COMPARATIVE EVALUATION OF SYNTHETIC DATA GENERATION METHODS

Ashish Dandekar, Remmy A. M. Zen, Stephane Bressan
(ashishdandekar, remmy)@u.nus.edu, steph@nus.edu.sg
NUS-Singtel Cyber Security Research & Development Laboratory

MOTIVATION

Unrestricted availability of datasets is important for researchers and decision makers to evaluate their strategies to solve certain problems. At the same time, the privacy of the respective data owners must be respected. Synthetically generated datasets provide a way to maintain the utility of the original data while keeping the privacy of the data owners.

We present a comparative study of synthetic data generation techniques using different data synthesizers: linear regression, decision tree, random forest and neural network. We comparatively evaluate the effectiveness of the four methods by measuring the amount of utility that they preserve and the risk of disclosure that they incur.

MULTIPLE DATA IMPUTATION

- Multiple data imputation [1] is a technique to repopulate missing values in the data.
- Rubin [2] proposes the use of multiple data imputation to synthetically generate values of sensitive features in the data.
- Partially synthetic dataset [3] generation uses multiple data imputation to synthetically generate records with high *risk of disclosure*.

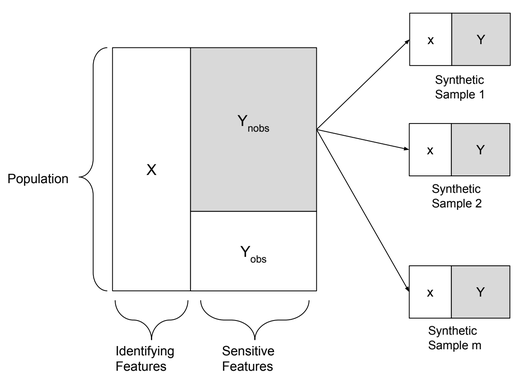

Figure 1: Multiple Data Imputation procedure

- Reiter et. al. [4] extend the idea of multiple imputation to the use of machine learning models to synthetically generate the data.

UTILITY EVALUATION

- We evaluate the similarity between distribution of values of sensitive features by calculating *normalized KL-divergence* between histogram of sensitive features in population and synthetic data.
- We calculate overlap between 95% confidence intervals of *mean estimator* of the sensitive feature using synthetically generated data and population.

DISCLOSURE RISK EVALUATION

- Disclosure risk estimated under the assumption that an intruder possess information about all insensitive records which she further uses to find the value of sensitive feature of a record.
- Let, **t** be the list of documents which an intruder is interested in. R be the set of records in **t** for which only one record in the dataset is matched with highest probability. R can be further split into two disjoint subsets T, a set of records in R with *true* matches and F, a set of records in R with *false* matches.
- True Match Rate (True MR) = |T| / |**t**|
  False Match Rate (False MR) = |F| / |R|

EXPERIMENTAL DESIGN

Dataset. A 1% microdata sample of US Census in 2003 provided by IPUMS International [5]. We focus on the sample of 316,276 heads of households in the population.

Experiment Design. We draw 1% sample from the population which is treated as the original sample dataset. We synthetically generate 5 datasets with synthetic values of income and age. We repeat this procedure for 500 original samples and mean of various metrics over 500 iterations is reported.

Disclosure Risk Estimation Scenario. An intruder who is interested in people who are born in US and have income more than 250, 000$.

RESULTS

DATA SYNTHESIZERS	ORIGINAL SAMPLE MEAN	SYNTHETIC MEAN	OVERLAP	NORM KL DIV
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

Table 1: Utility evaluation results for generation of "Income" attribute

Figure 2: Disclosure Risk evaluation

DATA SYNTHESIZERS	TIME (s)
Linear Regression	**0.040**
Decision Tree	0.048
Random Forest	3.350
Neural Network	0.510

Table 2: Efficiency evaluation

CONCLUSION

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. The analysis shows that neural networks are competitively effective compared to other methods in terms of utility and privacy. They achieve this effectiveness at the cost of running time.

ACKNOWLEDGEMENT

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] Donald B Rubin. 1986. Basic ideas of multiple imputation for nonresponse. Survey Methodology 12, 1 (1986), 37–47.
[2] Donald B Rubin. 1993. Discussion statistical disclosure limitation. Journal of Official Statistics 9, 2 (1993), 461
[3] Roderick JA Little. 1993. Statistical analysis of masked data. Journal of Official Statistics 9, 2 (1993), 407.
[4] Jerome P Reiter. 2003. Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 2 (2003), 181–188.
[5] 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