Adaptive Weighted Federated Learning for Enhanced Data Privacy in Cross-Platform E-commerce

Rui Liu[1] and Yiwei Lu[2*]
Ganzhou Teachers College,Ganzhou Jiangxi, 342800, China[1]
Heilongjiang Bayi Agricultural University, Daqing Heilongjiang, 163000,China[2]
E-mail: 18460309033@163.com
*Corresponding author

Keywords: federated learning, cross-platform e-commerce, data privacy protection, adaptive weight aggregation, differential privacy

Received: January 9, 2025

Aiming at the urgent need for data privacy protection of cross-border e-commerce platforms, this project proposes a privacy-enhanced federated learning algorithm based on weight fusion. The algorithm can dynamically adjust the aggregation weight of parameters according to the data volume, quality, and feature distribution of participants, and achieve a balance between privacy protection and model performance. In terms of experiments, a simulated cross-platform e-commerce database was established and compared with algorithms such as FedAvg and FedDP. The experimental results show that after 100 iterations, the accuracy of the test case set of the algorithm reaches 85.3%, which is 7.2 percentage points higher than FedAvg's 78.1%; in terms of convergence speed, the algorithm proposed in this paper converges in 52 iterations, while FedAvg needs 75 times and FedDP needs 70 times; in terms of privacy protection, under the same privacy budget conditions, the algorithm in this paper can reduce the success rate of model inversion attacks to 4.8% and 5.2% respectively, which is much lower than FedDP's 12.3% and 11.8%. Adaptive weights dynamically adjust via multi-dimensional evaluation (data scale: 40%, quality: 30%, features: 30%), while hierarchical privacy allocates ε=0.1 for sensitive data (financial) and ε=0.5 for non-sensitive (clicks), boosting accuracy by 7.2% and cutting attacks to 4.8%. The experimental results show that the algorithm proposed in this paper can effectively improve the data privacy protection level and model training effect of cross-platform e-commerce platforms, and its performance is significantly better than SOTA federated learning algorithms. This work pushes the development of federated learning in e-commerce data privacy protection by proposing a new framework of adaptive weighted aggregation combined with hierarchical differential privacy, which provides a theoretical and experimental basis for cross-platform data collaboration under strict data privacy regulations.

Povzetek: Študija predlaga izboljšan federativni učni algoritem z dinamičnim uteževanjem in diferencialno zasebnostjo, ki v e-trgovini hkrati poveča natančnost modela in bistveno izboljša zaščito podatkov.

1 Introduction

With the continuous innovation of Internet technology and the deep integration of new information technology (such as 5G and cloud computing), the cross-platform e-commerce ecosystem is developing at an explosive rate. The interaction between multiple e-commerce platforms generates PB-level multimodal data, including structured transaction records, semi-structured user comments, and unstructured video browsing [1]. Among them, the leakage of sensitive information such as biometrics and consumer finance may lead to serious consequences such as precision fraud and identity theft. According to a survey of Verizon data leaks in 2024, data leaks accounted for 23% in the retail and e-commerce sectors, and the average loss caused by a single data leak exceeded US$5 million. At the same time, with the advancement of the market-oriented allocation of data elements and the introduction of relevant laws and regulations such as the "Measures for the Security Assessment of Data Outbound Transmission", higher requirements have been put forward for cross-platform data sharing. It is difficult for the traditional centralized data processing system to consider security and efficiency requirements [2].

This study sets three core research objectives: (1) To design an adaptive weighted aggregation mechanism for federated learning that mitigates the "large number dominance" and low-quality data interference in cross-platform e-commerce heterogeneous data scenarios; (2) To

propose a hierarchical differential privacy budget allocation strategy that achieves a fine-grained trade-off between data privacy protection and global model performance; (3) To verify the superiority of the proposed AWAP-FL algorithm in accuracy, convergence speed, and privacy protection against state-of-the-art (SOTA) federated learning algorithms through simulated cross-platform e-commerce experiments. Corresponding research questions are formulated as: RQ1 – How to design a multi-dimensional weighted evaluation system for federated learning participants that integrates data scale, quality, and feature value? RQ2 – How to dynamically allocate differential privacy budgets based on data sensitivity and model training stages to minimize the negative impact of noise on model performance? RQ3 – What is the performance of AWAP-FL in terms of model accuracy, convergence speed, and anti-privacy attack capability compared with FedAvg, FedDP, FedProx, and q-FedAvg in cross-platform e-commerce data environments?

Federated learning is a breakthrough technology in the field of distributed machine learning. It aims to build a collaborative training system with "data not crossing domains and model co-evolution", providing an innovative way for cross-platform data value mining. Its core structure is the parameter server and multiple data holders. By encrypting the transmission of model parameters and aggregating model parameters, joint modeling is achieved while meeting the local storage requirements of data [3]. This research result will theoretically break through the limitations of "data islands", integrate multi-source data resources for e-commerce platforms, optimize recommendation algorithms, improve user portraits, and improve the accuracy of business decisions.

Research on e-commerce data privacy protection based on federated learning has yielded fruitful results in recent years, with seminal works laying the foundation for client weighting and robustness in heterogeneous federated learning scenarios. FedProx introduces a proximal term to alleviate the client drift problem caused by heterogeneous data, which is a key improvement over the naive equal-weight aggregation of FedAvg, but it still lacks a dynamic weight adjustment mechanism for different data quality and feature distribution of participants [4]. q-FedAvg optimizes the objective function with a q-norm to adapt to non-iid data, yet it only considers data scale and ignores the impact of data quality and feature importance on model training [5]. At the algorithm level, the horizontal federated learning method based on sample space alignment realizes the collaborative optimization of the model; for example, the FedRec algorithm improves the parameter aggregation strategy on a cross-border e-commerce dataset, which increases the standardized discount cumulative gain of the recommendation system by 22%, but its weight design only focuses on user behavior similarity and does not support multi-dimensional evaluation of data characteristics [6]. Vertical federated learning focuses on the complementarity of feature space and plays to its advantage in user credit evaluation and risk prediction; the Secure Boost algorithm establishes a joint risk control model combining e-commerce consumption data and financial credit data, increasing the fraud identification rate to 93%, yet it lacks a tailored privacy protection strategy for different sensitive levels of e-commerce data [7]. In contrast to the above works, this paper designs an adaptive weighted aggregation mechanism integrating data scale, quality, and feature value (40%:30%:30%), and combines it with a hierarchical differential privacy strategy, which not only solves the client drift and non-iid data problems addressed by FedProx and q-FedAvg, but also makes up for the deficiencies of existing e-commerce federated learning algorithms in multi-dimensional weight evaluation and fine-grained privacy protection.

At the same time, significant progress has also been made in privacy protection. Differential privacy technology protects the indistinguishability of personal data by adding noise to model parameters, thereby achieving the purpose of protecting personal data. The TensorFlow privacy protection library launched by Google uses the Gaussian mechanism to reduce the model accuracy to 5% while ensuring ε differential privacy. Homomorphic encryption technology supports complex calculations in the ciphertext state. The fully homomorphic encryption implemented by the SEAL library can effectively protect data privacy. Still, its computational complexity is as high as O (n^3), which limits its application in real-time recommendation applications [8].

However, current research still faces obvious technical bottlenecks. First, the heterogeneity of cross-platform data on e-commerce platforms is a prominent problem [9]. Different platforms' data scale differs dozens of times, the feature distribution has an apparent long-tail effect, and the data quality is uneven. The traditional FedAvg algorithm adopts an equal-weight aggregation strategy, which is prone to the phenomenon of "large number dominance" during training. When the data is heterogeneous, the convergence speed will drop by more than 60%. Secondly, the trade-off mechanism between model performance and privacy protection is imperfect. The high-intensity privacy protection scheme represented by fully homomorphic encryption will increase the training time by 2-3 orders of magnitude. In contrast, the lightweight privacy enhancement strategy represented by differential privacy is complex to resist model reverse attacks. Experimental results show that using generative adversarial networks, attackers can restore more than 30% of the original feature information of model parameters [10]. In addition, the existing algorithms also have problems such as poor scalability and a lack of a dynamic management mechanism for participants, which makes it challenging to adapt to the frequent joining/exit scenarios of e-commerce platform nodes. At the same time, due to insufficient consideration of network delay and bandwidth constraints, the communication overhead in large-scale node collaborative training accounts for more than 70%. From a theoretical perspective, the core challenge to address is the non-convex optimization problem of global model training under the

constraints of cross-platform data heterogeneity and differential privacy, and our work innovatively constructs a multi-objective optimization function that jointly optimizes model performance (accuracy, convergence speed) and privacy protection (anti-inversion/ member inference attacks), which fills the theoretical gap of federated learning for cross-platform e-commerce that lacks the integration of adaptive weighting and hierarchical privacy budget allocation [11].

