Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, 34, S. 396-402, Mart 2022
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 34, pp. 396-402, March 2022
Copyright © 2022 EJOSAT
Research Article

A Salp Swarm-Based Under-Sampling Approach for Medical Imbalanced Data Classification

Mohammed H. Ibrahim[1*]

[1*]Necmettin Erbakan University, Faculty of Engineering, Departmant of Computer Engineering, Konya, Turkey, (ORCID: 0000-0002-6093-6105), mibrahim@erbakan.edu.tr

(2nd International Conference on Applied Engineering and Natural Sciences ICAENS 2022, March 10-13, 2022)

(**DOI:** 10.31590/ejosat.1082451)

ATIF/REFERENCE: Ibrahim, M. H. (2022). A Salp Swarm-Based Under-Sampling Approach for Medical Imbalanced Data Classification. *European Journal of Science and Technology*, (34), 396-402.

Abstract

Data imbalance refers to the unequal distribution of classes within a dataset that directly affects the accuracy of machine learning classification algorithms. Although many resampling techniques have been proposed by researchers, learning from imbalanced data is still considered one of the contemporary challenges. The class imbalanced problem has been complicated as most of the existing techniques don't manage the similarity relationships between minority and majority classes well. In addition, due to the complex relationships among classes, most of the existing techniques do not focus on retaining valuable samples in the majority class(es) properly. In this article, a salp swarm optimization-based under-sampling technique (SSBUT) is proposed to address data class imbalance problems. Utilizing the proposed SSBUT, the similarity relationship among the samples of the majority class is well analyzed, and the samples that do not affect the accuracy of the classification algorithm are eliminated from the majority class. The performance of the proposed SSBUT has been tested on benchmark medical imbalanced datasets and the obtained results have been compared with state-of-the-art under-sampling techniques. The experimental results show that the proposed SSBUT consistently outperformed the state-of-the-art under-sampling techniques in terms of various evaluation criteria.

Keywords: Classification, Machine learning, Medical Imbalanced data classification, Salp swarm optimization, Under-sampling.

Dengesiz Tıbbi Veri Sınıflandırması İçin Salp Sürü Tabanlı Bir Aşağı-Örnekleme Yaklaşımı

Öz

Veri dengesizliği bir veri kümesi içindeki sınıfların eşit olmayan dağılımıdır ve makine öğrenmesi algoritmalarının başarısını doğrudan etkilemektedir. Araştırmacılar tarafından birçok yeniden örnekleme teknikleri önerilmiş olmasına rağmen, dengesiz verilerden öğrenme hala güncel zorluklardan biri olarak kabul edilmektedir. Mevcut tekniklerin birçoğu azınlık ve çoğunluk sınıflar arasındaki benzerlik ilişkilerini iyi bir şekilde yönetemediği için sınıf dengesizliği sorunu karmaşık hale gelmektedir. Ayrıca, sınıflar arasındaki karmaşık ilişkilerden dolayı mevcut tekniklerin birçoğu çoğunluk sınıf(lar)ında ki değerli örneklerin uygun bir şekilde veri kümesinde tutulmasına odaklanamaz. Bu makalede, veri sınıf dengesizliği problemini çözmek için salp sürüsü optimizasyon yöntemi kullanılarak bir aşağı örnekleme tekniği (SSBUT) önerilmiştir. Önerilen SSBUT çoğunluk sınıfına ait örnekler arasındaki benzerlik ilişkisini iyi analiz eder ve sınıflandırma algoritmasının doğruluğunu etkilemeyen örnekleri çoğunluk sınıfından çıkarır. Önerilen SSBUT'un performansı, tıbbi dengesiz veri kümeleri üzerinde test edilmiş ve elde edilen sonuçlar en güncel aşağı örnekleme teknikleri ile karşılaştırılmıştır. Deneysel sonuçlara göre, önerilen SSBUT tekniği birçok değerlendirme ölçütüne göre en güncel aşağı örnekleme tekniklerinden daha iyi performans sergilemiştir.

