INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Hybrid Data Resampling Algorithm Combining Leader and SMOTE for Classifying the High Imbalanced Datasets

S Karthikeyan[1]*, T Kathirvalavakumar[2]

1 Research Scholar, Research centre in Computer Science, V.H.N. Senthikumara Nadar College, Virudhunagar, Tamil Nadu, India
2 Associate Professor, Research centre in Computer Science, V.H.N. Senthikumara Nadar College, Virudhunagar, Tamil Nadu, India

OPEN ACCESS

Received: 20-01-2023
Accepted: 29-03-2023
Published: 24-04-2023

Citation: Karthikeyan S, Kathirvalavakumar T (2023) A Hybrid Data Resampling Algorithm Combining Leader and SMOTE for Classifying the High Imbalanced Datasets. Indian Journal of Science and Technology 16(16): 1214-1220. https://doi.org/10.17485/IJST/v16i16.146

* **Corresponding author**.

rgskarthi@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: The traditional classifiers are ineffective in classifying the imbalanced datasets. Most popular approach in resolving this problem is through data re-sampling. A hybrid resampling method is proposed in this paper that reduces the misclassification in all the classes. **Method:** The proposed method employs the Leader algorithm for under sampling and SMOTE algorithm for oversampling. It generates the desired number of samples in both the classes based on the problem that overcomes the over-fitting and under-fitting issues. **Findings:** To evaluate the performance of the proposed work, it is tested on 13 high imbalanced datasets obtained from the keel repository and the results are compared with the state-of-the-art hybrid data resampling methods such as SMOTE+Tomek Links, SMOTE+ENN, and SMOTE+RSB*. From the experiment it is observed that among the 13 high imbalanced datasets, the proposed method outperforms in 12 datasets and produces the same result in 1 dataset. The proposed method reduces the misclassification rates of minority and majority classes and is more suitable for the extreme imbalanced datasets. **Novelty:** This research work introduces a novel approach for classification by combining machine learning algorithms with domain-specific knowledge and resulting in significantly improved accuracy in classifying the extreme imbalanced datasets compared to the traditional methods. The uniqueness of the work is the utilization of the Leader algorithm and the SMOTE algorithm with a required resampling ratio instead of balancing and it improves the performance of the classification on the imbalanced data.

Keywords: Imbalanced Data; Leader; SMOTE; Hybrid Sampling; Resampling; Classification

1 Introduction

The class imbalance is becoming the most serious problem in machine learning applications. In general, the dataset with unequal class distribution is termed as an

imbalanced dataset. Class with more number of samples is called a majority class, and the class with less number of samples is termed a minority class. The ratio between the majority and the minority class is denoted as imbalance ratio which is given by the formula

$$Imbalance\ Ratio = \#\ of\ Majority\ Class\ Samples / \#\ of\ Minority\ Class\ Samples$$

Leevy et al. [1] have presented a survey that focuses on the high-class imbalance related to the data level and algorithm level methods and stated that data resampling methods are popular in addressing class imbalance and also find that Random oversampling method shows better results. They have concluded that algorithm level methods have some outstanding performers but there are inconsistent and conflicting results. Kraiem et al. [2] have conducted an analysis to identify the best resampling strategy that best fits for the dataset based on its characteristics. The authors have identified that the values of evaluation metrics and the statistical significance test results are used to identify some behaviors which were used to create models with association rules. They have concluded that the useful patterns involved in the models of high confidence rules select the best resampling strategy for every dataset.

Commonly used data resampling methods in the literature are categorized as oversampling, under-sampling, and hybrid sampling. Oversampling increases the samples of the minority class and under-sampling decreases the samples of the majority class but both methods have its own merits and demerits as well [3]. With random oversampling, instances are duplicated randomly from the available data, which creates more duplicates. In order to avoid duplicate samples, synthetic samples are generated using Synthetic Minority Oversampling Technique (SMOTE) [4]. This algorithm creates synthetic samples by selecting k-nearest neighbors among the samples. As this algorithm is extensible more variants are available and each one is designed for different purposes. Even though many enhanced versions of the SMOTE algorithm are available, the SMOTE algorithm itself is best suited for oversampling [5].

