Data Privacy Preservation in Database Metadata

Danning Zhan [1], Rihan Hai [1]

[1] Web Information Systems, TU Delft, Netherlands
{first.author,second.author}@tudelft.nl

In recent years there have been significant advancements in machine learning and its application in various areas, such as healthcare and finance. Alongside technical advancements in data processing, regulations such as GDPR [1] have been introduced. Data privacy is essential due to the increased digital presence in our daily lives, where data leakages can have detrimental effects. More importantly, securing data access is essential when the data owners want to exclude other parties from looking over their private sensitive data. However, knowing data from different sources could assist us in making better decisions through machine learning. The raw data is described using metadata such as the attribute names, the domain of the attributes, and relaxed functional dependencies (RFDs) [2]. So, knowing the metadata, allows us to make decisions regarding the raw data, such as identifying whether specific attributes should be selected [3] for machine learning.

Some metadata are already being leveraged in machine learning, such as data range and datatype [4]. The data ranges and the number of attributes can allow us to know the vector space with the suitable dimensions from which the raw data lies. The metadata that we will investigate are RFDs. The RFDs can tell us about more well-defined relationships between attributes within the data. As these relationships will indicate some implicit structure within the data. With this implicit structure, we can possibly generate data that is more similar to the source data. We will investigate if RFDs will allow us to generate better data than utilizing only the range and type. We will show the amount of privacy leakage with respect to different types of RFDs with a formal definition of privacy through set theory.

References

[1] General Data Protection Regulation (GDPR) – Official Legal Text — gdpr-info.eu. https://gdpr-info.eu/. [Accessed 01-11-2023].

[2] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. Relaxed functional dependencies—a survey of approaches. *IEEE Transactions on Knowledge and Data Engineering*, 28(1), 2016.

[3] Rihan Hai, Christos Koutras, Andra Ionescu, Ziyu Li, Wenbo Sun, Jessie van Schijndel, Yan Kang, and Asterios Katsifodimos. Amalur: Data integration meets machine learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023.

[4] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021.