Applied and Computational Mathematics
2024; 13(5): 165-177
http://www.sciencepublishinggroup.com/j/acm
doi: 10.11648/j.acm.20241305.15
ISSN: 2328-5605 (Print); ISSN: 2328-5613 (Online)

Pseudo Oversampling Based on Feature Transformation and Fuzzy Membership Functions for Imbalanced and Overlapping Data

Tingting Pan[1], Witold Pedrycz[2], Jie Yang[3, *], Dahai Zhang[1]

[1]Department of Basic Courses Teaching, Dalian Polytechnic University, Dalian, China
[2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[3]School of Mathematical Sciences, Dalian University of Technology, Dalian, China

Email address:
flower_bloom_inf@163.com (Tingting Pan), pedrycz@ee.ualberta.ca (Witold Pedrycz), yangjiee@dlut.edu.cn (Jie Yang), zhang_dh@dlpu.edu.cn (Dahai Zhang)
*Corresponding author

To cite this article:
Tingting Pan, Witold Pedrycz, Jie Yang, Dahai Zhang. (2024). Pseudo Oversampling Based on Feature Transformation and Fuzzy Membership Functions for Imbalanced and Overlapping Data. *Applied and Computational Mathematics, 13*(5), 165-177. https://doi.org/10.11648/j.acm.20241305.15

Received: 23 July 2024; **Accepted:** 20 August 2024; **Published:** 19 September 2024

Abstract: Class imbalance in data poses challenges for classifier learning, drawing increased attention in data mining and machine learning. The occurrence of class overlap in real-world data exacerbates the learning difficulty. In this paper, a novel pseudo oversampling method (POM) is proposed to learn imbalanced and overlapping data. It is motivated by the point that overlapping samples from different classes share the same distribution space, and therefore information underlying in majority (negative) overlapping samples can be extracted and used to generate additional positive samples. A fuzzy logic-based membership function is defined to assess negative overlaps using both local and global information. Subsequently, the identified negative overlapping samples are shifted into the positive sample region by a transformation matrix, centered around the positive samples. POM outperforms 15 methods across 14 datasets, displaying superior performance in terms of metrics of G_m, F_1 and AUC.

Keywords: Imbalanced Learning, Class Overlap, Feature Transformation, Oversampling

1. Introduction

Class imbalance data refers to at least one of its classes is usually outnumbered by the other classes. Many supervised learning tasks in real world applications involve imbalanced datasets, such as those encountered in disease diagnosis [1], financial risk prediction [2], bug recognition [3], and many others. In these domains, people have more interest in minority (commonly called positive) classes and expect a lower misclassification cost for positive samples after training process. Most studied imbalanced problems focus on oversampling which generates additional positive samples or undersampling which randomly selects a subset of the majority (commonly called negative) class without replacement [4]. However, the failure of conventional classifiers in imbalanced learning is not always caused by the skewed distributions between negative classes and positive classes solely [5]. Beyond that, class overlap makes learning imbalanced data even harder [6, 7].

Class overlap means that there are regions in the feature space where it is difficult or impossible to determine with certainty which class a sample belongs to, based on its feature values alone. Some studies demonstrate the significant impact of class overlap on the performance of classifiers that learn from imbalanced datasets [5]. In order to address the issue of class overlap, various undersampling approaches have been devised with the aim of improving the distinction within inter-class sample distributions, such as Neighbourhood-based Undersampling (NBU) [8], Density-Based Majority Under-Sampling Technique (DBMUTE) [9] and Overlap-Based

Undersampling (OBU) [10]. However, these efforts may entail a loss of information.

Although finding a clear classification boundary could be challenging in many real-world datasets due to overlapping regions [11], these regions also indicate that samples with different classes share similar information in the feature space. Due to the sufficient number of negative samples (N) and the insufficient number of positive samples (P) in imbalanced datasets, the information contained in N is often much greater than that in P [12]. However, this valuable information is not fully exploited by most existing rebalancing methods. Therefore, a natural approach is to mine information from negative overlapping samples and transfer it to P to enrich the feature expression.

In machine learning, the distribution of samples in a dataset is an important aspect of the information contained in the dataset [13, 14]. That is because the distribution information can provide critical insights into the underlying patterns and relationships between the features and the target variable.

As distribution information is typically carried by samples and expressed through their features, this paper introduces a novel pseudo oversampling method (POM) designed to extract the distribution information inherent in negatively overlapping samples, thereby generating additional positive samples. This method accomplishes this by converting negative overlapping samples into positive ones. By this transformation, the feature representation of the positive class can be enhanced, leading to an enhancement in the performance of classifiers trained on rebalanced datasets using POM.

The initialization of POM is to identify negative overlapping samples, which is often a time-consuming process [15, 16]. To address this issue, a novel membership function fuzzy set is introduced, which utilizes both local and global information derived from negative samples, as well as their distances to class centers. The membership function directly describes the degree to which a negative sample is an overlapping sample. By setting a threshold on this degree of belonging, negative overlapping samples can be identified efficiently.

The sample transformation of the proposed method is to generate additional positive samples through transforming the identified negative overlapping samples. The motivation for defining the transformation matrix is rooted in the statistical principle that features with distinctive differences between their respective means in a dataset are considered to have strong expressiveness, indicating that they contain significant distribution information. Therefore, the covariance matrix of negative features naturally becomes a key consideration in the proposed approach.

The main contributions of this paper are outlined as follows:

1. A new pseudo oversampling method is proposed to solve the skew distribution and class overlap problems which are important and frequent in imbalanced learning.
2. A membership function is proposed to quantitatively and directly measure the extent to which a negative sample is an overlapping sample.
3. A transformation matrix deduced in this paper provides a way to extract and compress the critical information of datasets.

The originality of POM stems from amount of explorations. Compared with existing literatures over oversampling methods [4, 17], POM strives to augment the distribution of the positive class by leveraging distribution information shared between the positive and negative classes, as opposed to solely relying on positive samples or negative samples without a solid foundation. POM aims to transform negative overlapping samples into positive samples, rather than generating new positive samples like most existing oversampling methods, which maximizes the shared information onto the generated samples. However, there have not been studies on measuring the overlapping degree of samples with the membership function from the perspective of fuzzy field. The membership function defined in POM is related to local information of nearest neighbors of every negative sample and the global information of class centers.

The paper is organized as follows: the proposed method POM is detailed in Section 2, introducing a novel membership function and sample transformation matrix. Section 3 reports the experimental results of POM compared with popular imbalanced methods and discusses on the effects on the performance of POM. Section 4 concludes the study. What's more, abbreviations defined upon the first appearance in the main body are listed in Table 1.

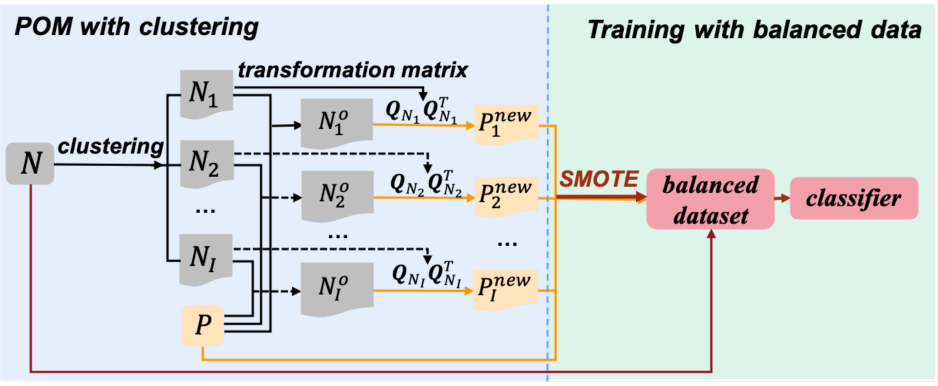

Figure 1. Overall flow of processing of the proposed method POM.

