BCONDS: Borderline Counterfactual Oversampling with Noise Elimination and Density Scoring

Asifa Mehmood Qureshi(✉), Abhishek Kaushik, Róisín Loughran, and Fergal McCaffery

Regulated Software Research Centre (RSRC), Dundalk Institute of Technology,
Dundalk, Ireland
{asifa.mehmood,abhishek.kaushik,roisin.loughran,
fergal.mccaffery}@dkit.ie

Abstract. Class imbalance in medical datasets may lead to the generation of biased Machine Learning models. Several methods are used to balance datasets but they do not consider the majority class samples while oversampling. Therefore, in this study, we proposed a novel technique called Borderline Counterfactual Oversampling with Noise elimination and Density Scoring (BCONDS). The method utilises isolation forest to remove the noisy samples from the majority class. Gower distance is used to find borderline minority class instances and extract their corresponding majority class neighbours. These neighbouring samples are then used to generate counterfactuals in order to enhance the separability of classes. The empirical analysis of four benchmark medical datasets indicates that our proposed technique outperforms other state-of-the-art techniques. On average, an improvement of 9.6% and 5.9% is recorded in the AUC and Gmean values of BCONDS when compared with other methods.

Keywords: counterfactual · borderline · density scoring · oversampling · gower distance · medical datasets

1 Introduction

The distribution of data samples in many real-world medical classification datasets is not uniform. This imbalance occurs when the majority class typically dominates the data, causing the classification algorithms to overlook or incorrectly categorise the minority class samples. In the healthcare domain, this skewness towards one class can cause discrimination in the output of automated decision-making systems that may have serious consequences [6].

The widely used method to handle this problem is oversampling [21]. Chawla et al. presented a method called the Synthetic Minority Oversampling Technique (SMOTE) that uses interpolation to generate synthetic samples of the minority

D. Cafolla et al. (Eds.): AIiH 2025, LNCS 16038, pp. 422–432, 2026.
https://doi.org/10.1007/978-3-032-00652-3_30

class [1]. SMOTE has shown successful results in many applications to mitigate the class imbalance issue [5].

However, the performance of SMOTE declines for real-world datasets with high dimensionality and complexity [14]. Therefore, various improved variants are proposed to address these limitations. Most of these variants are focused on generating synthetic samples using minority-class data points. However, neglecting the majority class entirely while generating new samples may lead to inaccurate data generation [22]. Moreover, generating data samples only in safe regions i.e., far from the sample boundary does not contribute to the separability between classes [2].

In this paper, we extend our model [13] by addressing its limitations. The previous model utilised a Support Vector Machine (SVM) classifier to detect borderline data samples, based on their distance from the decision boundary. Consequently, the model's overall performance was heavily reliant on the SVM classifier. Furthermore, the approach did not incorporate mechanisms for identifying and removing noisy samples, which led to the generation of outliers in the balanced dataset.

Therefore, we proposed an improved novel technique for the generation of synthetic data called counterfactuals. To define counterfactual, consider x is a sample in dataset X. The sample x is called a factual sample and its label y is called a factual label. If the value of factual sample x is changed to x' by simulating physical intervention then it can be referred to as a new sample generation called counterfactual sample while the rest of the samples remain unchanged [22]. The proposed model BCONDS is based on noise removal, borderline sample extraction and density scores. At first, the model takes an imbalanced medical dataset. Then, noisy samples are detected using the Isolation Forest (IF) algorithm [11]. To balance the training set, borderline instances of the minority class are identified. The majority-class neighbours of these borderline minority instances are extracted and further used to generate counterfactuals using density scores. BCONDS is evaluated on four benchmark medical datasets. These datasets consist of numerical, continuous and categorical feature values.

The key contributions of this study include: The implementation of IF algorithm to eliminate noisy samples from the dataset. This method prevents the distance calculation for each majority sample as done in most of the oversampling methods, thus decreasing the overall computational cost. A novel method to identify borderline minority samples using Gower distance is implemented as it efficiently handles mixed data types; then majority class neighbours of these borderline samples are extracted and used to generate counterfactuals. To enhance the separability of different classes and ensure guided sampling, a new method based on the Imbalance Ratio (IR) and density scores is used to estimate the number of samples to be generated from each majority neighbour. Also, the performance of the model is assessed on four benchmark medical classification datasets.

