88365_Auto_Edited-check.docx

WORD COUNT	2329	TIME SUBMITTED	24-OCT-2023 04:42PM
		PAPER ID	103691171

Name of Journal: *World Journal of Methodology*
Manuscript NO: 88365
Manuscript Type: EDITORIAL

Challenges and limitations of synthetic minority oversampling techniques in machine learning

Abstract

Oversampling is the most utilized approach to deal with class-imbalanced datasets, as seen by the plethora of oversampling methods developed in the last two decades. We argue in the following editorial the issues with oversampling that stem from the possibility of overfitting and the generation of synthetic cases that might not accurately represent the minority class. These limitations should be considered when using oversampling techniques. We also propose several alternate strategies for dealing with imbalanced data, as well as a future work perspective.

Key Words: Machine learning; Class imbalance; Overfitting; Misdiagnosis

Alkhawaldeh IM, Albalkhi I, Naswhan AJ. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World J Methodol* 2023; In press

Core Tip: Addressing class imbalance in medical datasets, particularly in the context of machine learning applications, requires a cautious approach. While oversampling methods like synthetic minority oversampling technique are commonly used, it is crucial to recognize their limitations. They may introduce synthetic instances that do not accurately represent the minority class, potentially leading to overfitting and unreliable results in real-world medical scenarios. Instead, consider exploring alternative approaches such as Ensemble Learning-Based Methods like XGBoost, Easy Ensemble which have shown promise in mitigating bias and providing more robust performance. Collaborating with data science specialists and medical professionals to design and validate these techniques is essential to ensure their reliability and effectiveness in medical applications.

INTRODUCTION

Imbalanced medical data may have a variety of issues that impede classification, such as the impact of noisy data and borderline samples, class overlapping, class imbalance,

or the presence of small disjuncts. When training a dataset, an imbalanced class distribution can occur when one class has significantly more samples than the other, resulting in a majority and minority class. This class imbalance can lead to prediction bias in machine learning models, which often translates to poor performance in the minority class(es). To address this issue, several techniques have been proposed in the literature. These techniques include increasing the number of samples from the minority class by obtaining more data from the source, modifying the loss function to assign a higher cost to misclassifications in the minority class, oversampling the minority class by replicating or generating synthetic samples, undersampling the majority class by reducing the number of instances, or using a combination of these approaches. By employing these techniques, the aim is to mitigate the class imbalance problem and improve the performance of machine learning models on imbalanced datasets[1]. There are benefits and drawbacks to each strategy. Many publications have used oversampling, particularly synthetic minority oversampling technique (SMOTE) approaches to create artificial samples from minority samples, to address the issue of class imbalance in the medical, biomedical and life sciences fields. This is evident from the abundance of oversampling methods developed in the last two decades. For instance, a PubMed search for the terms “oversampling” OR “smote” produced 2157 results from publications between 2000 and 2022, whereas a Web of Science search produced 2185 hits (but only for medical-related topics) (Figure 1). Additionally, when comparing this to undersampling using a PubMed search for (“undersampling” and “machine learning”) and (“oversampling” or “SMOTE” and “machine learning”) a noticeable difference is found (Figure 2). This essentially indicates the developing trend of oversampling research in the medical literature, which dealt with or simply discussed oversampling.

Although there has been a substantial increase, it is important to note that this does not automatically imply the effectiveness of the oversampling approach. The surge in oversampling research can be attributed to the significant prevalence of the class imbalance problem and the relative simplicity of oversampling solutions[2]. The concern

regarding oversampling methods arises from their potential to artificially increase the number of minority-class instances by generating new ones based solely on their similarity to existing minority examples. This raises concerns about the possibility of overfitting during the learning process. While oversampling techniques may yield favorable results in machine learning experiments, this does not necessarily translate to practical success. Additionally, a more significant issue with oversampling is that the synthetic examples created may actually belong to a different class in the real world, despite their similarity to the minority class examples. This is due to the fact that there are instances from class A that are closer to examples from a different class B, regardless of their similarity to the minority class examples[3].

