Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

A New Hybrid Under-sampling Approach to Imbalanced Classification Problems

Chun-Yang Peng & You-Jin Park

To cite this article: Chun-Yang Peng & You-Jin Park (2022) A New Hybrid Under-sampling Approach to Imbalanced Classification Problems, Applied Artificial Intelligence, 36:1, 1975393, DOI: 10.1080/08839514.2021.1975393

To link to this article: https://doi.org/10.1080/08839514.2021.1975393

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e1975393 (44 pages)
https://doi.org/10.1080/08839514.2021.1975393

Taylor & Francis
Taylor & Francis Group

A New Hybrid Under-sampling Approach to Imbalanced Classification Problems

Chun-Yang Peng and You-Jin Park

Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan, R.O.C

ABSTRACT

Among many machine learning applications, classification is one of the important tasks. Most classification algorithms have been designed under the assumption that the number of samples for each class is approximately balanced. However, if the conventional classification approaches are applied to a class imbalanced dataset, it is likely to cause misclassification and, as a result, may distort classification performance results. Thus, in this study, we consider imbalanced classification problems and adopt an efficient preprocessing technique to improve the classification performances. In particular, we focus on borderline noise and outlier samples that belong to the majority class since they may influence classification performance. For this, we propose a hybrid resampling method, called BOD-based under-sampling, which is based on density-based spatial clustering of applications with noise (DBSCAN) approach as well as noise and outlier detection methods, that is, borderline noise factor (BNF) and outlierness based on neighborhood (OBN) to divide majority class samples into four distinctive categories, i.e., safe, borderline noise, rare, and outlier. Specifically, we first determine the borderline noise samples in the overlapped region using the BNF method. Secondly, we use the OBN method to detect outlier samples and apply the DBSCAN approach to cluster the samples. Based on the results obtained from the sample identification analysis, we then segregate the safe category samples which are not abnormal samples while keeping the rest of the samples as rare samples. Finally, we remove some of safe samples by using the random under-sampling (RUS) method and verify the effectiveness of the proposed algorithm through the comprehensive experimental analysis with considering several class imbalance datasets.

ARTICLE HISTORY
Received 03 June 2021
Revised 23 August 2021
Accepted 27 August 2021

Introduction

It is very common that the majority samples (i.e., negative samples) dominate over minority samples (i.e., positive samples) in many practical areas such as fault or defect detection in semiconductor manufacturing, fraud detection in the financial sector, medical diagnosis, spam filtering, and so on. When the ratio of

CONTACT You-Jin Park yjpark@mail.ntut.edu.tw Department of Industrial Engineering and Management, College of Management, National Taipei University of Technology, Taipei, Taiwan, R.O.C.

This article has been republished with minor changes. These changes do not impact the academic content of the article.

observations (i.e., samples) in each class is disproportionate, a class imbalance problem arises and imbalanced classification refers to a classification problem when the distribution of classes is biased or skewed. Generally, it is known that the reliability of classification results often decreases as the imbalance ratio increases (Buda, Maki, and Mazurowski 2018; Ebenuwa et al. 2019; He and Garcia 2009; Janakiraman et al. 2014; Liu, Wu, and Zhou 2009). For example, when the class imbalance ratio is tremendously high, classifying all samples as the majority class may not affect the classification accuracy and so it may lead to a wrong conclusion (Bauder and Khoshgoftaar 2018; Liu et al. 2009; Loyola-González et al. 2016; Leevy et al. 2018). Thus, various useful approaches to imbalanced classification problems arising in diverse fields have been developed. He and Garcia (2009) reviewed the nature, technology, and evaluation methods (i.e., indicators) for learning performance of various imbalanced classification problems. They introduced the main opportunities and challenges of learning from the imbalanced data as well as potential research directions (He and Garcia 2009). Guo et al. (2017) provided a comprehensive review of the techniques and applications of imbalanced classification problems including various data preprocessing techniques, classification algorithms, and model evaluation methods which are widely used in chemical, biomedical engineering, financial management, security management, etc. (Guo et al. 2017). However, sometimes it may not be necessary to strike a perfect balance among the classes and most of the approaches developed for the imbalanced classification problems have not considered the intrinsic characteristics of data. So, it is necessary to provide an appropriate ratio of samples across the classes with understanding their own characteristics of data and class distributions in handling imbalanced classification problems (Buda, Maki, and Mazurowski 2018; Leevy et al. 2018). Thus, in this research, we propose a noble hybrid under-sampling method with considering noise and outlier detection methods as well as density-based clustering method to effectively separate majority class samples into four sub-categories and increase the performance of imbalanced classification problems. The rest of this paper is organized as follows: Section 2 presents several data preprocessing techniques for imbalanced classification problems. In section 3, we provide several related works on noise and outlier detection techniques as well as clustering methods for handling class imbalance problems. Section 4 describes in detail the proposed method and section 5 provides the experimental analysis results. Finally, in section 6, we discuss the conclusions and further researches.

Approaches to imbalanced classification problems

Approaches for handling imbalanced classification problems can be basically divided into three categories, i.e., data-level, algorithm-level, and hybrid approaches (Galar et al. 2011). Data-level approaches are to adjust the class distribution by using effective data preprocessing methods, such as resampling, feature selection, etc. The most well-known resampling methods used to

balance a biased (or skewed) distribution of data are under-sampling, over-sampling, and hybrid sampling. Algorithm-level approaches are to modify existing learning algorithms using various techniques such as SVM, neural network, decision tree, etc. (Guo et al. 2017; Krawczyk 2016).

