DOI: 10.23977/jeis.2025.100206
ISSN 2371-9524 Vol. 10 Num. 2

Cooperative Detection Algorithm of Malicious Nodes Based on Federated Learning

Zixuan Wan, Xiaosheng Wu

School of Computer Science and Information Technology, Harbin Normal University, Harbin, 150025, China

Keywords: Mobile Crowdsensing; Federated Learning; Malicious Node Detection; Gradient Similarity

Abstract: In Mobile Crowdsensing (MCS) networks, traditional malicious user detection methods typically rely on transmitting vast amounts of raw data to a central server for analysis. This approach not only incurs significant communication overhead, exacerbating network congestion, but also poses a high risk of exposing users' sensitive data. To address these challenges, this paper introduces an MCS malicious behavior detection framework that integrates the concepts of Federated Learning (FL) and edge computing. This framework employs a distributed architecture centered around edge servers, enabling multiple edge nodes to process data locally and collaboratively train detection models, thereby effectively safeguarding user privacy. Additionally, to counter potential malicious users in federated learning, a legitimate user identification method based on user contribution levels is designed using the gradient similarity principle. By excluding malicious users, the system can mitigate the risk of attacks, ultimately enhancing the accuracy and security of the system.

1. Introduction

Mobile Crowd-Sensing (MCS) networks serve as the backbone of intelligent transportation systems. However, the sensitive data collected during real-time information exchange-particularly location data-has raised privacy concerns and exposed vulnerabilities to malicious attacks. To address these security and privacy challenges, an AI-driven intrusion detection solution has been proposed. While AI-based malicious behavior detection technology can protect MCS networks from attacks, its application development remains constrained by data scarcity. The primary obstacle stems from public concerns about privacy breaches, which have significantly hindered the advancement of AI technologies in this critical field.

In this context, we propose leveraging edge computing, federated learning, and blockchain technologies to protect and build a malicious behavior detection system for mobile crowdsensing networks. The integration of edge computing, blockchain, and federated learning can enhance the security and privacy protection of these networks. Edge computing places security mechanisms closer to end devices, reducing communication latency while improving overall system security. Federated learning enables collaborative node learning while maintaining privacy. Blockchain ensures the credibility and reliability of data during federated learning training processes.

The main innovations of this paper include the following two aspects.

(1) A distributed architecture based on edge computing is constructed, and the federated learning mechanism is skillfully used to enable multiple sensing nodes to realize efficient collaboration and information sharing under the premise of ensuring network privacy and security.

(2) The proposed detection model deploys the training task to the edge node, which effectively improves the network performance while ensuring the detection efficiency of the network.

2. Related work

Federated learning distributes model training tasks across multiple endpoints, keeping participants' raw data locally while only uploading updated parameters or gradient information[1]. This mechanism effectively prevents sensitive data from being centrally stored on servers, thereby reducing privacy risks. Although training data is distributed across devices, the system can still aggregate parameters from local models to build a reliable global model[2]. This method not only guarantees the accuracy of the model, but also enhances the ability of user privacy protection, so that the MCS system can realize more secure data collaboration in the multi-participation environment. Compared to traditional centralized architectures that upload all data to central servers, federated learning mitigates single-point failure risks and reduces susceptibility to malicious attacks through distributed processing paradigms, significantly lowering security vulnerabilities associated with centralized storage[3].

Even so, federated learning still faces data security risks. Some scholars have proposed protection methods to address these risks. [3] A blockchain and federated learning hybrid approach was developed to implement privacy-preserving edge computing in federated learning. This method effectively resists 30% of poisoning attacks while maintaining high model accuracy. [4] An asynchronous federated learning system built on blockchain technology ensures data security. By combining local models with consensus algorithms, it constructs a complete, reliable, and secure dataset. It integrates all participants 'encrypted matrices and shared gradients through Newton iteration to solve optimal regression coefficients, enhancing encryption efficiency while addressing user privacy concerns. For vertical distribution datasets, Its workflow involves: First, the central server aligns all participants' samples using their identifier-generated CLK (Clock Lock), then calculates final results; Second, the central server generates public/private keys along with Taylor's approximation loss to determine encrypted gradients for two participants. To ensure data security, both parties must generate random masks that enable the central server to accurately identify users' private information. Wu et al. [7] proposed a joint learning solution that combines adaptive gradient descent with differential privacy mechanisms. This approach introduces differential privacy to counter inference attacks while providing more effective and quantifiable privacy protection for collaborative learning. The application of FL to crowd sensing system can enhance data security, but it also introduces the risk of malicious user-initiated model extraction attack and model reverse engineering attack, resulting in privacy vulnerabilities [1].

