaches, etc [19]. Nevertheless, they still have the following drawbacks: (1) Most methods easily introduce extra hyperparameters. (2) Most methods are ineffective at detecting suspect noise and prone to overgeneralization. (3) Most of the current sampling algorithms are based on the random linear sampling mechanism of SMOTE, which is not only limited by its blindness, but also the noise will exacerbates the performance degradation resulting from its blindness.

Given that oversampling plays an important role in Mixup, it improves the learning ability and robustness of traditional models to minority classes. For the absence of minority classes, oversampling compensates the traditional biased models by synthesizing minority samples. Oversampling not only enhances the diversity of the data, it mitigates the bias of the model that tends to predict common classes and avoids overfitting. Given the current challenges of learning difficulty exacerbated by imbalance and label noise, as well as the limitations of current sampling methods that require additional hyperparameter optimization and fail to effectively detect noise, resulting in performance degradation due to blind random sampling. we are committed to exploring an adaptive and robust oversampling method that guides sample synthesis to alleviate blind random sampling and effectively deal with imbalanced noisy learning.

To fill gaps, an adaptive and robust method (OOSI) for oriented oversampling with spatial information is proposed to deal with imbalanced noisy datasets. First, a dataset-specific adaptive spatial partitioning strategy is proposed to effectively fit the data distribution characteristics to obtain a dataset-specific adaptive clustering space. Then, by integrating the intra-cluster sparsity and multi-class density information, the spatial distribution information is adequately quantified and guides the reasonable sample generalization of the cluster space to alleviate both intra- and inter-class imbalance problems. Finally, to avoid noisy samples from introducing additional and chaotic generalization, sample synthesis paths are guided based on the level of multi-class information among non-noisy seed samples, keeping new samples away from chaotic regions. In conclusion, the proposed OOSI approach is expected to deal with imbalanced noisy datasets benefiting from oriented oversampling with spatial information and the innovative three-stage noise suppression strategy. Oriented oversampling with spatial information guides the rational allocation of the number of samples within the cluster and the improvement of the generation path of the synthesized samples to ensure the quality of the synthesized samples. The innovative three-stage noise suppression strategy consists of optimizing the clustering space and avoiding the chaotic expansion of noisy samples and guiding the improved synthesis. The main advantages of OOSI compared to existing methods are that a) It is a rare adaptive and robust oversampling method; b) it can prevent noise hazards with the innovative three-stage noise suppression strategy rather than removing them; c) it can create safe synthetic minority samples with spatial information to avoid overgeneralization and blindness of SMOTE. The following are the main contributions of this paper:

- A spatial partitioning strategy for dataset specificity is proposed to adaptively mine dataset-specific distribution information.
- The proposed OOSI is an adaptive and rare oversampling method. It not only guides reasonable sample generalization and sample synthesis path enhancement through spatial information, but also addresses the common and unavoidable imbalance and noise hazards at the same time.
- Extensive comparative experiments with 11 mainstream sampling algorithms demonstrate the effectiveness and superiority of the proposed OOSI on 11 datasets and 5 classifiers with varying noise levels.

The rest of the paper is organized as follows: Section II briefly reviews relevant literature. Section III presents the details and rationale for the proposed oversampling method. Section IV reports empirical results of extensively contrasting. Section V summarizes our work.

II. RELATED WORK

The oversampling technique, the most widely used strategy in imbalanced learning, enhances the data class distribution by generating new minority samples. In addition, the linear sampling mechanism built on SMOTE is currently the most effective resampling paradigm [20]. Numerous SMOTE-based variations have been developed due to the ubiquity and inevitability of imbalanced noisy applications. Representative approaches include noise-filtering approaches, region-emphasizing approaches, and clustering-based approaches [21].

Filtering-based approaches rely on various noise-filtering strategies to clean data. Based on the invasion of heterogeneous spaces by distinct classes, Batista et al. first proposed employing data-cleaning approaches for oversampling methods to generate balanced datasets with better-defined clusters [13]. The SMOTE-Tomek links and SMOTE-ENN delete samples of multiple categories based on the Tomek links and any sample misclassified by its three nearest neighbours, respectively. Moreover, Yang et al. rectified the sampling results of ant colony clustering by eliminating noisy and overlapping samples with Tomek links cleaning technology [22]. Ramentol et al. proposed SMOTE-RSB based on rough set theory and approximate editing under subsets, which iteratively filters noisy samples from original and synthetic data with similarity thresholds [23]. In addition, S'aez et al. and Ramentol et al. eliminate noisy samples iteratively by iterative partition filters [24] and distinct thresholding strategies based on instance selection in rough set theory [25], respectively. Proper data cleansing is feasible in the presence of noisy or improperly synthesised samples. However, the strategy of iterating or optimizing the threshold is vulnerable to high cost and hyperparameter optimization and limited by practical scenario applications.

To maintain the security of fresh samples and prevent the production of noisy samples, region-emphasizing approaches typically synthesise samples in particular regions. By carefully calculating the ratio of minority samples in the nearest neighbour, safe-level smote emphasises synthesising new samples near bigger safe-level samples, that is, the minority aggregation region [26]. The focus of MWMOTE is on synthesising new samples from informative minority samples near the decision boundary with assigned weights by their majority class distance [27]. Additionally, random space division sampling [28] and constrained oversampling [29] concentrate sampling on the boundary region through the random space division of the complete random forest and the minority class boundary defined by ant colony optimization, respectively. Nevertheless, region-emphasizing approaches are susceptible to over-generalization and might ignore the inherent characteristics of the data. In order to concentrate more on difficult-to-learn samples, He et al. dynamically modify the weights and distribute the number of new samples generated from each minority sample based on the data neighbourhood distribution [30]. Inspired by ADASYN, numerous methods employ similar mechanisms to regulate the number of new artificial instances associated with each minority sample or subset of minority samples [31], [32]. Pan et al. proposed an adaptive sampling method called adaptiveSMOTE. It improves the SMOTE by adaptively selecting the inner and danger areas from the minority class, thereby compiling new minority samples from the selected data, thus preventing the class boundary expansion and enhance the distribution characteristics of the original data [33]. Chen et al. proposed a robust method known as RSMOTE. It identifies non-noisy samples based on the locally salient characteristics of minority samples and reweights the synthetic number of new samples based on their degree of chaos [34].

Clustering-based approaches divide sub-clusters to guarantee the quality of synthetic samples by following the original distribution information. Bunkhumpornpat et al. proposed a sampling algorithm based on a density clustering strategy, DBSMOTE. It synthesizes new samples along the shortest paths between the minority and the pseudo-centroids of arbitrarily shaped clusters found by DBSCAN [35]. Although DBSMOTE has a certain noise resistance due to DBSCAN, dense synthetic samples near the centroid are prone to overfitting. Moreover, Iman et al. proposed A-SUWO, an adaptive semi-unsupervised weighted oversampling method. It clusters minority instances via semi-unsupervised hierarchical clustering and oversamples, considering the distance from the majority class to avoid generating overlapping samples [36]. Douzas et al. combined k-means clustering and SMOTE, namely kmeans-SMOTE. It detects secure clusters with non-overlapping classes across the entire data space by a high proportion of minority samples to prevent noise generation [37]. As well, NI-MWMOTE not only utilizes aggregated hierarchical clustering to prevent ignoring small minority sub-clusters but also eliminates real noise through iteratively suspected noise probabilities and misclassification errors [38]. Nevertheless, DBSMOTE, A-SUWO, NI-MWMOTE, k-means SMOTE require 3, 4, 6 and 9 parameters, respectively. Moreover, region-emphasizing approaches and clustering-based approaches cannot effectively detect and deal with suspicious noises.

Several efforts focus on improving the current mainstream sampling mechanisms. Geometric SMOTE (G-SMOTE) synthesized new samples around geometric regions of the input space as an enhancement to the current data generation mechanism [39]. The SW framework performed weighted sampling by calculating the chaos of the sample space to handle imbalanced noisy datasets [40]. In conclusion, current sampling algorithms continue to have deficiencies when coping with imbalanced, noisy data sets. (1) Additional hyperparameter optimisation restricts most methods. (2) Most methods fail to detect suspicious noise effectively and are prone to over-generalization. (3) Most of the current sampling algorithms are based on the random linear sampling mechanism of SMOTE, which is not only limited by its blindness, but also the noise will exacerbates the performance degradation resulting from its blindness. Therefore, this paper proposes an adaptive and robust method for oriented oversampling with spatial information to simultaneously address the aforementioned issues.

III. PROPOSED METHOD

A. MOTIVATION

The oversampling algorithms that exist now are basically improvements on the SMOTE algorithm. These improvements overcome some of the shortcomings of the SMOTE algorithm though. There are still several drawbacks as follows. (1). inability to fit the distribution characteristics of the data set. (2). large fluctuations in the sampling results since the selection of hyperparameters in the sampling performance. (3). blindness of random linear oversampling lead. (4). SMOTE and most of its variants are often restricted to specific application scenarios or datasets, such as high-dimensional datasets, large datasets or datasets with a large number of noisy samples. Therefore, an adaptive and robust method for oriented oversampling with spatial information is proposed. The purpose is to initially fit the data distribution characteristics by adaptive spatial partitioning of dataset characteristics, and to quantify spatial information to guide reasonable neighborhood generalization and sample synthesis path enhancement. The uncontrollability associated with random linear interpolation is avoided. Few of the sampling algorithms that have been proposed give mathematical models. For the sake of algorithmic soundness, we give the mathematical model and mathematical proof of the algorithm.

B. THE OOSI METHOD

An adaptive and robust method (OOSI) for oriented oversampling with spatial information is proposed to fill those gaps. As is depicted in Figure 1, the OOSI is divided into three stages.

- Adaptive data space partitioning.
- Quantification of spatial information.
- Sampling with oriented information.

TABLE 1. Notaions and definitions.

