Clustering-Based Federated Learning for Enhancing Data Privacy in Internet of Vehicles

Zilong Jin[1], Jin Wang[1], and Lejun Zhang[2*]
[1] School of Software, Nanjing University of Information Science and Technology
Nanjing 210044, China
[e-mail : zljin@outlook.com, wangjin991221@163.com]
[2] Cyberspace Institute Advanced Technology, Guangzhou University
Guangzhou 510006, China
[e-mail : zhanglejun@gzhu.edu.cn]
*Corresponding author : Lejun Zhang

Received December 11, 2023; revised February 25, 2024; accepted April 26, 2024; published June 30, 2024

Abstract

With the evolving complexity of connected vehicle features, the volume and diversity of data generated during driving continue to escalate. Enabling data sharing among interconnected vehicles holds promise for improving users' driving experiences and alleviating traffic congestion. Yet, the unintentional disclosure of users' private information through data sharing poses a risk, potentially compromising the interests of vehicle users and, in certain cases, endangering driving safety. Federated learning (FL) is a newly emerged distributed machine learning paradigm, which is expected to play a prominent role for privacy-preserving learning in autonomous vehicles. While FL holds significant potential to enhance the architecture of the Internet of Vehicles (IoV), the dynamic mobility of vehicles poses a considerable challenge to integrating FL with vehicular networks. In this paper, a novel clustered FL framework is proposed which is efficient for reducing communication and protecting data privacy. By assessing the similarity among feature vectors, vehicles are categorized into distinct clusters. An optimal vehicle is elected as the cluster head, which enhances the efficiency of personalized data processing and model training while reducing communication overhead. Simultaneously, the Local Differential Privacy (LDP) mechanism is incorporated during local training to safeguard vehicle privacy. The simulation results obtained from the 20newsgroups dataset and the MNIST dataset validate the effectiveness of the proposed scheme, indicating that the proposed scheme can ensure data privacy effectively while reducing communication overhead.

Keywords: Clustering, Federated Learning, Local Differential Privacy (LDP), Internet of Vehicles (IoV)

This work was sponsored by the National Natural Science Foundation of China (Grant Nos. 62271264, 42175194, and 61972207) and the Project through the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institution.

http://doi.org/10.3837/tiis.2024.06.003 **ISSN : 1976-7277**

1. Introduction

In the future, the realization of smart transportation systems relies on effective data sharing among intelligent vehicles, which enables the development of numerous innovative transportation applications. Autonomous driving is one of the most valuable applications of smart transportation systems. By communicating with each other and transport infrastructure, intelligent vehicles can share a set of sensor data, such as ultrasonic radar and camera data, and can enable cooperative driving by utilizing the data [1]. Cooperative driving enhances driving safety and the transportation system's efficiency, thereby reducing traffic accidents and optimizing the schedule of traffic signals. Furthermore, intelligent vehicles can share crucial information about road and weather conditions, enabling other vehicles to plan their routes for avoiding traffic congestion. However, driving data sharing among intelligent vehicles will result in certain risks [2]. Firstly, while centralized cloud servers possess the capability to process data and derive decisions using global information, the collection of data from distributed vehicles contributes to increased communication latency [3]. Secondly, as vehicle data is commonly deemed private and vehicles exhibit reluctance to share their data with others, the central cloud encounters challenges in acquiring an accurate model owing to the scarcity of training data [4].

According to the above concerns, how to train a learning model safely and efficiently, is a critical research issue in intelligent transportation. Centralized training is one of the most common methods where the model training process occurs on a centralized server [5]. In this approach, mobile devices act as data collection and transmission devices, sending data to the centralized server for training. However centralized training needs to upload users' sensitive data to a central server in which a global learning model is trained based on the uploaded data, so appropriate privacy protection approaches need to be considered in Internet of Vehicles (IoV).

Federated Learning (FL) is an emerging machine learning approach with the feature that training models in distributed manners [6]. Different from the centralized learning methods, a client performs FL tasks to transmit raw data to an edge server, which possesses sufficient computing power to execute the data learning process, instead of sending local training data. Therefore, the FL framework can provide an ideal solution for addressing the privacy challenges and mitigating data leakage in IoV. However, deploying FL in IoV needs to address new challenges. Firstly, the distribution of vehicles is notably dispersed in IoV, and the transmission and updates of model parameters between vehicles and RoadSide Units (RSUs) can lead to substantial communication cost. Secondly, while FL enhances data privacy through local model updates, potential privacy risks may persist. For example, it is possible that model parameters may inadvertently reveal some details about the local data distribution, leading opponents to infer that the updated model, which could compromise personal information related to the vehicle.

To address the above challenges, in this paper, we propose a clustering-based federated learning framework where the distribution of vehicles is often highly dispersed. Firstly, local data features are extracted from vehicles, aggregate similarity based on the extracted feature vectors, and divide the vehicles into different clusters. An optimal vehicle is selected as cluster head to communicate with RSUs. Then, during local training, the Local Differential Privacy (LDP) mechanism is integrated to ensure the privacy protection of local vehicle data. Our main contributions can be summarized as follows:

- We propose a clustering-based federated learning approach to tackle the issue of elevated transmission costs resulting from the dispersed distribution of users within

the Internet of Vehicles. In the pre-training phase, we extract local data features, conduct similarity clustering on the extracted feature vectors, and group vehicles into distinct clusters. Simultaneously, we establish communication channels between cluster heads and RSUs to reduce communication costs.

- In order to protect data privacy, we use LDP to encrypt local vehicles to solve the problem of data leakage during transmission.
- The experiments conducted on standard models and datasets demonstrate the significant efficacy of the proposed algorithms.

The rest of this paper is organized as follows. Section 2 outlines the related work concerning federated learning in IoV and distributed networks. Section 3 provides the problem definition. The proposed method is introduced in Section 4. Section 5 presents experimental evaluations of our scheme using two datasets. The concluding remarks are described in Section 6.

