Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 224-232, 3, 2024

Araştırma Makalesi

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 224-232, 3, 2024

Research Article

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

Novel SMote Based Ensemble Model for Medical Multi Class Imbalanced Data Set

Mouna Lamari[1], *Nabiha Azizi[2], Nadjette Dendani[2], Aya Lihoum[2], Nacer Eddine Hammami[3]

[1]*LRI Laboratory, Faculty of Technology, Badji Mokhat University of Annaba, 23000, Algeria*
[2]*Labge Laboratory, Faculty of Technology, Badji Mokhat University of Annaba, 23000, Algeria*
[3]*Computer Engineering Department, College of Engineering and Computer Science, Mustaqbal University, Saudi Arabia*
[*]*(azizi@labged.net)*

(Received:17 April 2024, Accepted: 25 April 2024)

(2nd International Conference on Scientific and Innovative Studies ICSIS 2024, April 18-19, 2024)

ATIF/REFERENCE: Lamari, M., Azizi, N., Dendani, N., Lihoum, A. & Hammami, N. E. (2024). Novel SMote Based Ensemble Model for Medical Multi Class Imbalanced Data Set. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(3), 224-232.

Abstract – Class imbalance, defined as a difference in the number of occurrences of the various classes in the issue, is present in many real-world classification datasets. Classifiers are known to suffer from this problem due to their accuracy-oriented design, which leads the minority class to be neglected. To overcome this issue, a number of balancing approaches have been widely adopted.

In this study, a classification system incorporating dynamic resampling approaches for the classification of imbalanced medical datasets is suggested.The major goal of our master thesis is to shed light on multi-class imbalanced data issues by adopting resampling techniques from SMOTE extensions for the classification of imbalanced multi-class data.

To beneficate of the different generated synthesis samples, our contribution consists on combining without duplication all datasets. The final decision will be ensured by the integration of a hybrid ensemble approach dubbed SMOTEBagging.

An empirical investigation on five imbalanced datasets was done to attain this objective and to evaluate the performance of the SMOTE extension methods before and after rebalancing data. As a first stage, a variety of resampling techniques were used to rebalance the data. Then, in a non-repetitive data fusion procedure, we integrate the results of each approach, and finally, we employ the resulting dataset to work with a hybrid ensemble method called SMOTEBagging.

The many tests we conducted on diverse datasets that the suggested approach performs quite well.

I. INTRODUCTION

Classification is a technique which needs machine learning algorithms and labelled training data to gain how to

designate class label to sample from the domain. Despite the development of several classification algorithms,

current work in machine learning has demonstrated that no classification algorithm can perform or surpass another,
and the choice of a classification approach is often dependent on the type of the problem we are dealing with.

There are two types of classification issues: binary-class problems and multi-class problems. The supplied data-set is classified into two classes in binary-class classifications, whereas the given data-set is classified into numerous classes in multi-class classifications depending on the classification rules. Alternatively, in real world applications, most datasets whether it is binary or multi-class suffer from the imbalance problem where some classes of data may have few training examples compared to other classes[1].

Indeed, most classification techniques fail to accurately classify minority class samples, implying the need for data balance in the construction of a competent classifier with a low error rate.

A variety of approaches have been developed to meet the goal of appropriately recognizing the minority class. In Algorithm level approaches, the idea is to attempt to modify current classifier learning algorithms in order to bias learning toward the minority class. On the other hand, Data level approaches resample the data space to rebalance the class distribution (or undersample the majority classes). We also have the third type of approach called ensemble-based methods where it generally consists of a mix of an ensemble learning algorithm and one of the previously mentioned ways[2, 3].

However, the difficulty is that we don't know which approach to use in order to achieve better classification results.

Recent study has discussed this subject and various strategies to work with, however it is important to note that the interest of these researches is always concentrated on the binary type of datasets, disregarding the multi-class dataset type.

