Financiado por la Unión Europea
NextGenerationEU

MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL

Plan de Recuperación, Transformación y Resiliencia

uc3m

6G-RIEMANN-DS Entregable E8

Definition of the usecases for privacy preserving data sharing

PROGRAMA DE UNIVERSALIZACIÓN DE INFRAESTRUCTURAS DIGITALES PARA LA COHESIÓN
UNICO I+D 5G 2021

Fecha: 31/7/2022

Versión: 1.0

Propiedades del documento

Id del documento	E8
Título	Definition of the usecases for privacy preserving data sharing
Responsable	UC3M
Editor	Albert Banchs

Equipo editorial	**Partner**	**Name**	**Surname**	**Sections**
	UC3M	Marco	Gramaglia	All
	UC3M	Mauro	Allegretta	All

Nivel de diseminación	Público
Estado del documento	Final
Versión	1.0

Historial

Revisión	**Fecha**	**Por**	**Descripción**
1.0	31/07/22	Editor	Final version

Revisor

Equipo revisor	**Partner**	**Name**	**Surname**	**Sections**
	UC3M	Albert	Banchs	All

Descargo de responsabilidad

This document has been produced in the context of the 6G-RIEMANN Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme. The information contained in this document is provided "as is" without any express or implied warranties, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. The document writer shall not be liable for any damages, whether direct or indirect, arising out of or in connection with the use of this information. The user of this document assumes all risks and liabilities associated with its use and shall indemnify and hold harmless the document writer from any and all claims, losses, damages, or expenses, including attorney's fees, arising from the use of this information.

Table of Contents

Lista de acrónimos

CTI: Cyber Threat Intelligence

PPDP: Privacy-Preserving Data Publishing

MIAs: Membership Inference Attacks

DRAs: Data Reconstruction Attacks

ML: Machine Learning

AI: Artificial Intelligence

NLP: Natural Language Processing

CV: Computer Vision

5G: 5th Generation (of wireless cellular networks)

Resumen ejecutivo

La privacidad en los modelos de aprendizaje automático (ML), tanto en el entrenamiento como en el tiempo de inferencia, se ha convertido en un tema importante en la comunidad de investigación de inteligencia artificial (AI), como lo demuestra la tendencia cada vez mayor de publicaciones relacionadas con la privacidad en ML. A lo largo de los años, se han descubierto y estudiado una gran variedad de posibles ataques contra un modelo de ML entrenado, como los ataques de inferencia de membresía, inversión de modelo, robo de modelo, inferencia de atributos, inferencia de propiedades y ataques de reconstrucción de datos.

En este trabajo, nos enfocamos en los métodos diseñados para proteger los datos crudos. Esta familia de técnicas permite el intercambio de datos de entrenamiento entre diferentes actores sin los riesgos de privacidad asociados, es decir, la publicación de datos que preservan la privacidad (PPDP). Por lo tanto, estas tecnologías abren la posibilidad de nuevas aplicaciones que antes solo eran posibles entre partes confiables en escenarios de aprendizaje federado (FL) complejos.

En el caso de uso de la puntuación crediticia, la protección de la privacidad de los datos personales del cliente es una consideración importante para las empresas que ofrecen servicios financieros. Las técnicas de PPDP permiten a estas empresas compartir datos de entrenamiento para la evaluación de la puntuación crediticia de los clientes de manera segura y privada.

En el caso de uso del intercambio de datos de inteligencia de amenazas cibernéticas (CTI), las organizaciones de seguridad pueden compartir información crítica de amenazas entre ellas de manera segura y privada mediante técnicas de PPDP. Esto permite una mejor colaboración entre las organizaciones en la lucha contra las amenazas cibernéticas y una respuesta más rápida y efectiva ante los incidentes de seguridad.

En resumen, las técnicas de PPDP son una herramienta importante para garantizar la privacidad y la seguridad de los datos de entrenamiento en escenarios de FL y compartir información crítica en diferentes campos de aplicación, como la evaluación de la puntuación crediticia y el intercambio de CTI. La investigación y el desarrollo continuos de técnicas de PPDP son cruciales para abordar los desafíos de privacidad y seguridad en el aprendizaje automático y la inteligencia artificial.

