Journal of Computer Applications
计算机应用, 2021, 41(6): 1546 - 1550
ISSN 1001-9081
CODEN JYIIDU
2021-06-10
http://www.joca.cn

文章编号:1001-9081(2021)06-1546-05
DOI:10.11772/j.issn.1001-9081.2020121912

高速车联网场景下分簇式无线联邦学习算法

王家瑞[1,2],谭国平[1,2*],周思源[1,2]
(1. 河海大学 计算机与信息学院 南京 211100;
2. 江苏智能交通及智能驾驶研究院 南京 210019)
(*通信作者:gptan@hhu.edu.cn)

摘　要:现有无线联邦学习框架缺乏对实际的分布式高速车联网(IoV)场景的有效支持。针对该场景下的分布式学习问题,提出了一种基于随机网络拓扑模型的分布式训练算法——分簇式无线联邦学习算法(C-WFLA)。首先,该算法基于高速公路场景下的车辆分布情况设计网络模型;其次,该算法考虑了用户端进行上行数据传输时的路径衰落、瑞利衰落等因素;最后,该算法设计了基于分簇式训练的无线联邦学习方法。利用所提算法对手写体识别模型进行了训练与测试,仿真结果表明:在信道状态较好、用户发射功率受限较小的情况下,传统无线联邦学习算法与C-WFLA在相同的训练条件下损失函数均能收敛至相近的数值,且C-WFLA收敛更快;而在信道状态较差、用户发射功率受限较大的情况下,C-WFLA损失函数收敛值相较于传统的集中式算法可以降低10%~50%。可见,C-WFLA更有助于高速IoV场景下的模型训练。

关键词:无线联邦学习;随机拓扑;车联网;分布式学习;分簇算法
中图分类号:TN929.5　　**文献标志码**:A

Clustered wireless federated learning algorithm in high-speed internet of vehicles scenes

WANG Jiarui[1,2], TAN Guoping[1,2*], ZHOU Siyuan[1,2]
(1. *School of Computer and Information, Hohai University, Nanjing Jiangsu* 211100, *China*;
2. *Jiangsu Intelligent Transportation and Intelligent Driving Research Institute, Nanjing Jiangsu* 210019, *China*)

Abstract: Existing wireless federated learning frameworks lack the effective support for the actual distributed high-speed Internet of Vehicles (IoV) scenes. Aiming at the distributed learning problem in such scenes, a distributed training algorithm based on the random network topology model named Clustered-Wireless Federated Learning Algorithm (C-WFLA) was proposed. In this algorithm, firstly, a network model was designed on the basis of the distribution situation of vehicles in the highway scene. Secondly, the path fading, Rayleigh fading and other factors during the uplink data transmission of the users were considered. Finally, a wireless federated learning method based on clustered training was designed. The proposed algorithm was used to train and test the handwriting recognition model. The simulation results show that under the situations of good channel state and little user transmit power limit, the loss functions of traditional wireless federated learning algorithm and C-WFLA can converge to similar values under the same training condition, but C-WFLA converges faster; under the situations of poor channel state and much user transmit power limit, C-WFLA can reduce the convergence value of loss function by 10% to 50% compared with the traditional centralized algorithm. It can be seen that C-WFLA is more helpful for model training in high-speed IoV scenes.

Key words: wireless federated learning; random topology; Internet of Vehicles (IoV); distributed learning; clustering algorithm

0　引言

随着5G技术的发展,物联网逐渐成为5G时代的研究热点。车联网作为物联网中一个有潜力的研究分支,有望成为智能交通系统中的重要的数据传输与控制平台。车联网是一种移动自组网络,可以有效地改善道路安全问题和驾驶者的驾乘环境。支撑这一功能的是用户及其车辆所带来的大量数据,但是车联网的规模巨大、所用无线信道较为开放缺乏保密性、车辆的运动轨迹容易被跟踪预测,这都使用户的安全隐私容易泄露。不法分子可能通过截取用户广播的信息、预测车辆轨迹等方式窃取同户的数据隐私,一旦车联网系统暴露了任何车辆或用户的隐私信息,将在很长一段时间内难以被公众广泛地接受。因此,用户的隐私安全问题逐渐成为限制车辆及用户参与数据提供程度的主要因素。为加强对用户隐私

