Evaluation of Synthetic Data for Privacy-Preserving Machine Learning

by Markus Hittmeir, Andreas Ekelhart and Rudolf Mayer (SBA Research)

The generation of synthetic data is widely considered to be an effective way of ensuring privacy and reducing the risk of disclosing sensitive information in micro-data. We analysed these risks and the utility of synthetic data for machine learning tasks. Our results demonstrate the suitability of this approach for privacy-preserving data publishing.

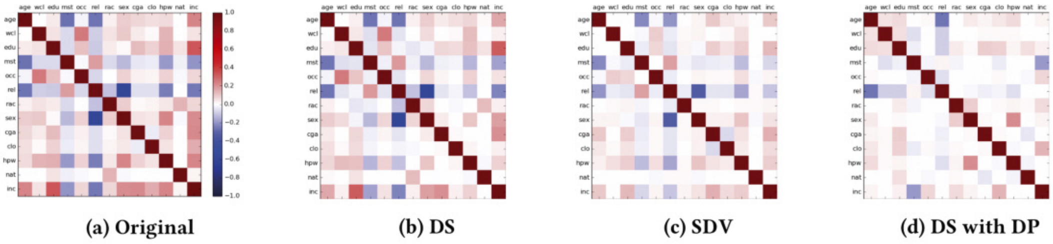
Recent technological advances have led to an increase in the collection and storage of large amounts of data. Micro-data, i.e. data that contains information at the level of individual respondents, is collected in domains such as healthcare, employment and social media. Its release and distribution, however, bears the risk of compromising the confidentiality of sensitive information and the privacy of affected individuals. To comply with ethical and legal standards, such as the EU's General Directive on Data Protection (GDPR), data holders and data providers have to take measures to prevent attackers from acquiring sensitive information from the released data.

Traditional approaches to compliance often include anonymisation of data before publishing or processing, such as using k-anonymity or differential privacy. Synthetic data offers an alternative solution. The process of generating synthetic data, i.e. data synthetisation, generally comprises the following steps:

- Data description: The original data is used to build a model comprising information about the distribution of attributes and correlations between them.
- Data generation: This model is then used to generate data samples. The global properties of the resulting synthetic dataset are similar to the original, but the samples do not represent real individuals.

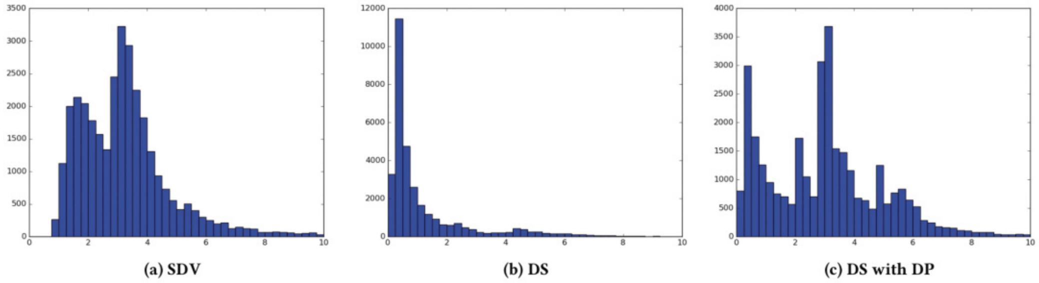
The goal of this technique is that analysis methods trained on the synthetic instead of the real data do not perform (notably) worse. The use of synthetic data should also reduce the risk of disclosure of sensitive information, as the artificially generated records do not relate to individuals in the original data in a one-to-one correspondence. Consequently, validating the utility and privacy aspects is crucial for trust in this method. We conducted an empirical evaluation, including three open-source solutions: the SyntheticDataVault (SDV) [L1], DataSynthesizer (DS) [L2] and synthpop (SP) [L3]. The SyntheticDataVault builds a model based on estimates of the distributions of each column. Correlations between attributes are learned from the covariance matrix of the original data. The model of the DataSynthesizer is based on a Bayesian network and uses the framework of differential pri-

Figure 1: Heatmaps for SyntheticDataVault and DataSynthesizer on the Adult Census Income dataset [L4].



