An Improved Unbalanced Data Classification Method Based on Hybrid Sampling Approach

2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI) | 978-1-6654-1270-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/BDAI52447.2021.9515306

Biru Xu
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
b.xu9@student.liverpool.ac.uk

Wenjia Wang
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
w.wang73@student.liverpool.ac.uk

Rui Yang*
School of Advanced Technology
Research Institute of Big Data Analytics
Xi'an Jiaotong-Liverpool University
Suzhou, China
r.yang@xjtlu.edu.cn
*Corresponding Author

Qi Han
College of Electrical Engineering and Automation
Shandong University of Science and Technology
Qingdao, China
15963243410@163.com

***Abstract*—The problem of data imbalance has received far-reaching concerns since they could affect the accuracy of classification problem in the area of machine learning. As the minority class instances can be ignored by traditional classifiers, it is necessary to improve the recognition rate of minority instances. Therefore, the paper proposes a new hybrid sampling method to solve the data imbalance problem by enlarging the proportion of minority instances. For the oversampling part, a variant of SMOTE is provided combining methods of LR-SMOTE and CCR (Combined Cleaning and Resampling Algorithm); for the under-sampling part, the Tomek-link method is utilized to complete the task. After the pre-processing stage, the data set is classified by Random Forest (RF). Experimental results show that the novel algorithm effectively enhances the performance of RF on the data set with a higher accuracy.**

Keywords—imbalanced dataset, hybrid sampling, smote, data mining

I. Introduction

Numbers of algorithms and techniques have been proposed for the classification problem in the field of machine learning. Algorithms' efficiency can be affected by the underlying features of datasets. These datasets are common to suffer from the data imbalance problem, which means the number of instances in one class overly outweighs the other. The problem of data imbalance has received far-reaching concerns.

The under-represented class is named as minority class, which has a lower number of instances, while the over-represented one is named as majority class, which owns a larger number of instances. The unbalanced datasets cause a bias toward the larger class when building the classification model, thus classifying most of new instances to the majority class. Some important minority instances have been ignored during the procedure. Therefore, it is necessary to improve the recognition accuracy for the minority class.

The solution to this issue lays on two levels: data level and algorithmic level. Data level technique aims to generate a more balanced dataset by modifying data distribution, generally includes over-sampling (adding the minority instances) and under-sampling (deleting the majority instances). Data resampling methods are more versatile since their independence of the classifier [1]. Algorithmic level technique aims to modify classifiers for better performance in minority class's accuracy. In this article, the focus has been put on the data level techniques.

There have been many oversampling methods proposed to produce new minority instances and balance the dataset. Random Oversampling [2] remains the most direct method for producing new instances, which randomly duplicates the minority instances achieving a balanced dataset. As it could generate a smaller and more specific decision area, it could cause overfitting problem [2]. Synthetic Minority Oversampling Technique (SMOTE) is the most well-known oversampling method proposed by Chawla et al.[3]. It generates new minority samples by creating them along the line joining any of the k minority class nearest neighbours to generalize the decision region of the minority class. Its disadvantage is that as it randomly operates on the entire dataset, it may cause noise within the area of majority class by ignoring them in new instances' generation process, especially problematic in the highly skewed datasets.

Many derivative SMOTE methods emerge for improving SMOTE. Borderline-SMOTE [4] extends the SMOTE by dividing the minority instances into danger, safe and noise and performs SMOTE only on danger instances. As there are trivial differences among the number of KNN neighbours, the classifications for noise and border-line instances are not accurate under the Borderline-SMOTE method. Safe-level SMOTE [2] creates new instances only in the area where more instances are belonging to the minority class. ADASYN (ADAptive SYNthetic Sampling) [5] adaptively generates more new instances that are harder to learn. It shifts the decision boundary to the instances that are difficult to learn from as well as balance the dataset. LR-SMOTE [6] is a newly proposed derivative of SMOTE, combining K-means to reasonably generate new instances of minority class as they are near the centre of minority class. It has deficiency in avoiding the risk of generating noises since a random ratio is still included in the synthesis process for new samples.

Under-sampling helps to remove duplicated majority class instances to achieve a balanced dataset. Random Under-sampling randomly selects the majority class instances for removal. Tomek-link under-sampling [7] method as a refinement of CNN technique eliminates boundary instances which have more possibilities to be misclassified. In addition, the Redundancy-driven modified Tomek-link based under-sampling method comprises the detection of outliner and noises other than the Tomek-link under-sampling, for a better accuracy in classification. The algorithm contains repetitive steps for noises removal, which can be simplified.