收稿日期:2020-11-04;**修回日期**:2021-03-31;**录用日期**:2021-04-06。　**基金项目**:国家自然科学基金资助项目(61701168,61832005,61571303);中国博士后科学基金资助项目(2019M651546);江苏省交通技术改造项目(2018Y45)。

作者简介:王家瑞(1998—),男,山东威海人,硕士研究生,主要研究方向:无线网络;　谭国平(1975—),男,湖南澧县人,教授,博士,CCF会员,主要研究方向:无线通信系统与网络;　周思源(1985—),男,江苏南京人,副教授,博士,CCF会员,主要研究方向:无线通信。

的保护,除差分隐私保护理论[1]、k匿名[2]等常用的隐私保护方法外,近几年,文献[3-7]中也提出了许多解决方案。与此同时,2016年谷歌提出了一种基于用户隐私保护的学习框架——联邦学习[8-10],其主要的特征是数据提供方的数据均保留在本地,没有进行数据传输,从源头上抑制了数据隐私的泄露。通过联邦学习,车联网系统可以在保护用户隐私不被泄露的条件下,使用大量用户数据进行模型训练。

现行的许多关于分布式联邦学习系统的研究[11-14]的用户拓扑通常为星型拓扑。但星型拓扑大多针对小范围的随机用户,并没有充分考虑车联网场景下车辆随道路分布的特殊性及其对联邦学习训练效果的影响,为此本文提出了一种分布式的分簇式联邦学习算法。从文献[15-16]中可以得知,目前车联网的发展存在以下两方面挑战:一方面,车联网场景下用户分布往往更为分散,采用单参数服务端进行用户的模型数据汇总、更新往往需要更长的时间;另一方面,用户距离参数服务端较远,用户所需的总功率相对较大。通过设计用户的分簇方案可以选择用户端总功率较小的分簇方式进行训练,从而对用户端进行功率控制。

1　高速路车联网模型

1.1　高速公路车辆分布模型

如图1所示,模型建立在双向四车道的高速公路上,路段长度为L,单车道宽为W,圆点表示车辆。在道路中间每隔距离i设置一个路侧元(Road Side Unit, RSU),用于完成用户模型的接收汇总与更新。

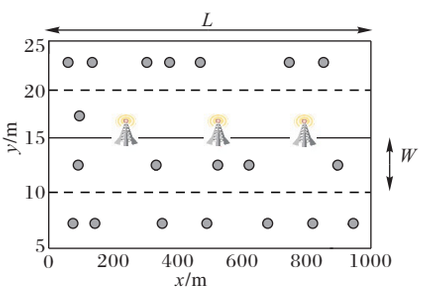

图1　高速公路车道模型

Fig. 1　Model of highway lanes

在车用无线通信技术的长期演进计划(Long Term Evolution-Vehicle to everything, LTE-V2X)系统级仿真中,设计车辆撒点及运动的内容包括五项:车辆数量、撒点方式、车速、行车方向、转向模型[17]。其中车辆数量N的计算式如下:

$$N = L/(P*T) \tag{1}$$

其中:P为车速;T为驾驶员安全反应时间。在上述模型的基础上,将在四条车道上随机撒点,使车辆散布于每条车道的中线上,并保证车辆之间的间距大于安全跟车距离l。

1.2　无线传输模型

考虑到RSU的发射功率可以满足数据的有效发送,而移动车辆的发射功率有限,假设RSU将数据下传至簇内用户的下行信道及RSU之间的信道均为无损信道,用户上传模型数据至RSU的信道为衰落信道。

