Comparative Evaluation of Synthetic Data Generation Methods

Ashish Dandekar
School of Computing,
National University of Singapore
Singapore, Singapore
ashishdandekar@u.nus.edu

Remmy A. M. Zen
School of Computing,
National University of Singapore
Singapore, Singapore
remmy@u.nus.edu

Stéphane Bressan
School of Computing,
National University of Singapore
Singapore, Singapore
steph@nus.edu.sg

ABSTRACT

Unrestricted availability of datasets is important for researchers and decision makers to evaluate their strategies to solve certain problems. Equally important is the privacy of the respective data owners. Synthetically generated datasets provide a way to retain the utility of the original data keeping the privacy of the owners invulnerable.

We perform a comparative study of synthetic data generation techniques using different data synthesizers like decision tree, random forest and neural network. We evaluate the effectiveness of these methods towards the amount of utility they preserve and the risk of disclosure.

KEYWORDS

Utility, Data Privacy, Synthetic Data Generation

ACM Reference Format:
Ashish Dandekar, Remmy A. M. Zen, and Stéphane Bressan. 2019. Comparative Evaluation of Synthetic Data Generation Methods. In *Proceedings of ACM Conference (Deep Learning Security Workshop).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Decision makers and researchers need data to validate their hypotheses. In order to have faithful analyses, publicly released datasets need to retain the utility from the original dataset; but the privacy of data owners is an equally important concern. In order to mitigate the risk of confidentiality breach, agencies employ different techniques such as reordering or recoding of sensitive variables, shuffling values among different records. In spite of these efforts by agencies, we have examples of confidentiality breaches in anonymized datasets [16][1]. If one keeps privacy of the data as the sole objective, then utility of the data is highly compromised. Therefore, there is a need for a way to generate datasets that can be made publicly available with minimum risk of data disclosure and maximum utility.

Fully synthetic datasets proposed by Rubin [14] and partially synthetic datasets proposed by Little [8] bridge the gap between these two contrasting issues of privacy and utility of open data. They use multiple imputation, a technique used to repopulate missing data, to generate synthetic records which preserve relationships in the population. Following up on these works of multiple imputation, Reiter et al. [1, 5, 11, 12] use different machine learning tools to generate synthetic data. These works treat values of synthetically generated attributes as missing values which are generated using models such as Decision Trees, Random Forest, Support Vector Machine, etc.

We comparatively evaluate synthetic data generation techniques using different data synthesizers: namely Linear Regression, Decision Tree, Random Forest and Neural Network. We evaluate their effectiveness in terms of how much utility is retained and their risk towards disclosure of individual data. We evaluate their efficiency in terms of time required to generate synthetic datasets. Given the tradeoff between the efficiency and effectiveness, we observe that neural networks are competitively effective towards reducing the risk of disclosure compared to other methods to generate the data.

In Section 2, we present the related work. Section 3 introduces the formalism of synthetic data generation using multiple imputation. The experiments and evaluation are presented in Section 4. Section 5 concludes this work.

2 RELATED WORK

Unlike the objective method of privacy estimation like ϵ-differential privacy [6], statistical disclosure limitation is an umbrella term given to different techniques which are applied on the dataset in order to reduce the risk of disclosure. Althought the aim of both differential privacy and statistical disclosure limitation techniques is same, SDL techniques approach is more specific to the dataset under consideration. Disclosure risk [17] is defined based on the assumptions of scenarios of data disclosure, that is, the scenarios under which an intruder might exploit the released data to reveal information of a record in the dataset. Generating a synthetic dataset is one of the techniques to reduce the risk of disclosure.

In [14], Rubin proposes a procedure to generate fully synthetic datasets, wherein values for certain attributes are replaced using multiple imputation for all datapoints in the dataset. Although it is advantageous to synthetically generate values for all datapoints, it is not always a necessity. Partially synthetic datasets, proposed by Little [8], are generated by synthetically generating the values of the attributes which are sensitive to public disclosure. Various data synthesizers such as decision tree [12], random forest [1], support vector machine [2] have been used to generate fully and partially synthetic data. Drechsler et al. [5] have performed an empirical comparative study between different data synthesizers. Comparison between fully and partially synthetic datasets can be found in [3]. Recently, Nowok et al. [9] have created an R package, *synthpop*, which provides basic functionalities to generate synthetic datasets and perform statistical evaluation.

