计算机科学
COMPUTER SCIENCE

**基于 DBSCAN 聚类的集群联邦学习方法**

鲁晨阳, 邓苏, 马武彬, 吴亚辉, 周浩浩

引用本文

鲁晨阳, 邓苏, 马武彬, 吴亚辉, 周浩浩. 基于 DBSCAN 聚类的集群联邦学习方法[J]. 计算机科学, 2022, 49(6A): 232-237.

LU Chen-yang, DENG Su, MA Wu-bin, WU Ya-hui, ZHOU Hao-hao. Clustered Federated Learning Methods Based on DBSCAN Clustering[J]. Computer Science, 2022, 49(6A): 232-237.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100059

# 基于 DBSCAN 聚类的集群联邦学习方法

鲁晨阳　邓　苏　马武彬　吴亚辉　周浩浩
国防科技大学信息系统工程重点实验室　长沙 410073
(luchenyang97@163.com)

**摘　要**　联邦学习(Federated Learning)是为了解决机器学习中以隐私保护为前提的数据碎片化和隔离问题。各客户端节点在本地训练数据,将训练的模型参数信息上传到中央服务器,由参数服务器聚合参数信息以达到共同训练的目的。由于现实环境中,各节点数据之间的分布往往不一致,通过分析非独立同分布数据对联邦学习准确率的影响,来证明传统联邦学习方法得到的模型精度较低。因此,采用多样化抽样策略模拟数据倾斜度分布,提出了基于 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)聚类的集群联邦学习算法(DBSCAN Based Cluster Federated Learning,DCFL),解决了联邦学习中不同节点的数据非独立同分布降低了学习准确率的问题。在 Mnist 和 Cifar-10 标准数据集上进行了实验,相比传统的联邦学习算法,基于 DBSCAN 聚类的集群联邦学习算法对模型的准确率有较大的提升。

**关键词**:联邦学习;聚类;数据分布;客户端选择;训练优化
**中图法分类号**　TP301

## Clustered Federated Learning Methods Based on DBSCAN Clustering

LU Chen-yang,DENG Su,MA Wu-bin,WU Ya-hui and ZHOU Hao-hao
Science and Technology on Information Systems Engineering Laboratory,National University of Defense Technology,Changsha 410073,China

**Abstract**　Federated learning is to solve the problem of data fragmentation and isolation in machine learning based on privacy protection. Each client node trains the data locally and uploads the model parameter information to the central server,which aggregates the parameter information to achieve the purpose of common training. In the real environment,the distribution of data among nodes is often inconsistent. By analyzing the influence of independent identically distributed data on the accuracy of federated learning,it is proved that the accuracy of the model obtained by the traditional federated learning method is low. Therefore,a diversified sampling strategy is adopted to simulate the data inclination distribution,and a Clustered Federated Learning Methods algorithm based on DBSCAN clustering(DCFL) is proposed,which solves the problem that the learning accuracy is reduced when the data of different nodes are not independently and identically distributed in federated learning. Through the experimental comparison of Mnist and Cifar-10 standard data sets,compared with the traditional federated learning algorithm,DCFL can greatly improve the accuracy of the model.

**Keywords**　Federated learning,Cluster,Data distribution,Client selection,Training optimization

## 1　引言

近年来随着机器学习领域算法的创新和计算机算力的巨大提升,以及大数据研究的兴起,人们广泛认为人工智能迎来了第三个研究高峰。

然而,训练一个成功的模型需要巨大的数据量,过往一些成功案例是伴随着大数据的发展而来的,随着大数据的进一步发展,对数据隐私和安全的重视成为了当前世界性的趋势[1]。各国都在加强对于公民隐私安全的保护,这就给人工智能领域带来了巨大的挑战。在满足数据隐私、安全和监管的前提下,设计一个机器学习框架,让人工智能系统可以获取所需的数据,一个可行的解决方式就是联邦学习。

联邦学习是一种新的分布式机器学习范式,它允许多个设备(在联邦学习中称为客户端)在不需要上传本地数据的情况下共同训练一个全局模型,每一个拥有数据的客户端组织训练一个模型,然后综合各节点模型,将其聚合得到一个全局的模型[2]。在这个过程中,各个客户端交换模型信息的过程将会被精心设计,使得没有组织能够猜测到其他组织的隐私数据内容,这便是联邦学习的核心思想[3]。联邦学习旨在建立一个基于分布数据集的联邦学习模型。在联邦学习模型训练的过程中,模型相关的信息能够在各方之间交换(或者是以加密的形式进行交换),这一类交换不会暴露每个站点上数据的任何受保护的隐私部分。已经训练好的联邦学习模型可以置与联邦学习系统的各参与方,也可以在多方之间共享[4]。

但是,传统的联邦学习应用到非独立同分布数据上时,效果并不理想。实验证明,当联邦学习的各节点的数据分布差异过大时,训练出来的模型精度会大大降低[5]。然而,各节点的数据在现实的产生过程中,可能会受到其他节点或者本地