在用户端进行上行模型数据传输时,采用模拟的方法进行传输,第i次迭代时,RSU接收到的信号$\boldsymbol{y}_i(t)$可表示为:

$$\boldsymbol{y}_i(t) = \sum_{m \in \boldsymbol{M}_i} \boldsymbol{h}_{m,i}(t)\boldsymbol{x}_{m,i}(t) + \boldsymbol{z}_i(t) \tag{2}$$

其中:$\boldsymbol{M}_i$为第i次迭代时当前簇内用户的集合;$\boldsymbol{h}_{m,i}(t)$~CN$(0, \boldsymbol{\sigma}_m^2)$为第$m$个设备在第$i$次模型迭代时与RSU之间的瑞利信道,$\boldsymbol{z}_i(t)$~CN$(0, \boldsymbol{\sigma}_m^2)$为加性高斯白噪声;$\boldsymbol{x}_{m,i}(t)$为$t$时刻第$m$个设备在第$i$次模型迭代完成后所需发送的信息。可以将$\boldsymbol{x}_{m,i}(t)$用式(3)表示:

$$\boldsymbol{x}_{m,i}(t) = \boldsymbol{\alpha}_{m,i}(t)\boldsymbol{g}_{m,i}(t) \tag{3}$$

其中:$\boldsymbol{g}_{m,i}(t)$为第i次迭代时的模型梯度值;$\boldsymbol{\alpha}_{m,i}(t)$表示功率控制向量。为满足发射功率的限制,该功率控制向量的表达式如下:

$$\boldsymbol{\alpha}_{m,i}(t) = \begin{cases} \dfrac{\boldsymbol{\gamma}(t)}{\boldsymbol{h}_{m,i}(t)}, & |\boldsymbol{h}_{m,i}(t)|^2 \geqslant \boldsymbol{\lambda}_{m,i}(t) \\ 0, & \text{其他} \end{cases} \tag{4}$$

其中,$\boldsymbol{\gamma}(t), \boldsymbol{\lambda}_{m,i}(t) \in \mathbf{R}$,为功率控制参数,调控$\boldsymbol{\lambda}_{m,i}(t)$与$\boldsymbol{\gamma}(t)$的值,可以使$\boldsymbol{\alpha}_{m,i}(t)$满足功率限制条件。

结合式(4),可以将RSU接收信号重新表达为:

$$\boldsymbol{y}_i(t) = \boldsymbol{\gamma}(t) \sum_{m \in \boldsymbol{M}_i(t)} \boldsymbol{g}_{m,i}(t) + \boldsymbol{z}_i(t) \tag{5}$$

假设信号需要传输的距离为d,考虑大尺度衰落,可以重新得到此时RSU处接收到的信号表达式:

$$\boldsymbol{y}_i(t) = \boldsymbol{\gamma}(t) \sum_{m \in \boldsymbol{M}_i(t)} \boldsymbol{g}_{m,i}(t) \left(\frac{d}{B}\right)^{\rho} + \boldsymbol{z}_i(t) \tag{6}$$

其中:B为与信号频率等条件相关的常数;ρ为信号距离衰落因子,控制信号衰落的幅度。

1.3　控制参数

由式(4)可知,可以通过调整λ的值来控制有效传输模型数据的数量,以完成对数据丢包情况的模拟。定义有效数据传输率β为有效传输的数据包数量J占模型数据完整传输时所需传输数据包数量H的比值,即:

$$\beta = \frac{J}{H} \tag{7}$$

它可以作为有效传输概率的估计,即:

$$\beta = \int_0^{\boldsymbol{\lambda}(t)} f(z)\mathrm{d}z \tag{8}$$

其中$f(z)$为瑞利分布的概率密度:

$$f(z) = \frac{z}{\delta^2}\exp\left(-\frac{z^2}{2\delta^2}\right) \tag{9}$$

其中δ为方差,由此,可以得到:

$$\lambda(t) = \sqrt{-2\ln(1-\beta)} \tag{10}$$

1.4　损失函数

第k个用户端处训练模型的损失函数可表示为:

$$F_k(\boldsymbol{w}) = \frac{1}{|\boldsymbol{D}_k|} \sum_{(\boldsymbol{x}_j, \boldsymbol{y}_j) \in \boldsymbol{D}_k} f(\boldsymbol{w}, \boldsymbol{x}_j, \boldsymbol{y}_j) \tag{11}$$