Anahtar Kelimeler: Aşağı-örnekleme, Makine öğrenmesi, Salp sürüsü optimizasyonu, Sınıflandırma, Tıbbi Dengesiz veri sınıflandırması.

[*] Corresponding Author: Necmettin Erbakan Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Konya, Türkiye ORCID: 0000-0002-6093-6105, mibrahim@erbakan.edu.tr

1. Introduction

In machine learning, classification is known as a training system that is trained with a dataset whose samples are labeled [1]. This system is used to classify new unseen samples to which class it belongs [2], however, the performance of this classification almost depends on the dataset [3]. Although the datasets contain too many samples, unfortunately, there is a lack of quality in such datasets. Machine learning classification algorithms assumed that classes were distributed in the same way [4]. However, this assumption is wrong in most real-world datasets, for example, diagnosis of diseases, fraud detection, and intrusion detection, because in such datasets, one class contains more samples, while the other class contains fewer samples. This reflects on the classification algorithm's precision, thus, in such a case, although a good accuracy can be achieved, however, we don't achieve good scores according to very important evaluation metrics, such as precision (PRE), recall (REC), specificity (SPE), F1-measure (F1-M), and area under curve (AUC) [5]. This is known as the class imbalance problem [6]. The minority class(es) and the majority class(es) are known as important terms in the class imbalance problem. In an imbalanced dataset, the minimum number of samples represents the minority class(es), on the other hand, a maximum number of samples represents the majority class(es) [7]. Recently, there is a large interest in the class imbalance topic by the researchers. Thus, it is considered a challenging issue and many techniques have been developed by researchers to solve this tricky problem [8,9]. Resampling techniques are considered one of the preprocessing techniques and commonly used approaches to making the dataset balanced [10]. Resampling techniques can be performed on the imbalanced dataset with under-sampling or over-sampling techniques [11]. The under-sampling process is applied to decrease the number of samples of the majority class(es) by eliminating the samples [12,13] and the random under-sampling, nearmiss-1, nearmiss-2, condensed nearest neighbor, and repeated edited nearest neighbor are examples of the under-sampling techniques [12]. On the other hand, the over-sampling process is implemented to increase the number of minority class(es) by generating new synthetic samples [14] and the random oversampling technique (ROS) [15], synthetic minority oversampling technique (SMOTE) [16], and Borderline-SMOTE [17] are examples of the oversampling techniques. Due to the effect of imbalanced datasets on the classification algorithms, many under-sampling techniques have been developed by researchers recently. For instance, Chih-Fong Tsai et al. proposed a cluster-based instance selection (CBIS) under-sampling method using the clustering approach to find the similarity relationship between the majority class and the minority class [18]. Pattaramon Vuttipittayamongkol and Eyad Elyan proposed a Neighborhood-based under-sampling technique based on the nearest neighbor approach. This under-sampling technique reduces the majority class by identifying and eliminating the overlapping data [19]. Debashree Devi et al. proposed an under-sampling technique called a boosting aided adaptive cluster-based under-sampling technique by using the AdaBoost ensemble learning model, the proposed under-sampling technique eliminates the insignificant data after clustering the data of the majority class [20]. In addition, a consensus clustering-based-undersampling technique [21], cluster-based under-sampling with random forest algorithm [22], cluster-based under-sampling with Random Forest classifier [23], and cluster-based majority under-sampling [24] techniques are proposed to imbalanced problems and depend on the standard clustering approach in the clustering process. In general, under-sampling techniques are based on standard clustering algorithms to find the centroid cluster, but as it is known, standard clustering algorithms are weak in the strategy of finding a cluster center of data, since they seek solutions in local search space [25]. Since the optimization algorithms have a global search space capability [26], in this article, the salp swarm optimization algorithm (SSA) [27] is exploited to find the similarity relationship among the majority class's samples.