In this work, we comparatively evaluate efficiency and effectiveness synthetic data generation techniques using different data synthesizers including neural networks.

[1] https://www.wired.com/2010/03/netflix-cancels-contest/

Deep Learning Security Workshop, December 2017, Singapore
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METHODOLOGY

Consider a population, $\mathcal{P}$, with a set of features. An imputer performs qualitative analysis of the features towards the risk of disclosure and divides the set of features into two disjoint subsets: X, a set of identifying features and Y, a set of sensitive features. Y contains sensitive data about records in the population; whereas X contains the data which can be publicly released or the data which is available to people from some alternative sources. Owner of the data divides the population into two sets: $\mathcal{P}_{obs}$ and $\mathcal{P}_{nobs}$. $\mathcal{P}_{nobs}$ is a holdout dataset which is never released and samples drawn from $\mathcal{P}_{obs}$ are released as datasets for public use. The problem is to conceive a mechanism to release samples of $\mathcal{P}_{obs}$ such that sensitive information is not disclosed.

Rubin [14] proposes the use of multiple data imputation [13] to generate fully synthetic dataset. Multiple data imputation techniques use a holdout data to fit a posterior distribution over the sensitive features given identifying features and fill in the missing values by generating samples from this posterior distribution. Error due to sampling in such an approach is twofold: error due to sampling from the population and error due to sampling from posterior. In order to alleviate these errors, Rubin proposes to release multiple synthetic datasets and he proposes to combine the results from individual datasets to get a final result.

Many times, different records have different risk of disclosure depending on their content. Records with a certain range of values for sensitive features possess very high risk of disclosure compared to other values of sensitive features. So, it is not always a necessity to synthetically generate values of sensitive features for all records in $\mathcal{P}_{obs}$. Little et. al. [8] propose partially synthetic data generation which uses multiple data imputation synthetically generates values for Y features for only those records in $\mathcal{P}_{obs}$ which have very high risk of disclosure.

Reiter et. al. [11] extend the idea of multiple imputation to the use of machine learning models to synthetically generate the data. We follow the steps of Reiter et. al. [1, 2, 11, 12] and extend the approach by using neural networks to synthetically generate the data. The general approach of using machine learning model is to train the model on the records in $\mathcal{P}_{nobs}$ and generate sensitive features given identifying features in $\mathcal{P}_{obs}$. One can use estimators defined in [10, 14] to calculate mean and variance of sensitive features using the released datasets. They are presented in Appendix B.

In this work, we use different machine learning models as data synthesizers to comparatively evaluate effectiveness of different models. We adapt and extend the works which use decision tree [12] and random forest [1] to neural network. We train a neural network with two hidden layers to generate synthetic values. We frame the problem as a K-class classification, where K is the number of unique values in each of the sensitive feature. There are K nodes in the output layer of the neural network with value at k-th neuron representing the probability of class k. To generate a value from a particular feature, we sample a class value using the output layer neuron values as a multinomial distribution.

4 EXPERIMENTAL EVALUATION

4.1 Dataset and Experimental Setup

We conduct experiments on a microdata sample of US Census in 2003 provided by IPUMS International [15]. The dataset consists of 1% sample of the original census data. It spans over 1.23 million households with records of 2.8 million people. It has several attributes of which not every single attribute is reported by all of the people. In order to avoid these discrepancies in the data, we follow the approach presented in [5] to consider the records of the heads of households. We treat this collection of the records of 316,276 households as the population. Please refer to Appendix A for the description of features in the dataset.

All programs are run on Linux machine with quad core 2.40GHz Intel® Core i7™processor with 8GB memory. The machine is equipped with two Nvidia GTX 1080 GPUs. Python® 2.7.6 is used as the scripting language.

4.2 Utility Evaluation

The utility of generated dataset needs to be evaluated at two different levels. Firstly, we need to evaluate differences between the distribution of original attribute values and generated attributes values. Secondly, we need to evaluate the difference between the quality of estimation of a certain estimand for synthetic data and generated data.