Table 1. *List of abbreviation and acronyms used in the paper.*

Abbreviation	Explanation
POM	Pseudo Oversampling Method proposed in this paper
NBU	Neighbourhood-based Undersampling [8]
DBMUTE	Density-Based Majority Under-Sampling Technique [9]
OBU	Overlap-Based Undersampling [10]
RUS	Random UnderSampling
TL	TomekLinks [21]
SM	SMOTE (Synthetic Minority Over-sampling Technique) [20]
SSM	SVMSMOTE sampling [22]
ROS	Random OverSampling
BSM	Borderline SMOTE [23]
ADA	ADAptive SYNthetic sampling [24]
SMT	SMOTE Tomek [25]
SME	SMOTE and Edited Nearest Neighbors [26]
BBC	Balanced Bagging Classifier [27]
EEC	Easy Ensemble Classifier [28]
BRFC	Balanced Random Forest Classifier [29]
RBC	RUS Boost Classifier [30]
REMDD	Resampling Ensemble Model based on Data Distribution [31]
OPF-US	Optimum-Path Forest UnderSampling [32]

2. Proposed Method

Consider a binary imbalanced classification problem with a training dataset $S = P \cup N$ with n features, where $P \in R^{|P|\times n}$ represents positive dataset and $N \in R^{|N|\times n}$ denotes negative dataset. For every sample $\boldsymbol{x} \in S$, if the class label $y = +1$, then $\boldsymbol{x} \in P$; if the class label $y = -1$, then $\boldsymbol{x} \in N$. Besides, $P \cap N = \phi$ and $|P| \ll |N|$. Figure 1 depicts an overall flow of processing of the proposed POM. To investigate the global distribution rather than on local information, clustering is performed as a preliminary step on N [18]. Specifically, samples of N are clustered into I subsets, denoted as $N_1, N_2, ..., N_I$, where I is determined using the C-H score [19]. POM is then operated on each dataset $N_i \cup P$ separately, where $i = 1, 2, ..., I$.

The proposed POM consists of three main steps. Firstly, in Section 2.1, a membership function is defined to identify negative samples that overlap with positive samples. Secondly, in Section 2.2, a transformation matrix is constructed to compress the information of negative data. Thirdly, in Section 2.3, the negative overlapping samples are transformed into positive samples using the transformation matrix. Finally, to clearly understand the role of each step, the process of POM on a toy dataset is visualized in Section 2.4.

2.1. Identification of Negative Overlapping Samples with Membership Function

A negative sample can be considered an overlapping sample if its neighbors contain both positive and negative samples. The concept of membership function naturally provides a way to address the subjectivity of this identification. Notice that the degree to which a negative sample $\boldsymbol{x}$ is an overlapping sample is positively correlated with the proportion of positive samples among its neighboring samples. Furthermore, the distance between $\boldsymbol{x}$ and class centers is considered to refine the definition. Based on these considerations, the membership function of $\boldsymbol{x}$ as an overlapping sample is defined as follows:

$$MF(\boldsymbol{x}) = e^{-\frac{1}{2}\cdot(\frac{K_P(\boldsymbol{x})/K-1}{D_{N_i}(\boldsymbol{x})/(D_{N_i}(\boldsymbol{x})+D_P(\boldsymbol{x}))})^2}, \tag{1}$$

where $K_P(\boldsymbol{x})$ is the number of positive samples belonging to the K nearest neighbors of $\boldsymbol{x}$. $D_P(\boldsymbol{x})$ (or $D_{N_i}(\boldsymbol{x})$) is the Euclidean distance between $\boldsymbol{x}$ and the center of P (or N_i). This membership function capitalizes on both the local information by considering the neighbors of $\boldsymbol{x}$ and the global information by taking into account the distances between $\boldsymbol{x}$ and the class centers.

As shown in Figure 2, negative samples $\boldsymbol{x}$, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ and $\boldsymbol{x_3}$ are distributed in different areas. For $\boldsymbol{x}$, its 10 nearest neighbors consist of 4 positive samples, i.e., the difference between the number of positive and negative samples in the neighbors is not large. Besides, $D_N(\boldsymbol{x})$ is close to $D_P(\boldsymbol{x})$ in Euclidean space. Therefore, $MF(\boldsymbol{x})$ is close to $e^{-\frac{1}{2}}$ according to (1). $\boldsymbol{x}_1$ is located in the inner of P and most of its nearest neighbors are positive samples, so $K_P(\boldsymbol{x}) \gg K - K_P(\boldsymbol{x})$. Besides, $\boldsymbol{x}_1$ is close to the center of P, so $D_N(\boldsymbol{x}_1) \gg D_P(\boldsymbol{x}_1)$. Therefore, $MF(\boldsymbol{x}_1)$ is close to e^0. Considering $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ located respectively in the inner and boundary of N, the neighbors of them are full of negative samples. Besides, $D_P(\boldsymbol{x}_2) \gg D_N(\boldsymbol{x}_2)$ and $D_P(\boldsymbol{x}_3) > D_N(\boldsymbol{x}_3)$. Therefore, $MF(\boldsymbol{x}_3) > MF(\boldsymbol{x}_2) > 0$ and $MF(\boldsymbol{x}_2) \to 0$. In general, both the neighbors of a negative sample and the distance between the negative sample and the central points are considered, i.e., both local information and global information can be incorporated in the membership function of the negative sample as an overlapping one.

Then the negative overlapping samples can be identified as follows:

$$N_i^o = \{(\boldsymbol{x}, y) \in N_i | MF(\boldsymbol{x}) > \epsilon\}. \tag{2}$$

where $\epsilon \in (0, 1)$ is threshold. With this rule, negative samples with large membership degrees to overlapping regions are highly likely to be transformed into positive samples.

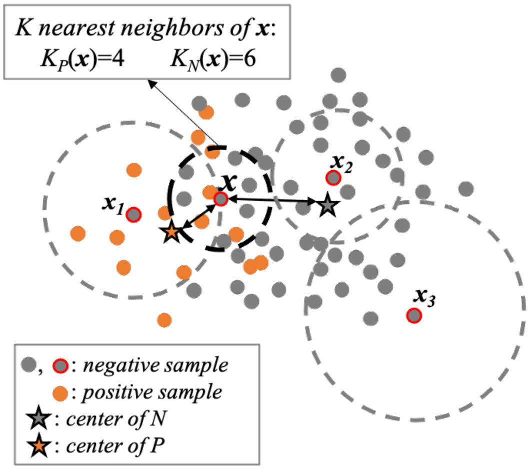

Figure 2. *Negative samples distributed in different areas have different overlapping degree.*

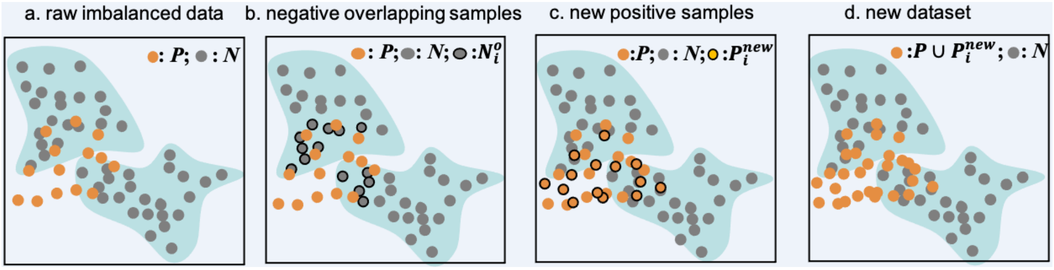

Figure 3. Oversampling with POM on a toy dataset.

2.2. Construction the Transformation Matrix

Given a cluster of negative training set N_i, it can be represented with its features as $N_i = (\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_n)$, where $\boldsymbol{F}_j \in R^{|N_i|\times 1}(j = 1, 2, \ldots, n)$ is the j-th feature vector. Then the covariance matrix of N_i is written as:

$$\boldsymbol{\Sigma}_{N_i} = \begin{bmatrix} cov_{1,1} & cov_{1,2} & \cdots & cov_{1,n} \\ cov_{2,1} & cov_{2,2} & \cdots & cov_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ cov_{n,1} & cov_{n,2} & \cdots & cov_{n,n} \end{bmatrix}, \tag{3}$$

where $cov_{j_1,j_2} = (\boldsymbol{F}_{j_1} - \overline{\boldsymbol{F}})^T \cdot (\boldsymbol{F}_{j_2} - \overline{\boldsymbol{F}})$ shows the covariance of $\boldsymbol{F}_{j_1}$ and $\boldsymbol{F}_{j_2}$ $(j_1, j_2 = 1, 2, \ldots, n)$, and $\overline{\boldsymbol{F}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{F}_j$. Because $\boldsymbol{\Sigma}_{N_i}$ is real symmetric and positive semi-definite, it has n linearly independent and orthogonal eigenvectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n \in R^{|N_i|}$ corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$.

