Combination of Oversampling and Undersampling Techniques on Imbalanced Datasets

Ankita Bansal, Ayush Verma, Sarabjot Singh, and Yashonam Jain

Abstract Many practical classification datasets are unbalanced, meaning that one of the classes is in the majority when compared to the others. In various real-world circumstances, class-imbalanced datasets arise, where the number of data samples in a class is not equal to the other class. To develop good classification models based on present level calculations, using these datasets is difficult, particularly for separating minority classes from the majority class. To address the issue of class imbalance, under/oversampling procedures are used to minimize and enhance the quantities of data examined in minority and majority class. This paper explores the utilization of combination of both undersampling and oversampling techniques mainly synthetic minority oversampling technique (SMOTE) and neighborhood cleaning rule (NCL) to balance the datasets. The performance has been evaluated using two machine learning algorithms. The results are then classified using recall measure and geometric mean which showed improved performance of the algorithms.

Keywords Class imbalance problem · Undersampling · Oversampling · SMOTE · NCL · Minority class · Majority class · Binary classification

A. Bansal (✉) · A. Verma · S. Singh · Y. Jain
Information Technology Department, Netaji Subhas University of Technology, Dwarka Sector-3, Delhi, India
e-mail: ankita.bansal06@gmail.com

A. Verma
e-mail: ayushv.it18@nsut.ac.in

S. Singh
e-mail: sarabjots.it18@nsut.ac.in

Y. Jain
e-mail: yashonamj.it18@nsut.ac.in

D. Gupta et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 492,
https://doi.org/10.1007/978-981-19-3679-1_55

1 Introduction

Imbalanced data refers to datasets in which the target class has an unequal distribution of observations; i.e., one class label has a large number of observations while the other has a small number. A large number of experiments in the domain of imbalance data on the behavior of a few standard classifiers have revealed that imbalance probabilities, which are defined as the ratio of the number of examples in the majority class to the number of models in the minority class, has the potential to produce misclassification since the majority data is more dominating on the minority data, resulting in a loss of accuracy [1]. With imbalanced datasets, standard characterization learning calculations are frequently one-sided toward the majority class (known as the "negative" class), resulting in a higher misclassification rate for minority class occurrences (known as "positive" class) [2]. One form of best-in-class method for creating a new dataset is sampling (or resampling) techniques. Undersampling can be defined as removing some observations of the majority class. Oversampling can be defined as adding more data samples to the minority class. The advantages of combining both the techniques are improved run time by reducing the number of training data samples, reduced storage problems when the training data set is huge, reduced data redundancy leading to perfect balancing of datasets. Most commonly used oversampling technique is synthetic minority oversampling technique (SMOTE) [3–5], and neighborhood cleaning rule (NCL) is commonly used undersampling technique [6–8]. In this study, we have suggested few modifications in original SMOTE and NCL to improve the performance. Then, we have used the combination of SMOTE and NCL together to different classification problems. The performance of the sampling approaches has been validated using linear support vector classifier and K-nearest neighbor. The results show the superior performance of the algorithms on the sampled/balanced datasets as compared to the imbalanced datasets.

The paper is organized as: Next section briefly states the literature review. Following this, next section explains the datasets used in this study. Proposed work is elaborated in the next section. Followed by the result evaluation. Finally, the work is concluded in the last section.

2 Literature Review

Cluster-based sampling approaches were created to minimize the amount of data samples in the majority class [9]. Cluster-based sampling approaches, in general, work by clustering a number of clusters from a given majority class dataset, then selecting a number of representative data samples from each of the clusters. However, cluster-based sampling approaches have significant drawbacks that have a direct impact on the decreased majority class dataset and ultimate classification performance. Before performing any synthetic sampling, [10] used the K-means clustering approach to cope with the noisy situation. They also employed the SMOTE to oversample clusters. On the sample dataset, the LR and ANN were employed to assess classification performance. The influence of overlapping was investigated in conjunction with other factors in [11], such as the breaking of minority class into sub classes which are small in size. The studies were conducted on two dummy linear datasets with more intricate nonlinear boundaries, and the findings revealed that class breakdown combined with overlapping makes learning extremely difficult. Evaluating methods for recognizing noisy samples is likewise extremely complicated [12]. The most often misclassified samples are evaluated as probable noise and deleted from the learning data progressively until a particular level of accuracy is reached. These approaches are dependent on a number of characteristics as well as the type of base classifier used. Furthermore, deleting the instances may be controversial, particularly among the minority class.

