Enhancing Predictive Performance through Optimized Ensemble Stacking for Imbalanced Classification Problems

Anthonia Oluchukwu Njoku
African Institute for Mathematical Sciences
Limbe, Cameroon
anthonia.njoku@aims-cameroon.org

Berthine Nyunga Mpinda
African Institute for Mathematical Sciences
Limbe, Cameroon
bmpinda@aimsammi.org

Olushina Olawale Awe
Global Humanistic University
Curacao
olawaleawe@gmail.com

Abstract

Ensemble methods have gained significant popularity in the field of machine learning due to their ability to improve predictive performance by combining multiple models. One ensemble technique that has shown promising results is ensemble stacking, which involves training a meta-model on predictions from multiple base models. This research focused on investigating and optimizing ensemble stacking while also incorporating tailored balancing techniques for imbalanced datasets. The study explored a variety of balancing strategies, including undersampling, oversampling, and hybrid approaches, to mitigate class imbalance. Two ensemble meta-learners were considered, evaluating their ability to capture the underlying class distributions and mitigate bias while maintaining overall model performance. The research findings will contribute to the development of optimized ensemble stacking techniques for addressing imbalanced classification challenges, enabling improved decision-making and performance in real-world applications.

1 Introduction

In the realm of machine learning, predictive performance plays a vital role in numerous real-world applications [18], ranging from medical diagnosis to bankruptcy prediction. However, most real world applications are usually imbalanced in their class distribution [9] [16] [6]. This has posed significant challenges in using traditional machine learning models when dealing with imbalanced classification problems, as many of these models tend to learn and better classify the class with the majority distribution [8][11]. These problems occur when the instances of one class greatly outnumber those of the other, leading to biased model performance and inaccurate predictions, particularly for the minority class [13]. To address these challenges caused by imbalance, this research is an exploration of various balancing techniques. Traditional balancing techniques such as undersampling and oversampling will be investigated, as well as hybrid approaches that combine the strengths of both.

Furthermore, we employ the use of ensemble stacking techniques which have emerged as a promising approach to enhance predictive performance by leveraging the collective power of multiple models [12]. Ensemble stacking involves training a meta-model on the predictions of a diverse set of base models [2], effectively combining their individual strengths to produce a more robust and accurate

5th Deep Learning Indaba Conference (DLI 2023).

prediction. This approach has shown remarkable success in various domains, including speech signals, medical diagnosis, and network securities [15] [12] [14].

We will examine different ensemble meta-learners within the ensemble stacking framework. By leveraging the diversity of base models and effectively combining their predictions, we aim to maximize the ensemble's predictive power while minimizing overfitting and bias. The selection of suitable meta-learners and base models will be guided by their ability to capture the intricacies of imbalanced datasets and optimize performance accordingly.

2 Related Literature

The ensemble stacking technique, which involves integrating the strengths of different machine learning models [10] [7], has been a focal point for numerous researchers aiming to enhance prediction performance. Several studies have successfully employed this approach to construct exceptionally effective models across a wide range of fields.

In their study, Esfar-E-Alam et al. (2022)[5] employed a multimodal methodology to classify emotions into four distinct classes. The research involved the utilization of six models for both audio and text domains, which were subsequently integrated through four heterogeneous ensemble techniques: hard voting, soft voting, mixing, and stacking. The results of this investigation indicate that the implementation of ensemble learning to combine modalities significantly improves classification accuracy. Among the ensemble strategies employed, stacking exhibited the highest performance, based on the selected set of models.

In order to address the challenge of imbalanced rockburst data, Yin et al. (2021)[17] employed the stacking ensemble learning technique to develop prediction models for rockburst occurrences. The data underwent preprocessing steps, including principal component analysis, local outlier factor, and expectation maximization algorithm. To mitigate the imbalance issue, stratified sampling was applied. The training process involved four traditional high-end models and four ensemble models constructed through stacking. The performance of all eight models was evaluated, and a comparative analysis was conducted to assess the disparities between the base models and ensemble models. Furthermore, the impact of class imbalance on prediction accuracy was examined, revealing that the ensemble stacking technique outperformed the individual models, making it an advantageous technique to use when dealing with imbalanced data.

Buyrukoglu and Savas (2023) [3], in their study, undertook the task of determining and classifying the positions of football players. They leveraged a stacked ensemble machine learning model, utilizing the FIFA'19 game dataset. To optimize performance, they employed four distinct feature selection algorithms to identify 10 relevant features. The individual base algorithms used in the stacked model included Deep Neural Networks, Random Forest, and Gradient Boosting, while Logistic Regression (LR) was employed as the meta-learner. The findings revealed that the combination of the Chi-square feature selection technique and the stacked-based ensemble learning model yielded the highest accuracy, perfoming with an 83.9% accuracy.