It has been established that using a group of classification algorithms produces better results than using a singular algorithm, because the fusion of the judgments of numerous complementary and diverse algorithms typically results in a reduction of error smaller than the result of a single classifier

Our goal is to achieve an effective classification system dedicated to imbalanced multiclass datasets classification problems by proposing an approach combining multiple balancing algorithms.To reach our goal, we use the extensions of the famous SMOTE algorithms for imbalanced data.

To accomplish an analysis of this approach, a comparative study has been done on a variety of multi-class imbalanced datasets. Where every SMOTE extension will generate different synthetic examples from the same dataset. For better classification results, we suggest to benefit the diversity of this generated data by integrating a hybrid method named SMOTEBagging.

II. BASIC CONCEPTS OF CLASS IMBALANCE

A. Data Imbalance

Any dataset with an unequal class distribution is technically unbalanced.A dataset, on the other hand, is called imbalanced when there is a significant, and in some severe circumstances, unequal distribution of samples for each type of problem. Class imbalance happens when the number of instances representing one class is much less than the number of instances representing the other classes. As a result, one or more classes may be overlooked in the dataset. and the minority class may represent the most interesting concept and may reveal primary concerns for determining the properties of the classification models.

B. Class Imbalance Problem

The efficacy of Software Fault prediction models is dependent by the classification process of the training data.The number of occurrences in the training data set is used to establish the Class distribution. If the number of instances belonging to one class is significantly greater than the number of instances belonging to another class, the problem is known as class imbalance.The class with more instances is known as the majority class, whereas the class with the fewest instances is known as the minority class. When the class beneath evaluation, i.e. the incorrect year, is represented by few instances, the problem becomes more severe. Several approaches for addressing this issue have been proposed.

Classifying problems connected to class imbalance into can be devided into six categories. These are the categories:

- "Absolute" absence of data: This is the most common cause of imbalance, and it occurs when the available data is insufficient to properly define the class borders. This is seen in Figure (1.1), where the observations in red are insufficient to adequately characterize the idea of this class.

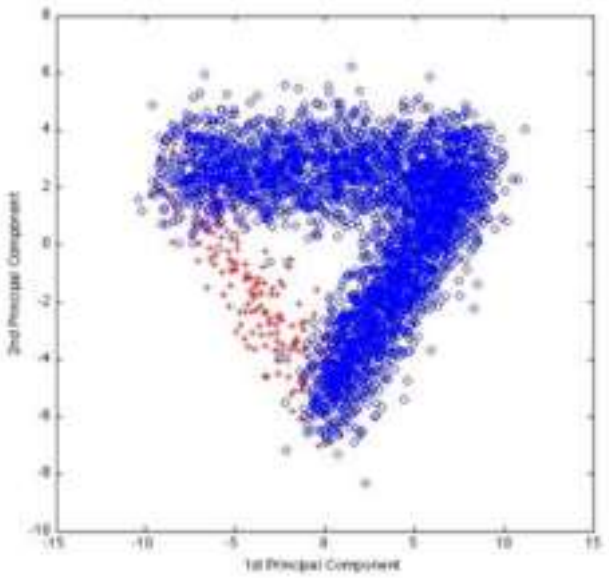

Figure 1. “absolute’ absence of data

- "Relative" lack of data: This problem is similar to the previous one, however the lack of data in this case is relative to the size of the majority data. Minority observations are uncommon in comparison to those of the opposite class (majority class). As demonstrated in Figure (2), the minority class objects (in red) are represented in the data with a proportion of 50% as compared to the majority class objects (in blue).

