INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

AnigeD

GENERATING SYNTHETIC DATA IS COMPLICATED: KNOW YOUR DATA AND KNOW YOUR GENERATOR

PSD2024: Privacy in Statistical Databases 2024,
26. September, 2024

Jonathan Latner, PhD
Dr. Marcel Neuenhoeffer
Prof. Dr. Jörg Drechsler

SECTION 1: INTRODUCTION

OVERVIEW

- Common perception that making synthetic data is easy
- We to show that its complicated
  - You need to know your data
    - Missing values, messy data, etc.
  - You need to know your synthetic data generator (SDG)
    - Compare 3 SDGs: DataSynthesizer, CTGAN, Synthpop
    - How does it deal with missing values?
    - How computationally efficient is it (in terms of duration in time)?
    - How does it meet privacy standards? (but not today)
- Conclusion - Every SDG has advantages/disadvantages (no one, correct solution)
  - Synthpop is good, but has problem with dimensionality
  - DataSynthesizer is not as good, but can set ϵ-DP
  - CTGAN is bad, but maybe the problem is CTGAN, not GANs in general

THE GOOD NEWS – MAKING SYNTHETIC DATA IS EASY

- `Gretel.ai`: The synthetic data platform for developers. Generate artificial datasets with the same characteristics as real data, so you can develop and test AI models without compromising privacy.
- `Mostly.ai`: Synthetic Data. Better than real. Still struggling with real data? Use existing data for synthetic data generation. Synthetic data is more accessible, more flexible, and simply...smarter.
- `Statice.ai`: Generating synthetic data comes down to learning the joint probability distribution in an original, real dataset to generate a new dataset with the same distribution. The more complex the real dataset, the more difficult it is to map dependencies correctly. Deep learning models such as generative adversarial networks (GAN) and variational autoencoders (VAE) are well suited for synthetic data generation.
- `hazy.com`: Synthetic data does not contain any real data points so can be shared freely. Say goodbye to lengthy governance processes associated with real data. Specifically, Hazy data is designed to preserve all the patterns, statistical properties and correlations in the source data, so that it can be used as a drop-in replacement for it.
- DataSynthesizer: The distinguishing feature of DataSynthesizer is its usability — the data owner does not have to specify any parameters to start generating and sharing data safely and effectively.

THE BAD NEWS – MAKING SYNTHETIC DATA IS HARD

- According to the Alan Turing Institute (Jordan et al., 2022)
- Synthetic data is not a replacement for real data. It is a distorted version of the real data.
  - Why are we creating synthetic data?
  - Agreeing on the goal will help us make decisions (synthesizer, measures, etc.)
- How does the synthesizer work? from complete black box to complete user choice
- How different should it be? How do we measure the difference (utility and fidelity)?
  - Utility and fidelity are sometimes called general/broad or specific/narrow measures within the single concept of utility (Snoke et al., 2018; Drechsler and Reiter, 2009).
- Computationally efficiency (i.e. duration in time) is important and often ignored. The algorithm should scale well with the dimension of the data space in a relational way, not exponential way.
- How do we evaluate privacy? (not today)

OUR GOAL IS TO ILLUSTRATE THE CHALLENGES

- Know your data (1 dataset)
  - Social Diagnosis 2011 (SD2011) - Cleaning/pre-processing (most evaluations use clean data)
- Know your generator
  - Evaluate 3 synthetic data generators (SDG): DataSynthesizer, CTGAN, Synthpop
    - How do they actually work? (only briefly described here)
- 4 utility measures
  - Propensity score mean-squared error (pMSE) - Append the original and synthetic datasets. Create an indicator variable for original/synthetic datasets. The probability of being in the synthetic dataset is computed for each record in the combined dataset (n); this is the propensity score (p). Lower scores are better. ($pMSE = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2$)
  - Ratio of counts/estimates (ROC/ROE) - Calculate the ratio of each value in a given variable for both synthetic/original datasets. Then, calculate the ratio of each value for each dataset, and divide the smaller of these two estimates by the larger one. Higher scores are better. ($ROE = \frac{min(y^1_{orig}, y^1_{synth})}{max(y^1_{orig}, y^1_{synth})}$)
  - Confidence interval overlap from 2 regression models (OLS, GLM)
  - Computationally efficient with respect to duration in time

SECTION 2: KNOW YOUR DATA (SD2011)

REAL DATA

- Social Diagnosis 2011 (SD2011)
- Loads with Synthpop
  - `http://www.diagnoza.com/index-en.html`
  - Not entirely clear how original data is created or cleaned to create data in Synthpop
    - No
- Like real data, has ‘quirks’ or unusual values/variables
  - Includes missings
    - Informative (i.e. for never worked abroad, `wkabdur` is missing)
    - Non-informative
  - Includes ‘errors’
    - `smoke` - Does smoke is NO, but `nociga` - 20/22 cigarettes per day
    - `bmi` = 451, but `height`(cm) = 149 and `weight`(kg) = NA (999)
  - Includes generated variables (Can be problematic for SDGs)
    - `bmi`, `agegr`