其中:$\boldsymbol{D}_k$表示在第k个用户处收集到的本地数据集;$f(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i)$表示模型$\boldsymbol{w}$基于训练集样本$\boldsymbol{x}_i$及其对应标签$\boldsymbol{y}_i$的误差损失函数。同时,一簇内的总体模型损失函数$F(\boldsymbol{w})$可以表示为如下形式:

$$F(\boldsymbol{w}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\boldsymbol{w}) \tag{12}$$

其中，K为该簇内参与模型训练的用户总数。

2 分簇式无线联邦学习算法

2.1 整体系统流程

图2为整体系统框图，后续实验也将据此进行相关仿真。在一次迭代中，当一簇用户的模型更新完成后，其模型将作为下一簇用户的初始模型进行训练，这种方式与传统联邦学习中模型值取平均的做法不同，但这也是针对分簇式联邦学习方法进行的一种尝试。

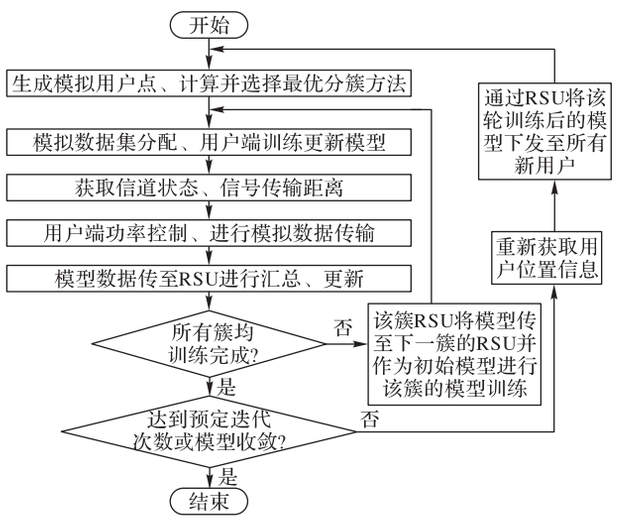

图2 C-WFLA流程

Fig. 2 Flow chart of C-WFLA

2.2 分簇算法

在每次随机撒点完成后，将根据个用户的车辆位置进行分簇，把模型中的N个用户分为C簇，控制用户端在上传数据时不要离RSU过远，具体的分簇方法基于二分k-means的思路，流程如下：

1) 计算N个用户位置坐标的质心。

2) 选择距离1)中质心最近的RSU作为初始中心点。

3) 随机选取2个用户位置做中心点，并由此将剩余用户分为两簇。

4) 选取步骤3)中未选择的用户点，分别计算其与步骤3)中选取两中心点欧氏距离的平方，并使其归于数值较小的一方，该用户点加入后，重新计算该簇用户位置坐标的质心。

5) 重复步骤4)直至所有点分簇完成，选择距离两簇质心最近的RSU作为该簇的中心点。

6) 分别计算两簇内用户点与中心点距离的平方和，选择数值较大的一簇重复步骤3)~4)直至模型内的总簇数达到设定值。

3 实验与结果分析

3.1 实验参数

在实验仿真中，图1中示意的高速公路的长度L定为1 000 m，单条道路宽定为7.5 m。

设置车辆数量时，取车辆速度P为120 km/h，驾驶员安全反应时间T取6 s，安全跟车距离l取20 m，确保同一车道两车间距大于20 m，根据式(1)，可得N = 20。因此，在每次迭代时将模拟生成20辆车的位置，以进行分簇。

