CIRUS: Critical Instances Removal based Under-Sampling - A solution for class imbalance problem*

Rekha Gillala[0000-0003-2688-2323][a**], V Krishna Reddy [b], Amit Kumar Tyagi[0000-0003-2657-8700] [c]

[a] Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Hyderabad, Telangana, India-500075
[b] Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
[c] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, 600127, Tamilnadu, India.

Abstract. The most critical issue in real world applications are class imbalance problems. Imbalanced data sets are common across different domain including banking, health care, finance and other. When such data sets are trained on typical classification algorithm they tends to be biased towards the majority class. The learning task becomes more challenging when there is also overlapping region with instances from different classes. In this paper, we propose an undersampling framework for binary classification datasets by removing data points in overlap region called Critical Instances Removal based Under-Sampling (CIRUS). Our method is designed to identify and eliminate majority class instances from the overlapping region. Accurate identification and elimination of these instances maximise the visibility of the minority/positive class instances and at the same time minimises excessive elimination of data, which in turn reduces loss of information. Extensive experiments using simulated and real-world datasets were carried out and the results show comparable performance with state-of-the-art methods across different common metrics with exceptional and statistically significant improvements in sensitivity.

Keywords: Imbalanced dataset· Undersampling· k-NN · Class overlap · Classification.

1 INTRODUCTION

In machine learning, the classification algorithms learn from previously known information for predicting the unknown events. However, most of the datasets from real world domain contains noisy instances [24]. Typical examples include finance, fraud detection, medical diagnosis, customer churn prediction and many

* Supported by KL University.
** Gillala Rekha. Email: gillala.rekha@klh.edu.in

more [10]. Training on these samples degrades the classification performance dramatically. It shows bias towards the over-represented class samples called majority class and ignores under-represented class samples called minority class. Moreover, imbalance occurs in binary classification problem and most of the time the minority samples are of great importance. This problem has been addressed by the machine learning research community over the past decades. The proposed solutions are broadly classified into data-level and algorithm-level techniques [21, 38] [2, 15].
Data level techniques consist of sampling methods to adjust the class distribution while algorithm-level techniques involve modification of existing or creation of new algorithm . Algorithm-level techniques need deep understanding of algorithms and are complicated to implement where as data-level techniques are simple and concentrate on resampling process which in turn can be applied to any classification algorithms. The most popular and commonly used resampling methods include random under-sampling, random oversampling and Synthetic Minority Oversampling TEchnique (SMOTE) [7]. Recently proposed resampling techniques include k-means clustering [12], density-based clustering [4] [6],and ensemble [40]. These techniques are meant to balance the data distribution before classification. However, a number of authors in the past argued that the performance of the classifier was not affected only by unequal class distribution but due may other reasons such as class overlapping, small disjunct and small sample size.

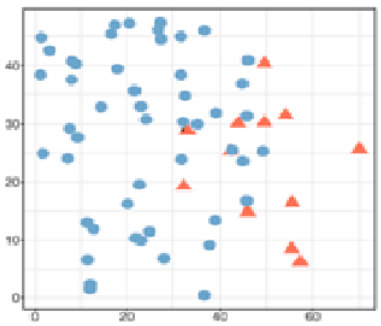

Fig. 1. Overlapping data regions

Consider Figure 1, shows the class distribution of two datasets, the data is overlapped between the classes and may be difficult for the classification algorithm to train on such data. In real-world application, datasets usually not only found imbalance but also overlapped. Therefore removal of majority samples from the overlapped region as shown in Figure 2 is a rational approach to improve the classification performance. In this work, we propose a nearest neighbor based undersampling approach for finding and eliminating the negative/ majority samples from overlapped region. By using this approach, we assume that

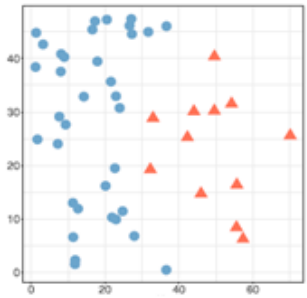

Fig. 2. After removal of overlapping data

most of the majority samples from the overlapped region are eliminated from the dataset. The two fold advantages of our approach is firstly the visibility of positive/minority samples increases in the complete dataset and next, more specific overlapped majority instances are identified using k-nearest neighbor which avoid unnecessary information loss. The main contribution of this paper is to propose a framework for handling overlapping data in the decision boundary of a skewed data distribution. An extensive experiments were carried out on highly imbalanced and overlapped datasets.

1.1 Imbalanced Data Classification Problem: An Overview

A dataset is said to skewed distributed when the number of samples of one class are larger in number than the ones from other classes. Moreover, the class with smaller number of samples is usually the class of interest from the learning point of view [8]. In many real world applications, this problem is of great interest, such as telecommunication customers churn [13], oil spills detection in satellite radar images [26], fraudulent telephone call detection [14], and specifically in medical diagnosis [30], [16].

