A Comparative Study of Classical and Modern Boosting Algorithms for Supervised Learning

Tony Luo, Will Kim, Sichen Li, Shang Peng, Jiaheng Guo

Hi team, I am using the template provided by Prof. Zhong. We can for now ignore the references in .bib file and focus on this single .tex file for synchronization.

1 Introduction

Briefly describe the problem you are addressing, why it matters, and what gap your work aims to fill.

2 Related Works

Review of existing studies in related fields.

3 Methodology

Please be aware that the paper structure Prof. Zhong provided in this template (Section 2 Related Works, Section 3 Methodology) is different from what we proposed in the group chat. Please move the mathematics explanations to Section 3, and reserve Section 2 for literature review, which covers the historical development of boosting algorithms, their applications, and comparisons in previous studies, with minimal mathematical details.

Describe the models, techniques, or algorithms you developed or compared.

4 Experiments Results

This section describes the experimental setup used to compare AdaBoost, GBM, XGBoost, and LightGBM. All models were evaluated under a unified and reproducible framework with consistent preprocessing, hyperparameter tuning, and cross-validation procedures.

4.1 Datasets and Preprocessing

We evaluate the four boosting algorithms on several publicly available datasets from *Kaggle* and *UCI Machine Learning Repository* spanning both classification and regression tasks across various domains. Table 1 and 2 summary the characteristics of the datasets we used.

dataset	#samples	#features	#classes	type	link
Adult Income	48,842	14	2	Finance	Kaggle
Heart Disease	303	13	2	Medical	UCI
Mushrooms	8,124	22	2	Medical	Kaggle
Telco Customer Churn	7,043	33	2	Commercial	Kaggle
Breast Cancer	569	30	2	Medical	Kaggle
Credit Card Fraud	284,807	30	2	Finance	Kaggle
IMDB Movie Reviews	50,000	1	2	NLP	Kaggle

Table 1: Characteristics of the Classification Datasets Used

dataset	#samples	#features	type	link
California Housing	20 640	8	housing	Kaggle
Ames House Prices	1 460	80	housing	Kaggle
Wine Quality	4 898	11	chemistry	UCI
Boston Housing	506	13	housing	UCI

Table 2: Characteristics of the Regression Datasets Used

All datasets were processed through a shared pipeline:

(1) Missing numerical features were imputed using medians; categorical features were imputed using the most frequent category.

(2) Categorical variables were one-hot encoded for AdaBoost, GBM, and XGBoost, while LightGBM additionally took advantage of its built-in categorical support.

(3) Numerical features were left unscaled, consistent with tree-based model characteristics.

(4) Training and testing splits were generated using fixed random seeds to ensure reproducibility across models.

4.2   Training and Model Selection

All datasets follows a 80/20 train-test split. For hyperparameter optimization, we performed hyperparameter optimization for each algorithm using grid search and randomized search to tune maximum depth, number of estimators, and learning rate. For classification problems, logistic or exponential losses were used as appropriate, and for regression tasks, squared-error loss was applied. Nested k-fold cross-validation with $k = 10$ was used to avoid information leakage and produce robust performance estimates.

4.3   Evaluation Metrics

Classification Metrics.   Accuracy, precision, recall, F1-score, and ROC–AUC were used.

Regression Metrics.   RMSE, MAE, and R^2 were used.

4.4   Comparative Analysis

The empirical analysis examines:

- **Predictive performance:** measurable improvements over classical methods.

- **Model complexity and overfitting:** effects of depth, shrinkage, and regularization.

- **Computational efficiency:** impact of parallelization, histogram-based training, and sampling strategies.

- **Feature importance stability:** consistency of learned feature rankings across folds and algorithms.

- **Scalability:** relative performance on larger datasets, highlighting differences in system design.

These analyses provide a comprehensive comparison of classical and modern boosting techniques, highlighting both algorithmic trade-offs and practical considerations.

5   Conclusion and Future Direction

Summarize key findings, limitations, and next steps. Propose meaningful future improvements or extensions that follow logically from your results.
[1]

References

[1] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.