2. Related Work

In this section, we introduce the background and applications of FL in IoV. The data generated by vehicles, containing substantial private information like location and trajectory, poses a significant challenge in preserving data privacy during IoV data communication.

Given the heterogeneous and dynamic nature of IoV, security has emerged as a paramount area of research. Federated Learning (FL) stands as a privacy-preserving distributed training framework, involving collaborative training of a unified Machine Learning (ML) model across diverse participants using their local datasets. The widely adopted Federated Average (FedAvg) algorithm [7] integrates local Stochastic Gradient Descent (SGD) on individual clients, which is then amalgamated by a central server through model averaging. FedAvg allows clients to execute multiple batch updates on their local data, exchanging updated weights rather than gradients. This approach notably amplifies communication efficiency and plays a crucial role in mitigating privacy risks.

In transportation systems, FL has showcased its capability to offer intelligent services including autonomous driving, route planning, safety prediction, and precise vehicle detection, all while ensuring high training accuracy and preserving data privacy. Liu *et al.* [8] introduced the Federated Gated Recurrent Unit neural network algorithm (FedGRU) for traffic flow prediction based on FL. FedGRU operates without direct access to scattered organizational data. Instead, it utilizes a security parameter aggregation mechanism to facilitate training global models in a distributed manner. This mechanism involves aggregating gradient information from all locally trained models to construct a comprehensive global model tailored for prediction purposes. Lim *et al.* [9] considered a perceptual and collaborative learning scheme based on FL, specifically examining its application in Unmanned Aerial Vehicle (UAV) scenarios within IoV. Acknowledging the misalignment of incentives between the UAV and the model owner, the authors introduced a multi-dimensional contract-matching incentive design. This design aims to assign the UAV with the minimum marginal cost for node coverage to each subregion, thereby ensuring efficient task completion.

FL has been applied to achieve distributed data sharing in vehicle networks. Samarakoon *et al*. [10] addressed the issue of joint power and resource allocation in Ultra-Reliable Low Latency Communication (URLLC), particularly within vehicle environments. They employed FL to estimate the tail distribution of the network-wide queue length, thereby gaining valuable insights into the network's overall state. Ye *et al*. [11] delved into FL's potential in the Internet of Vehicles (IoV) for image classification. They introduced a selective model aggregation method, which considers both local image quality and the computing power of each vehicle to

determine the most suitable local model for computation on the vehicle. Experiments conducted using real datasets have demonstrated that the proposed decentralized FL scheme exhibits superior accuracy and provides enhanced privacy protection compared to conventional centralized FL methods. Chai *et al*. [12] explored a federated learning framework for knowledge exchange within vehicle networks employing hierarchical blockchains. The proposed framework comprises two key chains: the ground chain and the top chain. The ground chain operates by employing multiple vehicles as FL clients, each conducting individualized learning processes using their respective hardware. Meanwhile, Roadside Units (RSUs) function as decentralized FL aggregators within the blockchain network, securely consolidating transactions within their coverage areas. Simultaneously, the top chain oversees multiple RSUs responsible for executing FL model computations. The results from FL are integrated into the block ledger, ensuring secure sharing between RSUs and vehicles, thereby upholding security and traceability. While FL transmits only updates to model parameters rather than raw data, thus reducing data transmission and conserving network bandwidth and energy consumption, there exists a potential risk of malicious attacks or data tampering by participants. Consequently, implementing adequate security mechanisms becomes imperative to uphold the integrity and credibility of the model. Zhao *et al.* [13] proposed an FL collaborative authentication protocol tailored for shared data. This protocol enables anonymous validation between vehicles, RSUs, and content servers (CS), ensuring simultaneous security of model parameters to protect privacy. However, the consolidation of these model parameters through a centralized CS introduces a potential single point of vulnerability.

FL has also made contributions to enhancing data privacy in IoV. Lu *et al*. [14] introduced a two-stage mitigation scheme focused on intelligent data conversion and collaborative data leak detection. Unlike existing solutions, the proposed vehicle FL solution allows participants (such as vehicles) to use their data to train models locally without centralized administrators, which significantly contributes to protecting their data privacy. Additionally, collaborative mapping techniques are employed across multiple vehicles to guarantee the effectiveness of the processed information. This method allows for the translation of unprocessed data from diverse origins into the learning data framework. To counter privacy risks in the Internet of Things (IoT), Zhao *et al*. [15] introduced a novel technique that merges FL with differential privacy. The objective is to disrupt gradients produced by vehicles while preserving their usefulness, thereby preventing adversaries from deducing raw data, even if they access altered gradients. On the other hand, Pokhrel *et al*. [16] integrate Federated Learning (FL) with blockchain technology to devise a decentralized solution for vehicle system planning. In this approach, each vehicle functions as an FL client, executing machine learning models and exchanging computed updates through the blockchain ledger to verify their associated rewards. By leveraging blockchain, this method potentially mitigates challenges associated with traditional FL, particularly in handling extended communication durations and security vulnerabilities linked to external entities. However, while these studies primarily focus on safeguarding data privacy, they do not delve into addressing concerns related to communication expenses.

Based on the characteristics of data privacy protection, FL has the potential to promote vehicle-to-vehicle (V2V) network resource management strategies. Zhang *et al*. [17] investigated an alternative strategy for resource allocation in vehicle-to-everything (V2X) communication. This method combines FL with Deep Reinforcement Learning (DRL) [18] to create a federated intelligent approach for allocating resources. The goal is to optimize the overall capacity of vehicle users while adhering to designated delay and reliability parameters.

