Qeios

Peer Review

Review of: "TabularGRPO: Modern Mixture-Of-Experts Transformer with Group Relative Policy Optimization GRPO for Tabular Data Learning"

Abdullahi Ibrahim[1]

1. Max-Planck-Institut für Intelligente Systeme, Tübingen, Tübingen, Germany

TabularGRPO: Modern Mixture-of-Experts Transformer with Group Relative Policy Optimization GRPO for Tabular Data Learning

The paper introduces TabularGRPO, a method that integrates:

- A Mixture-of-Experts (MoE) Transformer architecture designed to handle heterogeneous tabular data (continuous and categorical features).
- A Group Relative Policy Optimization (GRPO) procedure, presented as an extension or variant of Proximal Policy Optimization (PPO) to reduce variance in policy gradient updates, especially for imbalanced classification scenarios.

The paper reports performance gains over standard gradient-boosted decision tree models (such as XGBoost, CatBoost) on:

1. A synthetic dataset of only 150 samples.
2. The Census Income Dataset (Adult data).

The authors claim the model offers:

- Improved handling of data type heterogeneity (through gating networks/expert routing).
- More stable training using a “group advantage” calculation.
- Impressive gains on small or imbalanced datasets.

doi.org/10.32388/OZBEJX

While I found these interesting, some aspects of the paper need to be addressed.

Major Comments

1. The paper frames the work within RL, using policy-gradient notation (advantages, rewards, policy updates). For tabular classification, I believe it is not entirely clear how “actions,” “states,” and “rewards” are being concretely defined. While the text refers to the use of “contrastive advantage” and “group-wise advantage,” more detail is needed on how exactly a supervised label (class) is recast into a reward signal, and how policy objectives align with standard classification objectives.

2. The text suggests that the model’s “reward” is tied to classification correctness and “group advantage.” However, one would expect more explicit definitions: for instance, how is the advantage function computed from classification metrics, what is the reward function per sample, and how does that tie to the policy’s gradients?

3. The authors call their method “Group Relative Policy Optimization,” referencing PPO-like clipping, KL constraints, and group-wise normalization. However, the references on GRPO are either self-citations or not standard references in the RL literature. It would help to clarify precisely how GRPO differs from PPO—and show theoretical or empirical justifications for each modification. For instance, is it purely that the advantage is normalized per group? Are the “groups” simply batches or class bins? That crucial novelty needs clear exposition.

4. While MoE can be powerful for partitioning features or subtasks among specialized “experts,” it remains unclear if the gating function truly learns distinct “expert specializations.” The authors claim that some experts handle categorical features while others handle continuous features, yet the methods to ensure or encourage that specialized usage are only mentioned in passing. If the gating is purely learned without constraints, it would be nice to clarify how or why it converges to partition by feature type.

5. The paper emphasizes superior performance on a synthetic benchmark small dataset with 150 samples. Although small-sample success is an important scenario, near-perfect results (F1 = 1.0, AUC = 1.0) typically raise suspicions regarding overfitting or data leakage—especially for a flexible model like a transformer with multiple experts. I would advise the authors to:
    - provide clear details on how train/test splits are done (e.g., repeated cross-validation rather than a single 80–20 split)

- report standard deviations across multiple runs, given that small data can lead to high variance results

1. The authors compare against XGBoost and CatBoost. It is standard practice to tune these baseline models carefully (e.g., hyperparameter tuning for learning rate, tree depth, etc.). The paper does not specify how thoroughly the baselines were optimized. I encourage the authors to describe how you tuned XGBoost and CatBoost so readers can be confident that those baselines are well-optimized. The authors should consult this paper for example:

   Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature ***637****, 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-6*

2. For the RL-based approach, critical hyperparameters (clip parameter, group size, KL coefficients, etc.) are given, but there is no mention of how these were arrived at. A short table or subsection detailing hyperparameter tuning or rationale would be nice.
3. The paper states that TabularGRPO is suitable for highly imbalanced data, yet it only provides numeric results for the standard Census Income dataset (about 75% negative to 25% positive—moderately imbalanced). A more systematically imbalanced dataset (e.g., extreme ratio 1:10 or 1:20) could demonstrate TabularGRPO's advantage more forcefully. I strongly recommend additional experiments on highly skewed data to bolster claims of "group-based advantage" techniques.
4. The introduction highlights that "critical applications like credit scoring demand calibrated confidence estimates" and mentions that GBDTs "lack native uncertainty quantification." However, there is no explicit demonstration that TabularGRPO provides better-calibrated probabilities or confidence intervals. If this is one of the major motivations, I strongly suggest that the authors

should present

- a reliability diagram or calibration plots comparing TabularGRPO, XGBoost, and CatBoost
  - negative log-likelihood or Brier scores for measuring calibration.

