A Differential Privacy Noise Allocation Method Based on Attribute Sensitivity

Abstract

As AI technology advances and data utilization expands, the risk of privacy breaches caused by personal information exposure has intensified. Conventional data protection methods allocate noise based on arbitrarily defined attribute sensitivities, resulting in inconsistent protection levels and degraded data utility. To address this, the present study proposes a sensitivity-based differential noise allocation strategy incorporating attribute information diversity and inter-attribute correlations.

1 Introduction

Recent advances in artificial intelligence (AI) have led to an expansion in the use of large-scale data, increasing the risk of privacy violations due to personal information embedded in the data [1]. To address this, previous studies have proposed differential privacy techniques that vary the intensity of noise injection by attribute. However, their reliance on attribute sensitivity criteria based on subjective human judgment leads to inconsistent privacy protection levels and reduced data usability [2]. Therefore, in this paper, we calculate an attribute sensitivity score that considers both the information diversity of each attribute and the correlation between attributes, and then inject differential noise based on this to achieve a balance between model accuracy and privacy preservation.

2 Proposed Model

The structure of the proposed technique is illustrated in Figure 1. A higher entropy value indicates greater information diversity for a given attribute, while a stronger correlation with other attributes implies a higher exposure risk. Accordingly, the information entropy and inter-attribute correlations of the original dataset are computed and integrated to quantify the potential privacy risk of each attribute. A higher sensitivity score is assigned to attributes with larger calculated values, and differential noise is applied based on these scores. The proposed method preserves privacy while maintaining data utility by injecting stronger noise into highly sensitive attributes.

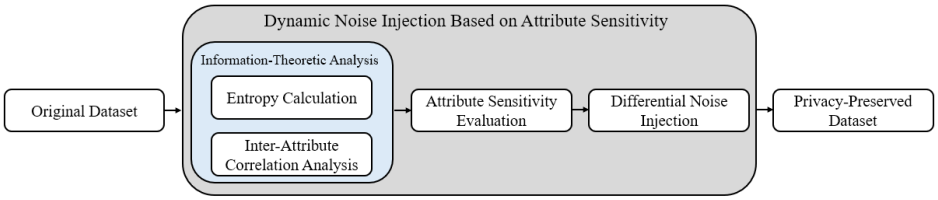

Figure 1: Attribute Sensitivity-Based Privacy Preservation Framework

Using the Adult dataset, we compared differential privacy strategies that adjust attribute-specific noise intensity based on XAI-based importance rankings, attribute sensitivity rankings, and subjective sensitivity rankings. The income classification accuracy was evaluated using a logistic regression model, and the defense rate was measured through attribute inference attacks. Experimental results indicate that the proposed method achieves approximately 2.4% lower accuracy than the XAI-based importance ranking approach, but improves attack resistance by up to 8.7%, thereby mitigating the utility–privacy trade-off. Furthermore, compared with the subjective sensitivity ranking method, the proposed approach maintains comparable accuracy while achieving up to 4.9% higher resistance to attribute inference attacks, demonstrating its overall effectiveness.

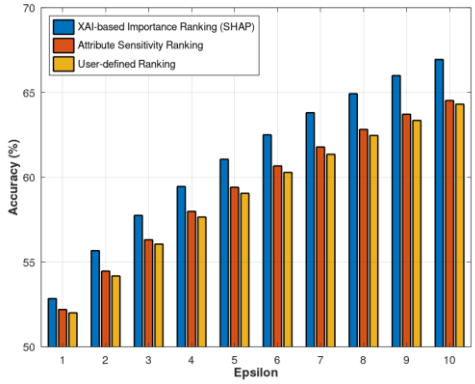

Figure 2: Accuracy by Model

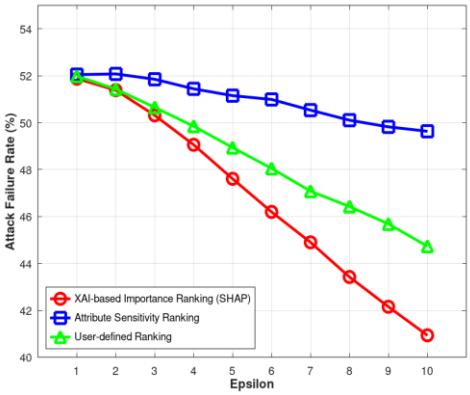

Figure 3: Attack Failure Rate by Model

3 Conclusion

This paper proposes a sensitivity-based noise allocation strategy that considers both attribute information diversity and inter-attribute correlations to enhance the utility-privacy tradeoff. Experimental results demonstrate that the proposed technique outperforms XAI-based importance ranking and subjective sensitivity ranking methods. Future research will explore differential privacy techniques that maintain model prediction accuracy similar to the original level even when injected noise.