In addition, the individual application of oversampling or under-sampling cannot modify some specific problems. For example, oversampling cannot resolve the redundancy of majority instances. As a result, a framework of hybrid resampling method should be put forward to handle problems in both majority and minority classes. A previous hybrid sampling method is SMOTE-ENN [1], which applies the SMOTE for oversampling and ENN for filtering noises for each class. A more comprehensive hybrid sampling method could be provided considering the improvement on both oversampling and under-sampling.

In the article, a novel hybrid algorithm has been proposed to deal with the imbalanced dataset problem in the data level. The purpose of the algorithm is to construct a complete and more reasonable hybrid algorithm for dealing with imbalance problem. It makes contribution by a promotion of SMOTE in oversampling, which can be better at reducing the randomness of synthetic samples and selectively generates synthetic samples, and for under-sampling, the Tomek-link under-sampling method for a better accuracy in the further classification problem.

The promotion of SMOTE lays on the combination of LR-SMOTE and CCR (Combined Cleaning and Resampling Algorithm) method. In order to achieve the reduced randomness of synthetic instances, the comparison of the distance from the minority instance to the synthetic one and the radius of calculated sphere area is in need for choosing a shorter one to reduce the randomness.

The remainder of this paper is divided into six sections. In Section 2, we provide a description of the problem discussed in the article. In Section 3 we describe how the proposed hybrid algorithm works. Experimental procedures and outcomes are shown in Section 4. The conclusion of the paper is included in Section 5.

II. Problem Description

One major problem for SMOTE is that synthetic instances are generated randomly between the two sample points. One sample point belongs to minority samples and the other belongs to its k nearest neighbours. If noisy points are included in the sample points, it may cause newly generated samples being noisy points or outliners [6]. The synthetic samples' location should be qualified within certain space to deal with the problem.

For SMOTE, the selection of minority samples for generation is random and the number of synthetic samples generated for selected sample is the same. Among all the minority samples, there are some samples that are difficult to learn [8]. These samples are the pressing issue for achieving improved overall accuracy due to the data imbalanced problem. An adaptive oversampling method for oversampling can reduce the misclassification ratio with limited randomness on selection and generation.

III. Preliminaries and Methods

To solve the classification problem of imbalanced dataset, a hybrid sampling CCR-SMOTETL has been proposed in the article to balance the samples between the minority class and the majority class. In the paper, the imbalanced dataset will be pre-processed by the CCR-SMOTETL algorithm and classified by the Random Forest classifier. In the following sections, the CCR algorithm is introduced and the LR-SMOTE algorithm is then illustrated. The Tomek Link concept is explained and Random Forest Classifier is then introduced. Finally, the CCR-SMOTETL algorithm is introduced in the section.

A. CCR Algorithm

The combined cleaning-resampling algorithm (CCR) is a novel data cleaning and oversampling method proposed by Koziarski and Wozniak [8]. The algorithm contains two operations, cleaning the minority samples neighbourhood and selectively generating synthetic samples. It enables the avoidance of the overlapping problems for synthetic instances and majority instances due to the cleaning process, and it shifts the decision boundary to difficult learning instances by adaptive oversampling like ADASYN.

1) Cleaning the minority samples neighbourhood

For the data cleaning step, the energy-based approach is designed for reducing the impact of noise in the majority class instances, within the minority class detection and accounts for minority class outliners. Every minority instance is associated with an energy budget and a spherical region. The sphere area consumes the available energy by enlarging its radius and the cost of growing the sphere is increased by every majority sample we encounter. Formally, for a minority instance denoted by xi, current radius denoted by ri, a function returning the number of majority samples within a sphere with radius r denoted by $fn(r)$, a target radius denoted by ri'. The energy change from radius ri' to ri is shown as following.

$$fn(r') = fn(r) + 1 \qquad (1)$$

$$\Delta e = -(r'i - ri) \cdot fn(r'i) \qquad (2)$$

During the expanding procedure, every minority instance depletes the available energy by increasing its radius. After the radius is fixed, all the majority observations are pushed to the outskirt region.