Each vehicle user operates as a DRL agent, utilizing the Deep Neural Networks (DNN) algorithm to select optimal modes and allocate resources. At the same time, BS (Base Station) aggregates the updated information unloaded by the user and constructs an undirected graph using channel gain information. Cao *et al*. [19] introduced a federated solution leveraging mobile edge computing for managing caching and computing resources in vehicle networks. Specifically, vehicles collaborate with RSUs to engage in federated learning. Within this framework, each entity computes sub-gradient descent updates, contributing to joint parameter optimization aimed at minimizing system costs.

The existing research efforts apply FL frameworks to IoV, solving the problem of model leakage that may occur during the upload process. However, the participation of each vehicle in communication leads to increased communication cost. Therefore, a clustered federated learning framework is proposed to reduce communication cost as well as protect data privacy.

3. The Problem Definition

In order to form clusters based on the data similarity, the features of the local dataset are extracted, and compare the distance between the obtained feature vectors. The set of vehicles is $V = \{v_1, v_2, \ldots, v_n\}$. The feature vectors $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_n\}$. The similarity between the feature vectors of vehicle v_i and vehicle v_j is:

$$\mathrm{sim}\left(\mathbf{R}_i, \mathbf{R}_j\right) = \left\|\mathbf{R}_i - \mathbf{R}_j\right\|_2 \tag{1}$$

where $\|\bullet\|_2$ represents the Euclidean norm of the vector.

It is assumed that there are several intersections in IoV, with RSUs distributed on both sides of the road. Based on the similarity of the vehicle feature matrix, vehicles are divided into different clusters to communicate with RSUs. Optimal vehicles in each cluster are selected as the cluster heads for communication with RSUs.

Each cluster head v_i maintains a local dataset $D_i = \{(a_1, b_1), \ldots, (a_m, b_m)\}$, where a_i represents the input data for initial models, and b_i denotes the corresponding expected output. For each cluster head v_i, the objective is to train a global model $Z = g(\theta, a)$ using the training set D. The loss function $F_i(\theta)$ for Dataset D_i is defined as:

$$F_i(\theta) = \frac{1}{|D_i|} \sum_{j \in D_i} f_j(\mathrm{g}(\theta, a), b) \tag{2}$$

where $f_j(g(\theta, a), b)$ is the loss function for the j-th data sample (a_j, b_j) with model parameters θ.

We define the objective global loss function $F(\theta)$ as:

$$F(\theta) = \frac{1}{|D|} \sum_{j \in D} f_j(\mathrm{g}(\theta, a), b) = \frac{1}{|D|} \sum_{j=1}^{K} \left|D_j\right| \cdot F_j(\theta) \tag{3}$$

In the phase of local model training, Local Differential Privacy (LDP) is employed to safeguard data privacy. LDP is a privacy protection technique that allows local devices to perform noise perturbation on gradient updates before uploading them. The advantage of this technology is that it can still provide meaningful contributions to global model updates while protecting data privacy. Compared with technologies such as homomorphic encryption and secure multi-party computing (SMC), LDP typically has lower computational costs and higher efficiency, as it only adds some noise on local devices without the need for complex encryption operations. Meanwhile, since it does not require large-scale modifications to algorithms or communication protocols, LDP can usually be more easily integrated into existing federated learning frameworks. In summary, although other encryption techniques can also be used for

data privacy protection, LDP may have more advantages in specific situations, especially when considering computational costs, implementation difficulties, and flexibility in federated learning. The Gaussian mechanism is utilized to introduce noise to the updated models of individual vehicles, aiming to perturb the parameters. The mathematical representation of this process is illustrated in the following equation:

$$\theta_i^{'}(t) = \theta_i(t-1) + \alpha_t \cdot \left(\nabla F_i\left(\theta_i(t-1)\right) + \zeta\left(0, \sigma^2 \cdot S_f^2\right)\right) \tag{4}$$

where $\zeta\left(0, \sigma^2 \cdot S_f^2\right)$ is the added Gaussian noise with mean 0 and standard deviation $\sigma \cdot S_f$. In cases where gradients significantly impact the global model, a lower privacy cost (i.e., less ε) should be introduced to safeguard privacy as the gradients approach convergence. Conversely, if the gradients fail to meet the necessary contribution level, a higher privacy cost is allocated, resulting in the addition of less noise to the gradients.

After completing the training of the local model, the subsequent phase involves training a global model capable of acquiring more comprehensive insights by amalgamating multiple clusters, thereby enhancing the overall performance of the global model. The aim of FL is to train a global model $Z = g(\theta, a)$ in IoV. This process constitutes an optimization problem that seeks to minimize $F(\theta)$ i.e.:

$$\begin{gathered} g\left(\theta\right) = \underset{\theta \in \{\theta(t): t<T\}}{\arg\min}\ F\left(\theta\right) \\ s.t. \Pr\left(\theta_i \in \delta_d\right) \le \exp\left(\varepsilon\right) \Pr\left(\theta_i^{'} \in \delta_d\right) \\ \forall v_i \in V, i \in \{1, 2, \ldots, K\} \end{gathered} \tag{5}$$

where $\theta(t)$ represents the aggregated model's parameters at round t and T denotes the maximum number of updating rounds. $\Pr\left(\theta_i \in \delta_d\right) \le \exp\left(\varepsilon\right) \Pr\left(\theta_i^{'} \in \delta_d\right)$ is the ε-privacy guarantee for update parameters θ_i, and $\theta(t)$ is formulated as follows:

$$\theta(t+1) = \theta(t) + \frac{1}{K} \sum_{i=0}^{K} \Delta\theta_i \tag{6}$$

where $\Delta\theta_i$ is the update from vehicle v_i in round t.

4. Proposed Method

In this section, a clustering method based on FL in IoV has been proposed. The first step is to pre-train the local data of the vehicles. The feature extraction in the preprocessing stage is the process of converting raw data into features that can be used by federated learning models, including extracting useful features from raw data such as vehicle sensor data, driving control data, GPS data, vehicle status, etc. The feature extraction process is shown in **Fig. 1**.