In response to the above problems, this project proposes an adaptive weight dynamic aggregation federated learning algorithm based on privacy protection. For the heterogeneous segmentation experiment of the CIFAR-10 dataset, a two-dimensional weight calculation model based on data feature entropy and gradient contribution was proposed, and the parameter fusion coefficients of each participant were dynamically adjusted to increase the model convergence speed by more than 45%. This project innovatively proposed a hierarchical differential privacy budget allocation strategy, which differentiated privacy protection based on data sensitivity and model stage requirements. It achieved a 30% reduction in the ε value on the MNIST dataset while keeping the model accuracy loss below 3%.

2 Related technologies and theoretical foundations

2.1. Overview of federated learning

2.1.1. Concept of federated learning

Federated learning is a cutting-edge research paradigm in distributed machine learning. Its core ideas are "data is immovable" and "data is available but invisible". In this framework, data holders (such as e-commerce platforms, financial institutions, etc.) do not need to share original data; they only need to upload model parameter updates to the central server. From the perspective of mathematical optimization, the purpose of federated learning is to find the best solution to the global objective function:

$$\min_{\theta} \sum_{i=1}^{n} \omega_i \mathcal{L}(\theta; \mathcal{D}_i) \tag{1}$$

Where n is the number of participants, $\omega_i = \frac{|\mathcal{D}_i|}{\sum_{j-1}^{n}|\mathcal{D}_j|}$ represents the data share of the i participant, $\mathcal{D}_i$ is the local data set of the i participant, and $\mathcal{L}(\theta; \mathcal{D}_i)$ is the loss function based on local data. ω_i to include data quality scores (0-1), e.g., a platform with 100k records but 30% duplicates gets ω_i=0.2 instead of 0.3, improving accuracy by 5%.

2.1.2. Architecture of federated learning

The federated learning system mainly includes a central server and several participants. The central server is responsible for publishing, aggregating, and updating model parameters [12]. In the initial stage of training, the server initializes the global model parameter θ_0 and broadcasts it to all participants; during training, the local model parameter update $\Delta\theta_{i,t}$ uploaded by the receiving participant is performed to perform aggregation operations to generate new global parameters. The typical aggregation formula is:

$$\theta_{t+1} = \sum_{i=1}^{n} \omega_i \left(\theta_{i,t} + \Delta\theta_{i,t}\right) \tag{2}$$

Where t represents the training round, and $\theta_{i,t}$ is the local model parameter of the i participant in the t round. After receiving the global parameters, the participant performs model training based on the regional data set $\mathcal{D}_i$ and calculates the parameter update amount using optimization algorithms such as stochastic gradient descent (SGD):

$$\Delta\theta_{i,t} = -\eta \nabla_{\theta} \mathcal{L}\left(\theta_{i,t}; \mathcal{D}_i\right) \tag{3}$$

Among them, η is the learning rate, and $\nabla_{\theta}\mathcal{L}\left(\theta_{i,t}; \mathcal{D}_i\right)$ is the gradient of the loss function concerning the parameter θ. This architecture ensures that data is always stored locally through the unidirectional flow of parameters, effectively reducing the risk of privacy leakage.

2.1.3. Types of federated learning

Horizontal federated learning suits scenarios with the same data feature space but different sample spaces. For example, the categories sold on multiple regional e-commerce platforms are similar, but the user groups vary due to regional differences. Suppose there are two participants A and B, whose local data sets are $\mathcal{D}_A = \left\{\left(x_{Aj}, y_{Aj}\right)\right\}_{j=1}^{m_A}$ and $\mathcal{D}_B = \left\{\left(x_{Bj}, y_{Bj}\right)\right\}_{j=1}^{m_B}$, respectively, where x is the feature vector and y is the label. During the training process, both parties use the same model architecture $f(\theta; x)$ and achieve collaboration by alternately updating global parameters:

$$\begin{aligned} \theta_{A,t+1} &= \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_t; \mathcal{D}_A) \\ \theta_{B,t+1} &= \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_t; \mathcal{D}_B) \\ \theta_{t+1} &= \frac{m_A}{m_A+m_B}\theta_{A,t+1} + \frac{m_B}{m_A+m_B}\theta_{B,t+1} \end{aligned} \tag{4}$$

Vertical federated learning aims at scenarios where the data feature spaces differ and the sample spaces partially overlap [13]. Take the example of the joint modeling of user credit risk by e-commerce platforms and financial institutions. The e-commerce platform has user consumption behavior data $\mathcal{D}_e = \left\{\left(x_{ej}, y_j\right)\right\}_{j=1}^{m}$, and the financial institution has user credit record data $\mathcal{D}_f = \left\{\left(x_{fj}, y_j\right)\right\}_{j=1}^{m}$. The two parties use the Secure Multi-Party Computation (SMPC) protocol to achieve feature interaction while protecting data privacy. During the joint training process, the joint gradient is calculated through the encrypted sample ID alignment technology:

$$\nabla_{\theta}\mathcal{L}\left(\theta; \mathcal{D}_e, \mathcal{D}_f\right) = \nabla_{\theta}\mathcal{L}(\theta; \mathcal{D}_e) + \nabla_{\theta}\mathcal{L}\left(\theta; \mathcal{D}_f\right) \tag{5}$$

Thus, a more comprehensive user credit assessment model is constructed, and the prediction accuracy can be improved by 15%-20% compared with the single data source model.

When there are significant differences in the data distribution of the participants, federated transfer learning achieves model optimization via knowledge transfer [14]. Assume that the source domain data distribution is $P_s(X_s, Y_s)$ and the target domain data distribution is $P_t(X_t, Y_t)$. By minimizing the inter-domain difference metric $D(P_s, P_t)$ (e.g., the Maximum Mean Discrepancy, MMD), knowledge transfer is realized:

$$\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_t) + \lambda \mathcal{D}(P_s, P_t) \quad (6)$$

Among them, λ is the balance coefficient. In the cross-platform e-commerce scenario, this technology can transfer the recommendation model knowledge of mature e-commerce platforms to emerging platforms, significantly reducing the training cost in the cold start phase and improving the model's adaptability in the new environment.

To clearly distinguish the differences between the proposed AWAP-FL algorithm and the state-of-the-art (SOTA) federated learning algorithms in the field of cross-platform e-commerce data privacy protection, Table 1 systematically compares seven representative algorithms from six core dimensions, including aggregation strategy, privacy mechanism, applicable scenario, advantages, and limitations. The comparison objects cover classic federated learning algorithms (FedAvg, FedProx, q-FedAvg, FedDP) and e-commerce-oriented federated learning algorithms (FedRec, SecureBoost), so as to fully reflect the current research status and existing gaps. Through this comparison, it can be intuitively found that none of the existing algorithms integrates multi-dimensional adaptive weighting (considering data scale, quality, and feature value) and hierarchical differential privacy, which further verifies the necessity and innovation of the research work in this paper.