5. Beyond XGBoost and CatBoost, there are several deep-learning-for-tabular-data baselines that I believe are relevant, e.g., FT-Transformer, TabNet, TabPFN, NODE, or TabTransformer. Even if only a subset are tested, referencing or briefly mentioning them would strengthen the significance of the approach. For example, TabPFN performs well on small-medium size data of < 10, 000.

6. The authors mention in "4.6.2 ablation study" where the gating network was removed, or the entire mixture-of-experts setup was removed, but the text only provides general statements like "we saw a 3-5% drop." I believe more detailed tables (with standard deviations) from multiple seeds/runs would be clearer and more scientific.

7. The text also references "LatentEncoder" or "LatentVoiceTransformer" as if borrowed from another domain (possibly from the authors' prior "VoiceGRPO" work). If the architecture is partially adapted from voice or speech tasks, that cross-domain adaptation should be described more explicitly.

8. The paper cites a "13.0% higher F1 score" on the Census dataset, then a "10% more precise and 10% higher F1 score" in the conclusion, so the exact numeric improvement is not entirely consistent. The final table states that TabularGRPO yields 0.8455 precision vs. 0.7694 or 0.7848 for baselines—so that's about 6–7 percentage points, not 10. This discrepancy should be clarified or the final text should stay consistent with the presented data table.

9. The results for the Census Income dataset appear to come from one "best training epoch" on a single train/test split, at least to my understanding. Since tabular benchmarks are often subject to random splits, a standard practice is to report average performance over multiple (i.e., 10) random

splits (and to give standard deviation). This would protect the results from a lucky or unlucky partition and avoid claims that might rely on a single random seed.

10. There are multiple instances of abrupt, almost bullet-point-style writing. Some sentences lack articles or conjunctions, which can impede comprehension (e.g., “tabular data learninig precisly,” “lack gradient-based fine-tuning,” “still challenging common task”). The authors should rewrite the article and use more sentences instead of bullet points.
11. References for PPO appear to be duplicated in the References list (both #2 and #3 cite Schulman et al. 2017). Please delete one.
12. Some references have minimal detail (e.g., “Tabular Synthetic Benchmark Small Dataset (150)” [11]). It would help to give details on how that dataset was generated, or at least a direct link to a repository if it is custom.
13. The authors provided a GitHub link to their code, but there is no confirmation in the text about how to access pretrained weights or replicate results. Detailed instructions (e.g., environment, scripts, a requirements file) should be included if reproducibility is claimed.
14. Figure 1 (the schematic of the proposed method) is helpful but might be too high-level. Readers would benefit from more detail on the gating mechanism or how each stage (MoE, advantage calculation, final classifier) interacts. Also, why is the dashed line passing through the gating network? Any meaning?
15. Let me give an additional suggestion. I encourage the authors to provide a stand-alone subsection thoroughly explaining how the classification setup is reinterpreted in an RL framework. Define the reward function, advantage estimation, and “group” concept with examples and formulas. Demonstrate that GRPO is not simply cross-entropy with some reweighting, but genuinely leverages RL-based policy gradient ideas.

Minor comments

1. Please revise the manuscript for grammar, clarity, and consistent style throughout.

2. Lastly, the authors should present the results in a more scientifically rigorous manner. For instance, they could include plots (e.g., ROC curves) to illustrate the AUC more effectively.

3. On page 2, the authors mention ‘KL divergence’ without defining ‘KL’ first. They should refer to it as ‘Kullback–Leibler divergence’ or define the acronym before using it.

4. On page 2, the authors claim that ‘over 60% of production...’ datasets are imbalanced. If no reference is provided, they could revise it to ‘some production datasets’ to avoid unsupported statistics.
5. The authors define ‘Mixture of Experts (MoE)’ multiple times. They should define it once and then consistently refer to it as ‘MoE’ thereafter.
6. On page 5, the final sentence needs appropriate punctuation (a comma or a period) to ensure grammatical completeness.
7. When defining metrics such as precision, F1 score, and ROC AUC, the authors should cite the original publications that introduced or popularized these measures.

Declarations

Potential competing interests: No potential competing interests to declare.