(a) Original (b) DS (c) SDV (d) DS with DP

Figure 2: Distance Plots for SyntheticDataVault and DataSynthesizer on the Adult Census Income dataset.



(a) SDV (b) DS (c) DS with DP

vacy. Finally, synthpop uses a classification and regression tree (CART) in its standard settings.

The utility of the generated synthetic data can be assessed by evaluating the effectiveness of machine learning tasks. Models that are trained on the synthetic data can be compared with models trained on the original data, and scored on criteria such as accuracy and F-score for classification problems. We studied classification [1] and regression [2] tasks on publicly available benchmark datasets. While the results vary depending on the number of attributes, the size of the dataset and the task itself, we can identify several trends. In general, models based on synthetic data can reach utility up to or very close to the original data. Models trained on data from the DataSynthesizer without Differential Privacy or on data from synthpop with standard settings tend to achieve utility scores that are close to those of the model trained on the original data.

On the other hand, the SyntheticDataVault seems to produce data with larger differences to the original, which usually leads to reduced effectiveness. The same is true for the DataSynthesizer when Differential Privacy is enabled. These trends also manifest in direct comparisons of the datasets' properties, e.g., in the heatmaps of pairwise correlations shown in Figure 1.

A basic assumption is that privacy is endangered if the artificial rows in synthetic data are very close or equal to the rows of actual individuals in the original data. Privacy risks could therefore by assessed by computing the distance between each synthetic sample and the most similar original record. Visualisations of these minimal distances can be seen in Figure 2 (the x-axis shows the distance, the y-axis counts the number of records). While the DataSynthesizer without Differential Privacy leads to many records with small distances to original samples, the SyntheticDataVault generates much larger differences.

We complemented this privacy analysis on synthetic data by establishing a baseline for attribute disclosure risks [3]. Attribute disclosure happens when an attacker knows the values of quasi-identifying attributes of their victim (such as birth date, gender or ZIP), and is able to use some data source to infer the value of sensitive attributes (such as personal health data). By considering several scenarios on benchmark datasets, we demonstrated how an attacker might use synthetic datasets for the prediction of sensitive attributes. The attacker's predictive accuracy was usually better for the DataSynthesizer without Differential Privacy and for synthpop than it was for the SyntheticDataVault. However, both the amount of near-matches in the analysis of Figure 2 and the computed attribute disclosure scores show that the risk of reidentification on synthetic data is reduced.

Our evaluations demonstrate that the utility of synthetic data may be kept at a high level and that this approach is appropriate for privacy-preserving data publishing. However, it is important to note that there is a trade-off between the level of the utility and the privacy these tools achieve. If privacy is the main concern, we recommend that samples are generated based on models that preserve fewer correlations. This reduces the attribute disclosure risk and ensures that the artificial records are not too similar to the originals.

Links:

[L1] https://github.com/sdv-dev/SDV
[L2] https://github.com/DataResponsibly/DataSynthesizer
[L3] https://cran.r-project.org/web/packages/synthpop/
[L4] http://archive.ics.uci.edu/ml/datasets/Adult

References:

[1] On the utility of synthetic data: An empirical evaluation on machine learning tasks, ARES '19 Proc., Canterbury, UK, https://doi.org/10.1145/3339252.3339281

[2] Utility and privacy assessments of synthetic data for regression tasks, IEEE BigData '19 Proc., Los Angeles, CA, USA, https://doi.org/10.1109/BigData47090.2019.9005476

[3] A baseline for attribute disclosure risk in synthetic data, CODASPY '20 Proc., New Orleans, LA, USA, https://doi.org/10.1145/3374664.3375722

Please contact:

Markus Hittmeir, Andreas Ekelhart, Rudolf Mayer
SBA Research, Austria
mhittmeir@sba-research.org, aekelhart@sba-research.org, rmayer@sba-research.org