IMPROVING RESULTS IN MULTI-CLASS CLASSIFICATION FOR IMBALANCED DATA BY ASSIGNING WEIGHTS TO CLASSES BASED ON MACHINE LEARNING MODELS

Sariyev Shokhrukh
Miqiyeva Gulchekhra
Samarkand State University named after Sharof Rashidov
Samarkand, Uzbekistan
sariyevshokhrukh@gmail.com

https://doi.org/10.5281/zenodo.16885125

Abstract. This study analyzed strategies for assigning weights to classes in the problem of imbalanced classification and elucidated the theoretical and practical aspects of applying these strategies in ensemble and neural network models. The imbalance in the proportion of classes for rare positive cases of medical diagnostics, the low result of the accuracy indicator, as a result, leads to a decrease in the results of the assessment criteria of the minority class recall in the models and is eliminated by assigning weight to the classes of the models to prevent the occurrence of an incorrect diagnosis. This study describes the weighted gradient-Hessian approach for gradient boosting family algorithms such as LightGBM and HistGradientBoosting, as well as weighted cross-entropy and, when necessary, threshold-moving and calibration methods for MLP. In mathematical form, the updating of leaf values using the Newton step, the calculation of cumulative gradients at the histogram bin level, and cross-entropy weighting for MLP based on output logits are presented. Additionally, boosting and weighted derivatives are included.

Keywords. LightGBM, HistGB, StochasticGB, MLP.

INTRODUCTION

Currently, machine learning technologies have begun to be widely applied in various fields, and their practical significance is increasing year by year. In classification problems, load data serves as a common tool in solving many real-life problems for various sins[1-2]. However, in practical settings, the dataset is often unbalanced, meaning that some classes have fewer observations than others. Such disproportionate distribution is especially widespread in areas such as healthcare, finance, security, and industrial quality control. This study consists of an unbalanced data set, as the number of patients with a medical diagnosis is much lower than that of healthy patients. Moreover, fraud in financial transactions accounts for a very small percentage of the total number of transactions.[3-4] The main problem when working with unbalanced datasets is that typical classification algorithms strive to maximize the overall accuracy and, as a result, favor the most common class. Furthermore, a rare class drastically reduces the precision and recall indicators for the class. As a result, the models tend to be poor at detecting cases that are practically important but statistically rare, and lead to erroneous results. There are a number of approaches to solving this problem, among which the method of assigning weights to classes is one of the simplest and most effective strategies. The reason is that the essence of assigning weights to classes is that less frequent class errors are given more weight in the loss function[5-6]. As a result, their influence leads to an equal amount of selection for each class of students. As a result, the model optimization process increases sensitivity to the rare class and improves the overall classification quality. In this study, weighting classes in unbalanced data is considered one of the methods that is not only statistically sound, but also has high value from the point of view of practical application.

WEIGHTING MACHINE LEARNING CLASSIFICATION ALGORITHMS

Machine learning models are used to classify classes on data sets that are unbalanced by assigning weights. The goal of weighting is not always to have equal or close classes. For example, in a medical dataset, the number of positive diagnoses is always less than the number of negative diagnoses, which leads to imbalance in the dataset[7-8]. To prevent this, the dataset is divided into equal or close classes using various methods. These methods include reducing multi-class data sets, artificially equalizing the number of lower classes to the number of higher classes, and assigning weights to classes. Among these methods, weighting of class classes in machine learning models is cited in this study.

MATHEMATICAL MODELS OF WEIGHTING MACHINE LEARNING MODELS.