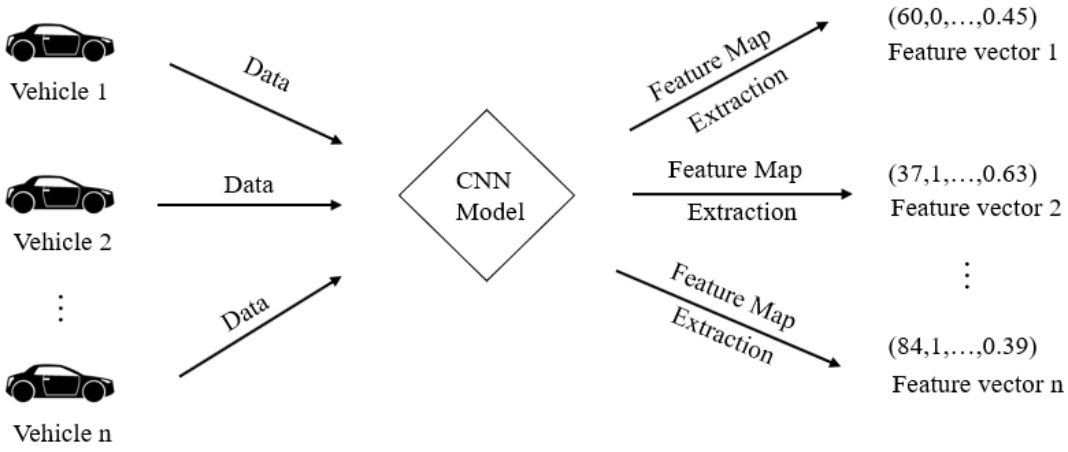

Fig. 1. Feature Extraction

After completion of the pretraining process, the subsequent step involves vehicles uploading their feature vectors to RSUs. Subsequently, the RSUs divide the vehicles into distinct clusters based on the similarity observed among the feature vectors.

Given two feature vectors $\mathbf{R}_i$ and $\mathbf{R}_j$, according to the definition of the Euclidean norm, the distance between $\mathbf{R}_i$ and $\mathbf{R}_j$ can be computed as:

$$\left\|\mathbf{R}_i - \mathbf{R}_j\right\|_2 = \sqrt{\left(R_{i,i} - R_{j,i}\right)^2 + \left(R_{i,j} - R_{j,j}\right)^2 + \ldots + \left(R_{i,n} - R_{j,n}\right)^2} \tag{7}$$

where $\mathbf{R}_{i,j}$ represents the value of the i-th row and j-th column element of the feature vectors.

When we use the feature vectors of vehicles as input, these vectors typically represent various attributes and characteristics of the vehicles. If two vehicle feature vectors are very close in each dimension, their positions in the feature space will also be very close, and we can consider them as similar in the feature space. Based on the similarity results obtained, we can divide the vehicles into different clusters.

Within a cluster, the vehicle exhibiting a stable speed and route will be chosen as the cluster head. The duration allocated for communication is determined by the period during which participating vehicles remain within proximity. Let the coverage area diameter of an RSU be represented as B. For each vehicle k, the time spent within the coverage area of the current RSU is defined by equation (8):

$$T_k = \frac{B - x_k}{q_k} \tag{8}$$

where x_k denotes the position of the k-th vehicle that represents the distance to the entrance, and q_k is the speed of the k-th vehicle.

To guarantee communication with the RSUs, the total time spent stationary by a vehicle k chosen as the cluster head should adhere to the condition $(t_k^{train} + t_k^{up} + t_{agg}) \leq T_{k.}$ Where, t_k^{train} and t_k^{up} represent the estimated training and upload times of vehicle k respectively, while t_{agg} signifies the time needed for aggregation. The standing time refers to the period a vehicle remains within the RSUs' coverage area, primarily influenced by the position and speed of connected vehicles. Extending the duration of stationary time within the coverage area ensures the completion of the training process and the timely delivery of its outcomes. Therefore, a vehicle that has the maximum standing time T_k will be elected as the cluster head.

Then the cluster heads will train local models and communicate with RSUs. 1) RSUs distribute the initial global model to the vehicles that are selected as cluster heads. 2) After receiving the global model, the vehicles iteratively train the model using local data, incorporating the LDP mechanism during the training process, and then generate local models through gradient descent. 3) Vehicles that are selected as cluster heads upload trained local models to RSUs. 4)The global model aggregation is performed in RSUs. The specific process is shown in **Fig. 2**.

