Analysis of Legal Protection Mechanism of User Privacy Information Based on Big Data and Machine Learning Algorithm

Xiangfen Ma

Faculty of Mathematics and Information Engineering, Puyang Vocational and Technical College, Puyang, Henan, 457000, China

Keywords: Big data, Machine learning algorithm, User privacy, Legal protection mechanism of information

Abstract: In recent years, due to its increasingly mature and vigorous development, a large number of companies have made breakthroughs in machine learning applications, such as medicine, spatial information security and other applications. Big data analysis and publishing technology enables data analysts to learn the common laws of big data, among which statistical information analysis and machine learning are popular application fields. Based on big data and machine learning algorithm, this paper studies the legal protection mechanism of users' privacy information. In machine learning, security is to confirm that users' data information is properly used, thus protecting the use and improvement of machine learning. Because security and privacy are two completely different and closely related categories, security is also the cornerstone of protecting users' privacy. For distributed big data machine learning, multi-data owners use their own data sets to perform machine learning locally, and then share their local model parameters to obtain the global model, so as to realize joint learning based on big data.

1. Introduction

In recent years, machine learning algorithms have received more and more attention and development. Its excellent data mining technology has been applied in a wide range of fields, such as disease detection, economic prediction, network optimization, and has rapidly gained popularity. Machine learning is also one of the key technologies of artificial intelligence. In recent years, due to its increasingly mature and vigorous development, a large number of companies have achieved breakthrough development in the field of machine learning applications, such as applications in medicine, space information security and other application fields [1-2]. Big data analysis and publishing technology enables data analysts to learn the common laws of big data, among which statistical information analysis and machine learning are hot application fields. However, all data analysis tasks may leak personal privacy information if appropriate privacy protection technologies are not added. In actual training, the machine learning algorithm needs as much sample data as possible, but the amount of data provided by a single data source is limited. Most of the data required by the algorithm come from multiple data sources, such as different people, companies, organizations or countries [3].

This paper studies the legal protection mechanism of user privacy information based on big data and machine learning algorithms. In machine learning, security is to confirm that user data information is properly used, thus protecting the use and improvement of machine learning. Because security and privacy are two completely different and closely related categories, security is also the cornerstone of protecting users' privacy. For distributed big data machine learning, multiple data owners use their own data sets to perform machine learning locally, and then share their local model parameters to obtain the global model, so as to achieve big data based joint learning, which we call federation learning [4]. Data providers do not want to disclose their private data to others, and models that have been trained by data from multiple data sources should not be published to any single participant, so we need to use privacy protection methods based on multiple data sources. The computing problem is privacy protection data mining and privacy protection deep learning, and there are differences in computing between the two. For example, for privacy protection data

Published by CSP © 2022 the Authors

mining, under the condition of ensuring the confidentiality of the data, the miners need to be able to learn the potential statistical rules between data items. Usually, the data calculations involved are addition operations and comparison operations [5-6]. The privacy attribute needs to be protected while the public attribute can be disclosed. However, according to later research, there is no clear boundary between the privacy attribute and the public attribute, because any combination of attributes may reveal the unique characteristics of individuals. This conclusion is particularly consistent with today's big data environment. Therefore, the right to privacy is also a complex category that lacks the definition of recognized norms. In machine learning, privacy is usually defined as the right of human beings to ensure that their private data information will not be disclosed.

2. Overview of the development of privacy technology

The right of privacy is a quite complex jurisprudential category, but there is still a lack of a recognized concept of jurisprudential norms. That is to say, attackers with rich background knowledge cannot obtain the privacy information of a single data in the dataset through the establishment of the target model through the dataset that is highly similar to the target dataset and is only one record different in extreme cases [7]. The existing solutions to privacy problems in machine learning algorithms are mainly based on four types of privacy protection technologies:

(1) Homomorphic encryption

Homomorphic encryption technology is to convert data into ciphertext, and directly perform the same basic computing processing as plaintext, such as addition and multiplication. It has been widely used in the practice of secure computing. Solutions based on homomorphic encryption usually require a trusted encryption service provider or rely on other privacy technologies.

(2) Secret sharing;

Secret sharing technology allows users to divide a secret s into n sub secrets, and then distribute them to n users. According to whether it has threshold property, we divide secret sharing technology into two categories: when k=n, it is common secret sharing; When k<n, it is threshold secret sharing.

(3) Garbled circuit

Scramble circuit technology is very successful in solving secure multi-party computation, symmetric encryption and inadvertent transmission problems based on digital circuits. The problem of security comparison is solved by using garbled circuits. The garbled code circuit scheme has weak scalability and is prone to generate high computational complexity.

(4) Differential privacy

Differential privacy protects data privacy by reducing the value of data in a single use. Although this method can effectively protect private data, the resulting reduction in the value of the data will lead to a decline in the accuracy of machine learning training based on small data sets.

