ISSN: 2664-9640

Building a User Data Protection Model by Combining Distributed Storage and Machine Learning

Shihan Zhao*

International College, Zhengzhou University, Zhengzhou, 450052, China

*zshzhao25@stu.zzu.edu.cn

Abstract

With the acceleration of the digitalization process, the risk of user data leakage has become increasingly prominent. Especially in fields such as healthcare and finance, the leakage of users' sensitive information may lead to serious consequences. This study aims to combine distributed storage with machine learning to build a new user data protection model. Therefore, this research will adopt federated learning as the framework, integrate lightweight deep learning models with differential privacy technology, to realize decentralized data storage and privacy-enhanced model training. This approach is intended to address the issues of traditional machine learning's reliance on centralized data and the hidden risks of privacy leakage. In addition, this study proposes a "edge-region-central" hierarchical architecture, a three-level privacy protection system, and a storage-computation collaboration mechanism. Finally, the advantages of the model in terms of privacy protection effect, accuracy, and efficiency will be verified through experiments, providing technical support for user data security.

Keywords

Distributed Storage; Federated Learning; Lightweight Model; User Data Protection Model.

1. Introduction

In the digital age, software, enterprises, and public service units are increasingly collecting user data. However, improper management leading to user data leakage, especially the leakage of sensitive information in fields such as healthcare and finance, is highly likely to cause identity theft, property losses, and social trust crises [1]. Therefore, it is urgent for various countries to accelerate the introduction of data protection policies [2].

Through participating in this scientific research project, I have gained interdisciplinary thinking - recognizing that although machine learning can identify risks through data analysis, it has privacy risks due to its reliance on centralized data and model vulnerabilities; although distributed storage can achieve decentralization to reduce the impact of single-point leakage, it lacks active protection capabilities. Thus, the two need to work together [3].

Therefore, in the following, I will propose to build a user data protection model with both storage security and intelligent protection capabilities through the integration of distributed storage and machine learning, so as to solve the limitations of a single technology.

2. Research Methods and Technical Route

2.1. Research Objectives

This study aims to provide security guarantees for the entire data lifecycle through decentralized storage and privacy-enhancing technologies, thereby reducing the risk of data leakage. While adhering to the bottom line of privacy protection, it also seeks to improve model

ISSN: 2664-9640

performance, ensuring that model accuracy remains unaffected while enhancing training efficiency. In addition, it hopes to optimize the solution according to the characteristics of edge devices such as intelligent medical terminals and industrial Internet of Things, so as to meet their lightweight requirements and thus achieve good adaptation to edge scenarios.

2.2. Research Content

In terms of model architecture design, this paper innovatively proposes a hierarchical architecture of "edge lightweight model + regional federated aggregation + central global optimization". Edge nodes are responsible for local training, regional nodes perform parameter aggregation, and central nodes optimize the global model, so as to adapt to the computing power of edge devices.

In terms of privacy enhancement mechanisms, differential privacy technology is integrated to build a three-level protection system [2,4]. Specifically, it is realized by edge nodes injecting noise, regional nodes filtering privacy information, and central nodes securely aggregating parameters. In the storage-computation collaboration strategy, a task scheduling algorithm is designed based on distributed storage metadata to dynamically match data storage locations with training nodes, thereby improving resource utilization.

2.3. Research Methods

In terms of research methods, first, an interdisciplinary approach is adopted, integrating federated learning and lightweight models in the field of machine learning, differential privacy in cryptography, and distributed system theories to build a cross-domain technical system. Second, model construction and simulation are carried out: based on Python frameworks (such as TensorFlow Federated and PyTorch), a federated learning platform is built, and EdgeSim is used to simulate the edge device environment to verify the performance of the algorithm. Finally, the experimental verification method is applied: typical scenarios such as intelligent medical terminals and industrial sensors are selected, and the proposed scheme is compared with combination schemes of traditional federated learning , single lightweight models, and differential privacy technologies[4,5]. Various indicators are then evaluated, including privacy protection effects (e.g., information leakage probability), model accuracy (e.g., AUC), and training time consumption.

2.4. Technical Route

The research process will start with demand analysis and literature research. While clarifying the privacy protection requirements in edge scenarios, it will sort out the pain points in the integration of distributed storage and machine learning, thereby laying a foundation for subsequent research [6,7]. On this basis, lightweight models suitable for edge devices (such as MobileNet variants) are screened, and then model parameters are optimized to reduce computing power requirements [8].

Next, entering the architecture design stage, a hierarchical federated learning architecture is constructed, differential privacy mechanisms are integrated, and a storage-computation collaborative scheduling algorithm is designed to form a complete technical solution. Using model construction and simulation methods, a federated learning platform based on Python frameworks (such as TensorFlow Federated and PyTorch) is built, and then EdgeSim is used to simulate the edge device environment. Combined with the experimental verification method, the model performance is tested in typical scenarios such as intelligent medical terminals and industrial sensors.

ISSN: 2664-9640

3. System Design and Implementation

3.1. Overall Architecture

The architecture is divided into three layers: the edge layer, the regional layer, and the central layer.

The edge layer consists of edge devices (such as medical terminals) and is responsible for storing local data. It completes initial training based on a lightweight model and uploads model parameters after injecting differential privacy noise.

The regional layer undertakes the task of aggregating parameters from edge nodes, and is also responsible for filtering abnormal parameters and performing secondary encryption, thereby further reducing the risk of privacy leakage.