The proposed methodology of this study uses a stacked ensemble of machine learning models to improve the predictive performance of an imbalanced classification problem. This paper also investigates the effect of various balancing techniques in managing the data's class imbalance.

3 Empirical Evidence

Experiments in this study were carried out with a population-based nutrition data comprising 671 adolescents. This data was gotten from the South African National Health and Nutrition Examination Survey (SANHANES) in 2011/12 and it classified the weight status of each adolescent[1]. For use with base models, the data had to undergo balancing to avoid biases in model training, as it was originally imbalanced with an imbalance ratio of 0.278 [4]. Three balancing techniques were employed in this work-oversampling, undersampling, and hybrid-sampling. Given the imbalanced nature of the dataset, relying solely on accuracy as a performance metric would be insufficient, as it fails to account well for the negative class in its calculation. Therefore, we also considered performance metrics that incorporate the negative class, such as specificity and balanced accuracy, to provide a more comprehensive evaluation of model performance.

3.0.1 Models built on Imbalanced Data

Using Imbalanced data leads to problems in classifying the minor class as some models are not able to handle imbalance. This can be seen from the results in Table 1. We see from the sensitivity results that the models were able to predict correctly the normal weight but with the overweight samples, they had very few accurate predictions, as seen in the specifiity results. Some models like support vector machine (SVM), neural network (NN) and random forest (RF) did not make any overweight prediction. This buttresses the importance of balancing.

Table 1: Performance Metrics of Models with Imbalanced data.

Metrics	SVM	Bagging	NN	RF	LR
Accuracy	0.781	0.761	0.791	0.791	0.746
Sensitivity	0.987	0.931	1.000	1.000	0.862
Specificity	0.000	0.119	0.000	0.000	0.310
F1	0.877	0.860	0.883	0.883	0.843
Balanced Accuracy	0.494	0.525	0.500	0.500	0.586

3.0.2 Models built on Oversampled Data

After oversampling the data and training the models, we see that with oversampling, our models have improved in the prediction of the overweight class. This is shown from the increase in specificity and balanced accuracy results in Table 2.

Table 2: Performance Metrics of Models with Oversampled data.

Metrics	SVM	Bagging	NN	RF	LR
Accuracy	0.697	0.642	0.677	0.731	0.652
Sensitivity	0.792	0.742	0.774	0.887	0.660
Specificity	0.333	0.262	0.310	0.143	0.619
F1	0.805	0.766	0.791	0.839	0.750
Balanced Accuracy	0.563	0.502	0.542	0.515	0.640

3.0.3 Models built on Undersampled Data

After building models with undersampled data, more of the overweight classes were accounted for, when compared to the oversampled and imbalanced classes as can be seen in Table 3.

Table 3: Performance Metrics of Models with Undersampled data.

Metrics	SVM	Bagging	NN	RF	LR
Accuracy	0.647	0.568	0.597	0.617	0.632
Sensitivity	0.642	0.591	0.597	0.610	0.610
Specificity	0.667	0.643	0.595	0.643	0.714
F1	0.742	0.701	0.701	0.716	0.724
Balanced Accuracy	0.654	0.617	0.596	0.626	0.662

3.0.4 Models built on Hybrid-sampled Data

With the Hybrid dataset, the models exhibited accurate classification for both the normal and overweight classes, surpassing the accuracy achieved by models constructed with undersampled data. This outcome strikingly balanced the accuracy and accountability of the negative class, as it consistently outperformed the undersampled dataset in terms of accuracy across all models, while also demonstrating superior performance in classifying negative instances compared to the oversampled dataset.

Consequently, we recommend adopting the hybrid sampling technique as a more effective approach for balancing imbalanced datasets. Table 4 shows results of models built with the hybrid-sampled data.

Table 4: Performance Metrics of Models with Hybrid-sampled data.

Metrics	SVM	Bagging	NN	RF	LR	NB	LDA	KNN
Accuracy	0.657	0.657	0.622	0.672	0.602	0.433	0.597	0.478
Sensitivity	0.711	0.698	0.648	0.717	0.623	0.327	0.629	0.459
Specificity	0.452	0.500	0.524	0.500	0.524	0.833	0.476	0.548
F1	0.766	0.763	0.730	0.776	0.712	0.477	0.712	0.582
Balanced Accuracy	0.582	0.599	0.586	0.608	0.573	0.580	0.553	0.503

3.0.5 Ensemble Stacking

Considering the performance range of the individual models, which fell between 43% and 67%, we employed our proposed ensemble stacked model to enhance overall performance. In this approach, we stacked eight models, using two meta models. The obtained results indicated that the ensemble model with RF as the meta model achieved the most significant improvement in accuracy. Specifically, when employing the LR meta model, the accuracy reached 85.6%. In contrast, utilizing the Random Forest meta model yielded an accuracy of 94.65%.

