Synthetic Data for Anonymization in Secure Data Spaces for Federated Learning

Cecilio ANGULO a,1 Cristóbal RAYA a

a Universitat Politècnica de Catalunya. IDEAI-UPC, Spain

Abstract. Federated learning implies the integration of shared data. Privacy-enforcing platforms should be implemented to provide a secure environment for federated learning. We are proposing the integration of real world data from local data lakes and the generation and use of general synthetic data to simplify, eventually avoid, encryption or differential learning and use general architectures for data spaces.

Keywords. federated learning, synthetic data, data spaces

1. Introduction

Despite a large number of rich datasets are gathered across Europe that would be invaluable in the creation of knowledge through novel AI tools, the obstacles for shared use of these data hubs for data-driven innovation are insurmountable [1,2].

Federated Learning (FL) has recently emerged as a disruptive privacy technology allowing data use for model learning or data visualization, without exporting the data across from the data owner’s hub [3,4]. This approach is being adopted as a key potential solution to the shared data use challenge. However, the FL approach is not mainly based on privacy protection, hence weakness issues has been pointed out for the general FL approach. Potential vulnerabilities include susceptibility to the man-in-the-middle attack and inference attacks aiming to re-identify data subjects [5].

Enhanced privacy methods being developed for federated learning includes processes such as Homomorphic Encryption (HE) [6], Differential Privacy (DP) [7] and Trusted Execution Environments (TEE) [8]. In particular, the latter one relates with the new directive about secure data spaces being promoted from the European Union (Data Governance Act).

In general, the term “data space” refers to a type of data relationship between trusted partners, each of whom apply the same high standards and rules to the storage and sharing of their data. Data are not stored centrally but at source and are therefore only shared (via semantic interoperability) when necessary. In this context, EU promoted initiatives...
as Gaia-X look for a trusted execution environment in the form of a secure data space where data is held exclusively be the members of the Association [9].

Data space’s participants can be data providers, users and intermediaries. Data sovereignty and trust are essential for the working of secure data spaces and the relationships between participants. In this sense, references architectures like IDSA, from the International Data Spaces Association (IDSA), a founding member of the Gaia-X AISBL, has been proposed.

Unfortunately, these resource-intensive privacy enhancement methods, in the form of local software processes, developed hardware architecture or safe communication, severely limit the scalability of federated learning in the form of secure data spaces. Moreover, beyond privacy concerns or resources limitations, some critical data solutions imply storing and processing personal data to infer knowledge and to know about user experience, leading to legal consequences claiming.

Hence, anonymization arises as a keypoint in personal data manipulation [10,11], a tool to mitigate risks when gathering and massively processing sensitive data. This process allows identifying and shadowing sensitive information contained in documents, allowing its disclosure, hence avoiding to violate data protection rights of people and organisations that can be referenced in them [12].

Data anonymization, in other side, as a method protecting sensitive information or the identity of the data owner, due to legal or ethical issues, is usually seen as a major problem in data analytic because it could lead to reduce so much the knowledge contained into the dataset.

2. Proposal

It is worth noting that obtaining data has a high cost in so different domains and, many times, information is very limited. Therefore, many research projects have worked on developing reliable methods for data augmentation with synthetic instances. Moreover, knowledge extraction implies an experience acquired through learning, detecting patterns, looking for behaviours, assessing risks, until reaching a diagnosis and being able to propose a solution indicated for each situation. However, often, experts are unable to fully consider the large amount of data obtained from several institutions or companies and use it to make decisions. By considering the total set of data hubs, even for a focused problem, and generating synthetic experiences from a seedbed emerged from real-world data, professionals can benefit from this valuable information better than buried within huge amounts of data.

New training generative procedures, such as generative adversarial networks (GANs), aim at learning representations that preserve the relevant part of the information (about regular labels) while dismissing information about the private labels which correspond to the identity of a person. The success of this approach has been demonstrated in [13], for instance. As a result of the GAN-based anonymization phase, a seedbed is obtained from the training data that allows not only to capture information from the original data avoiding privacy concerns, but to generate new synthetic information with a similar behaviour to the original one. This result is currently being applied in generative applications on speech [14], vision [15], natural language [16] or in the health domain [17], where data are scarce and missing values are everywhere.
Our proposal advocates for the use of synthetic data for anonymization in the framework of secure data spaces with the aim of produce federated learning.

Let’s us suppose $N$ data hubs are storing information $\mathcal{X}$ in the same domain. Currently, since data cannot be shared, models are being developed in local, $f_i = f(\mathcal{X}_i)$, for $i = 1, \ldots, N$. In order to share information and federate learning, models are exported to other data hubs, evaluated on local data there, $f_i(\mathcal{X}_j)$ and fine tuned $\bar{f}_j$, under the hypothesis that $\bar{f}_j \sim f_i(\mathcal{X}_i \cup \mathcal{X}_j)$. However, it usually does not work. The claim for secure data spaces is that you can effectively share this information because the connection is safe, secure, and reliable. Hence, under the assumption that nobody will access your data, because either, you are moving models, not data, or data is moved in secure form, federated learning iterates so that you obtain $f_{\{1,\ldots,N\}}$ from, for instance, $f_1, \ldots, f_N$. In fact, there exist several approaches, but all of them have relaxed trust conditions.

Our approach is defending that you can share data, but not the original one, but one that is synthetically following the same statistical properties, that is $\hat{\mathcal{X}} \sim \mathcal{X}$ because $P(\mathcal{X}) = P(\hat{\mathcal{X}})$ with $P$ indicating statistical properties. In this form, our hypothesis is that you can obtain $f_{\{1,\ldots,N\}} = f(\bigcup \mathcal{X}_j)$ by federating $f_i(\mathcal{X}_i \cup \bigcup_{j \neq i} \mathcal{X}_j)$ for $i = 1, \ldots, N$. In fact, as far as our approach is considering statistical properties, it can be noted from this domain in the form $P(y|\bigcup \mathcal{X}_j) \sim \Pi_i P(y|\mathcal{X}_i \cup \bigcup_{j \neq i} \mathcal{X}_j)$. This hypothesis is opening a new research line that we are starting to explore avoiding main privacy concerns.

References


