Qeios

Peer Review

Review of: "SampleLLM: Optimizing Tabular Data Synthesis in Recommendations"

Markus Bayer[1]

1. Technische Universität Darmstadt, Germany

The study “SampleLLM: Optimizing Tabular Data Synthesis in Recommendations" proposes a novel method for tabular data synthesis with an emphasis on recommendations. SampleLLM consists of using an LLM for data generation and clustering, as well as a feature attribution-based importance sampling method for distribution alignment.

I think the study is very well written! The problem is nicely motivated. Furthermore, the method is very reasonable, and the experiments show its effectiveness. However, as I am not an expert on tabular data, I do not know if it is sufficient to test the method with a simple DNN (I leave this to other reviewers). Other than that, I did not find any flaws in the study, and the following are mostly minor comments.

Strengths:

- Novel and beneficial method that is based on the distribution problem of generated data from LLMs
- Many experiments (datasets and baselines)
- A/B testing
- Well written and easy to follow

Weaknesses:

- Only DNN as a model for the experiments
- No limitations or outlook section

doi.org/10.32388/I1OLIL

1. Introduction

- I really like that you are motivating the problem of LLM generation with Figure 1! Maybe you could also reference some works that are providing more details on this topic.

2.2. SampleLLM Overview

- Figure 2 gives a very good overview of SampleLLM. I like that you also include examples! Still, I find that the figure might be improved by making it less cluttered (without removing the examples). Is there a reason for the gray and white arrows?
- “By combining the instruction and exemplars as inputs, the LLM generates synthetic samples using a few-shot learning strategy.” - what is meant here by few-shot learning strategy?

2.3. Designed Instruction

- “Furthermore, when it comes to textual features, these models typically use ID encodings as input, hindering their ability to capture the semantic associations between features. To address these limitations, we leverage the general knowledge and semantic understanding capabilities of LLMs for data synthesis via few-shot learning.” - How does your method address finding semantic associations between features for the downstream models?
- “Specifically, for each dataset, 1% synthetic samples are generated and incorporated into the training set. The instruction that delivers the best performance is then established as the standard for that dataset." - how did you ensure that the differences were not just random factors?
- I think the part of iteratively refining the instruction is a bit opaque. Do you prompt the model to refine the instruction?
- I was wondering how different the results are for different instructions or if it might also be possible to design a general instruction (working for every dataset)

2.5. Feature Attribution-based Importance Sampling

- This procedure makes a lot of sense, and you have described it very well. However, I was wondering (as with the clustering method) what amount of original data is required for your method to work.

3. Experiments

- I like that your experiments are guided by research questions
- In general, a very good analysis of the results of all experiments!

3.1.2. Baselines

- Minor: Formatting of REaLTabFormer
- While the choice of baselines seems reasonable, I am not an expert in this area and do not know if there are other approaches that need to be evaluated.

3.1.3. Implementation Details

- Minor: “module in SampleLLM. A simple” → “module in SampleLLM a simple”
- Again, I am not an expert in this field, but is it really enough to evaluate a simple DNN? I think it is important to use the most common method in the field in 2024 or 2025. If there is a better (more common model), then it is important to see if SampleLLM works for it.

3.2.1. Augmentation Utility

- Table 2: For ML-1M and Amazon, isn’t the Original the second-best?
  - For Table 3, I see why Original is not highlighted

3.2.1. MLE Utility

- I think it would be interesting to see how 100% synthetic data performs. However, I can see that this could be too time-consuming (could be discussed in the limitations).

4. Online A/B Test

- Interesting experiment!

5. Related Work

- The related works section lists the relevant works (note that I am not an expert on tabular data). However, I think you could consider moving the section further up, as there are many works relevant to the experiments, and it might be easier for the reader to follow.

Limitations Section

- Missing. I think it is very important to add a section on limitations, e.g., describing that you have not tested different tabular models and very small datasets.

Declarations

Potential competing interests: No potential competing interests to declare.