Leveraging Artificial Intelligence for Enhanced Data Generation in Addressing Imbalance in Binary Classification System

Srijita Bandopadhyay
Dept. of Electronics and Communication Engineering
University of Engineering and Management
Kolkata, India
srijitabando30@gmail.com

Srimonti Dutta
School of Agro and Rural Technology
Indian Institute of Technology Guwahati
Guwahati, India
srimonti@iitg.ac.in

Imran Haider*
Dept. of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
ihalig14@iitkgp.ac.in

Bhavaraju Anuraag
Dept. of Electronics and Communication Engineering
Vellore Institute of Technology
India
bhavarajuanuraag@gmail.com

Jerry Zhu
Faculty of Mathematics
University of Waterloo
Ontario, Canada
j25zhu@uwaterloo.ca

Saad Ahmed Bazaz
Dept. of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan
i180621@nu.edu.pk

***Abstract*—This paper delves into the challenges of binary classification using imbalanced datasets, particularly when instances of interest are infrequent. It explores a comprehensive approach that integrates Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs) to enhance classification outcomes. Traditional classification models tend to favor the majority class, while the impact of imbalanced misclassification costs is often overlooked. The integration of SMOTE, GANs, and VAEs in binary classification, or SMOTE-GAN-VAE, addresses these challenges by generating synthetic instances, refining data representations, and capturing latent features. To evaluate the effectiveness of various data generation methods, a credit card fraud dataset is used. The performance metrics considered include F0.5-score, F1-score, and F2-score, which account for both precision and recall. The results indicate that SMOTE-GAN-VAE outperforms individual methods, such as SMOTE, GANs, and VAEs, demonstrating its potential to enhance data representation and classification accuracy, and outperformed the β- VAE filtered approach employed in previous literature.**

***Index Terms*—Data generation, Imbalanced data, Binary classification**

I. Introduction

The binary classification task, which involves assigning instances to one of two possible classes, is a fundamental component in the fields of machine learning and data analysis. The achievement of accurate and robust binary classification holds significant importance in a wide range of practical scenarios, encompassing various fields such as medical diagnostics and fraud detection. Nevertheless, the efficacy of classification models is consistently impacted by intrinsic obstacles, notably class imbalance and the distinctive costs of misclassification associated with different classes.

The presence of class imbalance, which refers to a substantial difference in the number of instances between two classes, presents a significant challenge in classification tasks. Conventional classification algorithms are commonly formulated to maximize overall accuracy, resulting in an unintentional inclination towards the majority class. As a result, the effectiveness of these models greatly decreases when it comes to identifying instances belonging to the minority class. Furthermore, when the misclassification of instances

from one class carries greater implications than the other, traditional learning frameworks that do not consider the costs involved are inadequate in effectively managing the balance between the two classes.

Within this particular context, there has been a notable focus on combining the Synthetic Minority Oversampling Technique (SMOTE), Generative Adversarial Network (GAN), and Variational Autoencoder (VAE). The integration of Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GAN), and Variational Autoencoders (VAE) presents a robust methodology for addressing the issues associated with imbalanced datasets. The Synthetic Minority Oversampling Technique (SMOTE) enhances the dataset by generating synthetic instances of the minority class. Generative Adversarial Networks (GANs) produce realistic samples by training a generator network to mimic the distribution of the training data. Variational Autoencoders (VAEs) capture latent features of the data by learning a probabilistic model. The combined utilization of these techniques enhances the robustness and accuracy of models across diverse applications. This integration promotes a comprehensive representation of data and enhances the efficacy of machine learning outcomes.

Additionally, while several research studies have primarily concentrated on generating image data [1], [2], there exists a noticeable scarcity of research on the generation of tabular data. In addressing the challenge posed by imbalanced table data and aiming to enhance prediction accuracy, a limited number of studies have opted to amplify the presence of minority class data by leveraging data generation methods [3], [4]. Notable among these methodologies are SMOTE [5], GAN [6], and β-VAE [7]. However, it's worth noting that while SMOTE operates as an oversampling technique, it's incapable of producing new data instances with distinctive features beyond those present in the original dataset. In contrast, GAN, an extension of the GAN framework tailored for table data, could be susceptible to issues of learning instability and biased data generation. Conversely, the β-VAE approach introduces the hyperparameter β into the objective function of the Variational Autoencoder (VAE), facilitating the separation of features within the latent space.