Let $y \in Y$ be any sensitive feature which we synthetically generate from original data. We calculate the similarity between the overall distribution of values of y by calculating **normalized KL-divergence** between the distribution of values of y in population and the distribution of synthetically generated values. For m synthetic datasets, we consider mean of normalized KL-divergence with individual datasets. Closer the value to 1, more similar the synthetically generated values are to the original values.

Karr et al. [7] develop a mechanism based on **overlap** between confidence intervals to evaluate effectiveness of specific estimands. We estimate mean and variance of y using the point estimators described Section 3. We construct a 95% confidence interval around the estimator using the formulae in Appendix B. Let (L_s, U_s) be confidence interval for synthetically generate y and (L_o, U_o) be interval from original data. We compute intersection of these intervals denoted as (L_i, U_i). The overlap utility measure is calculated as given in Equation 1.

$$I = \frac{(U_i - L_i)}{2(U_o - L_o)} + \frac{(U_i - L_i)}{2(U_s - L_s)} \tag{1}$$

The value of I is close to one if the utility is preserved and $I = 0$ refers to the dissimilar confidence intervals.

4.3 Disclosure Risk Evaluation

We follow Reiter [4, 11] to estimate the **risk of disclosure** in the synthetically generated dataset. Let **t** be a vector of information possessed by an intruder. We assume that aside from the sensitive features, the intruder has complete information about an identifying feature, say region of birth. For instance, the intruder might be interested in an individual who is born in Nevada with more than 70, 000$ salary. The intruder tries to match every record in target in **t** with the record in the released datasets. For a record $j \in$ **t**,

Feature	Data Synthesizers	Original Sample Mean	Partially Synthetic Data		
			Synthetic Mean	Overlap	Norm KL Div.
Income	Linear Regression	27112.61	27117.99	0.98	0.54
	Decision Tree	27143.93	27131.14	0.94	0.53
	Random Forest	27107.04	27254.38	0.95	0.58
	Neural Network	27069.95	27370.99	0.81	0.54
Age	Linear Regression	49.83	24.69	0.50	0.55
	Decision Tree	49.83	49.83	0.90	0.56
	Random Forest	49.82	49.74	0.95	0.56
	Neural Network	49.87	49.78	0.90	0.56

Table 1: Utility Evaluation for Partially Synthetic Datasets

the intruder may find multiple records with an exact match on the features of interest. An intruder is said to be successful in identifying the record if the intruder finds only one record with exact match.

For a record $j \in \mathbf{t}$, an intruder may find multiple records with the same value of maximum probability. Let, R denotes the set of records in $\mathbf{t}$ for which only one record in the dataset is matched with highest probability. Set R can be decomposed into two mutually exhaustive sets T and F which denote set of records with true matches and false matches respectively. In order to evaluate the risk of disclosure, we calculate *true match rate* and *false match rate*. They are calculated as given in Equation 2. The smaller the true match rate, the better is the performance of data synthesizer and the opposite for false match rate.

$$\begin{aligned} \text{true match rate} &= \frac{|T|}{|\mathbf{t}|} \\ \text{false match rate} &= \frac{|F|}{|R|} \end{aligned} \tag{2}$$

For further details about the calculation please refer to [4, 11].

4.4 Evaluation

The process starts with drawing 1% sample from the population which is treated as the original sample dataset. We synthetically generate values for two attributes: income and age, in the same order. Interested readers can refer to [12] for a detailed discussion on choosing the order of synthesis. We generate 5 synthetic datasets for each original sample dataset. We repeat this procedure for 500 original samples and mean of various metrics over 500 iterations is reported.

In order to generate synthetic datasets, we need to define the cutoffs on the feature values which determine if the record contains sensitive information. We consider records that have more than 70000$ income value and less than 26 age value to be the ones with sensitive information. So, we synthetically generate values for *age* and *income* for the records which fit these criteria. Utility evaluation results are presented in Table 1. We observe that synthetic datasets show high extent of overlap with the original dataset. In case of linear regression, a large deviation from original sample mean is observed in case of *age* feature. Linear regression learns parameters such that squared loss on the training data is minimized. As specified in Section 3, the model is fit on original values of *income* and synthetically generated values of *age* are used for generating new *age* values. Decision tree and other models are not prone to overfitting the training data. Therefore, we do not observe this degradation in effectiveness due to order of variable synthesis.