One of the popular under-sampling methods is random under-sampling (RUS), which selects and removes majority class samples randomly to achieve an inter-class balance until the ratio of samples between classes reaches a predetermined level retaining the minority class samples (Gong et al. 2019). Since RUS depends on sample distribution without considering any other information, the operation is quite simple. However, as a disadvantage of the RUS, there is a chance that certain majority class samples including important information can be eliminated because existing information is not fully taken into account (Dubey et al. 2014). The removal of majority class samples through RUS can make the decision boundary between classes harder to learn or may result in a degradation of classification performance (Attenberg and Ertekin 2013). To alleviate the problem caused by class imbalance, Seiffert et al. (2010) proposed a new data sampling algorithm combined with AdaBoost algorithm, called RUSBoost. In particular, through the AdaBoost algorithm, they adjusted weights on samples iteratively and assigned a class to the unlabeled samples according to a weighted vote (Seiffert et al. 2010). In contrast with under-sampling, over-sampling is to generate minority class samples until the ratio of samples between classes reaches a certain level like under-sampling method. Random over-sampling (ROS) is the simplest over-sampling method, which achieves the balance between classes by randomly creating the minority class samples (Fotouhi, Asadi, and Kattan 2019). However, when synthetic samples exist in the majority class region, an overfitting problem could occur since the ROS may not create a clear decision boundary to separate the classes by generating synthetic samples near the original minority class samples (Zhu, Lin, and Liu 2017). To overcome the overfitting problem of ROS, Chawla et al. (2002) proposed a new method called SMOTE (synthetic minority over-sampling technique), which artificially generates minority class samples by interpolating neighboring samples rather than simply replacing samples. In this research, they showed that SMOTE has better classification performance (i.e., ROC) when combined with random under-sampling method rather than plain under-sampling method (Chawla et al. 2002). As extensions of the SMOTE, Han, Wang, and Mao (2005) proposed two novel over-sampling methods called borderline-SMOTE1 and borderline-SMOTE2, which focus on the samples close to the borderline between classes and generate minority class samples near the borderline since they can be misclassified as majority class samples than those far from the borderline. They showed that improved F-measure and true positive rate can be achieved through these methods compared to the SMOTE and ROS (Han, Wang, and Mao 2005). Galar et al. (2011)

reviewed several ensemble techniques such as bagging-, boosting-, and hybrid-ensemble for the binary imbalanced classification problems. In this research, they categorized the ensemble-based methods according to inner ensemble methodologies and the technique types and compared the performances of the combined methods (Galar et al. 2011). Liang et al. (2020) pointed out that some of the over-sampling techniques cannot generate samples near the center of minority class samples while avoiding noises. To overcome this drawback, they proposed a new SMOTE that limits the radius of sample generation, called LR-SMOTE, which consists of three preprocessing procedures, i.e., de-noising, over-sampling, and filtering (Liang et al. 2020). Xie et al. (2019) provided an advanced over-sampling method based on alien *k*-neighbors and random-SMOTE, called AKN-Random-SMOTE. In this research, to classify the minority class samples near the decision boundary more accurately, they used the alien *k*-neighbors to select support vectors and then applied SMOTE to create synthetic samples with considering only selected support vectors which belong to minority class (Xie et al. 2019). Wei et al. (2020) proposed a method that can effectively identify noises belonging to the minority class, called NI-MWMOTE (noise-immunity majority weighted minority oversampling technique). To improve the noise immunity of the conventional MWMOTE (i.e., to identify and eliminate the real noise more effectively), in this research, they considered an adaptive noise processing scheme and aggregative hierarchical clustering (AHC) method to prevent the generated samples from becoming noises and affecting the classification performance (Barua et al. 2012; Wei et al. 2020). Gnip, Vokorokos, and Drotár (2021) proposed a selective over-sampling method (SoA), which is based on the outlier detection method to retain representative original minority class samples and generates synthetic minority class samples by using SMOTE and ADASYN (Gnip, Vokorokos, and Drotár 2021).

Related works

Anomaly detection methods

In general, most real-world data include various types of noises that might affect the performance in learning. Particularly, in imbalanced classification problems, it is known that the decision boundary can make the samples distinguishable more clearly after identifying and eliminating noise samples in overlapped regions (Fotouhi, Asadi, and Kattan 2019; Guzmán-Ponce et al. 2020). Thus, several useful methods have been developed to identify and eliminate the noises in imbalanced classification problems, especially, which are close to the decision boundary. For this, Tomek (1976) proposed a distance-based method to determine whether a pair of different class samples can become a Tomek link or not. Since it is

quite helpful to identify samples from different classes around the decision boundary, some of the majority class samples which are identified as a Tomek link can be removed (Tomek 1976). Devi, Biswas, and Purkayastha (2017) proposed an under-sampling combined with the Tomek link method which focuses on detecting the samples that contribute less to the accurate estimation of class labels, that is, outliers or redundant and noisy samples (Devi, Biswas, and Purkayastha 2017).

Especially, for identifying borderline noises in the overlapped regions of imbalanced classification problems, Yang and Gao (2013) proposed a new evaluation method of noisy samples in overlapped areas, called BNF (borderline noise factor), and improved classification performance by eliminating borderline noises. In this method, when K_s represents the number of nearest neighbors for each sample within the same class in the training dataset S, the BNF value of sample x can be calculated as follows:

$$BNF(x) = \alpha\left(\frac{K_s + \delta}{|kNS(x)| + \delta}\right) + \beta|kND(x)| \tag{1}$$

where $kNS(x)$ and $kND(x)$ indicate the samples belonging to the same and different class of a sample x, and $|kNS(x)|$ and $|kND(x)|$ represent the numbers of $kNS(x)$ and $kND(x)$, respectively. Since there may not exist any mutual nearest neighbor of the same class sample x, i.e., $|kNS(x)| = 0$, an arbitrarily small positive value of δ (usually δ is set to be less than 1) is considered to prevent the denominator of the first term in BNF function from being zero. And, the parameters α and β ($0 \leq \alpha, \beta \leq 1$, and $\alpha + \beta = 1$) represent the weights assigned to each term in the BNF function, respectively. Through the investigation of the average G-mean of several datasets, the optimal value of α is found to be 0.3 (Yang and Gao 2013). For example, Figure 1 illustrates the procedure of finding $kNS(x)$ and $kND(x)$ for a majority class sample x when $k = 5$. The symbols "●" and "○" represent majority and minority class samples in two-dimensional space. Within a certain region of Θ_s, $|kNS(x)|$ (denoted by blue dots) and $|kND(x)|$ (denoted by red circles) for a majority class sample x are 5 and 6, respectively. The mutual nearest neighbors of majority class samples would be different since each majority class sample has different value of Θ_s. As shown in Figure 1 (b) to (f), the mutual nearest neighbors of the majority class sample x are x_1, x_2, x_3, x_6, and x_7. In other words, the other majority class samples (i.e., x_4, x_5, x_8, x_9, and x_{10}) are the nearest neighbors of sample x but not the mutual neighbors. The following section describes the outlier detection methods in detail.

