Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220389

Dataset Comparison Tool: Utility and Privacy

João COUTINHO-ALMEIDA[a,c,1], Ricardo João CRUZ-CORREIA[b] and Pedro Pereira RODRIGUES[b]
[a] *CINTESIS - Centre for Health Technologies and Services Research, University of Porto, Portugal*
[b] *MEDCIDS – Faculty of Medicine of University of Porto, Portugal*
[c] *Health Data Science PhD Program, Faculty of Medicine of the University of Porto, Portugal*

Abstract. Synthetic data has been more and more used in the last few years. While its applications are various, measuring its utility and privacy is seldom an easy task. Since there are different methods of evaluating these issues, which are dependent on data types, use cases and purpose, a generic method for evaluating utility and privacy does not exist at the moment. So, we introduced a compilation of the most recent methods for evaluating privacy and utility into a single executable in order to create a report of the similarities and potential privacy breaches between two datasets, whether it is related to synthetic or not. We catalogued 24 different methods, from qualitative to quantitative, column-wise or table-wise evaluations. We hope this resource can help scientists and industries get a better grasp of the synthetic data they have and produce more easily and a better basis to create a new, more broad method for evaluating dataset similarities.

Keywords. Synthetic data, utility evaluation, privacy evaluation

1. Introduction

Synthetic data can be defined as data that has no connection with a real-world phenomena or event. It was not originated from a process in the real world, but rather a synthetic one. The idea is that synthetic data can have similar properties with the real data, without needing to have an independent process for its generation.

Synthetic data has been used over the years for several usages, but in healthcare is still not very used. However, this scenario seems to be changing. It can be used for several use cases namely [1]; i) Software testing, ii) educational purposes, iii) machine learning, iv) regulatory, v) retention, vi) secondary and vii) enhanced privacy.

Software testing relates to using synthetic data to create use cases for software testing. This can be used for the development or pre-production stages for example. Often the data needed is not available on-demand and a synthetic generator of reliable data could be useful. Educational purposes relate to, at least, two different scenarios. One is for onboarding of employees [1], other is related to healthcare students for using health information systems and creating mechanisms for providing reliable data on-demand.

[1] Corresponding Author, João Coutinho-Almeida; E-mail: joaofcalmeida@outlook.com

Machine-learning is one of the areas where synthetic data has more widespread usage, where data augmentation through data synthesis is already common. It can be used for class-imbalance, sample-size boosting or machine-learning algorithms testing. Regulatory purposes could be important as well, with the rise of Artificial Intelligence (AI) as medical device systems and synthetic data could be used to properly evaluate these systems under controlled environments. Retention can be an important case for synthetic data as well, since personal data must not be kept more than it would be necessary. Synthetic data generators can be of use, where the original data can be deleted and a generator kept for further usage, given that the privacy mechanisms are properly employed. Secondary uses relate to using synthetic data to share data with academia or industry. Simulacrum [2] is a nice example of how the NHS uses these mechanisms to help scientists get a better grasp of data before having to fill documentation to query the real data. The same occurs for Integraal Kankercentrum Nederland (IKNL), which has a synthetic version of the cancer registry for scientific purposes [3] and the Medical and Healthcare products Regulatory Agency (MHRA) that uses synthetic data as well for its CPRD real-word evidence [4].

Finally, an aspect that is underlying to all these applications is the promise that synthetic data can be used to improve privacy. Even though specially tweaked data generators can be used to create more privacy-aware datasets, it will be inherently always at the cost of some utility [5]. So, even though synthetic data it is not the silver bullet as primarily thought, synthetic data generation can be undeniably used to help create more private data for all the use cases seen above at the cost of its utility. As for proper methods of evaluating security and utility, are, for now, open research questions. At the present time, it is still complicated to properly assess the utility of the generated data. We have qualitative and quantitative methods. Qualitative methods are related to plots, quantitative are related to some value that defines a evaluation metric. These quantitative metrics can be applied to equal columns from each data set, pair of columns from each dataset or applied over the whole datasets. As for privacy metrics, the metrics rely on duplicates. Full duplicates or membership inference related.

So, in this paper, we developed a data pipeline for data analysis in order to create a report for providing several metrics for data utility and privacy.

2. Methods

The pipeline relies on python and latex for creating the document. It relies also on several packages that implemented methods for evaluating data, namely *scipy* [6], *sdmetrics* [7] and *scikit-learn* [8] and *mlxtend* [9]. Its basis is related to uploading 2 datasets, and a report in pdf is produced. Being that is based on programmatic code, it can be easily converted into API.