LightGBM. The goal is to minimize the second derivative of the loss function. The Newton step for j for each leaf is calculated by the following formula (1).

$$G_j = \sum_{i\in leaf\ j} g_i,\ H_j = \sum_{i\in leaf\ j} h_i,\ g_i = w_i(p_i - y_i),\ h_i = w_i p_i(1 - y_i) \tag{1}$$

The optimal solution for the bar value is calculated using the following formula (2).

$$\gamma_j^* = -\frac{G_j}{H_j + \lambda} \tag{2}$$

HistGradientBoosting. The loss function is minimized by the histogram the G, H values collected over bins, and is calculated by the following formula (3).

$$G_b = \sum_{i\in b} w_i(p_{i,k} - y_{i,k}), \quad H_b = w_i p_{i,k}(1 - p_{i,k}) \tag{3}$$

MPL. Weighting in neural networks is trained with cross-entropy.

$$L(\theta) = -\sum_{i=1}^{n} w_i \sum_{k=1}^{K} y_{i,k} \log p_k(x_i;\theta), \quad p_k = soft\max_k(F(x_i;\theta)) \tag{4}$$

The weighting of each class is done as follows.

$\alpha_k = \dfrac{N}{Kn_k}$ for each class, for each sample $w_i = \alpha_{y_i}$.

ADVANTAGES AND LIMITATIONS

One of the advantages is that the dataset does not change, oversampling and undersampling are not required to balance the classes, and the true distribution is preserved. The models are fast and lightweight because only the loss function or split-gain formulas are weighted[9-11]. Cost-sensitive adaptation reduces FN by penalizing errors that are important to a class, such as disease, more. As a result, accuracy and recall improve.

Limitations. Excessively large weight selection will disrupt stability and may reduce overall accuracy. Decision threshold dependency: weights increase recall, but if the threshold is not adjusted, precision will drop sharply[12-15].

CONCLUSION

One of the least expensive and most commonly used strategies in unbalanced classification is to assign weights to classes, which focuses on loss, gradient, and split-gain optimization without changing the data structure[16-18]. It boosts its signals for rare classes[19-21]. However, this approach is not a solution, but rather the overall accuracy and stability may deteriorate if the weighting decision threshold and probability calibration are incorrectly chosen. For this reason, it is recommended that reweighting is always used in conjunction with the correct metrics macro-F1, accuracy, and recall, stratified validation, and threshold-tuning[22-24].

References:
Используемая литература:
Foydalanilgan adabiyotlar:

1. N. Fayzullo, S. Sariyev and Y. Sherzodjon, "Analyzing the Effectiveness of Ensemble Methods in Solving Multi-Class Classification Problems," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 788-793, doi: 10.1109/SmartIndustryCon65166.2025.10986248.

2. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.

3. Friedman, J. H. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis, 38(4), 367–378.

4. Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NeurIPS.

5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD.

6. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR, 12, 2825–2830.

7. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. ICCV.

8. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. CVPR.

9. Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. NeurIPS.

10. Mekhriddin Nurmamatov1, Shokhrukh Sariyev1. (2025). Intelligent data analysis and hyperparameter tuning using genetic algorithms in machine learning [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16325952

11. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 16, 321–357.

12. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE TSMC A, 40(1), 185–197.

13. Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. PKDD.

14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. ICML.

15. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. ICML.

16. Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. ICML.

17. M. Nurmamatov, S. Sariyev and B. Eshonkulov, "Application of Evolutionary Algorithms to Enhance the Efficiency of Neural Networks and Machine Learning Algorithms," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 533-537, doi: 10.1109/SmartIndustryCon65166.2025.10986257.

18. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3): e0118432.

19. Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochimica et Biophysica Acta, 405, 442–451. (MCC metrikasi)

20. Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. IJCAI.

21. M. Nurmamatov, S. Sariyev and I. Uddin, "Methods of Using Artificial Intelligence Algorithms in Human Resource Management," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 566-571, doi: 10.1109/SmartIndustryCon65166.2025.10986087

22. A. Axatov, M. Nurmamatov, F. Nazarov, and Sh. Sariyev, “Genetic algorithm application technology in multi-parameter optimization problems,” AIP Conf. Proc., vol. 3244, art. no. 030025, 2024, doi: 10.1063/5.0242074

23. Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. AAAI (arXiv:1908.07442).

24. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. Neural Networks, 106, 249–259.