References

1. Zheng, Lele, et al. "Sensitivity-Aware Differential Privacy for Federated Medical Imaging." *Sensors* 25.9 (2025): 2847.
2. Shi, Weiyan, et al. "Just fine-tune twice: Selective differential privacy for large language models." *arXiv preprint arXiv:2204.07667* (2022)

A Differential Privacy Noise Allocation Method Based on Attribute Sensitivity

Yu-Na Kim, Yeon-Jin Kim, and Il-Gu Lee : Sungshin Women's University

MobiSec 2025

1. Introduction

With recent advancements in artificial intelligence technology, the demand for large-scale data has rapidly increased. However, this data inherently contains sensitive personal information, increasing the risk of privacy violations during data utilization. In particular, the problem of inferring sensitive personal information by leveraging correlations between publicly available attributes has become a major concern. To address this issue, previous studies have proposed differential privacy techniques that vary the intensity of noise injection for each attribute, thereby preserving privacy. However, existing approaches rely on subjective human judgment in setting attribute-specific sensitivity assessment criteria, which can lead to inconsistent privacy protection levels or unnecessarily reduced data utility in real-world settings. In this paper, we calculate attribute sensitivity scores that simultaneously reflect both privacy and model contribution by considering the information diversity of each attribute and its correlations with other attributes. Based on this approach, we achieve a balance between model accuracy and privacy preservation by differentially injecting noise into the data.

- We propose a data protection technique that simultaneously considers privacy preservation and model performance based on attribute sensitivity evaluation criteria that reflect information diversity and correlation between attributes.
- We propose an evaluation framework that can compare and analyze the trade-off between privacy preservation level and model performance.

2. Proposed technique

Figure 1 illustrates the process of building a privacy-preserving dataset based on dynamic noise injection using attribute sensitivity scores. A higher entropy value indicates greater information diversity for an attribute, and the stronger the correlation with other attributes, the higher the exposure risk of that attribute. Accordingly, for each attribute in the original data, information entropy and correlation between attributes are calculated, and the two metrics are integrated to quantify the potential privacy risk for each attribute. A higher sensitivity score is assigned to a higher risk, and noise is injected differentially for each attribute based on this score. By injecting relatively strong noise into highly sensitive attributes, utility is maintained while ensuring privacy preservation.

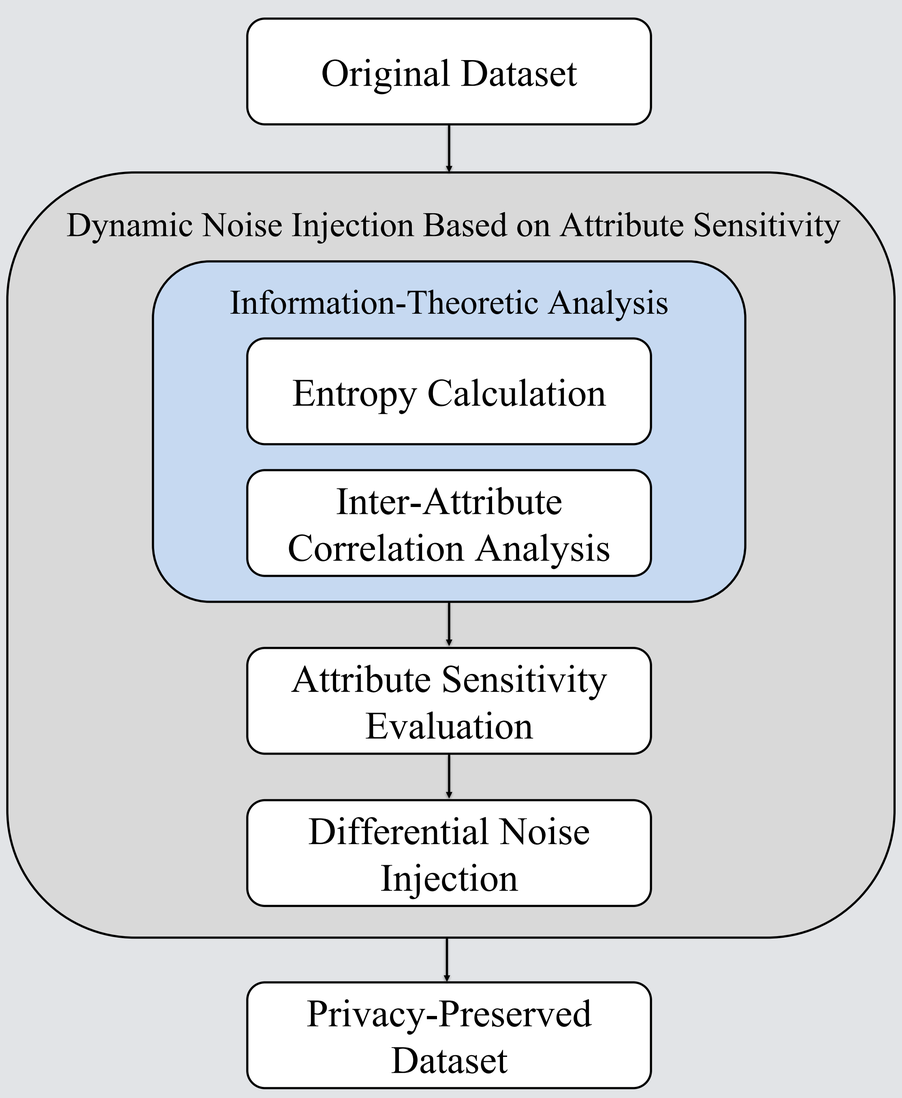

Fig.1. Structure of the original-synthetic optimal ratio data set.

