Enhanced Hybrid Sampling Technique for Classifying Unevenly Distributed Data to Forecast Student Performance

Mohamed Bellaj and Ahmed Ben Dahmane

[1,2] Abdelmalek Essadi University, High Normal School, Tetouan, Morocco
mohamed.bellaj@etu.uae.ac.ma

Abstract. Training an imbalanced dataset may cause classifiers to overfit the majority class, raising the risk of information loss for the minority class. One of the most common roadblocks to establishing successful classification and prediction systems is an uneven distribution of classes in the data obtained. Furthermore, accuracy may not be a good predictor of the classifier's performance. This study employed SMOTE, SMOTENC, SVM-SMOTE, Tomek Links, Edited Nearest Neighbors, and NearMiss. The Random Forest (RF) classifier is used for each resampling technique to explore its utility in dealing with the unbalanced data problem and predicting students' graduation grades based on their performance. In the first stage, four performance criteria are utilized to assess the influence of class imbalance on educational data. According to the findings, an RF model was employed to compare the outcomes before and after resampling the data. However, in terms of balanced accuracy and sensitivity, random forest with a random undersampling dataset outperformed all other techniques. When each performance parameter was studied separately, the findings indicated that the classifier in question performed better in terms of specificity, accuracy. Nonetheless, employing the complete resampled dataset enhanced RF classifier performance in terms of balanced accuracy and sensitivity.

Keywords: imbalanced dataset, Random Forest, undersampling, oversampling.

1 Introduction

Classification is the process of training a system with labeled datasets to categorize fresh, previously unseen data into certain classes. The increase in data volume is coupled with a lack of high-quality labeled data. Traditional machine learning algorithms frequently assume equal class distributions, which is wrong in many applications, including weather forecasting, sickness diagnosis, and fraud detection. In some circumstances, imbalanced datasets in which one class dominates cause models to favor the dominant class while ignoring the minority class. This imbalance, known as the class imbalance problem, has a severe effect on model performance, resulting in unsatisfactory results in assessment metrics such as precision, recall, F1-score, and ROC score.

There is a growing interest in tackling the class imbalance issue, which many researchers regard as a difficult subject that requires additional attention [1]. A typical

M. Khaldi et al. (eds.), Proceedings of the E-Learning and Smart Engineering Systems (ELSES 2024), Atlantis Highlights in Social Sciences, Education and Humanities 36,
https://doi.org/10.2991/978-2-38476-408-2_20

tactic is to use resampling techniques to balance datasets, which can be achieved by either undersampling or oversampling. Undersampling reduces the majority of class instances by using methods such as Tomek's linkages [2] and cluster centroids. Oversampling, on the other hand, entails increasing the number of cases of the minority class, which can be performed by creating new instances or repeating current ones. Borderline-SMOTE[3] is one example of an oversampling method.

The paper is structured as follows: This section presents the approaches used to address an unbalanced dataset. In the second section, we discuss relevant studies. And we present the methodology used in section3.Section 4 begins a discussion after presenting the experiment analysis and evaluating the unbalanced dataset. Section 5 marks the study's conclusion.

2 Related Works

Imbalanced data is one of the most impactful performance challenges for classifiers, particularly on large datasets. Several studies in recent years have given unique solutions to the problem of data imbalance in a hybrid fashion that incorporates more than one intelligent and statistical strategy.