基金项目:国家自然科学基金面上项目(61871388)
This work was supported by the General Program of National Natural Science Foundation of China(61871388).
通信作者:马武彬(wb_ma@nudt.edu.cn)

环境的影响，各节点的数据往往是非独立同分布的[6]，这就给联邦学习的应用带来了难题，即如何降低数据的非独立同分布带给联邦学习训练精度的影响。

针对以上问题，本文首先采用多样化抽样策略模拟不同分布的数据，探究数据分布倾斜程度对于联邦学习精度的影响。通过对比实验得出，随着数据分布倾斜的加深，联邦学习的模型训练精度变低。为解决在数据极端不平衡时联邦学习模型训练精度低的问题，将各客户端节点本地训练后的模型参数信息采用 DBSCAN 算法进行聚类，划分入不同的簇中，使簇内的节点分布具有更高的相似性，然后在各簇内分别进行联邦学习，最后得到多个适用于本簇的全局模型。本文在本地模拟多个节点和参数服务器，实验结果证明该方法可以有效降低数据非独立同分布给模型训练精度带来的影响，从而产生更精确的模型。综上，本文的主要贡献如下：

(1)设计了实验，证明了在联邦学习中，各节点数据的分布倾斜程度越深，联邦学习训练出来的全局模型精度就越低。

(2)提出了一种新的聚类方式，不需要知晓客户端的本地数据，采用 DBSCAN 的聚类算法对客户端节点的本地训练模型参数信息进行聚类。对比其他聚类算法，DBSCAN 具有可以发现任意形状簇、不需要提前确定簇的数量、可以发现异常点等优点。

(3)采用多样化的抽样策略在多个标准数据集上模拟出 non-IID 分布，对比分析了 DCFL 和 FedAvg 算法的性能。实验结果证明，由 DCFL 得到的模型具有更高的准确率。

## 2 相关工作

为解决日趋收紧的隐私保护要求与机器学习对于大量训练数据需求的矛盾，McMahan 等提出了一种基于迭代模型平均的深层网络联合学习方法[7]，该方法的学习任务通过由中央服务器协调的客户端的松散联合来进行，一个主要优点是将模型训练和直接访问原始训练数据的需求分离开来，这在对数据隐私有严格要求或者数据难以集中共享的场景下有着重大的意义。该算法的流程如下：初始化模型及各个参数，随机选择比例为 $C$ 的客户端($0<C<1$，$C=1$ 表示全部客户端都参与更新)，中央服务器将初始化的模型参数发给选中的客户端，这些客户端基于本地的数据根据接受到的模型参数使用随机梯度下降(Stochastic Gradient Descent，SGD)算法来实现优化。文献[7]第一次提出了联邦学习的概念，开启了联邦学习的相关研究。

随着联邦学习研究的兴起，大量的问题也随之浮现，文献[8]总结了目前联邦学习面临的急切挑战：1)数据的非独立同分布问题；2)个人数据的隐私保护问题；3)有限通信带宽下的训练问题；4)面对恶意节点和攻击的鲁棒性；5)联邦学习中的公平性问题。提高联邦学习的效率和有效性的一个基本挑战就是非独立同分布数据(non-IID)的存在。

数据的非独立同分布情况在现实条件中广泛存在，例如：1)非同分布的客户端分布，因为各客户端的数据是由本地产生的，不同客户端中的样本产生机制可能有差别(如不同国家或者地区)；2)特征分布倾斜(协变量漂移)，如手写字体的识别中，即便是同一个字，不同的人写法也不一样；3)标签分布倾斜(先验概率漂移)，例如中文的使用人群主要在中国，在外国使用的人较少；4)数量倾斜或者不平衡等。现实生活中各种情况都可能导致数据非独立同分布情况的出现。传统的机器学习都是基于数据独立同分布的假设，但是联邦学习不同于集中式的机器学习，在未将数据集中的情况下，每个节点的数据是非独立同分布的[9]。

为了解决联邦学习中数据的非独立同分布问题，Zhao 等对 FedAvg 算法进行了改进[10]，发现在数据处于非独立同分布时，应用 FedAvg 算法会有较高的精度损失。他们提出使用土方运算计算权重散度，能够提高联邦学习在 non-IID 数据中的准确度，并且提出了一种数据共享的联邦学习策略，通过在中央服务器创建所有客户端设备之间全局共享的一小部分数据来改进对 non-IID 数据的训练效果。在中央服务器创建所有客户端设备之间全局共享的数据来改进对 non-IID 数据的训练效果，虽然可以降低数据倾斜带来的影响，但这种方式相当于人为加入了误差，并且共享数据的方式本质上违背了联邦学习对于数据隐私保护的原则，在实施上有很大的困难。

文献[11]认为可以对训练模型进行个性化处理，以减轻异构性，并使每个模型获得高质量的个性化模型，即个性化联邦学习。个性化联邦学习可以分为两步：1)以协作的方式构建一个全局模型；2)使用客户端的私有数据为每个客户端进行全局模型的个性化处理。