Notaions	Definitions
$k_{(means)}$	the number of nearest neighbors
D	a data set
D_+	majority samples $\in D$
D_-	minority samples $\in D$
x_i	an sample
C_i	a cluster
D_{c+}	majority samples of a cluster
D_{c-}	minority samples of a cluster
Γ	a set of seed clusters $\in D$, referring to Eq.(6)
P	a condition referring to Def. (1)
Q	a condition referring to Def. (2)
ζ	the sparsity of a cluster, referring to Def. (3)
σ	the multi-class density information, referring to Def. (4)
Θ_C	synthesized numbers of a cluster, referring to Def. (5)
η	a seed sample, referring to Def. (6)

1) ADAPTIVE DATA SPACE PARTITIONING

The data space is divided adaptively according to the own characteristics of the dataset to fit the data space distribution characteristics initially. The initial spatial partitioning can be achieved by clustering, rather than simply dividing the data space into majority and minority classes. Any clustering method can be used for this step, e.g. k-means, dbscan, spectral clustering. Different distance metrics can also be chosen. Furthermore, the number of clusters of specificity is determined by the own characteristics of the dataset. The number of adaptive clusters k for dataset specificity is defined as following.

$$k = \log_2 \left\{ (|D_-| + |D_+|) * \frac{D_-}{D_+} \right\} \tag{1}$$

Therefore, the larger the dataset and the more unevenly distributed the data are, the more spatial partitioning is required to capture more detailed spatial information of the data. Taking kmeans and Euclidean distance as an example,the distance between two samples is $d(x_i, x_j)$. Then define the sum of the distances between the sample and the center of the cluster as the loss function. The loss function $W(C)$ is defined as following.

$$W(C) = \sum_{l=1}^{k} \sum_{C_i=l} ||x_i - \bar{x}_j||^2 \tag{2}$$

where $\bar{x}_l = (\bar{x}_{1l}, \bar{x}_{2l}, \cdots, \bar{x}_{ml})^{\mathrm{T}}$ is the mean or center of the i-th class, $n_l = \sum_{i=1}^{n} I(C(i) = l)$. $I(C(i) = l)$ is an indicator

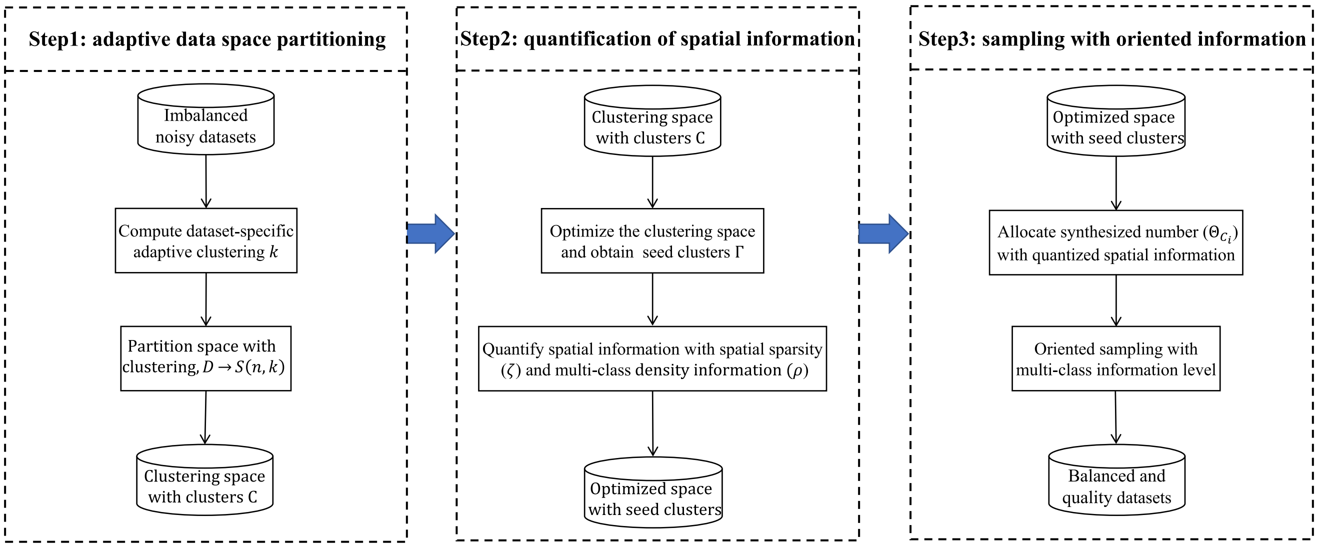

FIGURE 1. Three stages of oriented oversampling with spatial information.

function. The value of I is 1 or 0. Kmeans is to solve the following optimization model.

$$
\begin{aligned}
C^* &= \arg\min_C W(C) \\
&= \arg\min_C \sum_{l=1}^{k} \sum_{C(i)=l} \|x_i - \bar{x}_l\|^2 \qquad (3)
\end{aligned}
$$

When similar samples are clustered into the same cluster, the loss function gets the optimal solution. This is a combinatorial optimization problem. n samples are clustered into k clusters, and there are $S(n, k)$ clustering results.

$$S(n,k) = \frac{1}{k!}\sum_{l=1}^{k}(-1)^{k-l}\binom{k}{l}k^n \qquad (4)$$

In imbalanced data sets, the optimization model for clustering has some additional constraints. There are no overlapping regions for each cluster. In order to reduce the interference of outlier samples to clustering, the number of samples in each cluster is bigger than k. The final mathematical model is shown in Eq. (5).

$$
\begin{aligned}
\min_{C_1, C_2\cdots, C_k} & \sum_{i=1}^{k}\sum_{x\in C_i}\|x-c_i\|^2, \\
\text{s.t.} \quad & C_1 \cup C_2 \cup \cdots \cup C_k = \{x_1, x_2, \cdots, x_n\}, \\
& C_i \cap C_j = \varnothing, \quad \forall i \neq j, \\
& |C| > k. \qquad (5)
\end{aligned}
$$

2) QUANTIFICATION OF SPATIAL INFORMATION

To better quantify spatial information, after obtaining an adaptive and dataset-specific clustering space, a simple and effective optimization of the clustering space is necessary. If the imbalanced distribution of clusters (IR_C) has been mitigated or the number of minority class samples within clusters (D_{c-}) is insufficient to synthesize new samples, it will not be necessary or appropriate to synthesize new samples within these clusters. Therefore, the clustering space will be filtered to obtain seed clusters.

$$\Gamma = \{C | IR_C > 0.5 * IR \ \ and \ \ |D_{c-}| \geq k_{nn}|\} \qquad (6)$$

The seed clusters will serve as the optimized clustering space and will synthesize new samples in the seed clusters. Therefore, the filtering conditions for seed clusters are two as follows.

Definition 1 (Condition P): The proportion of minority samples in a cluster should be greater than half of the imbalanced proportion(IR) of the original dataset. The imbalance ratio of a cluster is denoted by IR_C.

$$IR_C > 0.5 * IR \qquad (7)$$

Definition 2 (Condition Q): The number of minority samples (D_{c-}) in a cluster must be bigger than the number of k nearest neighbor(k_{nn}) samples. In the stage of synthesizing samples, it needs to be used to interpolate between the seed sample and the neighbor samples. If there are too few samples in the cluster, overlapping samples will be synthesized, resulting in overfitting.

$$|D_{c-}| \geq k_{nn} \qquad (8)$$

Eligible clusters are called seed clusters. In order to effectively quantify and utilize spatial information, two quantitative techniques are proposed within the optimized clustering space. Specifically, to avoid overgeneralization of region-emphasizing sample synthesis and intra-class imbalance, the two quantitative techniques measure the distributional characteristics of clusters in terms of spatial sparsity and multi-class distributional information, respectively. Then the number of sample synthesis within clusters are reasonably allocated. These two techniques are sparsity and multi-class density information of the cluster. The corresponding definitions and calculation method has been given in the Def. (3) and Def. (4).

Definition 3 (Sparsity, ζ): Given a cluster C with center x_c and the minority set D_{c-} with n_- samples and the majority set D_{c+} with n_+ samples. $d(x_i, x_j)$ is the distance of any two samples x_i and x_j. The radius r of cluster C is the average of the deviations of all minority samples' distances to the cluster center. The sparsity of a cluster (ζ) is the reciprocal of the number of minority samples per unit area in the cluster, as defined below.

$$r = \frac{\sum d(x_c, x_i)}{n_-}, x_i \in D_{c-} \tag{9}$$

$$\zeta = \frac{\pi * r^2}{n_-} \tag{10}$$

Specifically, the space within a cluster is quantified as the simulated area based on the mean of the intra-cluster distance deviation, i.e., the radius. After that, the number of minority classes in a given space, i.e., the minority density, is quantized. Thus, intra-cluster sparsity based on spatial sparsity quantifies the sparsity of minority classes per unit of cluster space. The more minority classes per unit cluster space, the smaller the value of intra-cluster sparsity (ζ). The fewer the minority samples in the unit cluster space, the larger the value of intra-cluster sparsity (ζ). Therefore, intra-cluster sparsity (ζ) considers the difference in the spatial sparsity sparsity of clusters and requires allocating a reasonable number of samples for generalization. The sparser the samples within a cluster, the more samples can be allocated to generalize the space, and the denser the samples within a cluster, the fewer samples can be allocated to avoid overgeneralization or generation of overlapping samples with low values.

Definition 4 (Multi-Class Density Information, ρ): Given a cluster C with center x_c and the minority class $c-$ and the majority class $c+$. The absolute density information within a cluster, i.e., $\sigma(c.)$, indicates how densely the samples of a certain class are distributed within the cluster. The larger the absolute density information within a cluster, the more dense the distribution of such samples are. The multi-class density information, i.e., ρ, fully integrates the intra-class and inter-class distribution information within a cluster, as follows.

$$\sigma(c.) = \frac{n_-}{\sum d(x_c, x_i)}, x_i \in D_{c.} \tag{11}$$

$$\rho = \frac{\sigma(c-)}{\sigma(c+)} = \frac{n_- / \sum d(x_c, x_i)(x_i \in D_{c-})}{n_+ / \sum d(x_c, x_i)(x_i \in D_{c+})} \tag{12}$$

Specifically, according to Eq. 11, the absolute density information is the number of class per unit distance from the cluster center x_c to the given class. Thus for any given cluster, the farther the cluster center x_c is from the unit distance to the given class, the sparser the distribution of the given class within the cluster; conversely, the denser it is. Then, information about the spatial distribution of a given class within the cluster is quantized. In addition, based on the ratio of the cluster center x_c to the multiple classes of homogeneous and heterogeneous samples, the difference distribution information within the cluster space for multiple classes is quantified. The multi-class density information of the cluster ρ is a composite measure of the multi-class distribution information within the cluster to synthesize more samples in clusters far from the dense distribution of heterogeneous classes. By definition 4, the larger the value of $\sigma(c-)/\sigma(c+)$, the relatively denser the distribution of minority classes and the relatively sparser the distribution of majority classes within the cluster. Therefore, the larger the multi-class density information within a cluster (ρ) indicates, the closer it is to the densely distributed minority class and away from the densely distributed majority class.

3) SAMPLING WITH ORIENTED INFORMATION

During the sampling process, on the one hand, most current sampling algorithms emphasize synthesizing more samples in specific regions or synthesizing the same samples in distinct regions. It not only tends to lead to overgeneralization in specific regions, but also suffers from sample generalization blindness. Few sampling algorithms consider spatial information to guide the oriented allocation of the number of synthesized samples for different clusters. Therefore, after obtaining the simple optimized clustering space and the quantitative measures of cluster distribution characteristics, the sampling number within minority clusters is assigned reasonably to avoid overgeneralization of specific regions and to alleviate both intra- and inter-class imbalances. The number of synthesis per cluster with oriented information Θ_C is defined as follows.