[4] proposes a powerful preprocessing strategy that combines SMOTE with Tomek connections for imbalanced medical data from three disorders. When compared to eight classifiers, our technique outperformed SMOTE alone, demonstrating significant gains in 31, 27, and 30 out of 32 evaluation parameters for diabetes, Parkinson's, and vertebral column, respectively. [5] offer an undersampling method for addressing class imbalance in binary datasets by focusing on possibly overlapped occurrences. Our approaches accurately locate and delete majority class instances from overlapping regions, increasing minority class visibility while reducing information loss. The study describes four strategies for locating overlapped occurrences using different criteria based on neighborhood searching. Extensive trials on simulated and real-world datasets reveal comparable overall performance with cutting-edge approaches and exhibit extraordinary and statistically significant gains. SMOTE [6] is an oversampling technique that generates synthetic instances in the feature space from the original instance and its K-nearest neighbors. This approach prevents overfitting and helps classifiers determine decision boundaries between classes. The paper examines the challenges and issues associated with classifying imbalanced data, evaluates performance in such circumstances, analyzes recent extensions of SMOTE, and closes by noting current challenges and offering paths for future research in learning with imbalanced data. [7] Experiments show that models utilizing the SMOTE-ENC technique outperform those using SMOTE-NC, particularly in datasets containing a large number of nominal features and connections between categorical characteristics and the target class. Furthermore, our suggested method addresses a significant restriction of the SMOTE-NC algorithm, which is limited to mixed datasets with both continuous and nominal characteristics and cannot function when all features are nominal. Our unique approach has been extended to include both mixed datasets and datasets that are entirely made up of nominal features. In [8], the model uses a deep learning strategy to predict employee outcomes

while adapting to dataset balance. The algorithm entails data collection, exploration, pre-processing, and feature selection for training models. To solve class imbalance, a suggested deep learning method uses SVM SMOTE and incorporates a Bias initializer in the output layer. The achieved findings show remarkable metrics, with accuracy, recall, precision, and f1 Score reaching 89%, 92%, 94%, and 93%, respectively.

To solve class imbalance, the [9] study used SMOTE and SMOTE-Tomek algorithms. The results show that SVM with SMOTE-Tomek had the highest accuracy (98.52%), outperforming the DT model. These findings are useful for game companies and players looking for recommendations based on enhanced model performance. In addition, this work merged two sampling techniques[10], oversampling for the minority class and undersampling for the majority class, using SMOTE (Synthetic Minority Oversampling Technique) and ENN. This strategy, used in the Random Forest classifier, effectively balances imbalanced data, resulting in the highest classification accuracy of 93% for normal, suspect, and pathological data. The suggested technique beats other classifiers in parameters such as sensitivity and specificity, demonstrating its suitability for Cardiotocography data classification.

3 Methodology

3.1 Datasets Collection

Table 1. The data set attributes.

Attribute	Description
Gender	the student's gender {f or m}
Nationality	student Nationality
Relation	contact parent, specified as either "father" or "mother."
StageID	the stage or academic level the student belongs to, is categorized as "Lower level," "Middle level," or "High level."
GradeID	grade {G-01 to G-12}
SectionID	class division that the student belongs to "A," "B," or "C."
Topic	Student's enrolled course
StudentAbsenceDays	absence {>7, <7}
Semester	categorized as "1" or "2"
RaisedHands	raised hands in class (numeric: 0-100).
VisitedResources	visits of course content {numeric: 0-100}
AnnouncementsView	frequency of new announcement checks by the student (numeric: 0-100).
Discussion	number of discussion groups (numeric: 0-100).
ParentSchoolSatisfaction	parent satisfaction degree with school: "Good" or "Bad."
ParentAnsweringSurvey	parent participation in school surveys: "Yes" or "No."

Class	Total score classification: "Low-Level" (0-69), "Middle-Level" (70-89), "High" (90-100).

3.2 Tools and Instruments

The study conducted experiments in Python and Jupyter Notebook. The Scikit-learn library was used for machine learning, and the Imblearn module addressed class imbalance.

3.3 Preprocessing

Preprocessing, which is essential in data mining, improves data quality and modeling efficiency[11]. Initial feature filtering, using Python's Scikit-learn preprocessing tool, eliminates unneeded data. Categorical data is then translated to numerical values to assess its eligibility for modeling.

