Soft Computing (2022) 26:1141–1163
https://doi.org/10.1007/s00500-021-06532-4

DATA ANALYTICS AND MACHINE LEARNING

Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data

Jie Liu[1]

Accepted: 2 November 2021 / Published online: 21 November 2021

Abstract

highly unequal

one or more classes has fewer examples than other => less accurate

Synthetic minority oversampling methods have been proven to be an efficient solution for tackling imbalanced data classification issues. Different strategies have been proposed for generating synthetic minority samples. However, noisy samples which may cause the overlapping of minority and majority classes have not yet been properly treated for reducing their influence on the performance of a classification model. A new method, named Importance-SMOTE, is proposed in this paper. In this method, only borderline and edge samples in minority class are oversampled. The synthetic minority samples are generated proportionally to the importance of the minority samples which is calculated according to the composition and distribution of its nearest neighbors. The positions of the synthetic minority samples are determined by the relative importance of the paired neighbors. The proposed method is expected to obtain a more precise estimation of the true decision surface and reduce the influence of noisy samples. Various public imbalanced datasets and a real case study are considered in the experiments to prove the effectiveness of the proposed method.

Keywords Imbalanced data · Minority oversampling · Noisy samples · Sample importance · Overlapped distribution

1 Introduction

high prob => (-) Majority

low prob => (+) minority

Imbalanced data means that the prior probabilities of different classes are significantly different (López et al. 2013). In an imbalanced dataset, the class with a relatively high prior probability is named majority class or negative class, and the class with a relatively low prior probability is named minority class or positive class. In supervised learning, imbalanced data may deteriorate severely the performance of a standard classification model. The decision surface may be biased toward the majority class, resulting in a low classification accuracy on the minority class (Rivera 2017; Japkowicz 2000). Imbalanced data exists in many practical problems. For example, the collected health monitoring data from a braking system in a high-speed train include as much as 28,837 samples on normal conditions (i.e., majority class) and only 159 samples on faulty conditions (i.e., minority class) (Liu et al. 2017). The defectives of a flight software for one orbiting satellite occupy only 0.41% of the recorded events (Liu et al. 2014). In an imbalanced data, the minority class is usually of more interest for practitioners. For a practical problem, the sources of imbalance can be diverse. For example, the high reliability of the system brings low failure rate. The number and operation time of the system are limited; thus, few failures may occur.

In the last decade, many solutions have been proposed for classification issues related to imbalanced data. These methods can be categorized into (López et al. 2013):

- Data-level methods: In which the original dataset is processed to reduce the imbalance of the prior distribution and, then, the imbalanced dataset is fed into a conventional classification model (He and Garcia 2009; Shilaskar and Ghatol 2019).
- Algorithm-level method: In which a specific classification model is modified to make it less sensitive to class imbalance issues (Hassib et al. 2019; Zhai et al. 2018).
- Cost-sensitive learning: This method gives higher cost to the misclassification of a minority sample and lower cost to that of a majority sample and the overall cost is

✉ Jie Liu
liujie805@buaa.edu.cn

[1] School of Reliability and Systems Engineering, Beihang University, Beijing, China

minimized for building the classification model (Khan et al. 2018; Li and Maguire 2011).

The previous methods can be adopted individually or jointly with the others for tackling class imbalance issues. Re-sampling method is an efficient and effective data-level method as the resampled data can be used in any data-driven classification method. Oversampling minority samples and undersampling majority samples are two popular directions for re-sampling the original dataset. However, undersampling majority samples can lead to the loss of useful information. Thus, oversampling minority samples has been widely adopted. One of the most famous oversampling methods is the synthetic minority oversampling techniques (SMOTE) proposed in MacIejewski and Stefanowski (2011). Various modifications have been proposed for improving the efficiency and effectiveness of the original SMOTE method.

It has been pointed out in many published works that imbalance is not the only challenge, and that, disjuncts, noise and overlapping problems are also the main reasons for the degradation of traditional classification models with imbalanced data (He and Garcia 2009; Branco et al. 2016; Laurikkala 2001). Noise may introduce the overlapping and outliers. In the state-of-the-art SMOTE variants, the noisy samples, especially those in the borderline region, are categorized considering the composition of the nearest neighbors. This categorization is quite simple and trivial, and may not reflect the possible difference between different noisy samples. In this paper, a variant of SMOTE method, named Importance-SMOTE, is proposed for tackling noisy imbalanced data. As pointed out in Krawczyk (2016) and (Fernández et al. 2017), it would be interesting to analyze the structure of the class. Considering the number of samples from the same class among its nearest neighbors, each sample in the dataset is categorized into borderline, noise and safe samples, as shown in Fig. 1 Considering the distribution of the nearest neighbors, the safe samples are further divided into inner and edge samples. Two popular binary classification methods, i.e., k nearest neighbor (KNN) and classification and regression tree (CART), are adopted in this work as the classification model (Tuncer et al. 2020; Tuncer and Dogan 2019). The main contributions include:

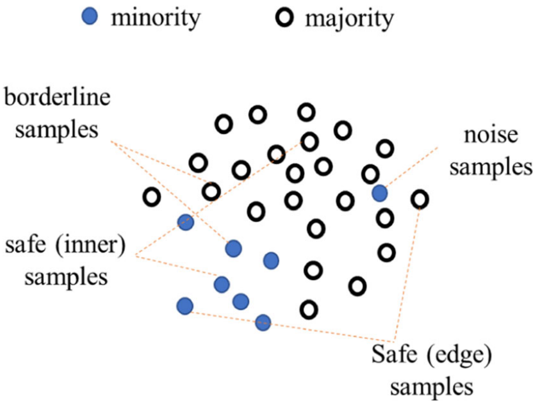


Fig. 1 Illustration of different types of samples in an imbalanced dataset

- Sample importance calculation method based on composition and distribution of its nearest neighbors;
- Improving the efficiency of minority oversampling process by oversampling only the borderline and edge minority samples;
- Generation of synthetic samples considering the relative importance of the paired neighbors from minority class or from both minority and majority classes;
- Comparisons with several SMOTE variants on various public datasets.

The remaining of the paper is structured as follows. Related works on SMOTE variants are reviewed in Section 2. The proposed method is detailed in Section 3 with the borderline and edge samples identification, importance calculation and synthetic samples generation. In Section 4, the proposed method is verified on 26 public datasets and one real dataset on high-speed train, in comparison with eight state-of-the-art synthetic minority oversampling methods. Some conclusions and perspectives are given in Section 5.

2 Related works

SMOTE generates synthetic minority samples along the line between a minority sample and one of its k nearest minority samples. Suppose $\boldsymbol{x}^+$ is a minority sample and $\boldsymbol{x}^+_{knn}$ is chosen from its k nearest neighbors in minority class with equal probability, a new synthetic sample $\boldsymbol{x}^+_{new}$ is generated by

$$\boldsymbol{x}^+_{\text{new}} = \boldsymbol{x}^+ + \alpha * \left(\boldsymbol{x}^+_{\text{knn}} - \boldsymbol{x}^+\right) \tag{1}$$

where α is a random value between 0 and 1. The oversampling can be carried out in the original feature space or the reduced feature space, e.g., reduced feature space of principal component analysis (Fernández et al. 2018).

However, SMOTE may causes over-generalization by oversampling the minority samples without any consideration of the nearest majority samples. And the over-generalization may lead to overlapping between classes (Wang and Japkowicz 2004). The other drawbacks of SMOTE include the creation of too many minority samples which do not facilitate the learning of the minority class, and the introduction of noisy minority samples in the area belonging to the majority class.

Improvements and modifications have been made in the following work to improve the effectiveness and efficiency of SMOTE. The improvements and modifications are generally made from three aspects:

- The first is by selecting the informative minority samples $\boldsymbol{x}^{+}$ and reduce the over-generalization problem.
- The second is by changing the selection process of $\boldsymbol{x}_{knn}^{+}$ to increase the usefulness of the synthetic samples in the data-driven models.
- The third is by modifying the generation rule of the random value α to increase the separability of minority and majority samples.
- The fourth is by integrating oversampling approaches with undersampling ones.
- The fifth is to relabel the majority samples.

Some of the successful and popular modifications of SMOTE are reviewed in this paragraph. Borderline-SMOTE proposed in Han et al. (2005) oversamples only the minority samples near the borderline. The borderline minority samples are identified by the number of majority samples among their k nearest neighbors. One limitation of Borderline-SMOTE is its capability for differentiating borderline and noisy samples. Some noisy samples may be judged as borderline samples and oversampled, reducing the classification accuracy. On the other hand, this strategy may not always be capable of identifying all the borderline samples. ADAptive SYNthetic sampling approach (ADASYN) proposed in He et al. (2008) generates more synthetic samples with the minority samples that are more difficult to classify. The classification difficulty of a minority sample is defined as the number of the majority samples among its k nearest neighbors. A precondition of ADASYN is that the minority samples do not contain any noise or outliers. Otherwise, these noise and outliers in minority class located close to or overlapped with the majority class are overgeneralized to deteriorate the model's performance. Safe-level SMOTE proposed in Bunkhumpornpat et al. (2009) generates synthetic samples along the line between a minority sample and one of its nearest minority neighbors. The synthetic samples are closer to the sample with larger safe-level. The safe-level of one minority sample is defined as the number of minority samples among its k nearest neighbors. In (Barua et al. 2014), majority weighted minority oversampling technique (MWMOT) is proposed for learning from imbalanced datasets. Informative minority samples to be oversampled are the minority samples among the nearest neighbors of the majority borderline samples. The possibility of an informative minority sample to be oversampled is the multiplication of a closeness factor and a density factor. Synthetic samples are generated between an informative minority sample and another minority sample from the same minority cluster. Reference (Nekooeimehr and Lai-Yuen 2016) proposes adaptive semi-unsupervised weighted oversampling (A-SUWO) for improving the performance of SMOTE. The minority samples are clustered with semi-supervised hierarchical clustering approach. The oversampling size of one minority sample is dependent on its Euclidean distance to the majority class. Reference (Piri et al. 2018) proposes a synthetic informative minority oversampling (SIMO) algorithm to improve the performance of SVM models on imbalanced dataset. The informative minority samples are the minority data points misclassified by a SVM model trained on the original imbalanced dataset. Then, the informative minority data points are oversampled to optimize the G-mean value on the training dataset. Similarly, critical data that are more frequently misclassified in validation set are considered more important for classification in Napierała and Stefanowski (2015). Reference (Xu et al. 2017) proposes fuzzy-SMOTE which applies oversampling in the minority samples according to their fuzzy membership degrees. The minority samples with small fuzzy membership degrees are more likely to be oversampled. As the fuzzy membership is calculated on the centroids of different classes, it may not suit the datasets with complex between-class boundary. And the noise and outliers are more likely to be oversampled, causing a high false alarm rate. Reference (Last et al. 2017) proposes to combine K-means method and SMOTE for oversampling the minority samples. K-means method can cluster the minority dataset into a proper number of clusters and, then, each cluster is oversampled with respect to its density. Similar idea is used in Cieslak et al. (2006), and the proposed method is named Cluster-SMOTE. SMOTE-IPF proposed in Sáez et al. (2015) considers the noisy and borderline examples influencing the classification performance on imbalanced data. As the first step, the imbalanced data are oversampled by SMOTE. Then, iterative-partitioning filter (IPF) is used to detect and eliminate the noisy samples. For this objective, an ensemble model is formed by IPF, and the samples that are misclassified by the ensemble with majority voting scheme are judged as noisy samples.