From the information theory perspective, the more scattered the dataset distribute, the more information it contains. Notice that if the absolute value of an eigenvalue λ_i is high, the projections of all negative samples on the eigenvector $\boldsymbol{\xi}_i$ of λ_i are scattered. It means that the larger the eigenvalue, the more information of N_i is stored in the corresponding eigenvector. It is assumed that not all eigenvalues of $\boldsymbol{\Sigma}_{N_i}$ are 0. Then some eigenvectors are selected to represent features of N_i:

$$\boldsymbol{Q}_{N_i} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T). \tag{4}$$

The number of eigenvectors T can be expressed as follows:

$$T = \arg\min_t \frac{\sum_{i=1}^{t} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \gamma (\lambda_t > 0), \tag{5}$$

where γ is a constant. For example, if $\gamma = 0.9$, one can think that there are T eigenvectors retain more than 90% information of N_i. Besides, $\lambda_t > 0$ confirms that the t eigenvectors selected into matrix $\boldsymbol{Q}_{N_i}$ are effective. So, $\boldsymbol{Q}_{N_i}$ is actually a compressed feature matrix of N_i. Finally, the transformation matrix is constructed as $\boldsymbol{Q}_{N_i}\boldsymbol{Q}_{N_i}^T$.

2.3. Sample Transformation

Because samples contained in N_i^o account for only a part of the overall negative samples and locate at the overlapping regions, the distribution characteristics of the negative class reflected by these samples are not obvious enough or biased. Therefore, the matrix $\boldsymbol{Q}_{N_i}\boldsymbol{Q}_{N_i}^T$ is used to transfer every sample $\boldsymbol{x}_{N_i^o} \in N_i^o$:

$$\boldsymbol{x}_i^{new} = \boldsymbol{Q}_{N_i}\boldsymbol{Q}_{N_i}^T \cdot \boldsymbol{x}_{N_i^o} + \overline{\boldsymbol{x}^P}, \tag{6}$$

where $\overline{\boldsymbol{x}^P}$ is the center of positive class. $\boldsymbol{Q}_{N_i}\boldsymbol{Q}_{N_i}^T$ is imposed on sample $\boldsymbol{x}_{N_i^o}$, which transfers the distribution information of the negative subset N_i to N_i^o. And $\overline{\boldsymbol{x}^P}$ shifts $\boldsymbol{Q}_N\boldsymbol{Q}_N^T \cdot \boldsymbol{x}_{N_i^o}$ to close the center of P, which makes $\boldsymbol{x}_i^{new}$ more similar with positive samples. Then all these $\boldsymbol{x}_i^{new}$ are collected into a set, namely P_i^{new}.

If the size of $(\bigcup_{i=1}^{I} P_i^{new}) \bigcup P$ is smaller than the size of N, Synthetic Minority Over-sampling Technique (SMOTE) [20] would be used to oversample on this union; otherwise, random undersampling would be operated on $\bigcup_{i=1}^{I} P_i^{new}$.

2.4. Visualization of POM on a Toy Dataset

The process of POM is visualized on a toy dataset shown in Figure 3, where the negative samples can be clustered into two clustering points with grey colors and positive samples are presented with yellow points (Figure 3a). Some negative samples are located within the positive class, which is clearly a case of sample overlap and can lead to confusion in conventional classifiers with respect to their labels. With (2), some negative samples are identified as negative overlapping samples (N_i^o), which are scatted by grey points with black rings as shown in Figure 3b.

Samples of two clusters in N_i^o are transformed into positive samples, respectively, as shown in Figure 3c. New generated positive samples (yellow points with black rings, noted as P_i^{new}) not only store the feature information of the negative class and the location information of the negative overlapping samples, but also try to get close to the negative class, making the data distribution characteristics more obvious. In Figure 3d, the distribution of new positive data ($P \bigcup P_i^{new}$) is strengthened and more consistent with peoples' visual perception.

3. Experiments

In this section, the performance of the proposed method is evaluated on 14 commonly used datasets in terms of three quality assessment metrics: G_m, F_1, AUC.

3.1. Experiment Setting

3.1.1. Datasets

In this paper, 4 binary imbalanced datasets are collected from Machine Learning Repository UCI[1] and 10 binary imbalanced datasets are coming from KEEL[2]. The details of these datasets are presented in Table 2. The size of data varies from 214 to 2000, the number of attributes ranges from 4 to 34, and the imbalanced ratio (IR) ranges from 2.9 to 85.88.

Table 2. *The details of 14 binary datasets used in experiments.*

ID	Datasets	Size	$\|N\|$	$\|P\|$	Attr.	IR
D1	vehicle1	846	629	217	18	2.9
D2	dermatology3_6	358	267	91	34	2.93
D3	Wireless Indoor Localization2	2000	1500	500	7	3
D4	Wireless Indoor Localization4	2000	1500	500	7	3
D5	Blood	748	570	178	4	3.2
D6	glass6	214	185	29	9	6.38
D7	led7digit-0-2-4-5-6-7-8-9_vs_1	443	406	37	7	10.97
D8	dermatology6	358	338	20	34	16.9
D9	yeast4	1484	1433	51	8	28.1
D10	winequality-red4	1599	1546	53	11	29.17
D11	kr-vs-k-zero_vs_eight	1460	1433	27	6	53.07
D12	winequality-white-3-9_vs_5	1482	1457	25	11	58.28
D13	poker-8-9_vs_6	1485	1460	25	10	58.4
D14	poker-8_vs_6	1477	1460	17	10	85.88

Table 3. *15 methods used to compare with POM.*

Group	Imbalanced methods
Group1: undersampling methods	RUS: Random UnderSampling TL: TomekLinks [21]
Group2: oversampling methods	ROS: Random OverSampling SM: SMOTE [20] SSM: SVMSMOTE [22] BSM: Borderline SMOTE [23] ADA: ADAptive SYNthetic sampling [24]
Group3: combined sampling methods	SMT: SMOTE Tomek [25] SME: SMOTEENN [26]
Group4: sampling combined with integrated learning methods	RUS: Random UnderSampling BBC: Balanced Bagging Classifier [27] EEC: Easy Ensemble Classifier [28] BRFC: Balanced Random Forest Classifier [29] RBC: RUS Boost Classifier [30]
Group5: advanced methods	REMDD [31] (2020): Resampling Ensemble Model based on Data Distribution OPF-US [32] (2022): Optimum-Path Forest UnderSampling

3.1.2. Evaluation Metrics

Accuracy (ACC) is a commonly employed metric for balanced datasets. For imbalanced datasets, there are four fundamental measures, TP, FP, TN and FN. $FP(TN)$ is the number of positive (negative) samples classified into negative class incorrectly (correctly), and $FN(TP)$ can be understood similarly. Three widely-used evaluation metrics, namely G_m, F_1, and AUC, are considered based on the aforementioned fundamental measures. The metric G_m is defined as the geometric mean between the true positive rate ($TPR = TP/(TP + FN)$) and the true negative rate ($TNR = TN/(TN + FP)$). The formula for G_m is given by the following equation:

$$G_m = \sqrt{TPR \times TNR}. \tag{7}$$

If there is any poor performance of the classifier, the values of G_m will become low. The metric F_1 is a harmonic mean between $Recall = TP/(TP + FN)$ and $Precision = TP/(TP + FP)$, and is defined in the form

$$F_1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{8}$$

It simultaneously assesses the performance of the positive class from the perspectives of both true positives and predicted positives.

By adjusting the threshold value of decision model, a ROC curve showing the relationships between the values of TPR and FPR is generated. AUC is the area under ROC, which shows the trade-off between TPR and FPR.

[1] https://archive.ics.uci.edu/ml/index.php
[2] http://sci2s.ugr.es/keel/imbalanced.php

3.1.3. Reference Methods

15 popular imbalanced methods are respectively combined with the same classifier SVM to compare with POM under same settings. Those approaches are divided into five groups: two undersampling methods, five oversampling methods, two combined sampling methods, four ensemble learning methods, and two advanced learning methods. The detailed information is listed in Table 3.