Definition 5 (Synthesized Number, Θ_C): Given a dataset $D = D_+ \cup D_-$, in which the majority and minority samples are $D+$ and $D-$, respectively. C_i represents one of the seed clusters ($i = 1, \ldots |\Gamma|$). The normalization of the sparsity and multi-class density information of C_i are $norm(\zeta(C_i))$, $norm(\rho(C_i))$. The specific calculation formula for the number of new samples of Ci (Θ_{C_i}) is as following:

$$G(D) = |D_+| - |D_-| \tag{13}$$

$$\Theta_{C_i} = G(D) * \frac{norm(\zeta(C_i)) * norm(\rho(C_i))}{\sum_1^{|\Gamma|} norm(\zeta(C_i)) * norm(\rho(C_i))} \tag{14}$$

During the sampling process, on the other hand, most current sampling algorithms either fail to detect and handle noise efficiently or rely on noise filtering mechanisms that require iteration and optimization. Not only do they suffer from noise-induced performance deterioration, but most of the current sampling algorithms are blind based on the smote random sampling mechanism. Blind generalization of seed samples and selection of nearest neighbors can extend the performance deterioration caused by noise generalization. Furthermore, the blind synthesis position between samples tends to introduce more chaotic samples. Thus, OOSI not only detects suspicious noise and prevents noise expansion by avoiding the selection of noisy samples as seed samples through neighborhood space information. The neighbor space of a suspicious noise sample tends to be distributed with more heterogeneous samples. It indicates that its neighbor space has been invaded by majority classes, then it is prone to synthesize new confusion samples or noise samples. Thus according to Eq. 15, the set of non-noise seed samples is obtained. Also, the oriented weights are assigned based on the multi-class information level of the seed samples to guide the generation path improvement of synthetic samples and avoid generating additional chaotic samples. Thus, the seed samples (η) and the synthetic new samples (s_{new}) are defined as follows.

Definition 6 (Seed Samples, η): Given a cluster C_i, each sample $s \in C_i$ has minority neighbors D_{k-} and majority neighbors D_{k+}. The seed samples η of cluster C_i and their oriented weights $\omega(s)$ are defined exactly as following.

$$\eta = \{s||D_{k-}| > |D_{k+}||\} \tag{15}$$

$$\omega(s) = \frac{\sigma(k-)}{\sigma(k+)} = \frac{|D_{k-}|/\sum d(s, s_j)(s_j \in D_{k-})}{|D_{k+}|/\sum d(s, s_j)(s_j \in D_{k+})} \tag{16}$$

Definition 7 (New Samples, s_{new}): Given any two seed samples ss and cs, where $\omega(ss) > \omega(cs)$, the reasonable synthesis of a new sample s_{new} with oriented weights is as follows. ξ is the random number between [0, 1] to maintain the randomness.

$$s_{new} = ss + (cs - ss) * \frac{min(\omega(ss), \omega(cs))}{\omega(ss) + \omega(cs)} * \xi \tag{17}$$

Specifically, according to Eq. 16, Specifically, the multi-class information level of the non-noise seed samples are based on their k-neighborhoods, which portray their local characteristics. According to Eq. 16, the multi-class information level of a non-noise seed sample is determined by calculating the ratio of the distances of the number of classes per unit distance between the k homogeneous and heterogeneous neighbors. The multi-class information level of the non-noise seed samples integrates the multi-class distribution information of the samples and fully reflects the information of homogeneous and heterogeneous samples within the neighborhoods in order to efficiently differentiate and guide diverse sampling. Therefore, the larger the value of $\omega(s)$, the closer the seed sample is to homogeneous samples, the further it is from heterogeneous samples, and the safer it is.

According to the multi-class information level of the seed samples, it is reasonable toguide the synthetic synthesis path improvement, which makes the new samples close to the safe region and away from the chaotic region. As in Figure 2 (a), ss is for any seed sample, and neighbors $nn = [n_1, n_2, n_3](k = 3)$ is its near neighbor sample. According to the information of the distribution of homogeneous and heterogeneous samples in the neighborhood, it is easy to see the discrepancy that exists in different seed samples. According to Eq. 16, the multi-class information level between samples can be calculated i.e. $\omega(n_1) < \omega(n_2) < \omega(ss) < \omega(n_3)$. When synthesizing new samples, if based on ss and n_1, since$\omega(n_1) < \omega(ss)$, the safer sample at this time is ss, the position of the synthesized sample is closer to the ss, as shown in Figure 2 (b). if based on ss and n_3, since$\omega(ss) < \omega(n_3)$, the safer sample at this time is n_3, the position of the synthesized sample is closer to the n_3. Therefore, the multi-class information level of the non-noise seed samples effectively guides the improvement of the synthetic sample generation path, ensures the synthetic quality of the new samples, and avoids the confusion introduced by blind generalization.

Algorithm 1 The OOSI Algorithm.

```
Input: Imbalanced dataset D=D+∪D−, |D−|: the minority samples, |D+|:
    the majority samples, n= |D−| + |D+|
Output: Balanced dataset D′.
    //FIRST : Adaptive data space partitioning.
    Computing dataset specificity's adaptive clusters k according to Eq. 1.
    Division space with the clustering strategy, D → S(n, k).
    clustering(k)
    //SECOND: Quantification of spatial information.
    Initialize Γ = ∅
    for C_i in S(n, k) do
        if IR_C > 0.5 ∗ IR and |D_c−| ≥ k_nn then
            Γ = Γ ∪ C_i
            Calculate sparsity ζ according to Def. (3).
            Calculate multi-class density information ρ according to Def. (4).
        end if
    end for
    //Third Sampling with oriented information.
    Initialize D_new = ∅
    for C_i in Γ do
        Calculate Θ_{C_i} by Eq. (14)
        for s in C_i do
            if s is η judge by Eq. (15) then
                Select nearest neighbor sample of s_j;
                Synthesize a new sample s_new between s and neighbours;
                D_new = D_new ∪ s_new;
            end if

        end for
    end for
    D′ = D+∪D−∪D_new
    Return Balanced dataset D′.
```

C. TIME COMPLEXITY ANALYSIS

The time complexity of the proposed method is determined by three main parts: adaptive data space partitioning,quantification of spatial information and sampling with oriented information. Given a dataset D containing N samples and n minority samples. For adaptively partitioning data space with specificity k, The time complexity of the kmeans clustering is $O(k * t * N)$, where t is the constant number of iterations. For quantifying spatial information, the time complexity of computing the cluster sparsity and multi-class density information are less than $O(k * n)$. For sampling with oriented information, The time complexity of distributing the number of synthetic samples within a cluster and synthesizing new samples is no greater than $O(k)$ and $O(n*n)$, respectively. Therefore, the overall time complexity of the proposed method is $O(k * t * N + n^2)$.

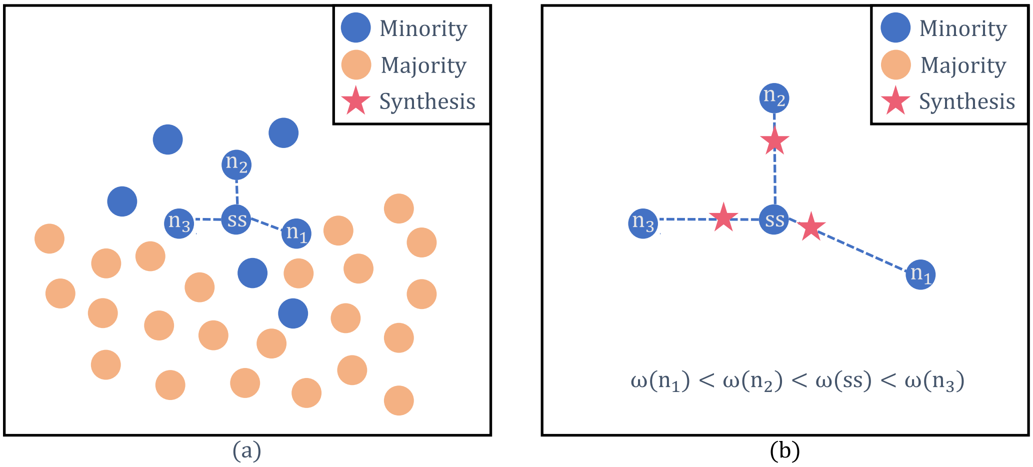

FIGURE 2. Schematic diagram of the synthesized new samples. (a) Original data distribution, (b) Location of sample synthesis.

IV. EXPERIMENTS

A. EXPERIMENTAL DATASETS

In order to evaluate the effectiveness of the proposed method, 11 real-world application datasets are obtained from the UCI and KEEL dataset repositories [41]. Moreover, the one versus others data processing strategy is employed to restructure the imbalanced multi-class datasets into binary classes. The specific information of all datasets is shown in Table 2. Among them, Minority, Majority, Features, and IR denote the number of minority class samples, the number of majority class samples, the number of attributes, and the imbalance ratio. The imbalance ratio equals Majority divided by Minority. In order to effectively evaluate the performance and stability of the method, hierarchical ten-fold cross-validation is used for data division to maintain the consistency of the distribution characteristics and imbalance ratio of each class. Furthermore, to validate the robustness of the proposed method, the dataset is manually added with varying levels of flip noise. Specifically, the same number of samples ($n * nl$) from minority and majority classes are randomly flipped into heterogeneous classes, where n is the number of minority class and nl is the noise level ($0\% \leq nl \leq 20\%$).

TABLE 2. The specific information of all datasets.

Datasets	Samples	Features	Majority	Minority	IR
ecoli	336	7	284	52	5.46
haberman	306	3	225	81	2.78
creditApproval	459	15	383	76	5.04
wisconsin	488	9	444	44	10.09
breastcancer	532	10	444	88	5.05
pima	768	8	500	268	1.87
segment	2310	16	1980	330	6.00
mushrooms	4628	112	4208	420	10.02
page-blocks0	5472	10	4913	559	8.79
svmguide1	4400	4	4000	400	10.00
magic	13565	10	12332	1233	10.00

B. EVALUATION METRICS

Frequently, the proportion of positive (minority) and negative (majority) samples is disproportionate in practical applications. Currently, traditional classification evaluation metrics such as accuracy rate and error rate may be misleading and cannot effectively reflect the model's true performance. Since these metrics are more biased toward the predictions of the majority classes, they are insensitive to the prediction errors of the minority classes [42]. Therefore, specific evaluation metrics for imbalanced datasets, such as F-measure, G-mean, and AUC, are required [43].