We require a scenario to evaluate *the risk of disclosure*. We select the scenario by performing an exploratory analysis on the population. The records which occur sparsely in the population, for instance records of people who are born in middle east with a certain income threshold, equally sparsely occur in a small sample. In order to statistically evaluate risk of disclosure, we need to have at least a handful of targets for evaluation.

Taking into account these requirements, we suppose that an intruder is interested in people who are born in US and have income more than 250, 000$. All these people are the targets of the intruder. Intruder tries to match every single target with the records in the released datasets. We consider two records perfectly match if people representing the records are born in US, they have income more than 250, 000$ and the age of the person in dataset in within the tolerance of 2 compared to target person.

The risk assessment is presented in Table 2. We see that neural networks are better at reducing the risk of disclosure than rest of the data synthesizers.

Data Synthesizers	True MR	False MR
Linear Regression	0.06	0.82
Decision Tree	0.18	0.68
Random Forest	0.35	0.50
Neural Network	0.03	0.92

Table 2: Disclosure Risk Evaluation

We comparatively analyse efficiency of both these techniques using different data synthesizers. The running time, in seconds, for generating 5 synthetic datasets is reported in Table 3. We observe that the neural networks achieve the low disclosure risk at the cost of a high value of running time.

Data Synthesizers	Linear Regression	Decision Tree	Random Forest	Neural Network
Time (s)	0.040	0.048	3.350	0.510

Table 3: Efficiency Analysis: Each cell shows running time to generate 5 synthetic datasets in seconds

5 CONCLUSION AND DISCUSSION

In this work, we present our preliminary results in comparative study of synthetic dataset generation techniques using different data synthesizers, namely linear regression, decision tree, random forest and neural network. We evaluate utility by using statistical estimators provided by Raghunathan et. al. [10]. We address privacy aspect by calculating the risk of disclosure for synthetically generated datasets. The analysis shows that neural networks are competitively effectives compared to other methods in terms of utility and privacy. They achieve this effectiveness at the cost of running time.

Following Rubin [14], we also conducted experiments to generate fully synthetic datasets. Fully synthetic dataset is generated by using $\mathcal{P}_{nobs}$ as a holdout dataset and synthetically generating values of sensitive features for all records in $\mathcal{P}_{obs}$. Therefore, fully synthetic datasets provide higher effectiveness towards reducing the risk of disclosure. Since one needs to synthetically generate entire population $\mathcal{P}_{obs}$, fully synthetic datasets require higher computation time and they show lower retention of utility.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Corporate Laboratory@University Scheme, National University of Singapore, and Singapore Telecommunications Ltd.

REFERENCES

[1] Gregory Caiola and Jerome P Reiter. 2010. Random Forests for Generating Partially Synthetic, Categorical Data. *Trans. Data Privacy* 3, 1 (2010), 27–42.

[2] Jörg Drechsler. 2010. Using Support Vector Machines for Generating Synthetic Datasets.. In *Privacy in Statistical Databases.* Springer, 148–161.

[3] Jörg Drechsler, Stefan Bender, and Susanne Rässler. 2008. Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. *Trans. Data Privacy* 1, 3 (2008), 105–130.

[4] Jörg Drechsler and Jerome P Reiter. 2008. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *International Conference on Privacy in Statistical Databases.* Springer, 227–238.

[5] Jörg Drechsler and Jerome P Reiter. 2011. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis* 55, 12 (2011), 3232–3243.

[6] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[7] Alan F Karr, Christine N Kohnen, Anna Oganian, Jerome P Reiter, and Ashish P Sanil. 2006. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60, 3 (2006), 224–232.

[8] Roderick JA Little. 1993. Statistical analysis of masked data. *Journal of Official statistics* 9, 2 (1993), 407.

[9] Beata Nowok, Gillian Raab, and Chris Dibben. 2016. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, Articles* 74, 11 (2016), 1–26.

[10] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. 2003. Multiple imputation for statistical disclosure limitation. *Journal of official statistics* 19, 1 (2003), 1.