3.2. Parameter Settings

As a commonly used and outstanding classifier, Support Vector Machine (SVM) [33] has the capability to generate posterior class probabilities for training samples. Therefore, SVM with Gaussian kernel is employed as the chosen classifier in this paper. To assess overall performance stable, POM and other baseline methods are evaluated by averaging the results from ten independent runs of 5-fold cross-validation on each dataset.

All experiments are implemented with Python 3.7 software and run on a computer with an i7-7700 CPU and 32 GB of RAM. Additionally, there are two parameters ϵ and γ in this paper considered as follows:

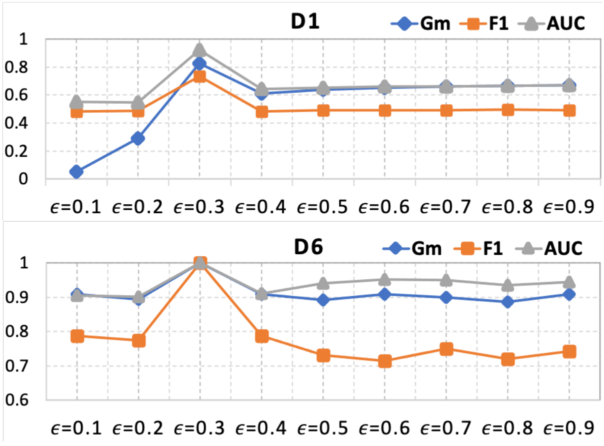

Figure 4. *The performance of POM related parameter ϵ on datasets D1 and D6.*

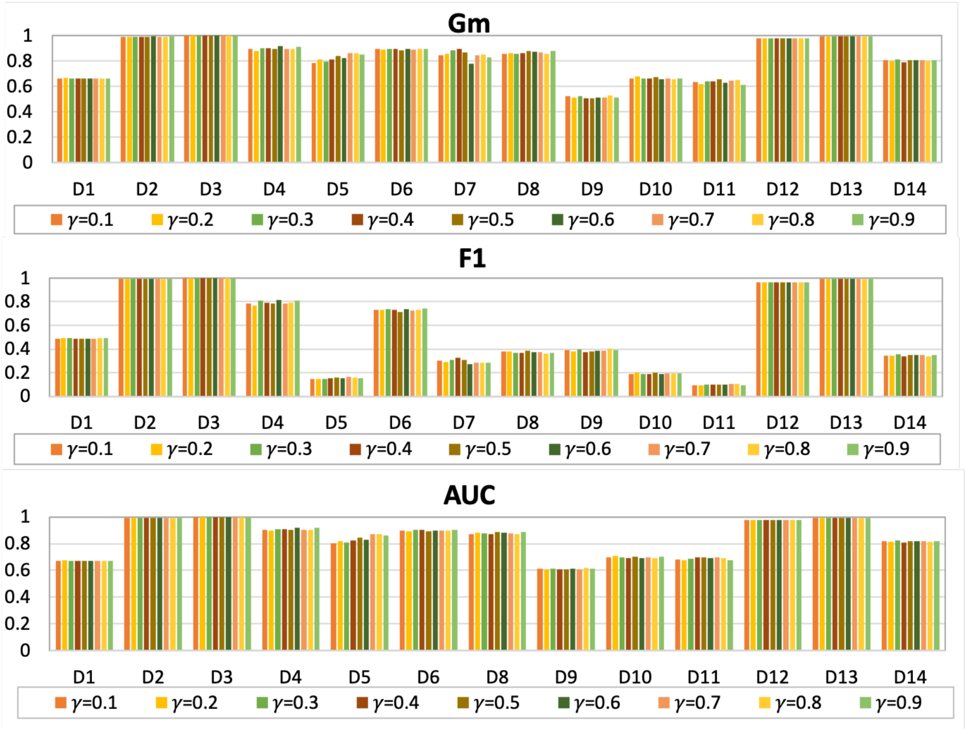

Figure 5. *The performance of POM referenced with γ.*

Every negative sample is assigned a membership value through (1). Parameter ϵ in (2) determines which negative sample distributed in the overlapping area. If parameter γ is fixed as 0.9 in (5) to ensure that a high rate of information is extracted from negative class, the effectiveness of ϵ can be observed and evaluated on the performance of POM. The experimental results of POM are presented in Figure 4, where the effect of parameter ϵ is investigated varying from 0.1 to 0.9 with an interval of 0.1 on two imbalanced datasets (D1 and D6). From (2), it is noteworthy that while the size of N^o may decrease as ϵ increases, it is observed that POM achieves peak values across three metrics simultaneously when ϵ is set to 0.3. This indicates that only positive samples demonstrating significant representation play a crucial role in effectively learning from imbalanced data. Based on this observation, ϵ is set as 0.3 for all subsequent experiments.

Experiments related to the parameter γ are conducted on the 14 datasets. As shown in Figure 5, the performance of POM on all imbalanced datasets is basically stable as γ increases from 0.1 to 0.9 with an interval of 0.1. Especially for the metrics of F_1 and AUC, the gap between the maximum value and the minimum value on each dataset is no more than 0.0521 and 0.0666, respectively. Furthermore, the fluctuations of G_m score for different γ on each dataset are within a range of 0.0414, except for D7 and D5, where the differences between the peak and trough values are 0.1171 and 0.0801, respectively. Overall, the performance of POM is not sensitive to the parameter γ, which validates that an appropriate ϵ contributes to the effectiveness of the information of negative overlapping samples. Therefore, γ is set as 0.9 to reduce information loss.

3.3. Quantitative Comparison

Figure 6 reports the performance of G_m on 14 datasets for all imbalanced methods. The proposed method POM achieves a remarkable peak value 12.9025, surpassing ensemble imbalanced method, BRFC, with a gap of 2.0923. Noted that the ensemble approaches are generally considered superior to other approaches. This claim is supported by the results shown in the table, where ensemble methods (RBC, BRFC, EEC, BBC) outperform other methods, particularly the two advanced sampling methods (OPF-US and REMDD), with the exception of POM. Besides, with the exception of POM, ensemble methods show strong competitiveness compared with other sampling methods, even surpass advanced methods (OPF-US, REMDD) no less than 3.6607.

Noticing that the metric G_m aims to maximize the accuracy for each class while maintaining balance between these accuracies, it can be inferred from the excellent performance of POM that it effectively handles both positive and negative samples, resulting in remarkable outcomes. This observation further supports the feasibility of extracting information from negative overlapping samples and transforming it for oversampling purposes.

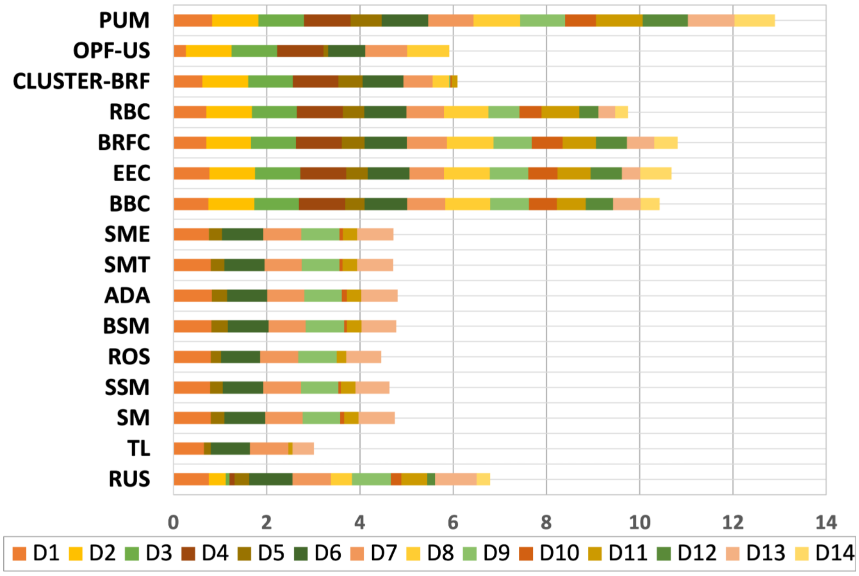

Figure 6. *A bar chart referring to average test G_m compares 15 popular imbalance approaches and POM on 14 datasets.*

Figure 7 illustrates the testing F_1 scores for 15 prevalent imbalanced methods and POM across 14 imbalanced datasets using a line graph. The proposed method POM achieves the best performance on all datasets, except D3. However, on D3 the performance difference between POM and the best method (OPF-US) are imperceptible (0.0037). Besides, some imbalanced methods becomes competitive occasionally. For examples, the advanced method (OPF-US) achieves good performance on datasets (D2, D3, and D4), where the gap of OPF-US to the best performance is no more than 0.0268. However, OPF-US performs worst on other datasets, especially on datasets with $IR \geq 28.1$. Combined sampling method (SME) is competitive among all kinds of traditional sampling methods, but the gap of SME to POM is large on all datasets (no less than 0.1262). Overall, POM achieves good performance to recognize positive samples under the metric of F_1.