The F-measure is the harmonic mean of precision and recall, which can reflect the model's predictive ability for the minority class. The precision refers to the proportion of truly positive samples in the predicted positive examples, and the recall refers to the proportion of truly positive samples predicted to be positive. The higher the F-measure value, the better the model can correctly classify the positive samples. The calculation of F-measure is as formula (18).

$$\mathrm{F-measure} = \frac{2 * \mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}, \tag{18}$$

The G-mean is the geometric mean of sensitivity (recall) and specificity, reflecting the model's predictive performance for positive and negative samples. Specificity refers to the proportion of truly negative samples predicted to be negative. The calculation method of the G-mean considers the true class rate and the true negative class rate. Therefore, the higher the G-mean, the more balanced the model's ability to identify positive and negative samples. The calculation of G-mean is as formula (19).

$$\mathrm{G-mean} = \sqrt{\mathrm{Sensitivity} * \mathrm{Specificity}}. \tag{19}$$

The AUC is the area under the ROC curve, which can reflect the ability of the model to distinguish positive and negative examples under different thresholds. Among them,

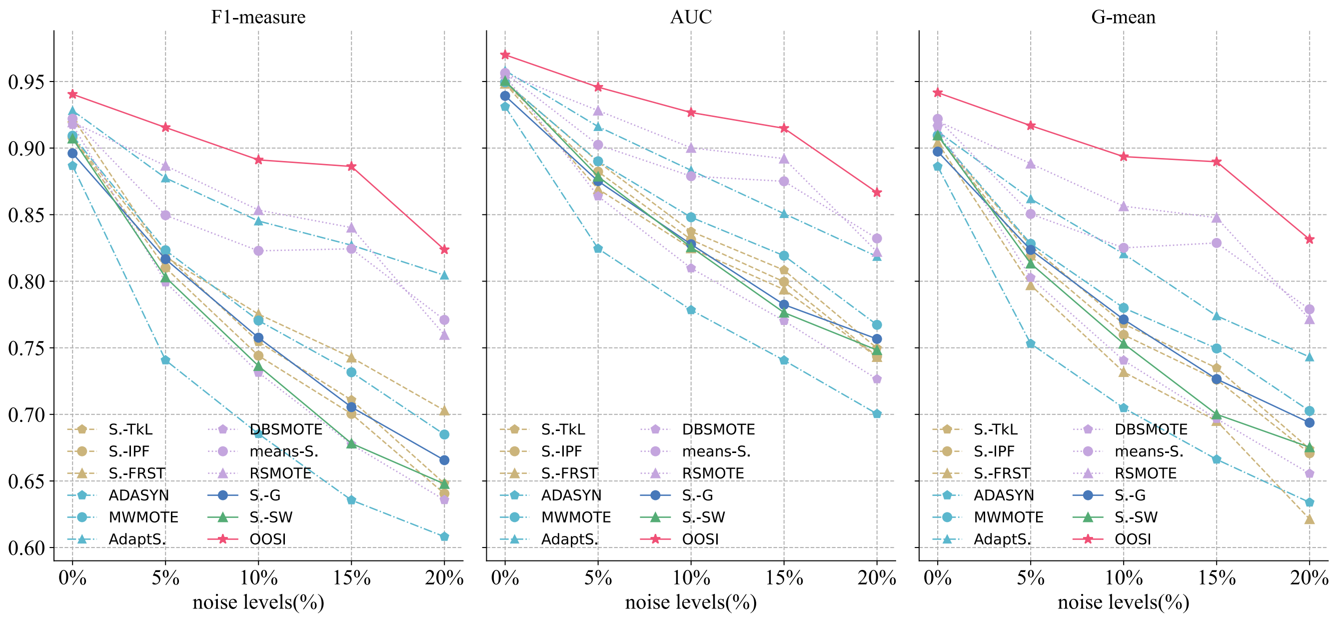

FIGURE 3. The average results of five classifiers on 11 real-world datasets for method comparison with varying noise levels ($0\% \le nl \le 20\%$).

the ROC curve is a curve with the false positive rate as the horizontal axis and the true positive rate as the vertical axis. The larger the AUC value, the better the classification performance of the model for different classes. The calculation of AUC is as formula (20).

$$\mathrm{AUC} = \frac{\mathrm{Sensitivity} + \mathrm{Specificity}}{2}. \tag{20}$$

C. COMPARATIVE OVERSAMPLING METHODS

To demonstrate the superiority of the proposed method, 11 related oversampling methods with distinct strategies are compared, and sample synthesis is conducted on twelve real-world application datasets. SMOTE-TomekLinks (S.-TkL) [13], SMOTE-IPF (S.-IPF) [24], and SMOTE-FRST-2T (S.-FRST) [25] are noise-filtering techniques. ADASYN [30], MWMOTE [27], and Adaptive-SMOTE (AdaptS.) [33] are region-emphasizing approaches. DBSMOTE [35], kmeans-SMOTE (means-S.) [37], and RSMOTE [34] are competitive clustering-based methods. Geometric SMOTE (G-SMOTE) [39] and SMOTE-SW (S.-SW) [40] are enhanced sampling mechanisms. The core idea of their algorithm has been presented in Section II. All sampling methods are provided with concrete implementations by the SMOTE-variant library [44] or respective authors. To verify the general validity and stability of the proposed method, five mainstream classifiers were used to evaluate the classification performance, including logistic regression (LR), support vector machine (SVM), adaptive boosting (AdaBoost), Gradient Boosted Decision Trees (GBDT), and Backpropagation Neural Networks (BPNN).

D. COMPARATIVE EXPERIMENTS ON REAL DATASETS

To verify the superiority of the proposed method when coping with imbalanced and noisy datasets, 11 sampling methods with distinct strategies are compared on five classifiers and 11 real-world datasets. These datasets involve different samples, features, and imbalance ratios to comprehensively evaluate the performance of different methods under various data distributions. Furthermore, to demonstrate the robustness of the proposed method, different levels of flip noise are randomly introduced into each experimental datasets ($nl \in \{0\%, 5\%, 10\%, 15\%, 20\%\}$).

The average results of method comparisons averaged over 5 classifiers and 11 datasets are shown in Figure 3. Each subplot in Figure 3 depicts the variation trend of the ten methods with increasing noise level for a particular metric. The solid red line with a five-pointed star represent the proposed method OOSI. Each color represents a comparative algorithm for a group of strategies.

In general, the solid red line with a five-pointed star is always above other lines for all metrics, noise levels, and methods. It demonstrates that OOSI consistently outperforms its competitors, regardless of the metrics and noise levels. It is common knowledge that the presence of noise impairs decision-making, leading to performance degradation. When noise is present, i.e., at 5% noise level, the performance of other lines (comparison method) decreases significantly, particularly for ADASYN. However, there is a slight decrease in OOSI performance and a larger performance improvement in OOSI compared to Noiseless. Notably, as the level of noise increases, OOSI achieves greater enhancement than most contrasting methods, particularly filtering-based methods, and region-emphasizing

TABLE 3. Mean and variance of 12 oversampling methods under 5 classifiers on each real-world dataset ($nl = 0\%$).

Metrics	Datasets	S.-TkL	S.-IPF	S.-FRST	ADASYN	MWMOTE	AdaptS.	DBSMOTE	means-S.	RSMOTE	S.-G	S.-SW	OOSI
F1-measure	ecoli	0.937±0.03	0.945±0.02	0.942±0.02	0.898±0.04	0.937±0.03	0.942±0.02	0.961±0.02	**0.965±0.02**	0.947±0.02	0.933±0.03	0.944±0.03	0.963±0.02
	haberman	0.687±0.07	0.693±0.08	0.721±0.07	0.638±0.08	0.693±0.08	**0.823±0.04**	0.685±0.06	0.725±0.06	0.682±0.06	0.653±0.07	0.637±0.09	0.785±0.07
	creditApproval	0.909±0.03	0.908±0.03	0.937±0.01	0.889±0.03	0.915±0.03	0.924±0.03	0.922±0.02	0.914±0.03	0.931±0.02	0.898±0.03	0.917±0.03	0.938±0.03
	wisconsin	0.988±0.01	0.988±0.01	**0.994±0.01**	**0.990±0.01**	**0.990±0.01**	0.987±0.01	0.988±0.01	0.985±0.01	0.971±0.01	0.985±0.01	**0.992±0.01**	0.989±0.01
	breastcancer	0.979±0.01	0.980±0.01	**0.990±0.01**	0.979±0.01	0.978±0.01	**0.985±0.01**	0.981±0.01	**0.983±0.01**	0.969±0.02	0.977±0.01	**0.992±0.01**	0.981±0.01
	pima	0.795±0.04	0.782±0.04	**0.838±0.02**	0.755±0.03	0.785±0.03	0.787±0.04	0.802±0.02	0.748±0.04	0.775±0.04	0.753±0.04	**0.818±0.04**	0.806±0.04
	segment	0.947±0.01	0.949±0.01	0.947±0.01	0.938±0.01	0.948±0.01	0.957±0.01	0.948±0.01	0.959±0.01	0.952±0.01	0.939±0.01	0.952±0.01	0.960±0.01
	svmguide1	0.975±0.01	0.974±0.01	0.973±0.01	0.960±0.01	0.968±0.00	0.975±0.00	0.977±0.00	0.980±0.00	0.984±0.00	0.960±0.01	0.974±0.00	0.986±0.00
	mushrooms	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	0.996±0.00	1.000±0.00	1.000±0.00	0.999±0.00	1.000±0.00	1.000±0.00
	page-blocks0	0.935±0.01	0.933±0.01	0.930±0.01	0.914±0.01	0.934±0.01	0.930±0.01	0.947±0.01	0.956±0.01	0.960±0.01	0.930±0.01	0.941±0.01	0.972±0.00
	magic	0.834±0.01	0.834±0.01	0.881±0.00	0.792±0.01	0.856±0.01	0.900±0.01	0.887±0.00	0.928±0.00	0.943±0.00	0.832±0.01	0.813±0.01	0.965±0.00
	Average	0.908±0.02	0.908±0.02	0.923±0.02	0.887±0.02	0.909±0.02	0.928±0.02	0.918±0.02	0.922±0.02	0.919±0.02	0.896±0.02	0.907±0.02	0.940±0.02
	Win-Lose	0–11	0–11	3–8	1–10	1–10	2–9	0–11	2–9	0–11	0–11	3–8	N/A
AUC	ecoli	0.971±0.02	0.974±0.02	0.973±0.01	0.945±0.03	0.971±0.02	0.975±0.02	0.981±0.02	0.979±0.02	0.976±0.02	0.966±0.02	0.969±0.02	0.988±0.01
	haberman	0.796±0.05	0.786±0.07	0.768±0.08	0.735±0.07	0.781±0.06	0.828±0.05	0.776±0.05	0.812±0.05	0.778±0.06	0.734±0.07	0.758±0.07	0.860±0.05
	creditApproval	0.953±0.02	0.953±0.02	0.963±0.02	0.945±0.02	0.962±0.02	0.968±0.02	0.973±0.02	0.958±0.02	**0.974±0.02**	0.946±0.02	0.965±0.02	0.973±0.02
	wisconsin	0.997±0.00	0.997±0.00	0.997±0.00	0.997±0.01	0.997±0.00	0.997±0.00	0.998±0.00	0.997±0.00	0.981±0.01	0.996±0.00	0.999±0.00	0.999±0.00
	breastcancer	0.995±0.01	0.996±0.01	0.998±0.00	0.993±0.01	0.996±0.01	0.996±0.00	0.997±0.00	0.994±0.01	0.985±0.01	0.994±0.01	**0.999±0.00**	0.998±0.00
	pima	0.872±0.03	0.867±0.03	0.873±0.03	0.835±0.03	0.858±0.03	0.864±0.03	0.873±0.02	0.831±0.04	0.863±0.03	0.845±0.03	**0.911±0.03**	0.886±0.03
	segment	0.978±0.01	0.980±0.00	0.978±0.00	0.969±0.01	0.977±0.01	0.985±0.00	0.982±0.01	0.988±0.00	0.988±0.00	0.976±0.01	0.981±0.00	0.989±0.00
	svmguide1	0.996±0.00	0.996±0.00	0.996±0.00	0.991±0.00	0.995±0.00	0.996±0.00	0.997±0.00	0.997±0.00	**0.999±0.00**	0.992±0.00	0.996±0.00	0.998±0.00
	mushrooms	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00
	page-blocks0	0.976±0.00	0.975±0.00	0.972±0.00	0.957±0.01	0.975±0.00	0.974±0.00	0.979±0.00	0.989±0.00	0.990±0.00	0.972±0.00	0.979±0.00	0.992±0.00
	magic	0.914±0.00	0.914±0.00	0.914±0.00	0.876±0.00	0.932±0.00	0.956±0.00	0.942±0.00	0.974±0.00	0.981±0.00	0.910±0.00	0.898±0.00	0.988±0.00
	Average	0.950±0.01	0.949±0.01	0.948±0.01	0.931±0.01	0.949±0.01	0.958±0.01	0.954±0.01	0.956±0.01	0.956±0.02	0.939±0.01	0.956±0.01	0.970±0.01
	Win-Lose	0–11	0–11	0–11	0–11	0–11	0–11	0–11	0–11	2–9	0–11	2–9	N/A
G-mean	ecoli	0.935±0.03	0.943±0.03	0.940±0.02	0.897±0.04	0.935±0.03	0.932±0.03	0.959±0.02	**0.964±0.02**	0.947±0.02	0.930±0.03	0.944±0.02	0.962±0.02
	haberman	0.706±0.06	0.716±0.07	0.687±0.08	0.646±0.08	0.713±0.07	0.699±0.06	0.678±0.08	0.739±0.05	0.698±0.06	0.675±0.06	0.666±0.08	0.796±0.06
	creditApproval	0.902±0.03	0.899±0.03	0.900±0.02	0.882±0.03	0.908±0.03	0.910±0.03	0.919±0.02	0.905±0.03	0.929±0.03	0.890±0.04	0.910±0.03	0.938±0.03
	wisconsin	0.988±0.01	0.988±0.01	**0.991±0.01**	**0.990±0.01**	**0.989±0.01**	0.985±0.01	0.988±0.01	0.984±0.01	0.970±0.01	0.985±0.01	**0.992±0.01**	0.988±0.01
	breastcancer	0.979±0.01	0.979±0.01	**0.985±0.01**	0.979±0.01	0.978±0.01	**0.982±0.01**	0.981±0.01	**0.983±0.01**	0.968±0.02	0.977±0.01	**0.992±0.01**	0.981±0.01
	pima	0.794±0.04	0.782±0.04	0.777±0.04	0.752±0.03	0.780±0.03	0.783±0.03	0.799±0.02	0.742±0.04	0.773±0.04	0.752±0.04	**0.820±0.04**	0.810±0.04
	segment	0.942±0.01	0.946±0.01	0.943±0.01	0.933±0.01	0.943±0.01	0.949±0.01	0.944±0.01	0.957±0.01	0.951±0.01	0.934±0.01	0.949±0.01	0.959±0.01
	svmguide1	0.974±0.01	0.973±0.01	0.973±0.01	0.959±0.01	0.967±0.00	0.972±0.00	0.976±0.01	0.980±0.01	0.984±0.00	0.959±0.01	0.974±0.00	0.986±0.00
	mushrooms	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	0.996±0.00	1.000±0.00	1.000±0.00	0.999±0.00	1.000±0.00	1.000±0.00
	page-blocks0	0.936±0.01	0.934±0.01	0.931±0.01	0.914±0.01	0.934±0.01	0.930±0.01	0.947±0.01	0.957±0.01	0.960±0.01	0.932±0.01	0.942±0.01	0.972±0.00
	magic	0.839±0.01	0.839±0.01	0.818±0.01	0.796±0.01	0.859±0.01	0.896±0.01	0.886±0.00	0.929±0.00	0.944±0.00	0.837±0.01	0.819±0.01	0.965±0.00
	Average	0.909±0.02	0.909±0.02	0.904±0.02	0.886±0.02	0.910±0.02	0.913±0.02	0.916±0.02	0.922±0.02	0.920±0.02	0.897±0.02	0.910±0.02	0.942±0.02
	Win-Lose	0–11	0–11	2–9	1–10	1–10	1–10	0–11	2–9	0–11	0–11	3–8	N/A