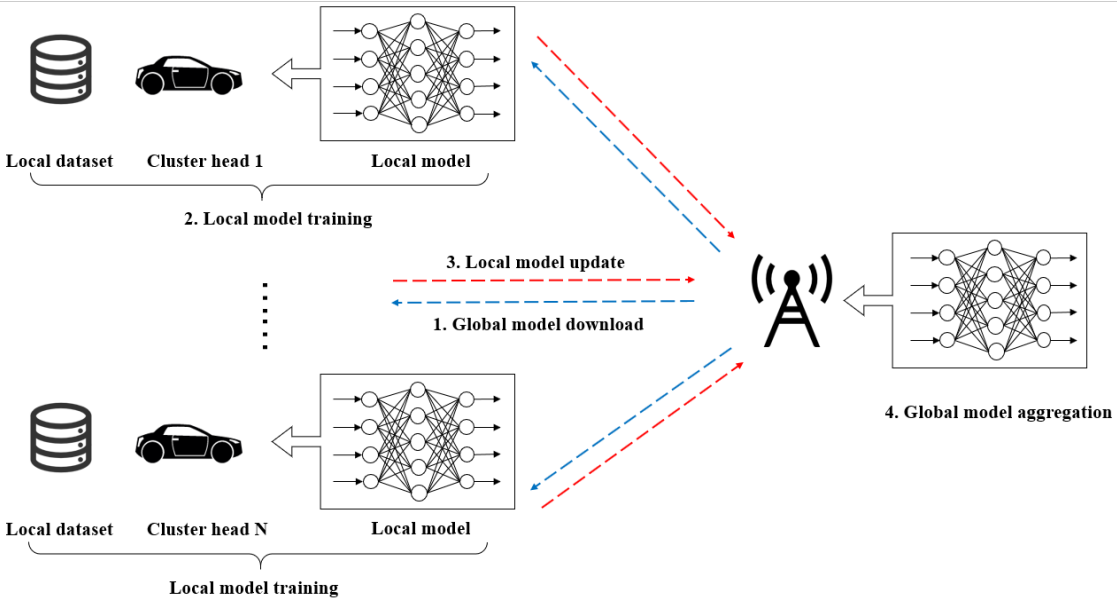

Fig. 2. Communication Process of Cluster FL in IoV

By utilizing communication between cluster heads and RSUs, we have realized the low-cost transmission of clustered FL, concurrently training the optimal global model with minimal loss. The specific algorithm details can be summarized in Algorithm 1, as shown in **Table 1**.

Table 1. Optimization algorithm

Algorithm 1	Clustering-based FL in IoV
Input	Participated vehicle i, number of vehicles n, dataset $D=\{d_i\}$, cluster head CH
Output	global model Z
	Initialize the global model, batch_size, n_cluster
1.	**Client executes:**
2	for each vehicle $i \in \{1, 2, \ldots, n\}$
3.	Pre-train the local datasets d_i and transform the raw data into feature vectors R_i through feature extraction
4.	Upload R_i to RSUs
5.	end for
6.	**RSUs execute:**
7.	calculate the similarity between feature vectors of the vehicle i and vehicle j $\mathrm{sim}\left(\mathbf{R}_i, \mathbf{R}_j\right) = \left\|\mathbf{R}_i - \mathbf{R}_j\right\|_2$
8.	divide vehicles into different clusters based on similarity

9. select the CHs which has the maximum standing time $T_k = \frac{B - x_k}{q_k}$
10. **Client executes:**
11. for each local epoch e < E
12. Extract the mini batch datasets randomly from the local dataset d_i
13. Train a new model based on its mini batch local dataset $di \subset D$
14. Compute the noise-added for local differential privacy $\zeta\left(0, \sigma^2 \cdot S_f^2\right)$
15. Gradient update $\theta(t+1) = \theta(t) + \frac{1}{K}\sum_{i=0}^{K} \Delta\theta_i$
16. Compute loss function $F_i(\theta) = \frac{1}{|D_i|}\sum_{j \in D_i} f_j(\mathrm{g}(\theta, a), b)$
17. end for
18. **RSUs execute:**
19. for each round *r*:
20. Receive models uploaded by CHs
21. Aggregation to generate global model *Z*
22. end for

In the above process, we utilize gradient descent, an efficient iterative optimization technique, to minimize the loss function and derive a global model. This optimal solution is achieved through multiple iterative training processes where we adjust the learning rate. Gradient descent operates by minimizing the objective function $F(\theta)$ via parameter updates in the opposite direction of the function's gradient, denoted as $-\nabla F(\theta)$. The aim of local training for a cluster head vehicle v_i is to compute the model parameters θ_i by progressing towards the direction of $-\nabla F_i(\theta)$, as specified in the following equation:

$$\nabla F_i(\theta) = \frac{\partial F\left(b_i, f\left(a_i\right)\right)}{\partial f\left(a_i\right)} \tag{9}$$

For the cluster head v_i in iteration t, a local update model $\theta_i(t)$ is computed based on the following equation:

$$\theta_i(t) = \theta_i(t-1) + \alpha_t \cdot \nabla F_i\left(\theta_i(t-1)\right) \tag{10}$$

where α_t is the step size for moving in the direction of the opposite gradient. By gradient descent update, we can obtain the minimum loss function and train an optimal global model Z for data security optimization of IoV.

5. Experimental Evaluation

We present numerical experiments to evaluate the proposed algorithm. The Pytorch-based simulations are implemented on a PC with a CPU (Intel(R) Core (TM) i5-10500). The memory of the PC is 32 GB. In the experiment, we set the number of vehicle users to 30 and the number of clusters to 5, with the aim of conducting effective clustered federated learning in a simulated urban vehicle network environment, and fully reflecting the data distribution and characteristics in the urban vehicle network. Choosing 30 vehicles as users can ensure that our dataset has a certain degree of diversity, which can better reflect the data distribution and characteristics in the urban vehicle network. This can ensure the reliability and repeatability

of the experimental results, and enable us to better understand and analyze the differences between different vehicles. Dividing 30 vehicles into 5 clusters can simplify the complexity of data analysis and result interpretation. A smaller number of clusters allows us to have a clearer understanding of the characteristics and behavioral patterns of each cluster, thereby better evaluating the performance and effectiveness of clustered federated learning frameworks. The other simulation parameters are summarized in **Table 2**.

Table 2. Simulation Parameters

Parameter	Value
batch size	100
learning rate	0.01
optimizer	SGD
the number of vehicle users	30
the number of clusters	5
momentum	0.9

5.1 Datasets

The performance of the proposed clustered federated learning framework is evaluated on the 20 newsgroups dataset [20] and the MNIST dataset [21].

1. The 20 newsgroups dataset: This dataset contains approximately 20000 articles and is divided into two parts: a training set and a testing set, with about 60% of the articles being used for training and 40% for testing. The data mentioned is utilized to simulate attribute-based unstructured data generated within urban vehicle networks. This encompasses configuration files and status log files of various vehicle applications.
2. The MNIST dataset: This dataset is a widely used handwritten digit recognition dataset that contains 60000 28x28 pixel grayscale images for training and 10000 for testing. We use this data to simulate the image information collected by vehicles in urban vehicle networks, such as road condition information and vehicle sign information.

