Open Source Code for Quality Evaluation of Synthetic Data

Abstract

Learning from personal data can have a great societal impact in many fields, but high sensitivity of these data often makes it difficult to learn from them. Synthetic data can alleviate this problem, by offering a way to keep the informative value of the sensitive data without revealing any personal information. But how can we reliably assess the quality of synthetic data? In this blog we introduce a tool to evaluate synthetic data based on four different criteria.

Data is being stored everywhere at a rapid pace, and there is often significant societal value that can be derived from this data. For instance, patient data on the effectiveness of various diabetes treatments can provide insights into which treatments work best. However, many of these datasets contain sensitive information, and therefore, they cannot be used or shared casually. By using synthetic data, value can be extracted from this data without compromising the privacy of individuals involved. This allows us to learn which diabetes treatments are effective without various organizations accessing your personal information. Due to the proliferation of options for generating synthetic data, evaluating the quality of these datasets can be challenging. However, this is crucial information for using them. Therefore, TNO has developed a method to directly compare different synthetic datasets in a single plot. The metrics we use for this purpose have been carefully selected, and we would like to share them with you. We have now open-sourced the code so that everyone can use it to assess their own synthetic data.

Synthetic Data

But first, what exactly is synthetic data? Synthetic data is artificially generated data based on real data. In the example of diabetes treatments, fake patients are generated who could have existed in the dataset. The trends present in the real data are retained, and this can be achieved through statistical approaches or Generative Artificial Intelligence (AI). Generative AI involves training a model to generate new fake individuals. This synthetic data can be used for testing, developing, and validating models. It can also generate data for situations that do not currently exist, allowing models to make better predictions. Similarly, it can create more balance in representing minority groups, ensuring that every diabetes patient has an equal chance of receiving an effective treatment, regardless of age, gender, or ethnicity.

Quality of Synthetic Data

The goal is for synthetic data to be used instead of real data, so it is important that synthetic data closely resembles the real data. To check the quality of these synthetic datasets, we test them on four different dimensions. Depending on the application of synthetic data, one dimension may be more important than another. First, we examine whether the distributions within a variable are preserved. We have named this dimension Univariate (1. Univariate). For example, is there an equal percentage of men and women as in the original dataset? Next, we check whether the correlations between pairs of variables remain the same (2. Bivariate). If real patients with overweight have higher blood sugar levels more often, is that also reflected in the synthetic data? Third, we train machine learning models to predict each variable using all other variables for training. The models are tested on the original data, and the accuracy of this machine learning task must remain consistent whether they are trained on synthetic or real data (3. Multivariate). If a prediction model has to recommend the best treatment for

the patient, that advice should remain the same. Finally, we train a machine learning classification model to distinguish the generated individuals from synthetic datasets from those of the original dataset (4. Distinguishability). There should be no distinction between real and fictitious patients.

Using the four calculated metrics, we create a spider plot, where multiple synthetic datasets can be directly compared. Figure 1 shows an example of a spider plot, with each dimension receiving a score between 0 (no similarity) and 1 (no distinction between synthetic and real data). The larger the area covered, the higher the quality. The spider plot can be used to compare the effect of different generation techniques and privacy levels on data quality. It provides a result in one overview, after which it may be desirable to evaluate individual metrics. We recommend defining certain thresholds in advance that the synthetic dataset must meet, such as "Machine learning predictions must not degrade by more than 10%."

Privacy and Synthetic Data

Note: using synthetic data is not a way to circumvent the GDPR. Its purpose is to, among other things, comply with the right to data minimization. The privacy of individuals in the data must always be protected, as intended by the legislation.

Measuring privacy assurance is a complex issue. There are empirical methods that simulate various attack scenarios. An important condition from the GDPR framework is to prevent individuals from being linked to the synthetic dataset (Risk of Re-identification) or that additional information can be derived from individuals in the synthetic dataset (Attribute Inference). For example, an attacker may try to train a machine learning model to reconstruct certain sensitive variables. This so-called Attribute Inference attack is trained on synthetic data and applied to (a part of) the real data. If variables of real individuals are predicted with certainty, privacy has been leaked.