TABLE 4. Mean and variance of 12 oversampling methods under 5 classifiers on each real-world dataset ($nl = 10\%$).

Metrics	Datasets	S.-TkL	S.-IPF	S.-FRST	ADASYN	MWMOTE	AdaptS.	DBSMOTE	means-S.	RSMOTE	S.-G	S.-SW	OOSI
F1-measure	ecoli	0.770±0.05	0.751±0.06	0.771±0.06	0.727±0.07	0.782±0.06	0.874±0.03	0.822±0.05	0.828±0.04	0.856±0.04	0.776±0.05	0.802±0.05	0.912±0.03
	haberman	0.626±0.08	0.604±0.07	0.708±0.05	0.616±0.07	0.626±0.06	**0.784±0.05**	0.672±0.06	0.707±0.07	0.636±0.10	0.624±0.06	0.642±0.07	0.755±0.05
	creditApproval	0.827±0.04	0.807±0.04	0.848±0.03	0.767±0.05	0.828±0.03	**0.884±0.03**	0.836±0.04	0.805±0.03	0.862±0.04	0.799±0.04	0.781±0.04	0.877±0.03
	wisconsin	0.767±0.04	0.765±0.05	0.803±0.03	0.704±0.05	0.814±0.04	0.888±0.03	0.670±0.06	0.884±0.03	0.894±0.03	0.767±0.05	0.722±0.05	0.939±0.03
	breastcancer	0.822±0.04	0.814±0.04	0.787±0.03	0.736±0.06	0.836±0.05	0.918±0.02	0.720±0.05	0.927±0.02	0.899±0.03	0.852±0.04	0.825±0.04	0.935±0.03
	pima	0.750±0.04	0.735±0.05	0.720±0.03	0.711±0.03	0.745±0.04	0.776±0.03	0.747±0.04	0.736±0.04	0.748±0.04	0.707±0.04	0.767±0.05	0.780±0.03
	segment	0.795±0.02	0.789±0.02	0.790±0.02	0.649±0.02	0.809±0.02	0.870±0.01	0.826±0.02	0.752±0.02	0.874±0.02	0.800±0.02	0.747±0.03	0.896±0.01
	svmguide1	0.716±0.02	0.700±0.02	0.696±0.02	0.558±0.02	0.667±0.02	0.731±0.01	0.703±0.02	0.771±0.02	0.905±0.01	0.675±0.02	0.680±0.02	0.927±0.01
	mushrooms	0.844±0.01	0.851±0.01	0.855±0.01	0.835±0.01	0.892±0.01	0.940±0.01	0.656±0.05	0.919±0.01	**0.950±0.01**	0.901±0.01	0.821±0.01	0.945±0.01
	page-blocks0	0.753±0.02	0.741±0.02	0.773±0.01	0.653±0.02	0.760±0.02	0.831±0.01	0.711±0.01	0.852±0.01	0.893±0.01	0.772±0.01	0.721±0.02	0.923±0.01
	magic	0.630±0.01	0.629±0.01	0.778±0.00	0.583±0.01	0.717±0.01	0.802±0.01	0.681±0.01	0.872±0.01	0.871±0.01	0.659±0.01	0.592±0.01	0.914±0.01
	Average	0.755±0.03	0.744±0.03	0.775±0.03	0.685±0.04	0.771±0.03	0.845±0.02	0.731±0.04	0.823±0.03	0.853±0.03	0.757±0.03	0.736±0.04	0.891±0.02
	Win-Lose	0–11	0–11	0–11	0–11	0–11	2–9	0–11	0–11	1–10	0–11	0–11	N/A
AUC	ecoli	0.867±0.04	0.857±0.04	0.876±0.04	0.808±0.05	0.879±0.03	0.919±0.03	0.906±0.04	0.871±0.04	0.922±0.04	0.860±0.05	0.896±0.04	0.952±0.02
	haberman	0.726±0.07	0.712±0.05	0.695±0.09	0.700±0.06	0.703±0.06	0.770±0.06	0.739±0.06	0.794±0.06	0.720±0.09	0.687±0.06	0.720±0.08	0.824±0.05
	creditApproval	0.899±0.03	0.881±0.03	0.885±0.03	0.860±0.04	0.892±0.02	0.922±0.03	0.889±0.04	0.876±0.03	0.919±0.03	0.853±0.04	0.861±0.04	0.925±0.03
	wisconsin	0.857±0.04	0.865±0.03	0.862±0.03	0.816±0.04	0.887±0.03	0.931±0.03	0.751±0.05	0.929±0.03	0.917±0.03	0.853±0.04	0.838±0.04	0.948±0.03
	breastcancer	0.896±0.03	0.895±0.03	0.853±0.03	0.850±0.04	0.912±0.03	0.947±0.02	0.793±0.04	0.945±0.02	0.932±0.02	0.898±0.04	0.892±0.03	0.953±0.03
	pima	0.810±0.04	0.798±0.05	0.794±0.04	0.776±0.03	0.816±0.04	0.820±0.03	0.822±0.04	0.812±0.04	0.814±0.04	0.790±0.04	0.839±0.04	0.859±0.03
	segment	0.863±0.02	0.859±0.02	0.852±0.02	0.739±0.02	0.874±0.02	0.910±0.01	0.887±0.01	0.809±0.02	0.928±0.01	0.863±0.02	0.830±0.02	0.937±0.01
	svmguide1	0.803±0.02	0.789±0.01	0.788±0.01	0.683±0.02	0.773±0.01	0.794±0.01	0.799±0.02	0.834±0.02	0.939±0.01	0.768±0.02	0.782±0.01	0.950±0.01
	mushrooms	0.924±0.01	0.930±0.01	0.923±0.01	0.897±0.01	0.940±0.01	0.953±0.01	0.768±0.03	0.951±0.01	**0.957±0.01**	0.929±0.01	0.914±0.01	0.953±0.01
	page-blocks0	0.829±0.01	0.822±0.01	0.808±0.01	0.749±0.02	0.844±0.01	0.884±0.01	0.817±0.01	0.920±0.01	0.936±0.01	0.845±0.01	0.808±0.01	0.949±0.01
	magic	0.738±0.01	0.737±0.01	0.735±0.01	0.684±0.01	0.809±0.01	0.866±0.01	0.736±0.01	0.925±0.01	0.920±0.01	0.759±0.01	0.704±0.01	0.943±0.01
	Average	0.837±0.03	0.831±0.03	0.825±0.03	0.778±0.03	0.848±0.02	0.883±0.02	0.810±0.03	0.879±0.03	0.9±0.03	0.828±0.03	0.826±0.03	0.927±0.02
	Win-Lose	0–11	0–11	0–11	0–11	0–11	0–11	0–11	0–11	1–10	0–11	0–11	N/A
G-mean	ecoli	0.779±0.05	0.760±0.05	0.779±0.06	0.716±0.07	0.788±0.05	0.837±0.05	0.825±0.04	0.821±0.04	0.859±0.04	0.788±0.05	0.811±0.05	0.911±0.03
	haberman	0.645±0.07	0.634±0.06	0.588±0.09	0.635±0.06	0.641±0.06	0.652±0.08	0.670±0.06	0.720±0.06	0.654±0.09	0.634±0.05	0.651±0.06	0.770±0.04
	creditApproval	0.826±0.04	0.808±0.04	0.809±0.04	0.777±0.05	0.826±0.03	0.864±0.04	0.826±0.04	0.802±0.03	0.858±0.04	0.800±0.03	0.780±0.04	0.877±0.03
	wisconsin	0.790±0.03	0.790±0.04	0.785±0.03	0.740±0.04	0.827±0.04	0.888±0.03	0.681±0.05	0.890±0.03	0.894±0.03	0.788±0.04	0.753±0.04	0.940±0.02
	breastcancer	0.835±0.04	0.827±0.03	0.755±0.04	0.760±0.05	0.845±0.04	0.915±0.03	0.721±0.05	0.928±0.02	0.901±0.03	0.861±0.03	0.837±0.03	0.936±0.03
	pima	0.742±0.05	0.730±0.05	0.718±0.03	0.705±0.04	0.741±0.04	0.750±0.03	0.747±0.04	0.734±0.04	0.743±0.04	0.712±0.04	0.768±0.04	0.780±0.03
	segment	0.799±0.02	0.794±0.02	0.794±0.02	0.667±0.02	0.811±0.02	0.845±0.02	0.827±0.02	0.747±0.02	0.876±0.02	0.807±0.02	0.759±0.02	0.896±0.01
	svmguide1	0.737±0.02	0.725±0.01	0.717±0.01	0.603±0.02	0.692±0.02	0.717±0.01	0.724±0.02	0.775±0.02	0.908±0.01	0.703±0.02	0.707±0.01	0.929±0.01
	mushrooms	0.856±0.01	0.862±0.01	0.861±0.01	0.844±0.01	0.899±0.01	0.941±0.01	0.715±0.04	0.922±0.01	**0.952±0.01**	0.907±0.01	0.836±0.01	0.946±0.01
	page-blocks0	0.773±0.01	0.764±0.01	0.719±0.01	0.689±0.02	0.777±0.01	0.822±0.01	0.738±0.01	0.859±0.01	0.897±0.01	0.792±0.01	0.749±0.01	0.925±0.01
	magic	0.665±0.01	0.664±0.01	0.527±0.01	0.619±0.01	0.732±0.01	0.795±0.01	0.672±0.01	0.877±0.01	0.878±0.01	0.692±0.01	0.633±0.01	0.917±0.00
	Average	0.768±0.03	0.760±0.03	0.732±0.03	0.705±0.03	0.780±0.03	0.821±0.03	0.741±0.03	0.825±0.03	0.856±0.03	0.771±0.03	0.753±0.03	0.893±0.02
	Win-Lose	0–11	0–11	0–11	0–11	0–11	0–11	0–11	0–11	1–10	0–11	0–11	N/A