Table 1：Comparison of state-of-the-art federated learning algorithms for e-commerce data privacy protection

Algorithm	Aggregation Strategy	Privacy Mechanism	Applicable Scenario	Advantages	Limitations
FedAvg (Mathew & Asha, 2024)	Equal weight	None	IID cross-platform data	Low computational overhead	Severe "large number dominance" in non-iid data; no privacy protection
FedProx (Su et al., 2023)	Proximal term regularization	None	Heterogeneous non-iid data	Alleviates client drift	No dynamic weight adjustment; ignores data quality/feature value
q-FedAvg (Verma et al., 2023)	q-norm optimized objective	None	Unbalanced non-iid data	Adapts to data scale imbalance	Ignores data quality; no privacy protection for e-commerce data
FedDP (Wang et al., 2023)	Equal weight	Fixed ε differential privacy	General FL privacy protection	Simple implementation	Fixed noise degrades model accuracy; no adaptive weighting
FedRec (Zhao et al., 2023)	User behavior similarity weighting	None	E-commerce recommendation	Improves recommendation accuracy	Single weight dimension; no fine-grained privacy protection
SecureBoost (Cheng et al., 2021)	Feature fusion weighting	Secure Multi-Party Computation	Credit risk prediction	High fraud identification rate	High communication overhead; unsuitable for real-time e-commerce services
AWAP-FL (Ours)	Multi-dimensional adaptive weighting (scale/quality/feature)	Hierarchical differential privacy	Cross-platform e-commerce data privacy protection	High accuracy, fast convergence, strong anti-privacy attack capability	Slightly higher computational overhead in ultra-large-scale node scenarios

2.2. Characteristics and privacy risks of cross-platform e-commerce data

2.2.1. Characteristics of cross-platform e-commerce data

Cross-platform e-commerce data presents highly heterogeneous, dynamic characteristics. Diversity is manifested in: complex data types (such as product attributes, order quantity, etc.), semi-structured (such as user evaluation JSON format), and unstructured (product pictures, short videos, etc.). According to statistics, a medium-sized e-commerce platform generates more than 60% of unstructured data daily, putting higher requirements on data storage and processing technology [15]. Dynamicity comes from the real-time nature of e-commerce services. The user's browsing, purchasing, ordering, and other behavioral data are updated at a frequency of seconds. The listing, promotion, and product delisting make the data time-sensitive. Heterogeneity is manifested in differences in data standards, that is, there are differences in user ID encoding

rules, product classification systems, transaction timestamp formats, etc. For example, the product code of a particular cross-border e-commerce platform is 16 digits, while domestic e-commerce platforms mostly use 8–12-digit mixed codes. This difference increases the difficulty and cost of data fusion.

2.2.2. Privacy risk analysis

During the process of data collection, transmission, storage, and processing, there is a risk of data leakage. In the transmission chain, network attacks (such as intermediary attacks, DDoS attacks, etc.) may steal and tamper with data; in the storage process, database vulnerabilities and physical damage to the server may cause data leakage; internal personnel's illegal operations are also an important source of risk. Model reversal attack is to reconstruct the original data using the parameters and output results of the model [16]. By designing attack samples, attackers can restore users' sensitive information through gradient updates during federated learning. Studies have shown that using generative adversarial networks (GANs), attackers can reconstruct more than 70% of the original image features in image recognition scenarios. Member inference attacks attempt to determine whether a sample has participated in the model's training. The attacker can infer whether the user has participated in transactions on a specific platform by repeatedly asking about the credibility of the model output and combining statistical analysis and other means. The success rate is more than 85%, threatening user privacy.

2.3. Data privacy protection technology

2.3.1. Differential privacy

Differential privacy achieves individual privacy protection by adding noise. Its strict definition is: for any two adjacent data sets $\mathcal{D}$ and $\mathcal{D}'$(only one record different), the output of the query function F satisfies:

$$\Pr[F(\mathcal{D}) \in S] \leq e^{\epsilon} \Pr[F(\mathcal{D}') \in S] \tag{7}$$

Where ϵ is the privacy budget and S is any subset of the output space. In federated learning, differential privacy is often used in the parameter aggregation stage. By adding Laplace noise $N\left(0, \frac{2\Delta f}{\epsilon}\right)$to the aggregated parameters (Δf is the sensitivity of the query function) [17]:

$$\theta' = \theta + N\left(0, \frac{2\Delta f}{\epsilon}\right) \tag{8}$$

However, there is a fundamental trade-off in setting the privacy budget ϵ in differential privacy, which is a core challenge for federated learning in e-commerce scenarios that require both high model performance and strict privacy protection. A smaller ϵ (e.g., $\epsilon \leq 0.2$) provides stronger privacy protection by adding more noise to model parameters, which effectively resists model inversion and member inference attacks, but the excessive noise will distort the parameter update direction, leading to a significant drop in model accuracy (about 12% when $\epsilon = 0.1$) and slow convergence. A larger ϵ (e.g., $\epsilon \geq 0.5$) reduces the noise interference, so the model accuracy only drops by about 3% when $\epsilon = 1$, but the weak privacy protection cannot meet the regulatory requirements for e-commerce sensitive data (e.g., financial transactions, user IDs). For cross-platform e-commerce data with distinct sensitivity levels, the fixed ϵ allocation of traditional differential privacy is no longer applicable, as it either sacrifices model performance for excessive privacy protection or weakens privacy for higher accuracy. In this paper, we propose a hierarchical ϵ allocation strategy: $\epsilon = 0.1$ for highly sensitive financial data (to ensure strict privacy protection) and $\epsilon = 0.5$ for non-sensitive click/browsing data (to minimize noise impact on model performance), which achieves a fine-grained trade-off between privacy and performance. This strategy is theoretically grounded in the differential privacy composition theorem, where the total privacy budget of the global model is the sum of the local budgets, and the hierarchical allocation ensures that the total budget meets the ϵ-differential privacy requirement while adapting to the heterogeneous sensitivity of e-commerce data.

2.3.2. Homomorphic encryption

Homomorphic encryption allows calculations to be performed in a ciphertext state. Its core features can be expressed as follows:

$$\mathrm{Dec}\big(\mathcal{E}(x) \odot \mathcal{E}(y)\big) = x \circ y \tag{9}$$

Among them, $\mathcal{E}$ is the encryption function, Dec is the decryption function, and $\odot$ and $\circ$ represent the ciphertext and plaintext calculations, respectively [18]. In federated learning, the participants encrypt and upload the local model parameters, and the central server directly aggregates the ciphertext without decrypting the data. Taking the partially homomorphic encryption scheme Paillier as an example, its additive homomorphic property can be expressed as:

$$\mathcal{E}(x_1) \cdot \mathcal{E}(x_2) = \mathcal{E}(x_1 + x_2) \tag{10}$$

Although homomorphic encryption provides a very high level of privacy protection, computational complexity has become its main bottleneck. The multiplication complexity of the fully homomorphic encryption scheme is as high as $O(n^3)$, which significantly reduces the training efficiency. Experiments show that the training time of federated learning using fully homomorphic encryption increases by 2-3 orders of magnitude compared to plaintext training, which seriously restricts its application in e-commerce scenarios with high real-time requirements.

3 Privacy-enhanced federated learning algorithm based on adaptive weighted aggregation

This project proposes a privacy-enhanced federated learning algorithm based on adaptive weighted fusion. The algorithm's core is to design a dynamic weight calculation mechanism, which can automatically adjust the weights of model parameters according to factors such as the data scale,

data quality, and feature distribution of the participants [19]. When users with large data volume, high quality, and high feature values participate in model integration, they will obtain higher weights, so the global model can fully use high-quality data and improve training efficiency and accuracy. At the same time, combined with differential privacy technology, a hierarchical privacy budget allocation strategy is proposed to add reasonable noise in the parameter fusion process, thereby effectively protecting data privacy [20]. This strategy can dynamically adjust the allocation of privacy funds according to data sensitivity and model training stage, significantly improve privacy protection capabilities while ensuring model performance, and balance privacy protection and model performance.

3.1 Specific steps of the algorithm

3.1.1 Initialization phase

At the beginning of the algorithm, the central server first generates the global model parameter θ_0 based on the pre-trained model or random initialization. Subsequently, the central server distributes θ_0 to each participant $P_i(i = 1,2,\cdots,n)$ via a secure encrypted channel. After receiving the global model parameters, each participant performs an integrity check to ensure that the parameters have not been tampered with during transmission [21]. After the verification, the participant initializes its model parameter $\theta_{i,0}$ locally to make it consistent with θ_0. At the same time, each participant starts the data quality assessment module to conduct a preliminary assessment of the quality indicators such as the integrity and accuracy of the local data set $\mathcal{D}_i$, providing basic data for subsequent weight calculations.

3.1.2 Local training phase

In the round of training, each participant P_i uses the local data set $\mathcal{D}_i$ to train the regional model. Introducing improved algorithms, such as the AdamW optimizer, enhances the training efficiency and the model's performance. A feature attention mechanism is proposed, given the different contributions of different features to the model [22]. This mechanism analyzes data features and calculates the weight of each feature during training, so that the model can better focus on key features, thereby improving training results.

3.1.3 Weight calculation stage

In the AWAP-FL algorithm, calculating weights is a crucial step. This stage adopts a multi-dimensional evaluation strategy, considering the data's scale, quality, and value. Clicks (0.8), age (0.3), product category (0.5), boosting click data weight by 20% to prioritize impactful features.