Some of the previous methods try to maximize the classification accuracy on the minority samples. For ADASYN and SIMO, in order to correctly classify all the minority samples, much accuracy on majority samples is sacrificed in the case of overlapping. Thus, it is very important to differentiate the importance of different minority samples as in fuzzy-SMOTE and safe-level SMOTE. In (Napierala and Stefanowski 2012) and (Stefanowski et al. 2014), minority samples are categorized to safe, borderline, rare and noise. Different types of minority samples impose different influence on the classification model in the experiment carried out in Skryjomski and Krawczyk (2017). Thus, it is very important to recognize and model the difference of minority samples' influence on the classification results. However, in these methods, minority samples are differentiated rather simply with respect to the composition of their k-nearest neighbors and may not reflect the precise importance of a minority sample. The noisy samples may make the situation even worse. In most of the previous minority oversampling methods, they assume that the data is noise-free. This assumption restricts the applications of the proposed methods. In this paper, a new synthetic minority oversampling method (named Importance-SMOTE) is proposed.

The work in Noorhalim et al. (2019) shows that sampling method may greatly benefit the performance of imbalanced data classification, by improving class boundary region. In the proposed method, as the first step, the borderline and edge samples from both minority and majority classes are identified. Unlike Borderline-SMOTE, in Importance-SMOTE, both the borderline and edge samples from minority class are oversampled with probabilities proportional to their importance. The importance of a minority/majority sample is calculated with respect to the composition and distribution of its nearest neighbors. The synthetic samples are generated between a minority sample and one of its k nearest neighbors, taking into consideration of their importance. Different from the previous minority oversampling methods, nearest neighbors from both the minority and majority classes are considered for generating new minority samples, i.e., $\boldsymbol{x}_{\mathrm{knn}}^{+}$ in Eq. (1) can be a minority sample or a majority sample. Different methods are considered for deciding the position of the synthetic minority sample considering the class label and importance of the randomly selected nearest neighbor. The generated minority samples are expected to represent more precisely the distribution boundary of the minority class.

The experiment concerns two synthetic imbalanced datasets and various public imbalanced datasets. KNN and CART are considered for training the classification model on the oversampled datasets. Statistical tests including Friedman test and Wilcoxon signed rank test are used for comparing the results of the proposed method and the benchmark methods.

3 Importance-SMOTE

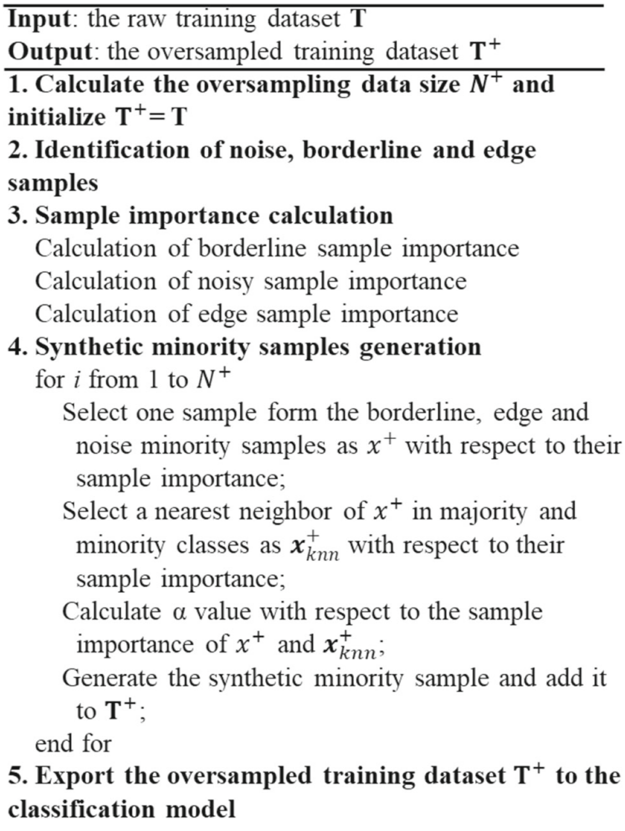
The proposed method is composed of three steps: borderline and edge samples identification, sample importance calculation and synthetic minority samples generation, as shown in Fig. 2. In this section, these three steps are explained in details.

3.1 Borderline and edge samples identification

The borderline samples are selected using k nearest neighbor method, as in Han et al. (2005). The majority and minority samples which have both minority and majority samples among their k nearest neighbors are judged as borderline samples.

Edge samples defining the border of a minority class are also very important to justify the distribution region of the minority class. They can be identified following the method proposed in Li and Maguire (2011). The edge samples are identified among the minority samples which have all the k nearest neighbors from minority class. Suppose a minority sample is $\boldsymbol{x}^{+}$ and its k nearest neighbors from minority class are noted, separately, as $\boldsymbol{x}_{\mathrm{knn},i}, i = 1, 2, \ldots, k$. The first step is to calculate the normal vector $\boldsymbol{v}^{n}$ of the tangent plane, as shown in Fig. 3.

The normal vector $\boldsymbol{v}^{n}$ is calculated as the sum of the unit vectors from $\boldsymbol{x}^{+}$ to its nearest neighbors. The equation is as follows:



```
Input: the raw training dataset T
Output: the oversampled training dataset $\mathbf{T}^+$
1. Calculate the oversampling data size $N^+$ and
initialize $\mathbf{T}^+ = \mathbf{T}$
2. Identification of noise, borderline and edge
samples
3. Sample importance calculation
  Calculation of borderline sample importance
  Calculation of noisy sample importance
  Calculation of edge sample importance
4. Synthetic minority samples generation
  for $i$ from 1 to $N^+$
    Select one sample form the borderline, edge and
      noise minority samples as $x^+$ with respect to their
      sample importance;
    Select a nearest neighbor of $x^+$ in majority and
      minority classes as $\boldsymbol{x}_{knn}^+$ with respect to their
      sample importance;
    Calculate α value with respect to the sample
      importance of $x^+$ and $\boldsymbol{x}_{knn}^+$;
    Generate the synthetic minority sample and add it
      to $\mathbf{T}^+$;
  end for
5. Export the oversampled training dataset $\mathbf{T}^+$ to the
classification model
```

Fig. 2 Pseudocode of Importance-SMOTE

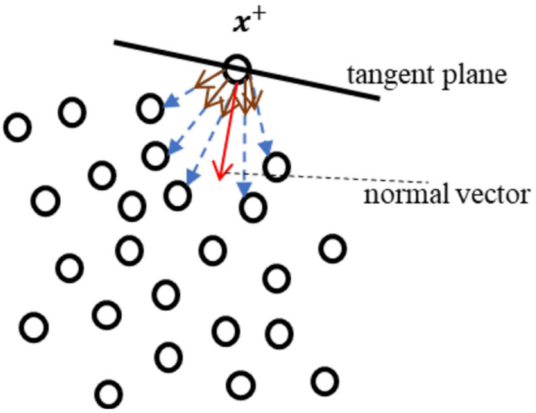


Fig. 3 Illustration of normal vector $\boldsymbol{v}^n$ of the tangent plane with short solid arrow being the unit vectors

$$\boldsymbol{v}^n = \sum_{i=1}^{k} \boldsymbol{v}_i^u \tag{2}$$

with $\boldsymbol{v}_i^u = \frac{\boldsymbol{x}_{\mathrm{knn},i} - \boldsymbol{x}^+}{\|\boldsymbol{x}_{\mathrm{knn},i} - \boldsymbol{x}^+\|}$ and $\|\blacksquare\|$ being the norm of a vector. The second step is to calculate the number of nearest neighbors located on the high-density side of the tangent plane. If the minority sample $\boldsymbol{x}^+$ is an edge sample, most of its nearest neighbors should be on the high-density side of the tangent plane, as shown in Fig. 3. In the figure, all its nearest neighbors are on the same side of the tangent plane, and, thus, $\boldsymbol{x}^+$ in the figure is an edge sample. By counting the number of nearest neighbors on the high-density side, one can identify the edge samples. If the dot product (noted as δ_i as in (3)) between $\boldsymbol{v}_i^u$ and the normal vector $\boldsymbol{v}^n$ is positive, the corresponding nearest neighbor is on the high-density side of the tangent plane.

$$\delta_i = \boldsymbol{v}_i^u . \boldsymbol{v}^n \tag{3}$$

Equation (4) gives the percentage of the k nearest neighbors with a positive dot product in (3). A threshold ρ (smaller than but close to 1) can be given as criterion for judging the edge samples. If $p \geq \rho$, the corresponding minority sample $\boldsymbol{x}^+$ is identified as an edge sample.

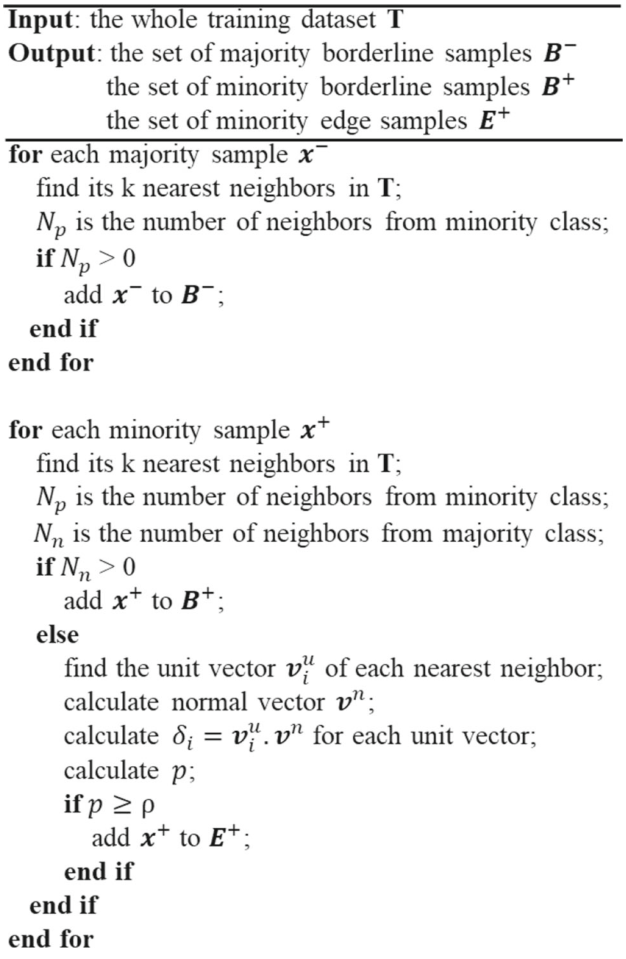
$$p = \frac{1}{k} \sum_{i=1}^{k} (\delta_i \geq 0) \tag{4}$$

The pseudo-code for identifying borderline and edge samples are given in Fig. 4.

3.2 Sample importance calculation

This part introduces the importance calculation for borderline samples, edge samples and noisy samples from minority and majority classes.

The importance calculation of borderline samples has been reported in a previous work (Liu and Zio 2018). In (Liu and Zio 2018), the objective is to properly weighting the samples that may become support vectors in a support vector machine model, and noisy samples are eliminated before training. In this work, edge samples and noisy samples are need to be properly treated. Edge samples are important for defining the distribution region of the minority class. To avoid the loss of information in minority class, the importance of noisy minority samples is not crudely assumed to be zero. However, the noisy majority samples are eliminated directly from the training dataset.



```
Input: the whole training dataset T
Output: the set of majority borderline samples $\boldsymbol{B}^-$
        the set of minority borderline samples $\boldsymbol{B}^+$
        the set of minority edge samples $\boldsymbol{E}^+$
for each majority sample $\boldsymbol{x}^-$
  find its k nearest neighbors in T;
  $N_p$ is the number of neighbors from minority class;
  if $N_p > 0$
    add $\boldsymbol{x}^-$ to $\boldsymbol{B}^-$;
  end if
end for

for each minority sample $\boldsymbol{x}^+$
  find its k nearest neighbors in T;
  $N_p$ is the number of neighbors from minority class;
  $N_n$ is the number of neighbors from majority class;
  if $N_n > 0$
    add $\boldsymbol{x}^+$ to $\boldsymbol{B}^+$;
  else
    find the unit vector $\boldsymbol{v}_i^u$ of each nearest neighbor;
    calculate normal vector $\boldsymbol{v}^n$;
    calculate $\delta_i = \boldsymbol{v}_i^u . \boldsymbol{v}^n$ for each unit vector;
    calculate $p$;
    if $p \geq \rho$
      add $\boldsymbol{x}^+$ to $\boldsymbol{E}^+$;
    end if
  end if
end for
```

Fig. 4 Pseudo-code for borderline and edge samples identification

3.2.1 Importance calculation for borderline samples

The importance of the selected borderline samples is calculated according to the composition and distribution of their nearest neighbors.