This article is organized as follows: the material and methods are explained in Section 2. Section 3 describes the proposed under-sampling technique. Experimental results are presented and discussed in Section 4. Finally, the conclusion of the article is given in Section 5.

2. Material and Methods

2.1. Medical Datasets

To evaluate the performance of the proposed optimization-based under-sampling technique, the proposed optimization-based under-sampling is applied to breast cancer, diabetes, and blood transfusion medical imbalanced datasets, and then the performance measures obtained are compared with the results of the state-of-the-art cluster-based under-sampling techniques. The medical imbalanced datasets were downloaded from the University of California, Irvine (UCI) machine learning repository [28], and the imbalanced rate (IR) for each medical imbalanced dataset was calculated according to Equation 1. The characteristics of the medical imbalanced datasets used in this article are given in Table 1 [20].

$$IR = \frac{\#\ of\ samples\ in\ majority\ class}{\#\ of\ samples\ in\ minority\ class} \tag{1}$$

Table 1. The characteristics of the medical imbalanced datasets

Dataset	Number of			Minority	Majority	IR
	instances	features	classes			
Breast cancer	683	10	2,4	4	2	1.86
Diabetes	768	8	0,1	1	0	1.87
Blood transfusion	748	4	0,1	1	0	3.2

2.2. Classification Algorithms and Resampling Techniques

Classification is a machine learning layered data mining method used to classify samples of a dataset. Commonly used classification algorithms such as decision trees, support vector machines, K-Nearest neighbor, and artificial neural networks perform the classification process by using their mathematical or statistical functions. The accuracy of the classification algorithms is directly proportional to the preparation of the dataset, and this accuracy generally increases in a well-prepared dataset. The main reason for this is that the training dataset used in the classification model is consistent and balanced. As the classification algorithms are used in many critical areas such as health, aviation, and information technologies, classification accuracy is very important. The class imbalance problem is found in the many datasets of real-world problems, and this problem seriously affects the performance of classification algorithms. To solve this problem, the imbalanced dataset has to be balanced with resampling techniques. In general, according to their function, the resampling technique is categorized into three categories: under-sampling, over-sampling, and hybrid sampling. When these techniques are applied to a dataset, they increase the important criteria values of the classification algorithm, because, in many classification problems, precision (PRE), recall (REC), specificity (SPE), F1-measure (F1-M), and area under curve (AUC) which are obtained from the confusion matrix may be more important than classification accuracy (CA) [29]. In the confusion matrix given in Figure 1 the TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively.

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Figure 1. Confusion matrix

The equations of the evaluation measurements: PRE, REC, SPE, F1-M, and AUC are given below as Equations 2, 3, 4, 5, and 6 respectively [9].

$$PRE = \frac{TP}{TP + FP} \tag{2}$$

$$REC = \frac{TP}{TP + FN} \tag{3}$$

$$SPE = \frac{TN}{TN + FP} \tag{4}$$

$$F1 - M = 2 * \frac{REC * PRE}{REC + PRE} \tag{5}$$

$$AUC = \frac{1}{2} \times \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right] \tag{6}$$

3. The Proposed Under-sampling Technique

In this section, the proposed SSBUT is explained in detail. The proposed SSBUT copes with the imbalanced problem by eliminating redundant samples of the majority class. Samples of the majority class need to be well analyzed before they can be eliminated, because, if the designed under-sampling technique causes to eliminate very important samples of the majority class instead of a redundant one, it will seriously affect the performance of the classification algorithms. Therefore, the designed under-sampling technique needs to discover the data distribution and similarity relationships among the majority class data very well. Many under-sampling techniques usually find the similarity relationship among samples of the majority class with traditional methods that have local searches such as k-mean and c-mean clustering [25]. However, the proposed SSBUT is used an optimization approach that has a global search to discover the similarity relationship among the samples of the majority class. SSA [27] developed by Ali is used as an optimization approach. The mathematical model of the SSA is inspired by salps behaver and movements. The steps of the proposed SSBUT are given below.