The issues with the oversampling and under sampling methods are eliminated when hybrid sampling approaches are used [6]. SMOTE+Tomek Links [7] balances the data distribution by oversampling using SMOTE but it creates an over fitting issue during the classification. This problem is rectified by removing the samples using the Tomek Links on both classes. SMOTE+ Edited Nearest Neighbor (ENN) [8] removes samples from both classes by considering the k-nearest neighbours among the data. But a problem with ENN is, it removes the samples which are necessary for modeling the data. SMOTE+RSB* [9] algorithm over samples the minority class samples by SMOTE. The similarity matrix is generated after re sampling the majority and minority class data. Based on the rough set theory, the similarity values lesser than the lower approximation threshold are used to form the final dataset. Wang et al. [10] improvise the SMOTE algorithm in such a way that the samples are generated in the center of the minority class data to prevent marginalization of the synthetic data. When inter-class distance and sample variance of the generated data are closer to the original data then it provides a high classification accuracy. Salunkhe and Mali [11] have presented a hybrid resampling model that uses SMOTE for oversampling and the samples in the borderline of the majority and minority classes are evaluated, and unwanted samples are removed. Further, the random under-sampling is applied on the majority class to make it balanced. Zhao et al. [12] have presented a weighted hybrid ensemble method that uses boosting algorithm to combine two data sampling methods and two base classifiers. Each classifier and each resampling method is assigned a weight that helps in producing better results. Liu and Hsieh [13] have proposed a model-based synthetic sampling to generate synthetic samples in a diversified manner. Prior to sample generation, the relationship between the data is identified through the regression models. Synthetic samples for each minority class sample are generated in the ratio 1:2, 1:3 and 1:5 and identify the ratio that gives better results.

It is observed from the literature that when oversampling or undersampling is performed on the imbalanced dataset, the misclassification rate of the majority or minority class gets increased. The proposed work combines the Leader algorithm for under-sampling and SMOTE algorithm for oversampling. It generates only the desired number of samples in both classes instead of balancing majority and minority classes. In this method, oversampling is performed if necessary. In this paper, section 2 describes the methodology, section 3 discusses the results and discussions, and section 4 draws a conclusion.

2 Methodology

Incremental clustering is an approach that addresses dynamically growing dataset. From the literature, it is found that the Leader algorithm [14] is a better choice for performing clustering in an incremental fashion [15]. Clusters are formed from the samples and the leaders of the clusters are considered for classification instead of considering all the samples in the training dataset. Considering only representatives of the clusters leads to having lesser training samples for classification and it is proven that it gives good classification accuracy. It shows that the representatives of the clusters are with significant attributes. Hence in the proposed work, the Leader algorithm is considered for undersampling the majority class data as it retains the samples

with significant attributes. Based on the requirement of the classification, cluster representatives are generated by changing the distance threshold of the Leader algorithm to get better accuracy. SMOTE algorithm is mostly used in the literature to generate synthetic samples for the minority class of the dataset. This algorithm uses interpolation for generating new unique samples from the minority class. It is popular in the literature because of its simplicity and extensibility. In the proposed work, SMOTE is used to oversample the minority class data when needed.

2.1 Leader Algorithm

The Leader algorithm works by keeping a random pattern as its initial leader. Subsequent patterns are compared against the existing leaders and if the distances between existing leaders are greater than the user-specified threshold then the newly arrived pattern becomes a leader of a new cluster. If the distance of the next pattern with any one of the existing leaders is less than the threshold then the pattern is included in the corresponding cluster that matches at the first instance. This process is repeated until all the existing patterns are clustered. In this algorithm, the number of samples going to be generated is not known to the user as the samples are generated dynamically based on the threshold value. If the generated samples are not enough to reduce the misclassification rate, then a new threshold value is to be selected to generate new samples. The correct choice of threshold helps in achieving better results.

2.1.1 Algorithm

1. Set a suitable threshold
2. Select any one of the training pattern as an initial leader for a cluster
3. Find the Euclidean distance for the next training pattern with the leaders one by one.
4. When the distance is less than the threshold then assign the training pattern to the corresponding cluster.
5. If the distance with all the existing leaders is greater than the threshold then consider the training pattern as a leader of a new cluster.
6. Repeat steps 3-5 for all the patterns in the training dataset

2.2 SMOTE Algorithm

1. Select a pattern from the minority class
2. For each attribute of the pattern,

    (a) Find Euclidean distance with corresponding attributes in other patterns.
    (b) Choose the smallest Euclidean distance.
    (c) Multiply the distance with a random number generated between 0 and 1.
    (d) Add the resultant with the corresponding attribute of the selected pattern.