On the other hand, there are methods that enhance privacy assurance. For example, by abstracting data, information about individuals is better protected, such as by using age categories instead of exact ages. Additionally, more complex methods intentionally add a degree of uncertainty, such as introducing noise. Differential Privacy is an important mathematically grounded method that limits information leakage within a pre-defined range, also known as the privacy budget. As expected, these methods reduce the quality of synthetic data, and this is referred to as the privacy vs. quality trade-off. With our open-source code, this is visualized in the spider plot in Figure 1. In this plot, you can see the influence of different privacy budgets on quality and which qualities deteriorate the most (Distinguishability) and the least (Bivariate). It is up to the user to thoroughly understand and minimize privacy risks appropriately according to the intended application.

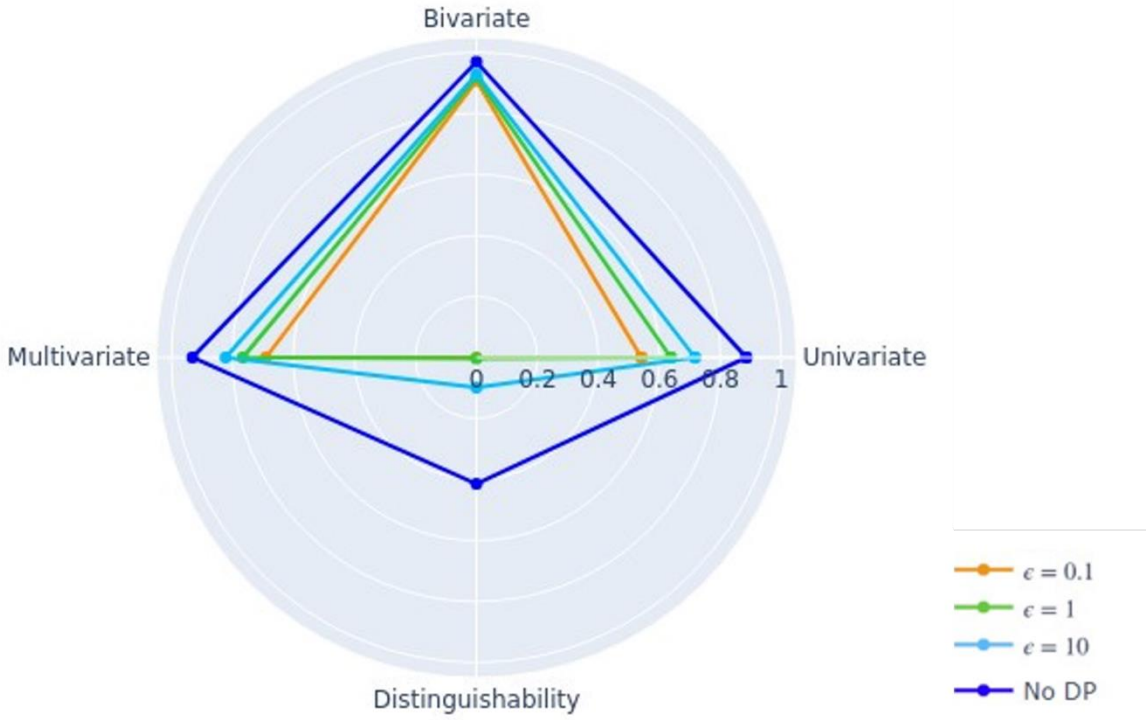

Figure 1: *Spider plot on the widely used benchmark Adult Income dataset generated with a generative AI CTGAN model trained with different privacy budgets (epsilon). Where epsilon=0.1 has the highest privacy assurance, and No DP (Differential Privacy) has the lowest. The lower quality at higher privacy is directly evident.*

So, the use of synthetic data has great potential to improve society. It can lead to better healthcare for diabetes patients. We are experiencing increasing interest in this topic and, consequently, in independently evaluating the quality of synthetic data. Our open-source code can contribute to this. Try it out for yourself at https://github.com/TNO-SDG/tabular.eval.utility_metrics. Let us know what you think!