In the process of obtaining user data, some users' secrets will be disclosed. Therefore, Google and Apple enterprises must save user data through differential privacy methods. In the actual process of using data, even if a single data information is meaningless, data analysis information is still useful. The goal of privacy protection is to allow data analysts to learn group rules without violating personal privacy. Therefore, how to define personal privacy disclosure is crucial [8]. From the perspective of information theory, the learning of group rules will inevitably lead to data analysts getting more information to guess personal privacy. In real data sets, the data characteristics are generally rational numbers. However, currently known homomorphic encryption algorithms generally only support integer and ring element encryption. Because the change of data accuracy will bring uncertainty to the data analysis results, in order to ensure the accuracy of analysis, we need to convert the data into a form suitable for encryption before encrypting rational data. Sharing does not only mean that by opening data information to other participants, all major participants can train their own models on various data analysis sets independently and share training results with other participants, thus indirectly sharing their training data information [9].

3. Research on Legal Protection Mechanism of User Privacy Information Based on Machine Learning Algorithm

3.1. Problems in Machine Learning

At present, there is still a lack of unified security evaluation standard and a lack of unified measurement standard for secret disclosure. It is an important link to build a complete evaluation system and standardize privacy protection principles to maintain the security and privacy of machine learning. The inadaptability of confrontation training makes it necessary to introduce enough confrontation samples to effectively prevent unknown confrontation threats, which is also a difficult problem in confrontation training and needs to be overcome. The typical structure of machine learning allows the server to aggregate the gradients calculated locally by each data provider, and then all data providers update the system model using the aggregated results returned by the server [10]. Obviously, any data provider may reveal the whole model. In machine learning, in order to protect the personal information of data owners and allow data analysts to analyze hidden patterns in data, there are two traditional privacy protection methods: non-interactive and interactive. The effective way to protect privacy is to use cryptography. However, the current homomorphic cryptography algorithms are too expensive, and some non-polynomial algorithms in computational machine learning cannot be implemented. This paper analyzes the problems existing in machine learning, and constructs the structure diagram, as shown in Figure 1.

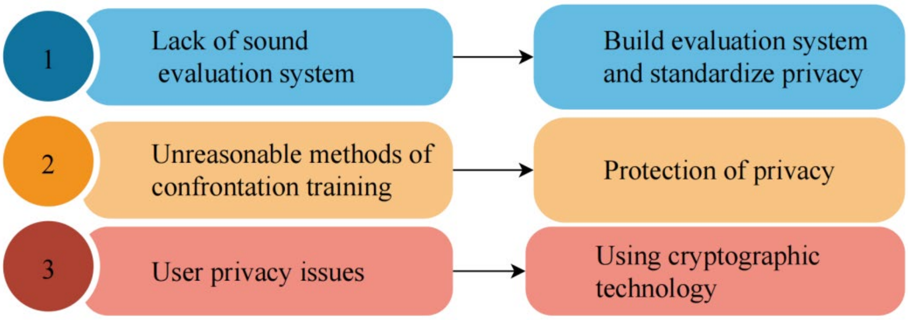
1
Lack of sound evaluation system
Build evaluation system and standardize privacy
2
Unreasonable methods of confrontation training
Protection of privacy
3
User privacy issues
Using cryptographic technology

Figure 1 Problems in machine learning

In the process of machine learning clustering, data does not need to be labeled, and the result of clustering is to find out the internal structure and representation of data. Data mining, including frequent itemsets mining and association rules mining, is also a kind of unsupervised learning. Reinforcement learning is based on the maximization of some scalar reward goals, data and some future rewards. Therefore, it is a major research issue to study effective encryption methods to protect users' privacy. In the face of information security threats, people must also deeply explore the threat of poisoning. In the face of information security threats, homomorphic encryption has always been regarded as the most important technical means of privacy protection machine learning.

3.2. Corresponding strategies for machine learning problems

The existing machine learning methods for privacy protection often assume that the ECS is a passive model, and fully consider the reliability of data information and the effectiveness of mechanical teaching when the ECS does not collude with each other; In addition, to push the higher security level to more heavy scenarios, it is also necessary to give consideration to fairness and consistency when encountering malicious attackers. With the further development of machine learning, deep learning has become an important means to find data rules. In general, machine learning fits data by building models and optimizing loss functions. However, if the machine learning model is used to fit personal sensitive data. According to the methods proposed at present, accuracy and effectiveness must be increased to reduce errors and take into account more dangerous situations, such as malicious scenes. Therefore, suggestions for future R&D direction include the following aspects. For this content, the flow chart of the corresponding strategy for machine learning problem is constructed, as shown in Figure 2.