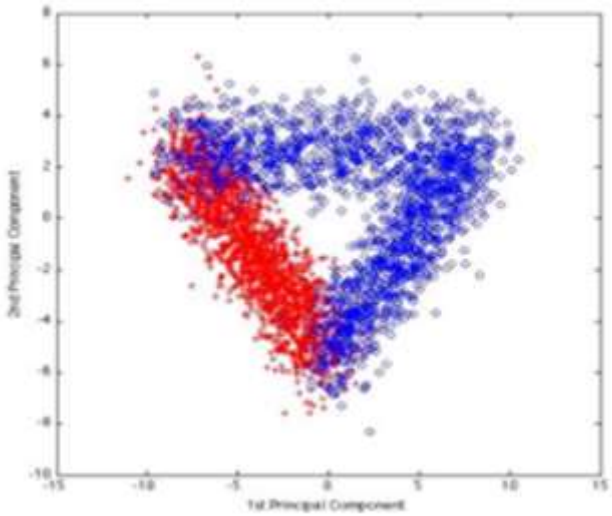

Figure 2. “Relativeé lack of data

- Inappropriate metrics: The metrics and assessments used to evaluate learning outcomes are not suited for the challenges of imbalanced courses.
- Data fragmentation: it is connected to top-down algorithms that start from the area of all individuals and repeatedly partition it into smaller and smaller subsets.

- Inappropriate induction margin: this is the margin used to generalize the learnt rule on the training data.

- noisy data: noise has a greater influence on uncommon classes than on frequent classes.

C. The imbalance ratio (IR) : Is the most commonly used measure to describe the imbalance extent of a dataset. IR is defined as :

$$\mathrm{IR} = \frac{N_{\mathrm{maj}}}{N_{\mathrm{min}}},$$

Where Nmaj represents the number of instances of the majority class and Nmin represents the number of instances of the minority class. When there are multi-classes(the number of classes is larger than 2) then Nmaj will represent the number of instances of the largest majority class and Nmin represents the number of instances of the
smallest minority class.
mathematically, when IR = 1, we have an exactly balanced dataset. When IR > 1, the larger the IR, the larger the imbalance extent of the dataset [8 , 9].

III. PROPOSED APPROCAH FOR IMBALANCED MULTI-CLASS DATASET LEARNING

The process of creating our system is made up of two sections :

balancing the data with four different hybrid multi-class sampling algorithms, producing the initial set of four balanced datasets, classifying each dataset with random forest, and comparing the results

We would also like to combine the advantages of each strategy with the other approaches. This leads us to the second phase of this system, where we describe the concept of combining the resulting dataset from the sampling ways and classifying it using a hybrid data method called SMOTEBAGGING.
The figure below describes the functioning of our system by displaying the succession of the several phases:

A. Data pre-processing
It is a critical phase since the performance and accuracy of the models created in the DM phase are dependent on how the data analyst structures and delivers the input to the following phase. On the other side, at this step, the data must be codified in order to be a valid input for the DM algorithms that will be utilized. In this manner, if the method to be used requires numerical input and the data chosen are categorical, or vice versa, the data will be changed at this step to acquire the proper format. Furthermore, additional variables are frequently derived at this point.
In this work, we have used both of the packages numpy and pandas to transform our dataset to a numerical convenient data to work on .

B. Balancing the dataset:
In real world classification situations, the classes are generally imbalanced, and the class of interest is the less numerous class. Because the minority class is rare, most algorithms penalize it far more.
A series of balancing techniques (see chapter 1, section 1.4.4.1) have been presented to overcome the class asymmetry problem (TomekLinks-SMOTE,SMOTE-OUT,TRIM-SMOTE,Gaussian-SMOTE). These data sampling strategies are meant to address multi-class imbalance.