Step 1: Determine the majority class and divided the samples of the majority class into a certain number of clusters (supply from the user).

Step 2: Random cluster centers are generated and the clustering process takes place according to the distance equation by assigning the sample to the cluster centroid having the smallest value. In this article, the Euclidean distance equation given in Equation 7 is used as a distance equation [30].

$$E^D(X_i, C_j) = \sqrt{\sum_{l=1}^{p} (X_{il} - C_{jl})^2} \tag{7}$$

Where $X_i; i = 1,2, \dots n$ where n a number of samples in the majority class, each sample with p attributes and $C_j; j = 1,2, \dots k$; where k is the number of clusters in the majority class.

Step 3: Solutions are produced according to the sum of the within-cluster distances fitness function as much as the population size, and the best solution is determined among the solutions. The sum of the within-cluster distances (S_w) index given in Equation 8 is used as internal cluster validation [31] and also represents the fitness function of the optimization algorithm.

$$S_w = \sum_{k=1}^{q} \sum_{i,j \in C_k \; and \; i<j} d(x_i, x_j) \tag{8}$$

In Equation 8, q is the number of clusters and $d(x_i, x_j)$ represents the distance between x_i and x_j samples in the cluster C_k.

Step 4: Based on Equations 9, 10, and 11, SSA attempts to improve the bad solutions by using the best solution.

$$x_j^1 = \begin{cases} F_j + C_1\left(\left(ub_j - Ib_j\right)C_2 + Ib_j\right) & C_3 \geq 0 \\ F_j - C_1\left(\left(ub_j - Ib_j\right)C_2 + Ib_j\right) & C_3 < 0 \end{cases} \tag{9}$$

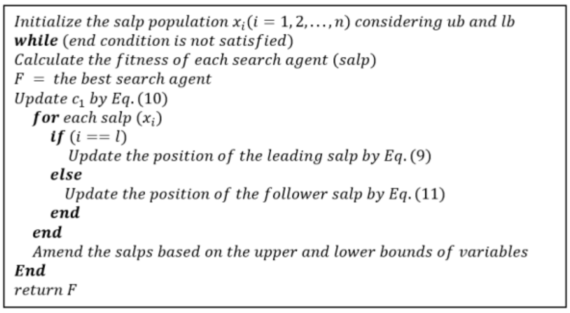
where x_j^1 shows the best cluster's center in the jth dimension, F_j is the cluster's center in the jth dimension, ub_j and Ib_j indicate the upper bound and the lower bound of jth dimension, respectively, and C_1, C_2, and C_3 are random numbers between [0,1]. In Eq. 10, the C_1 which helps to balance exploration is the most important parameter in SSA.

$$C_1 = 2e^{-\left(\frac{4l}{L}\right)^2} \tag{10}$$

where l and L are the current iteration and the maximum number of iterations, respectively. To obtain the best cluster centers the bad cluster centers in the swarm are updated according to the best cluster centers by Equation 11.

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \tag{11}$$

where i ≥ 2 and x_j^i indicate the new cluster's center of ith follower salp in jth dimension. The Pseudocode of the SSA algorithm is given in Figure 2.



```
Initialize the salp population $x_i (i = 1, 2, ..., n)$ considering ub and lb
while (end condition is not satisfied)
Calculate the fitness of each search agent (salp)
F = the best search agent
Update $c_1$ by Eq. (10)
   for each salp ($x_i$)
      if ($i == l$)
         Update the position of the leading salp by Eq. (9)
      else
         Update the position of the follower salp by Eq. (11)
      end
   end
   Amend the salps based on the upper and lower bounds of variables
End
return F
```

Figure 2. Pseudocode of the SSA algorithm

Step 5: Eliminate the redundant samples from the majority class.