To better analyse our proposed technique we formulated the following hypothesis and related research question:

Hypothesis: BCONDS with its noise removal technique and borderline counterfactual generation can help improve the classifier's performance on medical datasets.

RQ1. Can BCONDS improve the classification performance? If so, which classifier has better Area Under Curve (AUC) and Geometric mean (G-mean) scores?

The remainder of the paper is organised as follows:

Section 2 provides an overview of recent oversampling techniques. Section 3 describes the counterfactual generation methodology in detail. Section 4 provides an explanation of datasets and corresponding results. Lastly, Sect. 5 concludes the overall discussion with future directions.

2 Literature Review

This section provides a summary of recent techniques that used synthetic oversampling particularly focused on decision boundary enhancement and noise removal techniques.

Liang et al. proposed LR-SMOTE that used SVM classifier to identify misclassified samples and then generate samples near them. The SVM classifier has a major impact on the algorithm's performance. Therefore, to eliminate the dependency we implemented distance and density-based methods to identify borderline and hard to learn samples. Moreover, an adaptive semi-supervised weighted oversampling (IA-SUWO) technique is proposed in [23]. IA-SUWO uses a weighting mechanism based on least square support spectrum values and the Improving Majority Weighted Minority Oversampling (IMWMO) technique to locate hard to learn minority class data points and to generate samples near them. The method uses Euclidean distance which may not perform well for diverse datatypes. For this reason, BCONDS make use of Gower distance as it can handle diverse data types effectively.

Furthermore, a self-adaptive Robust SMOTE (RSMOTE) is presented by Chen et al. [2]. The method calculates relative density to split the minority class data samples into safe, noisy, and borderline regions. The new samples are generated only in borderline and safe regions only. Also, Range-Controlled SMOTE (RCSMOTE) is proposed by Soltanzadeh et al. [16]. The method uses a sample categorisation scheme to locate minority data points that lie in borderline and safe regions. The generation of new samples is controlled by range calculation to avoid intrusion into majority class points. The majority class samples are not considered by these approaches at any point throughout the data generation process. Contrary to the above, a counterfactual generation method is proposed by Wang et al. [22]. The method takes random samples from the majority class distribution and generates counterfactuals based on truncated probability distribution. The method does not take into account the noisy samples and also it doesn't explicitly identify borderline samples to pass only those to the generation process.

Therefore, we have presented an improved counterfactual generation process that considers both the majority and minority classes to generate new data points. It also explicitly identifies noisy or outlier and borderline samples. The borderline counterfactual generation process is elaborated in Fig. 1.

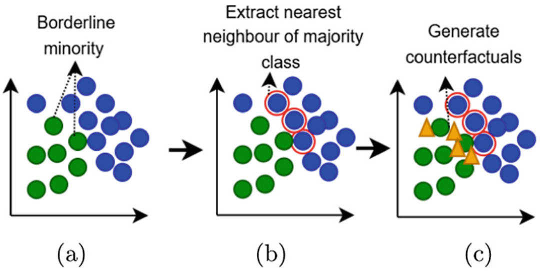

Fig. 1. BCONDS generation process (a) Identification of borderline minority instances (green presents minority class and blue presents majority class) (b) Extract majority class neighbours (c) generate counterfactuals as described in Sect. 3.2 (more generation near dense minority samples represented in yellow). (Color figure online)

3 Methodology

Figure 2 presents the methodology diagram of BCONDS. The method takes an imbalanced medical dataset and identifies noisy samples using IF or iForest algorithm [11]. The dataset is then divided into train and test sets. The training set is passed further to balance the dataset by generating synthetic data. Then, borderline minority instances are located and their neighbours, which belong to the majority class are extracted. The density score for each neighbouring majority sample is calculated to determine the number of counterfactuals to be generated from each data point. The generation process outputs a balanced training dataset which is further used to train ML classifiers and performance evaluation.