3. Performance Evaluation

Using the Adult dataset, we compared differential privacy strategies that adjust attribute-specific noise intensity based on XAI-based importance rankings, attribute sensitivity rankings, and subjective sensitivity rankings. We evaluated income classification accuracy using a logistic regression model predicting income level and measured attack defense rates by conducting attacks to infer the sensitive attribute "marital status."

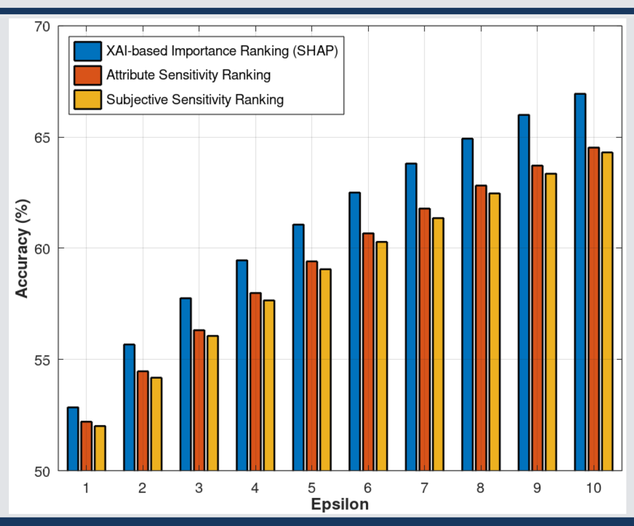

Fig. 2. Accuracy by model.

As shown in Figure 2, the proposed technique exhibited an accuracy approximately 2.4% lower than the XAI-based importance ranking method. This is because the XAI-based method ranks attributes based on their contribution to model performance and injects less noise as the ranking increases. Therefore, attributes that are important for the model's prediction accuracy are injected with relatively less noise, resulting in higher accuracy than the proposed technique.

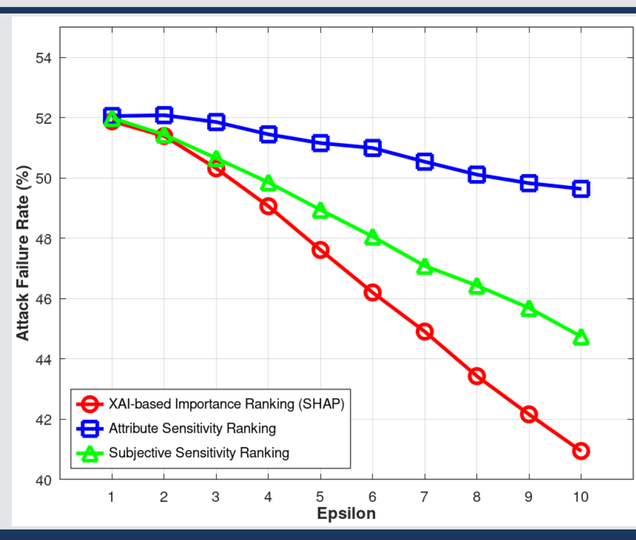

Fig. 3. Accuracy differences depending on the original-augmented data ratio

.Comparing the two experimental results, the proposed technique showed approximately 2.4% lower accuracy than the XAI-based importance ranking method, but up to 8.7% higher attack defense rate, mitigating the utility-privacy trade-off. Furthermore, compared to the subjective sensitivity ranking method, the proposed technique demonstrated up to 4.9% higher attack defense rate while maintaining similar accuracy, demonstrating the effectiveness of the proposed technique.

4. Conclusion

In this paper, we propose a sensitivity-based noise allocation strategy that considers both attribute information diversity and inter-attribute correlations. Experimental results demonstrate that the proposed technique outperforms XAI-based importance ranking and subjective sensitivity ranking methods. Future research will explore differential privacy techniques that maintain model prediction accuracy similar to the original level even when noise is injected.

Acknowledgments. This paper was conducted as a research result of the Industrial Innova-tion Talent Growth Support Project (RS-2024-00415520) of the Ministry of Trade, Industry and Energy and the Korea Institute for Advancement of Technology in 2025 and the ICT Innovation Talent 4.0 Project of the Ministry of Science and ICT and the National IT Indus-try Promotion Agency (No. IITP-2022-RS-2022-00156310).