Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230957

Machine learning for privacy-preserving: Approaches, challenges and discussion

Ziqi Pan

Department of Communication Engineering, Donghua University, Shanghai, 200051, China

190910107@mail.dhu.edu.cn

Abstract. Currently, advanced technologies such as big data, artificial intelligence and machine learning are undergoing rapid development. However, the emergence of cybersecurity and privacy leakage problems has resulted in serious implications. This paper discusses the current state of privacy security issues in the field of machine learning in a comprehensive manner. During machine training, training models often unconsciously extract and record private information from raw data, and in addition, third-party attackers are interested in maliciously extracting private information from raw data. This paper first provides a quick introduction to the validation criterion in privacy-preserving strategies, based on which algorithms can account for and validate the privacy leakage problem during machine learning. The paper then describes different privacy-preserving strategies based mainly on federation learning that focus on Differentially Private Federated Averaging and Privacy-Preserving Asynchronous Federated Learning Mechanism and provides an analysis and discussion of their advantages and disadvantages. By improving the original machine learning methods, such as improving the parameter values and limiting the range of features, the possibility of privacy leakage during machine learning is successfully reduced. However, the different privacy-preserving strategies are mainly limited to changing the parameters of the original model training method, which leads to limitations in the training method, such as reduced efficiency or difficulty in training under certain conditions.

Keywords: machine learning, cyber security, artificial intelligence

1. Introduction

On 3 November 2022, the European Union Agency for Cybersecurity, ENISA, released its cybersecurity threat landscape report, ENISA Threat Landscape 2022, covering records from July 2021 to July 2022 [1]. In general terms, the report notes that the geopolitical landscape in 2022 has had a significant impact on cyber security, with the volatile international situation contributing to the growth of malicious cyber activity and the challenges to the information security sector itself increasing year on year. Machine learning is a popular technology in recent years and is used in a wide range of different fields, and has had a significant impact in the field of cyber security [2, 3]. However, machine learning is vulnerable to attacks where attackers aim to replace the original training data with malicious data, resulting in incorrect training results for machine learning models [4]. Moreover, machine learning itself necessitates

DOI: 10.54254/2755-2721/18/20230957

the utilization of large amounts of data for training, a process that can potentially lead to the illegal disclosure of highly sensitive private information [5].

In recent years, there has been a proliferation of Artificial Intelligence (AI) application services for the average user. These include Novel-AI, a text-to-image generation program developed by Anlatan, ChatGPT, an AI chatbot program developed by OpenAI Artificial Intelligence Research Lab, New Bing, a new intelligent search engine integrated with ChatGPT services, and Microsoft 365 AI Copilot, a new intelligent office application, jointly launched by Microsoft and OpenAI. In practical use, the aforementioned application service presents two significant shortcomings of machine learning in the area of security. In the process of collecting data to build models, particularly Novel-AI's collection of drawing data to build models, there are web users who believe their privacy and copyright are being violated. During the dialogue phase of chat programs, there are also some users who prompt AI programs to generate responses that are contrary to ethics and law, including through malicious grooming, thus forcing application service providers to add restrictions on periodic data erasure for dialogue programs.

In previous studies, Carlini et al. attempted to quantitatively assess the risk of generative sequence models unconsciously remembering training data sequences and described a specific test method [6]; Lyu et al. attempted to protect privacy extracted from text and proposed a new method, Differentially Private Neural Representation, which quantifies extracted text privacy [7]; Ramaswamy et al. utilized the Differentially Private Federated Averaging technique in an attempt to avoid memorising user data in joint learning training data [8]; Xu et al. conducted a study on privacy-preserving Machine Learning (ML) and constructively proposed a Phase, Guarantee, and Utility (PGU) triad based on the Confidentiality, Integrity and Availability (CIA) triad to guide the Privacy-Preserving Machine Learning (PPML) solution [9]; Lu et al. carried out a study for artificial intelligence of edge network and proposed Privacy-Preserving Asynchronous Federated Learning Mechanism [10]. The method not only allows more efficient joint learning of edge nodes under the premise of data confidentiality, but is also suitable for situations where edge mobile devices need to join or exit arbitrarily.

These studies look at different technical areas, including federated learning, adversarial machine learning, Machine Learning as a Service, privacy-preserving deep learning, etc., and propose a variety of novel approaches and these researches have enhanced the security technologies in these areas. However, most of the research is mainly limited to the same technology area and cannot provide a systematic and comprehensive guide to security technologies. With the rapid development of the Internet, a large number of non-specialists are gradually moving into the professional field of information security through open courses and materials on the Internet, making it more urgent for them to read a text that covers a wide range of technologies. In addition, the development of security technology should be cross-cutting, and this article covers several different branches of technology, which is conducive to different security technologies influencing and promoting each other.