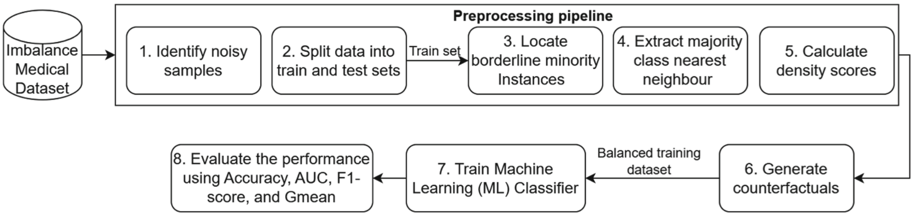
Imbalance Medical Dataset
Preprocessing pipeline
1. Identify noisy samples
2. Split data into train and test sets
Train set
3. Locate borderline minority Instances
4. Extract majority class nearest neighbour
5. Calculate density scores
6. Generate counterfactuals
Balanced training dataset
7. Train Machine Learning (ML) Classifier
8. Evaluate the performance using Accuracy, AUC, F1-score, and Gmean

Fig. 2. Overview of the proposed BCONDS to generate counterfactuals

3.1 Preprocessing

This section explains the preprocessing pipeline before counterfactual generation in detail.

In order to *identify noisy samples*, we implemented IF algorithm. It is a model-based approach to identify noise or anomalies in a dataset [11]. The model builds an ensemble of iTrees. Noise is defined as data points with a short average path length in the tree as shown in Fig. 3. This method is efficient for large datasets and eliminates the need to calculate the distance for each sample.

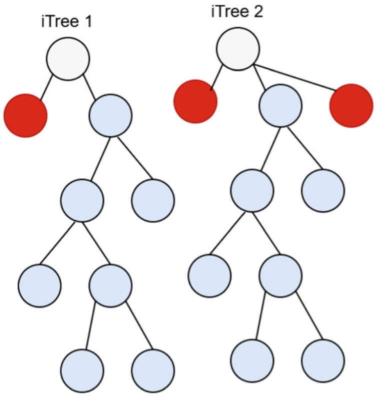

Fig. 3. Isolation Forest (IF) algorithm where red represents noise or anomalies with short average path length (Color figure online)

After identifying noisy samples from the majority class instances, we divide the original dataset into train and test sets of ratio 70:30. The training set is then passed to the *borderline detection method*. For each minority class data point, its m number of neighbours is extracted using Gower distance as given in Eq. 1. Gower distance can efficiently handle mixed data types and provide better measurement of dissimilarity [7].

$$D_G(A, B) = \frac{\sum_{s=1}^{n} |A_s - B_s|}{\sum_{s=1}^{n} |\max(A_s) - \min(A_s)|} \tag{1}$$

where A and B are the two data points. A_s and B_s are the features with $s = 1, 2, 3, ...\ n$. n indicates the total number of features in a dataset.

If all the neighbours of the minority class sample under consideration belong to another class, that point is not used in the data generation process. If the number of neighbours lies between 2 and m, then it is considered a borderline instance, and its majority class neighbours are extracted to generate counterfactuals. The number of the nearest neighbours (m) is kept 3 for PIDD and Haberman datasets and 5 for WDBC and Parkinsons dataset to get the best results.

Afterwards, the density score for each of the extracted majority class neighbour points is calculated using Eq. 2.

$$\text{Density Score} = \sum_{i=1}^{n} \frac{1}{D_G(A,B)_i + \epsilon} \tag{2}$$

where n is the neighbours, $D_G(A,B)_i$ is the distance of i^{th} neighbour and ϵ is the small distance to avoid division by zero. A higher density score indicates the presence of a minority sample within the dense majority points region. Further, in the final selection of majority class data points for counterfactual generation, samples that are also identified as noisy by IF are discarded so that they are not used in sample generation.

3.2 Generate Counterfactuals

The finally selected majority class data points are then used to generate counterfactuals. The adaptive sampling rate is determined by using Algorithm 1 to guide the resampling process.