The MNIST dataset is a widely used benchmark dataset in the field. In the context of the IoV, the MNIST dataset can be used to simulate image data collected by vehicles, such as road condition information or traffic sign recognition, showcasing the applicability of the framework to visual data processing tasks in vehicular environments. By using the 20 newsgroups dataset, the framework can simulate attribute-based unstructured data generated within urban vehicle networks, demonstrating its effectiveness in processing textual information in vehicular environments. The choice of these datasets aligns with the goal of demonstrating the framework's capability to address privacy concerns and enhance data processing in connected vehicle scenarios.

5.2 Models and Baselines

Convolutional Neural Networks (CNNs) are employed to train local models. In the case of the MNIST dataset, two two-dimensional convolutional layers are utilized to extract image features. Following this, the output of the convolutional layer undergoes activation via an activation layer, after which the feature map's size is reduced using a pooling layer via the Downsampling operation. Simultaneously, a Batch Normalization (BN) layer is introduced to standardize the data within each batch during the model training process. This standardization helps maintain relatively stable data distributions within each layer, accelerating the model's convergence. For the 20newsgroups dataset, the text is first converted into a word vector,

which is input into the model through the embedding layer. The word vector undergoes convolution via a one-dimensional convolutional layer to extract features. Subsequently, the output of the convolutional layer is activated and subjected to pooling, and various other operations. The outcome from the pooling layer is flattened and subsequently input into the fully connected layer.

We compare the method proposed in this paper with FedAvg [7], Fedrep [22], Fedprox [23], SCAFFOLD [24] and Text GCN [25]. FedAvg is used for training on multiple devices or data centers and merging model updates from all devices or data centers without sharing the original data. Fedrep is a dimensionality reduction algorithm that uses a base layer to learn global feature representations between data, in order to alleviate the impact of Non IID on model training. The personalized layer is used as the unique local head for each client to implement personalized optimization algorithms. Fedprox is a federated optimization algorithm that utilizes the proximal term and local training model deviation from the initial model in Non-IID situations, resulting in significant statistical heterogeneity differences. SCAFFOLD is a new federated optimization algorithm that overcomes gradient differences by introducing server control variables and client control variables, effectively alleviating client drift. Text GCN is a graph convolutional neural network model. It is an improvement on the traditional bag-of-words model and the sequence-based model. It can use the relationship between words to represent the Semantic information in the text better.

5.3 Experimental results

Fig. 3 shows a schematic diagram of the clustering results, showing the clustering situation of 30 vehicle clients at an intersection, where coordinates (50, 50) represent the intersection. The proposed algorithm is used to cluster the vehicle positions and select the optimal vehicle as the cluster head to communicate with the RSU. The colors of each cluster are different, and the red circle represents the cluster head of each cluster.

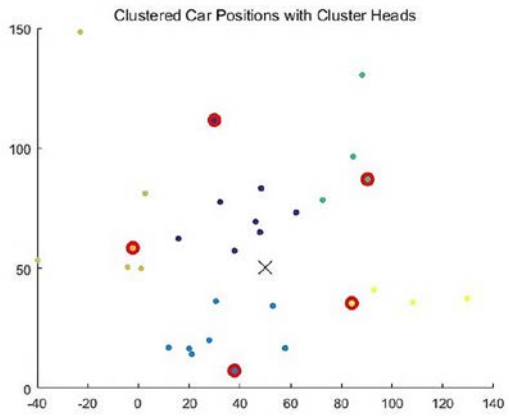

Fig. 3. Clustering Results

Fig. 4 shows the average throughput at speeds ranging from 20 to 70 km/h, and our proposed algorithm has an average throughput range of (0.57 to 0.59 Mbps). As shown in the figure, our proposed method performs better than the other two methods by randomly selecting small batches of data for updates, thereby reducing memory usage and computational time.

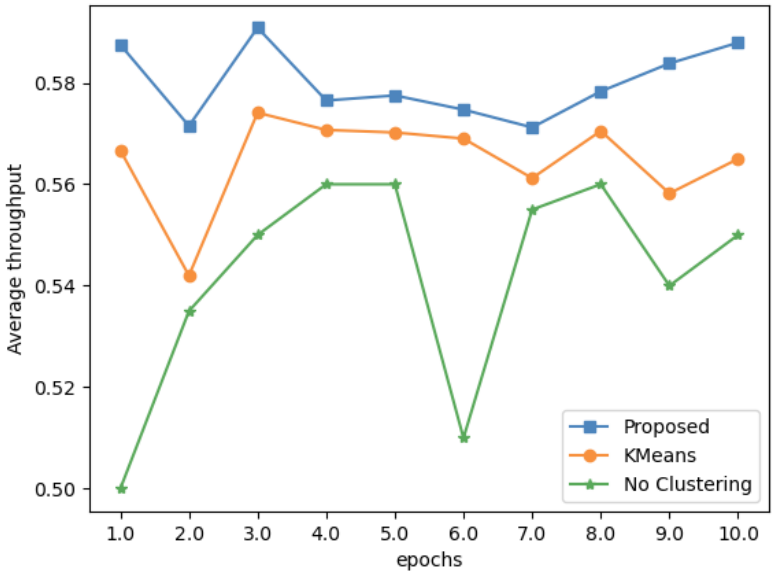

Fig. 4. Average Throughput

The circular representation in the figure illustrates that throughout the entire simulation duration when the Cluster Head (CH) attains the directional threshold point and a new CH is chosen, it results in an enhancement of the average throughput. The selection of new CHs at distinct points (threshold points) induces variations in the average throughput.