This paper describes several approaches to securing privacy information in the field of machine learning in Section 2. The advantages and disadvantages of these approaches are discussed in Section 3. The important aspects of the paper are summarised in Section 4.

2. Methodology

2.1. Overview of the method

By improving different modelling methods in the field of machine learning, such as parameter values, training thresholds, and variable ranges for important features, the aim is to optimise the model and reduce the possibility of unintentional invasion of private information during the training process, or the disclosure of raw private data due to malicious attacks. Exposure metric can be defined as the reduction in the number of guesses required to guess the exact value of the random variable, and this number of guesses is known as the guess entropy. The algorithm guided by the Exposure metric demonstrates the privacy security issues of model training and also shows that it is possible to use as a test metric by separating random sequences. These random sequences, called canaries, are manually created and

DOI: 10.54254/2755-2721/18/20230957

placed into the training data as secret information for testing. These sequences, which are unrelated to the target information w for machine learning, can test the exposure of privacy information.

2.2. Federal learning mechanisms

2.2.1. Differentially private federated averaging. During the process of model training, Differential Privacy (DP), which is able to scramble the extracted representations locally before the user publishes information [11], provides privacy protection for sensitive data in the training set, but it is not considered completely strict. To enhance this privacy protection, the method of federated learning can be employed, which is characterised as a distributed learning method in that it trains models without centralising user data [12, 13]. It is able to decentralise the data, with the server only getting targeted updates, and these updates are short-lived.

The Differentially Private Federated Averaging (DP-FedAvg) technique crops each user's update to a bounded L2 parametrization and introduces the Gaussian noise to its weighted average update. The algorithm avoids tuning the hyperparameters of the neural network on sensitive private data, and instead tunes them on the public dataset. The algorithm samples the client's devices, either using Poisson sampling methods or by randomly selecting a fixed number of users in each sampling round. After sampling, the same amount of noise is added to the cropped update average of the sample. This technique uses the detection method guided by the exposure metric, which inserts canaries into the data and detects the extent to which the algorithm has unintentional memory for canaries, from which the level of privacy protection of the algorithm is assessed.

2.2.2. Privacy-preserving asynchronous federated learning mechanism. The protection of private information is of utmost importance when processing data in the cloud, as not all data can be locally processed while maintaining privacy. In this context, edge network machine learning is a promising solution [14]. Privacy-Preserving Asynchronous Federated Learning Mechanism (PAFLM) allows for more efficient federation learning between multiple edge nodes without sharing private data between them. Each node is effectively trained independently using a local dataset, and its parameters are optimised by using models learned by other nodes, ultimately building a better model without sharing private information.

PAFLM optimizes traditional federation learning by Self-Adaptive Threshold Gradient Compression, which reduces the number of communications between nodes and servers, reducing the load on the network and making the federation learning process more efficient. The nodes of PAFLM are adaptive to each round of model training, adjusting the appropriate compression threshold and using variable thresholds to successfully avoid over-compression, thus avoiding the problem of difficult processing of subsequent training models. Gradient checking is involved in each round of iteration, and the accumulated information is eventually uploaded to the parameter server.

2.3. Defending against attacks

MLaaS has been increasingly employed by Internet companies such as Google as it facilitates the evaluation and learning of data and provides a cost-effective alternative for commercial entities [15]. However, MLaaS's have the potential risk of compromising personal and business privacy due to Membership Inference Attacks [16, 17]. Machine learning models often give different predictive responses to data in the training set and to new data when first exposed to it, which leads to their output being able to reflect such differences and thus being exploited by attackers.

MIASec's approach to defending against Membership Inference Attacks improves Data Indistinguishability precisely by reducing the difference between the results of training and test data. The important feature can have a significant impact on the accuracy of the machine learning model. The importance of the features can be determined by comparing the magnitude of the cross-entropy, which is inversely related to each other. Then, the range of values for this feature is narrowed. Since the attacker needs to compare the predicted approximation of the attack model with the prediction of the victim

DOI: 10.54254/2755-2721/18/20230957

model, all when the victim model has a smaller range of features, it will be more likely to resemble the attack model, and this confusion reduces the likelihood that the attacker will be able to determine the content of the victim model.

3. Discussion of algorithms

3.1. Federal learning mechanisms

3.1.1. DP-FedAvg. DP-FedAvg has proven its effectiveness in the training of Gboard input methods, which can be applied to various NWP tasks, providing privacy protection for search engines such as Google and Bing and input methods such as Gboard, preventing users from revealing their private information during the use of these tools. However, during random sampling, the server is restricted from allowing too many devices to be connected and the devices need to meet certain availability criteria in order to prevent system overload and crashes. This restriction makes it difficult for the server to accurately characterise the randomness of the sampling process, thus ensuring that only a smaller set of devices can be sampled.