Abstract

The fields of machine learning and cybersecurity are becoming increasingly intertwined, as the use of machine learning in cybersecurity applications becomes more prevalent. However, with the use of machine learning comes the risk of privacy breaches and malicious attacks on the models. Thus, ensuring privacy and security in machine learning models has become a major topic in the AI research community.

One way to protect the privacy of raw data in machine learning models is through Privacy Preserving Data Publishing (PPDP), which allows the sharing of training data among different actors without the associated privacy risks. This technology opens up the possibility of novel applications that were previously only possible among trusted parties in complex federated learning scenarios. One such use case is in credit scoring, where PPDP can be used to transform raw financial data while maintaining the accuracy of the machine learning model, and without compromising the privacy of the individual's financial data.

On the other hand, cybersecurity relies heavily on the sharing of information to prevent and mitigate cyber threats. The exchange of Cyber Threat Intelligence (CTI) data has become an important practice in the cybersecurity community. However, sharing CTI data also poses significant privacy and security risks. Techniques such as PPDP can be applied to CTI data to protect the privacy of the data while allowing for effective sharing and analysis of cyber threats.

Overall, the intersection of machine learning and cybersecurity has created a need for novel privacy-preserving technologies to ensure the security and privacy of sensitive data. The use of PPDP is a promising technique for protecting sensitive data while enabling the sharing and analysis of information in these critical fields.

1. Introduction

In recent years, there has been a significant increase in the number of AI applications being used in real-world environments. This is primarily due to remarkable advances made in various areas such as natural language processing, computer vision, and machine learning. These AI-based applications are being used to personalize services, offer improved healthcare to end-users, and automate network management in the new 5G architectures, among others.

However, these applications rely on input data from potentially diverse sources and platforms, which may not be fully trusted. This raises various privacy and confidentiality concerns. For instance, sensitive data may be leaked or used for unintended purposes, such as targeted advertising, which could lead to discrimination against specific groups of people. Additionally, if the data used to train these AI-based applications is not diverse enough, it could result in biased models that do not accurately represent the broader population.

To address these concerns, various solutions have been proposed. One such solution is k-anonymity[1], which ensures that individual data points cannot be identified by masking identifying features such as names or addresses. Another solution is l-diversity[2], which aims to preserve the privacy of individuals by ensuring that sensitive data is not overly concentrated in any one group. Yet another solution is t-closeness[3], which ensures that the distribution of sensitive data in a dataset is similar to the distribution in the entire population. While these solutions have been effective in preserving privacy, they are designed to keep the dataset in a human-readable format, without considering the implications of modifications on downstream machine learning tasks.

Other solutions, such as differential privacy[4], are designed to ensure the privacy of machine learning tasks mathematically, with strong theoretical guarantees. However, implementing these solutions can be challenging in practice, especially when sharing the entire dataset is necessary to complete a task. Furthermore, differential privacy may not apply when the data is generated by machines. For example, if different factories want to share data from their sensors to jointly train a predictive maintenance model, the aforementioned solutions may disclose the dataset as it is, revealing confidential business data.

In this document, we discuss the possible use cases for data transformation methods that maintain the ability of a machine to learn from the data while also preserving privacy. By transforming the data rather than keeping it in its original form, we can share it for specific machine learning tasks without compromising privacy and confidentiality.

A fundamental aspect that shall be taken into account is the privacy vs. accuracy trade-off that different data transformation methods yields.

2. A taxonomy of attacks

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International journal of uncertainty, fuzziness and knowledge-based systems 10.05 (2002): 557-570.

[2] Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 3-es.

[3] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." 2007 IEEE 23rd international conference on data engineering. IEEE, 2006.

[4] Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3. Springer Berlin Heidelberg, 2006.

Privacy has become a crucial concern in the field of machine learning, both during training and inference. This is evidenced by the growing number of publications in this area. Numerous potential attacks on machine learning models have been identified and studied over the years (Figure 2), including membership inference attacks, model inversion attacks, model stealing attacks, attribute inference attacks, property inference attacks, and data reconstruction attacks.