To address the limitations of existing solutions, developing a secure, trustworthy, and privacy-protected malicious behavior detection framework for mobile swarm sensing networks is crucial. This paper proposes a framework that leverages federated learning and edge computing technologies to protect network privacy and ensure reliability while enabling timely and accurate detection of malicious attacks in mobile swarm sensing networks. The proposed framework effectively countermeasures against attacks such as false location reporting.

3. System model

This study designs a secure system for detecting malicious nodes based on federated learning. By integrating federated learning models into crowd sensing networks, it ensures local storage of user data to mitigate data leakage risks. To achieve this, we developed a federated learning framework tailored for crowd sensing applications using a Zhang surface framework based on federated learning principles. Considering potential data leakage during interactions within federated learning systems, our goal is to establish a secure framework protecting data in crowd sensing applications[5]. To achieve this, we employ appropriate data encryption methods for transmission processes while accounting for typical data volumes encountered in crowd sensing application scenarios. Additionally, the primary threat in federated learning systems stems from malicious insiders involved in training processes. Each participant in federated learning possesses a dataset that allows malicious actors to manipulate datasets, training processes, and models, thereby compromising model performance. To mitigate internal attack risks, rigorous verification of training participants 'legitimacy is essential. This paper proposes a legitimate user screening method based on client contributions to global models. By calculating gradient similarity between different models, it identifies trustworthy users and evaluates their contribution levels to the server's global model[6].

3.1 Contribution-based legal user selection method

Data security threats originate from malicious users who attempt to compromise system model accuracy through abnormal training or uploading erroneous data. Furthermore, they may exploit vulnerabilities via gradient analysis to steal other users' data, thereby creating significant security risks. Typically, these malicious actors are few in number and contribute minimally or negatively to the overall model on servers. Leveraging this principle, we calculate the similarity between model gradients contributed by participating users and the total server model. Low similarity indicates unreliability, effectively filtering out legitimate users while reducing attack threats posed by malicious actors.

The proposed method measures the contribution level of different customers in a gradient-based aggregation model. It uses cosine similarity to calculate the similarity between the client model and the aggregation model, and uses this similarity to quantify the contribution level of the client.

The loss function of a simple linear regression model is calculated $L(w,b)$ $L(w,b)$ by formula (1). The loss function $\frac{1}{N}\sum_{N=1}^{i}$ measures the difference $\frac{1}{N}\sum_{N=1}^{i}$ between the model's predicted values and actual values. represents the average value across all samples,

This represents N N the number y_i y_i i of i samples. It denotes $f(w_{xi}+b)$ the i $f(w_{xi}+b)$ true label w of i the i-th sample b. is the $(y_i - f(w_{xi}+b))^2$ w model's y_i prediction b for $f(w_{xi}+b)$ $(y_i - f(w_{xi}+b))^2$ the i-th sample y_i, where $f(w_{xi}+b)$ are the weights of the features and represent the deviation. This indicates the difference between the actual value and the model's predicted value. It is commonly referred to as the mean squared error (MSE) loss function.

$$L(w,b)=\frac{1}{N}\sum_{i=1}^{N}(y_i - f(w_{xi}+b))^2 \ L(w,b)=\frac{1}{N}\sum_{i=1}^{N}(y_i - f(w_{xi}+b))^2 \tag{1}$$

A gradient is a vector whose components represent partial derivatives of the function $u(x,y)$ $u(x,y)$ with ∇u respect g ∇u to each variable g. x For two-dimensional $\frac{\partial u}{\partial x}i$ vector x functions $\frac{\partial u}{\partial y}j$, the $\frac{\partial u}{\partial x}i$ gradient is expressed $\frac{\partial u}{\partial y}j$ as formula (2). The gradient field represents the sum of its components: the y-component corresponds to the y-directional derivative, while the x-component corresponds to the x-directional derivative. This gradient vector characterizes the function in a two-dimensional plane

$u(x, y)$ $u(x, y)$ The rate and direction of change.