In most data collection processes, due to several causes such as variations in measurement methods, human negligence, or experimental errors, some samples are extremely different from the rest of the samples collected and they are

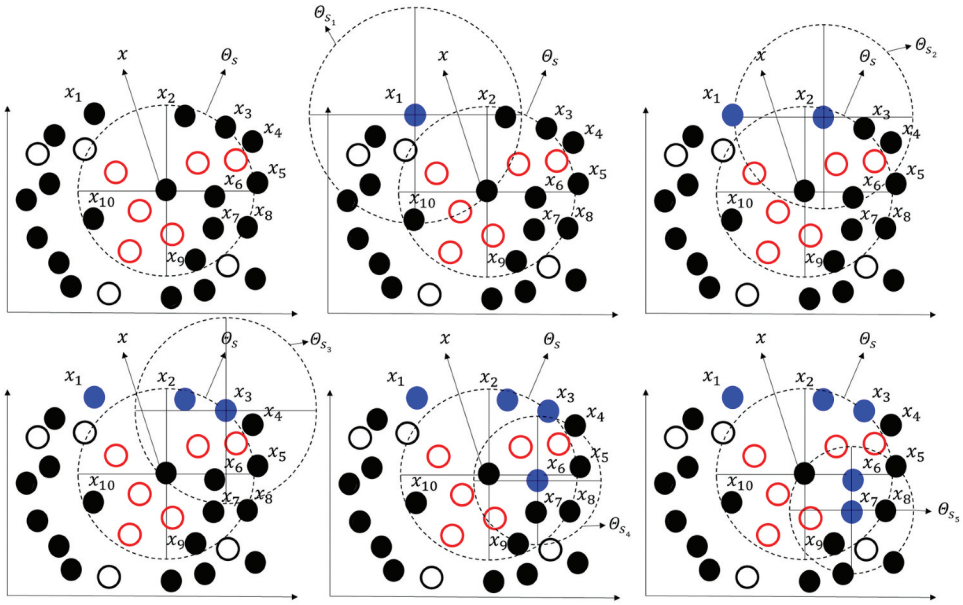

Figure 1. $kNS(x)$ and $kND(x)$ when $k = 5$.

referred to as outliers (Chen, Miao, and Zhang 2010; Li, Lv, and Yi 2016). Similar to noises, the outliers may affect the learning performance of the model and even make it difficult to interpret the analysis results (Shi et al. 2020). So, it is very important to identify and eliminate outliers in data and thus several useful approaches have been developed to handling outliers to improve the learning performance. Yuan, Zhang, and Feng (2018) proposed a neighborhood information entropy-based outlier detection method, called NIEOD, and investigated its measures. In this research, they determined the neighborhood information system through the heterogeneous distance and self-adapting radius and used the neighborhood information entropy and its three in-depth measures to illustrate data with uncertainty measurement. Then, they constructed the neighborhood entropy-based outlier factor (NEOF) (Yuan, Zhang, and Feng 2018). Gupta, Bhattacharjee, and Bishnu (2019) proposed a new neighborhood-based outlier detection and analysis technique that considers the weights of the neighbors of each sample. In this research, given a dataset $D = \{x_1, x_2, \ldots, x_n\}$, the set of r-neighbors is denoted as $N_r(x_i) = \{y_j | x_i, y_j \in D, dist(x_i, y_j) \leq r\}$ for a sample x_i where $dist(x_i, y_j)$ is the Manhattan distance of two samples x_i and y_j $(1 \leq i \leq n, 1 \leq j \leq N_r(x_i) \leq n)$ and "r" is a user-defined radius. And, the weights of neighborhood of the sample x_i and y_j are defined as $W[N_r(x_i)] =$

$\Sigma_{i=1}^{N_r(x_i)} dist(x_i, y_j)$ and $W[N_r(y_j)] = \Sigma_{j=1}^{N_r(y_j)} dist(x_i, y_j)$, respectively. So, the OBN (outlierness based on neighborhood) value for a sample x_i can be calculated as follows:

$$OBN(x_i) = \frac{W[N_r(x_i)]}{\Sigma_{j=1}^{N_r(x_i)} W[N_r(y_j)]} \tag{2}$$

Here, if $OBN(x_i)$ for a sample x_i is greater than the average OBN value of all the samples, the sample x_i is classified as an outlier (Gupta, Bhattacharjee, and Bishnu 2019). Shi et al. (2020) proposed a geodesic-based outlier detection algorithm that considers both the global disconnection score and the local realness which can evaluate the degree of outlier of each sample and connectivity between samples as the detection measure of outliers. In particular, they constructed a global disconnection score to incorporate appropriate distribution of the data and provided the local realness to consider the features of the samples effectively. Then, they determined the local outliers in smaller clusters with higher overall connectivity located in the minority class (Shi et al. 2020). Chen, Wang, and Yang (2021) extended the LOF (local outlier factor) method proposed by Breunig et al. (2002) and proposed a new outlier detection method, called CELOF (local outlier factor based on clustering and data extraction) (Chen, Wang, and Yang 2021). Wang et al. (2020) proposed a new outlier detection method based on the dynamic references nearest neighbors (DRNN) and local neighborhood outlier factor (LNOF), called LDNOD (Wang et al. 2020).

Clustering methods

The main purpose of clustering analysis is to group samples with similar characteristics into the same clusters (Rezaee et al. 2021). Particularly, for resolving imbalance classification problems, several different types of clustering analysis methods have been developed. Yen and Lee (2009) proposed a cluster-based under-sampling (SBC) method to select representative data for training. Since the behavior or characteristic of a cluster depends on the proportion of the majority and minority class samples in the cluster, they considered a ratio of the number of majority class samples to that of minority class samples in the clusters and randomly selected a proper number of majority class samples from each cluster for training (Yen and Lee 2009). Lin et al. (2017) presented two under-sampling strategies based on the K-means clustering and then compared performances of several combinations of the clustering-based under-sampling techniques with different types of classification methods to demonstrate the efficiency of the proposed methods (Lin et al. 2017). To improve the classification performance, Ofek et al. (2017)