The importance function $g(\boldsymbol{x})$ in (5) which is a monotone decaying function of a distance measure d calculates the importance of a borderline sample:

$$g(\boldsymbol{x}) = \frac{2}{1 + \exp(\beta d)} \tag{5}$$

with β being the steepness parameter of the decay and d being a distance measure including two elements, i.e., d^1 and d^2:

$$d = h\left(d^1, d^2\right) \tag{6}$$

with $h(\cdot,\cdot)$ describing the relation between d^1, d^2 and d, d^1 and d^2 characterizing the composition and distribution of the nearest neighbors of a borderline sample. The value of d represents the closeness of a sample to other samples of the same class. The importance function converts the distance to a smooth fuzzy value. From (5), one can observe that a smaller value of d derives higher importance of the corresponding borderline sample. Different from Borderline-SMOTE which considers only the composition of the nearest neighbors to categorize minority samples, the distribution of the nearest neighbors is also considered in Importance-SMOTE. An example is given in Fig. 5. The six nearest neighbors of two minority samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ include both three minority samples. Borderline-SMOTE judges these two samples in the same category. However, the distributions of their nearest neighbors show that it is more important to classify correctly $\boldsymbol{x}_1$ than $\boldsymbol{x}_2$, as the nearest neighbors of $\boldsymbol{x}_1$ from the majority and minority classes are clearly separated around $\boldsymbol{x}_1$.

The first element d^1 in (7) reflects the closeness of a borderline sample to the other samples from the same class. It is trivial to assign higher importance to the borderline sample with more samples from the same class among its nearest neighbors. Suppose the numbers of minority and majority samples among the k nearest neighbors of a borderline sample $\boldsymbol{x}$ are noted, separately, as N_p and N_n with $N_n = k - N_p$, the value of d^1 is calculated as follows:

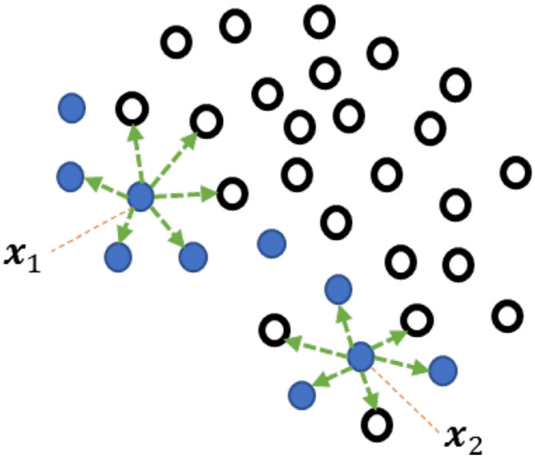


Fig. 5 Illustration of different local distributions of nearest neighbors (empty and solid circles represent data points from two classes, respectively)

$$d^1 = \begin{cases} \dfrac{eN_p}{N_n}, & \text{for a majority borderline sample} \\ \dfrac{N_n}{N_p}, & \text{for a minority borderline sample} \end{cases} \tag{7}$$

where e is a balancing factor for imbalanced data equal or larger than 1. For example, consider a data point $\boldsymbol{x}^+$ from the minority class and a data point $\boldsymbol{x}^-$ from the majority class have both only half of the k nearest neighbor from the minority class. The distance measure in (7) shows that the importance of $\boldsymbol{x}^+$ is higher than that of $\boldsymbol{x}^-$. This is for balancing the different prior probabilities in imbalanced datasets. And the correct classification of $\boldsymbol{x}^+$ is more important than $\boldsymbol{x}^-$.

The second part d^2 of (6) characterizes the distribution of the nearest neighbors of a borderline sample, including their separability and alignment, because two data points from the same class with the same composition of their near neighbors may not always have the same importance. In order to characterize quantitatively this difference, normal vectors introduced in the previous section is adopted. Suppose the k nearest neighbors of a borderline sample $\boldsymbol{x}$ are noted as $\boldsymbol{x}_j, j = 1, 2, \ldots, k$, and the first N_p neighbors are from the minority class and the rest are from the majority class: the normal vectors of the neighbors from the two classes are calculated as

$$\begin{aligned} \boldsymbol{v}^+ &= \frac{1}{N_p} \sum_{j=1}^{N_p} \frac{\boldsymbol{x}_j - \boldsymbol{x}}{\|\boldsymbol{x}_j - \boldsymbol{x}\|} \\ \boldsymbol{v}^- &= \frac{1}{k - N_p} \sum_{j=N_p+1}^{k} \frac{\boldsymbol{x}_j - \boldsymbol{x}}{\|\boldsymbol{x}_j - \boldsymbol{x}\|} \end{aligned} \tag{8}$$

Then, the value d^2 is calculated as

$$d^2 = \frac{\left\langle \frac{\boldsymbol{v}^+}{\|\boldsymbol{v}^+\|}, \frac{\boldsymbol{v}^-}{\|\boldsymbol{v}^-\|} \right\rangle + 1}{\|\boldsymbol{v}^+\| * \|\boldsymbol{v}^-\| + \varepsilon}, \tag{9}$$

with $\langle \blacksquare, \blacksquare \rangle$ being the inner product, ε being a small positive value. Thus, d^2 is always a positive value. The inner product of two normal vectors is the cosine of the angle between the two normal vectors. With the same norm of the two normal vectors, i.e., $\|\boldsymbol{v}^+\| = \|\boldsymbol{v}^-\|$, the nominator shows that higher separability of the samples from two classes gives a smaller value of d^2 and, thus, a larger importance value with (5). Similarly, with the same angle between these two normal vectors, the denominator shows that higher density the nearest neighbors from the same class gives a larger norm of the normal vector and a smaller value of d^2.

Finally, the distance d_i in (5) is expressed as

$$d_i = \gamma * d_n^1 + (1-\gamma) * d_n^2 \tag{10}$$

with d_n^1, d_n^2 being the normalized value of d^1, d^2 given by Eqs. (7) and (9) and γ being a positive value between 0 and 1, weighting the importance of the two elements.

3.2.2 Importance calculation for edge samples

Edge samples from majority classes are not considered in the work, since the main objective is to generating synthetically minority samples. Unlike some previous work where only borderline samples are oversampled, edge minority samples are also considered in this work, since they characterize the distribution region of minority class.

Since edge minority samples have no majority samples in its neighborhood, its importance is calculated with (5) but the distance d is determined by $\rho/(1+p)$ with p of (4). As $p \geq \rho$ and $\rho < 1$, the distance d is positive and smaller than 0.5.

3.2.3 Importance calculation for noisy samples

Noisy samples are the ones which are not among the nearest neighbors of any other sample of the same class. For example, a minority noisy sample does not fall into the nearest neighbors of any minority samples.

For a noisy minority sample, its importance should be very small and its distance d is fixed as a constant value (e.g. 2), while a noisy majority sample is eliminated directly from the training dataset. The different strategies for assigning distance values for noisy samples from different classes originate from the following considerations about imbalanced dataset: (1) noisy majority samples are highly possible to be true noise as discussed in Napierala and Stefanowski (2016); (2) noisy minority samples are not eliminated to keep the maximal information, but its importance is kept low to reduce its influence in case that it is a true noisy sample. The synthetic minority generation process introduced in the next section guarantees also the control of its influence.

The unknown parameters for importance calculation include k in KNN, the steepness parameter β in (5), the balancing factor e in (7) and the parameter γ in (10).

3.3 Synthetic minority samples generation

Synthetic minority samples are generated with Eq. (1). The important steps are to select proper $\boldsymbol{x}^+$ and $\boldsymbol{x}_{knn}^+$ and the calculation of α. In this work, $\boldsymbol{x}^+$ is selected from the minority borderline and edge samples in $\boldsymbol{B}^+$ and $\boldsymbol{E}^+$. However, the probability of each sample to be selected is not uniform. It is proportional to its importance given by Eq. (5). The benefit of oversampling minority edge samples is that the distribution range of minority class is tight and clear and, thus, the data-driven methods can capture easily the classification hyperplane.

As shown in Fig. 6, Fig. 6a shows the original imbalanced data, in which the solid blue dots represent minority samples and the empty dots are majority samples. In Figs. 6b an c, the red solid dots are the newly generated minority samples, and the same number of minority samples are generated in these two figures. In Fig. 6b, by oversampling all the minority samples, the newly generated samples are evenly distributed in the region of the minority class. In Fig. 6c, by oversampling only the borderline and edge samples in the minority class, the synthetic minority samples are located on the borderline and edge of the minority class. The minority samples generated with the borderline and edge samples are more useful in defining the region of the minority class, which may improve the classification accuracy.

In SMOTE and some other methods, $\boldsymbol{x}_{knn}^+$ is selected from the nearest minority neighbors of $\boldsymbol{x}^+$. In importance-SMOTE, $\boldsymbol{x}_{knn}^+$ is selected form the nearest neighbors in both minority and majority samples with equal probability. The minority samples are usually sparse, the borderline between minority and majority class can be more precise by taking into consideration the nearest neighbors in majority class for generating synthetic minority samples. This idea is illustrated in Fig. 7 where the red solid dots are the synthetic minority samples.

In Fig. 7, five nearest neighbors of the minority sample $\boldsymbol{x}$ are selected to generate six synthetic minority samples. Figure 7a considers only the minority samples as $\boldsymbol{x}_{knn}^+$ in Eq. (1), and Fig. 7b considers both minority and majority samples. One can observe that synthetic minority samples in Fig. 7b can characterize more clearly the borderline between minority and majority classes around $\boldsymbol{x}$, while in Fig. 7a, the newly generated minority samples may contradict the true borderline. Note that in Fig. 7b, the position of the newly generated minority sample is very important.

Thus, in this work, $\boldsymbol{x}_{knn}^+$ in Eq. (1) can be a majority or minority sample from the nearest neighbors of a borderline or edge sample of minority class, and a new strategy is proposed for the calculation of the position of synthetic minority samples. The positions of the synthetic minority samples are decided by α. If $\boldsymbol{x}_{knn}^+$ is a majority sample, the value of α is calculated as Eq. (10):

$$\alpha = \mu * rand\left(0, \min\left(1, g\left(\boldsymbol{x}^+\right)/g\left(\boldsymbol{x}_{\mathrm{knn}}^+\right)\right)\right) \tag{11}$$

with μ being a positive value smaller than 1. Otherwise, it is calculated as Eq. (11):

$$\alpha = rand\left(0, \min\left(1, g\left(\boldsymbol{x}_{\mathrm{knn}}^+\right)/g\left(\boldsymbol{x}^+\right)\right)\right) \tag{12}$$

