Measuring the Effectiveness of Anonymised Data

Tanja Šarčević and Rudolf Mayer (SBA Research)

Anonymising data has become increasingly important due to the legal constraints imposed by authorities such as the EU's GDPR and for ethical reasons relating to privacy. One large drawback of anonymised data is its reduced quality (utility). Therefore it is crucial to quantify and minimise the utility loss prior to data sharing. We take a closer look at the question of how well this utility loss can be estimated for a specific task, in terms of effectiveness and efficiency of the resulting dataset. Our evaluation shows that the most valuable utility metrics are also the most expensive to measure, and thus often, a suboptimal solution must be chosen.

With the rise of data-intensive computing applications, data is collected and used across different domains, such as healthcare, biomedicine, or for commercial purposes. One of the most valuable types of data in all these domains is personal data, which often comes in the form of 'micro-data', where each individual is represented with their own data record. However, the privacy of individuals in micro-data can be compromised even if direct personally identifiable information is removed (de-identification). The Netflix Prize from 2007[L1] is a famous example of how customer privacy can be threatened even if data without direct identifiers are shared, by linking based on other remaining attributes. Distributing personal data is highly regulated by law, especially within the European Union

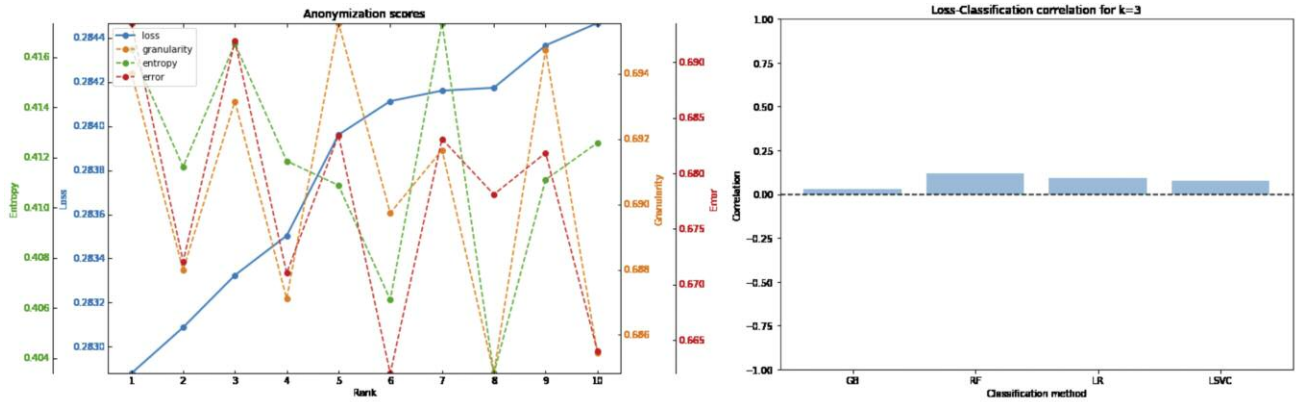

Figure 1: Utility results using different metrics, namely: loss, granularity, entropy and error (left) and the correlation between utility metric loss and machine learning performance metric F1 score (right). The comparison presented is among the anonymised datasets satisfying 3-anonymity.