3 Empirical Data Collection

In this study, we have used four open source datasets from different medical fields. The independent variables in all these datasets are different depending on the type of dataset. These are listed in Table 1. The dependent variable is a binary variable whose value is 0 if the disease is not present and 1 if the disease is present. All these datasets are imbalanced; i.e., the data belonging to one of classes is much more as compared to the data belonging to the other class. The amount of imbalance can be well defined by using imbalance ratio measure. Imbalance ratio is the ratio of size of minority class to the size of majority class. The details of the datasets including the name, number of attributes, total number of samples and imbalance ratio is specified in Table 1.

4 Proposed Work

The work in this paper has been carried out in three broad steps. Each of the steps has been explained in this section.

Step 1: The authors have proposed modifications in the original SMOTE [5] and NCL [7] algorithms of sampling. These modified algorithms are termed as $\text{SMOTE}_{\text{MODIFIED}}$ and $\text{NCL}_{\text{MODIFIED}}$ from here on. These proposed algorithms are applied on all the datasets to balance them. The algorithms are explained as below.

$\textbf{SMOTE}_{\textbf{MODIFIED}}$

Synthetic minority oversampling technique (SMOTE) is a technique in which new synthetic data is added to the minority class. To do this, a data point and its four neighboring points in minority class are considered. These are connected by lines, and one of the points of intersection if these lines are chosen as new data.

Table 1 Datasets used in the study

Dataset name	Total data samples	Independent variables	Imbalance ratio	Attributes	Source
Thyroid	215	5	8.6:1	Age, gender, DOB, survival status, total serum thyroxin, T3 level	[13]
Abalone	4177	8	9.7:1	NA	[13]
Arrhythmia	452	10	17:1	Age, height, weight, QRS duration, Pinterval, heartrate, QRSA, QRSTA	[14]
Mammography	11,183	5	42:1	Age, density, biophx, famhx, ptid	[15]

Pinterval is average duration of P wave in msec
Heart rate is number of heart beats per minute

$$Z_{i\text{new}} = Z_m + (Z_n - Z_m)\lambda(i++; Z_{i\text{new}} \neq Z_m)$$

Z_{inew} is newly created data point.
Z_m is data sample of class which is minor.
Z_n is data sample that is chosen from 4 kNN of Z_m in the minority class.
λ is an arbitrary constant which lie in range 0 to 1.

Originally, in SMOTE, each new data point becomes part of the population right away and is in contention to be chosen for the next nearest neighbors. In SMOTE$_{\text{MODIFIED}}$, when the number of new points generated is equal to the population of the minority class, then only the new points are added to data and become contentions for next nearest neighbor. To do this, a data point and its four (or n) neighboring points in minority class are considered.

NCL$_{\text{MODIFIED}}$

Neighborhood cleaning rule (NCL) is a technique that is used to remove the noisy or redundant data from the majority class. If the selected sample data (Z_m) belongs to the class which is in majority and the 4-nearest neighbor's classification of Z_m is in contrast to Z_m, then Z_m will be removed from the dataset. If Z_m is a member of the minority class, and the categorization of four closest neighbors is in conflict with Z_m, those four neighbors will be eliminated. In NCL$_{\text{MODIFIED}}$, during the cleaning neighborhood process, even if the majority of the neighbors of majority class sample belong to minority class, we eliminate that sample.

Step 2: The authors have used a combination of $\text{SMOTE}_{\text{MODIFIED}}$ and $\text{NCL}_{\text{MODIFIED}}$ ($\text{SMOTE}_{\text{MODIFIED}} + \text{NCL}_{\text{MODIFIED}}$) to improve the imbalance ratio.