Traditional classifiers when trained on such datasets have a bias towards the classes with larger number of instances (i.e, majority class). In turn, the minority class are usually ignored by considering them as noise. In this way, minority class samples are most often misclassified even though they are important in classification. The learning task does not hinder only by skewed data distribution but there are series of issue related to this problem like small size samples, overlapping between classes and small disjuncts. In Figure 3, we illustrate examples of the three kinds of imbalance class distribution.

a) Small size sample: It refer to the problem where not all the classes for a given dataset are represented equally. The high imbalanced ratio may lead to poor classification, resulting in complete uncountable for the said class [10].
b) Class overlapping: In the presence of overlap between the classes, the classifier tend to wrongly classify the minority instances [35]. Hence, combination

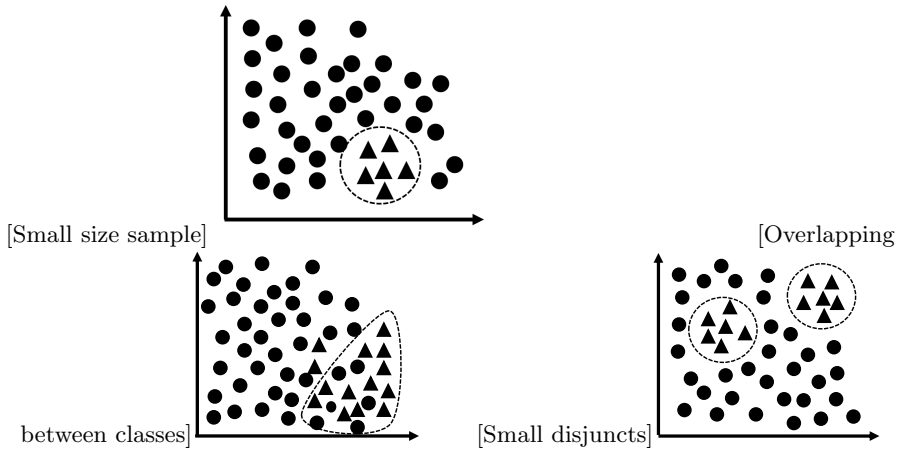

Fig. 3. Example of class distribution for two-class imbalance problems [38]

of overlapping between the classes with high imbalance ratio generally results in high misclassification rate for the minority class samples.

c) Small disjuncts: The presence of small disjuncts in a data-set occurs when the classes are constituted of smaller sub-concepts. The existence of small disjuncts also increases the complexity of the problem because of small fraction of the data instances, usually not balanced.

The rest of this paper is organized as follows: In section 2 we review the related work. Section 3 discuss the various challenges in handling skewed data distribution Section 4 presents the different evaluation metric for class imbalance problem domain. The proposed method is discussed in detail in Section 5. Section 6 discusses the experimental setup and results. Finally Section 7 presents the conclusion and discusses future scope.

2 Literature Survey

The most popular and common approach for balancing the skewed data is by using data level techniques. The data level solutions practice re-sampling method by either oversampling the minority class instances or under-sampling the majority class instances. At the algorithm level, a new algorithm or modification of existing algorithms are proposed to handle class imbalance problem. However, as data-level techniques, a learning algorithm cannot be changed once implemented. Ensemble based methods combines data-level techniques with algorithm level methods in solving the imbalanced datasets. As the scope of this paper focused on data-level techniques and for detail review on algorithm and ensemble technique, readers are suggested to refer the following papers [29] [34] [19] [20] [36] [18]. The class imbalance problem has attracted the research community and various data-level solutions have been proposed in literature. However, if the imbalance

dataset is linearly separable or sufficiently high, does not affect the results in spite of degree of imbalance. Recently few research studies showed that class overlap had a higher impact on classifier performance than skewed data distribution.

Thus, we broadly discuss the existing solutions for balancing the class distribution and class overlapping methods. The most popular and widely used data level technique is random resampling approach. It is based on undersampling the majority class instances or oversampling the minority class instances. However the main drawback of this two methods are undersampling may lead to loss of important information while oversampling may lead to overfitting.To substitute radom sampling methods, a new technique called SMOTE was introduced. This technique synthesis the minority samples based on linear interpolation using nearest neighbor concept. Various well-known extensions have been proposed such as Borderline-SMOTE [22] and SMOTE-IPF [39] [37], Safe-level-SMOTE [5] and DBSMOTE [6]. Other recent methods based on clustering [41] [33] and deep neural networks [25] have also been proposed.