本次实验，以数字手写体识别的模型训练为例，展示训练效果，优化器选择随机梯度下降(Stochastic Gradient Descent，SGD)，训练集大小r取5 000，经预实验迭代次数i取150，学习率lr选择如式(13)：

$$\begin{cases} lr_{i+1} = lr_i * 0.99, & i \geqslant 100 \\ 0.1, & 1 \leqslant i < 100 \end{cases} \tag{13}$$

基于每次迭代整体的效率与速度，分簇过少会使整体用户的发射功率增加，分簇过多会导致单次迭代内的训练区域较多，系统整体训练时间较长，因此选择将20个用户分为3簇。

根据图2介绍的流程，接下来通过一次仿真案例的执行情况，具体展示分簇算法运行结果细节：

1)根据用户位置，20个用户端的初始分簇情况如下：

[[1,2,5,6,12,14,17,18],[7,13,15,19],
[0,3,4,8,9,10,11,16]]

其中，数字0~19为用户端的标号，在分配训练集图片时，将给0号、1号用户端分配5 000张数字“0”的图片，以此类推18号、19号用户端将获得5 000张数字“9”的图片。

2)在根据β值的大小做好功率控制的情况下，通过当前簇内用户([1,2,5,6,12,14,17,18])的数据集进行模型学习，并通过RSU将汇总、更新后的模型参数传至下一簇([7,13,15,19])，并作为下一簇用户模型训练的初始模型。

3)重复2)中的操作，直至3簇用户均训练完成，第一次迭代结束。

4)在下一轮迭代开始之前，系统将重新生成用户的位置信息，并重新进行分簇。

5)重复2)~3)中的操作，直至迭代150次，模型损失值基本收敛，训练完成。

3.2 结果分析

图3为β值取20%、40%、60%、80%、100%时，模型经过150轮迭代，传统联邦学习(集中式)、分簇式联邦学习(分布式)两种训练方式下，模型损失函数的变化。

从图3可以看出：在β大于等于40%时，两种训练方式下的模型收敛值、收敛速度相近，但分簇式训练在模型收敛时的损失函数波动变大。当β值继续降低到20%时，传统联邦学习的收敛值剧增，整体模型训练效果变差。

表1为模型经过150轮迭代后，两种训练方式下损失函数的收敛值。

从表1中可以看出：β高于40%时，分簇式联邦学习训练后的模型收敛值略高于传统联邦学习；而当β值降低至20%，分簇式联邦学习的模型收敛值却更低，这说明在β值较低，即信道状态较差或者发射功率受限较大时，分簇式训练有着更好的抵抗性，因此获得了更好的模型训练效果。

对传统联邦学习模式在不同β值下的收敛情况进行了横向对比，如图4所示。在图4中可以观察到，β值为100%、80%

时曲线基本重合,当β值低于40%时,模型损失函数出现了类似门限效应的情况,随着β值的减小,损失函数的收敛值迅速变大,而分簇式联邦学习训练出的模型并没有出现类似情况。

表1　两种训练方式下模型收敛值对比

Tab. 1　Comparison of model convergence values under two training methods

β/%	分布式	集中式
20	3.479 094	8.334 863
40	3.012 653	3.116 485
60	2.873 888	2.400 096
80	2.875 693	2.369 822
100	2.875 526	2.371 209

这一现象,推测可能是随机拓扑网络的随机性产生的效果:

1)从模型参数的角度分析:假设有利于模型训练的关键参数位置基本不变,在随机网络引入之前,在通过功率控制进行模拟丢包后,关键位置的模型参数可能会丢失,从而导致模型不能正常收敛。而在分簇式联邦学习中,用户被分为了多簇,在每一次的迭代中,模型需要进行多次接力更新才能完成,而根据式(6)可知,联邦学习只关注模型更新时,所有用户发送的梯度矢量平均值。由于分簇式联邦学习的每一簇用户在上传模型数据时都需要进行一次功率控制,从概率上讲,模型中关键位置参数全部丢失的可能性相对减小,取而代之的是该位置上的参数值变小,这一变化提高了其模型数据在丢包较多的情况下,训练后模型的整体效果。