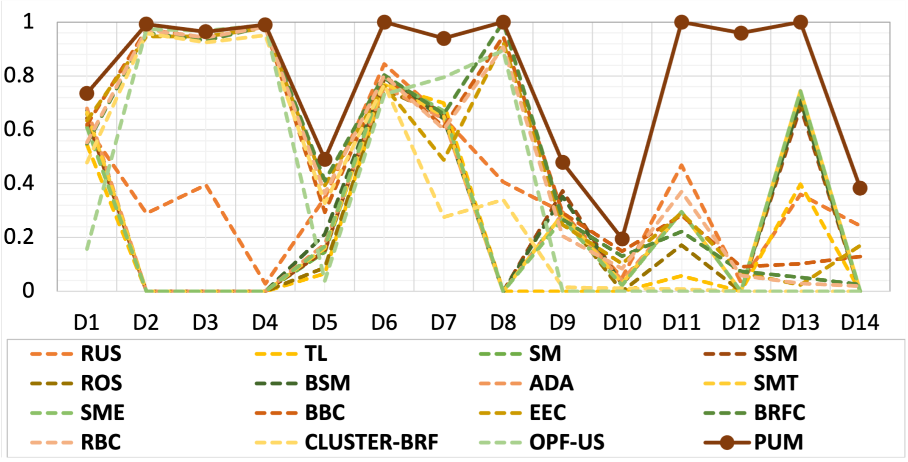

Figure 7. *Broken lines referring to average test F_1 compares 15 popular imbalanced approaches and POM on 14 datasets.*

Figure 8 shows a radar chart showcasing the test AUC of 15 popular imbalanced approaches and POM on 14 imbalanced datasets. The radar chart consists of 14 equi-angular spokes corresponding 14 imbalanced datasets and the performance of each imbalanced method is shown with a 14-side shape. Notably, the 14-side shape of POM (highlighted by a solid brown line) completely covers the other 14-side shapes with scores no less than 0.9956 on 9 datasets, which demonstrates the strong discriminatory power of POM between positive and negative instances. Furthermore, there is a substantial performance gap between POM and other methods on certain datasets, such as 0.1006 on D1, 0.1472 on D9, and 0.2984 on D12. Overall, POM gets an excellent trade-off between TPR and FPR on all datasets.

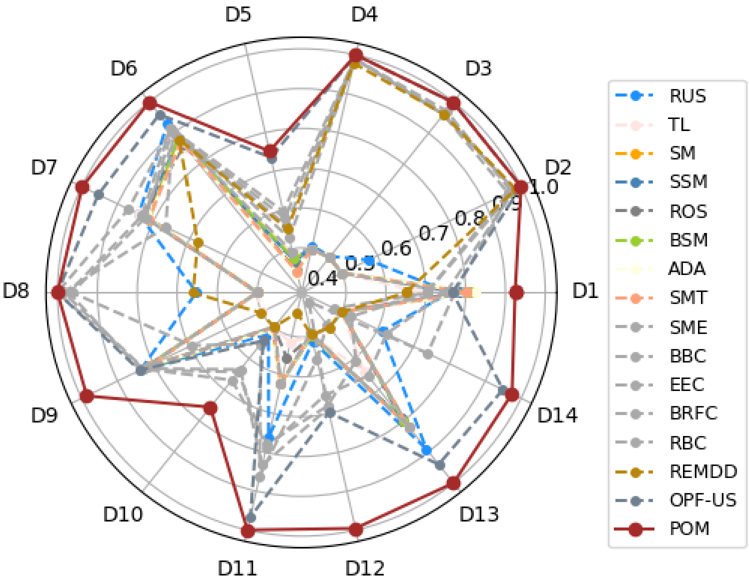

Figure 8. *A radar chart referring to average test AUC compares 15 popular imbalanced approaches and POM on 14 datasets.*

3.4. Statistical Analysis

Boxplots (shown in Figure 9) statistically reflects the learning performance of every method and facilitates a comparison among different approaches. Every box consists of five nodes: minimum, first quartile ($Q1$), median, third quartile ($Q3$) and maximum, where median is a main value to show the entire performance and $IQR = Q3 - Q1$ is an important quantity expressing the degree of dispersion. Firstly, POM outperforms the other methods referring to the median. For examples, compared with the second best methods, POM gets 0.9788 higher 0.1548 than BRFC for G_m, POM achieves 0.9646 higher 0.3115 than RUS for F_1, POM is 0.9987 higher 0.0448 than OPF-US for AUC. Besides, the small IQR shows the good stability of POM. Specifically, the IQR of POM achieves the best value 0.01975 on AUC, and is less 0.1456 than that of the second best method TL. This suggests that rebalanced dataset through POM provides sufficient information and captures meaningful patterns for classifiers.

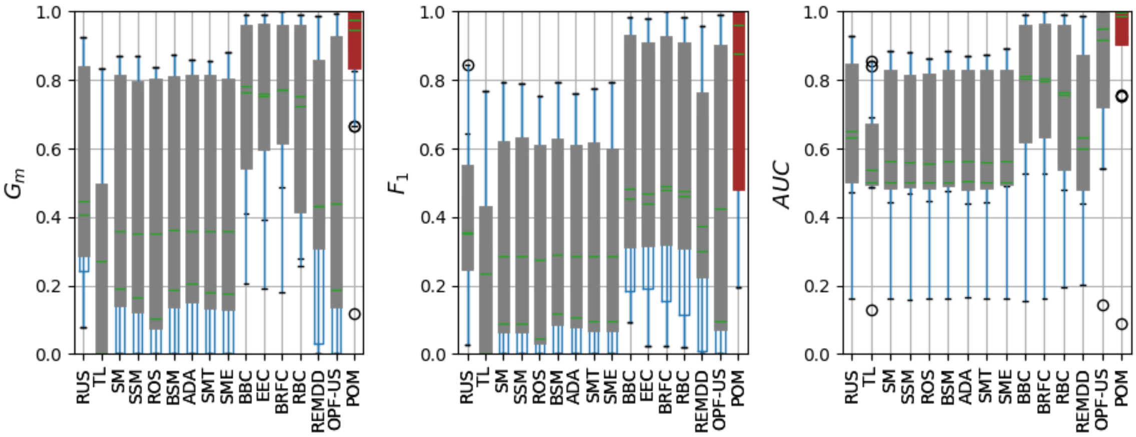

Figure 9. *Box-plots of the three metrics for the 15 popular imbalanced methods and POM on all datasets.*

Table 4. *Nemenyi test for 16 imbalanced methods based on the performance of G_m on all datasets.*

