ANITA

Anonymous big data A project funded by FFG

Quantifiable targets for accuracy and privacy

Deliverable 2.2

Author: Michael Platzer

Reviewer: Olha Drozd

Document version: 1.2
Date: 31.03.2020

Federal Ministry
Republic of Austria
Climate Action, Environment,
Energy, Mobility,
Innovation and Technology

Programme **"ICT of the Future"** -
the initiative of the Austrian Federal
Ministry for Climate Action,
Environment, Energy, Mobility,
Innovation and Technology (BMK)

Disclaimer

This deliverable describes the work and findings of the AI-Based Privacy-Preserving Big Data Sharing for Market Research (Anonymous Big Data (ANITA)) project.

The authors of this document have made every effort to ensure that its content was accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this deliverable are responsible for any possible errors or omissions as well as for any results and actions that might occur as a result of using the content of this document.

Table of contents

1 Introduction

Based on the use case requirements put forward by the consortium partners, a number of quantitative measures, both for accuracy and privacy of the generated synthetic data with respect to the original data, have been identified. These measures will be implemented and reported throughout the remainder of the project. Subsequently, we will assess the utility of the generated synthetic data.

Primary motivation to provide quantifiable targets is to allow the automation of quality assurance, as well as comparison across methods and approaches. The fundamental trade-off between accuracy and privacy for synthetic data will remain the same as for any other anonymization approach. Whereas, conceptionally, the case can be made that AI-generated synthetic data, given the same level of privacy, should allow for higher level of accuracy when compared to classic anonymization techniques that retain the 1:1 relationship to actual individuals (see Figure 1). The computation of hard measures is needed to allow for non-subjective validation.

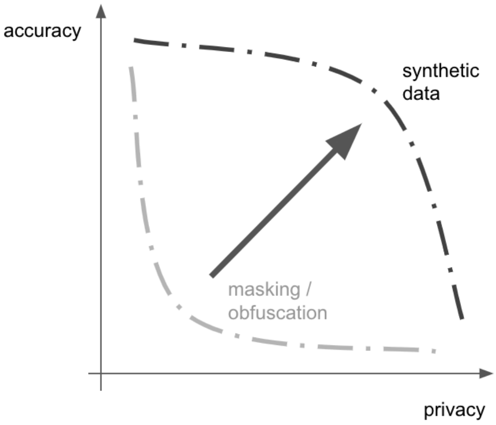

Figure 1: Classic anonymization vs synthetic data

Ideally, synthetic data captures all the characteristics, the patterns, and the correlations of the original data without disclosing private information of any single individual from the original data. To put it simply, synthetic data is expected to be *"as close to the original data as possible, but without being too close"*. This already points towards the need for a distance measure between two sets of data, as well as for a reference point to assess whether that distance is too small.

2 Quantifiable targets for accuracy

One can think of a data set to be the realization of independent and identically distributed random draws made from a multivariate probability distribution. Thus, statistical distance[1] measures, that quantify distance between probability distributions, provide a good list of candidates, with the total variation distance[2] (**TVD**) being among the most commonly used measures:

$$\delta(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

With the true probability distribution remaining unknown, we need to turn towards the empirical probability distribution, i.e., the relative frequency based on the observations. The advantage of the TVD is that it allows for easy interpretation, as it is the largest possible difference in probabilities that the two distributions can assign to the same event. Another candidate is the L1-Distance[3] (**L1D**) between the empirical distributions, that is the sum over all absolute deviations, as well as the Hellinger distance[4]. In particular L1D offers the advantage that it remains comparable across attributes, independent of their cardinality.

However, the problem is that the number of possible value combinations grows exponentially with the number of attributes - a phenomenon commonly referred to as the curse of dimensionality[5]. Thus, in the typical case of a high-dimensional data space, with a sheer unlimited number of potential outcomes, the probability of a single particular event quickly goes towards zero, rendering the absolute difference meaningless.

