Team 2
Multilabel classification using Gradient boost

Israel Prado, *2009107* Andre Dorantes, *2009045* Josue Gomez, *2009061* Jhair Cach, *2009012*
Jesus Casas, *2009024* Esteban Rodriguez, *2009116* and Angel Huerta, *2009071*

***Abstract*—Multi-label classification is a challenging task in machine learning, allowing the assignment of multiple labels to a single instance. This approach enables nuanced predictions, differing from traditional multi-class classification. Gradient Boosting, an ensemble learning method, sequentially constructs weak learners, rectifying errors iteratively and improving accuracy. XGBoost, a popular library, enhances gradient boosting with regularization terms, enhancing robustness.**

This research delves into fundamental aspects of gradient boosting, a significant ensemble learning technique. The study comprehensively explores its principles, strengths, and applications. It discusses related libraries like LightGBM and CatBoost, emphasizing their efficiency and handling of categorical features.

Addressing class imbalance in multi-label datasets, the research explores techniques such as class weighting and resampling.The research emphasizes the practicality and versatility of gradient boosting techniques, contributing to advancements in complex classification problems.

***Index Terms*—Multi-label Classification, Gradient Boosting, XGBoost, LightGBM, CatBoost, Ensemble Learning, Class Imbalance, Classification Metrics, Overfitting, Computational Efficiency, Machine Learning, Data Science, Weak Learners, Regularization, Categorical Features**

I. Introduction

Multilabel classification is a versatile machine-learning task where multiple labels can be assigned to a single instance, allowing for more nuanced and complex predictions. This is in contrast to traditional multi-class classification, which assigns a single label to an instance.

One powerful approach to tackling multi-label classification is Gradient Boosting, an ensemble learning method that sequentially constructs weak learners, typically in the form of decision trees. This iterative process rectifies errors from previous learners, resulting in improved accuracy. XGBoost, a popular gradient boosting library, excels in multi-label classification by incorporating regularization terms, enhancing both robustness and accuracy.

In this research endeavor, our primary focus is to conduct an in-depth examination of the fundamental characteristics and key attributes associated with gradient boosting. Gradient boosting stands as a significant and widely used ensemble learning technique within the realm of machine learning and data science. Our objective is to comprehensively delve into its various aspects, elucidating its underlying principles, strengths, and applications.

II. Gradient boost

Multilabel classification refers to the task of assigning multiple labels to a single instance. Unlike traditional multiclass classification, where an instance belongs to just one class, multilabel classification allows for the assignment of multiple classes simultaneously.

Gradient Boosting, an ensemble learning method, constructs a series of weak learners (typically decision trees) sequentially. It rectifies errors made by previous learners in subsequent iterations, continuing until a predetermined number of weak learners is reached or no further improvement can be achieved.

A. Generalization of Gradient Boosting

In traditional gradient boosting used for single-label classification, the loss functions are typically decomposable, meaning the loss can be computed and minimized for each label independently. However, this is not enough for multilabel classification problems where the relationship between labels can be important. In multilabel classification, non-decomposable loss functions, such as the subset 0/1 loss, take these label dependencies into account, but they pose a challenge for optimization.

The generalization of gradient boosting to multi-output problems involves extending the framework to work with both types of loss functions: decomposable and non-decomposable. This is crucial because it allows for the optimization of loss functions that better capture the complexity of multilabel tasks. For instance, while the Hamming loss is label-wise decomposable and relatively straightforward to minimize, the subset 0/1 loss is not and requires a more sophisticated approach [3].

B. Algorithm Customization

In multilabel classification, the choice of loss function is integral to the performance of the model. Different loss functions evaluate predictions in various ways, and optimizing for one may lead to deterioration with respect to another. Therefore, a learning algorithm should ideally be customizable towards a specific choice of performance measure. Gradient boosting algorithms, especially when generalized for multilabel classification, are appealing because they offer this flexibility. They can be tailored to optimize a particular loss function chosen based on the specific characteristics and requirements of the dataset and task at hand [3].