methods. One explanation for this could be that OOSI fully mines the data space information and adaptively obtains a specific clustering space with the characteristics of the dataset. Additionally, reasonable sample generalization is

TABLE 5. Mean and variance of 12 oversampling methods under 5 classifiers on each real-world dataset (nl = 20%).

Metrics	Datasets	S.-TkL	S.-IPF	S.-FRST	ADASYN	MWMOTE	AdaptS.	DBSMOTE	means-S.	RSMOTE	S.-G	S.-SW	OOSI
F1-measure	ecoli	0.684±0.07	0.666±0.07	0.684±0.07	0.664±0.07	0.692±0.07	0.827±0.03	0.729±0.06	0.765±0.05	0.759±0.06	0.697±0.06	0.711±0.06	0.837±0.06
	haberman	0.583±0.08	0.548±0.08	**0.702±0.04**	0.528±0.10	0.603±0.08	0.666±0.05	0.599±0.09	0.643±0.08	0.562±0.09	0.530±0.08	0.630±0.08	0.691±0.07
	creditApproval	0.693±0.05	0.687±0.06	0.799±0.03	0.686±0.06	0.736±0.04	**0.846±0.03**	0.768±0.05	0.696±0.06	0.765±0.04	0.713±0.06	0.723±0.05	0.801±0.04
	wisconsin	0.620±0.05	0.685±0.06	0.791±0.02	0.630±0.06	0.699±0.06	0.850±0.03	0.612±0.05	0.844±0.04	0.786±0.05	0.615±0.06	0.598±0.07	0.867±0.03
	breastcancer	0.733±0.05	0.696±0.05	0.744±0.04	0.612±0.05	0.754±0.04	**0.868±0.03**	0.596±0.06	0.859±0.04	0.815±0.04	0.762±0.04	0.726±0.05	0.865±0.04
	pima	0.675±0.05	0.666±0.05	0.661±0.06	0.660±0.04	0.670±0.04	**0.772±0.03**	0.699±0.04	0.693±0.04	0.695±0.05	0.675±0.04	**0.727±0.04**	0.720±0.05
	segment	0.660±0.03	0.652±0.03	0.667±0.03	0.583±0.03	0.707±0.03	0.825±0.01	0.687±0.02	0.722±0.02	0.765±0.02	0.697±0.02	0.622±0.03	0.827±0.02
	svmguide1	0.554±0.03	0.545±0.02	0.569±0.02	0.482±0.03	0.568±0.02	0.737±0.01	0.578±0.02	0.772±0.02	0.782±0.02	0.553±0.02	0.558±0.02	0.862±0.01
	mushrooms	0.750±0.02	0.745±0.02	0.796±0.01	0.744±0.01	0.809±0.01	0.876±0.01	0.472±0.05	0.824±0.02	**0.881±0.02**	0.826±0.01	0.712±0.02	0.876±0.01
	page-blocks0	0.639±0.02	0.625±0.02	0.793±0.01	0.584±0.02	0.654±0.02	0.799±0.01	0.622±0.02	0.825±0.01	0.795±0.01	0.672±0.02	0.609±0.02	0.861±0.01
	magic	0.536±0.02	0.530±0.01	0.526±0.01	0.517±0.02	0.642±0.01	0.784±0.01	0.632±0.01	0.837±0.01	0.752±0.01	0.583±0.01	0.509±0.02	0.853±0.01
	Average	0.648±0.04	0.640±0.04	0.703±0.03	0.608±0.05	0.685±0.04	0.805±0.02	0.636±0.04	0.771±0.03	0.760±0.04	**0.666±0.04**	**0.648±0.04**	0.824±0.03
	Win-Lose	0–11	0–11	1–10	0–11	0–11	3–8	0–11	0–11	1–10	0–11	1–10	N/A
AUC	ecoli	0.794±0.05	0.776±0.05	0.782±0.06	0.725±0.06	0.777±0.07	0.816±0.05	0.817±0.05	0.808±0.06	0.844±0.05	0.800±0.05	0.825±0.06	0.882±0.04
	haberman	0.643±0.07	0.619±0.07	0.581±0.10	0.573±0.09	0.657±0.08	0.659±0.06	0.645±0.08	0.713±0.08	0.633±0.07	0.615±0.08	0.656±0.08	0.762±0.07
	creditApproval	0.791±0.05	0.778±0.05	0.811±0.03	0.768±0.05	0.821±0.04	**0.868±0.03**	0.846±0.05	0.784±0.05	0.836±0.04	0.804±0.05	0.803±0.05	0.855±0.04
	wisconsin	0.772±0.04	0.819±0.04	0.778±0.04	0.759±0.05	0.794±0.05	0.885±0.03	0.681±0.04	0.881±0.04	0.831±0.04	0.751±0.05	0.749±0.04	0.895±0.03
	breastcancer	0.813±0.04	0.800±0.03	0.809±0.04	0.726±0.05	0.818±0.04	**0.895±0.03**	0.679±0.05	0.877±0.04	0.861±0.04	0.817±0.04	0.810±0.05	0.886±0.04
	pima	0.745±0.05	0.732±0.04	0.735±0.05	0.707±0.04	0.742±0.04	0.776±0.04	0.757±0.04	0.775±0.04	0.763±0.05	0.746±0.04	0.793±0.04	0.796±0.04
	segment	0.747±0.02	0.742±0.02	0.747±0.03	0.667±0.03	0.782±0.02	0.852±0.01	0.770±0.02	0.806±0.02	0.835±0.02	0.769±0.02	0.720±0.03	0.873±0.02
	svmguide1	0.691±0.02	0.681±0.02	0.677±0.02	0.623±0.02	0.678±0.02	0.720±0.01	0.701±0.02	0.852±0.02	0.846±0.02	0.685±0.02	0.690±0.02	0.893±0.01
	mushrooms	0.853±0.01	0.849±0.01	0.849±0.01	0.828±0.01	0.876±0.01	**0.903±0.01**	0.672±0.03	0.893±0.01	**0.904±0.01**	0.870±0.01	0.830±0.01	0.901±0.01
	page-blocks0	0.742±0.01	0.732±0.01	0.754±0.01	0.699±0.02	0.760±0.01	0.816±0.01	0.734±0.01	0.880±0.01	0.860±0.01	0.773±0.01	0.726±0.02	0.898±0.01
	magic	0.655±0.01	0.649±0.01	0.652±0.01	0.629±0.01	0.736±0.01	0.814±0.01	0.690±0.01	0.884±0.01	0.831±0.01	0.693±0.01	0.630±0.01	0.891±0.01
	Average	0.750±0.04	0.743±0.03	0.743±0.04	0.700±0.04	0.767±0.04	0.819±0.03	0.727±0.04	0.832±0.03	0.822±0.03	**0.757±0.05**	**0.748±0.04**	0.867±0.03
	Win-Lose	0–11	0–11	0–11	0–11	0–11	3–8	0–11	0–11	1–10	0–11	0–11	N/A
G-mean	ecoli	0.694±0.06	0.675±0.06	0.692±0.06	0.649±0.06	0.691±0.07	0.697±0.06	0.727±0.06	0.745±0.06	0.768±0.05	0.715±0.06	0.724±0.05	0.843±0.05
	haberman	0.590±0.07	0.585±0.06	0.261±0.10	0.544±0.09	0.612±0.07	0.604±0.05	0.598±0.08	0.661±0.07	0.582±0.06	0.566±0.07	0.622±0.08	0.711±0.06
	creditApproval	0.707±0.05	0.698±0.05	0.690±0.04	0.690±0.05	0.744±0.04	0.800±0.04	0.768±0.05	0.699±0.06	0.770±0.04	0.731±0.05	0.723±0.05	0.807±0.04
	wisconsin	0.666±0.04	0.724±0.05	0.602±0.04	0.672±0.05	0.728±0.05	0.828±0.03	0.613±0.04	0.851±0.03	0.795±0.04	0.663±0.05	0.648±0.06	0.874±0.03
	breastcancer	0.759±0.04	0.729±0.04	0.722±0.04	0.659±0.04	0.775±0.04	0.853±0.03	0.616±0.05	0.866±0.04	0.827±0.04	0.783±0.03	0.753±0.04	0.871±0.04
	pima	0.676±0.05	0.668±0.04	0.663±0.05	0.654±0.04	0.670±0.04	0.692±0.05	0.702±0.04	0.697±0.04	0.694±0.05	0.679±0.04	**0.734±0.04**	0.727±0.05
	segment	0.679±0.02	0.680±0.02	0.676±0.02	0.607±0.03	0.719±0.02	0.769±0.02	0.701±0.02	0.739±0.02	0.781±0.02	0.715±0.02	0.653±0.02	0.832±0.02
	svmguide1	0.605±0.02	0.598±0.02	0.599±0.02	0.539±0.02	0.607±0.02	0.599±0.02	0.617±0.02	0.790±0.02	0.798±0.01	0.604±0.01	0.608±0.02	0.869±0.01
	mushrooms	0.772±0.02	0.768±0.02	0.782±0.02	0.763±0.01	0.824±0.01	0.880±0.01	0.573±0.04	0.836±0.01	**0.887±0.01**	0.839±0.01	0.742±0.01	0.882±0.01
	page-blocks0	0.681±0.02	0.670±0.01	0.569±0.02	0.633±0.02	0.690±0.02	0.722±0.01	0.659±0.02	0.838±0.01	0.810±0.01	0.709±0.01	0.660±0.02	0.869±0.01
	magic	0.586±0.01	0.582±0.01	0.580±0.01	0.565±0.01	0.667±0.01	0.731±0.01	0.638±0.01	0.847±0.01	0.775±0.01	0.627±0.01	0.562±0.01	0.861±0.01
	Average	0.674±0.04	0.671±0.03	0.621±0.04	0.634±0.04	0.702±0.03	0.743±0.03	0.656±0.04	0.779±0.03	0.772±0.03	**0.694±0.03**	**0.675±0.04**	0.831±0.03
	Win-Lose	0–11	0–11	0–11	0–11	0–11	0–11	0–11	0–11	1–10	0–11	1–10	N/A