Imbalance datasets are first resampled by the $\text{NCL}_{\text{MODIFIED}}$ technique. This method eliminates outlier data from sample data in the majority class. After this, the dataset is fed into the $\text{SMOTE}_{\text{MODIFIED}}$ algorithm. This technique is a method for boosting minority class sample data by synthesizing new data from existing data. Two data points (Z_m, Z_n) will be chosen for the generating new data samples in minority class, and the distance between Z_m and Z_n will be calculated. This operation is performed till all the existing data points are exhausted, and number of samples are balanced out without including the newly created data points as neighbors. We repeat this technique till balanced datasets are generated.

Step 3: The performance of the sampling algorithms has been evaluated using two machine learning classifiers, linear support vector classifier (linear SVC) and K-nearest neighbor (kNN). Linear SVC converts the 2D space into a hyper plane, which is divided into two or more categories, to classify the data. In kNN algorithm, k-1 nearest neighbors are considered, then the classification of a majority of these k samples is chosen as the classification of the data sample.

5 Result Evaluation

In this section, we explain the results obtained in this study.

Analysis of the Datasets:

Table 2 summarizes the results of the dataset resampling phase of the project. The table lists the number of specimens of both classes for each of the four datasets. As shown in the table, NCL + SMOTE method produced a smaller training dataset with fewer amounts of data as compared to solo SMOTE technique.

The original dataset classification is compared with the results of KNN classification and the linear SVC. Table 3 lists the number of specimens of both classes for each of the four datasets according to the two classification.

Table 2 Training results for $\text{NCL}_{\text{MODIFIED}}$, $\text{SMOTE}_{\text{MODIFIED}}$ and $\text{NCL}_{\text{MODIFIED}} + \text{SMOTE}_{\text{MODIFIED}}$

Dataset names	Original dataset		$\text{NCL}_{\text{MODIFIED}}$		$\text{SMOTE}_{\text{MODIFIED}}$		$\text{NCL}_{\text{MODIFIED}} + \text{SMOTE}_{\text{MODIFIED}}$	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Thyroid	139	22	129	22	139	139	129	129
Abalone	2839	293	2372	293	2839	2839	2372	2372
Arrhythmia	320	19	283	19	320	320	283	283
Mammography	8192	195	7991	195	8192	8192	7991	7991

Table 3 Testing results for actual classification versus KNN and linear SVC classifiers

Dataset names	Original test dataset		KNN classification		Linear SVC classification	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Thyroid	46	8	45	9	46	8
Abalone	947	98	1003	42	617	428
Arrhythmia	107	6	76	37	106	7
Mammography	2731	65	2625	171	2485	311

Figure 1 shows us the original "thyroid" dataset. Figures 2 and 3 show us the results of individual undersampling and oversampling of the original dataset using $\text{NCL}_{\text{MODIFIED}}$ and $\text{SMOTE}_{\text{MODIFIED}}$ techniques, respectively. Finally in Fig. 4, we have oversampled the results of undersampling by applying $\text{SMOTE}_{\text{MODIFIED}}$, thus balancing the cleaned dataset. From these figures, we can observe we see that $\text{NCL}_{\text{MODIFIED}}$ has removed the samples of majority class that are surrounded by those of minority class, while $\text{SMOTE}_{\text{MODIFIED}}$ has basically balanced the classes. Combining these techniques has combined their benefits as well. Similar figures and interpretations were obtained for all the other datasets. Due to space constraints, we have not shown the figures.