The contribution of this paper is as follows:

- This work uses the three-stage framework of data generation algorithm (SMOTE-GAN-VAE) and is compared with the β-VAE filtering approach employed in [8].
- A random forest classifier is employed to evaluate the performance measures.

The rest of the paper is structured as follows: Section II discussed the related works that used different data generation algorithms, Section III discusses the solution methodology, followed by the results and discussions in Section IV. Finally, the study is concluded in Section VI.

II. Related works

[8] presented the concept of Filtered β-VAE as a method to improve the precision of adverse event prediction by generating patient data. The suggested technique integrates additional filtering measures into the data generation process of β-VAE. The purpose of incorporating these supplementary filtration measures is to effectively eliminate minority class data that possesses limited relevance in the context of adverse event prediction. The filtration process consists of two distinct stages: initially, reconstruction error assessment is employed, followed by the application of a machine learning technique. The Filtered β-VAE approach generates novel data instances with distinct attributes compared to the original dataset by eliminating low-quality data. This leads to an improvement in the predictive precision of the model.

Research gaps:

The evaluation of the suggested approach was conducted by [8] using a specific collection of machine learning algorithms. However, a need remains to assess its performance against other contemporary data generation techniques that represent the current state of the art.

III. Solution Methodology

The methodology adopted in this work is shown in Figure 1. A detailed description of each component in the workflow is mentioned in the subsections below.

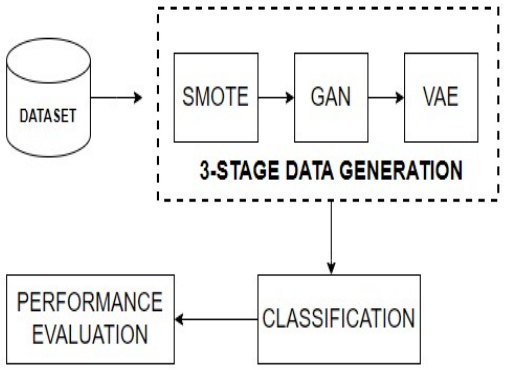

Fig. 1: Workflow Diagram

A. Dataset

In this paper, the imbalance dataset which is used for generating synthetic data is collected from Kaggle. A credit card fraud benchmark dataset is implemented to evaluate the three-stage data generation method analyzed in this work. The dataset is an imbalanced dataset containing information about credit cards along with the target column. The target column is seen to have more minority class than the majority ones hence this work is best suited for this type of dataset.

B. Data Generation

The presented work proposes a new approach for generating synthetic data. The framework represented three stages in a series. In the first stage, Synthetic Minority Over-Sampling technique is applied to the training data. It helps generate synthetic samples to balance the minority class over the majority ones. The second stage is developed using Generative Adversarial Network (GANs) which generates realistic samples. In the third stage, Variational AutoEncoder is applied to capture latent state representations.

C. Classification

Since the dataset is labeled, a supervised learning technique is employed. For this purpose, the Random Forest algorithm, an ensemble learning technique, is chosen. It is capable of handling complex datasets and mitigating issues such as overfitting. An ensemble of decision trees is constructed where each tree learns patterns within the data resulting in highly accurate and robust classifications.

D. Performance Evaluation

Choosing the right evaluation metrics is pivotal while building classification models. Accuracy can be a misleading metric in cases of imbalanced datasets. For the dataset utilized in this study which is related to credit card fraud, both precision and recall are important. Hence the performance of the classifier after the generation of synthetic data is tested using three parameters F1, F0.5 and F2 scores. The following metrics are calculated from the value obtained from T_p(True Positive), T_n(True Negative), F_p(False Positive) and F_n(False Negative). Equation (1), (2), and (3) describes the parameters which are harmonically weighted averages of precision and recall. The F-scores encapsulate the collective performance of both Recall and Precision [9].

$$Precision = \frac{T_\mathrm{p}}{T_\mathrm{p} + F_\mathrm{p}} \tag{1}$$

$$Recall = \frac{T_\mathrm{p}}{T_\mathrm{p} + F_\mathrm{n}} \tag{2}$$

$$F\beta = \frac{(1+\beta^2) * Precision * Recall}{\beta^2 * Precision * Recall} (\beta = 0.5, 1, 2) \tag{3}$$

IV. Results and Discussions

In the following section, we shall proceed to present the performance measures that have been derived from a range of data generation methods, all of which have been employed for the purpose of binary classification using random forest classifier. The evaluated measures encompass the F0.5-score, F1-score, and F2-score, which are extensively employed metrics for evaluating the efficacy of classification models, particularly in the context of imbalanced datasets.