2) Selectively generating synthetic samples

Random generation of synthetic minority samples is restricted to its sphere area and the proportion of samples is reversely proportional to its radius. Formally, let ri be the radius of i_th minority instance, N the number of minority samples and G the total number of synthetic samples. The number of synthetic points generated for i_th sample is g,

$$g = \frac{ri^{-1}}{\sum_{k=1}^{n} rk^{-1}} G \qquad (3)$$

The CCR method manages to remove noise from the minority samples and adaptively learn the difficult instances, with a reduced impact from minority outliners.

B. LR-SMOTE Algorithm

The algorithm is proposed by Liang et al. [6] and it is the modification of the SMOTE method. It first utilizes k-means and SVM for noise removal. In addition, the rule to generates new minority samples is located to the connection or extension line of the sample and sample centre, for generating samples more reasonably.

1) After the noise removal, the k-means method is used to find the sample centre xi . Calculate the Euclidean distance d from the centre to each minority sample, the average distance of all the distances is denoted as $dmean$.
2) Calculate the ratio between the average distance and Euclidean distance, names as M.
3) New samples are generated according to the formula until the dataset is balanced. Ui is each of the minority class samples, and Xi is the centre of minority class sample.
   $$xnew = ui + rand(0, M) * (xi - ui) \quad (4)$$
4) The balanced dataset is output as the final result. The LR-SMOTE algorithm can filter noises and prevents new minority samples being noise or outliners. It reduces the randomness of generating new samples to some extent while the random number between 0 and M can be further qualified.

C. Tomek Link Algorithm

Tomek Link [7] is defined as: two samples from different classes Xi and Xj form the Tomek Link if there doesn't exist a sample Xl such that $d(Xi, Xl) < d(Xi, Xj)$ or $d(Xj, Xl) < d(Xj, Xi)$. If two samples form a Tomek Link, they are considered noise samples or borderline samples along the data distribution. As a under sampling method, the majority class samples are eliminated and both class samples are removed as a data cleaning method. The SMOTE-TL is one of the classic hybrid sampling method as it could delete noise samples after the oversampling procedure.

D. CCR-SMOTETL Algorithm

The idea of CCR-SMOTETL is to improve the classification accuracy of the output dataset by denoising, under-sampling, and generating synthetic minority samples more reasonably. The improved oversampling procedure, combining CCR and LR-SMOTE reduces the randomness of new samples' location and raises a high proportion of new instances generated from easily misclassified instances.

The steps of the algorithm are shown below:

```
Algorithm: CCR-SMOTETL algorithm
Input: $X$ ← the imbalanced sample set;
  $Xi$ ←the majority set; $Xj$ ← the minority set;
  $Tnum$ ←the number of samples to be synthesized;
  $ki$← centre point in the minority set calculated by K-means
Output: The balanced data set $X'$

  for all minority points $xi$ do
      $ri$ ← energy-based radius calculated by CCR
      for all majority points $Mj$ within $ri$ of $xi$ do
          d ← $\|xj - mi\|_1$
          $tj \leftarrow tj + \frac{ri-d}{d} * (Mj - xi)$ {translation of $Mj$}
      end for
      $d_1$ ← the distance between xi and ki
      sum← sum + $d_1$
  end for
  apply accumulated translations to all points in $Xj$
  $D\text{-}mean \leftarrow \frac{\text{sum}}{\text{Num(Xi)}}$

  for all minority points $xi$ do
      M ← $\frac{D-mean}{d1}$
      $gi \leftarrow \frac{ri^{-1}}{\sum_{k=1}^{n} rk^{-1}} Tnum$
      If $rand(0, M) * (ki - xi) < ri$
          comparison ← $rand(0, M) * (ki - xi)$
          for $gi$ times do
              $xnew$ ← xi + $comparison$
      Else
          $xnew$ ← xi + comparison − ri
      {in the intersection of the line connecting ki and xi and the
      energy sphere with ri}
          for $gi$ - 1 times do
              p ← random point inside the energy sphere
              $xnew$ ← xi + p
          end for
      end if
  end for
  remove Tomek-Links pairs in $X$
  output $X'$
```

To briefly introduce the algorithm, the following explanation is attached. The first step of the algorithm utilizes the modified CCR method. Each minority class instance xi is assigned with a sphere area of radius ri, calculating using its available energy. The majority instances within every sphere area of radius ri are pushed outskirt of the sphere. Then, the LR-SMOTE method finds the sample centre for Xj and *D-mean*, the average distance of all samples to ki. The ratio between the average distance and Euclidean distance, denoted as M. The part of $\text{rand}(0, M) * (ki - xi)$ as comparison in the synthetic formula is calculated for each minority instance.