As this paper deals with data-level techniques, a brief introduction to various data-level techniques are described as follow. In data-level approach, the sample dataset is modified to balance the class distribution. The foremost aim is to maintain equality in the class distribution for the datasets using sampling methods such as over-sampling, under-sampling and combination of both. The oversampling and under-sampling techniques are the two popular techniques in sampling-based classification to address the imbalanced datasets. In the oversampling technique, some samples are added to the minority class to make it balanced when very less information is available for minority class samples. In the under-sampling technique, some samples of the majority class are eliminated to make the dataset balanced. Apart from above, the hybrid techniques usually come with a combination of both over and under-sampling methods. Figure 4 different approaches applied at data-level to address the class imbalance problem.

These methods balanced the class distribution based on the original data. But, a common drawback is it effects by degree of imbalance. If the class imbalance is high, a drastic loss of information may encounter and if imbalance is low, overfitting of samples may be generated. Since this paper concentrates on class overlapping regions, we reviewed literature work related to the same. Class overlapping deals with the samples near the borderlines and can be extend far from the class boundaries. Few existing literature shows the solution in address the class overlapping problem.

In [42], the author proposed oversampling based undersampling technique, based on negative instance removal from the overlapping region. They stated that the proposed method provide significant improvements over the state-of-the-art class distribution methods. In [4] the author proposed DBMUTE based on density-based clustering methods to identify and remove the majority instances from the overlapping boundaries. Another well established method ADAptive

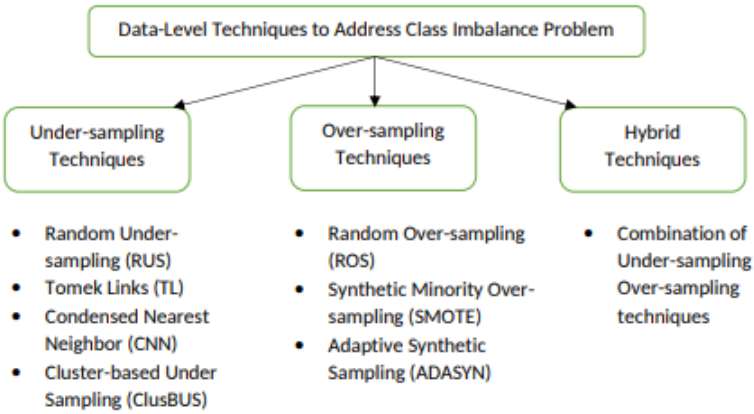

Fig. 4. Different Data-level Techniques Proposed for Handling Class Imbalance Problem [38]

SYNthetic sampling approach (ADASYN) [23], generates more minority samples surrounded by majority instances as its neighbours. Results showed a better sensitivity compared with other state-of-the-methods. However, the visibility of minority class were not sure by this method because the majority instances may still be present in the overlapping areas. Another methods called Edited Nearest Neighbour (ENN) [9], proposed to focus on boundary instances. It adopts k nearest neighbor (k=3) to remove majority class samples that lie in other class boundary. The author stated that setting of value k has significant impacts on the performance. The extension of ENN, Neighbourhood CLeaning rule (NCL) [27] considered both majority and minority k- nearest neighbours for discarding the majority samples and the results show a better performance over ENN. Later, combination of data cleaning and resampling approach has been proposed [39] such as SMOTE-IPF. In which noisy instances are removed before new samples are generated for minority class.

In [22] the author proposed BorderLine-SMOTE (BLSMOTE), to over sample the minority samples near the borderline. The author stated that their method behave better in terms of F-Measure compared to existing methods. Redundancy-driven modified Tomek-link based undersampling [11] to detect outlier, redundant and noisy instances having least contribution in estimating accurate class labels. Evolutionary undersampling [17], Majority Weighted Minority Oversampling TEchnique (MWMOTE) [3] works by identifying the minority class instances at boundary regions and assign weights based on the distance from majority class samples. Then, forms a cluster of these minority samples for generating synthetic data. Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) [32] consider minority samples closer to the boundary region and mark them as hard-to-learn samples. Those samples are not involved in generating new samples. Hence in this section, we discuss different data level

techniques proposed in literature for solving class imbalance problem. Now, next section will discuss about the different challenges in handling imbalance dataset/ skewed data distribution.

3 Challenges in Handling Skewed Data Distribution

Napierala and Stefanowski [31] proposed different method to analyze the minority samples by assigning it to predefined categories such as safe, borderline, rare and outliers. Such methods help in understanding the difficulties present in the data. Hence, some challenges are included here as:

a) As a future direction, it is important to propose new classification algorithms that incorporate the different difficulties in the data. Apart, while designing the classifier, attention should be needed towards individual minority samples. Another important issue is extreme class imbalance problems. The extreme imbalanced data sets exist in most of the real-world problems such as fraud detection with Imbalance Ratio (IR) approximately 1:3000. This poses a great challenge for classification algorithms to train on such extreme datasets. Third challenge is inefficient features extraction for some problems such as protein data, online transaction data. It is very much important for the classifier to be trained on such high- dimensional and sparse feature set.
b) Another way to solve class imbalance problem is by modifying the learning algorithm. However, a major drawback of such learning models gives much importance to minority samples, thus increasing the majority class misclassification. A technique needs to be proposed to select only uncertain samples and adjust the output accordingly.
c) Recently, ensemble learning became the most popular techniques for handling class imbalance. Algorithms like Bagging, Boosting, Stacking, and Random Forests were robust in handling data difficulties. Ensemble learning along with sampling techniques provides better performance to handle the skewed distribution of data. The main drawback is diversity among majority and minority class. There is no proper indication of how large the ensemble should be constructed as their size is selected arbitrarily.
d) Another problem is handling of multi-class imbalanced classification. The multi-class imbalance occurs when more than two classes with one majority class and multiple minority class exist. A deeper insight in handling multi-class imbalanced problems is needed.
e) Data pre-processing technique is highly important in balancing the imbalance datasets as there are independent of classifiers. The possible difficulties appear in the data are class overlapping, noise and small disjunct. Therefore, efficient data cleaning and sampling techniques are needed to balance the data. For multi class imbalance problems, efficient sampling techniques need to be proposed.
f) Multi class imbalance learning need special care while applying sampling techniques. Researchers should focus on developing algorithms which are robust in handling such skew distributions.

g) In ensemble learning algorithms such as bagging and boosting, there may be different level of uncertainty while sampling the data into bags. There may be a high probability of consisting samples from the same class within a single bag. The need for proper probability distribution techniques such as normal, binomial distribution can be used to check the balance distribution in each bag.
However, many difficulties may arise due to data distribution in each of the bags and also each bag may contain certain amount of noise which makes the classifier to perform poorly. So, efficient techniques need to be proposed in handling the size of the bags and the distribution of samples into each bags.

h) Another important and yet popular challenge in class imbalance is learning from continuous data. The process of learning from the continuous data is called data streaming. The need for active learning algorithms to address data streaming issue is still at infant. The general open issue will be based on sampling the streaming data and classifying it.

i) In last, extraction of efficient features and instance is also of major concern. Real time data such as bank data or genomic data are essentially have high-dimensional and sparse feature. The development of new approaches for high dimensional data is much needed, that will allow at the same time for an efficient processing and boosting discrimination of the minority class. Another interesting direction is to investigate the possibilities of using decomposition based solutions.
Hence in this section, we briefly discussed about the different challenges in handling imbalance dataset/ skewed data distribution. Next section presents the evaluation metric used in evaluation of classifier when trained on imbalanced dataset.

4 Evaluation Metrics in Skewed Data Distribution Domain

Most of the studies in skewed data distribution domain mainly concentrate on binary classification problem. By convention, positive class labels are considered as minority class and negative class labels as majority class. Table 1 illustrates

Table 1. Confusion Matrix

	Positive Prediction	Negative Prediction
Positive class	TP	FN
Negative class	FP	TN

a confusion matrix of a binary-class problem. TP and TN denote the number of

positive and negative examples that are correctly classified, while FN and FP denote the number of misclassified positive and negative examples respectively.

Accuracy is a well-known performance metric used in classification. It is defined as the ratio between the correctly classified samples to the total number of samples (1). In the imbalanced datasets, accuracy shows bias towards majority class and lead to wrong decisions. Therefore, different performance metrics are need to assess the performance of the classifier when trained on imbalanced datasets. The suitable metrics used are precision, recall, Area Under Curve (AUC) to measure the performance of classifier when trained on imbalanced datasets.

Precision is the proportion of true positive to the total number of true positive and false positive samples (2). Recall/sensitivity/True Positive Rate (TPR) represents how well the model detects the true positive samples (3). The F-Score/F-measure combines both recall and precision and defined as (4). Therefore, F-measure is suitable in imbalanced scenarios than the accuracy metric.

$$Accuracy = (TP + TN) \,/\, (TP + TN + FP + FN) \tag{1}$$

$$Precision = (TP) \,/\, (TP + FP) \tag{2}$$

$$Recall = (TP) \,/\, (TP + FN) \tag{3}$$

$$F - Score = 2 * (Precision * Recall) \,/\, (Precision + Recall) \tag{4}$$

In this paper, the various performance metrics used are AUC, F-Score and G-Mean.

5 Proposed method

This section describes the proposed method in detail. In Figure 1 we showed how class overlap makes the classification algorithms difficult with skewed data distribution. It affects the performance of the classifier when trained on highly overlapped imbalanced datasets. So, to overcome that we propose a framework for removing majority samples from the boundary regions and provide maximum visibility for minority samples. By using the proposed approach, we assume that most of the majority samples from the overlapped region are removed from the dataset. The two fold advantages of our approach is firstly the visibility of positive/minority samples increases in the complete dataset and next, more specific overlapped majority instances are identified using k-nearest neighbor which avoid unnecessary information loss.