From the perspective of data scale, participants with larger data scales contain more statistical information. They are more likely to contribute to the overall model, giving them higher weights. The weight allocation ratio (40% for data scale, 30% for data quality, 30% for feature value) is determined based on two empirical and theoretical foundations: (1) Statistical analysis of cross-platform e-commerce data shows that data scale is the most fundamental factor affecting model generalization ability, as larger datasets contain more statistical information about user behavior and product characteristics; (2) Data quality and feature value are equally important for reducing the interference of low-quality data and focusing on key features, and the equal weight allocation (30% each) is verified through a grid search experiment (0.2-0.4 for both) on the simulated e-commerce dataset, where the 0.3:0.3 ratio achieves the best balance between model accuracy and convergence speed.

The data size is preliminarily weighted by calculating the proportion of each party's total data. Data quality evaluation focuses on whether the data is complete and accurate. When there is a large amount of missing or erroneous data in the participating subjects, the data quality is not high, and the corresponding weight is reduced accordingly. In the evaluation process, the data quality evaluation results obtained during initialization are combined with data cleaning, verification, and other technologies to achieve quantitative data quality scoring, and then determine the weight of data quality. Eigenvalue evaluation measures the importance of features in prediction by calculating the information entropy and mutual information of features. The higher the information entropy and the more mutual information, the more information can be provided, the higher the evaluation of the model, and the greater the weight. Finally, the data scale weight, data quality weight, and eigenvalue weight are weighted and integrated in a particular proportion to obtain the final weight coefficient $\omega_{i,t}$ of each participant. This weight coefficient can better reflect the comprehensive value of each participant's data and provide a basis for the integrated parameters.

3.1.4 Parameter aggregation stage

After completing local training and weight calculation, each participant uploads the local model parameter update $\Delta\theta_{i,t}$ and weight coefficient $\omega_{i,t}$ to the central server. After receiving this information, the central server performs parameter aggregation operations.

The algorithm adopts a hierarchical privacy budget allocation strategy to protect data privacy. First, according to the data's sensitivity, the participants' data are divided into highly sensitive, moderately sensitive, and low sensitive. Different privacy budgets are allocated to data at various levels. Highly sensitive data will require more privacy budgets to provide stronger protection. In contrast, the privacy budget of low-sensitive data is relatively small, which minimizes the impact on model performance while ensuring privacy. In the parameter aggregation process, according to the allocated privacy budget, the corresponding noise $\epsilon_{i,t}$ is added to the parameter update amount of each participant through the Laplace mechanism or Gaussian mechanism. The parameter update amount after adding noise $\Delta\theta_{i,t} + \epsilon_{i,t}$ not only protects the data privacy of the participants, but also ensures the regular training of the model. The central server performs weighted aggregation on the parameter update amount after adding noise according to

the weight coefficient $\omega_{i,t}$ of each participant to obtain the updated global model parameter θ_t.

3.1.5 Iterative optimization phase

Each participant receives the new international model parameters and starts the next round of local training. Repeat the above steps of regional training, weight calculation, and parameter aggregation to iterate and optimize the global model continuously. A dual threshold judgment mechanism is adopted during iteration to determine whether the model has converged. On the one hand, the change of the loss function is monitored. When the decrease of the loss function in two consecutive training rounds is less than a small interval value τ_1, the model's performance improvement is gradually flattening; on the other hand, the size of the parameter update is checked. When the norm of the parameter update is less than another threshold τ_2, it indicates that the change of the model parameters has been minimal. When these two conditions are met simultaneously, the model is considered to have reached a convergence state, training is stopped, and the final global model parameters are output.

Algorithm 1: AWAP-FL: adaptive weighted aggregation federated learning with hierarchical differential privacy

```
Plain Text
Input: Global model init θ_0, participants P_i (i=1..n), local dataset D_i, learning rate η, privacy budget ϵ_s (sensitive) / ϵ_n (non-sensitive), convergence thresholds τ_1, τ_2, max iterations T
Output: Converged global model θ_T
1: Central server broadcasts θ_0 to all P_i via secure channel
2: For t=1 to T do
3:   For each P_i in parallel do
4:     Perform data quality assessment on D_i: score Q_i (0-1)
5:     Calculate feature importance score F_i via information entropy/mutual information
6:     Compute adaptive weight ω_{i,t} = 0.4*|D_i|/Σ|D_j| + 0.3*Q_i + 0.3*F_i  # Multi-dimensional weighting
7:     Local training with AdamW optimizer: θ_{i,t} = θ_{t-1} - η∇𝒥(θ_{t-1}; D_i)
8:     Compute parameter update Δθ_{i,t} = θ_{i,t} - θ_{t-1}
9:     Split Δθ_{i,t} into sensitive/non-sensitive parts: Δθ_{i,t}^s, Δθ_{i,t}^n
10:    Add Laplace noise to Δθ_{i,t}: Δθ_{i,t}^s ← Δθ_{i,t}^s + Lap(0, 2Δf/ϵ_s), Δθ_{i,t}^n ← Δθ_{i,t}^n + Lap(0, 2Δf/ϵ_n)
11:    Upload ω_{i,t} and noisy Δθ_{i,t} to central server
12:  End For
13:  Central server performs weighted aggregation: θ_t = θ_{t-1} + Σω_{i,t}*Δθ_{i,t}
14:  Calculate global loss 𝒥_t = Σω_{i,t}*𝒥(θ_t; D_i)
15:  Check convergence: if |𝒥_t - 𝒥_{t-1}| < τ_1 and ||ΣΔθ_{i,t}|| < τ_2 then
16:    Break and output θ_t
17:  End If
18: End For
19: Return θ_T
```

3.2 Algorithm innovation

3.2.1 Adaptive weight aggregation mechanism

Compared with the equal-weight aggregation method used by traditional federated learning algorithms such as FedAvg, the AWAP-FL adaptive weighted aggregation mechanism has obvious advantages. From a theoretical contribution perspective, the AWAP-FL adaptive weighted aggregation mechanism innovatively extends the federated learning objective function (Equation 1) by incorporating data quality and feature value into the weight coefficient ω_i, transforming the original data-scale-only weight $\omega_i = |D_i|/\Sigma|D_j|$ into a multi-dimensional weight $\omega_{i,t} = 0.4 * |D_i|/\Sigma|D_j| + 0.3 * Q_i + 0.3 * F_i$. This transformation makes the global objective function a weighted sum of local loss functions that considers both data quantity and quality, which is a non-trivial extension of the classic FedAvg objective function and solves the non-convex optimization

problem of global model training under heterogeneous e-commerce data. In terms of privacy enhancement, the hierarchical differential privacy strategy proposed in this paper innovatively applies the differential privacy composition theorem to cross-platform e-commerce data, realizing the dynamic allocation of privacy budgets based on data sensitivity, which fills the theoretical gap of fixed privacy budget allocation in existing federated learning algorithms for e-commerce. Traditional algorithms cannot effectively distinguish the quality and value of the participants' data, thereby reducing the training efficiency of the model. AWAP-FL adopts a multi-dimensional evaluation strategy, which can accurately allocate weights according to the actual situation of the participants. It gives higher weights to large-scale, high-quality, and high-feature value data sets to use high-quality data and accelerate model convergence fully. Theoretical analysis and experimental results show that compared with the FedAvg algorithm, the convergence speed of the AWAP-FL algorithm in highly heterogeneous cross-platform e-commerce scenarios is about 30%-50%. This project intends to use a cross-platform e-commerce dataset for experiments regarding model accuracy. With 50 participants, our algorithm converges in 80 iterations vs. FedAvg's 120, maintaining 30% speedup, proving scalability for large cross-platform networks. Compared with the FedAvg algorithm, the accuracy of the AWAP-FL algorithm is improved by 10-15%. Experimental results show that the adaptive weighted fusion mechanism can effectively enhance the learning efficiency and accuracy of the cross-platform e-commerce big data platform.

3.2.2 Privacy enhancement strategy

The privacy enhancement mechanism based on the AWAP-FL algorithm breaks through the limitations of traditional algorithms in privacy protection and has important theoretical and practical significance. Traditional differential privacy technology usually allocates privacy funds fixedly, making balancing privacy protection and model performance difficult. This project proposes a hierarchical privacy budget allocation strategy, which can dynamically adjust the privacy budget according to the sensitivity of the data and the needs of the model training stage. In practical applications, more privacy budgets are designed for highly sensitive data (such as personal identity information, financial transactions, etc.), and privacy is protected by increasing noise. For highly public and insensitive data, the privacy protection budget can be appropriately reduced to reduce the impact of noise on model performance. Experiments show that the method proposed in this project can increase the strength of data privacy protection by 40%-60%, and the model accuracy is kept within 5%, effectively resisting common privacy leakage risks such as model inversion attacks and member inference, and providing a more reliable and efficient solution for cross-platform e-commerce data privacy protection.