The report has a section for dataset description, columns removed due to high-null and brief variable overview. Then a null comparison is done by column and dataset. Following this is the utility subsection. Firstly, by visual methodologies: heat maps for the correlation, bar plots for categorical, density plots for continuous and a pair plot for an overview. As for the quantitative utility evaluation, we divided it column-wise, pairwise and table-wise. The first comprehends the *Kolmogorov–Smirnov* test for continuous and chi-squared test for categorical variables. Distance metrics were also applied to categorical columns. First, they are transformed into distributions and then distance metrics are applied. The results are a descriptive overview of the distance

metrics, having minimum value, average, max value and standard deviation. The distance metrics selected were *Jensen-Shannon Divergence*, *Wasserstein distance*, *Kullback–Leibler divergence* and entropy.

As for pair-wise metrics, we used a discrete and continuous *Kullblack-Leibler divergence*. In this, two pairs of continuous columns are compared using *Kullback–Leibler divergence*. For this, they are put into bins for further application. The same is applied to categorical columns without binning. As for table-wise metrics, first, we used likelihood metrics. We fitted several Gaussian Mixture models or Bayesian network models to the real data and then calculate the likelihood of the synthetic data belonging to the same distribution. The metrics are likelihood for Gaussian mixture and Bayesian models and log-likelihood for the Bayesian model as well.

Then we used machine-learning models (linear regression and decision trees) to assess how similar models behave on both datasets. First, we tested on the same dataset in order to compare evaluation metrics. Then we cross-tested, meaning that the training set was drawn from one dataset and the test set was drawn from the second dataset. Finally, data privacy constraints duplicate evaluation, duplicate existence by removal of a single column and a record linkage approach. With the record linkage, we define a record linkage blocking ("age" in the example) and then try to match rows from the synthetic dataset to the real, with varying known attributes. Then matrix, euclidean and cosine distance was also calculated. Even though it is used for privacy evaluation, by definition, we could also use it for utility assessment. For proper assessment, the continuous and categorical variables should be indicated at the start of the code. The metrics are listed in the table 1.

Table 1. Metrics Assessed

Metric	Method	Context
Bar Plot	Visual	Utility
KDE Plot	Visual	Utility
Heat-map	Visual	Utility
Pair-plot	Visual	Utility
KS test	Column-quantitative	Utility
ChiSquared Test	Column-quantitative	Utility
Kullback–Leibler divergence	Column-quantitative	Utility
Jensen-Shannon Divergence	Column-quantitative	Utility
Wasserstein distance	Column-quantitative	Utility
Entropy	Column-quantitative	Utility
DiscKLD	Table-quantitative	Utility
ContinuousKLD	Table-quantitative	Utility
BNLikelihood	Table-quantitative	Utility
BNLogLikelihood	Table-quantitative	Utility
GMLogLikelihood	Table-quantitative	Utility
Same dataset ratio	Table-quantitative	Utility
Support rules	Table-quantitative	Utility
Different dataset validation	Table-quantitative	Utility
Duplicates	Quantitative	Privacy
Duplicate minus 1	Quantitative	Privacy
Record Linkage	Quantitative	Privacy
Matrix distance	Quantitative	Privacy/utility
Cosine distance	Quantitative	Privacy/utility
Euclidean distance	Quantitative	Privacy/utility

3. Results

A trial example of comparing data is available for data in the UCI repository, namely the heart disease dataset [10]. The synthetic data was created by using the synthpop package [11]. The variables evaluated are listed in table 1. The code can be seen in https://github.com/joofio/dataset-comparasion-report. As an example, the images for visual analysis for categorical (figure 1) and continuous variables (figure 2) are presented below.