performed by merging the intra-cluster sparsity and multi-class density information of the samples. For clustering-based methods, particularly means-S. and RSMOTE, even though they combine clustering with synthetic sample size assignment. However, OOSI utilizes oriented weights containing multi-class information levels of samples to effectively guide the improvement of the sample synthesis path, generating new samples close to information-rich areas and preventing the development of extra noise samples and overlapping samples. These extensive results demonstrate the overall advantage of OOSI in handling imbalanced and noisy data.

Tables 3-5 present, due to limited space, the mean and standard deviation of each real dataset for the ten methods under five classifiers and three metrics. In each row of Tables 3-5, the method that outperform OOSI are highlighted in bold. The rows labeled "Average" contain the results for all datasets as a whole. The rows labeled "Win-Lose" represent the cumulative results of the method outperforming or underperforming OOSI across all datasets.

Overall, it can be seen from Tables 3-5 that for all noise levels, metrics, and datasets, the "Win-Lose" of the comparison methods is almost "0-11". It demonstrates that for all noise levels, metrics, and comparison methods, OOSI outperforms the comparison method on 11 datasets in most cases. Additionally, the bolded portion, which is slightly better than OOSI's comparison method, is almost

TABLE 6. The results of p-value for the friedman test.

Noise levels	F-measure	AUC	G-mean
0%	3.80E-02 *	1.81E-02 *	4.07E-02 *
5%	1.84E-03 *	1.26E-02 *	1.75E-03 *
10%	2.30E-05 *	4.43E-03 *	6.77E-05 *
15%	4.43E-05 *	2.76E-04 *	2.43E-05 *
20%	4.72E-07 *	1.24E-04 *	1.24E-06 *

concentrated in relatively small datasets, which shows that the OOSI performs better with larger datasets. The possible reason is that larger datasets contain more information, and the OOSI fully excavates and utilizes the data information to guide reasonable sampling. Even in the absence of noise, the OOSI method improves F-measure, AUC, and G-mean by 2.6%, 2%, and 3.2%, respectively. When there is noise, that is, the noise level is 10%, the OOSI method improves F-measure, AUC, and G-mean by 11.5%, 8.4%, and 11.7%, respectively. Particularly in comparison to ADASYN, the greatest improvement has been made. When there is no noise, the F-measure, AUC, and G-mean increase by 5.3%, 3.9%, and 5.6%, respectively, whereas they increase by 21.6%, 16.7%, and 19.2% when the noise level is 20%. It emonstrates that the OOSI method outperforms the comparison algorithms of 11 different strategies and achieves better performance improvements on the noisy imbalanced dataset.

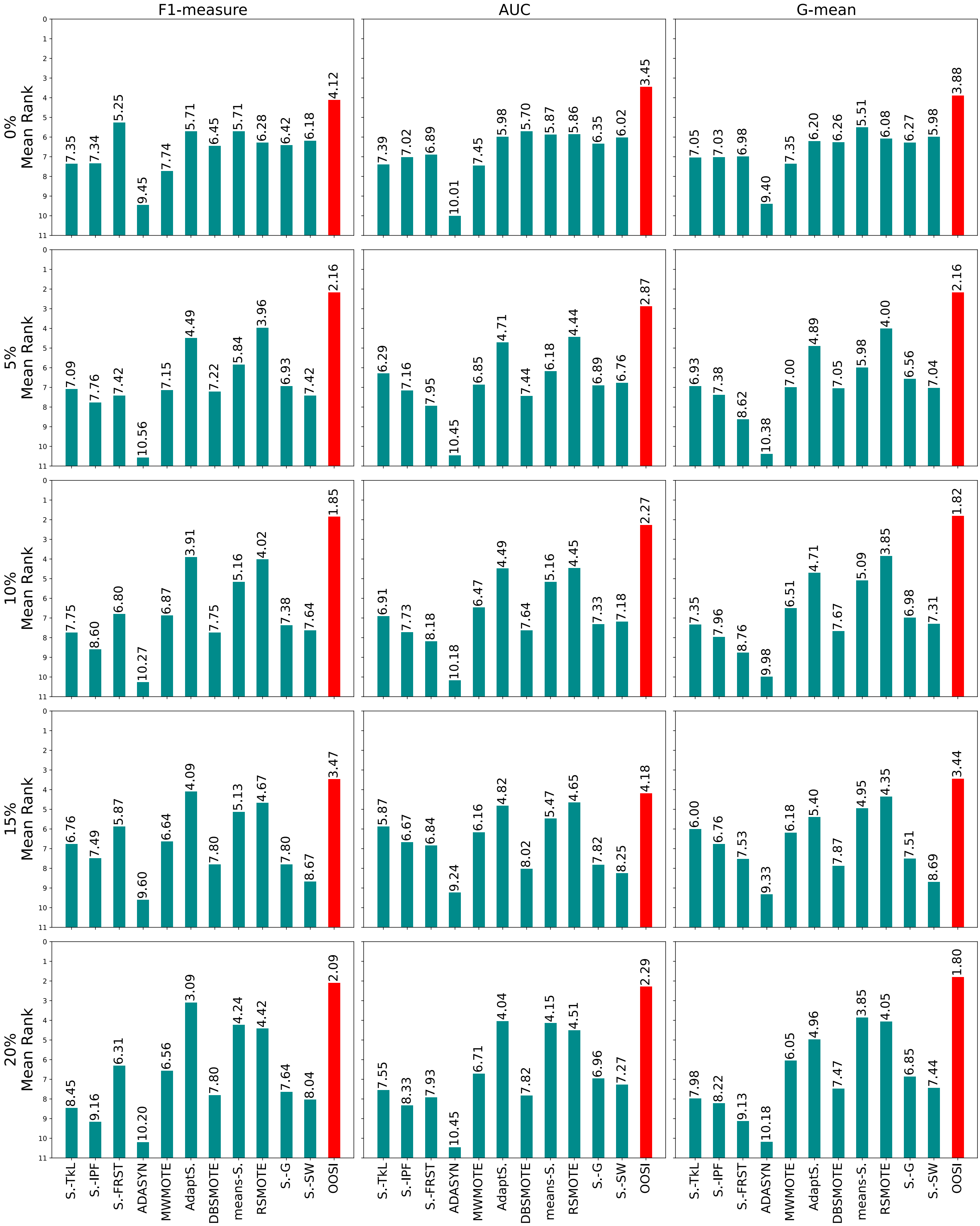

FIGURE 4. **The average rank sums of comparing algorithms on five classifiers and 11 datasets.**

TABLE 7. Average running time of comparative oversampling algorithms (Sec).

Datasets	S.-TkL	S.-IPF	S.-FRST	ADASYN	MWMOTE	AdaptS.	DBSMOTE	means-S.	RSMOTE	S.-G	S.-SW	OOSI
ecoli	0.401	0.605	0.499	0.411	0.492	0.456	0.391	0.444	0.370	0.150	0.769	**0.324**
haberman	0.547	0.524	0.422	0.592	0.253	0.629	0.443	0.485	0.374	0.370	0.588	**0.373**
creditApproval	**0.176**	0.504	0.305	0.514	0.419	0.955	0.257	0.323	0.347	0.241	0.152	0.279
wisconsin	**0.266**	0.391	0.461	0.602	0.413	0.600	0.411	0.403	0.380	0.290	0.758	0.478
breastcancer	**0.265**	0.384	0.441	0.325	0.420	1.276	0.347	0.403	0.429	0.256	0.813	0.419
pima	0.164	0.168	0.384	**0.132**	0.750	3.242	0.286	0.340	0.433	0.201	0.296	0.359
segment	0.671	1.313	**0.252**	0.342	0.782	18.77	0.492	0.336	0.726	0.180	0.578	0.634
svmguide1	0.696	0.667	0.949	0.347	0.979	15.02	1.504	0.663	0.634	0.200	0.692	**0.139**
mushrooms	9.987	**1.561**	6.456	0.602	6.809	483.8	0.209	0.154	1.562	0.353	5.018	4.524
page-blocks0	0.223	1.800	1.309	0.549	2.281	51.57	1.334	0.500	0.595	0.256	1.272	**0.186**
magic	0.902	10.13	23.88	**0.241**	19.14	357.1	0.806	0.299	0.341	0.731	5.333	0.635
Mean rank	5.182	7.455	7.636	5.455	8.636	11.45	5.545	4.818	5.727	2.364	7.909	**4.728**

E. NONPARAMETRIC STATISTICAL ANALYSIS

To verify whether the performance of different algorithms has a statistically significant difference, the non-parametric friedman test is used to analyze the experimental results of 12 methods on five classifiers and 11 datasets. The Friedman test contributes significantly to experimental analysis. It can be used to compare whether or not three or more related samples differ substantially without making any assumptions about the samples' distributions [45]. The fundamental concept of the Friedman test is to arrange each algorithm's experimental data on distinct datasets in ascending order and designate ranks. The ranks with the greatest and worst performance are 1 and 12, respectively, at this time. Then, statistics and p-values are calculated based on the rank sums determined for each algorithm. If the statistic exceeds the critical value or the p-value is less than the significance level (typically 0.05), the null hypothesis is rejected and the performance of distinct algorithms is deemed significantly different.