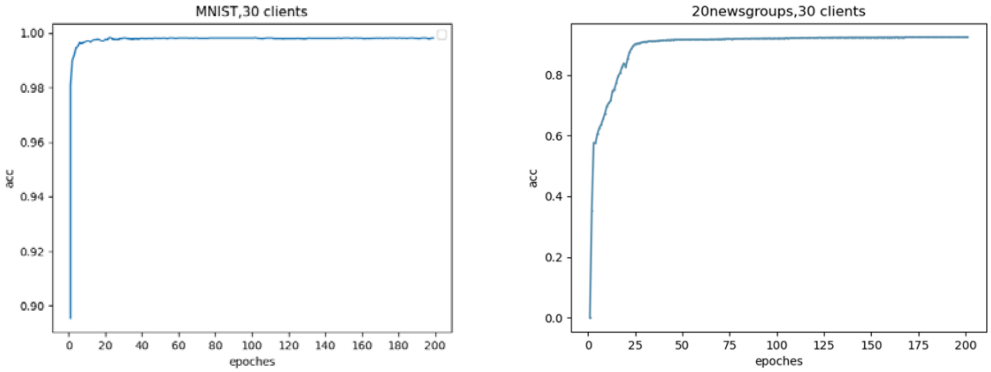

Fig. 5. Accuracy of MNIST dataset and 20newsgroups dataset

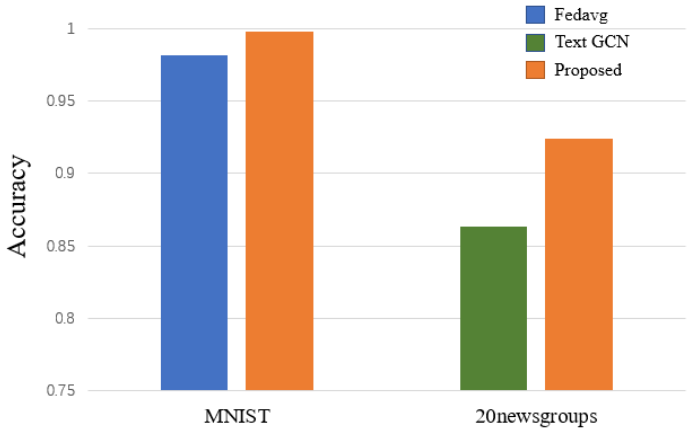

Fig. 6. Comparison of Accuracy on Different Datasets

We evaluated the performance of our proposed scheme on the MNIST dataset and the 20newsgroups dataset, respectively. **Fig. 5** shows the accuracy of various datasets, indicating that our proposed scheme exhibits good performance. Because clustering federated learning does not require the transmission of model parameters from all vehicle clients, only the parameters of the cluster head, it reduces communication overhead and improves training accuracy. The model training accuracy achieved 99.8% on the MNIST dataset and 92.42% on the 20newsgroups dataset. As shown in **Fig. 6** and **Fig. 7**, we also compared the accuracy of model training under different baselines. In the MNIST dataset, the training accuracy of our proposed algorithm improves with the increase of training rounds and consistently outperforms other baselines. In the 20newsgroups dataset, our proposed method has an accuracy improvement of 6.08% compared to Text GCN.

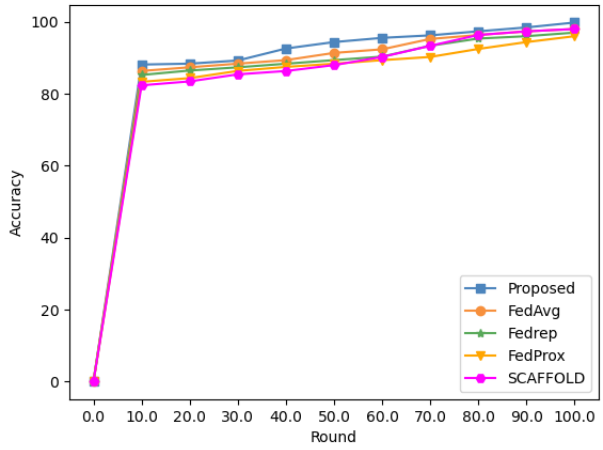

Fig. 7. Evaluations of accuracy over different algorithms

6. Conclusion

In this paper, we mainly explore a secure and efficient communication algorithm for wireless FL in IoV. Considering the communication bottleneck, a novel clustering-based FL framework is proposed in this paper. On the one hand, by dividing vehicles into different clusters based on the similarity of pre-trained feature vectors, it avoids the high communication cost caused by the transmission and update of model parameters between each vehicle and RSU. On the other hand, the LDP mechanism is incorporated during local training to safeguard vehicle privacy. The experimental results show that the accuracy of this scheme has been improved by 6.08% on the 20newsgroups dataset and 1.6% on the MNIST dataset, demonstrating its effectiveness and accuracy. This method can provide novel ideas and approaches for vehicle data processing and privacy protection.

We will focus on the proposed framework's ability to handle many vehicles and different data types in dynamic vehicle environments in future work. This will involve optimizing algorithms and system architecture to ensure effective transmission, processing, and updating of models in large-scale vehicle networks. We will conduct on-site tests or simulations in real driving scenarios to evaluate the performance of the framework, its impact on driving safety, and its practicality.

References

[1] Q. Wu, X. Wang, Q. Fan, P. Fan, C. Zhang and Z. Li, “High stable and accurate vehicle selection scheme based on federated edge learning in vehicular networks,” *China Communications*, vol. 20, no. 3, pp. 1-17, March. 2023. Article (CrossRef Link)

[2] H. Zhou, Y. Zheng, H. Huang, J. Shu and X. Jia, “Toward Robust Hierarchical Federated Learning in Internet of Vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5600-5614, May. 2023. Article (CrossRef Link)

[3] F. Liang, Q. Yang, R. Liu, J. Wang, K. Sato and J. Guo, “Semi-Synchronous Federated Learning Protocol With Dynamic Aggregation in Internet of Vehicles,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 5, pp. 4677-4691, May. 2022. Article (CrossRef Link).