focused on the samples that are difficult to identify (i.e., the majority class samples close to minority class region) and proposed a fast clustering-based under-sampling (Fast-CBUS) method which can cluster the minority class samples only (Ofek et al. 2017). As a spatial clustering technique, DBSCAN method has been widely used to cluster samples into three distinctive categories, namely, core, border, and noise samples. Through the DBSCAN, the outlier with a small density are treated as noise samples which are unreachable samples from any core point or that do not belong to any cluster. This method requires two important parameters, i.e., the radius (ε) of neighborhood around a sample and a threshold (*MinPts*) for the number of neighbors which indicates the total weight of a neighborhood for a core sample (Schubert et al. 2017). There have been many applications of the DBSCAN method to various imbalanced classification problems. He et al. (2014) presented a scalable DBSCAN algorithm based on MapReduce called MR-DBSCAN to enhance the performance of imbalanced classification problems as well as resolve the scalability problem in the DBSCAN algorithm (He et al. 2014). However, in the DBSCAN algorithm, it is very important but difficult to choose proper input parameters a priori, that is, the radius ε and *MinPts* even though they have a significant impact on the clustering results. Karami and Johansson (2014) provided an efficient hybrid clustering method called BDE-DBSCAN, which combines the binary differential evolution (BDE) method and DBSCAN algorithm to determine appropriate parameter values of ε and *MinPts* quickly and automatically (Karami and Johansson 2014). Guzmán-Ponce et al. (2020) proposed an under-sampling method called DBMIST-US that combines DBSCAN and minimum spanning tree (MST) algorithm for identifying noisy samples and cleaning borderline samples (i.e., the samples close to the decision boundary) sequentially (Guzmán-Ponce et al. 2020).

Proposed method

In this research, we propose a hybrid resampling method called BOD-based under-sampling to improve the performance of imbalanced classification problems, which combines BNF, OBN, and DBSCAN approaches. To evaluate the proposed method, we adopt a SVM classifier with RBF (radial basis function) kernel since the RBF kernel can deal with the linear non-separable problem (Liu et al. 2015). In this research, the DBSCAN method is used to detect outliers, that is, the samples with low density. However, when there is a significant difference in data distribution, since the DBSCAN method may often misjudge normal samples as outliers, the BNF and OBN methods are also considered simultaneously to overcome this drawback. Then, among the abnormal samples (i.e., borderline noises and outliers) extracted by BNF, OBN, and DBSCAN methods, the outliers only determined by DBSCAN

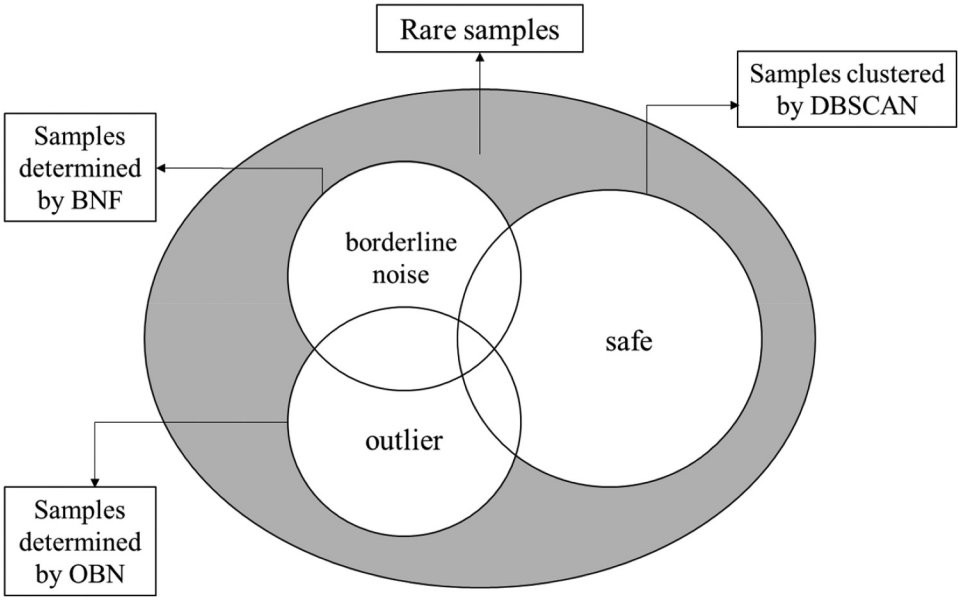

Figure 2. Categories of majority class samples.

method are classified as rare samples while the samples identified as normal are classified as safe samples sequentially. Then, some of the separated safe samples belonging to majority class are eliminated using RUS. Figure 2 and 3 illustrates the categories of samples resulted from the application of the proposed method and the detailed procedures of the proposed method, respectively. In particular, since there is no prescribed rule to set an exact radius value r in the original OBN function and many previous studies have considered kNN (k nearest neighbor)-based outlier detection methods to determine the neighbors of a sample (Angiulli et al. 2012; Angiulli, Basta, and Pizzuti 2006; Angiulli and Fassetti 2009; Angiulli and Pizzuti 2005), we considered a proper k nearest neighbor value in the application of OBN method instead of the radius r to search neighbors of a sample.

The Algorithm 1 and 2 present the pseudo codes for finding overlapped samples and for categorizing majority class samples into four distinct categories, respectively. In algorithms, the notations S_{tr}, S_{ts}, S_{OVR}, S_{BN}, S_O, S_{DBC}, S_a, S_s, x_{tr}, and x_{trn} represent the training dataset, test dataset, set of the samples in overlapped region, set of borderline noises, set of outliers, set of the samples clustered by DBSCAN, set of abnormal samples, set of safe samples, an individual sample in training dataset, and nearest neighbors of x_{tr}, respectively.

Algorithm 1. Pseudo code for finding overlapped samples.