Table I presents a comprehensive overview of the performance metrics attained by each respective method. The methods that have been assessed cover a range of data augmentation techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Network (GAN), Variational Autoencoder (VAE), and a hybrid approach that combines these techniques, referred to as SMOTE-GAN-VAE. Furthermore, we examine a modified version of the Variational Autoencoder (VAE) known as the β-filtered VAE employed in [8].

TABLE I: Performance measures obtained from different methods

Data generation methods	F0.5-score	F1-score	F2-sore
SMOTE	0.859	0.831	0.845
GAN	0.41	0.30	0.34
VAE	0.453	0.378	0.401
β-filtered VAE	0.453	0.378	0.401
GAN-VAE	0.849	0.74	0.671
SMOTE-GAN-VAE	0.867	0.834	0.849

The findings reveal significant discrepancies in the efficacy of these approaches. Among the various individual techniques, Synthetic Minority Over-sampling Technique (SMOTE) demonstrates favorable results, attaining an F0.5-score of 0.859, an F1-score of 0.831, and an F2-score of 0.845. In contrast, the performance of the GAN method is relatively inferior, as indicated by an F0.5-score of 0.41, an F1-score of 0.30, and an F2-score of 0.34. The aforementioned result implies that the utilization of GAN-based data augmentation may encounter difficulties in accurately representing the intricacies of the dataset, resulting in less-than-optimal classification outcomes.

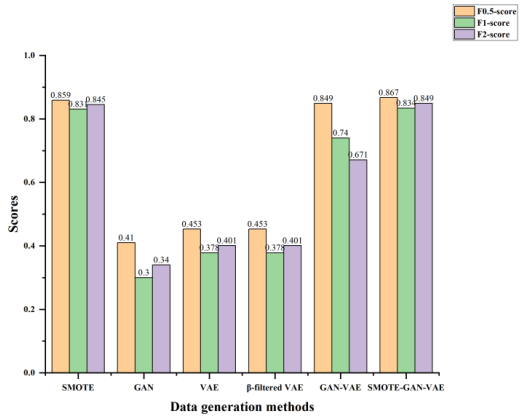

Fig. 2: Graphical Representation of the performance measures.

Similarly, the VAE approach produces intermediate outcomes, attaining an F0.5-score of 0.453, an F1-score of 0.378, and an F2-score of 0.401. It is noteworthy that the β-filtered variational autoencoder (VAE), which was developed to tackle specific challenges related to the latent space representations of VAEs, yields outcomes that align with those of the conventional VAE.

An ablation study is also performed using two-stage framework comprising of GANS and VAE. The method resulted in an average score of F0.5 score of 0.849, F1-Score of 0.74 and F2-SCore of 0.671. The study demonstrates the effect of results without using SMOTE.

However, the evaluation results indicate that the combined approach, SMOTE-GAN-VAE, demonstrates superior performance compared to other methods. The three-stage method demonstrates an F0.5-score of 0.867, an F1-score of 0.834, and an F2-score of 0.849, surpassing the performance of the individual methods and providing further evidence of the advantageous nature of combining multiple data generation techniques.

The aforementioned results underscore the significance of carefully choosing a methodology when addressing the issue of data imbalance. The effectiveness of individual techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and VAE (Variational Autoencoder) varies. However, combining these techniques in the suggested SMOTE-GAN-VAE integration shows improved performance. Moreover, the discrepancies that have been observed serve to highlight the importance of taking into consideration a variety of evaluation metrics. This is because different metrics place emphasis on different aspects of classification performance.