Multiple experimental papers have provided evidence to support the concerns regarding oversampling methods. Elreedy *et al*[4] conducted a study where they analyzed the probability distribution of the synthetic samples generated by the SMOTE method. Their findings led them to conclude that the synthetic data produced by SMOTE may not precisely match the original distribution of the minority class, which can have an impact on the classification performance. Similarly, Tarawneh *et al*[2] argue against the current forms and methodologies of oversampling, considering it a deceptive approach. They suggest that oversampling introduces falsified instances that are falsely classified as members of the minority class when they are more likely to belong to the majority class. Their conclusions were drawn from a recommended validation system that was applied to various class-imbalanced datasets, including medical datasets. The validation system involved hiding a number of majority examples and assessing the similarity between the synthesized examples generated by different oversampling methods and the hidden majority examples[3].

After conducting a detailed analysis of their findings, it becomes evident that all validated oversampling approaches suffer from errors in the synthesized samples. These approaches generate instances that are intended to represent the minority class but actually resemble the majority class or fall within the decision boundary of the majority class (Figure 3). The error rate varies across different validated methods and

oversampled datasets, ranging from 0% to 100%. None of the strategies achieve zero error on all datasets, indicating their inability to accurately oversample medical records. The oversampling techniques attempt to fill the feature space gap by creating new instances that are similar to one or more minority instances. However, these techniques wrongly assume that the synthesized examples belong to the minority class without providing any guarantee. Consequently, the training of these instances becomes misguided, increasing the risk of overfitting the classifier on false data. This poses a significant threat as the entire machine learning system may fail when applied in real-world medical applications, where even a single incorrectly generated example can have severe consequences. Therefore, oversampling structured medical datasets by synthesizing new instances solely based on their resemblance to the minority examples is a questionable practice, particularly in the context of medical data. It is crucial to ensure that the additional samples truly fall within the minority class. Moreover, since the model itself is flawed, any external validation, subsequent analysis, or conclusions based on it should be critically examined. The potential harm and consequences of a misdiagnosis, inaccurate prediction, or prognosis can be particularly detrimental for cancer patients[1].

When it comes to the currently existing methods for dealing with class imbalance problems, researchers suggested various strategies for analyzing data, which are classified as data level, algorithm level, and hybrid (Figure 4). The methods are dependent on the size of the data collection, distribution, imbalance ratio, and model performance criteria[5].

Under-sampling approaches, like oversampling, have some drawbacks on the data level, such as the loss of critical information for data distribution. Under-sampling can result in the loss of relevant information by removing valuable and significant patterns. Oversampling and undersampling approaches can be used independently or in hybrid methods. Because they are based on existing techniques, these hybrid methods built on them share the same limitations[5]. Recent research suggested random partitioning of data with a voting rule, a resampling method that works by randomly splitting the

imbalanced dataset into a number of smaller balanced sub-datasets. On each sub-dataset, a machine-learning model is subsequently trained. The final prediction is made by applying a voting mechanism to the individual model forecasts. Other resampling strategies were outperformed by this strategy. When tested using several machine learning classifiers on 33 benchmark class-imbalanced datasets. This approach has the potential to overcome the present limitations[6].

Hybrid techniques combine methods on multiple levels. For example, when data-level methods are used to process data externally and distribute classes to instances. The learning process is then carried out internally using algorithm-level methods. You can read for a thorough explanation of these techniques[7]. In algorithm-level methods, researchers have the ability to modify conventional machine learning models by assigning weights or costs to classifiers in order to mitigate bias towards the majority class. This approach ensures that the learning model remains unaffected by the class distribution. These methods can be categorized as recognition-based, cost-sensitive, or ensemble learning-based techniques.

In the absence of non-target class instances, recognition-based approaches such as One-class learning are employed. They model the classifier on the representation of the minority class and proceed to learn primarily from minority class instances rather than attempting to distinguish dissimilar patterns from majority class and minority class examples. One-class classification includes features such as outlier identification and novelty discovery. This approach performs well, particularly with high-dimensional data. One class learning may be used to build many models, including support vector machines and isolation forests, however it cannot be used to build other models, such as decision trees and Nave Bayes[5].