Set $S_{OVR} = \emptyset$

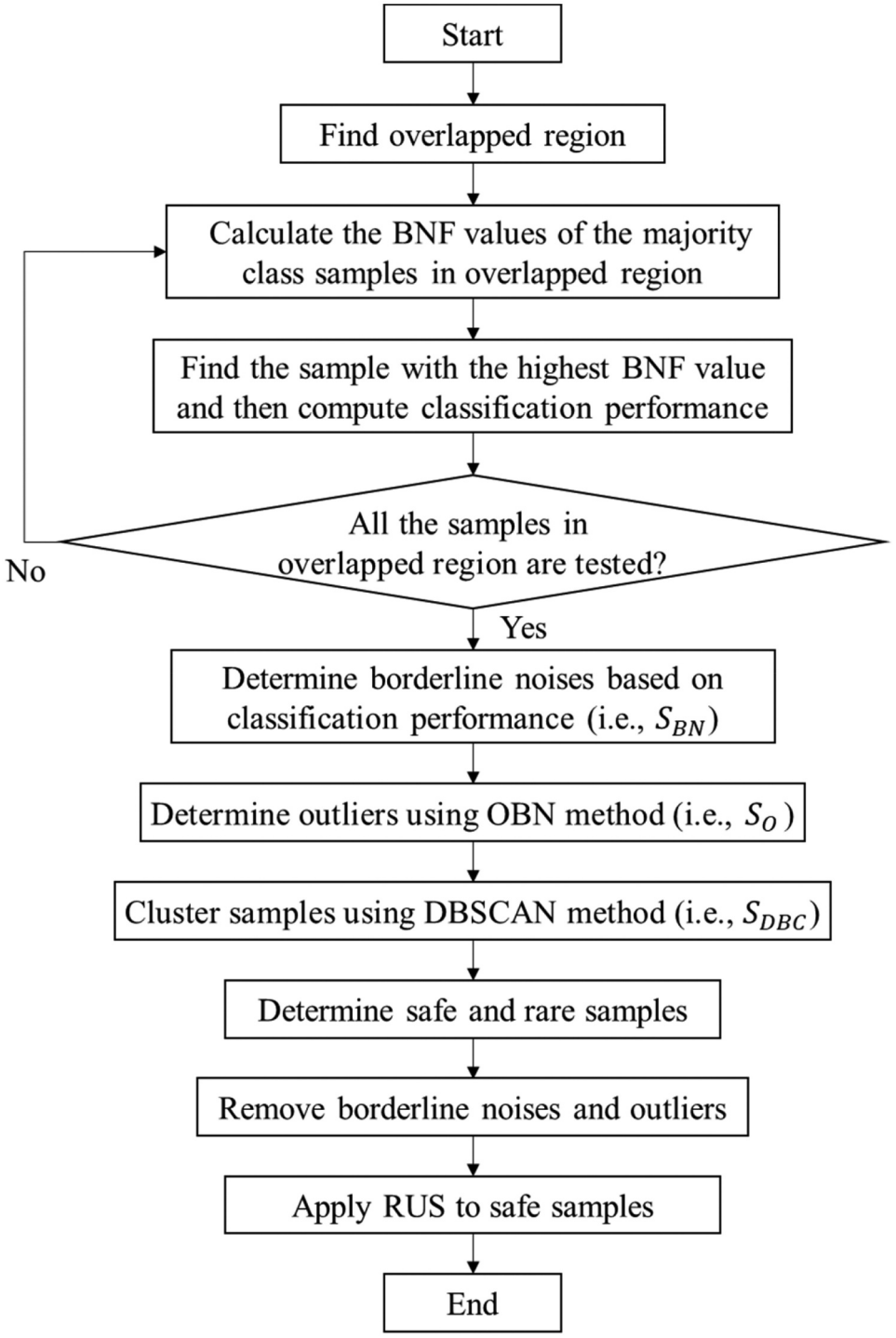

Figure 3. Procedures of the proposed method.

For each sample x_{tr} in the S_{tr}, find x_{trn} based on the Euclidean distance.
If x_{trn} belongs to the different class of the sample x_{tr}
Add x_{trn} to S_{OVR}.
Else
End

Algorithm 2. Pseudo code for categorizing majority class samples.
Set $S_{BN} = \emptyset$
For each sample x_{trn} in S_{OVR}, calculate BNF value.
Repeat
If Classification performance is maximized
Break
Else
Recalculate the remaining x_{trn} in S_{OVR}.
Store the majority class sample with highest BNF value in S_{BN}.
End Repeat
Set $S_O = \emptyset$
For each majority class sample in S_{tr}
Repeat
Calculate $OBN(x_{tr})$.
Calculate $E(OBN(x_{tr}))$.
If $OBN(x_{tr}) > E(OBN(x_{tr}))$
Add x_{tr} to S_O.
Else $OBN(x_{tr}) \leq E(OBN(x_{tr}))$
End Repeat
Set $S_{DBC} = \emptyset$ and $S_s = \emptyset$
Store the samples clustered by DBSCAN in S_{DBC}.
Remove the set of samples in $S_a = S_{DBC} \cap \left(S_{BN} \cup S_O\right)$.
Keep the remaining samples in S_s.
Use RUS to balance the class distribution in S_s.
End

Experimental analysis

In this research, we consider 15 imbalanced datasets selected from KEEL data repository (http://www.keel.es/dataset.php) to evaluate the classification performance of the proposed method. Table 1 contains basic information of the datasets such as the number of attributes, number of samples, percentages of the majority and minority class samples, and imbalance ratio.

In experimental analysis, we use RUS to eliminate some of the safe samples derived from BNF, OBN, and DBSCAN methods and the RBF kernel SVM classifier to assess the proposed method with considering two classification

Table 1. Information of datasets.

Dataset	NoA[1]	NoS[2]	PMa[3]	PMi[4]	IR[5]
glass1	9	214	64.54	35.46	1.82
yeast1	8	1484	71.10	28.90	2.46
haberman	3	306	73.54	26.46	2.78
ecoli1	7	336	77.06	22.94	3.36
segment0	19	2308	85.75	14.25	6.02
glass6	9	214	86.45	13.55	6.38
yeast2vs4	8	514	90.08	9.92	9.08
glass0146vs2	9	205	91.71	8.29	11.06
shuttle-c0-vs-c4	9	1829	93.28	6.72	13.87
yeast1vs7	7	459	93.46	6.54	14.30
glass4	9	214	93.93	6.07	15.47
glass016vs5	9	184	95.11	4.89	19.44
shuttle-c2-vs-c4	9	129	95.35	4.65	20.50
yeast5	8	1484	97.04	2.96	32.73
yeast6	8	1484	97.64	2.36	41.40

[1]Number of attributes, [2] Number of samples, [3] Percentage of the majority class samples, [4] Percentage of the minority class samples, [5] Imbalance ratio

Table 2. AUC of five classification models.