V. Conclusion

Within the domain of binary classification, where the accurate assignment of instances to separate classes holds significant significance, the persistent challenges of class imbalance and fluctuating misclassification costs continues to pose a challenge. The objective of this study was to investigate and compare various data generation algorithms in order to address the challenges associated with classification models. The study explored the complex interaction between Synthetic Minority Oversampling Technique (SMOTE), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs) with the aim of developing a comprehensive solution. In summary, this research conducted a comprehensive investigation into data generation algorithms in order to tackle the inherent difficulties associated with binary classification tasks, specifically when faced with class imbalance and diverse misclassification costs. The empirical assessment of various algorithms provides insights into their respective effectiveness in improving classification performance.

The study acknowledged the intrinsic bias of traditional classification algorithms towards the dominant class, consequently undermining their capacity to detect instances from the underrepresented minority class accurately. Furthermore, situations that involve imbalanced misclassification costs require more sophisticated methodologies than what conventional frameworks provide.

The empirical findings highlighted the variations in performance observed among the assessed data generation algorithms. The SMOTE technique demonstrated positive results by enhancing classification models by generating synthetic instances. On the other hand, Generative Adversarial Networks (GANs) encountered difficulties in accurately capturing complex characteristics of the dataset, resulting in relatively inferior performance in classification tasks. The variational autoencoder (VAE), situated between these two extremes, successfully captured latent features while not causing substantial modifications to classification results. The beta-filtered variational autoencoder (VAE), which was specifically developed for the purpose of refining the latent space, exhibited results that were comparable to those of the conventional VAE. Furthermore, the two stage framework using GANS and VAE shows considerable results in data generation.

Integrating SMOTE, GANs, and VAEs in the proposed SMOTE-GAN-VAE framework demonstrated improved classification performance, marking a significant breakthrough. The amalgamation exhibited significant

enhancements in multiple F-scores, suggesting its ability to improve data representation and performance measures.

This study presents a comprehensive framework for binary classification and highlights the importance of incorporating performance metrics beyond conventional measures to enhance our comprehension of these integrated methodologies. The utilization of the SMOTE-GAN-VAE methodology has the potential to revolutionize the management of imbalanced data, leading to the development of classification models that are characterized by enhanced accuracy and reliability.

This research significantly enhances our understanding and application of binary classification in imbalanced datasets, making valuable contributions to both theoretical and practical domains. Through a systematic examination of algorithms and the empirical validation of their performance, this research provides valuable insights that can inform future investigations and facilitate the implementation of these algorithms in real-world scenarios. The presented approach in this study not only tackles the challenges related to imbalanced data but also contributes to advancing knowledge in the field, indicating a more resilient framework for binary classification.

VI. Future Work

In contemplating the future trajectory of this research, several promising avenues emerge, particularly in the realms of optimization, domain-specific application, and real-time data augmentation. The current methodologies, which pivot around the SMOTE, GANs, and VAEs, present an opportunity for further refinement. The pursuit of enhanced efficiency in these techniques is not just a logical progression but a necessity to address the complexities inherent in imbalanced binary classification problems more effectively. Moreover, the potential applicability of the proposed framework within specific domains such as healthcare, finance, and cybersecurity warrants a thorough investigation. Tailoring the approach to these distinct fields could unveil unique challenges and opportunities, leading to more robust, specialized solutions that are acutely attuned to the nuances of each domain. Perhaps the most innovative direction for future exploration lies in the realm of real-time data augmentation. This would entail developing methodologies capable of generating synthetic data dynamically during the training phase of machine learning models. Such an approach, by allowing for an adaptive response to evolving data distributions, could significantly enhance the resilience and accuracy of classification outcomes. This future research trajectory, while building upon the foundational work presented, aims to push the boundaries of current data generation methodologies in imbalanced binary classification, thereby opening new horizons for practical applications and theoretical advancements in the field.

References

[1] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[2] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *arXiv preprint arXiv:1706.04987*, 2017.

[3] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE access*, vol. 9, pp. 39 707–39 716, 2021.

[4] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[6] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2016.

[8] Y. Yamasaki, C. Doi, S. Kitagawa, H. Seki, and H. Shigeno, "Data generation with filtered β-vae for the preoperative prediction of adverse events," *IEEE Access*, 2023.

[9] Y. J. Huang, R. Powers, and G. T. Montelione, "Protein nmr recall, precision, and f-measure scores (rpf scores): structure quality assessment measures based on information retrieval statistics," *Journal of the American Chemical Society*, vol. 127, no. 6, pp. 1665–1674, 2005.