Membership Inference Attacks[5] (MIAs) attempt to determine if a specific data point was part of the training set used to develop a machine learning model. This can be a serious breach of privacy, as it reveals information about the data used to train the model and potentially sensitive information about the individual associated with that data.

Model Inversion[6] attacks aim to reverse-engineer a machine learning model. Even if the only attack surface is the predicted class output of the model, model inversion can still be dangerous, for example, if reconstructing class feature or class representatives is sensitive information (e.g., identity recognition, Figure 1).

Figure 1: Popular model inversion attack example from [CIT]. Original face image (left) and restored one (right).

Model Stealing[7] attacks allow an adversary to either copy a deployed machine learning model or steal its exact architecture and parameters. This can be used to create a "shadow model" that can be manipulated to achieve the attacker's objectives.

Attribute Inference[8] attacks exploit machine learning models' output to infer sensitive attributes of the training set that were not the main prediction goal of the model. This can lead to the disclosure of personal information, such as medical conditions, sexual orientation, or political beliefs.

Property Inference[9] attacks, on the other hand, try to infer general statistics, patterns, and properties (considered sensitive) of the training data from the machine learning model's output. This can be used to deduce information about the data set, such as demographics, without actually revealing the data itself.

[5] Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE symposium on security and privacy (SP). IEEE, 2017.

[6] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015.

[7] Tramèr, Florian, et al. "Stealing Machine Learning Models via Prediction APIs." USENIX security symposium. Vol. 16. 2016.

[8] Yeom, Samuel, et al. "Privacy risk in machine learning: Analyzing the connection to overfitting." 2018 IEEE 31st computer security foundations symposium (CSF). IEEE, 2018.

[9] Melis, Luca, et al. "Exploiting unintended feature leakage in collaborative learning." 2019 IEEE symposium on security and privacy (SP). IEEE, 2019.

Data Reconstruction Attacks[10] [11] (DRAs) attempt to reconstruct the original input data from the output of the machine learning model. This can be a significant privacy concern, as it could reveal sensitive information about the individual associated with that data.

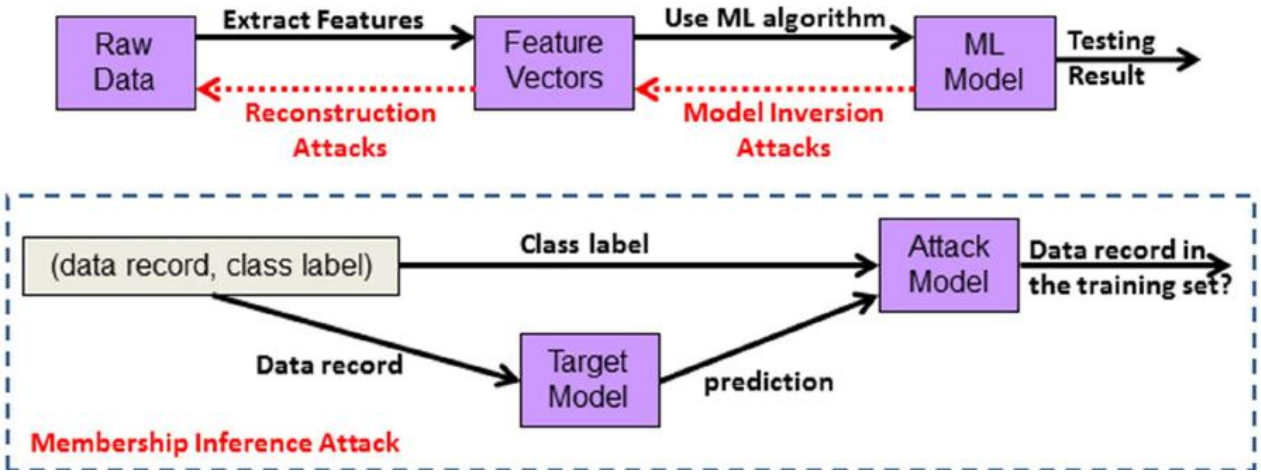

Figure 2: Some of the most popular ML privacy-related attacks and their attack surface in the typical ML pipeline

To mitigate these privacy concerns, various privacy-preserving machine learning techniques have been developed, including differential privacy, secure multi-party computation, homomorphic encryption, and federated learning. These techniques aim to enable the training of accurate machine learning models while protecting the privacy of the training data and ensuring that the model cannot be used to reveal sensitive information about the individuals associated with that data.