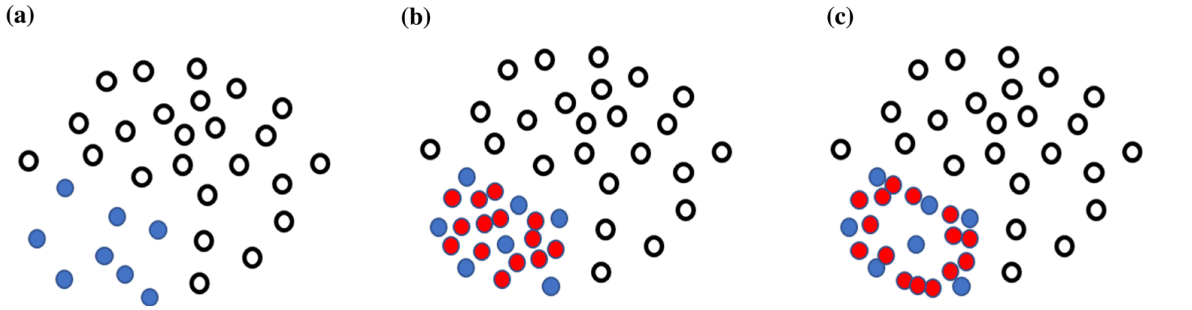
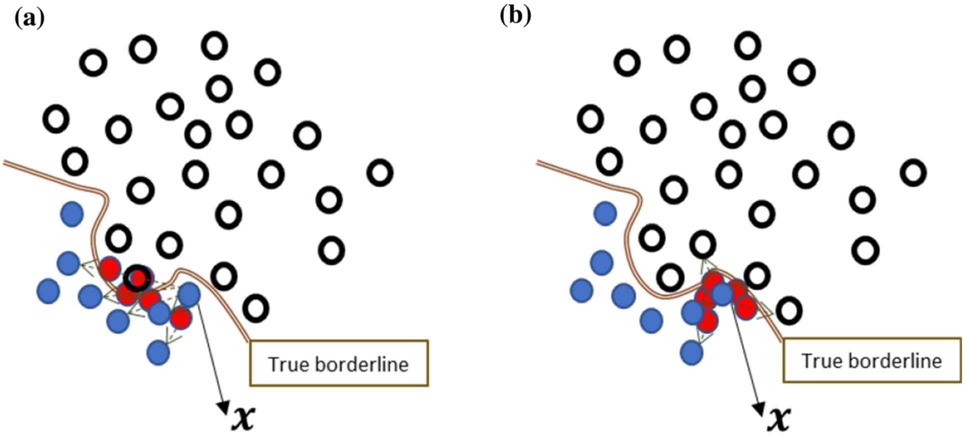


Fig. 6 Illustration of differences between oversampling all minority samples **b** and only borderline and edge samples in minority class **c**

Fig. 7 Illustration of minority samples generation with respect to **a** minority samples and **b** both minority and majority samples



Note that when $\boldsymbol{x}_{\text{knn}}^{+}$ is a majority sample and its importance is larger than that of $\boldsymbol{x}^{+}$, α would be much smaller than 1 and the synthetic minority sample would be closer to the minority sample $\boldsymbol{x}^{+}$. Even if the importance of $\boldsymbol{x}_{knn}^{+}$ is relatively smaller, the synthetic minority sample is still kept away from the majority sample with a proper distance, as α is always smaller than μ. This is to keep the separability of the majority and minority samples. When $\boldsymbol{x}_{knn}^{+}$ is a minority sample, the generated synthetic minority sample is closer to the sample with larger importance between the paired samples $\boldsymbol{x}_{\text{knn}}^{+}$ and $\boldsymbol{x}^{+}$.

Since noisy minority samples have small importance, they are less probable to be oversampled. Even if they are selected to be oversampled, the position of the synthetic minority sample is close to its nearest minority sample and far from the nearest majority sample.

4 Experiments

4.1 Experiment setup

In this section, two synthetic datasets and 26 Keel public datasets (Bach et al. 2017) with different Imbalance Ratios (IR) are adopted to test the effectiveness of the proposed approach, in comparison with the benchmark methods. Most datasets are from real-world problems. Table 1 summarizes the characteristics of Keel public datasets.

Except the classical IR, the adjusted IR is also listed in Table 1. The adjusted IR is calculated with Eq. (12), where N_{df} is the number of discriminative features determined by the Pearson correlation test, and λ is the parameter that controls the importance of the penalty term (Zhu et al. 2020). By considering the capability of the features in discriminating different classes, the adjusted IR is believed to better reflect the imbalanced data classification difficulty than conventional IR.

$$\text{Adjusted IR} = \text{IR} - \lambda \log(N_{\text{df}}) \tag{13}$$

Other characteristics listed in Table 1 include the number of borderline samples, the number of edge samples and the number of noise samples in majority and minority in majority and minority classes, respectively. They are calculated with a k value of 5 for the training datasets. These characteristics, especially those on minority class, can in a certain level reflect the distributions of the datasets. For example, in datasets with small numbers of borderline samples and noisy samples in minority class (such as *ecoli-0-1_vs_5*), the majority and minority classes are more separable; in datasets with a small number of edge samples and a large number of borderline samples (such as

Table 1 Characteristics of the keel datasets used in this paper

Dataset	#attributes	#instances	IR	Adjusted IR	#borderline samples in majority class	#borderline samples in minority class	#edge samples in majority class	#edge samples in minority class	#noise samples in majority class	#noise samples in minority class
03subcl5-600–5-30-BI-fivefold	2	600	5	5.00	118	64	199	5	1	10
ecoli-0–1-4-7_vs_2-3–5-6	7	336	10.59	9.56	21	16	215	4	1	3
ecoli-0–1-4-7_vs_5-6	6	332	12.28	11.25	15	14	221	4	1	2
ecoli-0-1_vs_5	6	240	11	9.84	6	3	163	10	0	2
ecoli-0–2-3-4_vs_5	7	202	9.1	7.94	8	4	132	10	0	2
ecoli-0–2-6-7_vs_3-5	7	224	9.18	8.22	17	11	137	4	0	3
ecoli-0–3-4-6_vs_5	7	205	9.25	8.09	10	3	132	11	0	2
ecoli-0–3-4_vs_5	7	220	9	7.84	8	3	132	11	0	2
ecoli-0–6-7_vs_3-5	7	222	9.09	7.93	16	11	138	3	0	4
ecoli067-5	6	220	10	8.90	16	10	137	4	0	3
ecoli2	7	336	5.46	4.50	31	16	186	22	2	3
glass-0–1-5_vs_2	9	172	9.12	9.12	43	9	81	0	0	5
glass-0–1-6_vs_2	9	192	10.29	10.29	36	8	104	0	0	6
glass4	9	214	15.47	14.31	16	9	144	0	0	1
haberman	3	306	2.78	2.78	110	51	62	0	1	14
led7digit-0–2-4–5-6–7-8-9_vs_1	7	443	10.97	9.58	12	12	313	0	0	18
paw02a-600–5-0-BI-fivefold	2	600	5	5.00	41	33	229	30	1	1
paw02a-600–5-30-BI-fivefold	2	600	5	5.00	82	42	214	17	1	6
pima	8	768	1.87	0.99	228	177	167	17	4	20
poker8-6	10	1477	85.88	85.88	14	10	1151	0	0	4
shuttle-c2-vs-c4	9	129	20.5	19.62	0	4	96	0	0	1
vehicle1	18	846	3.25	3.47	247	148	253	15	3	10
winequality-red-4	11	1599	29.17	29.17	133	11	1102	0	0	31
yeast-0–2-5–7-9_vs_3-6–8	8	1004	9.14	8.11	79	37	628	33	2	8

Table 1 (continued)

Dataset	#attributes	#instances	IR	Adjusted IR	#borderline samples in majority class	#borderline samples in minority class	#edge samples in majority class	#edge samples in minority class	#noise samples in majority class	#noise samples in minority class
yeast-0–3-5-9_vs_7-8	8	506	9.12	9.12	97	23	265	2	1	15
yeast-0–5-6–7-9_vs_4	8	528	9.35	9.35	74	31	303	1	0	9

haberman), the minority class is surrounded by majority class or is heavily overlapped with majority class.

The experimental process is composed of four steps as shown in Fig. 8. These steps are explained in detail in the following paragraphs.

1. Since the Keel public datasets include the data partition for fivefold cross-validation, the first step in Fig. 6 is to select onefold as test dataset and the rest as training dataset. The whole process is repeated five times to iteratively test each fold.
2. The second step is to oversample separately the minority samples with the proposed method, i.e., Importance-SMOTE, and the benchmark methods, including ROSE (random over sampling examples) (Menardi and Torelli 2014), SMOTE, ADASYN, A-SUWO, Borderline-SMOTE, Cluster-SMOTE, SOMTE-IPF and safe-level SMOTE. The parameters in these methods are tuned with grid search method. For parameters tuning, 20% of minority and majority samples in the training dataset are randomly selected, for testing the performance of each possible parameters' combination. The test is repeated for ten times to reduce the randomness, and the best parameters values are the combination with the best average performance characterized by the sum of F-measure and G-mean.

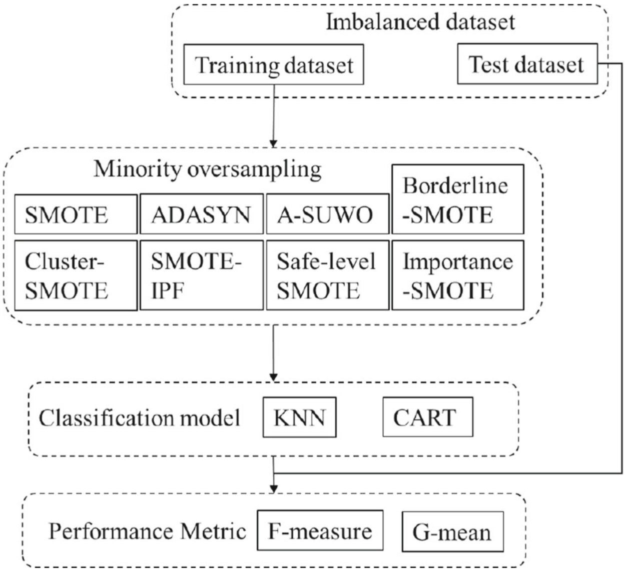


Fig. 8 Illustration of the experimental process

In precious works, the optimal parameters values are searched among few candidates (Nekooeimehr and Lai-Yuen 2016; Kovács 2019). They may even be pre-fixed in the experiments (He et al. 2008; Barua et al. 2014). For rigorousness and considering the searching space of previous work, in the experiments of this work, the best number of nearest neighbors for generating synthetic minority samples is selected from Japkowicz (2000); Liu et al. 2017; Liu et al. 2014; He and Garcia 2009; Shilaskar and Ghatol 2019; Hassib et al. 2019; Zhai et al. 2018; Khan et al. 2018; MacIejewski and Stefanowski 2011; Krawczyk 2016). Fifteen base classification and regression tree (CART) classifiers are used to evaluate the synthetic minority samples in SMOTE-IPF. The number of clusters in Cluster-SMOTE is chosen from Rivera (2017); Japkowicz 2000; Liu et al. 2017; Liu et al. 2014; He and Garcia 2009; Shilaskar and Ghatol 2019; Hassib et al. 2019; Zhai et al. 2018). The number of nearest neighbors for identifying noise and for determining the weight of each minority sample, and the number of folds for clustering minority samples in A-SUWO are from vectors (Japkowicz 2000; Liu et al. 2017, 2014; Shilaskar and Ghatol 2019; Zhai et al. 2018; Li and Maguire 2011; Branco et al. 2016; Krawczyk 2016) and (Rivera 2017; Japkowicz 2000; Liu et al. 2017, 2014; He and Garcia 2009; Shilaskar and Ghatol 2019), respectively. The value for density estimation in ADASYN is from vector (Japkowicz 2000; Liu et al. 2017, 2014; He and Garcia 2009; Shilaskar and Ghatol 2019; Hassib et al. 2019; Zhai et al. 2018; Khan et al. 2018; MacIejewski and Stefanowski 2011; Krawczyk 2016). For Importance-SMOTE, the steepness parameter β in (5), the balancing factor e in (7) and the parameter γ in (10) are tuned separately from vectors [0.1, 0.3, 0.5, 0.8, 1.2], (López et al. 2013; Rivera

2017; Japkowicz 2000; Liu et al. 2017) and [0.1, 0.2, 0.4, 0.6, 0.8].

1) The oversampled training dataset is fed to the classification models in the third step. The classification models considered in this paper include k-nearest neighbors' method (KNN) and standard CART provided by MATLAB with the functions *fitcknn* and *fitctree*. KNN and CART are two mature and popular classification approaches which have been used as benchmark methods in numerous works.
2) The fourth step is to calculate the performance of each classification model on test datasets with respect to F-measure and area under the precision-recall curve (noted as AUC(PRC)). F-measure is a popular and effective performance metric for characterizing classification performance on imbalanced datasets, reflecting the capability in balancing precision and recall (Branco et al. 2016). The AUC(PRC) is more informative than the AUC under the true positive rate–false positive rate curve in binary classification, as pointed out in Saito and Rehmsmeier (2015). AUC reflects the robustness of the method with respect to the classification decision boundary. Two performance metrics are considered in this paper, since single performance metric is not sufficient when handling imbalanced classification problem (He and Garcia 2009).