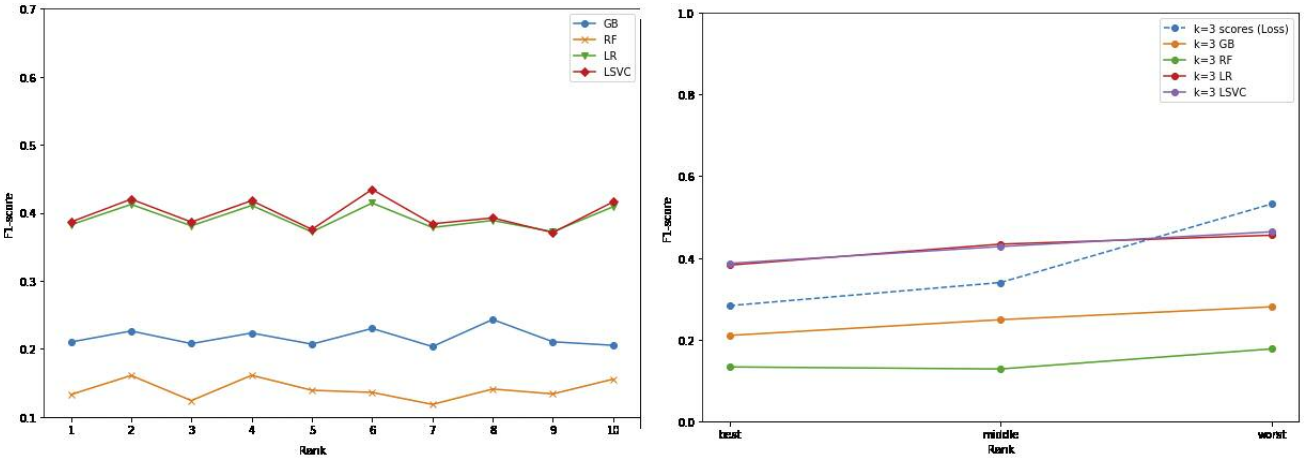

Figure 2: Utility in the notion of machine learning performance (F1 score) of optimal and suboptimal solutions satisfying 3-anonymity of k-anonymity algorithm (ARX [L2]) based on information loss metric on Adult Census Income dataset [L3]. Four classifiers are compared: gradient boosting (GB), random forest (RF), logistic regression (LR) and linear support vector classifier (LSVC). The difference in utility is shown for the top 10 k-anonymous solutions (left) and 3 solutions from different parts of the optimality spectrum: the best, middle and the worst solution (right).

with the General Data Protection Regulation (GDPR). For many purposes, datasets must be therefore anonymised before distribution.

K-anonymity is a privacy model that can be applied to sensitive datasets by obfuscating information that can be utilised to re-identify individual records in a dataset from which direct identifiers have been removed [1]. K-anonymity has certain privacy weaknesses, for which extensions have been proposed, such as l-diversity and t-closeness and other privacy models, such as differential privacy and synthetic data generation. However, k-anonymity as a model that facilitates easy data sharing is still considered in several settings.

In addition to privacy, another aspect to consider for datasets that have been sanitised is the utility of the resulting data. While anonymisation techniques provide a GDPR-compliant anonymity for the individuals in a dataset, they at the same time affect the utility of the data. This is because when sanitising a dataset via anonymisation or other approaches, some information at the level of individual records is invariably altered or removed.

Data utility can be evaluated by several approaches. One is to utilise quantitative measures of information loss [2]. Another is to measure the effectiveness of the final statistical analysis to be carried out on the data, such as the accuracy of a predictive machine learning model, compared to an analysis that would have been using the original, unabridged data. The latter is a very task-specific approach and is less efficient, as it is generally more resource-consuming (time, computing power) than the quantitative measures on the data itself. However, in many settings it provides a more useful insight into the utility of the data, given that such tasks are often carried out on the data. Without an exact knowledge of the final task, and with limited resources, it is therefore crucial to understand to what extent information loss can be used as a proxy measure for the other. In our analysis we estimated this in an experimental evaluation [3]. We utilised different machine learning models on different classification tasks and benchmark datasets and investigated how the performance of these classifiers correlate to the other utility loss metrics. The analysis has shown little correlation between the two types of utility evaluation, as shown in Figure 1, leading us to the conclusion that the estimation of the performance on a specific task cannot be replaced by more generic and faster utility metrics.

Another aspect of data utility is that there is generally not only one solution for achieving a sanitised version of a dataset that fulfils the desired level of privacy. Often a large number of candidate solutions exists, and finding the optimal solution is generally solved via heuristic approaches where implicitly one utility metric is used for finding an optimal solution. Our analysis showed that there is actually very little difference between optimal and suboptimal solutions (Figure 2), even between the optimal and worst solutions. In addition, depending on which utility metric is used in the heuristics, the optimal solution will also differ. This entails that the utility of resulting anonymised datasets are rather stable and not influenced by potentially minute aspects in the heuristic. This suggests that the data owner has a large solution space when deciding on anonymised data release. Relying on one, subjectively most appropriate utility metric will therefore not necessarily mean that the utility will be compromised based on other metrics.

The analysis showed that there is a large variety of estimates of the utility of an anonymised dataset, and no single anonymised version of a dataset that will score best across all investigated measures. Many possibly good solutions exist, assuming that the predefined level of privacy is achieved for all of them. Therefore, the choice of utility metric heavily depends on the actual use case for the data. The performance of a machine learning task is an example of such a specialised utility metric and can be used in scenarios when the usage of data can be foreseen. Using a variety of metrics can be advantageous for estimating the utility in the more general scenarios, but also needs to be put in relation to the cost of estimating these utility scores.

This work was partially funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078 (project "FeatureCloud").

Links:

[L1] https://kwz.me/h6Z
[L2] https://arx.deidentifier.org/
[L3] https://kwz.me/h0C

References:

[1] L. Sweeney: "k-anonymity: a model for protecting privacy", Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, https://doi.org/10.1142/S0218488502001648

[2] J. Eicher, K.A. Kuhn, F. Prasser: "An experimental comparison of quality models for health data de-identification", MEDINFO 2017: Precision Healthcare through Informatics, https://doi.org/10.3233/978-1-61499-830-3-704

[3] T. Šarčević, R. Mayer, D. Molnar: "An analysis of different notions of effectiveness in k-anonymity", Privacy in Statistical Databases (PSD) 2020 Proc., Tarragona, Spain. https://doi.org/10.1007/978-3-030-57521-2_9

Please contact:

Tanja Šarčević, Rudolf Mayer
SBA Research, Austria
tsarcevic@sba-research.org,
rmayer@sba-research.org