Algorithm 1: Find adaptive sampling rate

```
Input: nml, ds, tmas, tmis →neighbour list, density score, train majority, train minority
Output: usr, hps, counterfactuals        →uniform sampling rate, hard to predict samples
imbalance = $len$(tmas) - $len$(tmis)
avg_ds = AVG(ds)
hps=extract(nml)  with  ds > avg_ds        → points with density scores greater than
average

If imbalance > 0:
  usr = $\frac{len(\text{tmas})}{len(\text{tmis})} \times len(\text{nml})$
  For $i$ in range(usr):
    generate_counterfactuals(nml)
    imbalance = imbalance − 1
If imbalance > 0:
  while imbalance > 0:
    generate_counterfactuals(hps)
    imbalance = imbalance − 1
```

After calculating the sampling rate, counterfactuals are generated as detailed in [13]. For any majority sample s_n, the set of perturbations to generate counterfactual is defined as Eq. 3.

$$\mathcal{B}_{nm} = \{\Delta s_n \mid \Delta s_{nm} \sim \mathcal{F}_{nm},\, s'_n = s_n + \Delta s_n, s_n \in S_j,\, f_{\hat{w}}(s_n) = l,\, f_{\hat{w}}(s'_n) = k\} \tag{3}$$

where $\mathcal{B}_{nm}$ is the generated counterfactual, s_n is the majority class sample (factual sample), s_{nm} represents each feature, Δs_n is the change whereas l is the output label of majority class and k is the output label for minority class. $\mathcal{F}_{nm}$ represents the distribution on the perturbations Δs_{nm} estimated by using probability density function of truncated normal distribution. It is defined in Eq. 4.

$$\mathcal{F}_{nm}\left(\Delta s_{nm} \mid S_{nm}, S_n^-, S_n^+, \sigma\right)=\begin{cases} \dfrac{\frac{1}{\sigma}\psi\left(\frac{\Delta s_{nm}}{\sigma}\right)}{\Phi\left(\frac{S_n^+ - S_{nm}}{\sigma}\right)-\Phi\left(\frac{S_n^- - S_{nm}}{\sigma}\right)} & \text{if } S_n^- \leq s_{nm} + \Delta s_{nm} \leq S_n^+, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where S_n^+ and S_n^- are the maximum and minimum values of the m^{th} feature of the original dataset, σ and ψ is the standard deviation and probability density function of standard normal distribution respectively. Φ presents the cumulative distribution function. ΔS_{nm} is truncated to the range $[S_m^- - S_{nm}, S_m^+ - S_{nm}]$.

Further, a Random Forest (RF) classifier is trained on the original dataset to classify each generated counterfactual. If the newly generated sample is classified as a minority class then it is added to the dataset otherwise the sample is discarded and the next perturbation is performed. The process continues until the training set becomes balanced.

In practice, the oversampling is done by utilising the Gibbs sampling principles i.e., introducing a latent variable to modify features iteratively [3]. Also, it enables the method to explore possible counterfactual space using truncated normal sampling. This reduces the overall computational cost of the model.

4 Performance Evaluation

4.1 Datasets

We assessed our proposed method BCONDS on four benchmark open-sourced medical datasets including the Pima Indian Diabetes Dataset (PIDD) [8], Haberman dataset [19], Wisconsin Breast Cancer dataset (WDBC) [18], and Parkinsons Disease Dataset [20]. These datasets contain continuous, numerical, and categorical datatypes. The detail of each dataset is given in Table 1.

Table 1. Summary of each evaluated medical dataset where NoC = Number of Classes, NoS = Number of Samples, D = Dimension, MaC = Majority Class, MiC = Minority Class, and IR = Imbalance Ratio

Dataset	NoC	NoS	D	MaC	MiC	IR
Pima Indian Diabetes Dataset (PIDD)	2	768	8	500	268	1.866
Haberman	2	306	3	225	81	2.778
Wisconsin Breast Cancer Dataset (WDBC)	2	569	30	357	212	1.7
Parkinsons Dataset	2	195	23	147	48	3.060

4.2 Evaluation of Our Proposed Method

We trained and tested three ML classifiers: Logistic Regression (LR) as it is computationally efficient, C4.5 is a baseline standard for comparing oversampling techniques [4], and RF is used extensively to assess the quality of generated data. The number of samples generated for each dataset to balance the training sets is given in Table 2.

Table 2. Number of generated samples for each dataset.

Dataset	Number of Generated Samples
PIDD	161
Haberman	104
WDBC	100
Parkinsons	70

We used four evaluation metrics to assess the performance: Accuracy, AUC, F1-score, and Gmean. These metrics are extensively used to evaluate oversampling techniques [9,10,23]. Table 3 shows these metric values for the models trained on synthetically balanced datasets. The difference in the performance of BCONDS on each dataset depends on several factors including IR, distribution of minority and majority classes, and number of borderline instances and noise in each dataset.