DATA (SD2011)

Number	Variable	Description	Type	Observations	Unique.Values	Missings	Negative.values	Generated	Messy
1	sex	Sex	factor	5000	2	0	0		
2	age	Age of person, 2011	numeric	5000	79	0	0		
3	agegr	Age group, 2011	factor	5000	7	4	0	Yes	Yes
4	placesize	Category of the place of residence	factor	5000	6	0	0		
5	region	Region (voivodeship)	factor	5000	16	0	0		
6	edu	Highest educational qualification, 2011	factor	5000	5	7	0		
7	eduspec	Discipline of completed qualification	factor	5000	28	20	0		
				…					
10	income	Personal monthly net income	numeric	5000	407	683	603		
11	marital	Marital status	factor	5000	7	9	0		
12	mmarr	Month of marriage	numeric	5000	13	1350	0		
14	msepdiv	Month of separation/divorce	numeric	5000	13	4300	0		
15	ysepdiv	Year of separation/divorce	numeric	5000	51	4275	0		
				…					
22	nofriend	Number of friends	numeric	5000	44	0	41		
23	smoke	Smoking cigarettes	factor	5000	3	10	0		
24	nociga	Number of cigarettes smoked per day	numeric	5000	30	0	3737		Yes
				…					
28	wkabdur	Total time spent on working abroad	numeric	5000	33	0	4875		Yes
				…					
33	height	Height of person	numeric	5000	65	35	0		
34	weight	Weight of person	numeric	5000	91	53	0		
35	bmi	Body mass index (weight - kg/(height - cm^2)*10000)	numeric	5000	1396	59	0	Yes	Yes

SECTION 3a): KNOW YOUR GENERATOR (DATASYNTHESIZER)

“DataSynthesizer, a Python package, implements a version of the PrivBayes (Zhang et al., 2017) algorithm. DataSynthesizer learns a differentially private Bayesian Network which captures the correlation structure between attributes and then draws samples.” (Little et al., 2021)

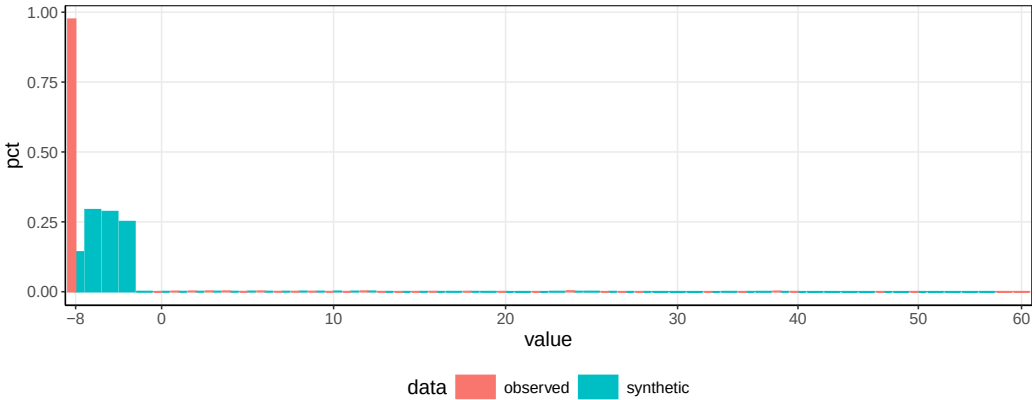
Variable type: The Bayesian network only works with discrete variables. One way to discretize continuous variables is by binning them.

HYPERPARAMETERS

- ϵ Differential Privacy (DP): we turn it off (default 0.1)
- *k*-degree Bayesian network (parents): 1 (independent), 2, 3, . . . (default=3)
  - In reality, *k* is not known, but maximum value of *k* = number of columns - 1
  - in PrivBayes, *k* is a function of tuples (rows), ϵ, and attributes (columns) (Default is 'greedy')
    - Computational Efficiency: The greedy algorithm makes the Bayesian network construction process faster by only considering the local best choice at each step, but if *k* is too large, the algorithm might take longer because it has more potential parent nodes to evaluate.
    - Balancing Utility and Privacy: A greedy approach with a high *k* could result in more complex networks that could more closely model the original data, but also reduce the privacy of the synthetic data.

VARIABLE: WKABDUR (WORK ABROAD DURATION)

Figure 1: Captures values < 0 as continuous, not missing/categorical



GENERATED VARIABLES (AGEGR)

Figure 2: Number of misclassified