4. Experimental

4.1. Results

In this section, to illustrate the effectiveness of the proposed SSBUT, the proposed SSBUT was compared with recent state-of-the-art cluster-based under-sampling techniques existing in [20] given in Table 2.

Table 2. The name of used under-sampling technique and classification algorithms

Technique / Classifier	Technical name	abbreviation
Technique	Sampling based on Clustering with Near Miss 1	SBCNM-1
	Sampling based on Clustering with Near Miss2	SBCNM-2
	Sampling based on Clustering with Near Miss2	SBCNM-3
	Sampling based on Clustering with Most Distance	SBCMD
	Sampling based on Clustering with Most Far	SBCMF
	SMOTE and Cluster Based under sampling	SCUT
	Clustering+OSS	ClusterOSS
	Boosting driven Cluster-based Under-sampling	BoostingCBU
Classifier	Decision Tree C4.5	C4.5
	Support vector Machine	SVM
	Nave Bayes	NB

All the experiments were conducted on a machine with an Intel Core i7@2.00 GHz processor and 8 GB memory, running on Microsoft Windows 10 OS. The proposed SSBUT was coded using Visual Studio 2019 with the C# language. The parameter setting of used classification algorithms is given in Table 3.

Table 3. The parameter setting of used classification algorithms

Classifier	Parameter Name	Parameter setting
Decision Tree	MergeLeaves	On
	MinLeaf	1
	MinParent	10
	Weight	Unit matrix, w of dimension [*l* χ *l*]; *l* is number of training instances
SVM	Kernel function	Gaussian Radial Basis function
	Method	Least squares (LS)
	Scaling factor (rbf-sigma)	0-1
NB	Distiribution	Kernael, normal
	prior	Empirical, uniform
	Distiribution	Kernael, normal

In all experiments, the performance measures PRE, REC, SPE, F1-M, and AUC are obtained using tenfold cross-validation. Tables 4, 5, and 6 list the complete experimental results of C4.5, SVM, and NB classification algorithms for medical imbalanced datasets, respectively and in each table, the best result is in boldface.

Table 4. The performance measurements of under-sampling techniques using the C4.5 classifier

Dataset	Technique	PRE	REC	SPE	F1-M	AUC
Breast cancer	SBCNM-1	71.10	69.71	74.97	79.14	0.56
	SBCNM-2	62.40	75.88	60.76	79.60	0.78
	SBCNM-3	70.16	63.35	60.99	65.37	0.73
	SBCNMD	66.06	80.85	68.83	73.10	0.51
	SBCNMF	67.25	83.35	73.06	85.87	0.77
	SCUT	78.15	84.24	79.98	84.66	0.77
	ClusterOSS	77.23	76.68	80.83	92.56	0.89
	BoostingCBU	88.59	73.87	81.55	92.56	**0.97**
	SSBUT	**94.38**	**93.86**	**93.73**	**95.48**	0.96
Diabetes	SBCNM-1	65.77	74.92	56.56	63.85	0.54
	SBCNM-2	56.60	69.43	55.07	66.62	0.55
	SBCNM-3	70.53	71.51	56.60	62.60	0.73
	SBCNMD	60.72	72.60	64.49	75.03	0.52
	SBCNMF	74.12	67.75	67.10	58.19	0.59
	SCUT	76.33	70.55	63.77	67.83	0.64
	ClusterOSS	63.68	64.80	68.35	60.99	0.88
	BoostingCBU	**83.05**	**83.90**	**74.82**	**74.79**	**0.88**
	SSBUT	79.48	78.83	78.17	78.79	0.82
Blood transfusion	SBCNM-1	71.38	69.39	60.24	66.74	0.61
	SBCNM-2	65.76	76.50	68.29	71.99	0.64
	SBCNM-3	67.48	74.03	68.94	74.81	0.60
	SBCNMD	69.98	78.87	68.98	77.74	0.60
	SBCNMF	77.00	82.95	71.02	75.75	0.75
	SCUT	71.60	84.43	70.07	81.62	0.74
	ClusterOSS	82.62	83.29	73.10	77.20	0.60
	BoostingCBU	77.46	**88.80**	**81.06**	**91.12**	**0.85**
	SSBUT	**86.38**	84.80	84.86	81.00	0.83

