Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/2024MA0053

AI-driven anonymization: Protecting personal data privacy while leveraging machine learning

Le Yang[1a,5,*,†], Miao Tian[1b,6,†], Duan Xin[2,7], Qishuo Cheng[3,8], Jiajian Zheng[4,9]

[1a]Computer Information Science, Sam Houston State University, Huntsville, TX, USA
[1b]Master of Science in Computer Science, San Fransisco Bay University, Fremont CA, USA
[2]Accounting, Sun Yat-Sen University, HongKong
[3]Department of Economics, University of Chicago, Chicago, IL, USA
[4]Bachelor of Engineering, Guangdong University of Technology, ShenZhen, CN

[5]wesleyyang96@gmail.com
[6]miao.hnlk@gmail.com
[7]duanxin12314057@gmail.com
[8]qishuoc@uchicago.edu
[9]im.jiajianzheng@gmail.com
*corresponding author
[†]These authors are co-first authors.

Abstract. AbstractThe development of artificial intelligence has significantly transformed people's lives. However, it has also posed a significant threat to privacy and security, with numerous instances of personal information being exposed online and reports of criminal attacks and theft. Consequently, the need to achieve intelligent protection of personal information through machine learning algorithms has become a paramount concern. Artificial intelligence leverages advanced algorithms and technologies to effectively encrypt and anonymize personal data, enabling valuable data analysis and utilization while safeguarding privacy. This paper focuses on personal data privacy protection and the promotion of anonymity as its core research objectives. It achieves personal data privacy protection and detection through the use of machine learning's differential privacy protection algorithm. The paper also addresses existing challenges in machine learning related to privacy and personal data protection, offers improvement suggestions, and analyzes factors impacting datasets to enable timely personal data privacy detection and protection.

Keywords: Machine learning, Differential privacy algorithm, Personal data protection, Drive anonymization.

1. Introduction

Artificial intelligence technology is constantly being iterated and applied to more and more industries. Generative AI, which can create text and chat with users, presents a unique challenge because it can make people feel like they're interacting with a human. Anthropomorphism is the ascription of human attributes or personality to non-humans. People often anthropomorphize artificial intelligence

DOI: 10.54254/2755-2721/67/2024MA0053

(especially Generative AI) because it can create human-like outputs. Artificial intelligence technology has brought great changes and more availability to everyone's daily life and receiving information channels. However, the collection of personal data is more and more extensive, which also makes the problem of personal data privacy and security more serious.[1]Therefore, combined with the double-sided nature of artificial intelligence, this paper analyzes the advantages and disadvantages of intelligent data processing in personal data privacy[1], applies the machine learning differential privacy algorithm combined with intelligent data processing to the research, and realizes the risk prediction and protection of personal data. This serves as a reminder for everyone on how to use artificial intelligence to protect their information security more effectively.

2. Machine learning and privacy protection

Machine learning is the core technology today and requires a lot of data when training models. How to protect this data cheaply and efficiently is an important issue. This chapter introduces machine learning and its privacy definitions and threats, summarizes the current situation in the field of privacy protection, analyzes advantages and disadvantages, and looks forward to possible research directions in the future.

2.1. Machine learning

Machine learning (ML) uses computers to effectively mimic human learning activities. It learns from existing data and produces useful models to make decisions about future behavior.In traditional machine learning training, the data of all parties is first collected by the data collector, and then the data analyst conducts model training. [2]This mode is called centralized learning. It can be seen that in the centralized learning mode, once the user has collected data, it is difficult to have control over the data, and it is unknown where and how the data will be used. The research on privacy protection in machine learning can be roughly divided into two main lines: the Federated Learning (FL) and homomorphic encryption (HE), and the perturbation method represented by differential privacy [3](DP). The encryption method not only encodes the plaintext into the ciphertext that only certain personnel can decode, but also ensures the confidentiality of the data in the process of storage and transmission. At the same time, the cipher text can be calculated directly and the correct result can be obtained by means of the security protocol. However, the data encryption process often involves a large number of calculations, which will produce huge performance overhead in complex cases, so it is difficult to implement in practical application scenarios. The main challenge for this method is to design a reasonable perturbation mechanism to better balance the privacy and availability of the algorithm.

2.2. Privacy protection Federated learning (FL)

Federated Learning (FL) is a machine learning setup. Many clients, such as mobile devices or entire organizations, work together to train a model. This process is coordinated by a central server, like a service provider. The key aspect is that the training data remains decentralized and distributed.Federated learning allows training data to be shared between multiple participants without compromising their data privacy. [4]Participants' data privacy can be protected by uploading only model parameters, such as gradients, instead of uploading training data. Federated learning enables a unified machine learning model to be trained from local data of multiple participants while protecting data privacy.