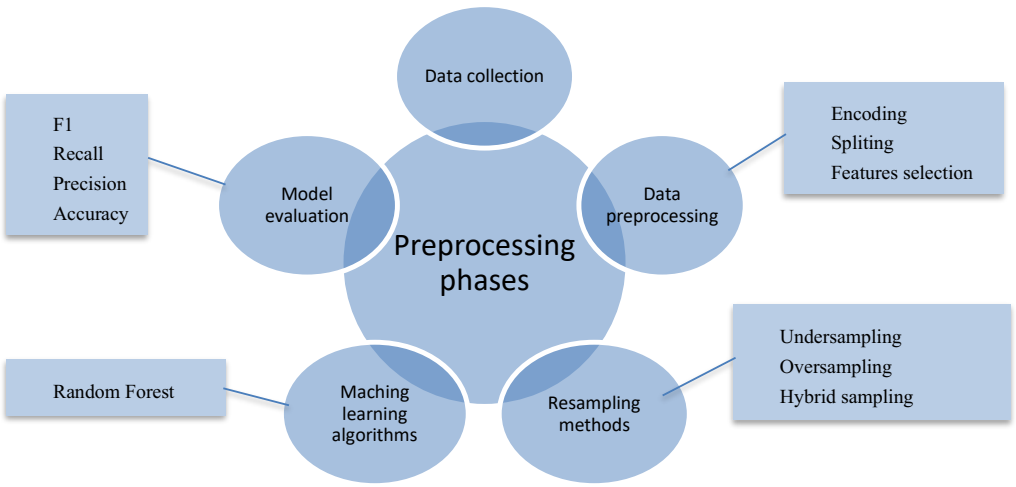

Fig. 1. The preprocessing phases.

3.4 Data encoding

Label encoding is essential in machine learning when working with datasets that contain labels in a variety of formats, such as letters or numerical values. This approach converts these labels into numerical representations, making them usable in machine-learning models [12]. This preprocessing phase is especially important for supervised learning techniques because it ensures proper communication between the model and the labeled data.

3.5 Feature Selection

In the context of LightGBM (Light Gradient Boosting Machine), feature importance is a critical tool for determining the impact of various features or variables on model predictions[13]. LightGBM uses the Gain technique, also known as Split Importance, to calculate each feature's proportional contribution. This is determined by computing the gain, which indicates how much a feature improves the model's accuracy on the branches it affects. A larger gain value for a certain trait suggests that it is more important for generating predictions. Essentially, gain quantifies how much a feature improves the model's predictions, with a focus on the quality of the splits it generates within the decision trees.

3.6 Resampling Technique

Resampling corrects the data imbalance by equating unequal classes. Three sampling methods [14] include:

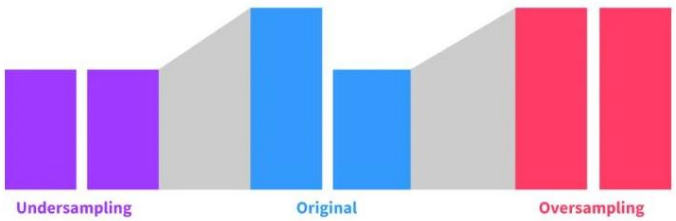

Fig. 2. Resampling methods explanation

Initially, the dataset is balanced by utilizing strategies for addressing imbalanced data, including oversampling, undersampling, and a hybrid approach. Machine learning techniques are then applied to the balanced dataset to generate results. Finally, the results of machine learning algorithms on the imbalanced dataset are compared to those on the balanced dataset, using the appropriate imbalanced dataset management strategies.

Oversampling Techniques.

Oversampling is a method for addressing imbalanced classes by adding points to the minority class, which is critical in fraud detection datasets. Many studies suggest oversampling strategies to address extremely imbalanced class difficulties, such as SMOTE, ADASYN, BorderlineSMOTE, SVM-SMOTE, and so on. Chawla[15] created SMOTE, a cutting-edge oversampling strategy that creates synthetic instances along line segments connecting positive examples to their k-nearest Neighbors (KNN) in the minority class. This entails selecting a random instance along a line segment connecting two distinct attributes.