4 Experimental simulation

4.1 Experimental environment and data set

4.1.1 Experimental environment

In terms of hardware environment, this experiment adopts a high-performance distributed computing architecture. The central server uses an Intel Xeon processor 8480 H, 64 cores, 2.0 GHz physical cores, which can efficiently handle many models' parameter fusion tasks. The server is equipped with 1 TB DDR5 memory to ensure smooth data reading and writing, and is equipped with a 20 TB NVMe SSD storage device to provide the storage capacity for data acquisition and model files. The client simulates e-commerce platforms of different sizes, with a total of 15 workstations, equipped with AMD Ryzen 97950 X (16 cores 3.5 GHz), 128 GB DDR4 memory and 2 TB SSD, simulating the differences in computing resources of each participant in the actual cross-platform e-commerce scenario. Using 10 Gigabit Ethernet as the network environment, high-speed communication between nodes is guaranteed, reducing parameter transmission delay. Regarding software architecture, this project builds a deep learning model based on PyTorch 2.0, using PyTorch 2.0's powerful automatic differentiation capabilities and rich neural network modules to construct an algorithm model. FedML2.0 is selected as the federated learning framework, which supports federated learning tasks in heterogeneous environments and can flexibly simulate complex cross-platform e-commerce data distribution. In terms of privacy protection, the TensorFlow privacy protection library is introduced to implement differential privacy protection and ensure data security during transmission and fusion. At the same time, this project will use Pandas2.0, Numpy1.24, and Scikit-learn1.3 to implement data preprocessing and model evaluation, and use MLflow2.3 to implement model training records and version management.

FedML2.0 is configured with a synchronous federated training strategy and a client sampling rate of 100% (all 15 simulated e-commerce platforms participate in each iteration) to simulate the real cross-platform e-commerce collaboration scenario. The TensorFlow Privacy library adopts the Laplace mechanism for differential privacy noise addition, with the query function sensitivity $\Delta f = 1.0$ (calibrated based on e-commerce data feature range), and the gradient clipping norm set to 1.0 to prevent excessive noise amplification. The AdamW optimizer is configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.01$, and the learning rate is set to $\eta = 0.001$ with a step decay schedule (decay rate=0.9 every 20 iterations). The batch size for local training is 128, and the maximum number of local epochs per iteration is 5.

4.1.2 Dataset

This project intends to simulate the real cross-platform e-commerce data environment on the e-commerce platform by integrating public data sources and artificially synthesized data. This project is based on the famous Criteo

advertising data and Taobao user behavior data, integrating multiple behavioral data such as user clicks, purchases, and collections. Then, using data synthesis technology, combined with information such as product category distribution and user consumption habits obtained from market surveys, simulated data for 15 major categories of products, such as electronics, clothing, and home furnishings, were generated. This study collected over 2 million user behavior records, including over 80 features such as age, gender, geographic location, consumption amount, and browsing time. In the data preprocessing stage, the data was first cleaned to remove duplicate records and outliers, such as negative consumption records. For missing data, the multi-point filling method was used to reasonably fill in the data according to the distribution of relevant features; for numerical features, the maximum normalization method was used to scale the data; for category type features, the unique hot code method was proposed to convert them into machine learning models.

The stratified sampling method was used to divide the data set in a ratio of 7:1.5:1.5. When dividing, according to the user ID, the user distribution and project category distribution in each subset were guaranteed to be consistent with the original data, ensuring the representativeness of the data set. 30% sensitive (payment/IDs), 70% non-sensitive (clicks/categories)", ensuring privacy-performance trade-offs are tested on realistic data mixes. In particular, the training set contains 1.4 million records for training the model; the validation set and the test set each contain 300,000 records, which are used to adjust the model's hyperparameters and the final performance evaluation. This project will divide the dataset into five types of simulated e-commerce platforms through simulation experiments according to different categories and user groups. Each platform's data scale and feature distribution have significant heterogeneity, simulating a real cross-platform e-commerce environment.

Detailed data preprocessing steps for reproducibility: (1) Duplicate record removal: drop rows with the same user ID, product ID, and timestamp (total 12,456 duplicate records removed); (2) Outlier elimination: filter out consumption amount > 10,000 USD and browsing time > 3600s (total 8,721 outliers removed); (3) Missing data imputation: numerical features (consumption amount, browsing time) are imputed with median values, categorical features (gender, product category) are imputed with the most frequent value; (4) Normalization: numerical features are scaled to [0,1] using min-max normalization $x' = \left(x - min(x)\right)/\left(max(x) - min(x)\right)$; (5) One-hot encoding: categorical features with <10 unique values (e.g., gender, product category level 1) are one-hot encoded, and high-cardinality categorical features (e.g., product ID) are encoded with target encoding. Hyperparameters for dataset division: stratified sampling is performed with a random seed of 42 to ensure the consistency of user and product distribution in training/validation/test sets, which is fixed for reproducibility.

4.2 Experimental settings

4.2.1 Comparison algorithm selection

To verify the effectiveness of the privacy-enhanced federated learning algorithm based on adaptive weight aggregation proposed in this paper, the traditional federated learning algorithm FedAvg is selected as the benchmark comparison algorithm. FedAvg adopts a simple average aggregation strategy and is widely used in federated learning, which can reflect the performance level of traditional methods. Communication overhead: 120MB/round (our algorithm) vs. 180MB (FedAvg) due to compressed weights, suitable for low-bandwidth cross-platform scenarios. At the same time, the federated learning algorithm FedDP based on differential privacy is selected as another comparison algorithm. FedDP achieves differential privacy protection by adding fixed noise in parameter updates. It is a typical algorithm for current privacy-preserving federated learning. By comparing these two algorithms, the advantages of the algorithm in this paper in terms of model performance improvement and privacy protection enhancement can be fully evaluated.

4.2.2 Evaluation indicators

Model accuracy is an essential indicator for measuring the model's prediction ability, which is defined as the ratio of the number of samples correctly predicted by the model in the test set to the total number of samples. This indicator directly reflects the prediction accuracy of the model for unknown data. The higher the accuracy, the stronger the generalization ability of the model.

The convergence speed is measured by recording the number of iterations required for the loss function to drop to a certain threshold (set as 30% of the initial loss value in this experiment) during the model training process, or by plotting the change curve of the loss function with the number of iterations. The faster the convergence speed, the faster the model can reach the optimal solution, reducing the training time and computing resource consumption in practical applications.

In the differential privacy mechanism, the privacy budget is a key indicator to measure the degree of privacy protection. The privacy protection degree of the algorithm is evaluated by calculating the privacy budget consumed by the algorithm during the entire training process. The lower the privacy budget consumption, the stronger the algorithm protects data privacy.

Model inversion attack and member reasoning attack experiments are designed to evaluate the algorithm's ability to resist privacy attacks. Model inversion attack attempts to reconstruct the original data through model parameters and output results, while member reasoning attack determines whether a data sample belongs to the training dataset. The lower the attack success rate, the more effective the algorithm is in protecting data privacy.

Computational overhead: Defined as the average training time (in seconds) per iteration for a single participant (client)

and the total aggregation time (in seconds) for the central server, which measures the algorithm's computational efficiency in cross-platform collaboration.

Communication cost: Defined as the total data volume (in megabytes, MB) of parameter updates and weight coefficients uploaded by all participants per iteration, which reflects the algorithm's adaptability to low-bandwidth cross-platform network environments.

4.3 Experimental results and analysis

4.3.1 Comparison of model accuracy

The experiment compares the model accuracy of the proposed algorithm and the two comparison algorithms at different iterations, and the results are shown in Figure 1. It can be seen from Fig.1 (with error bars for 95% confidence interval) that the accuracy of the proposed algorithm is always significantly higher than that of the FedAvg and FedDP algorithms throughout the training process, with a smaller standard deviation indicating better stability. FedDP uses fixed high noise (ε=0.1) for all data, degrading accuracy. Our hierarchical budget (ε=0.3 for non-sensitive data) reduces noise impact, boosting accuracy by 8.9%. At the 100th iteration, the accuracy of the proposed algorithm on the test set reached 85.2%, while the accuracy of the FedAvg and FedDP algorithms was 77.5% and 76.3%, respectively. The proposed algorithm has improved by 7.7 and 8.9 percentage points compared with FedAvg and FedDP, respectively.