The central layer is responsible for receiving parameters from the regional layer, optimizing the global model, and then distributing updates to the edge layer, thus realizing the principle of "models move while data remains stationary". Meanwhile, the research also incorporates a differential privacy mechanism and designs a scheduling algorithm for the collaboration between storage and computation, thereby forming a complete technical solution.

3.2. Core Module Design

Next, entering the architecture design stage, the research constructs a hierarchical federated learning architecture (edge layer, regional layer, central layer) and integrates multiple modules to form a complete solution. First, the differential privacy module. Laplacian noise injection is used in the edge layer, k-anonymity technology is adopted in the regional layer to filter sensitive parameters, and secure multi-party computation (SMPC) is employed in the central layer to aggregate parameters, forming an end-to-end privacy protection [2,4]. Second, the distributed storage collaboration module. Based on the Ceph distributed storage system, metadata is used to record data locations and access permissions [6-8], and a scheduling algorithm is matched to assign training tasks according to node computing power and network status, improving resource utilization efficiency.

4. Experiments and Results Analysis

4.1. Experimental Environment

Hardware: Edge nodes (Raspberry Pi 4B), regional nodes (server: Intel Xeon E5-2680 v4), central nodes (server cluster: 4×NVIDIA A100);

Software: TensorFlow Federated 0.20.0, EdgeSim 2.0, Python 3.8;

Datasets: For medical scenarios, the MIMIC-III electronic medical record data (desensitized) is used [4]; for industrial scenarios, the NASA bearing sensor dataset is adopted.

4.2. Evaluation Indicators

Privacy protection effect: Information entropy (measuring the risk of data leakage, with higher values indicating higher security), differential privacy budget (ε, with smaller values indicating stronger privacy protection);

Model performance: Classification accuracy (e.g., AUC), training time, model size;

System efficiency: Resource utilization (CPU/memory occupancy), task completion delay.

4.3. Experimental Results

Privacy protection effect: The information entropy of this model is 25% higher than that of traditional federated learning, and it still maintains 90% privacy security when ε=0.5 [2,4];

ISSN: 2664-9640

Model performance: On the MIMIC-III dataset, the AUC reaches 92% (close to 94% of centralized training), and the training time is 30% less than that of a single lightweight model [3,5];

System efficiency: Resource utilization is increased by 40%, task delay is reduced by 20%, which is suitable for edge devices with low computing power scenarios.

4.4. Results Analysis

The experimental results show that the proposed technical solution verifies the effectiveness of the integration of distributed storage and machine learning from multiple dimensions. Through the design of "edge lightweight model + regional federated aggregation + central global optimization", the computing power limitation of edge devices is solved, enabling edge devices to efficiently participate in machine learning with limited computing power, thus reflecting the deep adaptation of the distributed architecture to edge computing scenarios. In addition, the three-level privacy protection mechanism balances model accuracy while ensuring privacy security [2,4]. The design of end-to-end protection ensures data security while avoiding excessive protection that impairs accuracy, making the model's AUC close to the level of centralized training [3,5]. Moreover, the storage-computation collaboration strategy dynamically matches data with nodes and intelligently assigns tasks based on the Ceph system [6,8], which improves system efficiency, intuitively demonstrates the collaborative effect, and further verifies the value of integration in enhancing the overall efficiency of the system.

5. Conclusion and Prospects

The integrated model of distributed storage and machine learning constructed in this study achieves a balance between privacy protection and model performance in edge scenarios through a hierarchical architecture, privacy-enhancing technologies, and collaborative strategies, providing a new solution for user data security. In the future, we can explore more efficient lightweight models (such as Transformer variants) to further improve accuracy and efficiency; expand to more scenarios (such as smart homes and Internet of Vehicles) to verify the universality of the model; combine blockchain technology to enhance the immutability of distributed storage and improve the credibility of the system.

References

[1] Metwally, A., Snyder, M., & McLaughlin, T. (2022) Machine learning predicts type 2 diabetes subtypes. Nature Biomedical Engineering, 6: 210–218.

[2] Zhao, B., Singhal, R., Raghavan, M., et al. (2022) Learning from Multiple Jurisdictions: A Data Sharing Framework with Local Differential Privacy and Fairness Guarantees. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. pp. 1-14.

[3] Zhou, Z. H. (2020) New machine learning theories and methods for ambiguous objects. Scientia Sinica (Informationis), 50: 1407–1420.

[4] Germanidis, I., Lampropoulos, P., Mavromoustakis, C. X., et al. (2021) Swarm Learning for decentralized and confidential clinical machine learning. Nature, 593: 265–270.

[5] Wu, C., Zhang, Y., Wang, X., et al. (2022) A federated graph neural network framework for privacy - preserving personalization. Nature Communications, 13: 1–13.

[6] Georgiou, C., Musial, P., Shvartsman, A. (2009) Reconfigurable distributed storage for dynamic networks. Journal of Parallel and Distributed Computing, 69: 100–116.

[7] Tian, C. (2020) On a class of Distributed Storage Systems with Prioritized Data Sources. In: Proceedings of the 2020 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. pp. 415-427.

ISSN: 2664-9640

[8] Yang, X., Zhang, Y. (2006) Research and performance evaluation of data replication technology in distributed storage systems. Computers & Mathematics with Applications, 52: 779-793.