The main contribution of this paper is to propose a framework for handling overlapping data in the decision boundary of a skewed data distribution. This is implemented by removing the majority samples that are most nearer to that of the minority class samples. The nearest neighbors of minority samples are computed based on k-nearest neighbor algorithm. The value setting for K is vital in identifying the samples to be discarded. Here, we empirically set the

value of k by considering the imbalance ratio with that of the size of the dataset. So Equation 5 shows the computation for k.

$$K = \sqrt{N} \times \sqrt{ImbRatio} \tag{5}$$

Where, N is the number of samples in the dataset and ImbRatio is the imbalance ratio i.e, proportion of majority samples towards minority samples. Unlike existing methods, k values is defined based the real world datasets and its imbalance ratio rather manual assignment. In this paper, we proposed a boundary region based undersampling method as mentioned in algorithm 1. The method vary with the existing algorithms in terms of identification and elimination of majority samples which are overlap with minority samples. As class imbalance problems is not a problem by itself but existence of overlap classes, sample disjunct and small sample size which in turn make the classifier complicated to perform better.

In this work, we concentrate on the boundary regions and identify the majority samples which are in the overlapping class region. This process eliminate the samples with out disturbing the data and may not lead to loss of information. This method showed a better accuracy for minority class samples as will be discussed in section 5. The algorithm works by identifying and eliminating the majority class samples from the boundary regions. The undersampled data are used for training the classification algorithm.

Hence in this section, we discussed the solution to handle the class overlapping problem in imbalance dataset. Next section presents the experimental results of the proposed method trained on 15 real world datasets.

Algorithm 1: Critical Instances removal based Under-Sampling (CIRUS)

Data: Training set N, K
Result: Removal of overlapped majority samples
begin:
T: training data
T_{pos}: positive or minority instances
T_{neg}: negative or majority instances
For each instance in minority class computes its nearest neighbours based on K (as defined in Eq(5)).
Remove the majority instances that are nearest to most of the minority samples (the samples with more than 2 neighbour) from the majority class.
After removal of overlapped samples combine T_{pos} and T_{neg} samples as final undersampled dataset D*.

6 Experimental setup

Extensive experiments using 15 public real-world datasets were carried out for evaluation. Experimental results were compared with state-of-the-art namely, SMOTE [7], Clustering-based undersampling [28], BLSMOTE [22] and ENN [9]. Support Vector Machine (SVM), Decision Tree (J48) and Random Forest (RF) were chosen as the learning algorithms because of widely used algorithms for skewed data distribution.

6.1 Data Set

We evaluate the proposed algorithm using 15 datasets from Keel repository with different imbalance ratio (IR) [1]. Table 2 shows the details of the imbalanced datasets with number of features and imbalance ratio.

Table 2. Datasets Used

Dataset	**features**	**Sample Size**	**Minority sample size**	**IR%**
Wisconsin	9	683	239	1.86
Pima	8	768	268	1.87
Glass0	9	214	70	2.06
Vehicle1	18	846	217	2.9
Ecoli1	7	336	77	3.36
New-thyroid2	5	215	35	5.14
Segmemt0	19	2308	329	6.02
Yeast3	8	1484	163	8.1
Vowel0	13	988	90	9.98
Yeast1vs7	7	459	30	14.3
Page-blocks13vs2	10	472	28	15.86
Abalone09-18	8	731	42	16.4
Yeast4	8	1484	51	28.1
Ecoli0137vs26	7	281	7	39.14
Yeast6	8	1484	35	41.4

6.2 Results

In this section, we compared the results using different state-of-the-art techniques namely, SMOTE, clustering-based undersampling, BLSMOTE and ENN on different classification algorithms such as SVM, J48 and RF. The metric used are AUC, F-Score and G-Mean. Table 3-11 shows the results of different state-of-the-art techniques compared with our proposed method on different classification algorithms. The experiments are carried out using 3 classification algorithms.

From the experimental results we observe that the performance of the proposed method is consistent across different algorithms and datasets. Figure 5-7

Table 3. The AUC performance measure on different datasets using SVM

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	96.66	96.66	96.66	96.66	96.66
Pima	60.02	61.43	64.6	67.8	61.43
Glass0	68.14	64.29	74.23	80.34	64.29
Vehicle1	48.93	51.54	55.91	70.4	51.54
Ecoli1	70.11	80.85	80.85	84.95	77.46
New-thyroid2	100	92.58	98.6	95.74	98.6
Segmemt0	97.54	98.85	97.67	97.95	97.67
Yeast3	67.68	70.04	74.14	90.11	69.9
Vowel0	62.9	54.77	70.71	88.47	63.25
Yeast1vs7	84.98	94.28	94.28	94.28	94.28
Page-blocks13vs2	62.9	62.9	62.9	64.17	62.9
Abalone09-18	100	89.44	99.43	97.12	99.43
Yeast4	35.1	66.52	66.52	66.52	66.52
Ecoli0137vs26	100	100	100	96.23	100
Yeast6	75.46	37.67	53.27	74.8	37.8

Table 4. The F-Score performance measure on different datasets using SVM

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	91.02	91.02	91.06	91.06	91.06
Pima	57.77	55.73	57.42	50.28	55.73
Glass0	77.29	85.38	86.61	68.14	85.37
Vehicle1	67.96	66.73	54.65	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	0	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

shows the comparison based on AUC on the proposed method with state-of-the-methods using three different classifiers.