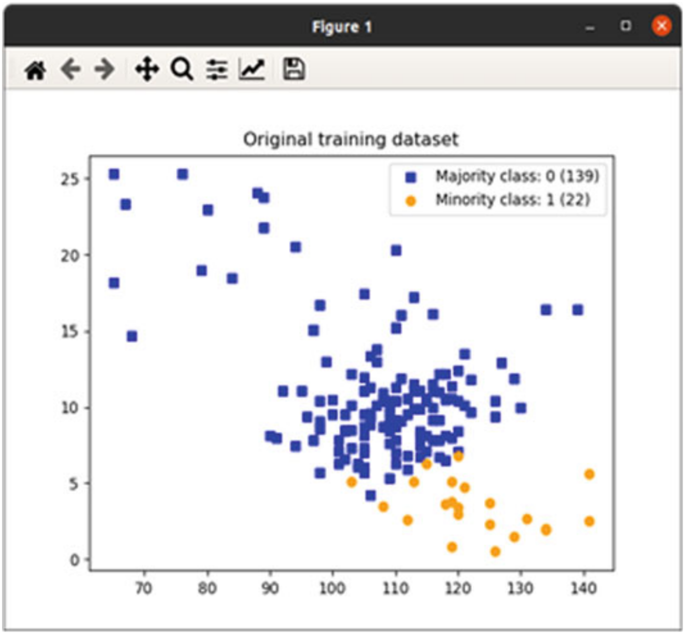

Fig. 1 Thyroid training data

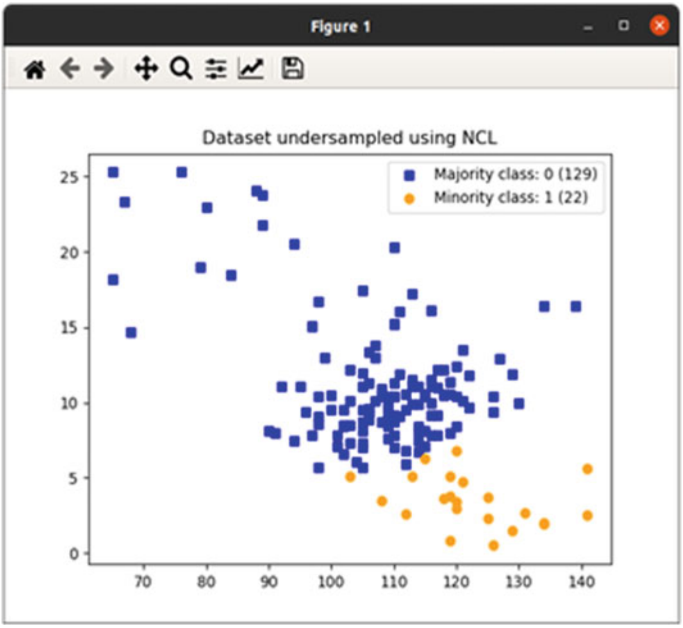

Fig. 2 Thyroid data after undersampling

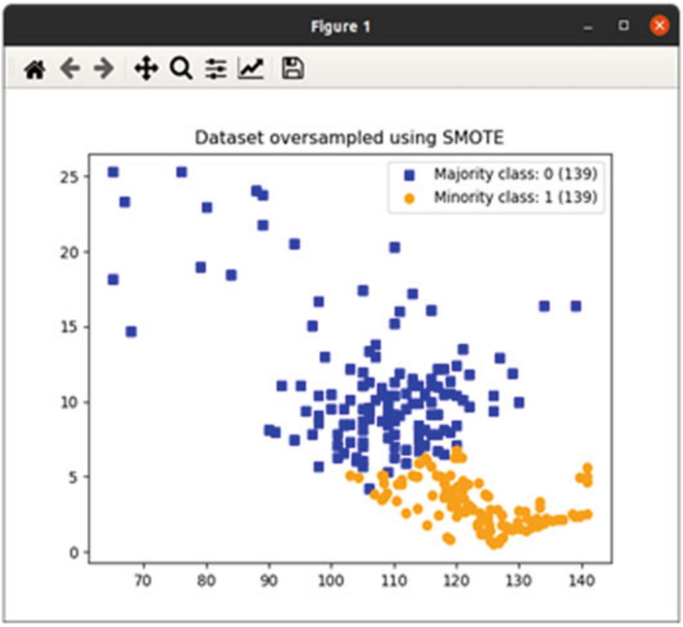
Figure 1
Dataset oversampled using SMOTE
■ Majority class: 0 (139)
● Minority class: 1 (139)
25
20
15
10
5
0
70
80
90
100
110
120
130
140