	RUS	TL	SM	SSM	ROS	BSM	ADA	SMT	SME	BBC	EEC	BRFC	RBC	REMDD	OPF-US	POM
RUS	—	*0.0316*	0.8407	0.2924	0.3688	0.7662	0.9000	0.6669	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.8034	0.0876
TL		—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.8655	*0.0022*	0.0057	*0.0010*	*0.0277*	0.9000	0.9000	*0.0010*
SM			—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.3828	0.5427	0.1567	0.8159	0.9000	0.9000	*0.0010*
SSM				—	0.9000	0.9000	0.9000	0.9000	0.9000	*0.0462*	0.0927	*0.0112*	0.2695	0.9000	0.9000	*0.0010*
ROS					—	0.9000	0.9000	0.9000	0.9000	0.0666	0.1274	*0.0172*	0.3414	0.9000	0.9000	*0.0010*
BSM						—	0.9000	0.9000	0.9000	0.3042	0.4666	0.1160	0.7414	0.9000	0.9000	*0.0010*
ADA							—	0.9000	0.9000	0.4666	0.6172	0.2078	0.8904	0.9000	0.9000	*0.0010*
SMT								—	0.9000	0.2174	0.3550	0.0749	0.6420	0.9000	0.9000	*0.0010*
SME									—	0.5675	0.7165	0.2924	0.9000	0.9000	0.9000	*0.0010*
BBC										—	0.9000	0.9000	0.9000	0.5179	0.3414	0.4392
EEC											—	0.9000	0.9000	0.6669	0.5054	0.2808
BRFC												—	0.9000	0.2475	0.1342	0.7041
RBC													—	0.9000	0.7786	0.0982
REMDD														—	0.9000	*0.0010*
OPF-US															—	*0.0010*
POM																—

Table 5. *Nemenyi test for 16 imbalanced methods based on the performance of F_1 on all datasets.*

	RUS	TL	SM	SSM	ROS	BSM	ADA	SMT	SME	BBC	EEC	BRFC	RBC	REMDD	OPF-US	POM
RUS	—	0.1274	0.9000	0.9000	0.3969	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.7165	0.9000	0.0627
TL		—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	*0.0212*	0.2475	*0.0259*	0.1414	0.9000	0.9000	*0.0010*
SM			—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.8407	0.9000	0.8779	0.9000	0.9000	0.9000	*0.0010*
SSM				—	0.9000	0.9000	0.9000	0.9000	0.9000	0.5799	0.9000	0.6172	0.9000	0.9000	0.9000	*0.0010*
ROS					—	0.9000	0.9000	0.9000	0.9000	0.1038	0.5799	0.1223	0.4252	0.9000	0.9000	*0.0010*
BSM						—	0.9000	0.9000	0.9000	0.8407	0.9000	0.8779	0.9000	0.9000	0.9000	*0.0010*
ADA							—	0.9000	0.9000	0.7414	0.9000	0.7786	0.9000	0.9000	0.9000	*0.0010*
SMT								—	0.9000	0.6669	0.9000	0.7041	0.9000	0.9000	0.9000	*0.0010*
SME									—	0.8159	0.9000	0.8531	0.9000	0.9000	0.9000	*0.0010*
BBC										—	0.9000	0.9000	0.9000	0.3282	0.6296	0.2808
EEC											—	0.9000	0.9000	0.8904	0.9000	*0.0259*
BRFC												—	0.9000	0.3688	0.6669	0.2475
RBC													—	0.7414	0.9000	0.0555
REMDD														—	0.9000	*0.0010*
OPF-US															—	*0.0010*
POM																—

Table 6. *Nemenyi test for 16 imbalanced methods based on the performance of AUC on all datasets.*

	RUS	TL	SM	SSM	ROS	BSM	ADA	SMT	SME	BBC	EEC	BRFC	RBC	REMDD	OPF-US	POM
*R*US	—	0.2174	0.7910	0.6420	0.6544	0.9000	0.7786	0.5303	0.9000	0.9000	0.9000	0.9000	0.9000	0.7786	0.3550	*0.0383*
*T*L		—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	*0.0226*	*0.0491*	*0.0172*	0.4392	0.9000	*0.0010*	*0.0010*
*S*M			—	0.9000	0.9000	0.9000	0.9000	0.9000	0.9000	0.2808	0.4392	0.2374	0.9000	0.9000	*0.0010*	*0.0010*
*S*SM				—	0.9000	0.9000	0.9000	0.9000	0.9000	0.1647	0.2808	0.1342	0.8655	0.9000	*0.0010*	*0.0010*
*R*OS					—	0.9000	0.9000	0.9000	0.9000	0.1729	0.2924	0.1414	0.8779	0.9000	*0.0010*	*0.0010*
*B*SM						—	0.9000	0.9000	0.9000	0.5179	0.6669	0.4666	0.9000	0.9000	*0.0016*	*0.0010*
*A*DA							—	0.9000	0.9000	0.2695	0.4252	0.2273	0.9000	0.9000	*0.0010*	*0.0010*
*S*MT								—	0.9000	0.1038	0.1893	0.0838	0.7538	0.9000	*0.0010*	*0.0010*
*S*ME									—	0.4252	0.5799	0.3688	0.9000	0.9000	*0.0010*	*0.0010*
*B*BC										—	0.9000	0.9000	0.9000	0.2695	0.8655	0.3042
*E*EC											—	0.9000	0.9000	0.4252	0.7165	0.1806
*B*RFC												—	0.9000	0.2273	0.9000	0.3550
*R*BC													—	0.9000	0.1647	*0.0112*
*R*EMDD														—	*0.0010*	*0.0010*
*O*PF-US															—	0.9000
*P*OM																—

To investigate the significant differences in performance among various methods, a Nemenyi test [34] is conducted on 16 imbalanced methods. The p-values referring to G_m, F_1, and AUC are 2.73×10^{-9}, 1.16×10^{-4}, and 2.30×10^{-12}, respectively. Since p-values are far less than 0.05, the performance of 16 imbalanced methods is significantly different. Specifically, the performance of G_m is reported in Table 4, it is meaningless to research on $\mathcal{A}$-vs-$\mathcal{A}$, so the diagonal positions in the table is ”—”, where $\mathcal{A}$ traverses through 16 imbalanced methods. Besides, p-values of between $\mathcal{A}_i$-vs-$\mathcal{A}_j$ and $\mathcal{A}_j$-vs-$\mathcal{A}_i$ is no difference, where $\mathcal{A}_i$ and $\mathcal{A}_j$ are two different methods traveling through 16 imbalanced methods, respectively. Therefore, Table 4 is filled with an upper triangle, where all p-values less than 0.05 are marked bold. Firstly, for POM-vs-$\mathcal{A}$, there are 10 p-values (all of them are 0.0010) smaller than 0.05 among 15 p-values. Combined with the experimental results of Section 3.4, POM significantly outperforms 10 imbalanced methods, except for TL and 4 ensemble methods. Besides, there are 7 p-values totally less than 0.05 for the rest 15 imbalanced methods without considering the p-values involved in POM. For the most competitive ensemble method, BRFC, only three p-values of BRFC-vs-$\mathcal{A}$ (BRFC-vs-TL, BRFC-vs-SSM, BRFC-vs-SSM) are no more than 0.05 (0.001, 0.0112, 0.0172 respectively). Overall, the proposed method POM achieves good performance referring G_m, where POM is significantly better than most popular methods.

Similarly, the Nemenyi tests about performance F_1 and AUC are reported in Table 5 and Table 6, respectively. There are 11 p-values of POM-vs-$\mathcal{A}$ no more than 0.0259 for 15 imbalanced methods, except for three ensemble methods and RUS, where p-values of POM-vs-RUS (0.0627) and POM-vs-RBC (0.0555) is close to 0.05. Without considering POM, there are two p-values of $\mathcal{A}_i$-vs-$\mathcal{A}_j$ are smaller than 0.05. Therefore, for the evaluation criterion F_1, POM is significantly better than 11 imbalanced learning methods, but there is no significance among 15 methods basically.

The results in Table 6 shows that the performance of POM and other method is similar to that of in Tables 4-5. However, there is significance between the advanced method OPF-US and 9 imbalanced methods. Besides, there is no significance between POM and OPF-US. Therefore, POM and OPF-US show excellent trade-off between TPR and FPR and significantly outperform than other classical imbalanced methods.