[11] Jerome P Reiter. 2003. Inference for partially synthetic, public use microdata sets. *Survey Methodology* 29, 2 (2003), 181–188.

[12] Jerome P Reiter. 2005. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* 21, 3 (2005), 441.

[13] Donald B Rubin. 1986. Basic ideas of multiple imputation for nonresponse. *Survey Methodology* 12, 1 (1986), 37–47.

[14] Donald B Rubin. 1993. Discussion statistical disclosure limitation. *Journal of official Statistics* 9, 2 (1993), 461.

[15] Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. Integrated Public Use Microdata Series: Version 6.0 [dataset]. (2015). https://doi.org/10.18128/D010.V6.0

[16] Latanya Sweeney. 1997. Computational disclosure control for medical microdata: the Datafly system. In *Record Linkage Techniques 1997: Proceedings of an International Workshop and Exposition.* 442–453.

[17] Matthias Templ. 2008. Statistical disclosure control for microdata using the R-package sdcMicro. *Transactions on Data Privacy* 1, 2 (2008), 67–85.

APPENDIX

A DATASET DESCRIPTION

Table 4 lists the features in our dataset.

Attribute Name	Variable Type
House Type	Categorical
Family Size	Ordinal
Sex	Categorical
Age	Ordinal
Marital Status	Categorical
Race	Categorical
Educational Status	Categorical
Employment Status	Categorical
Income	Ordinal
Birth Place	Categorical

Table 4: Dataset Description

B ESTIMATORS OF MEAN AND VARIANCE

Let $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^m\}$ denote set of m synthetic datasets generated using the technique of multiple imputation. Suppose, we want to estimate an estimand Q. Let q_l and v_l denote sample mean and sample variance of Q using dataset D^l. Let $\bar{q_m}$, b_m and $\bar{v_m}$ denote mean of sample means, "between imputation variance" and "within imputation variance" over m datasets respectively. They are calculated as defined in Equation 3. Actual estimation of estimand Q depends upon the synthetic data generation procedure. We use these for the estimation purpose as described in the later section.

$$\begin{aligned} \bar{q_m} &= \frac{1}{m}\sum_{l=1}^{m} q_l \\ b_m &= \frac{1}{m-1}\sum_{l=1}^{m}(q_l - \bar{q_m})^2 \\ \bar{v_m} &= \frac{1}{m}\sum_{l=1}^{m} v_l \end{aligned} \tag{3}$$

$\bar{q_m}$ serves as an estimator of mean of Q. For fully synthetic dataset T_f is the estimator of variance in Q with degrees of freedom v_f; whereas for partially synthetic dataset T_p is the estimator of various in Q with degrees of freedom v_p. They are given in Equation 4 and Equation 5 respectively.

$$\begin{aligned} T_f &= b(1 + m^{-1}) - \bar{v_m} \\ v_f &= (m-1) * (1 - \frac{v_m}{b_m(1 + m^{-1})})^2 \end{aligned} \tag{4}$$

$$\begin{aligned} T_p &= \bar{v_m} + \frac{b}{m} \\ v_p &= (m-1) * \left(1 + \frac{v_m}{b_m}\right)^2 \end{aligned} \tag{5}$$

C CURRICULUM VITAE

C.1 Ashish Dandekar

Ashish Dandekar is Ph.D. candidate in the Department of Computer Science of the School of Computing (SoC) of the National University of Singapore (NUS). His research interest is in data-driven simulation and generation.

C.2 Remmy A. M. Zen

Remmy A. M. Zen is Ph.D. student in the Department of Computer Science of the School of Computing (SoC) of the National University of Singapore (NUS). Remmy received his master degree from Faculty of Computer Science of the University of Indonesia in 2014. His research interest is in Deep Learning.

C.3 Stéphane Bressan

Stéphane Bressan is Associate Professor in the Department of Computer Science of the School of Computing (SoC) of the National University of Singapore (NUS). Stéphane is Track leader for Maritime Information Technologies at NUS Centre for Maritime Studies (CMS). He is Affiliate Professor at NUS Business Analytics Centre. He is researcher at Image & Pervasive Access Lab (IPAL) (Singapore-France CNRS UMI 29255). Stéphane's research interest is the integration, management and analysis of data from heterogeneous, disparate and distributed sources.