Cost-sensitive methods are crucial in medical applications because of the significance of minimizing false positive and false negative instances. These methods involve adjusting the misclassification cost to achieve a balance between the majority and minority classes. For instance, assigning a higher weight or cost to false negative predictions compared to false positive predictions can be an effective approach. This

practical solution enables cost-sensitive learning in the context of medical applications. In the literature, there are several cost-based approaches available to address class imbalance in data, including weighted cross-entropy, multiclass dice loss function, and focal loss. There are other recent methods such as weighted extreme learning machine, cost-sensitive decision tree ensemble methods, and cost-sensitive deep neural networks[5,8].

Ensemble learning methods have gained significant attention in various fields, including machine learning and medical applications. These methods combine multiple weak learners to create a more robust model with improved performance. Some popular ensemble learning techniques include voting and boosting Ensemble learning models, such as XGBoost and Easy Ensemble, have been found to outperform individual learning models and exhibit greater resistance to noise and outliers[3,8]. However, it is important to consider the potential drawbacks of ensemble learning. These models often require significant training time and may be prone to overfitting in certain scenarios. To address the limitations of ensemble learning, researchers have introduced new techniques. Ensemble pruning methods aim to reduce the complexity and training time of ensemble models while maintaining their performance. Regularization techniques, such as dropout and bagging, have also been incorporated into ensemble models to mitigate the risk of overfitting[8].

Feature selection approaches use are growing in addressing data imbalance. These techniques reduce computational and storage costs, eliminate redundant information, and facilitate data visualization. Feature selection methods can be grouped into filter, wrapper, and embedded methods. Filter methods select variables using statistical measures, while wrapper methods assess features based on model performance and selection criteria. Embedded methods, like least absolute shrinkage and selection operator regression, perform feature selection as an integral part of the learning process[5]. Additionally, it is important to consider diverse performance metrics, such as the area under the precision-recall curve, Matthew's correlation coefficient, F-score, and geometric mean, to effectively evaluate model performance in the presence of class

imbalance. These metrics provide a comprehensive assessment of model performance, taking into account precision, recall, and the balance between sensitivity and specificity[5,9,10].

To effectively address class imbalance, it is advisable to consider various approaches at different levels, including data-level, algorithm-level, and hybrid-level techniques. These approaches aim to mitigate bias and improve classifier algorithms by combining different methods. There is a growing body of research focusing on hybridization techniques that integrate sampling, feature selection, and classifier building to gain a better understanding of class representation and achieve more accurate classification results. For instance, evolutionary computing can be employed in feature selection, while ensembles can be constructed to tackle the challenges associated with class imbalance. By adopting these practical and reliable solutions, the issue of class imbalance in medical datasets can be significantly improved, as these approaches are based on sound assumptions[5,9].

To gain insights into the impact of oversampling and other methods on real-world medical applications, it is crucial to collaborate with data science specialists and medical professionals. By working together, these experts can create and evaluate these approaches, taking into account their unique perspectives and expertise. This collaboration can provide valuable guidance and influence the selection or adaptation of appropriate techniques. The joint efforts of medical doctors and data scientists can lead to the development of more reliable and efficient solutions in the field of healthcare.

CONCLUSION

It is recommended to address class imbalance at different levels considering data-level, algorithm-level, and hybrid-level approaches that can mitigate bias. At the algorithmic level Ensemble Learning-Based Methods such as XGBoost and Easy Ensemble, have proved to have a better performance than individual learning models, and they provide more resistance to noise/outliers. Another hybrid method is Random Under-Sampling

Boost which is not without limitations either, noting that there isn't a one-size-fits-all approach and exercising caution should be taken when addressing class imbalances. The adoption of such more practical and trustworthy solutions would improve the class imbalance issue in medical datasets more because these approaches have no wrong assumptions[2-6]. It is vital to understand how oversampling and other methods may affect real-world medical applications. Collaborating with data science specialists and medical professionals can enhance the development and testing of reliable and effective solutions, as their insights can provide valuable advice and influence the selection or modification of relevant techniques.

88365_Auto_Edited-check.docx

ORIGINALITY REPORT

5%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	67 words — 3%
2	Ahmad Hassanat, Ghada Altarawneh, Ibraheem M. Alkhawaldeh, Yasmeen Jamal Alabdallat et al. "The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets", 2023 IEEE Symposium on Computers and Communications (ISCC), 2023 Crossref	52 words — 2%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 15 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 15 WORDS