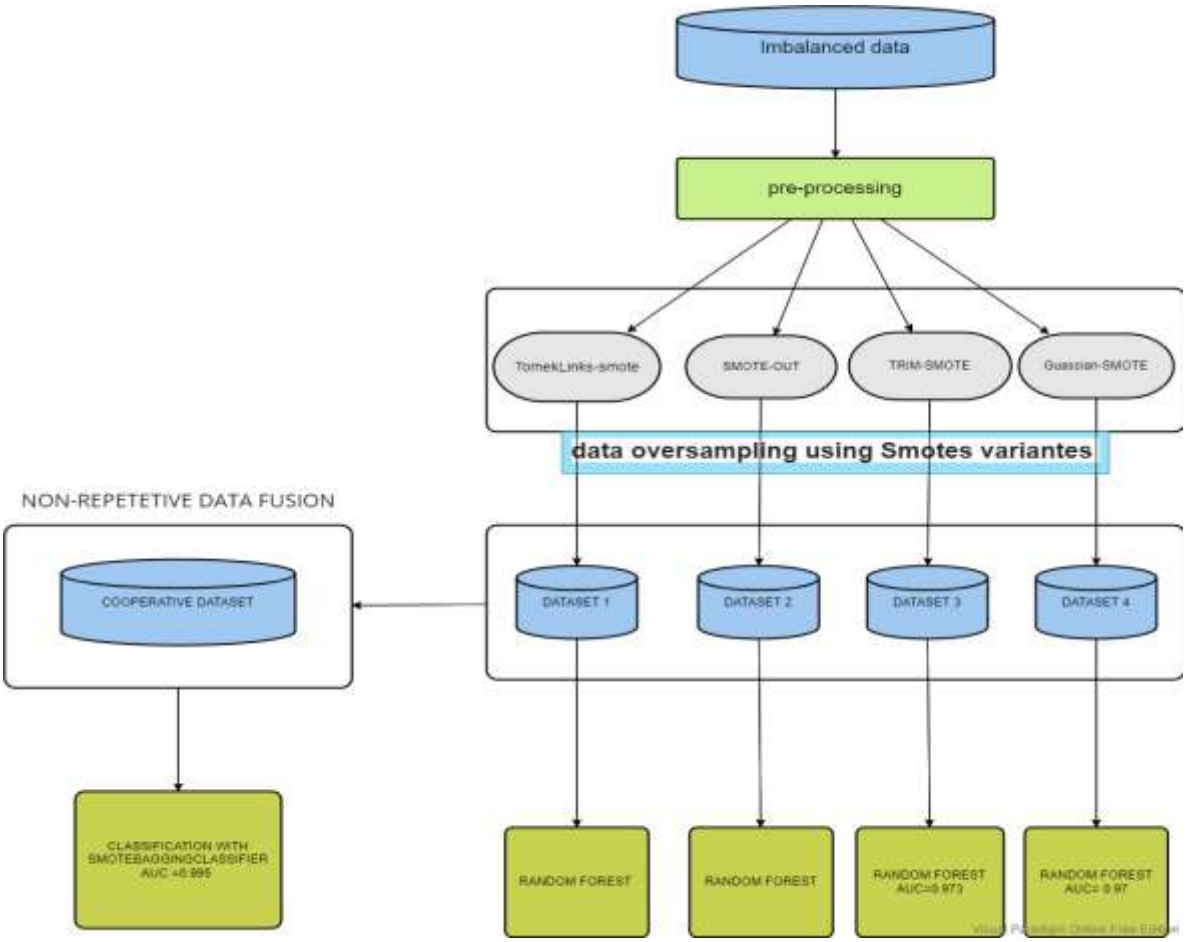

Figure 3. General diagram of the multi-class system classification for the classification of imbalanced data.

C. Used techniques for the data balancing

We employed four hybrid sampling procedures to redefine the data distribution: SMOTE-OUT, TomekLinks-SMOTE, TRIM-SMOTE, and Gaussian-SMOTE.

The goal of this study is to investigate the impact of different strategies on data distribution as well as the efficiency of the final selection of the set of selected classifiers.

- TomekLinks-SMOTE :

the approach TomekLinksSMOTE combines the two methods tomek links and SMOTE, the first technique is an undersampling method that works on deleting the mutual “tomek links pairs” from both the minority and majority class, then SMOTE oversampling the output .

- SMOTE-OUT :

this method generates synthetic examples in the outer region of the desk line to avoid the issue of having two vectors near together

- TRIM-SMOTE ;

TRIM-SMOTE is a solution for the overgeneralization problem where the boundaries of the classes are confused together. its works on producing synthetic minority class samples with higher quality than SMOTE

- Gaussian-SMOTE :

The gaussian-SMOTE approach provides diversity in generating synthetic samples which will solve the overgeneralization problem Once the data is balanced, we move on to the classification phase of balanced data provided by each SMOTE variant.