3. The new values generated in step 2 for each attribute forms a synthetic pattern
4. Repeat the above steps for all the patterns in the minority class to increase the pattern size.

2.3 Proposed Algorithm

1. Generate clusters for the majority class dataset using Section 2 1 1
2. Form a new dataset with the leaders generated by step 1 and the existing minority class samples of size N.
3. Classify the new dataset using a classifier
4. Iterate the following steps if the misclassification rate is not minimized

    (a) Generate synthetic patterns for the initial minority class dataset using Section 2 2
    (b) Form a new dataset with the leaders generated by step 1 and all samples generated so far by Section 2.2. Now the size of the minority sample is N*(#iteration+1).
    (c) Classify the dataset using the classifier

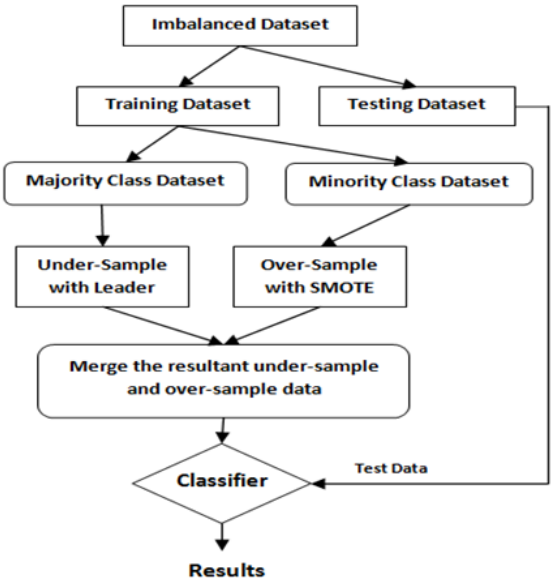

Fig 1. Experimental Setup

3 Results and Discussion

The extremely Imbalanced datasets for this experiment are collected from the Keel repository[16]. The selected datasets and their information are shown in Table 1. The imbalance ratio of the selected datasets ranges from 15 to 130. With an Imbalanced dataset, classification accuracy may not be the only deciding criteria to claim the efficiency of the work. The misclassification of the minority class data is not reflected much on the classification accuracy so it has been decided to use AUC value as an evaluation metric. AUC value is a measure which gives equal importance to the classes regardless of the class distribution. The classifier preferred for this experiment is C4.5 since it is less affected by class imbalance[17]. To demonstrate the working of the proposed method, the imbalanced dataset is divided into a training dataset with 80 percent of the majority class instances and 80 percent of the minority class instances, and the remaining instances forms the testing dataset.

Table 1. Dataset Information

Dataset	#Attributes	#Samples	Imbalance Ratio
Abalone9-18	8	731	16.4
Abalone19	8	4174	129.4
Ecoli0137vs26	7	281	39.14
Glass016vs5	9	184	19.44
Glass4	9	214	15.47
Glass5	9	214	22.78
Pageblocks13vs4	10	472	15.86
Shuttle-c2-vs-c4	9	129	20.5
Yeast1289vs7	8	947	30.57
Yeast1458vs7	8	693	22.10
Yeast2vs8	8	482	23.10
Yeast5	8	1484	32.73
Yeast6	8	1484	41.4

The information regarding the number of leaders obtained after under sampling and the selected threshold value for each dataset is given in Table 2. After executing the under sampling process, the minority class samples (N) are over-sampled to the size of 2xN, 3xN, 4xN, and 5xN step-by-step. Table 3 lists the number of misclassified instances of the majority and minority classes, the number of misclassified instances, and its AUC scores are given in Table 4. From the results it is observed that,

Table 2. Number of leaders generated and threshold value

Dataset	Selected Threshold Value	# Leaders
Abalone9-18	0.025	500
Abalone19	0.1	246
Ecoli0137vs26	15	186
Glass016vs5	0.5	65
Glass4	0.5	75
Glass5	0.5	74
Pageblocks13vs4	2	332
Shuttle-c2-vs-c4	5	68
Yeast1289vs7	0.160	124
Yeast1458vs7	0.150	160
Yeast2vs8	0.09	183
Yeast5	0.125	315
Yeast6	0.15	226

significance is not observed when oversampling the samples more than twice. It has been observed that no misclassification is found in the minority class of 7 datasets and the majority class of 4 datasets.