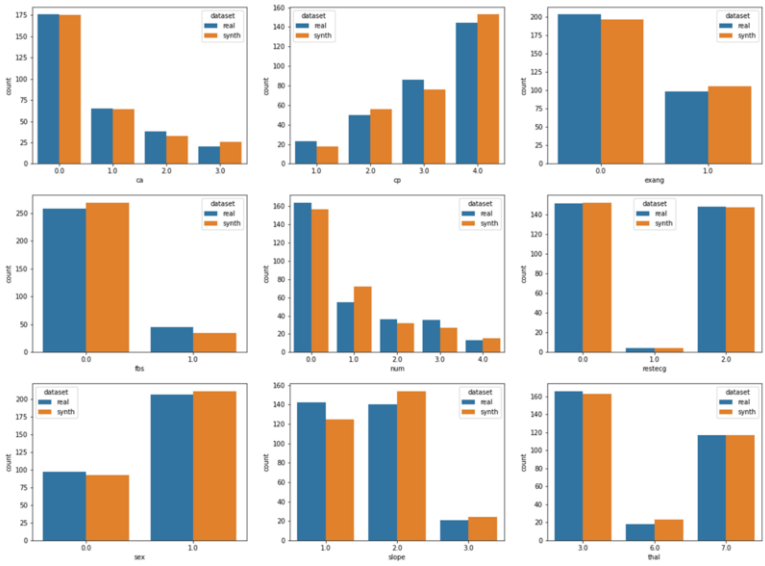

Figure 1. Categorical Variables Plotted

4. Discussion and Conclusions

The compiled evaluation metrics between two datasets are important not only for synthetic vs real datasets evaluation. For example, in distributed learning, where different silos exist, with similar or even equal features, a method for evaluating the similarities can be useful for understanding how the populations are similar between them, trying to shed light on the most similar among them, or different in order to understand the differences in the silos or data acquisition inside them. Furthermore, the differences can be assessed on a more granular level. The column-wise similarities can be different from the inter-columns' similarities and differences between these two metrics can be a metric of interest in itself regarding the quality of the synthetic data and its generator.

With this work, we hope to help institutions and academics getting to access to a benchmark of the datasets provided in order to leverage synthetic data in the healthcare space. Finally, we hope this work helps other researchers reach an evaluation metric that could be a unique and clear response to the question of how similar two datasets are.

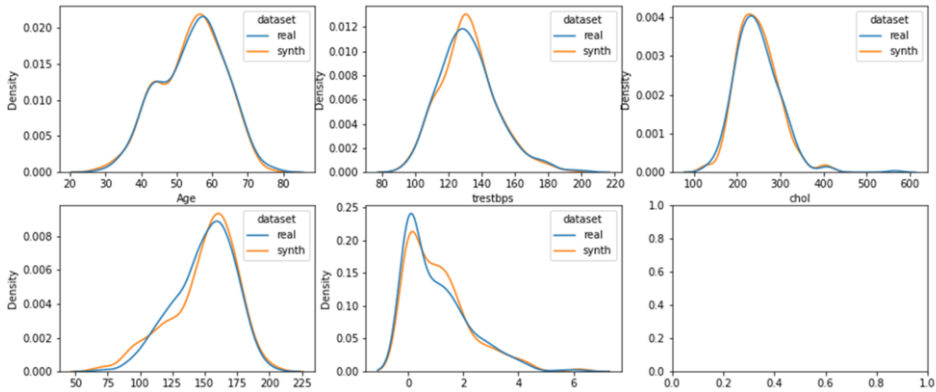

Figure 2. Continuous Variables Plotted

Acknowledgement

This work has been done under the scope of - and funded by - the PhD Program in Health Data Science of the Faculty of Medicine of the University of Porto, Portugal - heads.med.up.pt.

References

[1] James S, Harbron C, Branson J, et al. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell. 2021;1:15.
[2] The Simulacrum [Internet]. healthdatainsight.org.uk. 2022 [cited 2022 Jan 21]. Available from: https://healthdatainsight.org.uk/project/the-simulacrum/.
[3] Synthetische dataset NKR beschikbaar voor onderzoekers [Internet]. [cited 2022 Jan 21]. Available from: https://iknl.nl/nieuws/2021/synthetische-data-nkr-beschikbaar-voor-onderzoeker.
[4] Clinical Practice Research Datalink. CPRD cardiovascular disease synthetic dataset [Internet]. Clinical Practice Research Datalink; 2020 [cited 2022 Jan 21]. Available from: https://cprd.com/cprd-cardiovascular-disease-synthetic-dataset.
[5] Stadler T, Oprisanu B, Troncoso C. Synthetic data – A privacy mirage. arXiv. 2020;
[6] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17:261–272.
[7] Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. 2016 IEEE Int Conf Data Sci Adv Anal DSAA [Internet]. Montreal, QC, Canada: IEEE; 2016 [cited 2022 Jan 20]. p. 399–410. Available from: http://ieeexplore.ieee.org/document/7796926/.
[8] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.
[9] Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw [Internet]. 2018;3. Available from: http://joss.theoj.org/papers/10.21105/joss.00638.
[10] Janosi AMD, Steinbrunn W, Pfisterer M, Detrano R &MD. Heart Disease. 1988.
[11] Nowok B, Raab GM, Dibben C. synthpop: Bespoke Creation of Synthetic Data in R. J Stat Softw. 2016;74:1–26.