2)分簇式联邦学习在每次迭代时,用户的位置与用户的分组方式发生了改变,这相当于在模型训练的过程中引入了一定的随机性,从而优化了整体模型训练的效果,而也正是由于这种随机性的引入使模型收敛时会出现一定的波动。

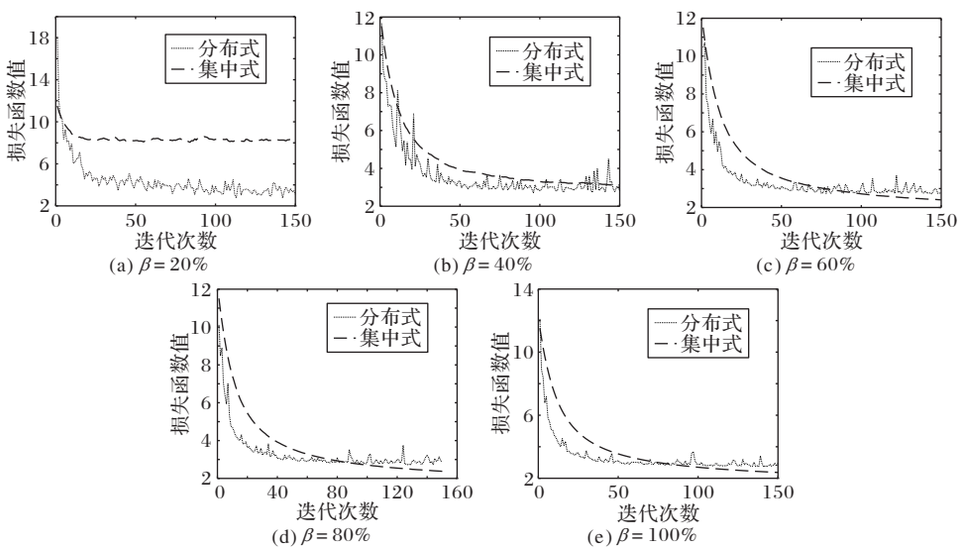

图3　不同β值下的损失函数变化

Fig. 3　Change of loss function under different β values

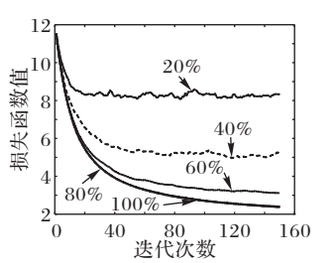

图4　传统联邦学习在训练时的损失函数值

Fig. 4　Loss function values during training of traditional federated learning

4　结语

针对基于高速公路模型的车联网场景,本文提出了一种分布式的分簇式无线联邦学习算法(C-WFLA)。通过仿真实验对该算法的训练性能进行的分析可知,本文提出的分簇式训练方式能有效应对无线系统中的数据丢包状况,即在相应的丢包率低于一定的阈值时,本文提出的分布式算法依然能够取得较好的训练效果。但本文所提出的算法还存在很多问题值得探讨:1)目前只考虑了数字手写体识别模型训练,对一些更复杂的模型有待实验验证;2)对于无线信道的仿真还不够实际,没有考虑多径效应、多普勒效应等实际情况;3)对模型随机性的考虑还不够完备,分簇方法也还有待优化;4)在诸如城市道路、乡村道路等不同车联网模型下的训练效果还有待验证。以上问题都将是我们后续的重点研讨方向。

参考文献(References)

[1] DWORK C. Differential privacy: a survey of results [C]// Proceedings of the 2008 International Conference on Theory and Applications of Models of Computation, LNCS 4978. Berlin: Springer, 2008: 1-19.