Table 5. The performance measurements of under-sampling techniques using the SVM classifier

Dataset	Technique	PRE	REC	SPE	F1-M	AUC
Breast cancer	SBCNM-1	76.29	87.12	62.54	78.27	0.66
	SBCNM-2	74.49	66.46	69.18	73.20	0.72
	SBCNM-3	68.44	78.31	75.84	79.19	0.81
	SBCNMD	75.08	60.82	79.53	81.15	0.67
	SBCNMF	74.15	70.69	71.15	82.35	0.84
	SCUT	80.42	74.76	84.25	70.52	0.71
	ClusterOSS	73.77	72.63	80.31	80.90	0.86
	BoostingCBU	86.28	87.64	73.28	85.39	0.92
	SSBUT	**91.53**	**91.81**	**90.57**	**91.00**	**0.93**
Diabetes	SBCNM-1	60.54	55.92	56.94	71.47	0.51
	SBCNM-2	68.11	58.42	69.12	55.64	0.68
	SBCNM-3	68.57	70.15	69.86	62.84	0.65
	SBCNMD	69.84	60.54	72.74	74.01	0.81
	SBCNMF	66.33	73.74	71.88	74.39	0.82
	SCUT	79.14	69.71	**76.01**	62.84	0.78
	ClusterOSS	75.81	62.44	66.96	73.67	0.63
	BoostingCBU	**81.29**	**83.12**	67.54	**83.27**	**0.88**
	SSBUT	78.92	77.67	77.23	76.48	0.83
Blood transfusion	SBCNM-1	70.94	65.93	74.89	63.78	0.76
	SBCNM-2	77.68	67.20	67.56	69.20	0.61
	SBCNM-3	74.48	67.78	70.05	71.59	0.66
	SBCNMD	80.80	76.90	75.42	77.04	0.74
	SBCNMF	78.39	85.70	78.80	75.45	0.78
	SCUT	72.76	73.60	76.26	70.74	0.57
	ClusterOSS	76.03	71.14	73.60	71.85	0.78
	BoostingCBU	90.35	86.20	81.32	84.24	0.83
	SSBUT	**92.61**	**92.13**	**90.69**	**91.17**	**0.91**

Table 6. The performance measurements of under-sampling techniques using the NB classifier

Dataset	Technique	PRE	REC	SPE	F1-M	AUC
Breast cancer	SBCNM-1	78.99	63.51	69.63	60.60	0.66
	SBCNM-2	75.03	65.10	70.12	73.98	0.90
	SBCNM-3	79.60	81.10	72.04	63.05	0.90
	SBCNMD	67.54	69.38	79.29	69.32	0.78
	SBCNMF	81.29	69.87	83.59	72.00	0.63
	SCUT	68.93	70.02	77.32	74.47	0.69
	ClusterOSS	72.03	81.62	76.71	75.99	0.93
	BoostingCBU	84.65	69.88	68.57	79.95	**0.97**
	SSBUT	**92.63**	**91.81**	**92.18**	**90.75**	0.94
Diabetes	SBCNM-1	57.80	62.35	61.68	64.69	0.65
	SBCNM-2	65.64	66.32	59.14	65.20	0.55
	SBCNM-3	69.83	62.44	61.78	67.48	0.53
	SBCNMD	77.28	66.13	70.53	64.03	0.52
	SBCNMF	75.03	65.10	70.12	73.98	0.71
	SCUT	78.36	79.39	73.21	67.08	0.81
	ClusterOSS	**79.66**	69.38	81.73	71.30	0.73
	BoostingCBU	87.93	72.15	**83.93**	**83.84**	**0.87**
	SSBUT	75.81	**75.27**	74.84	74.61	0.79
Blood transfusion	SBCNM-1	69.74	68.72	68.94	66.13	0.79
	SBCNM-2	71.64	71.68	76.90	76.51	0.71
	SBCNM-3	74.66	70.68	77.17	72.99	0.74
	SBCNMD	71.56	79.21	78.39	74.15	0.63
	SBCNMF	74.34	73.81	68.12	69.52	0.70
	SCUT	74.59	71.23	79.82	70.73	0.87
	ClusterOSS	71.84	70.90	71.87	75.41	0.73
	BoostingCBU	71.80	75.51	86.82	70.61	0.86
	SSBUT	**83.29**	**83.58**	**82.83**	**80.86**	**0.93**