Undersampling Techniques.

.

Due to the vast quantity of the dataset, previous researchers used a variety of undersampling strategies to analyze students' academic achievement. However, utilizing undersampling may result in data and information loss. Researchers advise against undersampling for severely skewed datasets of student performance, instead recommending oversampling strategies to avoid information loss [16]. Random sampling, TomkeLinks, and other undersampling approaches are commonly employed to overcome unbalanced dataset concerns, however, they should be avoided for extremely imbalanced datasets due to the danger of underfitting and information loss[17].

Hybrid Techniques

To overcome the disadvantages of both under- and over-sampling, the hybrid technique combines the two. Under-sampling risks losing vital information, whilst oversampling may result in overfitting. Techniques such as SMOTE-ENN combine SMOTE for over-sampling and ENN for under-sampling, whereas SMOTE-Tomek uses SMOTE for over-sampling and Tomek links for under-sampling. These strategies try to balance the training dataset, remove noisy points, refine clusters, and improve the models' generalization abilities.

3.7 Classification using Machine Learning Classifiers

Random Forest

Pooling various DT in the RF ensemble reduces the model's variance [19]. Averaging forecasts in an RF can help generate predictions for unknown samples.

$$I = \frac{1}{N}\sum_{N=1}^{N} \int (x) \quad (1)$$

The Random Forest (RF) technique examines data with a series of Decision Trees (DTs), integrating predictions to find the best course of action. Operating within an ensemble learning framework and applying the bagging technique, RF demonstrates robustness and can efficiently manage datasets with missing values.

Performance Measure

The approach's assessment measures include accuracy, the F-measure, and the ROC curve. Accuracy is the successful identification of instances across all performance classes (low, medium, and high). F-measure combines precision and recall to maximize their benefits. Precision is the fraction of recovered instances that are relevant, whereas recall is the proportion of retrieved and relevant examples divided by the total number of relevant instances [20].

$$\text{Accuracy}=(TP+TN)/(TP+FP+FN+TN) \quad (2)$$

$$\text{Precision}=TP/(TP+FP) \tag{3}$$

$$\text{Recall}=TP/(TP+FN) \tag{4}$$

$$\text{F1-score}=(2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall}) \tag{5}$$

4 Result and Discussion

The study proceeded through four experimental stages: (1) feature selection, (2) testing with different machine learning algorithms, (3) testing with oversampling and undersampling approaches, and (4) testing with a hybrid approach (oversampling-undersampling). Initially, each method's performance was evaluated using the Random Forest classifier. Testing was then performed on the three resampling data types chosen (oversampling, undersampling, and hybrid sampling).

4.1 Feature selection

Feature Selection is a strategy for limiting model input variables to useful data, hence reducing noise. The LGBM classifier was initially used to evaluate variable correlation strength and establish relevance rankings, hence decreasing the total number of features and comparing training times. Following feature selection, the emphasis switched to the eight most important parts of training and testing.

4.2 Distribution of the data

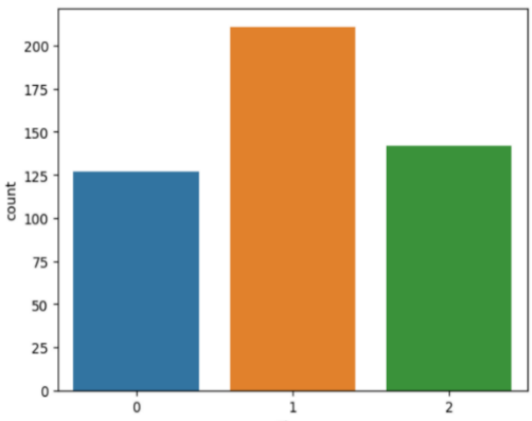

Fig. 3. Data distribution

The graphic depicts the dataset's distribution before applying any resampling algorithm. Non-resampled data contains both the majority and minority classes. In the highly imbalanced dataset, "0" indicates students at the low level (127 students), "1" for the intermediate level (211 students), and "2" for the high level (142 students).