Dataset	Model [1]	Model [2]	Model [3]	Model [4]	Model [5]
glass1	0.6319	0.5787	0.6759	**0.6792**	0.6343
yeast1	0.6692	0.7596	0.7387	0.7428	**0.7684**
haberman	0.5477	0.5647	0.5477	**0.6562**	0.6386
ecoli1	0.7933	0.8702	0.8245	0.8462	**0.8894**
segment0	**0.9924**	0.9899	**0.9924**	**0.9924**	**0.9924**
glass6	0.8333	0.9459	0.8333	0.8333	**0.9869**
yeast2vs4	0.5000	0.9437	0.5000	0.9328	**0.9491**
glass0146vs2	0.5000	0.5034	0.5000	**0.8513**	0.5439
shuttle-c0-vs-c4	0.9792	0.9861	0.9792	**0.9985**	**0.9985**
yeast1vs7	0.5000	0.7093	0.5000	0.6977	**0.7209**
glass4	0.6542	0.8875	0.8333	**0.9688**	0.9500
glass016vs5	0.5000	0.9857	0.9714	**1.0000**	**1.0000**
shuttle-c2-vs-c4	0.7500	0.9583	0.7500	0.7500	**1.0000**
yeast5	0.7205	0.9340	0.7760	0.9449	**0.9736**
yeast6	0.5714	0.8793	0.7126	0.8304	**0.9131**

[1]SVM without any data preprocessing, [2] SVM with RUS, [3] SVM with DBSCAN, [4] SVM with SMOTE, [5] SVM with BOD

Table 3. G-mean of five classification models.

Dataset	Model [1]	Model [2]	Model [3]	Model [4]	Model [5]
glass1	0.5774	0.5528	0.6526	**0.6769**	0.6236
yeast1	0.6106	**0.7592**	0.7102	0.7384	0.7512
haberman	0.3392	0.5636	0.3392	**0.6550**	0.6366
ecoli1	0.7752	0.8702	0.8131	0.8407	**0.8893**
segment0	**0.9924**	0.9899	**0.9924**	**0.9924**	**0.9924**
glass6	0.8165	0.9444	0.8165	0.8165	**0.9864**
yeast2vs4	0.0000	0.9430	0.0000	0.9325	**0.9483**
glass0146vs2	0.0000	0.4350	0.0000	**0.8383**	0.4577
shuttle-c0-vs-c4	0.9789	0.9860	0.9789	**0.9985**	**0.9985**
yeast1vs7	0.0000	0.6470	0.0000	0.6627	**0.6862**
glass4	0.5701	0.8834	0.8165	**0.9683**	0.9487
glass016vs5	0.0000	0.9856	0.9710	**1.0000**	**1.0000**
shuttle-c2-vs-c4	0.7071	0.9574	0.7071	0.7571	**1.0000**
yeast5	0.6655	0.9317	0.7441	0.9338	**0.9732**
yeast6	0.3780	0.8710	0.6535	0.8223	**0.9113**

[1]SVM without any data preprocessing, [2] SVM with RUS, [3] SVM with DBSCAN, [4] SVM with SMOTE, [5] SVM with BOD

performance measures, i.e., AUC (area under the ROC curve) and G-mean. The AUC measures the ability of a classifier to distinguish classes while the G-mean comprehensively measures the accuracy rates of positive samples and negative samples. The AUC and G-mean measures are as follows (Loyola-González et al. 2016; Yang and Gao 2013):

$$AUC = \left(1 + \frac{TP}{FN + TP} - \frac{FP}{TN + FP}\right)/2 \tag{3}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{4}$$

We set the termination condition of the proposed method as IR = 1 of the safe category samples while all the borderline and outlier samples are removed but the entire rare samples are kept in the final training dataset. Tables 2 and 3 include classification performances resulted from applying the SVM classifier combined with RUS, DBSCAN, SMOTE, and the proposed method to 15 imbalanced datasets.

From the experimental results, we can see that the proposed method outperforms the other three traditional resampling approaches, that is, RUS, DBSCAN, and SMOTE, as well as the pure SVM for most of the considered imbalanced datasets with respect to AUC and G-mean. However, it seems that the proposed method may perform poorly for the datasets with a low imbalance ratio (i.e., IR < 9). Specifically, for 'glass1′ (IR = 1.82), 'yeast1′ (IR = 2.46), and "haberman" (IR = 2.78) datasets, the AUC and G-mean of the proposed method seem to be lower than those of the SVM combined with SMOTE, DBSCAN, and RUS sampling methods. We also found that the SVM combined with SMOTE outperforms the proposed method for 'glass0146vs2′ and 'glass4′ datasets. Compared to the under-sampling, since most over-sampling methods (e.g., SMOTE) may not ignore noise and outlier samples belonging to the minority class, they can classify minority class samples more effectively and better classification performance can be obtained. However, in this research, since we aim to eliminate both borderline noise samples which belong to the majority class in the overlapped region and outliers which are far from the majority class sample cluster, it is very difficult to correctly classify certain unlabeled samples having nearest neighbors of the same class but located in the majority class region as the minority class. Thus, we can conclude that the SVM combined with SMOTE slightly outperforms the proposed method because most of the minority class samples in 'glass0146vs2′ and 'glass4′ datasets could be located inside the majority class region as well as a few samples of the same class. However, we can see that the proposed method performs quite well for the datasets with a high imbalance ratio (IR > 9). For example, there are almost 34% and 54% of improvements in AUC

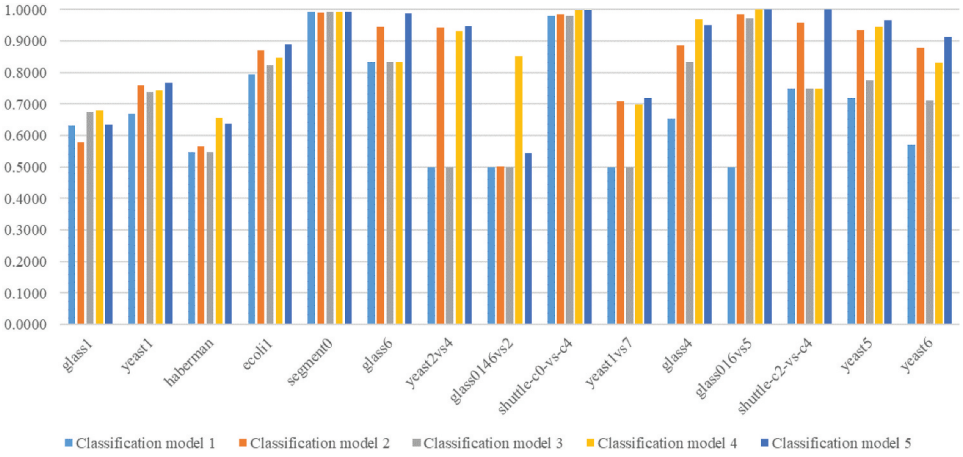

Figure 4. AUCs of 15 imbalance datasets.