Table 3. Misclassified instances

Dataset	Misclassifications when the minority class size is				
	N	2xN	3 x N	4 x N	5 x N
Abalone9-18	6/138:5/8	7/138:5/8	9/138:6/8	10/138:3/8	8/138:4/8
Abalone19	59/828:3/6	50/828:4/6	75/828:4/6	67/828:4/6	97/828:4/6
Ecoli0137vs26	0/55:0/1	0/55:0/1	1/55:0/1	1/55:0/1	1/55:0/1
Glass016vs5	1/35:0/2	1/35:0/2	1/35:0/2	1/35:0/2	1/35:0/2
Glass4	3/40:1/3	3/40:1/3	1/40:1/3	0/40:1/3	0/40:0/3
Glass5	0/41:0/2	0/41:0/2	0/41:0/2	0/41:0/2	0/41:0/2
Pageblocks13vs4	0/89:2/6	0/89:1/6	0/89:0/6	0/89:0/6	0/89:0/6
Shuttle-c2-vs-c4	0/25:0/1	0/25:0/1	0/25:0/1	0/25:0/1	0/25:0/1
Yeast1289vs7	39/183:1/6	46/183:4/6	74/183:0/6	48/183:1/6	51/183:3/6
Yeast1458vs7	24/133:2/6	28/133:2/6	26/133:2/6	29/133:4/6	28/133:3/6
Yeast2vs8	2/92:2/4	2/92:1/4	3/92:2/4	7/92:0/4	8/92:1/4
Yeast5	2/288:0/9	3/288:3/9	4/288:3/9	2/288:3/9	4/288:3/9
Yeast6	22/290:3/7	10/290:1/7	21/290:**1/7**	19/290:**1/7**	18/290:**1/7**

It is observed from the datasets that the performance degradation of a classifier happens in two situations. In the first case, the feature values of a dataset are the combination of integer and real values or having more number of zero elements. In the second case, the distances between the patterns are large. The AUC scores of existing hybrid re-sampling approaches namely SMOTE+Tomek, SMOTE+ENN, and SMOTE+RSB* are compared with the proposed method and are shown in Figure 2. The results show that the proposed method is better than other methods in 12 datasets and produces same result in 1 dataset.

The methods considered for comparison generates an equal number of samples in both major and minor classes for classification. But with the proposed method it is obvious that the extremely imbalanced datasets need not be converted into a balanced form for classification, instead, it is necessary to generate the required number of samples. The oversampling through SMOTE algorithm along with the Leader algorithm for undersampling helps in achieving the better performance of the classifier.

Table 4. AUC score

Dataset	AUC score when the minority class size is				
	N	2 x N	3 x N	4 x N	5 x N
Abalone9-18	0.666	0.662	0.592	0.776	0.721
Abalone19	0.714	0.636	0.621	0.626	0.608
Ecoli0137vs26	1	1	1	1	1
Glass016vs5	0.986	0.986	0.986	0.986	0.986
Glass4	0.796	0.796	0.821	0.833	1
Glass5	1	1	1	1	1
Pageblocks13vs4	0.833	0.917	1	1	1
Shuttle-c2-vs-c4	1	1	1	1	1
Yeast1289vs7	0.810	0.541	0.798	0.786	0.611
Yeast1458vs7	0.743	0.728	0.736	0.558	0.645
Yeast2vs8	0.739	0.864	0.734	0.962	0.832
Yeast5	0.997	0.828	0.826	0.830	0.882
Yeast6	0.748	0.911	0.892	0.896	0.898