$$g = \frac{\partial u}{\partial x} i + \frac{\partial u}{\partial y} j \quad g = \frac{\partial u}{\partial x} i + \frac{\partial u}{\partial y} j \tag{2}$$

To achieve the minimum loss function in neural networks, gradient descent is employed to compute the gradient of the weight matrix's loss function. This process enables iterative updates of weights, ultimately yielding the final weight matrix. Therefore, the similarity calculation between gradients corresponds directly to the similarity measurement of the weight matrix. The computation of the parameter matrix can be performed using the cosine similarity formula (3).

$$\cos\theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times b_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3}$$

The cosine similarity formula measures the similarity between two vectors B A A by B calculating the cosine of their angle. A cosine value close to 1 indicates high similarity, while values near-1 signify dissimilarity. Cosine similarity approaching zero θ suggests no correlation $A \cdot B$ θ between A vectors B $A \cdot B$ $\|A\|$. Here $\|B\|$ A, B represents B A the $\|A\|$ $\|B\|$ angle B_i A_i between A A two B B vectors; denotes A_i the B_i dot product A of B vectors and; and are the norms (or lengths) of vectors and; and are the components of vectors and. In gradient-based learning, high-quality client gradients typically show strong similarity to the overall model gradient, indicating minimal model loss. Conversely, low-quality client gradients demonstrate lower similarity to the overall model gradient, suggesting greater model loss.

The customer contribution level serves to identify and eliminate potential malicious users, as their contributions are typically minimal. By comparing each customer's similarity with the overall model, participants exhibiting low or negative contributions are flagged as unreliable or malicious and removed from the system. During each model update cycle, participants with high similarity to the overall model are excluded from subsequent rounds beyond a certain threshold. These removed participants cease to participate in model training, thereby reducing the threat of malicious user attacks during the training process

3.2 System design

The specific implementation steps are as follows:

(1) The server sends the task and initial model to the client.

(2) The client sends the model parameters to the server.

(3) After receiving the parameters sent by all clients, the server performs fusion according to the average value of the user model (the weight of each client excluding malicious clients).

(4) After obtaining the global model of the server, the cosine similarity algorithm was used to calculate the similarity between all clients and the global model. This similarity score was then used to determine each client's contribution level (highly contributing clients were considered reliable users, while those with low contributions were deemed unreliable users), and participants with unreliable contributions were subsequently removed.

(5) The server sends the summary model to all clients for further training.

(6) Repeat steps (2) to (5) until the server model converges.

4. Conclusion

This paper proposes a novel malicious node detection framework for Manufacturing Control Systems (MCS) networks. By integrating federated learning technology, the framework establishes an efficient and secure joint learning architecture. It introduces a data security protection method based on user contribution levels, which effectively filters legitimate users by evaluating their contributions to the overall model. This approach reduces risks of user privacy data leakage and minimizes the threat posed by malicious users.

References

[1] Zhao B, Liu X, Chen W-N. When crowdsensing meets federated learning: Privacy-preserving mobile crowdsensing system [J]. arXivpreprint arXiv: 210210109, 2021.

[2] Zhang M, Chen S, Shen J, Zhang M, Chen S, Shen J, et al. Privacyeafl: Privacy-enhanced aggregation for federated learning in mobile crowdsensing [J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5804-5816.

[3] Fang Chen, Guo Yuanbo, Wang Yifeng, Hu Yongjin, Ma Jiali, Zhang Han, and Hu Yangyang. Privacy-preserving edge computing methods based on blockchain and federated learning [J]. Journal of Communications, 2021,42(11):28-40.

[4] Gao Sheng, Yuan Liping, Zhu Jianming, et al. A blockchain-based privacy-preserving asynchronous federated learning [J]. Science in China: Information Science, 2021,51(10):20.

[5] Xie W ,Wang Y ,Boker M S , et al. PrivLogit: Efficient Privacy-preserving Logistic Regression by Tailoring Numerical Optimizers.[J].CoRR,2016,abs/1611.01170

[6] S H ,K H ,M C , et al.Necrotising Scleritis and Peripheral Ulcerative Keratitis Associated with Rheumatoid Arthritis Treated with Rituximab.[J].Klinische Monatsblatter fur Augenheilkunde,2017,234(4):567-570.

[7] MA J, CHEN L, XU J, et al. FedCrow: Federated Learning-Based Data Privacy Preservation in Crowd Sensing [J/OL]. Applied Sciences, 2024,14(11):4788.