4.2.1 Evaluating results using machine learning algorithms without resampling methods

Comparison of the performance metrics after oversampling.

This subsection compares the performance of the RF algorithm without resampling against the three resampling approaches.

Table 2. The performance results after using undersampling methods.

Model		F1	Precision	Accuracy	Roc-Auc	Recall
RF	Test	0.77	0.77	0.77	0.92	0.77
	TomekLinks	0.74	0.73	0.73	0.91	0.73
	EditedNearest-Neighbours	0.69	0.72	0.69	0.91	0.69
	NearMiss	0.74	0.77	0.72	0.90	0.72

According to the results in Table 2, the RF ensemble model performs better than undersampling methods. The employment of the undersampling approach with RF results in a lower outcome than using solely the RF algorithm during the test phase. This implies that RF has the best performance across all measurements, but the NearMiss technique has the same result in terms of precision metric (0.77).

Comparison of the performance metrics after undersampling.

Table 3. Performance results after employing oversampling methods.

Model		F1	Precision	Accuracy	Roc-Auc	Recall
RF	Test	0.77	0.77	0.77	0.92	0.77
	SMOTE	**0.80**	**0.80**	**0.80**	**0.93**	**0.80**
	SMOTENC	0.75	0.74	0.74	0.92	0.74
	SVMSMOTE	0.76	0.76	0.76	**0.93**	0.76

This study compares the performance of three oversampling techniques: SMOTE, SMOTENC, and SVMSMOTE. The strategy without resampling, which means no established method of processing imbalance, was also employed to compare baseline performance. According to the results in Table 3, SMOTE paired with RF performed better on all measures, followed by SVMSMOTE.

SMOTE underperformed on the moderately unbalanced dataset, as seen by its highest overall performance measure scores when using RF (accuracy = 0.80, precision = 0.80, recall = 0.80, F1-score = 0.80, and ROC-AUC = 0.83). Using RF yields a poorer performance metric for SMOTENC (accuracy = 0.74, precision = 0.74, recall = 0.74, ROC-AUC = 0.92, and F1-score = 0.75).SVMSMOTE produces modest results in all parameters except (ROC-AUC = 0.93).

Comparison of the performance metrics after hybrid sampling.

Table 4. The performance results after using hybrid sampling methods.

Model		F1	Precision	Accuracy	Roc-Auc	Recall
RF	Test	0.77	0.77	0.77	0.92	0.77
	SMOTETomek	0.75	0.74	0.74	**0.93**	0.74

SMOTEENN	0.65	0.74	0.65	0.91	0.65

Table 4 displays all of the results achieved with the RF algorithm alone, without any resampling approach, to compare them to the other outputs. After utilizing the hybrid resampling method SMOTETomek and SMOTEENN, the lowest metric results were achieved by SMOTEENN combined with RF (accuracy = 0.65, precision = 0.74, recall = 0.65, F1-score = 0.65, ROC-AUC = 0.91). When RF is used in conjunction with SMOTETomek, the Roc-Auc=0.93 measure improves the most compared to other metrics (accuracy = 0.75, precision = 0.74, recall = 0.74, and F1-score = 0.74).

The study found that oversampling, undersampling, and hybrid sampling had a substantial impact on categorization performance. The RF classifier performed well when SMOTE was used for oversampling. However, undersampling strategies such as NearMissb, EditedNearestNeighbours, and NearMiss produced the lowest results when compared to non-resampled data. Furthermore, using all hybrid techniques produced moderate results with no noticeable performance improvement.

5 Conclusion and Future Works

This research demonstrated the efficiency of data sampling strategies in developing prediction models for unbalanced water quality data. The approaches largely use preprocessing techniques such as under-sampling, oversampling, and hybrid resampling to convert an unbalanced dataset into a balanced dataset. The balanced datasets are then used to train three machine-learning classification models.