4.2 Results on synthetic imbalanced datasets

Two synthetic imbalanced are generated for testing the effectiveness of Importance-SMOTE. Four hundreds samples are generated in the space of $[0, 1] \times [0, 1]$. With predefined borderlines between the two classes, the samples inside the borderlines are labeled $+1$, i.e., minority samples, and the others labeled -1. The synthetic *imbalanced dataset 1* is a conventional dataset with minority samples located at the center of the figure, as shown in Fig. 9. And for the synthetic *imbalanced dataset 2*, the minority samples are in four disjoint regions, as shown in Fig. 10. Since these are synthetic data, the borderlines between the majority and minority samples are known and marked black in these Figures. With a noise level of 10%, a total of 40 samples from majority and minority classes are injected with noise.

The experiment in this part repeats only the second step of Fig. 8, i.e., minority oversampling. The aim is to show the effectiveness of different oversampling method for noisy imbalanced data.

Table 2 reports the information (including numbers and percentages (in brackets) of synthetic minority samples inside and outside the true borderlines) of the synthetic minority samples generated by different oversampling methods. Note that ROSE generates both synthetic minority and majority samples from the original training datasets, but synthetic majority samples are not counted in the table. From the table, one can observe that in general, safe-level SMOTE, SMOTE-IPF and Importance-SMOTE proposed in this work outperform the other benchmark methods with a low percentage of synthetic minority samples outside the borderlines.

Safe-level SMOTE may classify the minority samples into different safe levels with respect to the number of majority samples among their nearest neighbors. Noisy samples are given a low safe level and, thus, few synthetic minority samples are generated around the noisy ones.

SMOTE-IPF may eliminate the noisy samples in the synthetic dataset. Thus, the influence of noisy samples in the original datasets can be partly neutralized.

ROSE generates synthetic samples with respect to a probability distribution centered on the selected sample and dependent on a matrix of scale parameters. In case of noise, the probability distributions of a pair of a minority sample and a majority sample which are close enough may overlap. Thus, the generated minority and majority samples are also overlapped. As shown in Figs. 9 and 10, the synthetic minority and majority samples are overlapped and may exceed the definition region of the original training dataset, i.e., $[0\ 1] \times [0\ 1]$. With the borderlines composed of straight-line segments, the probability distribution which is normally isotropic cannot properly model the local property. Thus, ROSE does not work properly on these synthetic imbalanced datasets, as shown in Figs. 9 and 10.

SMOTE may generate synthetic minority samples with respect to noisy samples outside the borderlines and, thus, producing useless synthetic samples. These samples may even reduce the performance of the classification model trained on the oversampled training dataset.

Borderline-SMOTE has similar problem with SMOTE for noisy imbalanced data. It may determine the noisy samples as borderline ones and, thus, generates minority samples deteriorating the classification models.

The performance of Cluster-SMOTE is influenced by noisy samples, especially in the synthetic imbalanced dataset 2. As shown in Fig. 10, many synthetic minority samples are located outside the borderlines. This is caused by the fact that the clusters with noisy minority samples are very likely to generate minority samples around the noisy ones.

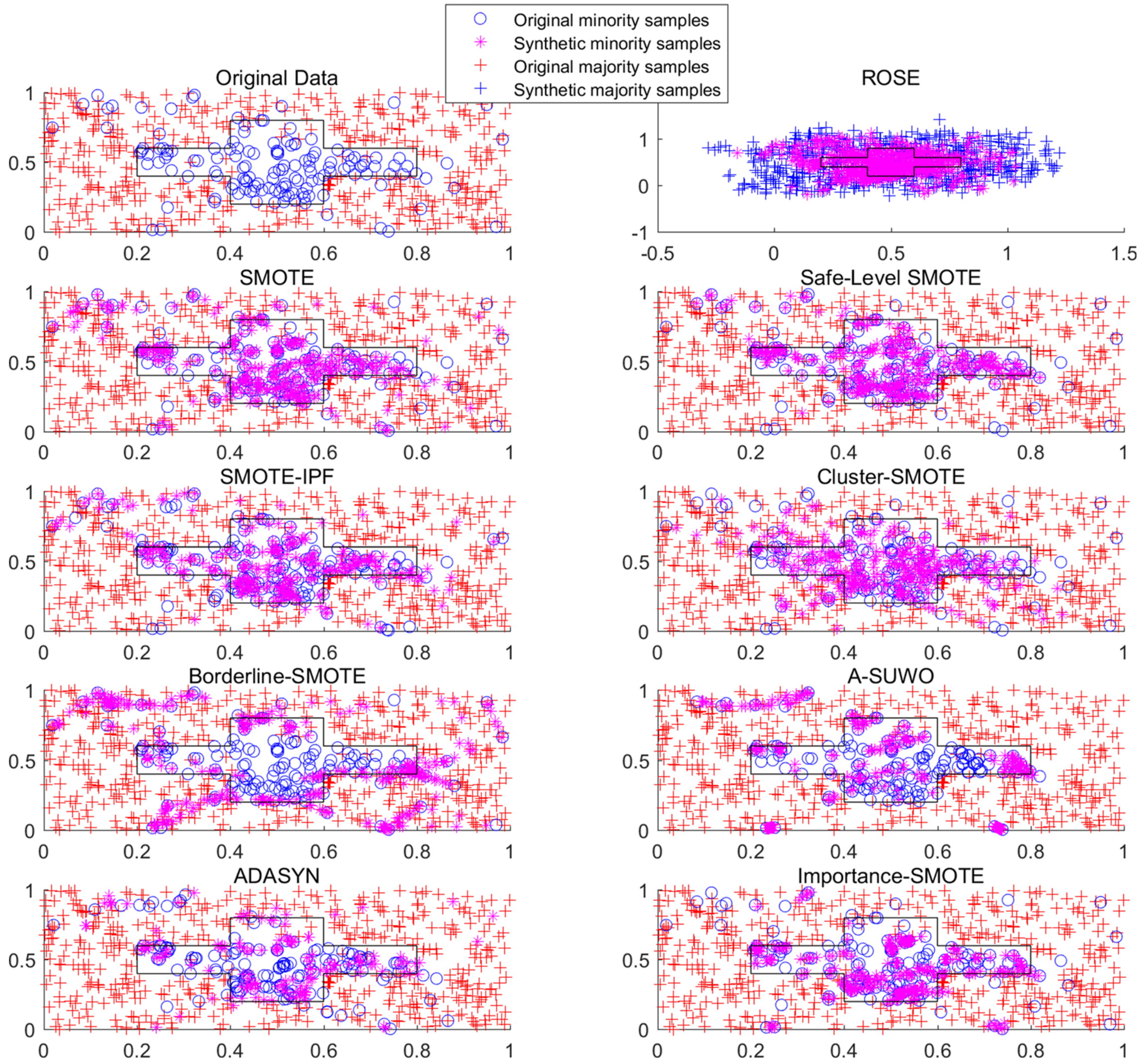


Fig. 9 Experimental results on synthetic imbalanced dataset 1

ADASYN may generate more synthetic minority samples around the noisy minority ones, in order to correctly classify the noisy ones.

A-SUWO faces similar problem as Cluster-SMOTE. When two or more noisy minority samples are accidentally clustered, A-SUWO tends to generate more minority samples close to the noisy ones. Thus, the performance of the classification model can be reduced.

Importance-SMOTE proposed in this work follows a similar idea of safe-level SMOTE and sample importance is introduced. Different from safe-level SMOTE, sample importance in this work is determined by the composition of its nearest neighbors and their distribution around the sample. Like Borderline-SMOTE where samples inside the class are not oversampled, the borderline and edge samples are oversampled in Importance-SMOTE, and they are selected randomly with probabilities proportional to their sample importance. Thus, Importance-SMOTE does not have the same problem of noisy imbalanced data as Borderline-SMOTE. Most of the synthetic minority samples are generated with true minority samples inside the borderlines. Few synthetic samples are generated around the outliers (single minority sample surrounded by majority ones). This is benefited from the fact that outliers have a low probability to be selected during the synthetic minority samples generation process.

Another interesting phenomenon that can be observed from Figs. 9 and 10 is that synthetic minority samples

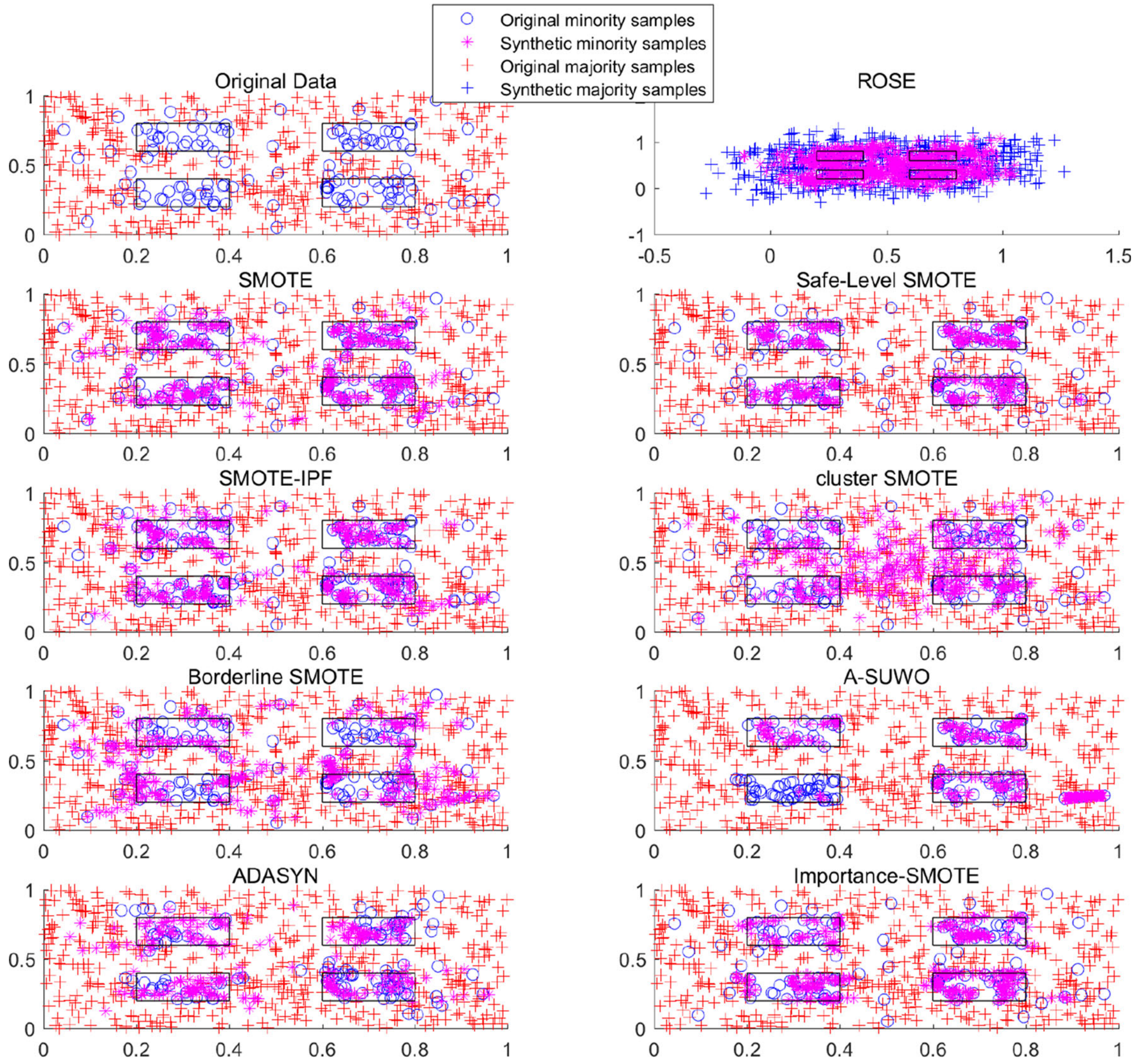


Fig. 10 Experimental results on synthetic imbalanced dataset 2

Table 2 Numbers and percentages (%) of synthetic minority samples inside and outside the borderlines generated by different oversampling methods