Privacy-preserving data publishing[12] (PPDP) techniques are designed to protect the privacy of raw data while allowing the sharing of data among different parties. These techniques are particularly relevant in complex federated learning scenarios, where multiple parties with sensitive data collaborate to train a machine learning model.

PPDP techniques typically involve applying a transformation to the raw data before sharing it with other parties. The transformation is carefully designed to protect the privacy of the data, while preserving its utility for machine learning purposes.

One widely used PPDP technique is differential privacy, which adds random noise to the data to prevent re-identification attacks. Another popular technique is homomorphic encryption, which allows computation on encrypted data without requiring decryption. This allows parties to perform machine learning operations on the data without having to reveal the original data to each other.

In the optimal scenario, the transformed data should preserve the accuracy of the ML model that is trained on it, compared to the accuracy obtained using the raw data. Additionally, the transformed data should not

[10] Al-Rubaie, Mohammad, and J. Morris Chang. "Reconstruction attacks against mobile-based continuous authentication systems in the cloud." IEEE Transactions on Information Forensics and Security 11.12 (2016): 2648-2663.

[11] Feng, Jianjiang, and Anil K. Jain. "Fingerprint reconstruction: from minutiae to phase." IEEE transactions on pattern analysis and machine intelligence 33.2 (2010): 209-223.

12 Fung, Benjamin CM, et al. "Privacy-preserving data publishing: A survey of recent developments." ACM Computing Surveys (Csur) 42.4 (2010): 1-53.

allow the reconstruction of the raw data, meaning that even if an attacker gains access to the transformed data, they should not be able to infer any exact information about the original data.

Overall, PPDP techniques are an important tool for enabling collaboration on sensitive data, while protecting the privacy of the individuals whose data is being used. As the demand for machine learning models that respect user privacy continues to grow, it is likely that we will see continued development and refinement of these techniques in the coming years.

3. Use cases

In the following, we discuss two potential use cases related to this topic.

3.1. Credit scoring

Credit scoring is a fundamental process in the financial industry that determines a borrower's creditworthiness based on their financial history, including their credit card usage, payment history, and employment status. Credit scoring models are developed using machine learning algorithms that are trained on sensitive data such as credit reports, payment histories, and employment records. However, this data contains personally identifiable information (PII) that can lead to privacy risks if it falls into the wrong hands.

To mitigate these risks, privacy-preserving data publishing techniques can be used to protect the raw data while still allowing the development of accurate credit scoring models. For instance, a financial institution could use differential privacy techniques to add noise to the raw data to protect individual records' privacy. Another approach is homomorphic encryption, which allows computation on encrypted data, making it possible to train models without revealing any sensitive information.

In this scenario, the financial institution applies a transformation to the raw data, and the transformed data is then shared with other actors, such as data scientists, who may use it to train a credit scoring model. The transformed data may also be combined with other transformed data to increase the sample size, which can improve the accuracy of the model.

In the optimal scenario, the accuracy of the credit scoring model remains similar to the accuracy obtained using the raw data, and the transformed data does not allow the reconstruction of the raw data. Moreover, the recipient of the transformed data should not be able to infer any exact information about the original data.

Overall, privacy-preserving data publishing techniques enable financial institutions to develop accurate credit scoring models while ensuring the privacy and security of sensitive data. This has the potential to increase access to credit for borrowers who might otherwise be excluded due to privacy concerns, while still maintaining the privacy of their sensitive financial data.

3.2. Cybersecurity

Cybersecurity threats continue to increase in frequency and complexity, and the need for timely and accurate threat intelligence has become more important than ever. Cyber Threat Intelligence (CTI) is the information that is collected, analyzed, and distributed to help organizations identify and respond to cyber threats. However, sharing CTI data among organizations can be challenging due to privacy concerns and the potential risks associated with sharing sensitive information. To address these challenges, Privacy Preserving Data Publishing (PPDP) techniques can be used to protect the confidentiality and privacy of CTI data. In this use

case, we explore how PPDP techniques can be applied to exchange CTI data between organizations while maintaining data privacy and confidentiality.