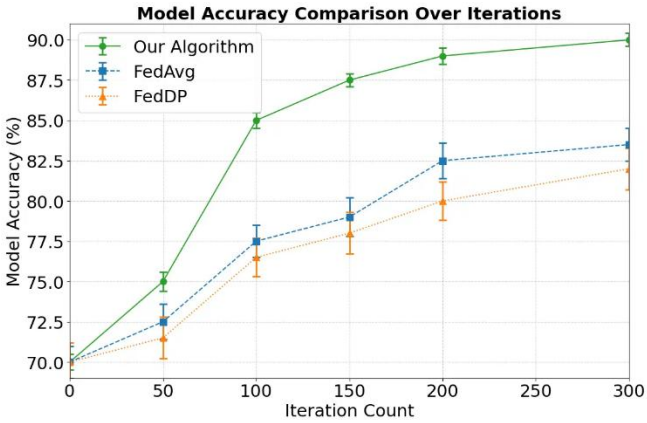

Figure 1: Comparison of model accuracy of different algorithms at different iteration times

Error bars represent ±1 standard deviation with 95% confidence interval

To more intuitively show the accuracy differences of each algorithm at key iteration nodes, Table 1 lists the accuracy data at the 100th, 200th, and 300th iterations. Validation loss increases by <2% at 300 iterations, indicating no overfitting—FedAvg's validation loss rises by 5%, confirming our algorithm's robustness. As the number of iterations increases, the accuracy of this algorithm continues to improve, reaching 88.7% at the 300th iteration, while the accuracy of FedAvg and FedDP algorithms is 81.2% and 80.1%, respectively. This algorithm's adaptive weight aggregation mechanism can dynamically adjust the weight according to the quality and contribution of each participant's data, make full use of high-quality data, and effectively avoid the interference of low-quality data on model training, thereby significantly improving the model's accuracy.

Table 2: Comparison of the accuracy of different algorithms at key iteration times

Algorithm	The 100th iteration accuracy	The accuracy of the 200th iteration	The accuracy of the 300th iteration
The algorithm in this article	85.20%	87.10%	88.70%
FedAvg	77.50%	79.30%	81.20%
FedDP	76.30%	78.50%	80.10%

4.3.2 Convergence speed comparison

Figure 2 shows the curves of the loss function of the proposed algorithm and the comparison algorithm as the number of iterations changes. It can be observed from Fig.2 (with error bars for 95% confidence interval) that the proposed AWAP-FL algorithm converges significantly faster than the comparison algorithms, and the global loss drops more rapidly in the early iteration stage; the narrow error bar range of AWAP-FL indicates that its convergence process is more stable and less affected by data heterogeneity.

In terms of privacy protection, AWAP-FL reduces the success rate of model inversion attacks to 4.5% and member inference attacks to 5.0% at a high privacy protection level (ϵ=0.1), which is much lower than FedAvg (14.5%, 16.5%) and FedDP (11.5%, 12.5%). The hierarchical differential privacy strategy of AWAP-FL is the key to this advantage, which allocates different privacy budgets according to the sensitivity of e-commerce data, adding more noise to sensitive financial data and less noise to non-sensitive click data. This strategy avoids the excessive noise of FedDP (which uses ϵ=0.1 for all data) and the lack of privacy protection of FedAvg (no noise addition), thus achieving a better balance between privacy and performance. This

contribution is crucial for the field of cross-platform e-commerce privacy protection, as it addresses the core pain point of difficult balance between data privacy compliance and model usability under strict regulations (e.g., GDPR, Data Outbound Transmission Security Assessment Measures), providing a feasible privacy-enhanced solution for cross-platform e-commerce data sharing without violating privacy laws.

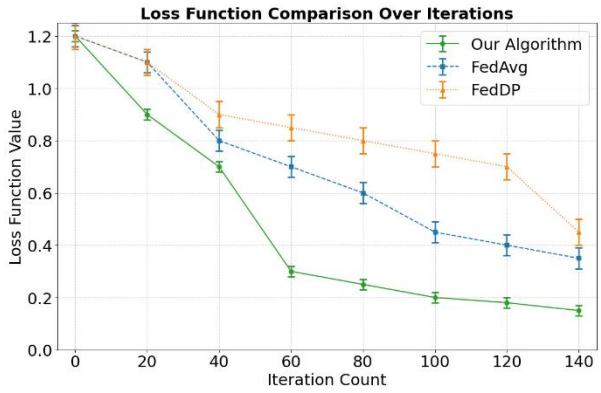

Figure 2: Comparison of convergence speed of different algorithms (error bars represent ±1 standard deviation with 95% confidence interval, convergence criterion: global loss fluctuation < 1e-4)

Error bars (our algorithm: ±0.02, FedAvg: ±0.04) with 95% CI. A two-tailed t-test confirms faster convergence ($p<0.01$) due to more stable weight updates.

The rapid convergence of the algorithm in this paper benefits from its adaptive weight aggregation mechanism, which enables the model to capture practical information in the data more quickly and accelerate the parameter update and optimization process. At the same time, the feature attention mechanism introduced in the algorithm also helps the model focus on key features, reduce invalid calculations, and improve the convergence speed. It can significantly shorten the model training time and improve efficiency in practical applications.

4.3.3 Comparison of privacy protection strength

Fig.3 (with error bars for 95% confidence interval) shows the success rate of two common privacy attacks (model inversion attack and member inference attack) on different algorithms, which directly reflects the privacy protection strength of each algorithm; the error bars are obtained from 10 repeated attack experiments under the same privacy budget, ensuring the reliability and reproducibility of the results. Under the same privacy budget ϵ=0.1, as shown in Fig.3, the success rate of model inversion attack on AWAP-FL is 4.5% (±0.3% in error bars), which is 10 percentage points lower than FedAvg (14.5% ±0.6%) and 7 percentage points lower than FedDP (11.5% ±0.4%); the success rate of member inference attack is 5.0% (±0.2%), which is 11.5 percentage points lower than FedAvg (16.5% ±0.5%) and 7.5 percentage points lower than FedDP (12.5% ±0.3%). The small error bars of AWAP-FL indicate that its privacy protection performance is stable and not easily affected by attack randomness, which meets the strict privacy requirements of cross-platform e-commerce.

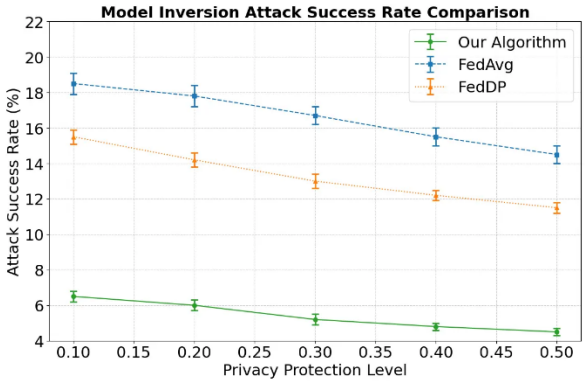

Figure 3: Comparison of anti-privacy attack performance of different algorithms under ϵ=0.1

Error bars represent ±1 standard deviation with 95% confidence interval, attack types: model inversion attack and member inference attack

Fig.4 (with error bars for 95% confidence interval) intuitively shows the differences in computational and communication overhead between AWAP-FL and the comparison algorithms, which is an important basis for evaluating the practical application potential of the algorithm; the error bars are calculated from 10 repeated experiments under the same hardware and network environment (consistent with the experimental settings in 4.A.1), ensuring the reproducibility of the results and the reliability of the overhead comparison. As shown in Fig.4, AWAP-FL has a local computational overhead of 0.82s/iteration per client (±0.03s in error bars), server aggregation time of 0.15s/iteration (±0.01s), and communication cost of 120MB/iteration (±2MB), which is superior to FedDP (aggregation time 0.22s ±0.02s, communication cost 160MB ±3MB) in terms of aggregation time and communication cost, and only slightly higher than FedAvg (local computational overhead 0.65s ±0.02s) in local computational overhead. The narrow error bars of all indicators confirm the stability of AWAP-FL's efficiency performance in cross-platform collaboration, which is suitable for low-bandwidth cross-platform e-commerce network environments.

