Synthetic Data
Utility Evaluation for Machine Learning

Markus Hittmeir, Andreas Ekelhart, Rudolf Mayer

COMET
Competence Centers for Excellent Technologies
www.ffg.at/comet

SBA Research

Problem & Motivation

Due to technological advances, an increasing amount of micro-data, i.e., data that contains information about individuals, is collected. To comply with ethical and legal standards, data holders have to take privacy-preserving measures. Traditional concepts like k-Anonymity and Differential Privacy prevent intruders from learning sensitive information about individuals in the dataset. An alternative is the generation of *synthetic data*, which usually consists of the following steps:

1. **Data Description**: The original data is used to build a model which comprises information about the distribution of attributes and correlations between them.
2. **Data Generation**: The model is 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.

Ultimate Goal: Machine Learning models trained on synthetic data instead of the real data perform nearly as well. Simultaneously, the use of synthetic data reduces the risk of disclosure of sensitive information.

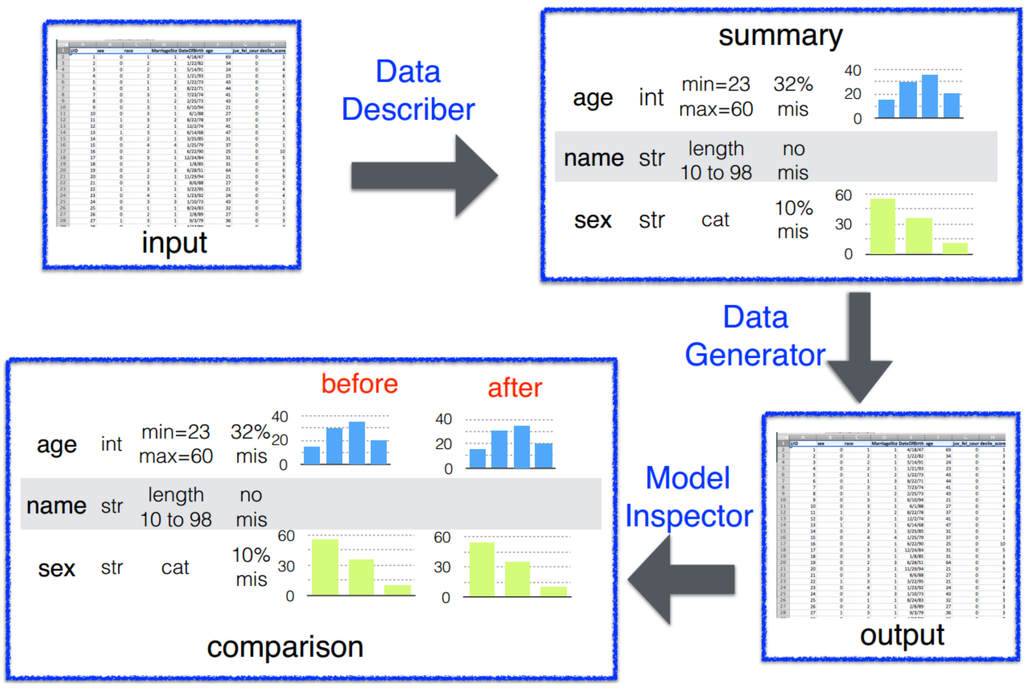

Figure 1: Workflow of the DataSynthesizer (Figure from Ref. 2)

Synthetic Data Generation Tools

- **Synthetic Data Vault (SDV)**[1]: This tool builds a model based on distribution estimates for each column and the covariance matrix for preserving correlations.
- **DataSynthesizer**[2]: The DataSynthesizer's model is based on a Bayesian network learned from the original data. For statistical disclosure control, the user is able to turn on Differential Privacy. In our experiments, we have evaluated the DataSynthesizer both with enabled (**DSP**) and disabled (**DS**) Differential Privacy.
- **synthpop (SP)**[3]: The default synthesis method is the CART algorithm. However, the user may specify a large number of parameters. Moreover, the implementation comes with its own function for statistical disclosure control.

Utility Evaluation for Classification

We performed classification tasks on standard benchmark data, such as the Adult Census dataset from the UCI Machine Learning Repository:

1. We split the data into a training (80 %) and a test (20 %) dataset.
2. We applied the synthesizers to the original training data to generate synthetic training data of equal length.
3. We trained machine learning classification models on both the original and the synthetic training data.
4. We compared the accuracy scores of these models on the test data.

General Findings

The **DS** and **SP** with standard settings achieve accuracy scores close to the models trained on original data. Using the **SDV** or the **DSP** usually leads to a loss of performance. Figure 2 shows an example, where **O** is the accuracy score of the original data and **B** the baseline score given by a Zero Rule classifier.

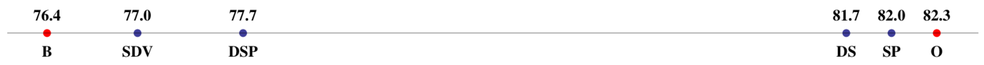

Figure 2: Accuracy scores of Logistic Regression on the Adult Census dataset

Utility Evaluation for Regression

Using a similar experimental setup as for classification and benchmark data such as the Bike Sharing dataset from Kaggle, we evaluated the utility of synthetic data for solving regression tasks. In such tasks, the target variable is continuous and the problem is not to predict a category, but a numerical value. The goal is to be as close to the sample's real value as possible.

General Findings

We compared the results of multiple utility measures, such as the mean average error (MAE) and the R2 score. In accordance with our analysis on classification tasks, the **DS** and **SP** with standard settings usually achieve scores close to the models trained on original data. Using the **SDV** or the **DSP** still leads to a performance loss. However, Figure 3 is an example of the **SDV**'s tendency to perform much better on regression than on classification problems, as its MAE is only slightly larger than the MAE for **SP** and **DS**.

Figure 3: MAE for Support Vector Regression on the Bike Sharing dataset

1. N. Patki, R. Wedge, K. Veeramachaneni, *The Synthetic Data Vault*, In: Proceedings of the 3rd DSAA, 2016.
2. H. Ping, J. Stoyanovich, B. Howe, *DataSynthesizer: Privacy- Preserving Synthetic Datasets*, In: Proceedings of the 29th SSDBM, 2017.
3. B. Nowok, G. M. Raab, C. Dibben, *synthpop: Bespoke Creation of Synthetic Data in R*, In: Journal of Statistical Software, 2016.
4. M. Hittmeir, A. Ekelhart, R. Mayer, *On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks*, In: Proceedings of the 14th ARES Conference, 2019.
5. M. Hittmeir, A. Ekelhart, R. Mayer, *Utility and privacy assessments of synthetic data for regression tasks*, In: Proceedings of the 7th IEEE Big Data Conference, 2020.

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
Bundesministerium Digitalisierung und Wirtschaftsstandort
FFG Forschung wirkt.
wirtschaftsagentur wien Ein Fonds der Stadt Wien
Europäische Kommission
FWF Der Wissenschaftsfonds.
netidee OPEN INNOVATIONS

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.