For generating new synthetic instance, the radius ri is compared with the $rand(0, M) * (ki - xi)$ for each minority class instance. The number of synthetic samples for each instance is calculated as g in the CCR method. If $rand(0, M) * (ki - xi) < ri$, the synthetic instance is generated according to the LR-SMOTE formula. If $rand(0, M) * (ki - xi) >= ri$, one synthetic instance is generated in the intersection point of the connection line for the centre and the sample and the sphere. Others are generated according to the CCR method. Tomek-Links pairs are found on X' and are removed majority instances from the dataset.

The adaptive hybrid sampling absorbs the advantages of CCR method, as it utilizes the energy-based method for noise detection and selective generation for minority samples. The combination of CCR and LR-SMOTE qualifies the location of new synthetic samples from random points between two points to a specific and limited line, which reduces the randomness of generation. Further, the improved algorithm can remove both noise from minority neighbourhood and from Tomek Links, which is a more complete noise removal process compared to other pre-processing methods.

IV. Experiments

A. Data Set

The experiment uses 2 datasets from UCI Machine Learning Repository. The dataset detailed information is shown below, which includes the number of samples, the number of attributes, the number of majority class instances, the number of minority class instances and the imbalanced ratio between majority and minority classes (IR). The scope of the paper is restricted to the binary classification problem.

The reasons for selecting these two datasets are that their IR ratio is relatively large and the datasets are easy for manipulation. First, their IR ratios are all above 1, which can be regarded as imbalanced datasets in the binary classification problem. Furthermore, their attributes types are number, which is convenient for data processing. In addition, there is no missing value in the datasets. Thus, additional operations dealing with missing value can be saved.

The implementation details of the experiment can be found on[1] the open-source website[1] for further exploration.

TABLE 1. Dataset Information

Dataset	Samples	Attributes	Minority Samples	Majority Samples	IR
Haberman	306	3	81	225	2.78
Blood	748	4	178	570	3.20

B. Performance Evaluation

The evaluation indicator for traditional classification problem is normally the accuracy, which can effectively represent the classification effect as a whole, but cannot reflect the accuracy for minority classes. In the unbalanced dataset problem, the recognition of minority class is important so that the classification accuracy is not an appropriate measurement.

TABLE II. Confusion Matrix

Predicted Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The confusion matrix shown in the table is an effective tool to evaluate the algorithm prepared for imbalanced data classification. According to the definition, TP represents the number of positive class samples that are correctly classified, TN represents the number of negative class samples that are correctly classified, FP represents the number of positive class samples that are misclassified, and FN represents the number of negative class samples that are misclassified. Several basic indicators for evaluation can be defined based on the above definition:

1. Precision represents the proportion of instances which are correctly labelled as positive samples in all the samples that are classified to be positive. $Precision = \frac{TP}{TP + FP}$

2. Recall represents the proportion of correctly labelled positive samples in the actual positive samples, which reflects the classification accuracy of the positive samples. $Recall = \frac{TP}{TP + FN}$

3. TNR represents the proportion of correctly labelled negative samples in the actual negative samples, which reflects the classification accuracy of the negative samples. $TNR = \frac{TN}{TN + FN}$

4. F1 is the harmonic mean of Precision and Recall. F1 shows a high value only when the values of Precision and Recall are both high.

5. G-means is the geometric mean of TPR and TNR. When the values of TPR and TNR increase, the overall classification performance improves

TABLE III. Performance of Random Forest Classification under Different Pre-processing Methods

	Algorithm	Recall (TPR)	TNR	F-1	G-means
Haberman	Original	0.760695	**0.842963**	**0.799719**	0.471185
	SMOTE	0.796687	0.783704	0.790142	0.59018
	CCR	0.792537	0.786667	0.789591	0.580242
	SMOTE_TomekLink	0.711538	0.755319	0.696203	0.604244
	CCR-SMOTETL	**0.802432**	0.765926	0.792198	**0.603116**
Blood	Original	0.805629	**0.887135**	0.84442	0.528298
	SMOTE	**0.831721**	0.754386	0.791168	0.601023
	CCR	0.8246	0.874269	0.848708	0.594674
	SMOTE_TomekLink	0.705128	0.754098	0.69158	0.603992
	CCR-SMOTETL	0.829457	0.876023	**0.852105**	**0.608894**