Datasets		ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
Dataset 1	Inside	270 (45.0)	251 (80.5)	270 (86.5)	240 (87.0)	242 (77.6)	92 (29.5)	115 (38.2)	107 (74.3)	**274 (87.8)**
	Outside	330 (55.0)	61 (19.5)	42 (13.5)	36 (13.0)	70 (22.4)	220 (70.5)	186 (61.8)	37 (25.7)	**38 (12.2)**
Dataset 2	Inside	156 (26.0)	262 (76.6)	320 (93.6)	262 (86.8)	150 (43.9)	125 (36.6)	201 (58.8)	205 (79.5)	**322 (94.2)**
	Outside	444 (74.0)	80 (23.4)	22 (6.4)	40 (13.2)	192 (56.1)	217 (63.4)	131 (41.2)	53 (20.5)	**20 (5.8)**

The bolded values are the results given by the proposed method

generated with the noisy minority samples, although inevitable in the proposed method, are quite close to the noisy samples. This is due to the fact that the generation of synthetic minority samples with Importance-SMOTE considers not only the minority samples in the neighborhood but also the majority samples, and that the location of a synthetic minority sample is determined by the relative importance of the paired samples. Thus, for a noisy minority sample, its sample importance is relative much lower than the majority samples in its neighborhood and, with (10), the position of the synthetic minority sample is located closely to the noisy minority sample. Thus, from Figs. 9 and 10, one may observe clearly that the synthetic minority samples generated with Importance-SMOTE are centered closely to the noisy minority samples, which means that their influence on the distribution region of majority samples is reduced. By contrary, the concentration effect around the noisy minority samples are not that clear in the benchmark methods.

In conclusion, Importance-SMOTE is an effective oversampling method for these two synthetic noisy imbalanced datasets.

4.3 Results on Keel public datasets

In this part, 26 Keel public imbalanced datasets are retained for evaluate statistically the performance of the proposed oversampling method. Statistical comparisons with the benchmark methods are also carried out.

Friedman test (Demšar 2006) is adopted for testing statistically if the performance of different methods are significantly different, considering their ranks. Since Friedman test considers only the mean rank of different methods, not the relative difference in classification performance metrics. Wilcoxon signed rank test (Rey and Neuhäuser 2011) is, then, adopted for comparing statistically the performances of the proposed method and each benchmark method considering the F-measure and AUC (PRC).

In Friedman test, the null hypothesis is that the performance differences of all the considered methods are not significant. And the alternative hypothesis is that the performance differences are significant. For k methods and n rank results, Friedman test calculates the statistic value F_F which is the dependent of k, n and mean ranks of all methods which are shown in Table 3. The calculation process can be found in Demšar (2006). If the statistic value is larger than the critical value $\mathrm{F}(\mathrm{k}-1,(\mathrm{k}-1)(\mathrm{n}-1))$ at a certain significance level, the alternative hypothesis is accepted. If the null hypothesis is rejected, the critical difference is adopted for pairwise comparisons. If the mean rank difference of two methods are higher than the critical difference, significant difference exists among their performance in the experiments.

Since two classification methods (i.e., KNN and CART) and two performance metrics (i.e., F-measure and AUC (PRC)) are considered in this experiment, comparisons are separately carried out for KNN with F-measure, KNN with AUC(PRC), CART with F-measure, CART with AUC (PRC). The statistic values F_F in this experiment are 18.315, 18.425, 18.305 and 18.390, respectively. In this experiment, the significance level is chosen to be 0.05, and the critical value is 0.3388. It is obvious that significant differences exist among the oversampling methods in this experiment. The critical difference value is 2.378 for the experiments in this work. From Table 3, one may observe that the proposed method is significantly better than ROSE, SMOTE, SMOTE-IFP and A-SUWO with respect to F-measure and KNN in the experiments. The proposed method performs better with respect to F-measure, in comparison with AUC(PRC). Considering AUC(PRC), the proposed method is not significantly better than the benchmark methods.

Considering that Friedman test does not consider the performance difference of the pairwise methods for comparison, Wilcoxon signed rank test is adopted with a significance level of 0.05. The null hypothesis is that the performance difference between Importance-SMOTE and the corresponding benchmark method on the experiment datasets follows a symmetric distribution around zero. The results are shown in Tables 4 and 5. In the tables, the h value "true" means that null hypothesis is accepted, and otherwise, the alternative hypothesis is accepted. The p

Table 3 Average ranks of all methods in the experiment with respect to F-Measure and AUC(PRC)

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
KNN									
F-measure	5.462	5.481	3.865	5.462	4.923	4.519	8.423	4.077	**2.788**
AUC (PRC)	5.462	5.346	4.885	4.731	4.462	5.404	5.154	5.635	**3.923**
CART									
F-measure	4.731	5.462	4.346	5.308	5.596	4.942	8.096	4.058	**2.462**
AUC (PRC)	6.885	4.423	3.865	5.692	5.404	5.115	4.423	5.423	**3.769**

The bolded values are the results given by the proposed method

Table 4 Results of wilcoxon signed tank test between importance-SMOTE and benchmark methods with KNN

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN
F-measure								
h value	**True**	**True**	**True**	**True**	**True**	**True**	**True**	False
p value	0.002	0.000	0.048	0.001	0.002	0.047	0.000	0.056
AUC(PRC)								
h value	False	False	False	**True**	False	False	**True**	**True**
p value	0.334	0.051	0.063	0.500	0.390	0.111	0.049	0.045

The bolded ones are the positive outcome of the statistical tests

Table 5 Results of wilcoxon signed tank test between importance-SMOTE and benchmark methods with cart

Methods	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN
F-measure								
h value	**True**	**True**	**True**	**True**	**True**	**True**	**True**	**True**
p value	2.76e-4	5.34e-5	8.92e-4	6.27e-5	2.16e-5	3.65e-4	6.54e-6	0.011
AUC(PRC)								
h value	**True**	False	False	False	**True**	**True**	False	**True**
p value	0.001	0.352	0.481	0.053	0.026	0.046	0.109	0.026

The bolded ones are the positive outcome of the statistical tests

value means the probability that the accepted hypothesis is violated.

The comparison results of Wilcoxon signed rank test show that Importance-SMOTE achieves at least as good performance as the benchmark methods in the experiment. The proposed method can give significantly better results than the benchmark methods with respect to one/two of the performance metrics.

4.4 Results analysis and exploration

By exploring the experimental results in more details, several empirical conclusions and remarks can be drawn from the experiments:

(1) The proposed oversampling method may achieve better F-measure values and comparable AUC(PRC) in comparison with the considered benchmark methods.

This means that the proposed method has advantages in balancing sensitivity and specification in imbalanced data classification, as indicated by F-measure. However, it does not show significantly superior robustness than the benchmark methods in the experiments. One motivation of the work is to strength the boundary of the minority class by generating synthetic minority samples near the borderline and edge samples. When there is overlapping between the majority and minority classes with noise, the oversampling method increase the true positive rate (recall value) by sacrificing the accuracy on majority class, i.e., low true negative rate or low precision value. The proposed method may increase the true positive rate with a possibly limited loss on the true negative rate through oversampling the borderline and edge samples with larger importance. If the samples with less importance as the noise are oversampled without difference, the decision boundary will be biased to the minority class, and the precision-recall curve will be like benchmark

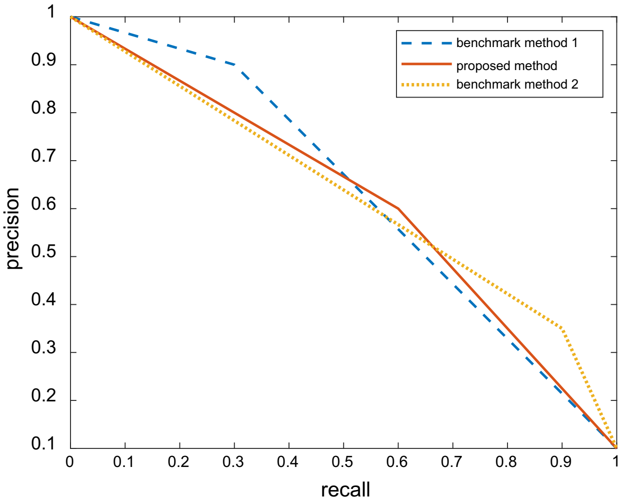


Fig. 11 An illustrative figure of the contradiction between F-measure and AUC(PRC) as performance metrics

method 2 in Fig. 11. On the other hand, if the oversampled borderline and edge samples are not dense enough to represent clearly the distribution space of the minority class, the decision boundary will be biased to the majority class as the precision-recall curve of benchmark method 1 in Fig. 11. For the analysis above, one main advantage of the proposed method is, with comparable AUC(PRC) values, the better balance between the true positive rate and true negative rate in the presence of noise, i.e., significantly better F-measure values.

(2) The performance of the proposed oversampling method is less sensitive to the number of noisy samples, with respect to F-measure.

With a negative correlation between the number of noisy samples and the rank of the F-measure value achieved by the proposed method, its rank on F-measure is better as the number of noise increases, verifying the effectiveness of the proposed method for noisy imbalanced data. By assigning an importance value to each sample and generating synthetic minority samples with respect to the sample importance, the proposed oversampling method may assign small importance for noisy samples and generates as less as possible synthetic minority samples around them.

(3) The performance of KNN is enhanced by oversampling more borderline samples.

With negative correlations between the percentage of borderline samples in minority class and the rank of the proposed oversampling method with KNN with respect to both F-measure and AUC (PRC), the performance of the KNN method is relatively more improved with Importance-SMOTE than the benchmark minority oversampling methods as the percentage of borderline samples in minority class increases. This can be explained by the fact that the proposed method tries to generate more synthetic minority samples close to the borderline. With more synthetic borderline samples, the oversampled minority class is dense on the borderline region, making the minority samples less likely to be misclassified. Since the oversampled samples are close to the minority borderline samples with larger importance, the majority borderline samples are not severely misclassified in the oversampled dataset.

(4) The performance of CART is enhanced by oversampling more edge samples

Similar to point 3), considering the percentage of edge samples in minority class, the performance of the CART method is relatively more improved with Importance-SMOTE than the benchmark oversampling methods. It can be explained by the fact that the proposed method tries to oversample the minority edge samples, and that synthetic minority samples generated around the edge samples are proportional to the percentage occupied by the minority edge samples in the minority class. CART method adopted in this work takes mean squared error as the objective for optimizing the tree structure. Different from KNN that depends highly on local characteristics of the oversampled dataset, CART takes into account the general characteristic of the whole oversampled dataset. The edge samples and the generated synthetic minority samples around them are relatively far from the majority ones and, thus, they may favor in optimizing the CART decision tree structure.

(5) The influence of IR and number of discriminative features on the classification performance are coherent with the previous work

By comparing the performance of the proposed method, it can be observed that the classification performance degrades as the IR increases. This is caused by the fact that limited minority samples cannot represent the distribution of minority class. The proposed method relies on the local characteristics of nearest neighbors for oversampling. While the sample size in minority class decreases, the statistical characteristics of the nearest neighbors are less representative. Similarly, IR is not the only factor to consider for describing the influence of between-class imbalance on the classification performance, the influence is reduced with more discriminative features in the imbalanced datasets. That’s also why adjusted IR is more suitable for describing the difficulty in classification modeling of imbalanced datasets.

