Contrastive Learning-Based privacy metrics in Tabular Synthetic Datasets

Milton Nicolàs Plasencia Palacios[1,2]

[1] University of Trieste, Trieste, Italy
[2] AINDO, Trieste, Italy,
plasencia.milton@gmail.com

Abstract. Synthetic data has garnered attention as a Privacy Enhancing Technology in sectors such as healthcare and finance. When using synthetic data in practical applications, it is important to provide protection guarantees. We introduce a contrastive method that improves privacy assessment of synthetic datasets by embedding the data in a more representative space. This overcomes obstacles surrounding the multitude of data types and attributes. It also makes the use of intuitive distance metrics possible for similarity measurements and as an attack vector. Our results show that relatively efficient, easy to implement privacy metrics can perform equally well as more advanced metrics explicitly modeling conditions for privacy referred to by the GDPR.

Keywords: Contrastive Learning · Tabular Data · Privacy Metrics.

1 Introduction

Synthetic data is widely used as a Privacy Enhancing Technology to address privacy concerns in fields like healthcare and finance. However, assessing privacy risks in synthetic datasets is challenging due to the complexity of tabular data. The General Data Protection Regulation (GDPR) identifies Singling Out, that is, identifying a specific individual within a dataset, as one of the key privacy risks. To tackle this, inspired by [2], we propose a contrastive learning-based method that embeds synthetic data into a lower-dimensional and more representative space. This approach improves the detection and assessment of singling-out risks, addressing limitations in existing privacy evaluation techniques.

2 Method and Results

The proposed contrastive learning method uses a random masking function to process each record in a dataset D, creating two masked versions of the same record. The goal is to train a neural network f that maps rows of D to an embedding space, ensuring that embeddings of different masked versions of the same record are similar, while embeddings of masked versions of different records are dissimilar, as shown in Figure 1. Categorical variables are pre-processed through

an embedding layer before entering the main network while numeric variables are scaled using MinMax scaler. The implemented contrastive learning network all comprise three hidden layers, each consisting of 1024 neurons with dropout and GELU activation. The network uses cosine similarity in the normalized embedding space and cross-entropy loss for training. To perform the singling-out attack, we identify outliers in the embeddings and use the corresponding original rows to create queries for individual identification.

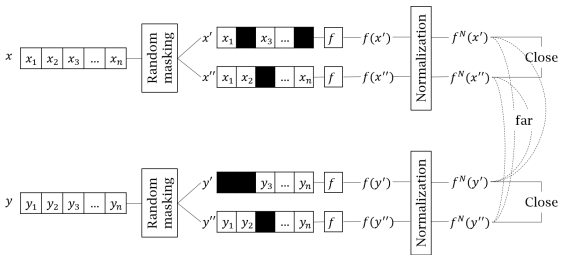

Fig. 1. Contrastive Learning Approach.

The method has been tested on three different datasets and compared to the current SOTA method [1] using REalTabFormer [3] as the synthetic data generator.

Dataset	SO [1]	SO + CL
Adult	0.02783 ± 0.031	0.01984 ± 0.028
Texas	0.03103 ± 0.026	0.04340 ± 0.029
Census	0.02695 ± 0.024	0.04650 ± 0.027

Table 1. Measured risks for the two approaches. SO: Singling out attack from [1]; SO + CL: Contrastive Learning-based Singling Out.

Acknowledgments. This research is the result of a joint collaboration between the University of Trieste (UniTS) and AINDO.

References

1. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data (2022)
2. Shenkar, T., Wolf, L.: Anomaly detection for tabular data with internal contrastive learning. In: International Conference on Learning Representations (2021)
3. Solatorio, A.V., Dupriez, O.: Realtabformer: Generating realistic relational and tabular data using transformers. arXiv preprint arXiv:2302.02041 (2023)