It is clear from the experimental results that our proposed method is better than SMOTE, clustering-based undersampling, BLSMOTE and ENN for most of the datasets. CIRUS produces better performance in finding and eliminating the overlapping majority samples from the boundary region.The proposed model produces better performance using F-Score, AUC for majority datasets with

Table 5. The G-Mean performance measure on different datasets using SVM

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	92.06	92.06	92.06	92.06	92.06
Pima	57.77	55.83	57.42	50.28	55.83
Glass0	77.29	85.37	88.61	68.14	85.37
Vehicle1	57.96	66.73	53.64	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	37.92	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

Table 6. The AUC performance measure on different datasets using J48

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	96.66	96.66	96.66	96.66	96.66
Pima	60.02	61.43	64.6	67.8	61.43
Glass0	68.14	64.29	74.23	80.34	64.29
Vehicle1	48.93	51.54	55.91	70.4	51.54
Ecoli1	70.11	80.85	80.85	84.95	77.46
New-thyroid2	100	92.58	98.6	95.74	98.6
Segmemt0	97.54	98.85	97.67	97.95	97.67
Yeast3	67.68	70.04	74.14	90.11	69.9
Vowel0	62.9	54.77	70.71	88.47	63.25
Yeast1vs7	84.98	94.28	94.28	94.28	94.28
Page-blocks13vs2	62.9	62.9	62.9	64.17	62.9
Abalone09-18	100	89.44	99.43	97.12	99.43
Yeast4	35.1	66.52	66.52	66.52	66.52
Ecoli0137vs26	100	100	100	96.23	100
Yeast6	75.46	37.67	53.27	74.8	37.8

SVM, Random Forest. From the experiments, we conclude that the proposed method is superior to state-of-the-methods on most of the datasets.

7 Conclusion

In this paper, we proposed a novel framework for undersampling the critical majority instances from the boundary regions. The proposed CIRUS method

Table 7. The F-Score performance measure on different datasets using J48

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	92.06	92.06	92.06	92.06	92.06
Pima	57.77	55.83	57.42	50.28	55.83
Glass0	77.29	85.37	88.61	68.14	85.37
Vehicle1	57.96	66.73	53.64	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	0	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

Table 8. The G-Mean performance measure on different datasets using J48

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	92.06	92.06	92.06	92.06	92.06
Pima	57.77	55.83	57.42	50.28	55.83
Glass0	77.29	85.37	88.61	68.14	85.37
Vehicle1	57.96	66.73	53.64	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	37.92	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

effectively identified and removed the majority instances in the boundary regions. Extensive experiments using real-world datasets were carried out. The proposed methods were compared against state-of-the-art methods with good performance. This method can be applied to imbalanced datasets with any classification algorithm in general.

Table 9. The AUC performance measure on different datasets using RF

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	96.66	96.66	96.66	96.66	96.66
Pima	60.02	61.43	64.6	67.8	61.43
Glass0	68.14	64.29	74.23	80.34	64.29
Vehicle1	48.93	51.54	55.91	70.4	51.54
Ecoli1	70.11	80.85	80.85	84.95	77.46
New-thyroid2	100	92.58	98.6	95.74	98.6
Segmemt0	97.54	98.85	97.67	97.95	97.67
Yeast3	67.68	70.04	74.14	90.11	69.9
Vowel0	62.9	54.77	70.71	88.47	63.25
Yeast1vs7	84.98	94.28	94.28	94.28	94.28
Page-blocks13vs2	62.9	62.9	62.9	64.17	62.9
Abalone09-18	100	89.44	99.43	97.12	99.43
Yeast4	35.1	66.52	66.52	66.52	66.52
Ecoli0137vs26	100	100	100	96.23	100
Yeast6	75.46	37.67	53.27	74.8	37.8

Table 10. The F-Score performance measure on different datasets using RF

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	92.06	92.06	92.06	92.06	92.06
Pima	57.77	55.83	57.42	50.28	55.83
Glass0	77.29	85.37	88.61	68.14	85.37
Vehicle1	57.96	66.73	53.64	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	0	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