(6) The computational complexity is almost stable with respect to the value k of nearest neighbors

The experimental process includes mainly the sample importance calculation, oversampling with respect to sample importance and the classification modeling process. Taking the *Poker8-6* dataset as an example, the change of computation time with respect to the *k* value of nearest neighbors is shown in Fig. 12. One may observe that the computational complexity is almost stable for different steps of the

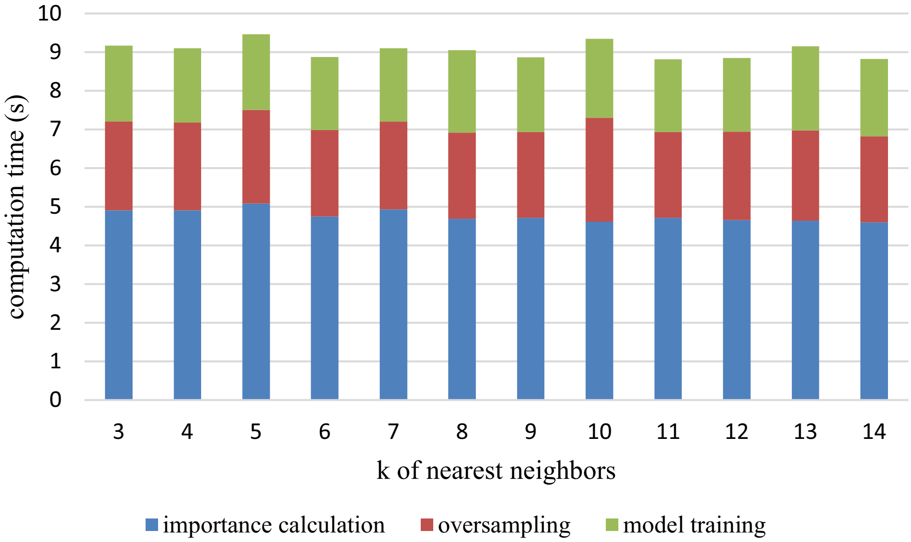


Fig. 12 Computation complexity of SMOTE-Importance for dataset *poker8-6*

experimental process. The computational complexity of the experiment mainly depends on the training data size, the number of synthetic samples and the data size of the oversampled training dataset. Thus, for the same dataset, the computational complexity is almost stable for different k values of nearest neighbors.

5 Conclusions

Noisy samples may reduce the performance of a synthetic minority oversampling method by introducing overlapping and outliers. Considering the composition and distribution of the k nearest neighbors of a borderline or edge sample in minority class, this paper proposes a new minority oversampling method, i.e., Importance-SMOTE. In the proposed method, the minority samples with nearest neighbors from different classes well-separated and with more samples from its own class among its nearest neighbors are given larger importance. The minority samples with larger importance are more likely to be selected and oversampled. Since the borderline and edge samples are more likely to be misclassified, the proposed method oversamples only the borderline and edge sample in minority class. Two popular classification models, i.e., KNN and CART, are integrated in the experiment for binary classification of public imbalanced datasets. The proposed method obtains always highest mean rank in comparison with the benchmark methods. Wilcoxon singed rank test on the experiment results show that the proposed method gives significantly better results than the benchmark methods for most of the comparisons. This work shows the benefits for exploring the local characteristics (e.g., composition and distribution of the nearest neighbors) of the imbalanced training datasets for improving the classification accuracy.

One drawback of the proposed method is that the outliers are not considered. Since outliers can be wrongly judged as noise in the proposed method, one interesting future work is to distinguish noise from outliers and to propose effective importance calculation and oversampling methods. Another research direction is to focus on a cost-sensitive balance between recall and precision in hyperparameters tuning. Considering that the published work shows already that oversampling in high-dimensional Hilbert space with nonlinear kernel functions improves the performance of classical SMOTE method (Mathew et al. 2018), the future work may also extend the current Importance-SMOTE to the Hilbert space for identifying different types of samples, calculating sample importance and generating synthetic minority samples. Considering the randomness brought by the oversampling strategy of the proposed method, the ensemble model has a good potential to improve the robustness and stability of the imbalanced data classification results, especially for dealing with difficult data. While since the samples are given different importance, it should be carefully discussed about the diversity of the sub-models and the combination methods of results from sub-models. Boosting and Bagging in combination with sampling methods are popular strategies for generating different training datasets (Liu et al. 2021; Chen et al. 2021). The sub-models should be assigned a different weight considering the sample importance.

Appendix

See Tables 6, 7, 8, 9

Table 6 Experimental results with respect to F-Measure of KNN

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
03subcl5-600–5-30-BI-fivefold	0.617	0.636	0.615	0.618	0.624	0.585	0.378	0.631	0.639
ecoli-0–1-4-7_vs_2-3–5-6	0.708	0.727	0.756	0.660	0.708	0.713	0.501	0.716	0.743
ecoli-0–1-4-7_vs_5-6	0.744	0.781	0.791	0.704	0.712	0.702	0.575	0.766	0.790
ecoli-0-1_vs_5	0.724	0.711	0.801	0.741	0.758	0.741	0.599	0.779	0.763
ecoli-0–2-3-4_vs_5	0.800	0.771	0.805	0.794	0.783	0.754	0.506	0.805	0.794
ecoli-0–2-6-7_vs_3-5	0.751	0.765	0.696	0.765	0.746	0.718	0.496	0.720	0.753
ecoli-0–3-4-6_vs_5	0.838	0.855	0.855	0.838	0.855	0.838	0.631	0.855	0.855
ecoli-0–3-4_vs_5	0.764	0.778	0.819	0.800	0.779	0.800	0.595	0.800	0.800
ecoli-0–6-7_vs_3-5	0.675	0.711	0.829	0.759	0.715	0.765	0.479	0.812	0.721
ecoli067-5	0.721	0.715	0.776	0.721	0.698	0.721	0.512	0.736	0.763
ecoli2	0.797	0.803	0.875	0.745	0.803	0.805	0.528	0.857	0.851
glass-0–1-5_vs_2	0.380	0.388	0.451	0.448	0.389	0.444	0.197	0.358	0.486
glass-0–1-6_vs_2	0.428	0.279	0.338	0.314	0.362	0.413	0.188	0.379	0.345
glass4	0.665	0.698	0.681	0.705	0.705	0.717	0.620	0.719	0.752
haberman	0.338	0.294	0.306	0.332	0.347	0.357	0.427	0.329	0.339
led7digit-0–2-4–5-6–7-8-9_vs_1	0.352	0.352	0.342	0.352	0.401	0.741	0.536	0.340	0.362
paw02a-600–5-0-BI-fivefold	0.928	0.935	0.914	0.928	0.877	0.928	0.601	0.919	0.928
paw02a-600–5-30-BI-fivefold	0.725	0.708	0.737	0.693	0.697	0.739	0.512	0.707	0.713
pima	0.576	0.522	0.590	0.567	0.603	0.583	0.451	0.570	0.600
poker8-6	0.838	0.877	0.665	0.887	0.761	0.906	0.109	0.843	0.731
shuttle-c2-vs-c4	0.933	0.933	0.933	0.933	0.933	0.933	0.867	0.933	1.000
vehicle1	0.476	0.465	0.456	0.466	0.505	0.465	0.441	0.455	0.476
winequality-red-4	0.165	0.166	0.183	0.169	0.160	0.155	0.146	0.230	0.207
yeast-0–2-5–7-9_vs_3-6–8	0.703	0.688	0.770	0.653	0.717	0.718	0.618	0.761	0.776
yeast-0–3-5-9_vs_7-8	0.338	0.324	0.395	0.334	0.373	0.359	0.306	0.406	0.445
yeast-0–5-6–7-9_vs_4	0.479	0.480	0.519	0.474	0.485	0.503	0.380	0.542	0.545

Table 7 Experimental results with respect to F-Measure of CART

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
03subcl5-600–5-30-BI-fivefold	0.577	0.648	0.649	0.621	0.629	0.629	0.391	0.687	0.615
ecoli-0–1-4-7_vs_2-3–5-6	0.638	0.585	0.685	0.640	0.683	0.638	0.458	0.724	0.691
ecoli-0–1-4-7_vs_5-6	0.694	0.667	0.765	0.642	0.676	0.646	0.548	0.787	0.840
ecoli-0-1_vs_5	0.764	0.716	0.791	0.646	0.701	0.668	0.660	0.701	0.804
ecoli-0–2-3-4_vs_5	0.820	0.669	0.719	0.688	0.712	0.734	0.236	0.797	0.784
ecoli-0–2-6-7_vs_3-5	0.537	0.614	0.639	0.649	0.615	0.615	0.575	0.607	0.624
ecoli-0–3-4-6_vs_5	0.787	0.731	0.663	0.821	0.724	0.729	0.323	0.749	0.821
ecoli-0–3-4_vs_5	0.711	0.756	0.752	0.683	0.737	0.715	0.554	0.747	0.705
ecoli-0–6-7_vs_3-5	0.662	0.569	0.710	0.623	0.628	0.567	0.154	0.723	0.664
ecoli067-5	0.663	0.684	0.688	0.632	0.587	0.643	0.456	0.721	0.775
ecoli2	0.718	0.750	0.745	0.693	0.697	0.722	0.275	0.773	0.762
glass-0–1-5_vs_2	0.333	0.279	0.268	0.351	0.359	0.448	0.202	0.346	0.390
glass-0–1-6_vs_2	0.337	0.266	0.367	0.387	0.246	0.308	0.208	0.274	0.450
glass4	0.739	0.700	0.660	0.752	0.667	0.667	0.468	0.700	0.710
haberman	0.397	0.296	0.266	0.364	0.367	0.323	0.385	0.354	0.387
led7digit-0–2-4–5-6–7-8-9_vs_1	0.763	0.770	0.790	0.773	0.750	0.772	0.441	0.677	0.805
paw02a-600–5-0-BI-fivefold	0.911	0.902	0.894	0.847	0.873	0.912	0.647	0.913	0.902
paw02a-600–5-30-BI-fivefold	0.724	0.685	0.721	0.687	0.685	0.701	0.548	0.715	0.717
pima	0.548	0.582	0.612	0.569	0.593	0.547	0.533	0.543	0.610
poker8-6	0.886	0.731	0.656	0.789	0.600	0.586	0.074	0.325	0.900
shuttle-c2-vs-c4	0.933	0.933	0.933	0.933	0.933	0.933	1.000	0.933	1.000
vehicle1	0.552	0.462	0.521	0.461	0.523	0.551	0.470	0.486	0.523
winequality-red-4	0.079	0.094	0.143	0.112	0.094	0.116	0.137	0.106	0.161
yeast-0–2-5–7-9_vs_3-6–8	0.723	0.735	0.784	0.712	0.723	0.793	0.299	0.796	0.792
yeast-0–3-5-9_vs_7-8	0.294	0.351	0.282	0.309	0.363	0.308	0.182	0.351	0.379
yeast-0–5-6–7-9_vs_4	0.442	0.458	0.400	0.488	0.449	0.491	0.177	0.477	0.478