Fig. 3 Thyroid data after oversampling

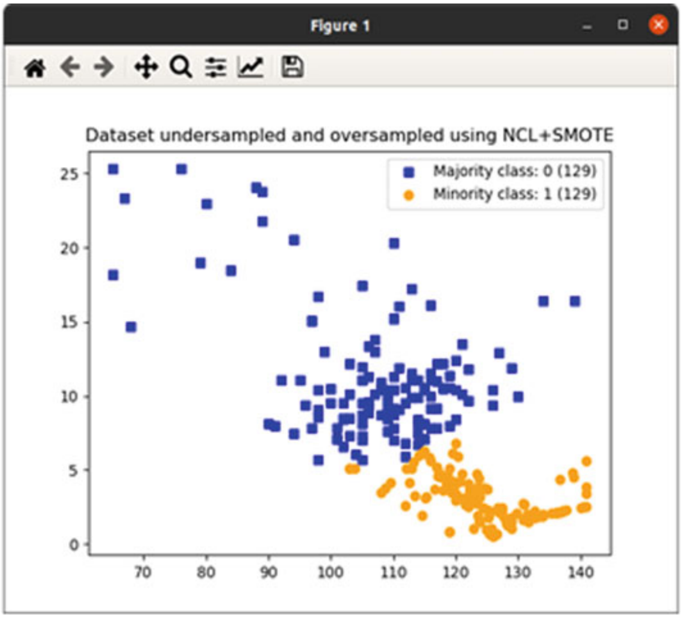

Fig. 4 Thyroid data after combination

6 Results of Empirical Validation

The results have been evaluated using two performance metrics, recall measure (sensitivity of minority class) and geometric mean. The mathematical formulae of recall measure and geometric mean are as follows:

$$\text{Recall measure} = \text{TP}/(\text{TP} + \text{FN}),\ \text{Geometric mean} = \sqrt{}(\Pi\ \text{sensitivity}_{\text{i}})$$

where TP = true positive, FP = false positive, TN = true negative, FN = false negative, sensitivity$_i$ is the recall measure of i_{th} class.

The geometric mean measure aims to improve the precision of each class while keeping them properly calibrated. The best value is 1, and the most egregiously bad value is 0. In most cases, G-mean resolves to zero if the classifier misses one or more classes. The validation used in hold-out validation in which training and testing data is divided in the ratio 4:1.

Tables 4 and 5 contrast review and performance measures between the kNN classifier and linear SVC classifier. Here, we have calculated the recall measure and geometric mean score of the individual classification of each dataset. We can observe from the tables that both recall and geometric mean values have increased after the datasets have been balanced using NCL$_{\text{MODIFIED}}$ + SMOTE$_{\text{MODIFIED.}}$ The percentage increase of each algorithm is shown in Table 5. We can observe that the range of percentage increase is 2.89% to 425.42%, which is significantly high. Thus,

Table 4 Recall measure for original datasets

Dataset names	Recall measure		Geometric mean	
	KNN	Linear SVC	KNN	Linear SVC
Thyroid	0.6250	0.8750	0.7906	0.9354
Abalone	0.1633	0.5306	0.3985	0.1428
Arrhythmia	0.1667	0.2387	0.4005	0.3986
Mammography	0.4615	0.2462	0.6785	0.4959

Table 5 Recall measure for $\text{NCL}_{\text{MODIFIED}} + \text{SMOTE}_{\text{MODIFIED}}$ technique

Dataset names	Recall measure				Geometric mean			
	KNN	%inc	Linear SVC	%inc	KNN	%inc	Linear SVC	%inc
Thyroid	0.9781	56.49	0.9126	4.29	0.9891	25.10	0.9625	2.89
Abalone	0.7347	349.90	0.9082	71.16	0.7322	83.73	0.7503	425.42
Arrhythmia	0.6667	299.94	0.333	39.50	0.6790	69.53	0.5637	41.41
Mammography	0.8769	90.01	0.9077	268.68	0.9167	35.10	0.9108	83.66

the authors in this study promote the use of $\text{NCL}_{\text{MODIFIED}} + \text{SMOTE}_{\text{MODIFIED}}$ for sampling the datasets.