D. CLASSIFICATION STAGE

- **Using randon forest :** At the end of the learning phase we will have four balanced bases in four different ways
- **Non repetitive-data fusion/** After performing the classification step during the learning phase, we move on to non-repetitive data fusion.
  It consists of combining the four balanced datasets in a way where no duplicates are found in the new dataset.

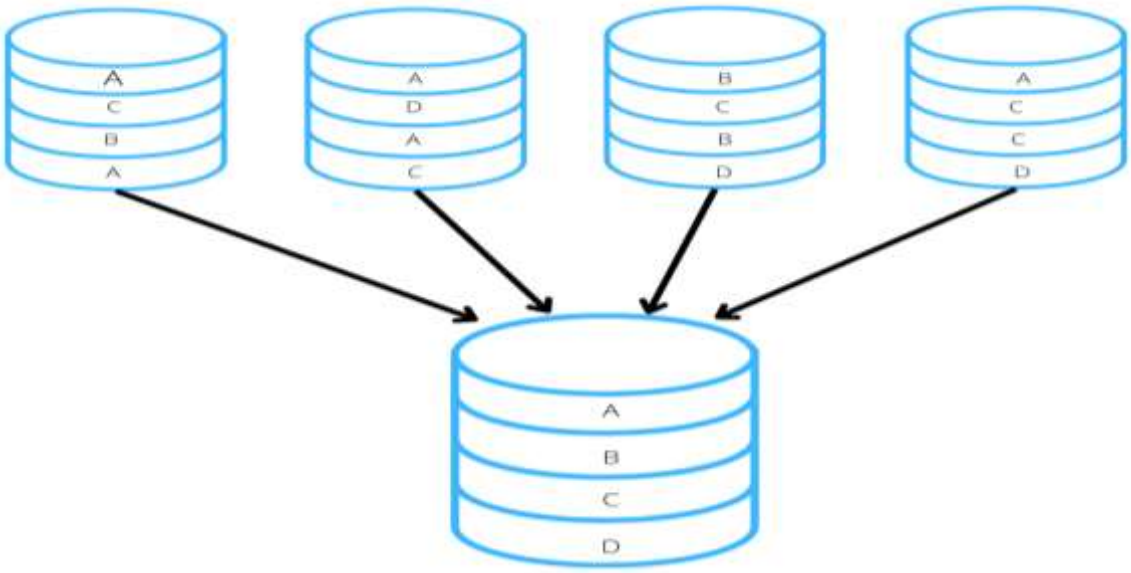

Figure 4. Generated dataset cooperation

E. Classification Using Smote Bagging

SMOTEBagging is a hybrid of the SMOTE and Bagging algorithms.

SMOTEBagging consists of creating synthetic instances throughout the subset creation process. SMOTE generates synthetic data until the amount of minority data equals the amount of majority data. The object's characteristic and k-nearest neighbor are used to generate synthetic data. Bagging is an abbreviation for Bootstrap Aggregating, which was created by Breiman (1996) to minimize predictor variance. According to Zhou [38], the main notion of the ensemble technique is to bootstrap a fresh dataset to develop a classifier in multiple versions. The goal of this combination is to produce a strong model for identifying imbalanced data while maintaining overall accuracy.

Each subset is produced via the Bootstrap process balanced by SMOTE prior to modeling, according to SMOTEBagging. SMOTE requires the selection of two parameters: N is the total number of oversampling from the minority class multiplied by k-nearest neighbors. The sum of over-sampling determined in terms of the amount of majority and minority class is balanced.