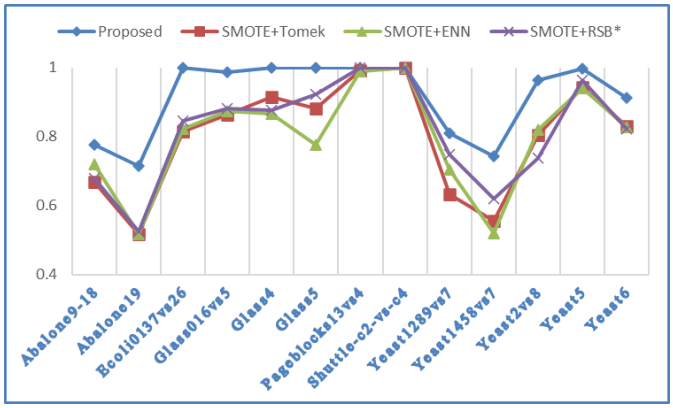

Fig 2. Comparison of AUC score of different hybrid sampling methods

4 Conclusion

This study highlights the importance of handling the imbalanced class distribution problem in classification tasks, and the importance of data resampling in improving the classification performance on the imbalanced datasets. By carefully selecting and combining different approaches, handling of challenges posed by imbalanced data is improved. The hybrid re-sampling method proposed in our work gives better classification results and reduced misclassification rate in both major and minor classes than the existing SMOTE-based hybrid sampling approaches in 12 datasets out of the experimented 13 datasets. The number of samples generated through the SMOTE algorithm need not be in a balanced form. The re-sampling procedure adopted in the proposed work is having the ability to solve problems with an extreme imbalance ratio. In this work, the proposed procedure is applied in the two class classification problem. In the future the procedure has to be identified for the multi class problems with imbalanced dataset to achieve better results.

References

1) Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. *Journal of Big Data*. 2018;5(1):1–30. Available from: https://doi.org/10.1186/s40537-018-0151-6.
2) Kraiem MS, Sánchez-Hernández F, Moreno-García MN. Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models. *Applied Sciences*. 2021;11(18):8546. Available from: http://dx.doi.org/10.3390/app11188546.
3) Wongvorachan T, He S, Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*. 2023;14(1):54. Available from: https://doi.org/10.3390/info14010054.

4) Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357. Available from: https://doi.org/10.1613/jair.953.
5) Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*. 2013;14(1):1–6. Available from: https://doi.org/10.1186/1471-2105-14-106.
6) Xu Z, Shen D, Nie T, Kou Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*. 2020;107:103465. Available from: https://doi.org/10.1016/j.jbi.2020.103465.
7) Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*. 2004;6(1):20–29. Available from: https://doi.org/10.1145/1007730.1007735.
8) Wilson DL. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*. 1972;SMC-2(3):408–421. Available from: https://doi.org/10.1109/TSMC.1972.4309137.
9) Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*. 2012;33(2):245–265. Available from: https://doi.org/10.1007/s10115-011-0465-6.
10) Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*. 2021;11(1):1–1. Available from: https://doi.org/10.1038/s41598-021-03430-5.
11) Salunkhe UR, Mali SN. A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling. *International Journal of Intelligent Systems and Applications*. 2018;10(5):71–81. Available from: https://doi.org/10.5815/ijisa.2018.05.08.
12) Zhao J, Jin J, Chen SJ, Zhang R, Yu B, Liu Q. A weighted hybrid ensemble method for classifying imbalanced data. *Knowledge-Based Systems*. 2020;203:106087. Available from: https://doi.org/10.1016/j.knosys.2020.106087.
13) Liu CLL, Hsieh PYY. Model-Based Synthetic Sampling for Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. 2020;32(8):1543–1556. Available from: https://doi.org/10.1109/TKDE.2019.2905559.
14) Hartigan J. Wiley Series in Probability and Mathematical Statistics. 1975.
15) Vijaya PA, Murty MN, Subramanian DK. An efficient incremental protein sequence clustering algorithm. *TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region*. 2003;1:409–413. Available from: https://doi.org/10.1109/TENCON.2003.1273355.
16) Imbalanced Classification Datasets. . Available from: https://sci2s.ugr.es/keel/datasets.php.
17) Mahmudah KR, Indriani F, Takemori-Sakai Y, Iwata Y, Wada T, Satou K. Classification of Imbalanced Data Represented as Binary Features. *Applied Sciences*. 2021;11(17):7825. Available from: https://doi.org/10.3390/app11177825.