To overcome this issue, we propose to combine value bucketing (to reduce number of possible outcomes per attribute) with computing L1D only for small subsets of attributes. So, instead of analyzing the full multi-variate distribution across the set of all (mixed-type) attributes, we restrict to *n*-way cross-tabulations where every attribute is binned into a maximum of *m* buckets. For numeric and date/time attributes, the boundaries for the buckets are chosen to match the quantiles (for *m=10*, these are the deciles), which ensures similar-sized buckets. For categorical attributes, one would pick *m-1* most common values, and group the remaining into a common *Other* bucket. As single condensed measures, we propose then to report on the average value across all calculated TVDs.

Example. For a dataset with *r=100* attributes (e.g., age, gender, country, employment, etc.), *m=10* and *n=2,* there are *100^2 / 2 = 5'000* combinations of

[1] Statistical distance. https://short.wu.ac.at/statisticaldistance
[2] Total variation distance of probability measures. https://short.wu.ac.at/tvd
[3] L1-Distance. https://en.wikipedia.org/wiki/Taxicab_geometry
[4] Hellinger distance. https://en.wikipedia.org/wiki/Hellinger_distance
[5] Curse of dimensionality. https://short.wu.ac.at/cd

2-way interactions (e.g., age x gender, age x country, etc.), with each combination having a maximum of $m^n = 10^2 = 100$ of possible outcomes (e.g., 18y ≤ age < 24y & country = Austria, etc.) For each outcome, the absolute difference between the relative frequency of the original data and the relative frequency of the synthetic data is calculated. The sum thereof is the L1D for that particular 2-way combination. The final reported accuracy measure is then the average across all 2-way combinations. The same approach can be extended to 3-way, 4-way and higher interactions, whereas one needs to resort to a random subset of variable combinations, as the computational effort would otherwise grow exponentially and become infeasible.

For sequential data, it is key to retain the correlation and interdependencies not only for attributes of the same time period, but also for attributes across different time periods. To incorporate this into an accuracy measure, we propose to reshape the event data to a single flat table, by picking two random subsequent events for each subject. That single table can then be merged with all subject-level attributes and be used for the above described computation.

3 Quantifiable targets for privacy

A naïve requirement towards the synthetic data might be that the synthetic data must not consist of any records that are identical to actual records. However, this is neither a sufficient nor a necessary condition to protect individual's privacy. Let us look at the example of a data set consisting of only two attributes: year of birth and gender. Just as we expect multiple records of the original data to be identical in their values (e.g., male person born in 1978), we would also expect the synthetic data to have records identical to actual ones. On the other hand, if we consider a data set with dozens or more attributes, the probability of having two identical records becomes typically astronomically small, so that we would never expect a synthetic record to match exactly an actual record. However, even if we exclude identical matches, a match in all but one attribute can already leak enough of individual's information to be considered problematic.

Given that we have defined a distance measure for individual records (e.g., by bucketing each attribute into a maximum of *m* buckets, and counting matching attributes), we can determine for each synthetic record the closest actual record, and thus calculate the so-called distance to closest record (**DCR**).

The key question that remains is: what are the expected values for DCR, i.e., what are the expected distances between the synthetic data records and the original records, so that they are not considered to be "too close"? A readily available reference point can be calculated by splitting the original data into a training and a holdout data set, and computing the DCR for each holdout record with respect to the training data. The distribution thereof serves as a lower boundary, i.e., the DCRs of the synthetic data must not be systematically any smaller than the DCRs of a holdout data.

While this already provides a test for "being too close", it is possible to construct scenarios where the test passes while privacy is still exposed. One such scenario would be to add an average level noise to outliers, i.e. to actual records that exhibit a bigger distance to any neighbor. While for a typical record the level of obfuscation might be sufficient, it is not enough to obscure the identity of an outlier. A stronger test that we therefore propose is to assess the distribution of nearest-neighbor distance ratios (**NNDR**), which is defined as the ratio between the nearest and the second nearest neighbor. One can think of it as a DCR, that is normalized by the local density. Again, we would require to see that the synthetic data is no closer to the original data with respect to the NNDR distribution, than an actual holdout data set would be.

4 Conclusion

Based on the provided requirements by the industry partners, we proposed measures for accuracy and privacy, each with underlying concepts that are easy to communicate as well as easy to reason about. The empirical case studies on actual data are expected to then yield further insights in terms of their applicability and usefulness in practice.