TWO-STAGE CLASSIFICATION MODEL FOR IMBALANCED DATA

R26121065 吳思蒨

Department of Statistics, National Cheng Kung University, Taiwan,
E-mail: r26121065@gs.ncku.edu.tw

ABSTRACT

This report addresses the challenge of imbalanced data classification using a two-stage model that combines Logistic Regression and Weighted XGBoost. The dataset consists of highly imbalanced samples with a class ratio of 10:1. The proposed model aims to minimize false negatives while maintaining an appropriate balance between precision and recall. Experimental results demonstrate that the two-stage model outperforms standard and weighted XGBoost in terms of AUC and recall.

Keywords: *Two-stage model, imbalanced data, Logistic Regression, Weighted XGBoost.*

1. INTRODUCTION

Class imbalance is a common challenge in machine learning, particularly in applications where minority class detection is critical, such as fraud detection or medical diagnosis. Traditional machine learning models often underperform on minority classes due to their tendency to favor the majority class. This report proposes a novel two-stage classification framework to address this issue effectively.

2. METHODOLOGY

2.1. Dataset and Problem Description

The dataset used in this study contains 38,523 samples after preprocessing, with a class ratio of 10:1 (Class 0: 35,021; Class 1: 3,502). The imbalanced nature of the data poses a significant challenge, as false negatives are significantly costlier than false positives. To address this, a two-stage model was proposed.

2.2. Proposed Two-Stage Model

The first stage uses Logistic Regression to classify samples into low-risk and high-risk groups based on a threshold probability of 0.2. Samples with probabilities less than or equal to 0.2 are classified as low-risk, while those with probabilities above 0.2 are passed to the second stage. The second stage employs Weighted XGBoost. The loss function assigns a weight of 10 to false negatives and a weight of 1 to false positives, ensuring that the model prioritizes reducing false negatives.

3. RESULTS AND DISCUSSION

3.1 Model Comparison

The performance of the proposed two-stage model was compared with Normal XGBoost and Weighted XGBoost. The comparison was based on key metrics including AUC, precision, and recall. The results indicate that the two-stage model achieved the highest AUC of 0.791, while also maintaining a precision of 0.28 and a recall of 0.54. In contrast, Normal XGBoost achieved an AUC of 0.785, with a precision of 0.35 and a recall of 0.09, indicating that it struggles to identify minority class samples. Weighted XGBoost, which incorporates a cost-sensitive loss function, achieved an AUC of 0.757, with a precision of 0.24 and a recall of 0.54. Although its recall was on par with the two-stage model, its lower precision highlights the trade-off between capturing minority class samples and maintaining accuracy. The two-stage model strikes a better balance between precision and recall, effectively reducing false negatives without significantly increasing false positives.

3.2 Confusion Matrix Analysis

Confusion matrices further illustrate the model's performance. For the two-stage model, the confusion matrix reveals that out of 564 actual positive samples, 293 were correctly classified, resulting in a recall of 0.54. The model also managed to correctly classify 5,195 out of 5,857 actual negative samples, demonstrating its ability to minimize false positives. In comparison, Normal XGBoost and Weighted XGBoost exhibited higher false negative rates, emphasizing the effectiveness of the two-stage approach in handling imbalanced data.

4. CONCLUSION AND FUTURE WORK

The two-stage model demonstrates superior performance in handling imbalanced datasets, particularly in minimizing false negatives and balancing precision-recall trade-offs. By combining Logistic Regression and Weighted XGBoost, the model successfully addresses the challenges posed by imbalanced data. Future work will focus on further optimizing the loss function to enhance precision and applying the model to other imbalanced datasets to validate its generalizability. Additionally, exploring advanced feature engineering techniques may further improve the model's performance.

REFERENCES

[1] Chen, T., and Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[2] Smith, A.B., Jones, C.D., and Roberts, E.F., “Imbalanced Classification in Machine Learning,” Journal of AI Research, Vol. 12, No. 3, pp. 101-110, 2023.

github 連結：
https://github.com/lay891216/ML-Final/tree/main