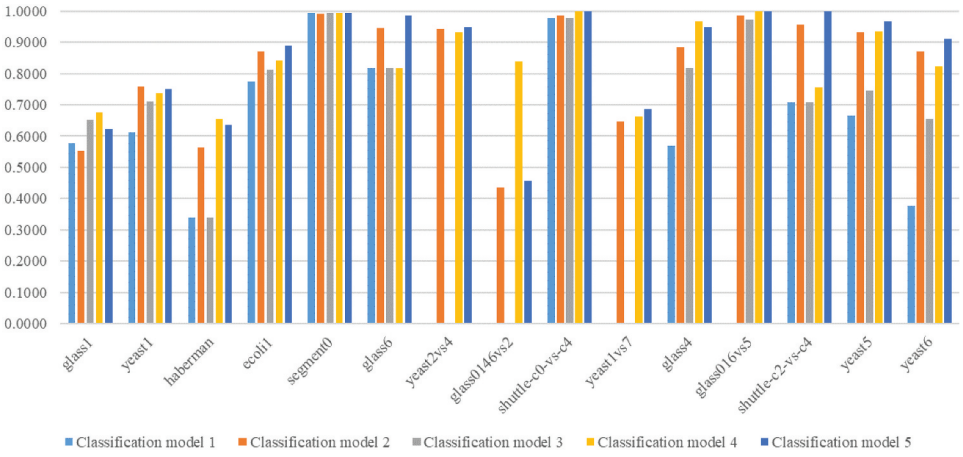

Figure 5. G-means of 15 imbalance datasets.

and G-mean for 'yeast6' dataset. Figure 4 and 5 present a comparison of AUCs and G-means of the five classification models graphically. The x-axis and y-axis in figures represent imbalanced datasets considered and the corresponding classification performances, respectively.

Conclusions and further works

It has been known that class imbalance and abnormal observations such as noises and outliers have a considerably huge influence on the classification performance (Attenberg and Ertekin 2013; Shi et al. 2020; Tomek 1976). Thus, in this research, we considered an imbalanced binary classification problem and applied the DBSCAN method as well as efficient noise and outlier detection methods, i.e., BNF and OBN, to improve classification

performances. The borderline noise samples determined by the BNF method are eliminated until the maximum level of AUC and G-mean values are achieved and then outliers (i.e., abnormal samples with large OBN values) are identified and eliminated. Finally, the safe and rare samples are determined by the DBSCAN method and some of the safe samples are eliminated using the RUS. Through the comprehensive experimental analysis, we showed that the proposed data preprocessing method for imbalanced classification problems can effectively determine borderline noises and outliers in the majority class and, by removing these abnormal samples, better classification performance can be achieved than the classification models combined with simple RUS, DBSCAN, and SMOTE. However, in this research, since we focused on the elimination of only majority class samples while keeping minority class samples that might have certain critical information, there is a limitation to improve the classification performances more significantly. Thus, it is necessary to develop efficient hybrid resampling methods considering both under-sampling and over-sampling. And, since the proposed method involves two critical procedures, that is, (i) calculation procedure of the BNF values for all the borderline noise samples and (ii) removal procedure of the sample with the maximum BNF value iteratively to obtain the best classification performance, a moderately large amount of computation time is required for the complex and large-scale imbalanced classification problems. Thus, it is also necessary to develop an efficient and appropriate stopping criterion to reduce the computation time for the BNF values of majority class samples in overlapped region.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been supported by the General Research Program funded by the Ministry of Science and Technology, Taiwan, R.O.C. [MOST 110-2221-E-027-106-MY3].

ORCID

Chun-Yang Peng http://orcid.org/0000-0002-3345-4493
You-Jin Park http://orcid.org/0000-0002-1006-5380

References

Angiulli, F., S. Basta, S. Lodi, and C. Sartori. 2012. Distributed strategies for mining outliers in large data sets. *IEEE Transactions on Knowledge and Data Engineering* 25 (7):1520-1532. doi:10.1109/TKDE.2012.71.

Angiulli, F., S. Basta, and C. Pizzuti. 2006. Distance-based detection and prediction of outliers. *IEEE Transactions on Knowledge and Data Engineering* 18 (2):145-160. doi:10.1109/TKDE.2006.29.

Angiulli, F., and F. Fassetti. 2009. Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets. *ACM Transactions on Knowledge Discovery from Data* 3 (1):1-57. doi:10.1145/1497577.1497581.

Angiulli, F., and C. Pizzuti. 2005. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering* 17 (2): 203-215. doi:10.1109/TKDE.2005.31.

Attenberg, J., and S. Ertekin. 2013. Class imbalance and active learning. In *Imbalanced Learning: Foundations, Algorithms, and Applications*, ed. H. He, and Y. Ma, 101–49. Piscataway, New Jersey, United States: Wiley-IEEE. doi:10.1002/9781118646106.ch6.

Barua, S., M. M. Islam, X. Yao, and K. Murase. 2012. MWMOTE-Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering* 26 (2): 405-425. doi: 10.1109/TKDE.2012.232.

Bauder, R. A., and T. M. Khoshgoftaar. 2018. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Information Science and Systems* 6 (1):1–14. doi:10.1007/s13755-018-0051-3.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander. 2002. LOF: Identifying density-based local outliers. Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA. ed. W. Chen, J. Naughton, and P. A. Bernstein, 29 (2): 93–104. New York, United States. doi:10.1145/335191.335388.

Buda, M., A. Maki, and M. A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106:249–59. doi:10.1016/j.neunet.2018.07.011.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic smaller over sampling technique. *Journal of Artificial Intelligence Research* 16:321–57. doi:10.1613/jair.953.

Chen, L., W. Wang, and Y. Yang. 2021. CELOF: Effective and fast memory efficient local outlier detection in high-dimensional data streams. *Applied Soft Computing* 102:107079. doi:10.1016/j.asoc.2021.107079.

Chen, Y., D. Miao, and H. Zhang. 2010. Neighborhood outlier detection. *Expert Systems with Applications* 37 (12):8745–49. doi:10.1016/j.eswa.2010.06.040.

Devi, D., S. K. Biswas, and B. Purkayastha. 2017. Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. *Pattern Recognition Letters* 93:3–12. doi:10.1016/j.patrec.2016.10.006.

Dubey, R., J. Zhou, Y. Wang, P. M. Thompson, and J. Ye. 2014. Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage* 87:220–41. doi:10.1016/j.neuroimage.2013.10.005.