The baseline models selected for comparison are the original dataset, the SMOTE algorithm, the CCR algorithm and the SMOTE-TomekLinks methods. The CCR model and SMOTE model testing on datasets shows the superiority of the proposed algorithm on the modification of oversampling. The SMOTE-TomekLink algorithm uses TL for under-sampling following the traditional SMOTE method. Although it shows progress on G-means value comparing to the SMOTE in the result, its performance is worse than the CCR-SMOTETL method as it is lack of the improvement on the oversampling

[1] The source code can be found on https://github.com/BeatrizXu/An-Improved-Unbalanced-Data-Classification-Method-Based-on-Hybrid-Sampling-Approach

method. It provides the superiority of the proposed methods as an improved hybrid model.

1) Comparison of F-1 and G-means

The use of pre-processing methods improves the *G-means* value to classify both datasets by RF. Compared by the original dataset that classified by RF, the CCR-SMOTETL has the obvious growth in *G-means*. It proves that the overall classification performance is enhanced by the proposed method.

Compared with other algorithms used for pre-processing, the *F-1* value in table 3 is relative better with the proposed method. Although it is acknowledged that the *F-1* value shows variations in the selected two datasets, the proposed method has a better performance comparing to the SMOTE and CCR methods, yet 0.007 lower than the original dataset. For the Blood dataset, the CCR-SMOTETL algorithm is better than all other algorithms for pre-processing in *F-1* value. It presents the improved ability of the CCR-SMOTETL algorithm to identify positive instances with the enhancement of *F-1* value in most circumstances.

2) Comparison of TPR and TNR

By evaluating the performance of classification used by the values of *TPR* and *TNR*, it can be detected that the CCR-SMOTETL algorithm improves *TPR* the top two highest. The CCR-SMOTETL algorithm performs best on *TPR* for the Haberman dataset among all the algorithms. It remains the top two algorithms for the testing of the Blood dataset. There is no obvious improvement on *TNR,* which could be improved in further research.

In summary, the CCR-SMOTETL method can effectively improve the overall classification performance and the accuracy of identifying positive samples using the Random Forest classifier by solving the imbalanced data set problem.

V. Conclusion

Data imbalanced problem remains one of the difficulties for many classification algorithms. A hybrid sampling CCR-SMOTETL method based on CCR, SMOTE and Tomek Links is proposed to solve the problem. The CCR-SMOTETL selects minority samples for generation using the CCR method and it generates every synthetic sample in a limited range of space by comparing the energy sphere radius and the calculated distance regarding to the centre of minority samples used in LR-SMOTE. The CCR-SMOTETL algorithm solved the problem of data imbalance by hybrid sampling and noise removal. The results from experiments present that by the CCR-SMOTETL algorithm, the classification problem of imbalanced data is effectively solved.

Despite this method shows a good performance according to the Table III, it has some limitations as well. Firstly, its performance on TPR and TNR is not stable comparing to previous algorithms. Some adjustments should be placed on the algorithm for a stable improvement on TPR and TNR values. Secondly, the algorithm is not designed for datasets that have attributes other than number. In the experiment, it has only been tested on two number-based datasets without any other types of attributes. Finally, it can only deal with binary classification problem, which has a restriction

Acknowledgment

This work is partially supported by: National Natural Science Foundation of China (61603223), Jiangsu Provincial Qinglan Project, Research Development Fund of XJTLU (RDF-18-02-30, RDF-20-01-18), Key Program Special Fund in XJTLU (KSF-E-34) and The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJB520034).

References

[1] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," vol. 6, no. 1, pp. 20-29, 2004.

[2] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," Springer Berlin Heidelberg, 2009, pp. 475-482.

[3] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," vol. 16, pp. 321-357, 2002.

[4] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*, 2005, pp. 878-887: Springer.

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008, pp. 1322-1328: IEEE.

[6] X. Liang, A. Jiang, T. Li, Y. Xue, and G. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Systems,* vol. 196, p. 105845, 2020.

[7] I. Tomek, "Two modifications of CNN," 1976.

[8] M. Koziarski, M.Wożniak, and C. Science, "CCR: A combined cleaning and resampling algorithm for imbalanced data classification," vol. 27, no. 4, pp. 727-736, 2017.