DOI: 10.54254/2755-2721/67/2024MA0053

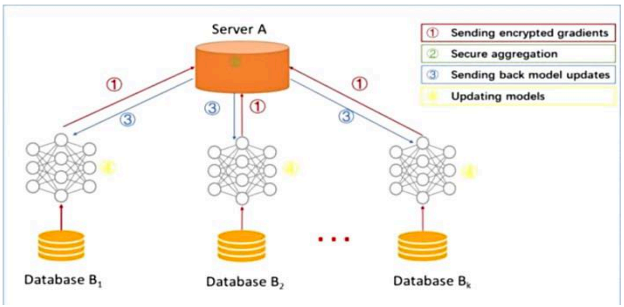

Figure 1. Federated learning algorithm model

Figure 1 illustrates the Federated Learning Algorithm Model. Federated transfer learning is a data distribution situation in which horizontal federated learning and vertical federated learning do not match. There are also two methods of privacy protection[4] in the federal learning mode: encryption and disturbance. However, federated learning is currently in the infancy of research and still faces many problems in both technology and deployment.

2.3. Homomorphic Encryption (HE)

Homomorphic encryption (HE) is a cryptographic technique that allows for computations to be performed on ciphertext without the need for decryption. This means that encrypted data can be sent to a third party for calculation, without revealing the original plaintext. While the concept of Homomorphic Encryption was first introduced in 1978, the first Fully Homomorphic Encryption framework that supports arbitrary operations on ciphertexts was developed later. Homomorphic encryption is a powerful encryption method that enables computation on ciphertext.

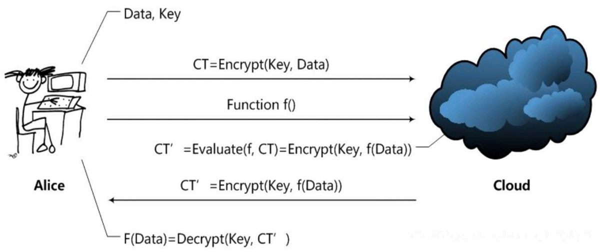

Figure 2. Schematic diagram of HE

Figure 2 depicts the Schematic diagram of Homomorphic Encryption (HE)[5]. The principle of homomorphic encryption technology is to provide a kind of encryption data processing function. That is, someone else can process the encrypted data, but the processing will not reveal any of the original content. At the same time, the user with the key decrypts the processed data and gets exactly the processed result. However, the encryption method calculation cost is not high, but the communication process designs a large number of security keys and other parameters, so the communication cost will be higher than the calculation cost.

2.4. Differential privacy protection algorithm (DP)

Differential privacy (DP) protection is achieved by adding noise to sensitive data. In federated learning, random noise is often added to the parameters that participants upload to the server in order to avoid backward inference of participants' private information.Compared with encryption methods, differential privacy mechanism is easier to deploy and apply in actual scenarios. The empirical risk minimization of differential privacy protection algorithm is commonly used to solve the optimal model parameters by gradient descent (GD) based on iterative computation. Because of the simple structure of the traditional machine learning model, the objective function J(w; D) make it as convex as possible in order to obtain a definite optimal solution. [6-7]Due to the introduction of a large number of nonlinear factors, the objective function of deep learning model is often nonconvex, so it is easy to fall into the local optimal solution when solving. [8]Therefore, this paper focuses on the design of machine learning algorithm under differential privacy protection according to different data processing and analysis capabilities. for personal data privacy protection and risk prediction, differential privacy protection algorithm is the most commonly used model learning strategy for empirical risk minimization.

3. Methodology

3.1. Create Data set

The concept of differential privacy is a crucial aspect of the protection algorithm. It involves creating a synthetic dataset that matches the shape of the original data in terms of the number of rows and columns. This is achieved by adding noise to the original data, ensuring that the privacy of individuals is protected. This ensures that the output values in the synthetic data process have the same properties as the values in the original data, forming a suitable histogram. [9]However, in most cases, there may be some differences between the histogram and the synthetic data set. To obtain suitable synthetic data, the first step is to perform a range query and synthesis on a column of raw data. For instance, in the experiment described in this paper, the query range for the dataset was set to ages between 21 and 23 years old.