[4] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang and H. Du, “Federated Multi-Agent Deep Reinforcement Learning for Resource Allocation of Vehicle-to-Vehicle Communications,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8810-8824, Aug. 2022. Article (CrossRef Link).

[5] L. F. Da Costa, L. S. Furtado, P. H. G. Rocha, P. A. L. Rego and F. A. M. Trinta, “Time series prediction in IoT: a comparative study of federated versus centralized learning,” in *Proc. of CCNC*, Las Vegas, USA, pp. 993-994, 2023. Article (CrossRef Link)

[6] H. Xiao, J. Zhao, Q. Pei, J. Feng, L. Liu and W. Shi, “Vehicle Selection and Resource Optimization for Federated Learning in Vehicular Edge Computing,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11073-11087, Aug. 2022. Article (CrossRef Link).

[7] McMahan B, Moore E, Ramage D, et al., “Communication-efficient learning of deep networks from decentralized data,” in *Proc. of Artificial Intelligence and Statistics*, Florida, USA, vol. 54, pp. 1273-1282, 2017. Article (CrossRef Link)

[8] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, “Privacy-preserving traffic flow prediction: A federated learning approach,” *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, August. 2020. Article (CrossRef Link)

[9] Lim W Y B, Huang J, Xiong Z, et al., “Towards federated learning in UAV-enabled internet of vehicles: A multi-dimensional contract-matching approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 5140-5154, August. 2021. Article (CrossRef Link)

[10] S. Samarakoon, M. Bennis, W. Saad and M. Debbah, "Distributed Federated Learning for Ultra-Reliable Low-Latency Vehicular Communications," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 1146-1159, February. 2020. Article (CrossRef Link)
[11] D. Ye, R. Yu, M. Pan, and Z. Han, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23920–23935, January. 2020. Article (CrossRef Link)
[12] H. Chai, S. Leng, Y. Chen and K. Zhang, "A Hierarchical Blockchain-Enabled Federated Learning Algorithm for Knowledge Sharing in Internet of Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 3975-3986, July. 2021. Article (CrossRef Link)
[13] P. Zhao, Y. Huang, J. Gao, et al., "Federated Learning-Based Collaborative Authentication Protocol for Shared Data in Social IoV," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 7385-7398, April. 2022. Article (CrossRef Link)
[14] Y. Lu, X. Huang, Y. Dai, S. Maharjan and Y. Zhang, "Federated Learning for Data Privacy Preservation in Vehicular Cyber-Physical Systems," *IEEE Network*, vol. 34, no. 3, pp. 50-56, June. 2020. Article (CrossRef Link)
[15] Y. Zhao et al., "Local Differential Privacy-Based Federated Learning for Internet of Things," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8836-8853, June. 2021. Article (CrossRef Link)
[16] S. R. Pokhrel and J. Choi, "Federated Learning With Blockchain for Autonomous Vehicles: Analysis and Design Challenges," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734-4746, August. 2020. Article (CrossRef Link)
[17] X. Zhang, M. Peng, S. Y an, and Y. Sun, "Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6380–6391, July. 2020. Article (CrossRef Link)
[18] D. C. Nguyen, P. N. Pathirana, M. Ding and A. Seneviratne, "Privacy-Preserved Task Offloading in Mobile Blockchain With Deep Reinforcement Learning," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2536-2549, December. 2020. Article (CrossRef Link)
[19] J. Cao, K. Zhang, F. Wu and S. Leng, "Learning Cooperation Schemes for Mobile Edge Computing Empowered Internet of Vehicles," in *Proc. of WCNC*, Seoul, Korea, pp. 1-6, 2020. Article (CrossRef Link)
[20] T. Mitchell, "20 newsgroups dataset," 2019. [Online]. Available: http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz
[21] Y. LeCun and C. Cortes, "MNIST Handwritten Digit Database," 2010. [Online]. http://yann.lecun.com/exdb/mnist/
[22] Collins L, Hassani H, Mokhtari A, et al., "Exploiting shared representations for personalized federated learning," in *Proc. of Machine Learning Research*, pp. 2089-2099, 2021. Article (CrossRef Link)
[23] Li T, Sahu A K, Zaheer M, et al., "Federated optimization in heterogeneous networks," in *Proc. of Machine learning and systems*, pp. 429-450, 2020. Article (CrossRef Link)
[24] Karimireddy S P, Kale S, Mohri M, et al., "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. of Machine Learning Research*, pp. 5132-5143, 2020. Article (CrossRef Link)
[25] Yao L, Mao C, Luo Y., "Graph convolutional networks for text classification," in *Proc. of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, pp. 7370-7377, Hawaii, USA, 2019. Article (CrossRef Link)

Zilong Jin received the B.E. degree in computer engineering from Harbin University of Science and Technology, China, in 2009, and the M.S. and Ph.D. degrees in computer engineering from Kyung Hee University, Korea, in 2011 and 2016, respectively. He is currently an associate professor at School of Software at Nanjing University of Information Science and Technology, China. His research interests include mobile wireless networks, cognitive radio networks, and mobile edge networks.

Jin Wang received her B.E. degree in software engineering from Qufu Normal University, China, in 2021. Now she is a master student in School of Software, Nanjing University of Information Science and Technology. Her main research interests include federated learning and internet of vehicles.

Lejun Zhang received his M.S. degree in computer science and technology in Harbin Institute of Technology and the Ph.D. degrees in computer science and technology at Harbin Engineering University. Now he is currently a professor and Ph.D. Supervisor of the Cyberspace Institute of Advanced Technology, Guangzhou University. He was a Visiting Scholar with Carnegie Mellon University. His research interests include Cyberspace Security, blockchain and information security.