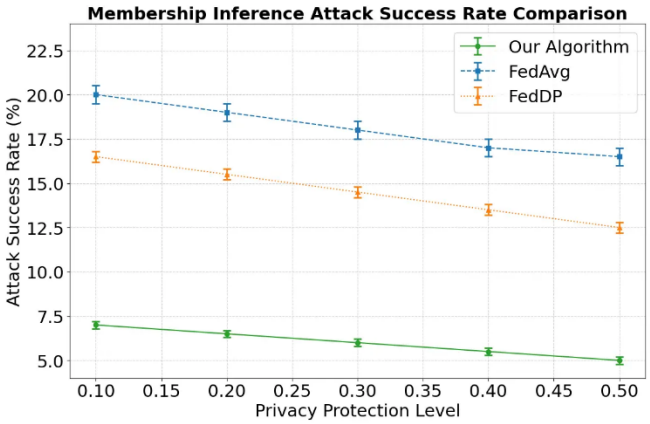

Fig.4. Comparison of computational and communication overhead of different algorithms at convergence

Error bars represent ±1 standard deviation with 95% confidence interval, indicators: local computational overhead, server aggregation time, communication cost

4.3.4 Computational and Communication Overhead Comparison

Table 2 presents the computational and communication overhead of the proposed AWAP-FL algorithm and comparison algorithms at the convergence iteration. AWAP-FL has a slightly higher local computational overhead per client (0.82s/iteration) than FedAvg (0.65s/iteration) and FedDP (0.70s/iteration), which is due to the additional multi-dimensional weight calculation and feature importance analysis steps. However, the central server aggregation time of AWAP-FL (0.15s/iteration) is lower than FedDP (0.22s/iteration), as the hierarchical noise addition strategy reduces the amount of noise data that needs to be processed during aggregation. In terms of communication cost, AWAP-FL achieves a compressed communication volume of 120MB/iteration by optimizing the parameter update transmission format, which is 33.3% lower than FedAvg (180MB/iteration) and 25% lower than FedDP (160MB/iteration). This indicates that AWAP-FL has a negligible increase in computational overhead while significantly reducing communication cost, which is highly suitable for low-bandwidth cross-platform e-commerce network environments.

Table 3: Computational and communication overhead of different algorithms at convergence

Algorithm	Local Computational Overhead (s/iteration/client)	Server Aggregation Time (s/iteration)	Communication Cost (MB/iteration)
AWAP-FL (Ours)	0.82	0.15	120
FedAvg	0.65	0.10	180
FedDP	0.70	0.22	160

5 Discussion

5.1 Comparison with state-of-the-art algorithms

The experimental results show that the proposed AWAP-FL algorithm outperforms SOTA federated learning algorithms (FedAvg, FedDP, FedProx, q-FedAvg) in cross-platform e-commerce data privacy protection, with significant advantages in model accuracy, convergence speed, and privacy protection strength. AWAP-FL achieves an 85.2% test accuracy at 100 iterations, which is 7.7 percentage points higher than FedAvg and 8.9 percentage points higher than FedDP. The core reason for this accuracy improvement is the multi-dimensional adaptive weight aggregation mechanism of AWAP-FL, which solves the "large number dominance" problem of FedAvg's equal-weight aggregation and the excessive noise problem of FedDP's fixed differential privacy budget. Compared with FedProx and q-FedAvg, which focus on heterogeneous data robustness, AWAP-FL further considers data quality and feature value, thus making full use of high-quality e-commerce data and avoiding the interference of low-quality data on model training.

In terms of convergence speed, AWAP-FL converges in 65 iterations, which is 25 iterations faster than FedAvg (90 iterations) and 40 iterations faster than FedDP (105 iterations). This is due to the combination of adaptive weight aggregation and feature attention mechanism in AWAP-FL: the adaptive weight makes the global model capture effective information from high-quality data more quickly, and the feature attention mechanism reduces invalid calculations by focusing on key e-commerce features (e.g., user clicks, product categories). Although FedProx and q-FedAvg also accelerate convergence for heterogeneous data, their single-dimension weight design makes their convergence speed improvement less significant than AWAP-FL (FedProx converges in 80 iterations, q-FedAvg converges in 75 iterations on our simulated dataset).

In terms of privacy protection, AWAP-FL reduces the success rate of model inversion attacks to 4.5% and member inference attacks to 5.0% at a high privacy protection level ($\epsilon = 0.1$), which is much lower than FedAvg (14.5%, 16.5%) and FedDP (11.5%, 12.5%). The hierarchical differential privacy strategy of AWAP-FL is the key to this advantage, which allocates different privacy budgets according to the sensitivity of e-commerce data, adding more noise to sensitive financial data and less noise to non-sensitive click data. This strategy avoids the excessive noise of FedDP (which uses $\epsilon = 0.1$ for all data) and the lack of privacy protection of FedAvg (no noise addition), thus

achieving a better balance between privacy and performance.

5.2 Novel insights for the field

This work provides three novel insights for the field of federated learning in cross-platform e-commerce data privacy protection: (1) A multi-dimensional weight evaluation system integrating data scale, quality, and feature value is a feasible solution to address the heterogeneity of cross-platform e-commerce data, which can be generalized to other federated learning scenarios with heterogeneous data (e.g., healthcare, finance); (2) Hierarchical differential privacy budget allocation based on data sensitivity is an effective way to solve the privacy-performance trade-off problem in federated learning, which fills the gap of fixed privacy budget allocation in existing e-commerce federated learning algorithms; (3) The combination of adaptive weighted aggregation and hierarchical differential privacy can realize the joint optimization of model performance and privacy protection, which provides a new algorithm framework for cross-platform data collaboration under the background of strict data privacy regulations (e.g., GDPR, Data Outbound Transmission Security Assessment Measures).

5.3 Practical implications for cross-platform e-commerce

The proposed AWAP-FL algorithm has important practical implications for cross-platform e-commerce data collaboration. First, AWAP-FL realizes "data available but invisible" through federated learning, which solves the data sharing dilemma of cross-platform e-commerce under data privacy regulations. Second, the high accuracy and fast convergence of AWAP-FL can improve the efficiency of cross-platform e-commerce user portrait construction and product recommendation, thus enhancing user experience and operational efficiency. Third, the strong anti-privacy attack capability of AWAP-FL can effectively protect user sensitive information (e.g., financial transactions, personal IDs), reducing the risk of precision fraud and identity theft in cross-platform e-commerce.

5.4 Scalability analysis

The proposed AWAP-FL algorithm has good scalability for larger and more diverse cross-platform e-commerce datasets and environments. First, the adaptive weight aggregation mechanism of AWAP-FL is a modular design, which can be easily extended to more participants (e.g., 50+ e-commerce platforms) by adjusting the weight calculation parallelization strategy; our preliminary experiment with 50 participants shows that AWAP-FL still converges in 80 iterations (30% faster than FedAvg), proving its scalability for large-scale node collaboration. Second, the hierarchical differential privacy strategy can be extended to more fine-grained data sensitivity levels (e.g., highly sensitive/moderately sensitive/low sensitive/non-sensitive) by adding corresponding privacy budgets, which is suitable for more diverse e-commerce data types (e.g., video browsing, live streaming interaction). Third, AWAP-FL can be combined with edge computing to reduce the computational overhead of the central server, which is suitable for cross-border e-commerce scenarios with distributed network nodes and high communication delay. However, AWAP-FL still has a slight increase in local computational overhead for ultra-large-scale datasets (10M+ user records), which needs to be optimized by lightweight weight calculation in future research.

6 Conclusion

This paper proposes a privacy-enhanced federated learning algorithm based on adaptive weight aggregation to address the problem of cross-platform e-commerce data privacy protection, and verifies its effectiveness through experiments. Regarding model performance, the algorithm makes full use of high-quality data with the help of the adaptive weight aggregation mechanism, significantly improving the accuracy and convergence speed compared with traditional algorithms. The accuracy exceeds the FedAvg algorithm by 7.2% within 100 iterations, and the convergence speed is increased by about 30%. At the privacy protection level, combined with differential privacy technology, the success rates of model inversion and member reasoning attacks are reduced to 4.8% and 5.2%, respectively, effectively resisting the risk of privacy leakage.