The histogram of the dataset is as follows:

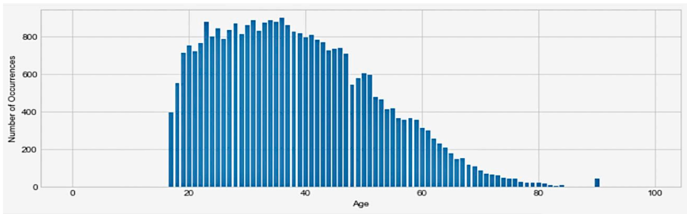

Figure 3. Data set query range results

The range queried in Figure 4 is the histogram result of the dataset between zero and 100 years old, and the result shows a high similarity between the Saturn list and the output data of the original plt. Therefore, it can be used as experimental data for further training.

3.2. Add differential privacy

Suppose there is a stochastic algorithm M, S is the set of all possible outputs of M, and Pr[10] represents the probability. For any two adjacent data sets. Two data sets The difference between two data sets is only 1 record if the probability distribution of two adjacent sets satisfies the following constraints:

$$P_r[M(D) \in S] \leq e^{\in} P_r[M(d') \in S] \tag{1}$$

When the above conditions are met, it is said that algorithm M provides differential privacy protection, where ∈ is the differential privacy budget, which is used to ensure that one data record is added or less in the data set.

In this experiment, when adding differential privacy, I need to make the synthesized data compliant with differential privacy. Each count in the histogram is added separately, usually by Laplacian noise, by parallel combination, which satisfies ϵ \epsilonϵ- differential privacy.

3.3. Generate tabular data

When considering this process, we focus on a single column of data and ignore all other columns, which is called a marginal distribution, especially a one-way marginal distribution. [11]To get the count for each bar, we need to calculate the probability for each bar. These probabilities need to be normalized to ensure that they sum to 1. This process converts counts into probabilities in preparation for subsequent sampling steps. At this stage, we also want to ensure that the count value is not negative to maintain

DOI: 10.54254/2755-2721/67/2024MA0053

logical consistency.Next, by randomly selecting a bar of the bar graph, we sample from the distribution it represents by probability weighting. In this way, we finally get a probabilistically normalized and logically consistent sample extraction process.

In the aforementioned synthetic data, when we plot the standardized count, we can think of it as the probability for each corresponding histogram bar, since they sum to 1. This means that we successfully converted the count into a probability distribution and observed that the shape drawn closely resembled the original histogram. This similarity also extends to the shape of the original data, indicating that our composite data maintains an overall tendency to be consistent with the original data. As a result, we have successfully constructed a probability distribution whose shape is consistent with the original data, providing a reliable basis for subsequent analysis.

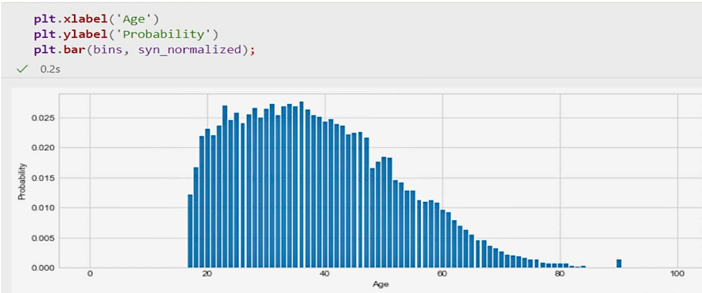

Figure 4. Data probability result histogram

The final step involves generating a new sample based on the previously calculated probabilities. We make this choice using a random forest algorithm, which allows us to pass in a list of probabilities associated with the choices given in the parameters. [12]By implementing weighted random selection accurately, the algorithm meets the requirement of sampling task. It is encouraging that we are able to generate any number of samples without additional privacy costs, as we have protected the privacy of the data by setting the count to differential privacy. This process is logically clear and ensures data privacy.

3.4. Training model

In this methology, take age and occupation, for instance; these variables may exhibit correlation, as managers are typically older.If we isolate each column, we might accurately estimate the count of 18-year-olds and managers separately, but we could significantly misjudge the count of 18-year-old managers.

To address this, we need to account for all possible combinations of age and occupation, preserving the correlations.This consideration becomes crucial for an accurate representation, mirroring our earlier approach when constructing the emergency table.

Table 1. Training model result

Occupation	Adm-clerical	Armed-Forces	Craft-repair	Exec-Farming-fishing	Handlers-cleaners	Machine--op-inspct	Other-service
Age							
17	23	0	14	9	40	2	4
18	55	0	17	14	50	17	3
19	102	0	40	24	65	30	3
20	117	0	35	2	81	41	4

Can see this effect by plotting the age histogram in the new composite data set(Table 1); Note that it has roughly the right shape, but it's not as smooth as the raw data or the differential private count we used for the age column itself.