Table 8 Experimental results with respect to AUC(PRC) of KNN

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
03subcl5-600–5-30-BI-fivefold	0.755	0.760	0.735	0.735	0.785	0.745	0.740	0.750	0.795
ecoli-0–1-4-7_vs_2-3–5-6	0.684	0.868	0.876	0.868	0.860	0.852	0.884	0.868	0.876
ecoli-0–1-4-7_vs_5-6	0.775	0.992	1.000	0.984	0.984	0.984	1.000	0.992	1.000
ecoli-0-1_vs_5	0.852	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625
ecoli-0–2-3-4_vs_5	0.764	0.708	0.722	0.708	0.722	0.708	0.722	0.708	0.847
ecoli-0–2-6-7_vs_3-5	0.825	0.838	0.850	0.963	0.963	0.963	0.750	0.850	0.850
ecoli-0–3-4-6_vs_5	0.973	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ecoli-0–3-4_vs_5	0.972	0.736	0.736	0.736	0.736	0.736	0.750	0.736	0.861
ecoli-0–6-7_vs_3-5	0.950	0.975	0.988	0.963	0.988	0.975	0.988	0.975	0.988
ecoli067-5	0.775	0.725	0.738	0.725	0.738	0.725	0.750	0.713	0.738
ecoli2	0.877	0.974	0.982	0.974	0.974	0.930	0.991	0.956	0.982
glass-0–1-5_vs_2	0.613	0.602	0.634	0.634	0.634	0.651	0.500	0.618	0.651
glass-0–1-6_vs_2	0.586	0.790	0.457	0.762	0.776	0.595	0.500	0.581	0.471
glass4	1.000	0.988	1.000	1.000	1.000	1.000	1.000	0.988	1.000
haberman	0.572	0.569	0.547	0.578	0.514	0.547	0.567	0.610	0.612
led7digit-0–2-4–5-6–7-8-9_vs_1	0.981	0.571	0.571	0.922	0.643	0.571	0.571	0.571	0.571
paw02a-600–5-0-BI-fivefold	0.875	0.950	0.945	0.950	0.935	0.950	0.945	0.950	0.950
paw02a-600–5-30-BI-fivefold	0.803	0.864	0.869	0.874	0.884	0.854	0.849	0.879	0.869
pima	0.578	0.659	0.665	0.655	0.663	0.674	0.692	0.631	0.665
poker8-6	0.877	1.000	1.000	1.000	1.000	0.833	0.833	1.000	0.833
shuttle-c2-vs-c4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
vehicle1	0.551	0.684	0.634	0.657	0.641	0.688	0.634	0.653	0.638
winequality-red-4	0.698	0.569	0.590	0.569	0.668	0.572	0.545	0.564	0.539
yeast-0–2-5–7-9_vs_3-6–8	0.864	0.903	0.891	0.935	0.929	0.880	0.914	0.900	0.920
yeast-0–3-5-9_vs_7-8	0.712	0.618	0.651	0.612	0.640	0.612	0.667	0.612	0.673
yeast-0–5-6–7-9_vs_4	0.784	0.692	0.713	0.742	0.661	0.737	0.629	0.697	0.713
03subcl5-600–5-30-BI-fivefold	0.755	0.760	0.735	0.735	0.785	0.745	0.740	0.750	0.795
ecoli-0–1-4-7_vs_2-3–5-6	0.684	0.868	0.876	0.868	0.860	0.852	0.884	0.868	0.876

Table 9 Experimental results with respect to AUC(PRC) of CART

	ROSE	SMOTE	Safe-level SMOTE	SMOTE-IPF	Cluster-SMOTE	Borderline-SMOTE	A-SUWO	ADASYN	Importance-SMOTE
03subcl5-600–5-30-BI-fivefold	0.675	0.730	0.810	0.840	0.825	0.760	0.715	0.730	0.755
ecoli-0–1-4-7_vs_2-3–5-6	0.492	0.960	0.976	0.984	0.852	0.976	0.984	0.768	0.992
ecoli-0–1-4-7_vs_5-6	0.867	0.984	1.000	0.984	0.984	0.934	0.992	0.984	1.000
ecoli-0-1_vs_5	0.750	0.750	0.750	0.750	0.727	0.614	0.750	0.750	0.750
ecoli-0–2-3-4_vs_5	0.375	0.861	0.847	0.833	0.708	0.694	0.847	0.722	0.722
ecoli-0–2-6-7_vs_3-5	0.500	0.725	0.750	0.700	0.738	0.738	0.750	0.738	0.738
ecoli-0–3-4-6_vs_5	0.716	0.986	0.986	0.973	0.848	0.861	0.986	0.973	0.973
ecoli-0–3-4_vs_5	1.000	1.000	0.875	0.875	0.986	0.750	0.875	0.861	1.000
ecoli-0–6-7_vs_3-5	0.500	0.988	0.988	0.950	0.850	0.963	0.975	0.950	1.000
ecoli067-5	0.688	0.838	0.875	0.813	0.863	0.713	0.488	0.850	0.863
ecoli2	0.509	0.874	0.874	0.865	0.856	0.847	0.882	0.865	0.932
glass-0–1-5_vs_2	0.645	0.618	0.667	0.468	0.618	0.634	0.484	0.484	0.634
glass-0–1-6_vs_2	0.476	0.638	0.471	0.443	0.443	0.610	0.471	0.638	0.486
glass4	0.425	0.988	0.988	0.975	0.975	1.000	0.988	0.988	1.000
haberman	0.512	0.585	0.639	0.630	0.523	0.565	0.599	0.516	0.565
led7digit-0–2-4–5-6–7-8-9_vs_1	0.944	0.994	0.916	0.994	0.916	0.994	0.922	0.922	0.922
paw02a-600–5-0-BI-fivefold	0.825	0.990	0.990	0.910	0.930	0.840	0.940	0.920	0.980
paw02a-600–5-30-BI-fivefold	0.808	0.839	0.824	0.784	0.824	0.854	0.839	0.829	0.824
pima	0.557	0.605	0.666	0.639	0.686	0.687	0.697	0.700	0.669
poker8-6	0.837	0.490	0.486	0.490	0.495	0.493	0.667	0.667	0.500
shuttle-c2-vs-c4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
vehicle1	0.701	0.761	0.614	0.676	0.711	0.644	0.700	0.661	0.672
winequality-red-4	0.588	0.516	0.539	0.508	0.555	0.616	0.545	0.555	0.535
yeast-0–2-5–7-9_vs_3-6–8	0.623	0.917	0.963	0.954	0.946	0.878	0.931	0.864	0.963
yeast-0–3-5-9_vs_7-8	0.500	0.562	0.651	0.512	0.628	0.640	0.573	0.573	0.667
yeast-0–5-6–7-9_vs_4	0.500	0.658	0.668	0.663	0.732	0.692	0.508	0.597	0.608
03subcl5-600–5-30-BI-fivefold	0.675	0.730	0.810	0.840	0.825	0.760	0.715	0.730	0.755
ecoli-0–1-4-7_vs_2-3–5-6	0.492	0.960	0.976	0.984	0.852	0.976	0.984	0.768	0.992

Acknowledgements This work was supported in part by National Natural Science Foundation of China (No.52005027).

Funding The author declares he has no financial interests.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

Human and animal rights This article does not contain any studies with human participants performed by any of the authors.

References

Bach M, Werner A, Żywiec J, Pluskiewicz W (2017) The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. Inf Sci (Ny) 384:174

Barua S, Islam MM, Yao X, Murase K (2014) MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl Data Eng 26:405

Branco P, Torgo L, Ribeiro RP (2016) (不平衡数据综述) A survey of predictive modeling on imbalanced domains. ACM Comput. Surv. 49(2):1

Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2009) Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class Imbalanced Problem. Pacific-asia Conference on Advances in Knowledge Discovery & Data Mining, Springer-Verlag, pp 475–482

Chen Z, Duan J, Kang L, Qiu G (2021) A hybrid data-level ensemble to enable learning from highly imbalanced dataset. Inf Sci (Ny) 554:157

Cieslak DA, Chawla NV, Striegel A (2006) "Combating imbalance in network intrusion datasets.," in *GrC*, pp. 732–737

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Fernández A, del Río S, Chawla NV, Herrera F (2017) An insight into imbalanced Big Data classification: outcomes and challenges. Complex Intell. Syst. 3:105

Fernández A, García S, Herrera F, Chawla NV (2018) SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. J Artif Intell Res 61:863

Han H, Wang W, Mao B (2005) "Borderline-SMOTE : A New Over-Sampling Method in," in *International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23–26 Proceedings, Part I*, 2005

Hassib EM, El-Desouky AI, Labib LM, El-kenawy ESM (2019) WOA + BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network. Soft Comput. 24:5573

He H, Bai Y, Garcia EA, Li S (2008) "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans. Knowl. Data Eng 21(9):1263–1284

Japkowicz N (2000) The class imbalance problem: significance and strategies," in Proceedings of the 2000 International Conference on Artificial Intelligence

Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R (2018) Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans. Neural Netw Learn. Syst. 29 (8):3573

Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput. J. 83:105662

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5:221

Last F, Douzas G, Bacao F (2017) "Oversampling for Imbalanced Learning Based on K-Means and SMOTE,"

Laurikkala J (2001) "Improving identification of difficult small classes by balancing class distribution," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 63–66

Li Y, Maguire L (2011) Selecting critical patterns based on local geometrical and statistical information. IEEE Trans. Pattern Anal. Mach. Intell. 33:1189

Liu J, Zio E (2018) A scalable fuzzy support vector machine for fault detection in transportation systems. Expert Syst Appl 102:36

Liu M, Miao L, Zhang D (2014) Two-stage cost-sensitive learning for software defect prediction. IEEE Trans. Reliab 63:676

Liu J, Li YF, Zio E (2017) A SVM framework for fault detection of the braking system in a high speed train. Mech. Syst. Signal Process 87:401

Liu X, Yi GY, Bauman G, He W (2021) Ensembling imbalanced-spatial-structured support vector machine. Econom. Stat. 17:145

López V, Fernández A, García S, Palade V, Herrera F (2013) "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics,." Inf Sci (Ny) 250:113–141

MacIejewski T, Stefanowski J (2011) "Local neighbourhood extension of SMOTE for mining imbalanced data," in *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining*

Mathew J, Pang CK, Luo M, Weng HL (2018) Classification of imbalanced data by oversampling in kernel space of support vector machines. Neural Netw Learn Syst IEEE Trans 29 (9):4065–4076

Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28(1):92–122

Napierala K, Stefanowski J (2012) "Identification of different types of minority class examples in imbalanced data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*

Napierała K, Stefanowski J (2015) Addressing imbalanced data with argument based rule learning. Expert Syst Appl 42:9468

Napierala K, Stefanowski J (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst 46:563

Nekooeimehr I, Lai-Yuen SK (2016) Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. Expert Syst Appl 46:405

Noorhalim N, Ali A, Shamsuddin SM (2019) "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*

Piri S, Delen D, Liu T (2018) A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. Decis Support Syst 106:15

Rey D, Neuhäuser M (2011) Wilcoxon-signed-rank test. In: Lovric M (ed) International encyclopedia of statistical science. Springer, Berlin, Heidelberg, pp 1658–1659. https://doi.org/10.1007/978-3-642-04898-2_616

Rivera WA (2017) "Noise reduction a priori synthetic over-sampling for class imbalanced data sets,." Inf Sci (Ny) 408:146–161

Sáez JA, Luengo J, Stefanowski J, Herrera F (2015) SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci (Ny) 291:184

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10:e0118432

Shilaskar S, Ghatol A (2019) Diagnosis system for imbalanced multi-minority medical dataset". Soft Comput 23:4789

Skryjomski P, Krawczyk B (2017) "Influence of minority class instance types on SMOTE imbalanced data oversampling," in *Proceedings of Machine Learning Research LIDTA 2017*

Stefanowski J, Napierała K, Trzcielińska M (2014) Local characteristics of minority examples in pre-processing of Imbalanced Data. In: Andreasen T, Christiansen H, Cubero J-C, Raś ZW (eds) Foundations of intelligent systems (ISMIS 2014 Roskilde, Denmark, June 25–27, 2014 Proceedings) . Springer, Cham, pp 123–132

Tuncer T, Dogan S (2019) A novel octopus based Parkinson's disease and gender recognition method using vowels. Appl. Acoust. 155:75

Tuncer T, Dogan S, Acharya UR (2020) Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybern. Biomed. Eng. 40:211

Wang B, Japkowicz N (2004) "Imbalanced data set learning with synthetic samples," in *InProc. IRIS Machine Learning Workshop*

Xu Y, Wu C, Zheng K, Niu X, Yang Y (2017) Fuzzy-Synthetic minority oversampling technique: oversampling based on fuzzy set theory for android malware detection in imbalanced datasets. Int J Distrib Sens Netw. https://doi.org/10.1177/1550147717703116

Zhai J, Zhang S, Zhang M, Liu X (2018) Fuzzy integral-based ELM ensemble for imbalanced big data classification. Soft Comput 22:3519

Zhu R, Guo Y, Xue JH (2020) Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognit Lett. 133:217