4.2. Discussion

In this section, the results of Tables 4, 5, and 6 are discussed to demonstrate the superiority of the proposed SSBUT over the considered state-of-the-art under-sampling techniques. When considering the C4.5 classification algorithm, the proposed SSBUT in the breast cancer dataset achieved better results than all compared under-sampling techniques with PRE, REC, SPE, and F1-M values of 94.38, 93.86, 93.73, and 95.48, respectively. However, in terms of the AUC criterion, the BoostingCBU method performed better than the proposed method. In the diabetes dataset, the proposed SSBUT outperformed the other under-sampling techniques, except the BoostingCBU technique. In the blood transfusion dataset, the proposed SSBUT performs well only in the PRE criteria with a value of 86.38, the BoostingCBU technique outperformed all the techniques in other criteria. When the proposed SSBUT was evaluated in terms of the SVM classification algorithm, the proposed SSBUT achieved better success than other techniques with PRE, REC, SPE, F1-M, and AUC values of 91.53, 91.81, 90.57, 91.00, and 0.93, respectively, in the breast cancer dataset. In the diabetes dataset, the BoostingCBU technique demonstrated better performance from the proposed SSBUT and the other under-sampling techniques. However, in the blood transfusion dataset, the proposed SSBUT outperformed all used under-sampling techniques with PRE, REC, SPE, F1-M, and AUC values of 92.61, 92.13, 90.69, 91.17, 0.91, respectively. When we examine the results of the BoostingCBU technique, it can be observed that the BoostingCBU technique gets good results, especially in the diabetes dataset, but the PRE, REC, and SPE values obtained by the BoostingCBU technique are fluctuating. As a result, the proposed SSBUT boosts the classification accuracy on the medical datasets by eliminating unimportant samples from the majority class.

5. Conclusion

Machine learning classification algorithms are widely used in critical fields such as medicine, aviation, and banking, thus the accuracy of these classification algorithms depends on the consistency and balance of the dataset. In this article, an optimization-based under-sampling technique named SSBUT is proposed to solve the class imbalance problem. The proposed SSBUT is used to balance the dataset that has class imbalanced by eliminating the redundant samples from the majority class. There is a very important role of the SSA in discovering the similarity relationship among the majority class's samples. The proposed SSBUT is applied to medical imbalanced datasets. The performance of the proposed SSBUT is demonstrated by comparing it with the existing state-of-the-art under-sampling techniques in [20] on medical imbalanced datasets. The experiments illustrate that the proposed SSBUT achieves better performance measurements in breast cancer and blood transfusion medical imbalanced datasets. For future work, the number of clusters in the majority class will be determined using the entropy of the majority class.

References

1. Han J, Pei J, Kamber M. (2011). Data mining: concepts and techniques. Elsevier.
2. Sen PC, Hajra M, Ghosh M. (2020). Supervised classification algorithms in machine learning: A survey and review. In:

Emerging technology in modelling and graphics. Springer, pp 99-111.