Table 11. The G-Mean performance measure on different datasets using RF

Dataset	SMOTE	BLSMOTE	ENN	Cluster-based Under-sampling	CIRUS
Wisconsin	92.06	92.06	92.06	92.06	92.06
Pima	57.77	55.83	57.42	50.28	55.83
Glass0	77.29	85.37	88.61	68.14	85.37
Vehicle1	57.96	66.73	53.64	41.45	66.73
Ecoli1	66.91	91	91	70.81	100
New-thyroid2	100	100	87.5	70	87.5
Segmemt0	98.43	95.57	100	94.09	100
Yeast3	71.79	76.51	75.33	60	73.08
Vowel0	100	100	100	100	100
Yeast1vs7	16.54	16.54	16.54	16.54	16.54
Page-blocks13vs2	100	100	84.73	52.6	84.73
Abalone09-18	34.29	34.29	34.29	11.54	34.29
Yeast4	37.92	37.92	37.92	13.85	100
Ecoli0137vs26	100	100	100	25.64	100
Yeast6	79.96	33.27	49.93	39.93	100

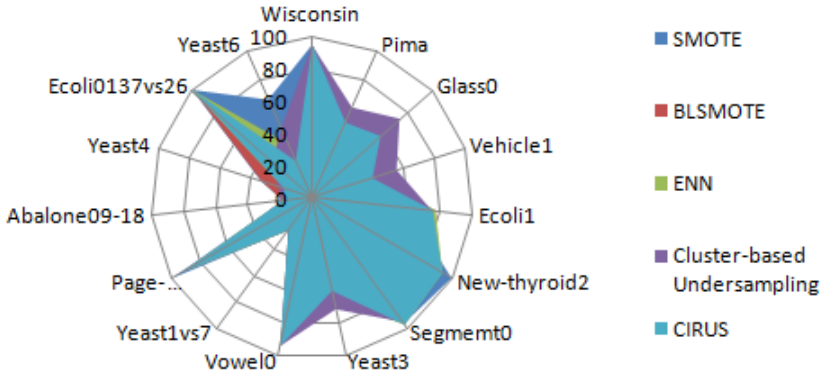

Fig. 5. Comparison of the proposed method with state-of-the-methods using J48 classifier (AUC metric)

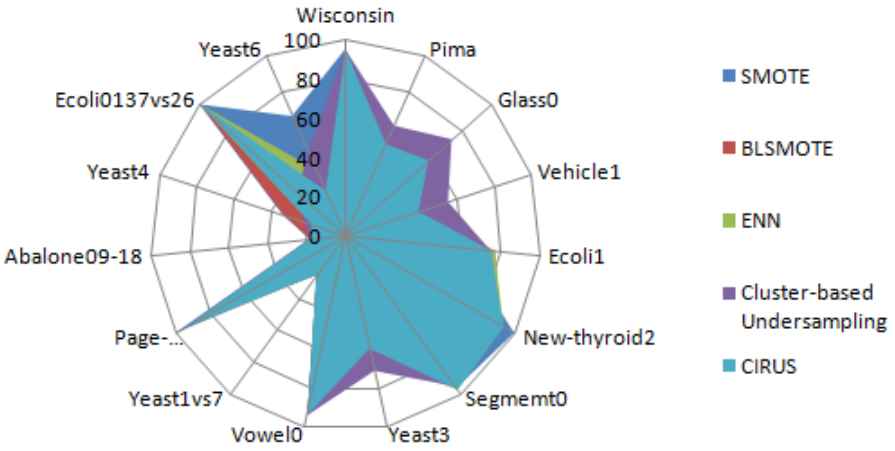

Fig. 6. Comparison of the proposed method with state-of-the-methods using SVM classifier (AUC metric)

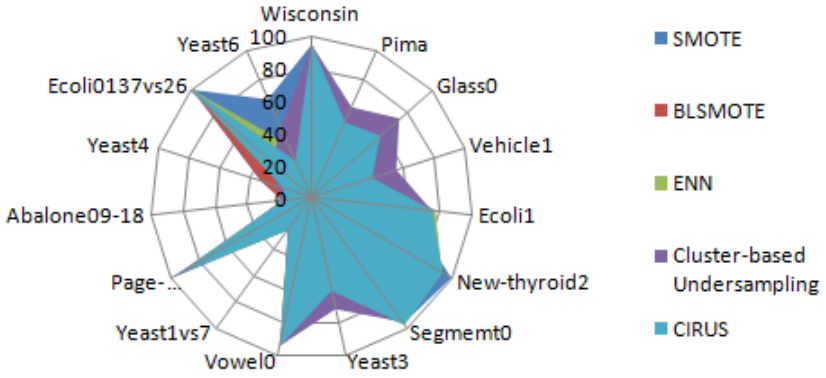

Fig. 7. Comparison of the proposed method with state-of-the-methods using Random Forest classifier (AUC metric)