Figure 4 depicts the average rank sums of the contrasted algorithms on five classifiers and 11 datasets. The red and blue columns, respectively, represent OOSI and comparison algorithms. The crimson bars are consistently larger than the blue bars for all noise levels and metrics. It demonstrates that the OOSI algorithm's average rank sum is superior to the comparison algorithm's. Additionally, Table 6 displays the p-value results of the Friedman test on five classifiers and eleven data sets. The * indicators reject the null hypothesis at a significance level of 0.05. For all noise levels and metrics, the calculated p-values are less than 0.05, indicating that the performance of the various algorithms differs significantly. Likewise, the smaller the p-value, the more significant the difference. It can be observed from the table that as the noise level increases, the p-value continues to decrease, indicating that the performance difference between the algorithms is also intensified.

F. VALIDATING AVERAGE RUNNING TIME

The average running time of ten executions of the compared oversampling algorithms is shown in Table 7. Each dataset's least time-consuming algorithm is highlighted in bold. In the column designated "Mean rank", the friedman test's mean rank sum is analyzed. The method with the lowest rank is the fastest, and the algorithm with the lowest average rank sum is highlighted in red bold. From Table 7, it can be seen that OOSI achieves four of the quickest running efficiencies across 11 datasets, as well as the most victories. Although the OOSI algorithm is not the least time-consuming on every data set, in most cases, the difference between it and the least time-consuming algorithm is tiny, and it is almost always ranked highly in terms of time-consumption relative to other comparable algorithms. Consequently, by integrating all eleven data sets, the OOSI algorithm achieves the lowest average time-consuming ranking. As a whole, the OOSI algorithm's average running time is competitive.

V. CONCLUSION

To cope with both imbalance and noise problems, an adaptive and robust oriented oversampling method with spatial information (OOSI) is proposed. It is an adaptive and rare sampling method that can guide rational sample generalization and sample synthesis path boosting with spatial information. First, the dataset-specific clustering space is adaptively partitioned to mine the data distribution information. After that, OOSI integrates intra-cluster sparsity and multi-class density information to quantify spatial information to guide reasonable sample generalization in different clusters, which not only prevents over-generalization in specific regions but also effectively alleviates intra-class inter-class imbalance. Finally, to avoid noisy samples from introducing deteriorating generalization, sample synthesis paths are guided according to the level of multi-class information among non-noisy seed samples, avoiding the uncontrollability associated with random linear interpolation. The main advantages of OOSI compared to existing methods are that a) It is a rare adaptive and robust oversampling method; b) it can prevent noise hazards with the innovative three-stage noise suppression strategy rather than removing them; c) it can create safe synthetic minority samples with spatial information to avoid overgeneralization and sampling blindness of SMOTE.

Extensive comparative experiments were performed with 11 sampling algorithms with different strategies. Experiments demonstrated that (a) OOSI outperforms comparative sampling algorithms in 5 baseline classifiers on extensive real-world datasets with varying noise levels; (b) OOSI with the lowest average rank is statistically superior to the comparison algorithms; (c) the average running time of OOSI is competitive due to its lowest average time-consuming ranking.

In the future, these results encourage the development of OOSI as a useful tool for improving the sampling mechanism to rebalance the dataset by generating high quality artificial data. In addition, OOSI may not have optimal runtime on certain complex datasets. To overcome this limitation, efficient clustering improvement strategies will be further explored. Furthermore, while oversampling equalizes class imbalances and improves the learning ability of unrepresented classes and reduces overfitting. However, the potential drawbacks that have received little attention i.e., local ambiguity and unnaturalness may introduce ambiguous and deviating samples from the data distribution. Although the proposed OOSI not only prevents the intrusion of noisy or unsuitable samples through a three-stage noise suppression strategy, but also guides rational and data-distribution-compliant sampling through the quantization of spatial information. While this mitigates local ambiguity and unnaturalness to some extent, further attention and exploration of more solutions are critical.

REFERENCES

[1] C.-R. Wang and X.-H. Shao, "An improving majority weighted minority oversampling technique for imbalanced classification problem," *IEEE Access*, vol. 9, pp. 5069–5082, 2021.

[2] H. Zhou, X. Dong, S. Xia, and G. Wang, "Weighted oversampling algorithms for imbalanced problems and application in prediction of streamflow," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107306.

[3] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019.

[4] S. Roy, U. Roy, D. Sinha, and R. K. Pal, "Imbalanced ensemble learning in determining Parkinson's disease using keystroke dynamics," *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119522.

[5] R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, and H. Faris, "Sentiment analysis of Customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022.

[6] A. Shangguan, G. Xie, L. Mu, R. Fei, and X. Hei, "Abnormal samples oversampling for anomaly detection based on uniform scale strategy and closed area," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 25, 2021, doi: 10.1109/TKDE.2021.3130595.

[7] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021.

[8] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.

[9] C. Cao and Z. Wang, "IMCStacking: Cost-sensitive stacking learning with feature inverse mapping for imbalanced problems," *Knowl.-Based Syst.*, vol. 150, pp. 27–37, Jun. 2018.

[10] J. Zheng, X. Wang, D. Wei, B. Chen, and Y. Shao, "A novel imbalanced ensemble learning in software defect predication," *IEEE Access*, vol. 9, pp. 86855–86868, 2021.

[11] Z. Wang and H. Wang, "Global data distribution weighted synthetic oversampling technique for imbalanced learning," *IEEE Access*, vol. 9, pp. 44770–44783, 2021.

[12] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021.

[13] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, p. 20–29, Jun. 2004.

[14] 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, no. 1, pp. 321–357, Nov. 2002.

[15] M. Dudjak and G. Martinović, "An empirical study of data intrinsic characteristics that make learning from imbalanced data difficult," *Expert Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115297.

[16] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[17] Q. Kang, X. Chen, S. Li, and M. Zhou, "A noise-filtered under-sampling scheme for imbalanced classification," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4263–4274, Dec. 2017.

[18] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6651–6672, Jul. 2023.

[19] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, Dec. 2019.

[20] Z. Zhang and J. Li, "Synthetic minority oversampling technique based on adaptive local mean vectors and improved differential evolution," *IEEE Access*, vol. 10, pp. 74045–74058, 2022.

[21] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[22] G. Yang and L. Qicheng, "An over sampling method of unbalanced data based on ant colony clustering," *IEEE Access*, vol. 9, pp. 130990–130996, 2021.

[23] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RS*B*∗: 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, pp. 245–265, Nov. 2012.

[24] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci.*, vol. 291, pp. 184–203, Jan. 2015.

[25] E. Ramentol, I. Gondres, S. Lajes, R. Bello, Y. Caballero, C. Cornelis, and F. Herrera, "Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: The SMOTE-FRST-2T algorithm," *Eng. Appl. Artif. Intell.*, vol. 48, pp. 134–139, Feb. 2016.

[26] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2009, pp. 475–482.

[27] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, Feb. 2014.

[28] S. Xia, Y. Zheng, G. Wang, P. He, H. Li, and Z. Chen, "Random space division sampling for label-noisy classification or imbalanced classification," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10444–10457, Oct. 2022.

[29] C. Liu, S. Jin, D. Wang, Z. Luo, J. Yu, B. Zhou, and C. Yang, "Constrained oversampling: An oversampling approach to reduce noise generation in imbalanced datasets with class overlapping," *IEEE Access*, vol. 10, pp. 91452–91465, 2022.

[30] 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.

[31] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowl.-Based Syst.*, vol. 248, Jul. 2022, Art. no. 108839.

[32] Y. Yan, Y. Jiang, Z. Zheng, C. Yu, Y. Zhang, and Y. Zhang, ‘‘LDAS: Local density-based adaptive sampling for imbalanced data classification,’’ *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116213.

[33] T. Pan, J. Zhao, W. Wu, and J. Yang, ‘‘Learning imbalanced datasets based on SMOTE and Gaussian distribution,’’ *Inf. Sci.*, vol. 512, pp. 1214–1233, Feb. 2020.

[34] B. Chen, S. Xia, Z. Chen, B. Wang, and G. Wang, ‘‘RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise,’’ *Inf. Sci.*, vol. 553, pp. 397–428, Apr. 2021.

[35] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, ‘‘DBSMOTE: Density-based synthetic minority over-sampling technique,’’ *Appl. Intell.*, vol. 36, no. 3, pp. 664–684, 2012.

[36] I. Nekooeimehr and S. K. Lai-Yuen, ‘‘Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets,’’ *Expert Syst. Appl.*, vol. 46, pp. 405–416, May 2016.

[37] G. Douzas, F. Bacao, and F. Last, ‘‘Improving imbalanced learning through a heuristic oversampling method based on k-means and smote,’’ *Inf. Sci.*, vol. 465, pp. 1–20, Jun. 2018.

[38] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, ‘‘NI-MWMOTE: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems,’’ *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113504.

[39] G. Douzas and F. Bacao, ‘‘Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE,’’ *Inf. Sci.*, vol. 501, pp. 118–135, Oct. 2019.

[40] M. Li, H. Zhou, Q. Liu, and G. Wang, ‘‘SW: A weighted space division framework for imbalanced problems with label noise,’’ *Knowl.-Based Syst.*, vol. 251, Sep. 2022, Art. no. 109233.

[41] A. Asuncion and D. Newman, ‘‘UCI machine learning repository,’’ Tech. Rep., 2007.

[42] T. Zhu, X. Liu, and E. Zhu, ‘‘Oversampling with reliably expanding minority class regions for imbalanced data learning,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 6167–6181, Jun. 2023.

[43] B. Mirzaei, B. Nikpour, and H. Nezamabadi-pour, ‘‘CDBH: A clustering and density-based hybrid approach for imbalanced data classification,’’ *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 114035.

[44] G. Kovács, ‘‘An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets,’’ *Appl. Soft Comput.*, vol. 83, Oct. 2019, Art. no. 105662.

[45] D. W. Zimmerman and B. D. Zumbo, ‘‘Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks,’’ *J. Exp. Educ.*, vol. 62, no. 1, pp. 75–86, 1993.

YI DENG received the master’s degree from Southwestern University, China, in 2008. He is currently a Senior Experimentalist with the School of Computer and Information Science, Chongqing Normal University. His current research interests include network engineering and networking technology, big data application, and cloud computing.

MINGYONG LI (Member, IEEE) received the B.S. degree from Central China Normal University, in 2003, and the Ph.D. degree from the School of Computer Science and Technology, Donghua University, in 2021. He is currently a Professor with the School of Computer and Information Science, Chongqing Normal University. His current research interests include cross-modal big data processing, large-scale data retrieval, and deep learning.