Overall, while DNS analysis can be a powerful tool for detecting phishing and other cyber attacks, it must be performed in a responsible and privacy-preserving manner to ensure that individuals and organizations are not put at risk.

Imagine a group of companies in the same industry, such as financial services, who are interested in sharing cyber threat intelligence to improve their overall security posture. Traditionally, these companies might be hesitant to share their sensitive data, such as details about their internal network architecture and security practices, due to concerns about privacy and intellectual property protection.

To address this issue, the companies could use privacy-preserving data publishing techniques to transform their raw CTI data before sharing it with other members of the group. This transformation could involve techniques such as differential privacy, homomorphic encryption, or secure multi-party computation.

For example, the companies could use Secure Multi-party Computation[13] [14] (SMC) to jointly compute statistics on their CTI data, such as the number of attacks they have seen in a particular time period, without revealing any sensitive information about their specific security incidents. This would allow the companies to gain valuable insights into the overall threat landscape without revealing any confidential details about their own internal security posture.

Alternatively, the companies could use Homomorphic Encryption[15] [16] to encrypt their CTI data before sharing it, allowing other members of the group to perform computations on the data without being able to see the underlying raw data itself. This would enable the companies to collaborate on threat analysis and incident response without risking the exposure of their sensitive data.

Overall, the use of privacy-preserving data publishing techniques would enable these companies to share CTI data in a secure and privacy-preserving manner, reducing their overall cybersecurity risk while improving their collective ability to detect and respond to threats.

4. Conclusion

In conclusion, privacy and security are crucial aspects of any system involving sensitive data, and machine learning is no exception. The research community has been actively developing new techniques and methods to address the privacy and security concerns of machine learning models, both at training and inference time. In particular, the focus has been on protecting raw data and ensuring that the sharing of such data does not pose any privacy risks. This has led to the development of technologies such as differential privacy, federated learning, and homomorphic encryption, which offer a range of benefits for various use cases.

[13] Mohassel, Payman, and Yupeng Zhang. "Secureml: A system for scalable privacy-preserving machine learning." 2017 IEEE symposium on security and privacy (SP). IEEE, 2017.

[14] Liu, Jian, et al. "Oblivious neural network predictions via minionn transformations." Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. 2017.

[15] Gilad-Bachrach, Ran, et al. "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy." International conference on machine learning. PMLR, 2016.

[16] Gentry, Craig. "Fully homomorphic encryption using ideal lattices." Proceedings of the forty-first annual ACM symposium on Theory of computing. 2009.

As machine learning continues to evolve and become more ubiquitous, it is clear that privacy and security will remain important topics for the research community to tackle. It will be essential to ensure that machine learning models are built in a way that protects sensitive data and users' privacy while maintaining high levels of accuracy and functionality. With the continued development of new technologies and techniques, we can look forward to a future where machine learning is even more powerful and trustworthy, with privacy and security built into the very fabric of the models.

In addition to the points previously mentioned, it is worth noting that as machine learning continues to be integrated into various industries and sectors, the importance of privacy and security in these models will only increase. Privacy concerns can arise not only from the data itself but also from the models' outputs and how they are used. Therefore, it is crucial to ensure that machine learning models are designed with privacy and security in mind from the outset. This can involve the use of privacy-preserving techniques, careful data handling practices, and regular security audits and updates.

Moreover, as the amount of data being exchanged between organizations and individuals continues to grow, the need for secure and efficient methods of data exchange will become increasingly important. Whether it is for credit scoring or CTI data sharing, privacy-preserving technologies provide an avenue for data exchange while minimizing the privacy risks associated with data sharing. The development and adoption of these technologies will not only provide benefits in terms of privacy and security but also enable new opportunities for collaboration and innovation.

In conclusion, privacy-preserving technologies are becoming increasingly important in the fields of machine learning and data exchange. As we continue to collect and analyze vast amounts of data, it is vital to ensure that we do so in a way that protects the privacy and security of individuals and organizations. By adopting privacy-preserving techniques and regularly updating and auditing our systems, we can create a safer and more secure data ecosystem that benefits everyone involved.