References

1. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic & Soft Computing **17** (2011)
2. Aşkan, A., Sayın, S.: Svm classification for imbalanced data sets using a multiobjective optimization framework. Annals of Operations Research **216**(1), 191–203 (2014)
3. Barua, S., Islam, M.M., Yao, X., Murase, K.: Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on knowledge and data engineering **26**(2), 405–425 (2012)
4. Bunkhumpornpat, C., Sinapiromsaran, K.: Dbmute: density-based majority undersampling technique. Knowledge and Information Systems **50**(3), 827–850 (2017)
5. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 475–482. Springer (2009)
6. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Dbsmote: density-based synthetic minority over-sampling technique. Applied Intelligence **36**(3), 664–684 (2012)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
8. Chawla, N.V., Japkowicz, N., Kotcz, A.: Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter **6**(1), 1–6 (2004)
9. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers. Multiple Classifier Systems **34**(8), 1–17 (2007)
10. Das, S., Datta, S., Chaudhuri, B.B.: Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition **81**, 674–693 (2018)
11. Devi, D., Purkayastha, B., et al.: Redundancy-driven modified tomek-link based undersampling: a solution to class imbalance. Pattern Recognition Letters **93**, 3–12 (2017)
12. Douzas, G., Bacao, F., Last, F.: Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Information Sciences **465**, 1–20 (2018)
13. Ezawa, K.J., Singh, M., Norton, S.W.: Learning goal oriented bayesian networks for telecommunications risk management. In: Proceedings of the International Conference on Machine Learning. pp. 139–147 (1996)
14. Fawcett, T., Provost, F.J.: Combining data mining and machine learning for effective user profiling. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. pp. 8–13 (1996)
15. FernáNdez, A., LóPez, V., Galar, M., Del Jesus, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems **42**, 97–110 (2013)
16. Freitas, A., Costa-Pereira, A., Brazdil, P.: Cost-sensitive decision trees applied to medical data. In: International Conference on Data Warehousing and Knowledge Discovery. pp. 303–312. Springer (2007)

17. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation **17**(3), 275–306 (2009)
18. Gillala Rekha, V Krishna Reddy, A.K.T.: Chaotic salp swarm optimization using svm for class imbalance problems. In: 19th International Conference on Hybrid Intelligent Systems (HIS 2019). Springer (2019)
19. Gillala Rekha, V Krishna Reddy, A.K.T.: A novel approach for solving skewed classification problem using cluster based ensemble method. Mathematical Foundations of Computing (2020)
20. Gong, J., Kim, H.: Rhsboost: Improving classification performance in imbalance data. Computational Statistics & Data Analysis **111**, 1–13 (2017)
21. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
22. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887. Springer (2005)
23. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. IEEE (2008)
24. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis **6**(5), 429–449 (2002)
25. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 27 (2019)
26. Kubat, M., Holte, R.C., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. Machine learning **30**(2-3), 195–215 (1998)
27. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Conference on Artificial Intelligence in Medicine in Europe. pp. 63–66. Springer (2001)
28. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Information Sciences **409**, 17–26 (2017)
29. López, V., Del Río, S., Benítez, J.M., Herrera, F.: Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. Fuzzy Sets and Systems **258**, 5–38 (2015)
30. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks **21**(2-3), 427–436 (2008)
31. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems **46**(3), 563–597 (2016)
32. Nekooeimehr, I., Lai-Yuen, S.K.: Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. Expert Systems with Applications **46**, 405–416 (2016)
33. Onan, A.: Consensus clustering-based undersampling approach to imbalanced learning. Scientific Programming **2019** (2019)
34. Patel, H., Thakur, G.S.: Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. International Journal of Intelligent Engineering and Systems **10**(1), 56–64 (2017)

35. Prati, R.C., Batista, G.E., Monard, M.C.: Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence. pp. 312–321. Springer (2004)
36. Rekha, G., Tyagi, A.K.: Necessary information to know to solve class imbalance problem: From a user's perspective. In: Proceedings of ICRIC 2019, pp. 645–658. Springer (2020)
37. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. International Journal of Hybrid Intelligent Systems (Preprint), 1–10 (2019)
38. Rekha, G., Tyagi, A.K., Krishna Reddy, V.: A wide scale classification of class imbalance problem and its solutions: A systematic literature review. Journal of Computer Science **15**, 886–929 (2019)
39. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences **291**, 184–203 (2015)
40. Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., Zhou, Y.: A novel ensemble method for classifying imbalanced data. Pattern Recognition **48**(5), 1623–1637 (2015)
41. Tsai, C.F., Lin, W.C., Hu, Y.H., Yao, G.T.: Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Information Sciences **477**, 47–54 (2019)
42. Vuttipittayamongkol, P., Elyan, E., Petrovski, A., Jayne, C.: Overlap-based undersampling for improving imbalanced data classification. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 689–697. Springer (2018)