4 Results and Discussion

Imbalanced datasets pose a common challenge in machine learning, especially in real-life applications. The presence of imbalanced data can result in high false negative rates and reduced accuracy when predicting minority classes. To mitigate this issue, it is crucial to employ balancing techniques. In this study, we evaluated and compared the effectiveness of various balancing techniques for an imbalanced dataset related to overweight classification.

We applied oversampling, undersampling, and hybrid sampling methods to the dataset and assessed their performance. From the results presented in Tables 1 to 4, it is evident that models trained on the undersampled dataset exhibited favorable performance in terms of specificity and balanced accuracy. However, it had the lowest overall accuracy among the three balanced datasets. On the other hand, the oversampled dataset achieved good accuracy but performed poorly in detecting the negative class, as indicated by lower values of specificity and balanced accuracy. In contrast, the hybrid dataset achieved a balance between accuracy and accountability of the negative class. It demonstrated higher accuracies than the undersampled dataset across all models while also outperforming the oversampled dataset in classifying negative instances. Based on these findings, we recommend adopting the hybrid sampling technique as a superior balancing approach for imbalanced datasets.

Furthermore, we explored the impact of ensemble learning on enhancing performance and accuracy compared to using individual models in isolation.

Considering that the data balanced using the hybrid sampling method exhibited superior performance to the oversampling and undersampling methods, we utilized it to train an ensemble model by stacking eight individual models. The stacked ensemble model achieved an accuracy of 85.63% when LR was employed as the meta model. Remarkably, when RF was used as the meta model, the accuracy considerably improved to 94.65%. The performances of these two stacked ensembles highlight the efficacy of ensemble stacking, as they outperformed each of the individual models. Additionally, the choice of meta model for stacking demonstrated an influence on the overall accuracy of the model.

5 Conclusion

In summary, our study demonstrates the effectiveness of hybrid sampling as a balancing technique for imbalanced datasets. Additionally, we have shown that the integration of multiple high-performing individual models through stacking can further enhance predictive performance. These findings have practical implications for the development of machine learning models in real-life applications, and they also provide a foundation for future research endeavors in this domain.

References

[1] O. O. Awe, N. Dukhi, and R. Dias. Shrinkage heteroscedastic discriminant algorithms for classifying multi-class high-dimensional data: Insights from a national health survey. *Machine Learning with Applications*, 12:100459, 2023.

[2] J. Brownlee. *Ensemble learning algorithms with Python: Make better predictions with bagging, boosting, and stacking*. Machine Learning Mastery, 2021.

[3] S. Buyrukoglu and S. Savaş. Stacked-based ensemble machine learning model for positioning footballer. *Arabian Journal for Science and Engineering*, 48(3):1371–1383, 2023.

[4] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

[5] A. M. Esfar-E-Alam, M. Hossain, M. Gomes, R. Islam, and R. Raihana. Multimodal emotion recognition using heterogeneous ensemble techniques. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 1033–1037, 2022.

[6] A. Fernández, S. García, and F. Herrera. Addressing the classification with imbalanced data: Open problems and new challenges on class distribution. In E. Corchado, M. Kurzyński, and M. Woźniak, editors, *Hybrid Artificial Intelligent Systems*, pages 1–10, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[7] M. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[8] J. Johnson and T. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27, 2019.

[9] B. Krawczyk, M. Wozniak, and G. Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. 5 2014.

[10] G. Kumar, K. Thakur, and M. R. Ayyagari. Mlesidss: machine learning-based ensembles for intrusion detection systems—a review. *The Journal of Supercomputing*, 76:8938–8971, 2020.

[11] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar. Classification of imbalanced data:review of methods and applications. *IOP Conference Series: Materials Science and Engineering*, 1099(1):012077, mar 2021.

[12] V. Lingampally and K. Radhika. A cascading ensemble with custom subset generation and multi-level fusion for enhanced breast cancer detection. *IJFMR-International Journal For Multidisciplinary Research*, 5(3).

[13] K. Napierala and J. Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46:563–597, 2016.

[14] S. Rajagopal, P. P. Kundapur, and K. S. Hareesha. A stacking ensemble for network intrusion detection using heterogeneous datasets. *Security and Communication Networks*, 2020:9, 2020.

[15] M. Tanveer, A. Rastogi, V. Paliwal, M. Ganaie, A. Malik, J. Del Ser, and C.-T. Lin. Ensemble deep learning in speech signal tasks: a review. *Neurocomputing*, page 126436, 2023.

[16] S.-J. Yen and Y.-S. Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006*, pages 731–740. Springer, 2006.

[17] X. Yin, Q. Liu, Y. Pan, X. Huang, J. Wu, and X. Wang. Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: Comparison of eight single and ensemble models. *Natural Resources Research*, 30:1795–1815, 2021.

[18] D. Zhang and J. J. Tsai. *Advances in machine learning applications in software engineering*. Igi Global, 2006.