3.5. The Properties of POM

Table 7. *Average running time(s).*

	D1	D5	D6	D7
*R*US	0.001	0.0009	0.0008	0.0009
*T*L	0.01	0.0026	0.0017	0.002
*S*M	0.002	0.0015	0.0013	0.0013
*S*SM	0.0212	0.0128	0.0045	0.0042
*R*OS	0.0008	0.0007	0.0005	0.0005
*B*SM	0.0058	0.0029	0.0022	0.0023
*A*DA	0.0048	0.0026	0.0025	0.0023
*S*MT	0.023	0.0051	0.0037	0.004
*S*ME	0.0257	0.0056	0.0047	0.0049
*B*BC	0.0512	0.0317	0.0303	0.031
*E*EC	0.8052	0.6749	0.6965	0.6901
*B*RFC	0.2451	0.2367	0.2339	0.2368
*R*BC	0.1436	0.1288	0.1243	0.1273
*R*EMDD	0.8851	0.7599	0.744	0.7484
*O*PF-US	27.1199	20.8531	1.8012	4.7754
*P*OM	0.6986	0.6622	0.3054	0.4544

To measure the efficiency of the proposed POM, the running time of 15 imbalanced methods is measured on four datasets by running the codes on Python 3.9.4 (the computer parameters: CPU i7-7700, RAM 32.0 GB) shown in Table 7. Experimental results show that the cost of ensemble methods (such as BBC, EEC, BRFC and RBS) is higher than that of oversampling (undersampling) methods, more than 10 times, on every dataset. For example, on D1, the running time of oversampling (undersampling) methods is no more than 0.001s but the running time of ensemble methods is no less than 0.1436s. The cost of advanced methods (such as REMDD and OPF-US) are more expensive than that of classical methods. The cost of the proposed method POM is lower than that of popular methods and some of ensemble methods (such as EEC). However, POM takes the promoting performance compared with other methods on all datasets. The reason for the high efficiency of POM is that no complex parameters need to be stored and calculated during oversampling. Besides, there is no complex integration algorithm involved in training process.

Table 8. *The quantity of N^o of membership function in training process.*

	$\mathbf{\|S\|}$	$\mathbf{\|P\|}$	$\mathbf{\|N\|}$	$\mathbf{\|N^o\|}$	$\frac{\mathbf{\|N^o\|}}{\mathbf{\|P\|}}$	$\frac{\mathbf{\|N^o\|}}{\mathbf{\|N\|}}$
D1	677	174	503	153	0.88	0.30
D2	286	73	214	3	0.04	0.01
D3	1600	400	1200	32	0.08	0.03
D4	1600	400	1200	21	0.05	0.02
D5	598	142	456	188	1.32	0.41
D6	171	23	148	4	0.17	0.03
D7	354	30	325	13	0.43	0.04
D8	286	16	270	3	0.19	0.01
D9	1187	41	1146	78	1.90	0.07
D10	1279	42	1237	160	3.81	0.13
D11	1168	22	1146	20	0.91	0.02
D12	1186	20	1166	41	2.05	0.04
D13	1188	20	1168	42	2.10	0.04
D14	1182	14	1168	40	0.35	0.03

To take an insight into how many negative overlapping samples there are in each dataset, all negative overlapping samples are collected as $N^o = \bigcup N_i^o$ according to (2). List values of $|N^o|$ and related variables ($|S|$, $|P|$, $|N|$, $\frac{|N^o|}{|P|}$, $\frac{|N^o|}{|N|}$) during oversampling as shown in Table 8. Firstly, a remarkable phenomenon is the number of N^o ranging from 3 to 188, which is not close to the size of S, P, or N. Besides, the ratio between $|N^o|$ and $|N|$ is no more than 0.07 on most datasets except for D1, D5, and D10. There are two inspirations: i). Only a few negative samples in overlapping regions play a supporting role in imbalanced learning; ii). The dataset with the higher or lower ratio between $|N^o|$ and $|N|$ has no clear relation to the performance of POM. Finally, the ratio between $|N^o|$ and $|P|$ is relatively large on most datasets (more than 0.17) except for D2, D3, and D4 (smaller than 0.08). It is easy to notice that if the ratio between $|N^o|$ and $|N|$ is less than 0.08 on datasets, such as D2, the performance of POM on it would be close to compared methods, and vice versa. Interestingly, the performance of POM is not directly proportional to the size of N^o or the ratio between $|N^o|$ and $|N|$, but closely related to the ratio between $|N^o|$ and $|P|$. This indicates that POM can better capture and represent the characteristics and patterns of both positive and negative classes.

Notice that the inequality $|P|+|N^o| < |N|$ holds true on all datasets from Table 8. This illustrates that every dataset is not balanced by the transformed samples and SMOTE also works during the rebalance procedure. However, by comparing the performance of POM and SM (where only SMOTE is used to rebalance) based on results shown in Figures 6-9, POM consistently outperforms SM on all datasets. This indicates that the newly generated samples P^{new} transformed from N^o exactly play a crucial role in providing essential support points for strengthening the representation of the distribution of the original positive class P. Subsequently, SMOTE strengthens the connection between the support points and P. There is a large difference between the distribution of the data rebalanced by POM and the data rebalanced by SMOTE only.

To verify the generalization of POM, the results of the learning model about training and testing process on 7 datasets are shown in Figure 10. First of all, the datasets leveraged in this experiment are diverse, such as IR ranging from 2.9 to 85.88, data size varying from 443 to 2000. Besides, the performance of four metrics is considered, where the solid line and the dotted line of the same color represent the training and testing results of one metric, respectively.

It is evident that the scores of training results and testing results exhibit a close resemblance (the gaps between them smaller than 0.0459) across various metrics for most datasets, except for D14. The reason for this phenomenon on D14 can be attributed to its high imbalance ratio (IR) of 85.88 and an extremely small number of positive samples (17). During the cross-validation process, only a limited number of 3 to 5 positive samples are randomly assigned to the testing set, where the characteristics of these small samples can significantly influence the results. Consequently, there is a notable disparity between the training and testing results, with the training process yielding a G_m score that is 0.2544 higher than that of the testing process.

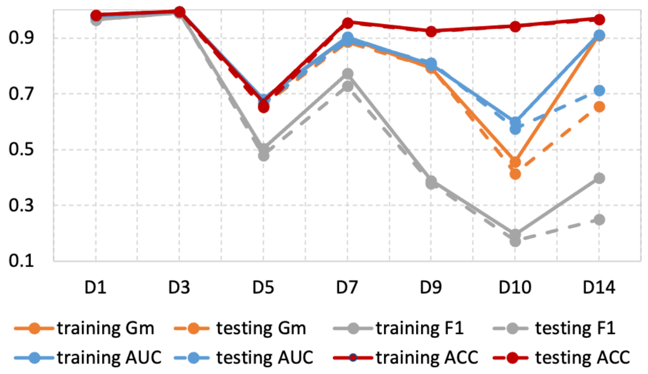

Figure 10. *The training and testing performance of classifiers learned from dataset balanced through POM.*

4. Conclusions

In this paper, a new pseudo oversampling method POM is proposed, which strengthens feature expression of positive class through transferring features of negative overlapping samples that exhibit similar distribution with positive samples.

During the identification of negative overlapping samples, a novel membership function has been introduced to quantitatively and explicitly evaluate the degree to which a negative sample overlaps with positive samples. Additionally, a transformation matrix has been devised to distribute the negative samples over the negatively overlapping samples. Subsequently, these samples are adjusted by the center of the positive class to create new positive samples. The suite of quantitative experiments shows the outperformance of POM compared with popular imbalanced learning methods. In future endeavors, deep convolutional networks are considered to construct a transfer matrix and acquire comprehensive and underlying distribution information.

Funding

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0100300, in part by the National Natural Science Foundation of China under Grant 62176040, 62172073, and 62076182, in part by the China Association for Higher Education under Grant 23BR0203, in part by the Liaoning Provincial Leading Group Office of Educational Science Planning under Grant JG22DB073.