3.1.2. PAFLM. PAFLM provides privacy protection for edge network machine learning that must be uploaded to a cloud server, but Asynchronous learning is affected by many factors. For example, the nodes in an edge network are often mobile, which makes PAFLM susceptible to problems such as differences in learning progress or uneven samples. Other influencing factors need to be discussed separately with different decay functions.

3.2. Defending against attacks

MIASec was tested on the Census Income dataset, the Shopping dataset and the Location dataset and can be applied to Random Forest, Xgboost and Support Vector Machine (SVM) models. miasec reduced attack accuracy by an average of 11.7% and attack recall by 15.4%, both of which were tested maximum values can reach 18.6% and 21.8%. The weakness of this method is that changes in the range of feature values can easily lead to a decrease in the prediction accuracy of the model. The attacker's idea of attack is to compare the degree of approximation of the data, therefore, the number of output categories of the model is proportional to the degree of information leakage. When the number of output categories is high, changes in the range of feature values are more likely to lead to a drop in test accuracy, which can be as much as 20% in extreme cases.

4. Conclusion

This paper offers a thorough analysis of the current state of privacy security issues in the field of machine learning, covering both the purposeful extraction of privacy information from raw data by third-party attackers and the unintentional violation of privacy information during machine training. The study gives a description of several privacy-preserving techniques, an analysis of them, as well as a discussion of their benefits and drawbacks. It also introduces the validation criteria in privacy-preserving strategies. The likelihood of privacy leakage during the machine learning process is successfully decreased by enhancing the original machine learning approach, such as increasing the parameter values and restricting the range of features. However, the various privacy-preserving solutions are generally restricted to altering the parameters of the original model training approach, which results in restrictions in the manner in which models are taught, such as a reduction in effectiveness or difficulty in training under specific situational circumstances. In the current thinking on modifying established models, reducing privacy leakage and ensuring training efficiency are often difficult to combine, and in some cases one or the other will have to be discarded. In future research, it may be necessary to consider additional modifications to the model or building more appropriate models, rather than relying solely on parameter adjustments or range restrictions.

DOI: 10.54254/2755-2721/18/20230957

References

[1] ENISA 2022 ENISA Threat Landscape [J/OL] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

[2] Goodfellow I J Shlens J Szegedy C 2014 Explaining and harnessing adversarial examples arXiv preprint arXiv:1412.6572, 2014.

[3] Biggio B Roli F 2018 Wild patterns: Ten years after the rise of adversarial machine learning Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security pp 2154-2156

[4] Xue M Yuan C Wu H et al. 2020 Machine learning security: Threats, countermeasures, and evaluations IEEE Access 8 74720-74742

[5] Li J 2018 Cyber security meets artificial intelligence: a survey Frontiers of Information Technology & Electronic Engineering 19(12) pp 1462-1474

[6] Carlini N Liu C Erlingsson Ú et al. 2019 The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks USENIX Security Symposium 267

[7] Lyu L He X Li Y 2020 Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness arXiv preprint arXiv:2010.01285

[8] Ramaswamy S Thakkar O Mathews R et al. 2020 Training production language models without memorizing user data arXiv preprint arXiv:2009.10031

[9] Xu R Baracaldo N Joshi J 2021 Privacy-preserving machine learning: Methods, challenges and directions arXiv preprint arXiv:2108.04417

[10] Lu X Liao Y Lio P et al. 2020 Privacy-preserving asynchronous federated learning mechanism for edge network computing IEEE Access 8: pp 48970-48981

[11] Dwork C Roth A 2014 The algorithmic foundations of differential privacy Foundations and Trends® in Theoretical Computer Science 9(3–4) pp 211-407

[12] McMahan B Moore E Ramage D et al. 2017 Communication-efficient learning of deep networks from decentralized data Artificial intelligence and statistics PMLR pp 1273-1282

[13] Kairouz P McMahan H B Avent B et al. 2021 Advances and open problems in federated learning Foundations and Trends® in Machine Learning 14(1–2) pp 1-210

[14] Shi W Cao J Zhang Q et al. 2016 Edge computing: Vision and challenges IEEE internet of things journal 3(5): pp 637-646

[15] Wu Y et al. 2017 Big data and computational intelligence in networking CRC Press

[16] Shokri R Stronati M Song C et al. 2017 Membership inference attacks against machine learning models 2017 IEEE symposium on security and privacy (SP) 2017 3-18

[17] Salem A Zhang Y Humbert M et al. 2018 Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models arXiv preprint arXiv:1806.01246