Qeios

Peer Review

Review of: "SafeSynthDP: Leveraging Large Language Models for Privacy-Preserving Synthetic Data Generation Using Differential Privacy"

Bhavani Malisetty[1]

1. Computer Science, University of Nebraska, Omaha, United States

1. Introduction and Research Significance

The paper SafeSynthDP presents a novel Large Language Model (LLM)-driven framework that integrates Differential Privacy (DP) mechanisms to generate synthetic datasets while preserving privacy guarantees. This research is particularly relevant given the growing need for privacy-preserving machine learning (ML) applications in domains such as healthcare, finance, and social media.

With the emergence of strict privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), organizations face significant challenges in utilizing real-world datasets while ensuring compliance with data privacy laws. Traditional data anonymization techniques have proven insufficient against modern privacy attacks, such as membership inference and model inversion attacks, highlighting the need for differentially private synthetic data generation.

The study effectively addresses this issue by demonstrating how DP noise injection mechanisms (Laplace and Gaussian distributions) can be applied at the data generation stage to produce synthetic data that maintains statistical similarity to real data while safeguarding sensitive information. The paper also provides a privacy-utility trade-off analysis, evaluating how different privacy budgets (ε values) impact model performance when trained on synthetic vs. real datasets.

2. Strengths and Contributions of the Research

doi.org/10.32388/MSINQ7

The study makes several significant contributions to the field of privacy-preserving machine learning and synthetic data generation:

2.1. Integration of Differential Privacy in LLM-Based Synthetic Data Generation

The authors propose a training-free synthetic data generation approach, eliminating the need for fine-tuning (which itself poses privacy risks). Instead, LLMs use in-context learning (ICL) prompts to generate privacy-preserving synthetic datasets. By injecting DP noise at the data generation stage, the study ensures that synthetic datasets retain statistical utility while mitigating privacy risks.

Real-World Application Example:

- **Healthcare**: SafeSynthDP can be used to generate synthetic patient records that preserve key medical trends for AI-driven disease prediction models while ensuring HIPAA compliance.

2.2. Empirical Evaluation of Privacy-Utility Trade-Off

The paper rigorously evaluates the impact of DP noise (Laplace, Gaussian) on synthetic data utility across different privacy budgets (ε values), quantifying data degradation under stricter privacy constraints.

Real-World Application Example:

- **Finance**: Banks can generate synthetic transaction datasets for fraud detection models, ensuring compliance with CCPA and GDPR while maintaining transaction pattern integrity.

2.3. Generalization Across ML Architectures

The study evaluates SafeSynthDP on multiple machine learning architectures, demonstrating its broad applicability:

Traditional ML models: Multinomial Naïve Bayes (MNB), Support Vector Machines (SVM)

Deep learning models: Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM)

LLM-based In-Context Learning (ICL): GPT-4o-mini, Gemini-1.5-Flash

Real-World Application Example:

- Social Media & NLP Applications: Platforms such as Twitter and YouTube could use SafeSynthDP to generate synthetic social media conversations for AI content moderation models, ensuring real user interactions remain private.

2.4. Mitigation of Membership Inference Attacks (MIA)

A major concern in synthetic data generation is the potential exposure of individual records through membership inference attacks. The study evaluates SafeSynthDP's robustness against MIA, showing that differential privacy mechanisms significantly reduce the risk of re-identification.

Real-World Application Example:

- **Government Agencies**: Statistical organizations can use SafeSynthDP to generate privacy-preserving synthetic census data for public research and policymaking without exposing individual records.

3. Methodological Rigor and Experimental Evaluation

The study follows a rigorous experimental framework, incorporating:

3.1. Dataset Selection

- The AGNews dataset was used for a controlled multi-class text classification task, allowing for direct comparison between models trained on real vs. synthetic data.

3.2. Comparative Model Performance

Findings show:

- Minimal degradation (3-10%) in traditional ML models (MNB, SVM) → indicating strong statistical feature retention.
- Greater accuracy reduction (15-20%) in deep learning models (GRU, LSTM) → suggesting synthetic data struggles with complex temporal dependencies.
- LLM-based In-Context Learning (ICL) performed well, with higher context examples (2-shot, 4-shot learning) improving performance.

4. Recommended Additional Evaluation Metrics

While the study primarily evaluates classification accuracy, additional privacy and statistical similarity metrics would strengthen the analysis:

Statistical Similarity Metrics

- Wasserstein Distance, KL Divergence, Jensen-Shannon Divergence → measure how closely synthetic data mimics real data distributions.

Privacy Leakage Metrics

- Membership Inference Attack (MIA) Success Rate → measures how well privacy is preserved.
- Attribute Inference Attack Success Rate → evaluates risk of reconstructing private data from synthetic data.

Fidelity vs. Privacy Trade-Off Analysis

- Precision, Recall, F1-score → evaluates how much useful information is retained in synthetic data.
- Feature Correlation Coefficients (Pearson, Spearman, Kendall) → assess how well feature relationships are preserved.

Downstream ML Performance & Generalization

- Cross-Dataset Transferability Test → train on synthetic data, test on real data.
- Perplexity (for text-based data) → evaluates synthetic text coherence.

5. Limitations and Future Research Directions

Despite its strengths, the study identifies several areas for future research:

Semantic Preservation in Deep Learning Models

- LSTM and GRU models experience greater performance degradation, suggesting the need for better context retention in synthetic data.

Exploring Alternative Noise Mechanisms

- Laplace and Gaussian noise are used, but adaptive DP mechanisms could optimize privacy-utility balance dynamically.

Extending to Multi-Modal Data

- SafeSynthDP currently focuses on text → future work could extend it to images, videos, and structured datasets.

Robustness Against Advanced Privacy Attacks

- Future Work: Evaluating SafeSynthDP against GAN-based attacks, model inversion, and adversarial perturbations.

6. Conclusion: Impact of Study's Findings on Real-World Scenarios

The study’s findings have far-reaching applications across multiple industries:

Healthcare → Synthetic patient records for privacy-compliant AI in disease prediction. Finance → Synthetic banking transactions for fraud detection while ensuring GDPR compliance.

Social Media → Privacy-preserving AI training for chatbot and content moderation systems.

Government & Census → Synthetic demographic datasets for policy research.

Declarations

Potential competing interests: No potential competing interests to declare.