The unbalanced dataset was used to examine and assess the issue. The dataset was examined with a variety of resampling strategies, including oversampling (SMOTE, SMOTENC, and SVMSMOTE), undersampling (TomekLinks, EditedNearestNeighbours, and NearMiss), and hybrid resampling (SMOTETomek and SMOTEENN). The classifiers were tested using a variety of criteria, and SMOTE consistently outperformed the Random Forest (RF) classifier when paired with numerous resampling methods. Our plans involve comparing various deep-learning approaches and resampling strategies.

References

1. V. Sampath, I. Maurtua, J. J. Aguilar Martin, and A. Gutierrez, “A survey on generative adversarial networks for imbalance problems in computer vision tasks,” J. big Data, vol. 8, pp. 1–59, 2021.
2. G. Yang and L. Qicheng, “An over sampling method of unbalanced data based on ant colony clustering,” IEEE Access, vol. 9, pp. 130990–130996, 2021.
3. H. Al Majzoub, I. Elgedawy, Ö. Akaydın, and M. Köse Ulukök, “HCAB-SMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification, ” Arab. J. Sci. Eng., vol. 45, no. 4, pp. 3205–3222, 2020.

4. M. Saul and S. Rostami, "Assessing performance of artificial neural networks and re-sampling techniques for healthcare datasets," Health Informatics J., vol. 28, no. 1, p. 14604582221087108, 2022.
5. R. FENG, H. JI, Z. ZHU, and L. WANG, "A Wasserstein Distance-Based Cost-Sensitive Framework for Imbalanced Data Classification," Radioengineering, vol. 32, no. 3, p. 451, 2023.
6. A. Kanwal, "Handling Class Imbalance Data using Class Label Prediction." CAPITAL UNIVERSITY, 2020.
7. M. Awasthi, V. Sajwan, P. Awasthi, A. Goel, and R. Kumar, "Performance Efficacy of Cost-Sensitive Artificial Neural Network: Augmenting the Results of Imbalanced Datasets in Supervised and Unsupervised Learning," in Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2022, 2022, pp. 305–322.
8. D. A. Neu, J. Lahann, and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," Artif. Intell. Rev., pp. 1–27, 2022.
9. E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," Sensors, vol. 22, no. 9, p. 3246, 2022.
10. A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Synthetic minority oversampling using edited displacement-based k-nearest neighbors," Appl. Soft Comput., vol. 148, p. 110895, 2023.
11. I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to continuous quality management," J. Big Data, vol. 8, no. 1, pp. 1–41, 2021.
12. J. Rajala, "Multi-label classification for industry-specific text document sets." 2021.
13. H. Padalko, V. Chomko, S. Yakovlev, and D. Chumachenko, "Ensemble machine learning approaches for fake news classification," Radioelectron. Comput. Syst., no. 4, pp. 5–19, 2023.
14. J. Ren, Y. Wang, M. Mao, and Y. Cheung, "Equalization ensemble for large scale highly imbalanced data classification," Knowledge-Based Syst., vol. 242, p. 108295, 2022.
15. D. Dablain, B. Krawczyk, and N. V Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," IEEE Trans. Neural Networks Learn. Syst., 2022.
16. L. Sha, M. Raković, A. Das, D. Gašević, and G. Chen, "Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education," IEEE Trans. Learn. Technol., vol. 15, no. 4, pp. 481–492, 2022.
17. K. M. Ghori, M. Awais, A. S. Khattak, M. Imran, and L. Szathmary, "Treating class imbalance in non technical loss detection: An exploratory analysis of a real dataset, " IEEE Access, vol. 9, pp. 98928–98938, 2021.
18. K. Piotr and L. Rongbing, "Comparative Analysis of Sampling Methods for Imbalanced Classification." uis, 2023.
19. G. Garg and R. Garg, "Brain tumor detection and classification based on hybrid ensemble classifier," arXiv Prepr. arXiv2101.00216, 2021.
20. A. M. Carrington et al., "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 329–341, 2022.