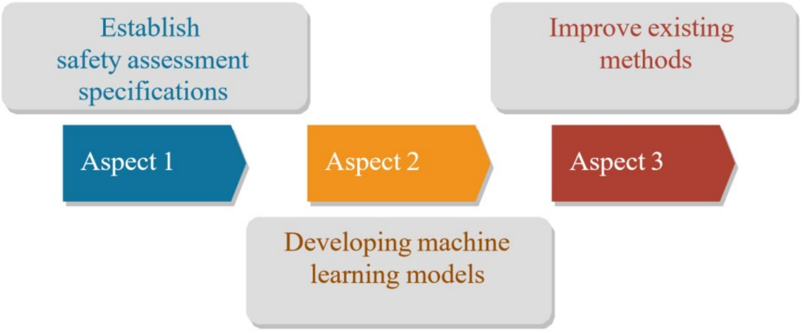
Establish safety assessment specifications
Improve existing methods
Aspect 1
Aspect 2
Aspect 3
Developing machine learning models

Figure 2 Flow chart of corresponding strategies for machine learning problems.

The flow chart of the corresponding strategies for machine learning problems is mainly divided into three aspects:

① Establish safety evaluation specifications.

In view of the fact that there is no reasonable and complete security assessment standard in China at present, and there is no reasonable and unified management standard for the application scope and acquisition process of private data in various organizations, which inevitably leads to huge risks of information security, it is imperative and urgent to establish a reasonable and complete and unified security assessment standard.

② Develop machine learning model.

The economic loss caused by the leakage of the content of the model to organizations and institutions will be immeasurable, so it will be an important work in the future to develop a highly robust machine learning model that can counter stronger threats.

③ Improve the existing mode.

At present, most of the privacy protection methods are homomorphic encryption, secure multiple operations and differential privacy. However, at present, the transmission, operation and other expenses of these technical means are quite large, which reduces the effectiveness of calculation and causes unnecessary waste of resources.

All along, the privacy security and functional security of machine learning models are in two relatively parallel research lines. Whether the model reveals personal privacy or not, there is also a kind of security that refers to the functional security of the model, such as anti-sample attack, sample poisoning, etc., which refers to the misjudgment of the data model that is difficult to distinguish by malicious attackers with naked eyes. The above content is a series of analysis on the legal protection mechanism of users' privacy information based on big data and machine learning algorithm.

4. Conclusions

With the rapid development of mobile Internet, mobile intelligent terminals are rapidly popularized and become an indispensable daily auxiliary tool for people. At the same time, they also expose relatively serious security problems. When there are many types of client data sets or the data sets are small, it is difficult to successfully train such attack models. Furthermore, the machine learning attack model based on shared gradient training is not interpretable. One of them is to damage the interests of users by stealing their private data. Machine learning based on big data makes up for the lack of quantity and diversity of training data under a single data source, which has broad application prospects and practical significance. The vulnerability of machine learning itself leads to the inevitable threat to its security, and at the same time, it is easy to expose users' secrets. In recent years, the security of machine learning has received widespread attention. However, as far as all the current researches on the privacy disclosure of federated learning are concerned, it is not clear how the gradient in federated learning leaks data privacy. The existing security analysis either aims at the shallow learning model or uses an additional machine learning model to learn the privacy information about the original dataset from the gradient. The privacy protection technology under machine learning algorithm directly affects the development and promotion of this machine learning scheme in the real society, which is of great significance. At present, the research and development of machine learning security protection and privacy protection technology is still in its infancy, and further research is needed in the future.

References

[1] Li Y,Ma L,Li X. IDR Privacy Protection Based on Database Digital Watermarking[J]. Recent advances in electrical & electronic engineering, 2020(1):13-38.

[2] Li P,Xu C,Xu H, et al. Research on Data Privacy Protection Algorithm with Homomorphism Mechanism Based on Redundant Slice Technology in Wireless Sensor Networks[J]. China Communications: English, 2019, 16(5):13-28.

[3] Yang S. Research on K-anonymous Privacy Protection Mechanism Based on Spatial Location[J]. Modern Information Technology, 2018, 39(8):11-19.

[4] He Y,Chen J. User location privacy protection mechanism for location-based services[J]. Digital communication and network: English, 2021, 7(2):13-19.

[5] Li J,Fan X,Wang Y, et al. DESIGN OF SHARING ECONOMY PRIVACY PROTECTION MECHANISM BASED ON BLOCK CHAIN[J]. Computer Applications and Software, 2019, 33(11):18-22.

[6] Wang X,Wu C. Research on Legal Protection of Personal Data Privacy of E-Commerce Platform Users[J].2021, 52(17):26-48.

[7] Zhou Qiang, Yue Kaixu, Duan Yao. Privacy Cost Analysis and Privacy Protection Based on Big Data[J]. Journal of Donghua University, 2019, 36(01):99-108.

[8] Kaori, Ishii. A Study of Adjacent Legal Fields to Laws on Privacy and Personal Data Protection[J]. Journal of Information and Communications Policy, 2019, 3(1):47-72.

[9] Jiang N,Qing-Chuan G U,Yang H Y, et al. Machine learning Algorithm Under Big Data[J]. Computer and Information Technology, 2019, 36(15):25-39.

[10] Guan W,Zhang L. MAH-ABPRE based privacy protection algorithm for big data[J]. Computer Engineering and Design, 2018, 11(2):11-19.