Master's Thesis

Towards Data Privacy: Evaluating different synthetic data approaches with the data from Covid-19 Trends and Impact Surveys

Department of Statistics
Ludwig-Maximilians-Universität München

Yue Xiong

Munich, Month Dayth, 2022

Submitted in partial fulfillment of the requirements for the degree of M. Sc.
Supervised by Dr. Anna-Carolina Haensch

Abstract

To be completed...

Contents

1 Introduction

? introduced this and that. Another statement that needs a reference, but the authors are not named directly (**?**). Another statement where the reference is just one possible source (see, e.g., **?**).

2 Related Work

In this chapter, general definition of data privacy and the necessity to maintain data privacy are described. Furthermore, in order to keep data privacy, this chapter also presents an overview of methods to achieve the goal of privacy preserving data analysis and publication, which have been adopted in several domains. Here, we want to emphasize the use of synthetic data and one of its most popular applications, i.e., differentially private synthetic data, that are able to prevent disclosure in the process of synthetic data generation.

2.1 Data Privacy

It is well-acknowledged that we have entered a data-driven world and data are often regarded as significant constituents for our society. At the same time, an open society can also learn from these data so as to develop feasible and practical policy guidelines (**?**). Especially during the outbreak of coronavirus disease 2019 (COVID-19), more and more increased concerns are raised that it is essential for a society to utilize such data, which are widely-spread in the population and analyzed with regard to various perspectives, to advance sophiscated planning and develop more concrete social welfare benefits for the citizens. Consequently, both seen from the public health perspective and the economy perspective, the on-going COVID-19 global pandemic serves as a rigid reminder that detailed data are urgently needed to assist in decision making, damage control scenarios. Regardless of prevalent consensus reached to leverage more microdata, the inappropriate use of such information can cause harm in data confidentiality and privacy as sometimes the attacks from an intruder may result in the leakage of an individual's sensitive information, e.g., identity, address and salary, etc.

On the premise of possible outcomes brought by the misuse of microdata, it is crucial that we encourage proper and legal use of the collected datasets. Holding this motivation, researchers have developed a variety of strategies aiming to avoid the disclosure of sensitive records while revealing these specific information to the public (**?**). In the early times, several traditional methods have been proposed to limit data disclosure with strategies like top-coding, swapping or data suppression. Nevertheless, with increased computing power and more data access demanded by the public, the risks of data disclosure are often seen as underestimated using simply these traditional protection strategies, where instances of privacy breaches can be found simultaneously in the public and private sectors (**?**).

Alternatively, with the purpose to reach the trade-off between data disclosure protection and broad data access, the idea of synthetic data has been released. When using this approach, we make a model fitted to the microdata and the corresponding outputs from the fitted model are then used to replace the original values in the previous information.

2.2 An Overview of Data Synthesis Approaches

The field of employing synthetic data to avoid data statistical disclosure has been introduced by **?** and **?** in the context of learning multiple imputation (MI) for nonresponse (**?**). In their work, they showed the possibility to get these sentitive values replaced with "imputed" values rather than impute data for those missing records in the original dataset.

With the application of this multiple imputation framework, ramdom draws sampled from these imputed populations can then be circulated to the public. In more extreme cases, when it is necessary to avoid the release of original data completely, those instances from the original sample can also be displaced by samples from the imputation model.
Depending on the level of protection prone to specific scenarios, data synthesis methods are categorized to two groups, i.e., partial synthesis and full synthesis. For partially synthetic data, only some parts of the original records are synthesized. On the contrary, the entire dataset are replaced by synthetic values with the utilizatio of full synthesis methods specifically. It is evident to infer that the desired protection level is high when applying fully synthetic data methods, as original instances are completely excluded.

2.2.1 Computer Science Approaches

Despite the origin and early development of synthetic data, in computer science, the data Synthesis approach did not raise much interest in the study of data privacy until the advent of various privacy standards. In order to cater to people's requirement of privacy protection, scientists have defined several popular data privacy standards such as $k-$anonymity (**?**), $l-$diversity (**?**) and $t-$closeness (**?**). In the following context, we try to shortly introduce the key ideas of the three popular standards.

$k-$anonymity

$l-$diversity

$t-$closeness

2.2.2 Statistical Approaches

2.2.3 Differentially Private Data Synthesis

3 Synthetic Data and Differential Privacy

? introduced this and that. Another statement that needs a reference, but the authors are not named directly (**?**). Another statement where the reference is just one possible source (see, e.g., **?**).

4 Evaluation of Utility and Risk Assessment of Remaining Disclosure

? introduced this and that. Another statement that needs a reference, but the authors are not named directly (**?**). Another statement where the reference is just one possible source (see, e.g., **?**).

5 Evaluating Different Synthetic Datasets generated from CTIS

Additional material goes here

6 General Discussion

A concise summary of contents and results

7 Conclusion

A concise summary of contents and results

A Appendix

Additional material goes here

B Electronic appendix

Data, code and figures are provided in electronic form.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Location, date

Name