Ebenuwa, S. H., M. S. Sharif, M. Alazab, and A. Al-Nemrat. 2019. Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access* 7:24649–66. doi:10.1109/ACCESS.2019.2899578.

Fotouhi, S., S. Asadi, and M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics* 90:103089. doi:10.1016/j.jbi.2018.12.003.

Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetic (Part C: Applications and Reviews)* 42 (4):463–84. doi:10.1109/TSMCC.2011.2161285.

Gnip, P., L. Vokorokos, and P. Drotár. 2021. Selective oversampling approach for strongly imbalanced data. *PeerJ Computer Science* 7:e604. doi:10.7717/peerjcs.604.

Gong, L., S. Jiang, L. Bo, L. Jiang, and J. Qian. 2019. A novel class-imbalance learning approach for both within-project and cross-project defect prediction. *IEEE Transactions on Reliability* 69 (1):40–54. doi:10.1109/TR.2019.2895462.

Guo, H. X., Y. J. Li, J. Shang, M. Y. Gu, Y. Y. Huang, and G. Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* 73:220–39. doi:10.1016/j.eswa.2016.12.035.

Gupta, U., V. Bhattacharjee, and P. S. Bishnu. 2019. A New Neighborhood-Based Outlier Detection Technique. *Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems*, ed. V. Nath and J. K. Mandal, 556:527–534. Springer Nature, Singapore. doi:10.1007/978-981-13-7091-5_43.

Guzmán-Ponce, A., R. M. Valdovinos, J. S. Sánchez, and J. R. Marcial-Romero. 2020. A new under-sampling method to face class overlap and imbalance. *Applied Sciences* 10 (15):5164. doi:10.3390/app10155164.

Han, H., W. Y. Wang, and B.-H. Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Proceedings of International Conference on Intelligent Computing: Advances in Intelligent Computing*, ed. D. S. Huang, X. -P. Zhang, G. -B. Huang, 3644:878–887. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/11538059_9.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21 (9):1263–84. doi:10.1109/TKDE.2008.239.

He, Y., H. Tan, W. Luo, S. Feng, and J. Fan. 2014. MR-DBSCAN: A scalable MapReduce-based DBSCAN algorithm for heavily skewed data. *Frontiers of Computer Science* 8 (1):83–99. doi:10.1007/s11704-013-3158-3.

Janakiraman, V. M., X. Nguyen, J. Sterniak, and D. Assanis. 2014. Identification of the dynamic operating envelope of HCCI engines using class imbalance learning. *IEEE Transactions on Neural Networks and Learning Systems* 26 (1):98–112. doi:10.1109/TNNLS.2014.2311466.

Karami, A., and R. Johansson. 2014. Choosing DBSCAN parameters automatically using differential evolution. *International Journal of Computer Applications* 91 (7):1–11. doi:10.5120/15890-5059.

Krawczyk, B. 2016. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence* 5 (4):221–32. doi:10.1007/s13748-016-0094-0.

Leevy, J. L., T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5 (1):42. doi:10.1186/s40537-018-0151-6.

Li, X., J. Lv, and Z. Yi. 2016. An efficient representation-based method for boundary point and outlier detection. *IEEE Transactions on Neural Networks and Learning Systems* 29 (1):51–62. doi:10.1109/TNNLS.2016.2614896.

Liang, X. W., A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang. 2020. LR-SMOTE - An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems* 196:105845. doi:10.1016/j.knosys.2020.105845.

Lin, W.-C., C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang. 2017. Clustering-based undersampling in class-imbalanced data. *Information Sciences* 409–410:17–26. doi:10.1016/j.ins.2017.05.008.

Liu, X.-Y., J. Wu, and Z.-H. Zhou. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics (Part B: Cybernetics)* 39 (2):539–50. doi:10.1109/TSMCB.2008.2007853.

Liu, Z., M. J. Zuo, X. Zhao, and H. Xu. 2015. An analytical approach to fast parameter selection of Gaussian RBF Kernel for support vector machine. *Journal of Information Science and Engineering* 31 (2):691–710. doi:10.6688/JISE.2015.31.2.18.

Loyola-González, O., J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto. 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing* 175:935–47. doi:10.1016/j.neucom.2015.04.120.

Ofek, N., L. Rokach, R. Stern, and A. Shabtai. 2017. Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing* 243:88–102. doi:10.1016/j.neucom.2017.03.011.

Rezaee, M. J., M. Eshkevari, M. Saberi, and O. Hussain. 2021. GBK-means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game. *Knowledge-Based Systems* 213:106672. doi:10.1016/j.knosys.2020.106672.

Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems* 42 (3):1–21. doi:10.1145/3068335.

Seiffert, C., T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano. 2010. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics (Part A: Systems and Humans)* 40 (1):185–97. doi:10.1109/TSMCA.2009.2029559.

Shi, C., X. Li, J. Lv, Y. Yin, and I. Mumtaz. 2020. Robust geodesic based outlier detection for class imbalance problem. *Pattern Recognition Letters* 131:428–34. doi:10.1016/j.patrec.2020.01.028.

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

Wang, R., Q. Zhu, J. Luo, and F. Zhu. 2020. Local dynamic neighborhood based outlier detection approach and its framework for large-scale datasets. *Egyptian Informatics Journal (Available online)*. doi:10.1016/j.eij.2020.06.001.

Wei, J., H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang. 2020. NI-MWMOTE: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems. *Expert Systems with Applications* 158:113504. doi:10.1016/j.eswa.2020.113504.

Xie, W., G. Liang, Z. Dong, B. Tan, and B. Zhang. 2019. An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. *Mathematical Problems in Engineering* 2019:3526539. doi:10.1155/2019/3526539.

Yang, Z., and D. Gao. 2013. Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. *Applied Mathematics & Information Sciences* 7 (1L):375–81. doi:10.12785/AMIS/071L50.

Yen, S.-J., and Y.-S. Lee. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* 36 (3):5718–27. doi:10.1016/j.eswa.2008.06.108.

Yuan, Z., X. Zhang, and S. Feng. 2018. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures. *Expert Systems with Applications* 112:243–257. doi:10.1016/j.eswa.2018.06.013.

Zhu, T., Y. Lin, and Y. Liu. 2017. Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognition* 72:327–40. doi:10.1016/j.patcog.2017.07.024.