3. Özkaya, U., Öztürk, Ş., Barstugan, M. (2020). Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach (pp. 281-295). Springer, Cham.
4. Kwon O, Sim JM. (2013). Effects of data set features on the performances of classification algorithms. Expert Systems with Applications 40 (5):1847-1857.
5. Atomi WH. (2012). The effect of data preprocessing on the performance of artificial neural networks techniques for classification problems. Universiti Tun Hussein Onn Malaysia.
6. Rout N, Mishra D, Mallick MK. (2018). Handling imbalanced data: a survey. In: International proceedings on advances in soft computing, intelligent systems and applications. Springer, pp 431-443.
7. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data 5 (1):1-30.
8. Singh A, Purohit A. (2015). A survey on methods for solving data imbalance problem for classification. International Journal of Computer Applications 127 (15):37-41.
9. Ibrahim MH. (2021). ODBOT: Outlier detection-based oversampling technique for imbalanced datasets learning. Neural Computing and Applications 33 (22):15781-15806.
10. Hasib KM, Iqbal M, Shah FM, Mahmud JA, Popel MH, Showrov M, Hossain I, Ahmed S, Rahman O. (2020). A survey of methods for managing the classification and solution of data imbalance problem. arXiv preprint arXiv:201211870.
11. Abd Elrahman SM, Abraham A. (2013). A review of class imbalance problem. Journal of Network and Innovative Computing 1 (2013):332-340.
12. More A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:160806048.
13. Vuttipittayamongkol P, Elyan E, Petrovski A, Jayne C. (2018). Overlap-based undersampling for improving imbalanced data classification. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp 689-697
14. Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access 4:7940-7957.
15. Chowdhury A, Alspector J. (2003). Data duplication: an imbalance problem? In: ICML'2003 Workshop on Learning from Imbalanced Data Sets (II), Washington, DC.
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321-357.
17. Han H, Wang W-Y, Mao B-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing, 2005. Springer, pp 878-887.
18. Tsai C-F, Lin W-C, Hu Y-H, Yao G-T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Information Sciences 477:47-54.
19. Vuttipittayamongkol P, Elyan E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Information Sciences 509:47-70.
20. Devi D, Namasudra S, Kadry S. (2020). A boosting-aided adaptive cluster-based undersampling approach for treatment of class imbalance problem. International Journal of Data Warehousing and Mining (IJDWM) 16 (3):60-86.
21. Onan A. (2019). Consensus clustering-based undersampling approach to imbalanced learning. Scientific Programming 2019.
22. Arafat MY, Hoque S, Farid DM. (2017). Cluster-based under-sampling with random forest for multi-class imbalanced classification. In: 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE, pp 1-6.
23. Miah MO, Khan SS, Shatabda S, Farid DM. (2019). Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests. In: 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT), 2019. IEEE, pp 1-5.
24. Zhang Y-P, Zhang L-N, Wang Y-C. (2010). Cluster-based majority under-sampling approaches for class imbalance learning. In: 2010 2nd IEEE International Conference on Information and Financial Engineering, IEEE, pp 400-404
25. IBRAHIM MH. (2020). WBBA-KM: a hybrid weight-based bat algorithm with K-means algorithm for cluster analysis. Politeknik Dergisi:1-1.
26. Khishe M, Mosavi MR. (2020). Chimp optimization algorithm. Expert systems with applications 149:113338.
27. Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. (2017). Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software 114:163-191.
28. Asuncion A, Newman D. (2007). UCI machine learning repository. Irvine, CA, USA.
29. Gorunescu F. (2011). Data Mining: Concepts, models and techniques, vol 12. Springer Science & Business Media.
30. Giancarlo R, Bosco GL, Pinello L. (2010). Distance functions, clustering algorithms and microarray data analysis. In: International Conference on Learning and Intelligent Optimization, Springer, pp 125-138.
31. Charrad M, Ghazzali N, Boiteux V, Niknafs A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set| Charrad| Journal of Statistical Software.