[2] SWEENY L. *k*-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[3] KANG J, YU R, HUANG X, et al. Privacy-preserved pseudonym scheme for fog computing supported internet of vehicles [J]. IEEE

Transactions on Intelligent Transportation Systems, 2018, 19(8): 2627-2637.

[4] SUN Y, WU L, WU S, et al. Security and privacy in the internet of vehicles [C]// Proceedings of the 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things. Piscataway: IEEE, 2015:116-121.

[5] KANG J, YU R, HUANG X, et al. Location privacy attacks and defenses in cloud-enabled internet of vehicles [J]. IEEE Wireless Communications, 2016, 23(5): 52-59.

[6] SUN G, SUN S , SUN J, et al. Security and privacy preservation in fog-based crowd sensing on the internet of vehicles [J]. Journal of Network and Computer Applications, 2019, 134: 89-99.

[7] JOY J, GERLA M. Internet of Vehicles and autonomous connected car - privacy and security issues [C]// Proceedings of the 26th International Conference on Computer Communication and Networks. Piscataway: IEEE, 2017: 1-9.

[8] AMIRI M M, GÜNDÜZ D. Federated learning over wireless fading channels [J]. IEEE Transactions on Wireless Communications, 2020, 19(5): 3546-3557.

[9] YANG Q, LIU Y, CHENG Y, et al. Federated Learning [M]. San Rafael: Morgan & Claypool Publishers, 2020: 1-207.

[10] ZHAO Y, LI M, LAI L, et al. Federated learning with non-IID data [EB/OL]. [2019-06-29]. https://arxiv. org/pdf/1806. 00582. pdf.

[11] SAMARAKOON S, BENNIS M, SAAD W, et al. Distributed federated learning for ultra-reliable low-latency vehicular communications [J]. IEEE Transactions on Communications, 2020, 68(2): 1146-1159.

[12] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency [EB/OL]. [2019-06-29]. https://arxiv. org/pdf/1610. 05492. pdf.

[13] JIANG R, ZHOU S. Cluster-based cooperative digital over-the-air aggregation for wireless federated edge learning [C]// Proceedings of the 2020 IEEE/CIC International Conference on Communications in China. Piscataway: IEEE, 2020: 887-892

[14] SATTLER F, WIEDEMANN S, SAMEK W, et al. Robust and communication-efficient federated learning from non-i. i. d. data [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020,31(9):3400-3413.

[15] 牛志升,SHEN S,张钦宇,等. 面向沉浸式体验的空天地一体化车联网体系架构与关键技术[J]. 物联网学报,2017,1(2):17-27. (NIU Z S, SHEN S, ZHANG Q Y, et al. Space-air-ground integrated vehicular network for immersive driving experience [J]. Chinese Journal on Internet of Things, 2017, 1(2): 17-27.)

[16] 尉志青,马昊,张奇勋,等. 感知-通信-计算融合的智能车联网挑战与趋势[J]. 中兴通讯技术,2020,26(1):45-49. (WEI Z Q, MA H, ZHANG Q X, et al. Challenge and trend of sensing, communication and computing integrated intelligent internet of vehicle [J]. ZTE Technology Journal, 2020, 26(1): 45-49.)

[17] 许瑞琛,王俊峰,张莎. LTE-V2X 测试与仿真从入门到精通[M]. 北京:人民邮电出版社,2018:127-128. (XU R C, WANG J F, ZHANG S. LTE-V2X Testing and Simulation: From Initiation to Mastery [M]. Beijing: Posts & Telecom Press, 2018: 127-128.)

This work is partially supported by the National Natural Science Foundation of China (61701168, 61832005, 61571303), the China Postdoctoral Science Foundation (2019M651546), the Jiangsu Province Transportation Technology Transformation Project (2018Y45).

WANG Jiarui, born in 1998, M. S. candidate. His research interests include wireless network.

TAN Guoping, born in 1975 , Ph. D., professor. His research interests include wireless communication system and network.

ZHOU Siyuan, born in 1985, Ph. D., associate professor. His research interests include wireless communication.