The customization towards specific performance measures is facilitated by the development of algorithms such as BOOMER, which allows for the optimization of a wide range of loss functions, both decomposable and non-decomposable. This aligns the training process with the ultimate goal of the classification task, whether it's to minimize the number of

incorrect label assignments (as with the Hamming loss) or to ensure that the set of predicted labels matches the true set as closely as possible (as with the subset 0/1 loss) [4].

C. XGBoost

A widely-used gradient boosting library, excels in multilabel classification. It enhances the traditional gradient boosting algorithm by incorporating regularization terms into the objective function, enhancing its robustness and accuracy, also optimizing the the memory usage and computation speed supporting in addition to handle massive datasets. Theon speed XGboost is highly applied because of its benefits that includes a short evaluation time required, having a faster model in comparison to some neural networks models, and it has also the possibility to still have missing values without requiring compensation for them, so the model can continue working with no problems of that cause. XGBoost is also helpful when preventing the Overfitting of a model, monitoring the model's performance on the validation dataset, detecting then if its performance stops improving or start degrading, therefore the training can be stopped early before it gets over-fitted.

In multilabel classification problems, the XGBoost helps with the imbalanced class distributions, allowing the assignment of different wights for the present categories to handle that data, considering in a fair way all classes in place of giving increasing favor to a majority class.

D. LightGBM

Developed by Microsoft, is another gradient-boosting framework designed for efficiently handling vast datasets and supporting multilabel classification. Its use of a histogram-based learning method ensures both speed and accuracy.

E. CatBoost

An open-source gradient boosting library by Yandex, stands out for its efficient handling of categorical features. It is well-suited for multilabel classification tasks and automatically manages categorical variables, eliminating the need for manual preprocessing.

Multilabel datasets often face class imbalance, where certain labels are more prevalent than others. To counter this, techniques such as class weighting and resampling can be applied to mitigate the issue.

Selecting appropriate evaluation metrics is crucial in multilabel classification. Common metrics include Hamming Loss, F1 Score, and Subset Accuracy. The choice of metric depends on the specific problem and the relative importance of false positives and negatives.

Incorporating gradient boosting techniques like XGBoost, LightGBM, and CatBoost provides a robust approach for tackling intricate classification problems. By comprehending the intricacies of multilabel data and employing suitable algorithms and evaluation metrics, researchers and practitioners can develop accurate and reliable multilabel classification models.

III. Conclusion

In conclusion, a comprehensive overview of gradient boosting has been provided in this research, revealing an ensemble learning technique that has been proven to be invaluable in the domain of machine learning and data science. Throughout the study, its fundamental characteristics, underlying principles, and applications have been thoroughly examined, shedding light on its role in constructing a series of weak learners, typically in the form of decision trees, sequentially. The iterative error rectification process, which aims to enhance predictive accuracy and model performance, has been explored in detail. Additionally, prominent gradient boosting implementations, such as XGBoost, LightGBM, and CatBoost, have been scrutinized, highlighting their unique features and advantages. This research has aimed to serve as a valuable resource for both novice and experienced practitioners, facilitating a deeper understanding of gradient boosting and its relevance in solving complex problems within the field.

References

[1] A. W. Wong, W. Sun, S. V. Kalmady, P. Kaul and A. Hindle, *Multilabel 12-Lead Electrocardiogram Classification Using Gradient Boosting Tree Ensemble*, 2020 Computing in Cardiology, Rimini, Italy, 2020, pp. 1-4, doi: 10.22489/CinC.2020.128.

[2] M. R. Choirulfikri, Adiwijaya and A. A. Suryani, *Comparison of Bagging and Boosting in Imbalanced Multilabel of Al-Quran Dataset*, 2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), Bandung, Indonesia, 2022, pp. 01-05, doi: 10.1109/ICADEIS56544.2022.10037462.

[3] M. Rapp, *BOOMER — An algorithm for learning gradient boosted multi-label classification rules*, Software Impacts, vol. 10, pp. 100137, 2021, doi: https://doi.org/10.1016/j.simpa.2021.100137. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963821000567

[4] M. Rapp, E. Loza Mencía, J. Fürnkranz, V.-L. Nguyen, and E. Hüllermeier, *Learning Gradient Boosted Multi-label Classification Rules*, arXiv preprint arXiv:2006.13346, 2020. [Online]. Available: https://arxiv.org/abs/2006.13346