Acknowledgments

The authors would like to thank the Associate Editor and the reviewers for their insightful comments and suggestions.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Yuan, X., Xie, L., Abouelenien, M. A. Regularized Ensemble Framework of Deep Learning for Cancer Detection from Multi-Class, Imbalanced Training Data. Pattern Recognition. 2018, 77, 160-172. https://doi.org/10.1016/j.patcog.2017.12.017

[2] Serguieva, A., Ishibuchi, H., Yager, R. R., Alade, V. P. Guest Editorial Special Issue on Fuzzy Techniques in Financial Modeling and Simulation. IEEE Transactions on Fuzzy Systems. 2017, 25(2), 245-248. https://doi.org/10.1109/TFUZZ.2017.2682542

[3] Chen, R., Guo, S. K., Wang, X. Z., Zhang, T. L. Fusion of Multi-RSMOTE with Fuzzy Integral to Classify Bug Reports with an Imbalanced Distribution. IEEE Transactions on Fuzzy Systems. 2019, 27(12), 2406-2420. https://doi.org/10.1109/TFUZZ.2019.2899809

[4] Jiang, Z., Zhao, L., Lu, Y., Zhan, Y., Mao, Q. A Semi-Supervised Resampling Method for Class-Imbalanced Learning. Expert Systems with Applications. 2023, 221, 119733. https://doi.org/10.1016/j.eswa.2023.119733

[5] Vuttipittayamongkol, P., Elyan, E., Petrovski, A. On the Class Overlap Problem in Imbalanced Data Classification. Knowledge-Based Systems. 2021, 212, 106631. https://doi.org/10.1016/j.knosys.2020.106631

[6] Soltanzadeh, P., Feizi-Derakhshi, M. R., Hashemzadeh, M. Addressing the Class-Imbalance and Class-Overlap Problems by a Metaheuristic-Based Under-Sampling Approach. Pattern Recognition. 2023, 143, 109721. https://doi.org/10.1016/j.patcog.2023.109721

[7] Ren, J., Wang, Y., Cheung, Y. M., Gao, X. Z., Guo, X. Grouping-Based Oversampling in Kernel Space for Imbalanced Data Classification. Pattern Recognition. 2023, 133, 108992. https://doi.org/10.1016/j.patcog.2022.108992

[8] Vuttipittayamongkol, P., Elyan, E. Neighbourhood-Based Undersampling Approach for Handling Imbalanced and Overlapped Data. Information Sciences. 2020, 509, 47-70. https://doi.org/10.1016/j.ins.2019.08.062

[9] Bunkhumpornpat, C., Sinapiromsaran, K. DBMUTE: Density-Based Majority Under-Sampling Technique. Knowledge and Information Systems. 2017, 50, 827-850. https://doi.org/10.1007/s10115-016-0957-5

[10] Vuttipittayamongkol, P., Elyan, E., Petrovski, A., Jayne, C. Overlap-Based Undersampling for Improving Imbalanced Data Classification. In Intelligent Data Engineering and Automated Learning-IDEAL 2018: 19th International Conference, Madrid, Spain, November 21-23, 2018, Proceedings, Part I 19 (pp. 689-697). Springer International Publishing. https://doi.org/10.1007/978-3-030-03493-1_72

[11] Dai, Q., Liu, J. W., Shi, Y. H. Class-Overlap Undersampling Based on Schur Decomposition for Class-Imbalance Problems. Expert Systems with Applications. 2023, 221, 119735. https://doi.org/10.1016/j.eswa.2023.119735

[12] Lango, M., Stefanowski, J. What Makes Multi-Class Imbalanced Problems Difficult? An Experimental Study. Expert Systems with Applications. 2022, 199, 116962. https://doi.org/10.1016/j.eswa.2022.116962

[13] Li, Z., Xie, H., Cheng, G., Li, Q. Word-Level Emotion Distribution with Two Schemas for Short Text Emotion Classification. Knowledge-Based Systems. 2021, 227, 107163. https://doi.org/10.1016/j.knosys.2021.107163

[14] Yu, H., Sun, C., Yang, X., Zheng, S., Zou, H. Fuzzy Support Vector Machine with Relative Density Information for Classifying Imbalanced Data. IEEE Transactions on Fuzzy systems. 2019, 27(12), 2353-2367. https://doi.org/10.1109/TFUZZ.2019.2898371

[15] Tao, X., Zheng, Y., Chen, W., Zhang, X., Qi, L., Fan, Z., Huang, S. SVDD-Based Weighted Oversampling Technique for Imbalanced and Overlapped Dataset Learning. Information Sciences. 2022, 588, 13-51. https://doi.org/10.1016/j.ins.2021.12.066

[16] Dai, Q., Liu, J. W., Liu, Y. Multi-Granularity Relabeled Under-Sampling Algorithm for Imbalanced Data. Applied Soft Computing. 2022, 124, 109083. https://doi.org/10.1016/j.asoc.2022.109083

[17] Shi, H., Zhang, Y., Chen, Y., Ji, S., Dong, Y. Resampling Algorithms Based on Sample Concatenation for Imbalance Learning. Knowledge-Based Systems. 2022, 245, 108592. https://doi.org/10.1016/j.knosys.2022.108592

[18] Bui, Q. T., Vo, B., Snasel, V., Pedrycz, W., Hong, T. P., Nguyen, N. T., Chen, M. Y. SFCM: A Fuzzy Clustering Algorithm of Extracting the Shape Information of Data. IEEE Transactions on Fuzzy Systems. 2020. 29(1), 75-89. https://doi.org/10.1109/TFUZZ.2020.3014662

[19] Ünlü, R., Xanthopoulos, P. Estimating the Number of Clusters in a Dataset via Consensus Clustering. Expert Systems with Applications. 2019, 125, 33-39. https://doi.org/10.1016/j.eswa.2019.01.074

[20] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 2002, 16, 321-357. https://doi.org/10.1613/jair.953

[21] Tomek, I. Two Modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics. 1976, SMC-6(11), 769-772, https://doi.org/10.1109/TSMC.1976.4309452

[22] Tang, Y., Zhang, Y. Q., Chawla, N. V., Krasser, S. SVMs Modeling for Highly Imbalanced Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008, 39(1), 281-288. https://doi.org/10.1109/TSMCB.2008.2002909

[23] Han, H., Wang, W. Y., Mao, B. H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In International Conference on Intelligent Computing. 2005, 878-887. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91

[24] He, H., Bai, Y., Garcia, E. A., Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008, 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

[25] Zeng, M., Zou, B., Wei, F., Liu, X., Wang, L. Effective Prediction of Three Common Diseases by Combining SMOTE with Tomek Links Technique for Imbalanced Medical Data. In 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS). 2016, 225-228. https://doi.org/10.1109/ICOACS.2016.7563084

[26] Fitriyani, N. L., Syafrudin, M., Alfian, G., Yang, C. K., Rhee, J., Ulyah, S. M. Chronic Disease Prediction Model Using Integration of DBSCAN, SMOTE-ENN, and Random Forest. In 2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). 2022, 289-294. https://doi.org/10.1109/ICETSIS55481.2022.9888806

[27] Wang, S., Yao, X. Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. In 2009 IEEE Symposium on Computational Intelligence and Data Mining. 2009, 324-331. https://doi.org/0.1109/CIDM.2009.4938667

[28] Liu, X. Y., Wu, J., Zhou, Z. H. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008, 39(2), 539-550. https://doi.org/10.1109/TSMCB.2008.2007853

[29] Asim, Y., Malik, A. K., Raza, B., Shahid, A. R., Qamar, N. Predicting Influential Blogger's by a Novel, Hybrid and Optimized Case Based Reasoning Approach with Balanced Random Forest Using Imbalanced Data. IEEE Access. 2020, 9, 6836-6854. https://doi.org/10.1109/ACCESS.2020.3048610

[30] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2009, 40(1), 185-197. https://doi.org/10.1109/TSMCA.2009.2029559

[31] Niu, K., Zhang, Z., Liu, Y., Li, R. Resampling Ensemble Model Based on Data Distribution for Imbalanced Credit Risk Evaluation in P2P Lending. Information Sciences. 2020, 536, 120-134. https://doi.org/10.1016/j.ins.2020.05.040

[32] Passos, L. A., Jodas, D. S., Ribeiro, L. C., Akio, M., De Souza, A. N., Papa, J. P. Handling Imbalanced Datasets through Optimum-Path Forest. Knowledge-Based Systems. 2022, 242, 108445. https://doi.org/10.1016/j.knosys.2022.108445

[33] Dong, Z., Xu, C., Xu, J., Zou, B., Zeng, J., Tang, Y. Y. Generalization Capacity of Multi-Class SVM Based on Markovian Resampling. Pattern Recognition. 2023, 142, 109720. https://doi.org/10.1016/j.patcog.2023.109720

[34] Friedman, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. The Annals of Mathematical Statistics. 1940, 11(1), 86-92. https://www.jstor.org/stable/2235971