Table 3. Performance metrics for each dataset and classifiers using BCONDS

Datasets	Classifiers	Accuracy	AUC	F1-score	Gmean
PIDD	LR	0.7749	0.8448	0.7174	0.7857
	C4.5	0.8442	0.8817	0.7857	0.8395
	RF	**0.8528**	**0.9173**	**0.8046**	**0.8579**
Haberman	LR	0.7174	0.6713	0.8060	0.6145
	C4.5	0.7282	0.6969	0.8148	0.6202
	RF	**0.7391**	**0.7628**	**0.8235**	**0.6258**
WDBC	LR	0.9591	0.9913	0.9674	0.9577
	C4.5	0.9415	0.9586	0.9524	0.9469
	RF	**0.9708**	**0.9981**	**0.9765**	**0.9735**
Parkinsons	LR	0.7288	0.8727	0.8000	0.7303
	C4.5	0.8983	0.9015	0.9302	0.8876
	RF	**0.9153**	**0.9530**	**0.9425**	**0.8987**

4.3 Comparison with State-of-the-Art Techniques

Based on the analysis, the answer to our devised research question is as follows:

RQ1. Can BCONDS improve the classification performance? If so, which classifier has better Area Under Curve (AUC) and Geometric mean (G-mean) scores?

Table 4 compares BCONDS on those metrics that were common with other state-of-the-art methods. The comparison indicates that our proposed method outperforms others except for the Gmean value of SymProD and IA-SUWO techniques, which are slightly higher on Haberman and Parkinsons datasets. Furthermore, BCONDS is more accurate with an average improvement of 9.6% and 5.9% for AUC and Gmean respectively as compared to other oversampling methods. This is mainly because BCONDS incorporates the inherent information of the majority class along with the minority class unlike most of these methods. Also, the noise removal and borderline counterfactual generation help to enhance

Table 4. Comparison of BCONDS with other state-of-the-art techniques

Dataset	Technique	AUC	Gmean
PIDD	LWNB+SMOTE [15]	70.7	67.9
	RCSMOTE [16]	78.4	76.2
	NI-MWMOTE [24]	74.6	74.5
	SMOTE-LOF [12]	81.4	–
	KSMOTE [17]	77.8	–
	RSMOTE [2]	88.6	82.5
	IA-SUWO [23]	73.7	73.6
	SymProD [9]	82.6	76.4
	BCONDS	**91.7**	**85.8**
Haberman	LWNB+SMOTE [15]	67.2	57.0
	RCSMOTE [16]	71.5	58.2
	NI-MWMOTE [24]	62.2	59.9
	SMOTE-LOF [12]	73.1	–
	KSMOTE [17]	67.2	–
	IA-SUWO [23]	62.4	59.9
	SymProD [9]	71.4	**67.3**
	BCONDS	**76.3**	62.6
WDBC	LWNB+SMOTE [15]	97.8	92.4
	Counterfactual [22]	97.8	97.0
	BCONDS	**99.8**	**97.3**
Parkinsons	RCSMOTE [16]	88.9	85.9
	IA-SUWO [23]	90.6	**90.2**
	BCONDS	**95.3**	89.9

the separability of the classifier. Therefore, we can conclude that BCONDS help to improve the performance of the classifier.

To statistically identify the best-performing classifier, we applied a one-way ANOVA test on each dataset using 5 fold-cross validation. The p-values were less than 0.05 (significance level) except for the Haberman dataset because it is a small dataset with less number of features. Therefore, in general, we reject the null hypothesis which states that there is no significant difference between the classifier performance. Keeping in view the AUC scores, we conclude that the RF classifier performs best as compared to LR and C4.5.

Consequently, we fail to reject our hypothesis stating that BCONDS help to improve the performance of the classifier on medical datasets.

5 Conclusion and Future Work

In this study, we proposed a new borderline counterfactual generation method called BCONDS. The method incorporates noise elimination by employing IF algorithm. Furthermore, the borderline minority instances are identified and then their majority-class neighbours are extracted to generate counterfactuals. An adaptive sampling rate based on density scores is calculated to implement guided oversampling. The comparison shows that BCONDS outperforms other state-of-the-art methods. However, finding an optimal value for the majority nearest neighbour to be extracted is challenging as it depends on the nature of the dataset. In future, we will upgrade our model to find adaptive nearest neighbour value by taking into account both intra and inter-class imbalance. Also, we will extend our experiment to multi-class, large and high-dimensional medical datasets.