Training:
1. Let S be the original training set.
2. Construct subset S_k containing instances from all classes with same number by executing the following:
   2a. Re-sample class C with replacement at percentage 100%.
   2b. For each class $i\,(1, \ldots, C-1)$:
   Re-sample from original instances with replacement at the rate of $(N_C/N_i) \cdot b\%$.
   Set $N = (N_C/N_i) \cdot (1 - b\%) \cdot 100$.
   Generate new instances by using SMOTE (k, N).
3. Train a classifier from S_k.
4. Change percentage $b\%$.
5. Repeat step 2 and 3 until k equals M.

Testing on a new instance:
1. Generate outputs from each classifier.
2. Return the class which gets the most votes.

IV. EXPERIMENTAL STUDY

A. Used Datasets

In order to evaluate the performance of the proposed classification system, we used a multi-class datasets chosen from the KEEL dataset repository [10]
A detailed description of the database used is presented in the following table:

Table1 : Main results based on 05 imbalanced dataset

	number of features	number of instances	number of classes	imbalanced ratio
Thyroid dataset	21	2666	3	20.04
contraceptive	10	1472	3	0.74
glass	10	213	6	0.59
hayes-roth	5	131	3	0.63
shuttle	10	2174	5	3.63

For the evaluation of the performance of the methods on imbalanced data, we divided our experiments into two main parts: before and after balancing the databases. In each step the results of the basic classifiers are presented.

V. CONCLUSION

This study was dedicated to the investigation and implementation of a system for the classification of imbalanced multi-class data based on sophisticated machine learning paradigms such as the usage of sets of resampling techniques and a hybrid ensemble approach.

The databases utilized in this work were preprocessed and balanced using SMOTE variations entitled TRIM-SMOTE, SMOTE-OUT, SMOTE TOMEKLINKS, and GAUSSIAN-SMOTE. Once datasets will be balanced, Random Forest classifier will be applied toachive classification stage .

After the validation of SMOTE variants on the five databases, we found that the use of these methods for data balancing gives better results, the results of the AUC measurement to which we have arrived are: …

For better classification results, we combined the output of each SMOTE variant used in this study in a singular dataset to work with SMOTEBagging , the results of the AUC measurement to which we have arrived are: …

The results of our empirical investigation are really encouraging, and they urge us to continue with this area of research. Here are some future views to consider:

- Other data balancing strategies may be used, either by modifying the classification algorithms or by employing ensemble approaches tailored to imbalance concerns.

REFERENCES

[1] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *International conference on intelligent computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[2] Brandt, Jakob, and Emil Lanzén. "A comparative review of SMOTE and ADASYN in imbalanced data classification." (2021).

[3] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." *IEEE Transactions on knowledge and data engineering* 21.9 (2009): 1263-1284.

[4] Cheriguene, S., Azizi, A., & Ziani, A. (2016, December). A two stage Classifier Selection Ensemble based on mRMR Algorithm and Diversity Measures. In the 2nd Conference on Computing Systems and Applications, Algiers, Algeria

[5] Rout, Neelam, Debahuti Mishra, and Manas Kumar Mallick. "Handling imbalanced data: a survey." *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016*. Springer Singapore, 2018.

[6] Cheriguene, S., Azizi, N., Dey, N., Ashour, A. S., Mnerie, C. A., Olariu, T., &a Shi, F. (2016, August). Classifier Ensemble Selection Based on mRMR Algorithm and Diversity Measures: An Application of Medical Data Classification. In International Workshop Soft Computing Applications (pp. Bibliographie 155 375-384). Springer, Cham

[7] LAMARI, Mouna, AZIZI, Nabiha, HAMMAMI, Nacer Eddine, *et al.* SMOTE–ENN-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification. In : *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020*. Springer Singapore, 2021. p. 37-49.

[8] Tanha, Jafar, et al. "Boosting methods for multi-class imbalanced data classification: an experimental review." *Journal of Big Data* 7 (2020): 1-47.

[9] Krawczyk, Bartosz, et al. "Undersampling with support vectors for multi-class imbalanced data classification." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

[10] https://sci2s.ugr.es/keel/datasets.php