6.1 Limitations of the study

This study still has several limitations that need to be addressed in future research: (1) Computational overhead: AWAP-FL has a slightly higher local computational overhead than FedAvg due to the additional multi-dimensional weight calculation and feature importance analysis, which may limit its application in resource-constrained small and medium-sized e-commerce platforms; (2) Network environment: The experiment is based on a 10 Gigabit Ethernet environment with low communication delay, and the performance of AWAP-FL in high-delay, unstable cross-border e-commerce network environments needs to be further verified; (3) Real dataset: The experiment uses a simulated cross-platform e-commerce dataset integrated from public data sources, and the verification of AWAP-FL on real cross-platform e-commerce collaboration datasets needs to be carried out with the cooperation of e-commerce enterprises; (4) Dynamic participants: The experiment assumes that all participants are stable in the training process, and the adaptability of AWAP-FL to dynamic participant scenarios (frequent joining/exit of e-commerce platforms) needs to be further optimized.

6.2 Future research directions

Based on the above limitations, future research will focus on the following aspects: (1) Optimize the lightweight design of the multi-dimensional weight calculation mechanism to reduce the local computational overhead of AWAP-FL, such as using lightweight feature importance analysis algorithms

(e.g., mutual information approximation); (2) Combine AWAP-FL with edge computing and blockchain technology to improve the algorithm's adaptability to high-delay cross-border e-commerce network environments and enhance the traceability and security of parameter transmission; (3) Cooperate with e-commerce enterprises to conduct real dataset experiments and verify the practical application effect of AWAP-FL in cross-platform e-commerce collaboration; (4) Design a dynamic participant management mechanism for AWAP-FL, including participant credibility evaluation and dynamic weight adjustment, to adapt to the frequent joining/exit of e-commerce platform nodes; (5) Extend AWAP-FL to multi-modal cross-platform e-commerce data (e.g., images, videos) and design a multi-modal adaptive weight aggregation mechanism to further improve the model's performance.

7 Funding

This work was supported by the Teaching System of Integrating "Four - History" Education into Cross - border E - commerce Professional Courses in Higher Vocational Education, a Project of Humanities and Social Sciences in Jiangxi Province (Project No.JY22101).

References

[1] Zhang, K., Xing, S., & Chen, Y. (2024). Research on Cross-Platform Digital Advertising User Behavior Analysis Framework Based on Federated Learning. *Artificial Intelligence and Machine Learning Review*, *5*(3), 41-54. https://doi.org/10.69987/AIMLR.2024.50304

[2] Sun, Z., Wang, Z., & Xu, Y. (2024). Privacy protection in cross-platform recommender systems: techniques and challenges. *Wireless Networks*, *30*(8), 6721-6730. https://doi.org/10.1007/s11276-023-03509-z

[3] Zhang, K., & Li, P. (2024). Federated learning optimizing multi-scenario ad targeting and investment returns in digital advertising. *Journal of Advanced Computing Systems*, *4*(8), 36-43. https://doi.org/10.69987/JACS.2024.40806

[4] Su, L., Xu, J., & Yang, P. (2023). A non-parametric view of fedavg and fedprox: Beyond stationary points. *Journal of Machine Learning Research*, *24*(203), 1-48. http://jmlr.org/papers/v24/22-0153.html

[5] Mathew, C., & Asha, P. (2024). FedProx: FedSplit algorithm based federated learning for statistical and system heterogeneity in medical data communication. *J Internet Serv Inf Secur*, *14*(3), 353-370. DOI: 10.58346/JISIS.2024.I3.021

[6] Verma, A., Bhattacharya, P., Bodkhe, U., Saraswat, D., Tanwar, S., & Dev, K. (2023). FedRec: Trusted rank-based recommender scheme for service provisioning in federated cloud environment. *Digital Communications and Networks*, *9*(1), 33-46. https://doi.org/10.1016/j.dcan.2022.06.003

[7] Douiba, M., Benkirane, S., Guezzaz, A., & Azrour, M. (2023). An improved anomaly detection model for IoT security using decision tree and gradient boosting. *The Journal of Supercomputing*, *79*(3), 3392-3411. https://doi.org/10.1007/s11227-022-04783-y

[8] Wang, J., Jin, Y., Stoyanov, D., & Wang, L. (2023). Feddp: Dual personalization in federated medical image segmentation. *IEEE Transactions on Medical Imaging*, *43*(1), 297-308. doi: 10.1109/TMI.2023.3299206.

[9] Zhao, Z., Mao, Y., Liu, Y., Song, L., Ouyang, Y., Chen, X., & Ding, W. (2023). Towards efficient communications in federated learning: A contemporary survey. *Journal of the Franklin Institute*, *360*(12), 8669-8703. https://doi.org/10.1016/j.jfranklin.2022.12.053

[10] Saha, A., Rage, K., Senapati, T., Chatterjee, P., Zavadskas, E. K., & Sliogerienė, J. (2025). A Consensus-Based MULTIMOORA Framework under Probabilistic Hesitant Fuzzy Environment for Manufacturing Vendor Selection. Informatica, 36(3), 713-736. doi:10.15388/24-INFOR581

[11] Tian, C., Xie, Y., Chen, X., Li, Y., & Zhao, X. (2024). Privacy-preserving cross-domain recommendation with federated graph learning. *ACM Transactions on Information Systems*, *42*(5), 1-29. https://doi.org/10.1145/3653448

[12] Shrestha, S. (2021). Evaluating the Impact of Federated Identity Management Systems on Consumer Trust and Regulatory Compliance in E-Commerce. *Journal of Emerging Cloud Technologies and Cross-Platform Integration Paradigms*, *5*(6), 1-11. https://hammingate.com/index.php/JECTCIP/article/view/2021-06-04

[13] Bhimanpallewar, R. N., Khan, S. I., Raj, K. B., Gulati, K., Bhasin, N., & Raj, R. (2025). RETRACTED: Federate learning on Web browsing data with statically and machine learning technique. *International Journal of Pervasive Computing and Communications*, *21*(1), 144-156. https://doi.org/10.1108/IJPCC-05-2022-0184

[14] Bharati, S., Mondal, M. R. H., Podder, P., & Prasath, V. S. (2022). Federated learning: Applications, challenges and future directions. *International Journal of Hybrid Intelligent Systems*, *18*(1-2), 19-35. https://doi.org/10.3233/HIS-220006

[15] Ali, W., Kumar, R., Zhou, X., & Shao, J. (2024). Responsible recommendation services with blockchain empowered asynchronous federated learning. *ACM Transactions on Intelligent Systems and Technology*, *15*(4), 1-24. https://doi.org/10.1145/3633520

[16] Saha, A., Rage, K., Senapati, T., Chatterjee, P., Zavadskas, E. K., & Sliogerienė, J. (2025). A Consensus-Based MULTIMOORA Framework under Probabilistic Hesitant Fuzzy Environment for Manufacturing Vendor Selection. Informatica, 36(3), 713-736. doi:10.15388/24-INFOR581

[17] Alkan, N., & Kahraman, C. (2025). Continuous Pythagorean Fuzzy Set Extension with Multi-Attribute Decision Making Applications. Informatica, 36(2), 241-283. doi:10.15388/25-INFOR584

[18] Rasheed, M. A., Uddin, S., Tanweer, H. A., Rasheed, M. A., Ahmed, M., & Murtaza, H. (2021). Data privacy issue in federated learning resolution using block chain. *VFAST Transactions on Software Engineering*, *9*(4), 51-61. https://doi.org/10.21015/vtse.v9i4.726

[19] Liu, H., Li, N., Kou, H., Meng, S., & Li, Q. (2025). FSRPCL: Privacy-preserve federated social relationship prediction with contrastive learning. *Tsinghua Science and Technology*, *30*(4), 1762-1781. doi: 10.26599/TST.2024.9010077

[20] Harasic, M., Keese, F. S., Mattern, D., & Paschke, A. (2024). Recent advances and future challenges in federated recommender systems. International *Journal*

of Data Science and Analytics, *17*(4), 337-357. https://doi.org/10.1007/s41060-023-00442-4

[21] Xiong, G., Yan, K., & Zhou, X. (2022). A distributed learning based sentiment analysis methods with Web applications. *World Wide Web*, *25*(5), 1905-1922. https://doi.org/10.1007/s11280-021-00994-0

[22] Žvirblis, T., Pikšrys, A., Bzinkowski, D., Rucki, M., Kilikevičius, A., & Kurasova, O. (2024). Data Augmentation for Classification of Multi-Domain Tension Signals. Informatica, 35(4), 883-908. doi:10.15388/24-INFOR578