Acknowledgments. This publication has emanated from research conducted with the financial support of Research Ireland (RI) under Grant number 21/FFP-A/9255.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
2. Chen, B., Xia, S., Chen, Z., Wang, B., Wang, G.: RSMOTE: a self-adaptive robust smote for imbalanced problems with label noise. Inf. Sci. **553**, 397–428 (2021)
3. Damien, P., Walker, S.G.: Sampling truncated normal, beta, and gamma densities. J. Comput. Graph. Stat. **10**(2), 206–215 (2001)
4. Drummond, C., Holte, R.C., et al.: C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on Learning from Imbalanced Datasets II, vol. 11 (2003)
5. Fernández, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)

6. Gesi, J., Shen, X., Geng, Y., Chen, Q., Ahmed, I.: Leveraging feature bias for scalable misprediction explanation of machine learning models. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), pp. 1559–1570. IEEE (2023)
7. Kadhim, M.N., Al-Shammary, D., Sufi, F.: A novel voice classification based on Gower distance for Parkinson disease detection. Int. J. Med. Inf. **191**, 105583 (2024)
8. Kaggle: Pima Indian diabetes dataset (1988). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed 10 Jan 2025
9. Kunakorntum, I., Hinthong, W., Phunchongharn, P.: A synthetic minority based on probabilistic distribution (SyMProD) oversampling for imbalanced datasets. IEEE Access **8**, 114692–114704 (2020)
10. Liang, X., Jiang, A., Li, T., Xue, Y., Wang, G.: LR-SMOTE–an improved unbalanced data set oversampling based on K-means and SVM. Knowl.-Based Syst. **196**, 105845 (2020)
11. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
12. Maulidevi, N.U., Surendro, K., et al.: SMOTE-LOF for noise identification in imbalanced data classification. J. King Saud Univ.-Comput. Inf. Sci. **34**(6), 3413–3423 (2022)
13. Qureshi, A.M., Kaushik, A., Regan, G., McDaid, K., McCaffery, F.: Handling class imbalance via counterfactual generation in medical datasets. In: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, pp. 102–113 (2024)
14. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering. Inf. Sci. **291**, 184–203 (2015)
15. Sağlam, F., Cengiz, M.A.: Local resampling for locally weighted Naïve Bayes in imbalanced data. Computing **106**(1), 185–200 (2024)
16. Soltanzadeh, P., Hashemzadeh, M.: RCSMOTE: range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf. Sci. **542**, 92–111 (2021)
17. Thejas, G., Hariprasad, Y., Iyengar, S., Sunitha, N., Badrinath, P., Chennupati, S.: An extension of synthetic minority oversampling technique based on Kalman filter for imbalanced datasets. Mach. Learn. Appl. **8**, 100267 (2022)
18. UCI: Breast cancer Wisconsin (diagnostic) (1995). https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. Accessed 10 Jan 2025
19. UCI: Haberman's survival (1999). https://archive.ics.uci.edu/dataset/43/haberman+s+survival. Accessed 10 Jan 2025
20. UCI: Parkinsons disease detection dataset (2008). https://archive.ics.uci.edu/dataset/174/parkinsons. Accessed 10 Jan 2025
21. Verbiest, N., Ramentol, E., Cornelis, C., Herrera, F.: Preprocessing noisy imbalanced datasets using smote enhanced with fuzzy rough prototype selection. Appl. Soft Comput. **22**, 511–517 (2014)
22. Wang, S., et al.: Counterfactual-based minority oversampling for imbalanced classification. Eng. Appl. Artif. Intell. **122**, 106024 (2023)
23. Wei, J., Huang, H., Yao, L., Hu, Y., Fan, Q., Huang, D.: IA-SUWO: an improving adaptive semi-unsupervised weighted oversampling for imbalanced classification problems. Knowl.-Based Syst. **203**, 106116 (2020)
24. Wei, J., Huang, H., Yao, L., Hu, Y., Fan, Q., Huang, D.: NI-MWMOTE: an improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems. Expert Syst. Appl. **158**, 113504 (2020)