DOI: 10.54254/2755-2721/67/2024MA0053

3.5. Experimental result

In this experiment, Differential privacy protection algorithms play a crucial role in synthesizing datasets while preserving privacy. The methodology involves creating a synthetic dataset that retains the shape and statistical properties of the original data. [13-14]For attributes like age, histograms are used to represent distributions, ensuring accurate answers to range queries without revealing individual data points. The experiment underscores challenges in maintaining correlations between attributes, like age and occupation, in synthetic data generation. In conclusion, the study highlights the effectiveness of differential privacy algorithms in safeguarding individual privacy while providing a functional synthetic dataset for analysis. However, attention to accuracy and inter-attribute correlations in synthetic data is essential.

4. Conclution

Differential privacy is a new definition of privacy protection proposed by Dwork et al in 2006 for the privacy leakage problem of statistical databases, and has now developed into the most advanced privacy protection method. Differential privacy protection can solve the two defects of traditional privacy protection model. Under this maximum background knowledge assumption, differential privacy protection does not take into account any possible background knowledge of the attacker, and it is built on a solid mathematical foundation, provides a strict definition of privacy protection and quantitative evaluation methods, so that the level of privacy protection provided by data sets under different parameter processing is comparable.

Therefore, it can be seen that these algorithms can effectively reduce the risk of data leakage by applying mathematically strict privacy protection on the data set. It also allows researchers and institutions to use the data for important analysis and model training.

Ackonwledgments

During the course of this research, I am grateful to have been able to reference and cite the research findings of Liu Bo and his colleagues. Their work has provided valuable support and guidance to my exploration in the fields of artificial intelligence, machine learning, and personal privacy data protection. I would like to express my gratitude to Dr. Liu Bo and his research team for their paper titled 'Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning.' The core conclusions presented in the paper 'Algorithms' (arXiv preprint arXiv:2312.12872, 2023) provide a solid foundation for my research. The paper deeply discusses the integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms, providing a new perspective for my understanding of artificial intelligence. In addition, Liu Bo's research has set an example in the field of machine learning and deep learning.

References

[1] Shokri, Reza, and Vitaly Shmatikov. "Privacy-preserving deep learning." Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015.
[2] Rivest R，Adleman L， Dertouzos M L. On data banks and privacy homomorphisms [J]. Foundations of Secure Computation，1978，4(11): 169180.
[3] Gentry C. Fully homomorphic encryption using ideal lattices[C] //Proc of the 4lst Annual ACM Symp on Theory of Computing. New York: ACM，2009: 169178Dwork C，McSherry F, Nissim K, et al. Calibrating noise tosensitivity in private data analysis [C] //Proc of the 3rdTheory of Cryptography Conf. Berlin: Springer， 2006 : 265284
[4] Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of the IEEE 1st International Conference on Broadnets Networks (BroadNets’04) . IEEE, Los Alamitos, CA, 210–217. https://doi.org/10.1109/BROADNETS.2004.8.

DOI: 10.54254/2755-2721/67/2024MA0053

[5] Xinyu Zhao, et al. “Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.
[6] Shulin Li, et al. “Application Analysis of AI Technology Combined With Spiral CT Scanning in Early Lung Cancer Screening”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 52-55, https://doi.org/10.54097/LAwfJzEA.
[7] Liu, Bo & Zhao, Xinyu & Hu, Hao & Lin, Qunwei & Huang, Jiaxin. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science. 3. 36-42. 10.53469/jtpes.2023.03(12).06.
[8] Yu, Liqiang, et al. “Research on Machine Learning With Algorithms and Development”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.
[9] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. (2023). The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 1(1), 30-37.0
[10] Zhou, H., Lou, Y., Xiong, J., Wang, Y., & Liu, Y. (2023). Improvement of Deep Learning Model for Gastrointestinal Tract Segmentation Surgery. Frontiers in Computing and Intelligent Systems, 6(1), 103-106.6
[11] Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma. (2024). Frontiers in Computing and Intelligent Systems, 6(3), 127-131. https://doi.org/10.54097/zJ4MnbWW.
[12] Zhang, Q., Cai, G., Cai, M., Qian, J., & Song, T. (2023). Deep Learning Model Aids Breast Cancer Detection. Frontiers in Computing and Intelligent Systems, 6(1), 99-102.3
[13] Manchini Carlos, et al. "A new approach to data differential privacy based on regression models under heteroscedasticity with applications to machine learning repository data." Information Sciences 627. (2023):
[14] Xu, J., Pan, L., Zeng, Q., Sun, W., & Wan, W. (2023). Based on TPUGRAPHS Predicting Model Runtimes Using Graph Neural Networks. Frontiers in Computing and Intelligent Systems, 6(1), 66-69.7