7 Conclusion

In this research, we used a combination of both undersampling and oversampling techniques mainly SMOTE and NCL as they are most commonly used in the literature. We eradicated class imbalance problem in the datasets by combining both techniques, by stacking $\text{SMOTE}_{\text{MODIFIED}}$ on top of $\text{NCL}_{\text{MODIFIED}}$. We proposed that how neighborhood cleaning rule ($\text{NCL}_{\text{MODIFIED}}$) undersamples our datasets by combining ENN algorithm to clean datasets and CNN to some redundant samples. We also proposed how $\text{SMOTE}_{\text{MODIFIED}}$ oversamples the dataset by creating new data points by using samples from original data. We evaluate the balanced datasets using two classification algorithms, namely kNN and linear SVC. We use two performances metric to measure the effectiveness of our resampling technique, namely recall measure and geometric mean score. Both performance measures showed significant percentage increase in the range of 2.89–425.42% when the datasets are sampled as compared to their values on the imbalanced datasets.

There are also some limitations to our approach as this method is not suitable for mono dimensional data containing medium to high data imbalance level. Since there are few number of sample data it is possible to ignore serious data. The model resampling time on larger training datasets increased, and some information loss may also occur.

References

1. Choirunnisa S, Lianto J (2018) Hybrid method of undersampling and oversampling for handling imbalanced data. Int Seminar Res Inf Technol Intell Syst (ISRITI) 2018:276–280. https://doi.org/10.1109/ISRITI.2018.8864335
2. Jiang C, Liu Y, Ding Y, Liang K, Duan R (2017) Capturing helpful reviews from social media for product quality improvement: a multi-class classification approach, pp 3528–3541
3. Bej S, Davtyan N, Wolfien M, Nassar M, Wolkenhauer O (2021) LoRAS: an oversampling approach for imbalanced datasets. Mach Learn 110(2):279–301
4. Hassan M, Amiri N (2019) Classification of imbalanced data of diabetes disease using machine learning algorithms. In: IV international conferences on theoretical and applied computer science and engineering. Istanbul, Turkey
5. Mohammed A, Hassan M, Kadir D (2020) Improving classification performance for a novel imbalanced medical dataset using SMOTE method. Int J Adv Trends Comp Sci Eng 9:3161–3172. https://doi.org/10.30534/ijatcse/2020/104932020
6. Beckmann M, Ebecken N, Lima B (2015) A KNN undersampling approach for data balancing. J Intell Learn Syst Appl 7:104–116. https://doi.org/10.4236/jilsa.2015.74010
7. Napierala K, Stefanowski J (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst 46:563–597. https://doi.org/10.1007/s10844-015-0368-1
8. Agustianto K, Destarianto P (2019) Imbalance data handling using neighborhood cleaning rule (NCL) sampling method for precision student modeling 86–89. https://doi.org/10.1109/ICOMITEE.2019.8921159
9. 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. Inf Sci 477:47–54
10. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A (2015) A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inf 58(2015):49–59
11. Stefanowski J (2013) Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In: Ramanna S, Jain L, Howlett R (eds) Emerging paradigms in machine learning. Smart innovation, systems and technologies, vol. 13. Springer, Berlin. https://doi.org/10.1007/978-3-642-28699-5_11
12. Sáez JA, Luengo J, Stefanowski J, Herrera F (2015) SMOTE–IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci 291:184–203. https://doi.org/10.1016/j.ins.2014.08.051
13. https://archive.ics.uci.edu
14. https://www.kaggle.com
15. https://www.bcsc-research.org