Muhammad 等为了提高联邦学习的训练效率，将联邦学习和推荐系统相结合，提出了 FedFast 算法[12]，该算法是 FedAvg 算法的改良版，其基本流程与联邦平均算法相似，主要针对联邦学习的两个重点步骤，对客户端选择和模型聚合进行了改良。在客户端选择上，他们提出了 ActvSAMP 方法，该算法首先采用 $K$-means 方法对不同节点的推荐系统的相似度进行聚类，将所有节点分为不同的类，然后在不同的簇内随机抽取一定数量的节点参与训练。在对被选中的节点进行参数更新的同时，他们提出了 ActvAGG 算法，利用每轮参与了训练的节点更新的梯度信息去更新同簇类未参与训练的节点信息，以达到更快收敛的目的，FedFast 算法提出的主要目的是提高训练的效率，其聚类的方式是对推荐系统的信息进行聚类，对于非结合推荐系统的联邦学习方式不具有普适性，并且使用 $K$-means 方法聚类无法排除离群点的干扰，可能会受到恶意节点的攻击[13]。该方法需要事先指定簇的数量，在实际情况下，中央服务器并不知道客户端的数据分布情况，也就无法事先指定将客户端聚成的簇数量。

文献[14]提出了一种根据节点梯度或者更新信息进行动态划分的算法。该文提出传统的联邦学习都遵循一个核心假设：可以用一个模型满足所有客户端的要求。但事实上这并不准确，首先这个模型不一定足够精确地满足所有客户端的要求，其次各客户端的数据分布不一定相同，因此该文提出了一个新的假设：存在一个合理的划分，使每个划分中的节点都满足传统的联邦学习核心假设。利用各参与者的余弦相似度来进行划分，对于一个分类问题，首先求出所有节点的余弦相似度矩阵，然后将相似矩阵按从小到大的顺序将索引排序，每次取最小且处于不同分组的节点合并，直到最后只剩指定分类的组。这种方式同样也有需要事先指定聚类簇的数量、无法排除离群点干扰的问题。

针对上述研究中存在的问题，本文首先依据不同的抽样策略在多个数据集上模拟出不同分布的数据，实验证明数据分布的倾斜程度会影响模型精度。本文在数据分布极度不均衡的情况下，提出了基于 DBSCAN 聚类的集群联邦学习方法，对客户端模型参数进行聚类，通过先期聚类的方式，使簇内的客户端拥有较高的相似度，以降低数据非独立同分布给模型精度带来的影响。

DBSCAN 是一种基于密度的聚类算法，它将簇定义为密度相连的点的最大集合，将具有足够高密度的区域划分为簇[15]，与 *K*-Means，BIRCH 等只适用与凸样本集的聚类相比，DBSCAN 可以在有噪声的空间数据中发现任意形状的聚类。DBSCAN 聚类方法与 *K*-Means 等方法相比有几点好处：1）DBSCAN 不需要事先知道要形成的簇类数量；2）DBSCAN 可以发现任意形状的簇类；3）DBSCAN 可以发现噪声点，剔除某些恶意攻击的节点的影响。对比上述论文聚类的方法，本文方法不需要事先指定簇的数目，并且可以排除离群点的干扰，在实际中有更广泛的应用。

## 3 基于 DBSCAN 聚类的集群联邦学习方法

### 3.1 联邦学习机制[16]

常见的联邦学习机制由一个参数服务器和多个节点组成，参数服务器负责收集节点上传的梯度信息，并运行优化算法对模型各参数进行更新，计算全局模型和参数；各节点独立地对本地数据进行学习，每轮学习结束后，将学习的梯度信息上传到参数服务器，参数服务器对收集到的梯度信息加权求和，然后更新全局模型参数，将新的全局模型参数分发给下一轮参与训练的客户端[17]，在学习的过程中，数据只会在本地计算，将计算好的梯度信息上传到参数服务器，除了节点共同维护的全局参数，节点不能得到其他节点的任何信息，这样就维护了数据的机密性[18]。

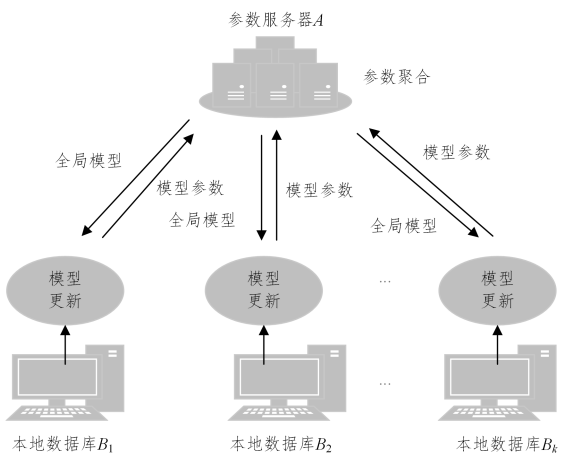



图 1 联邦学习模型

Fig. 1 Model of federated learning

### 3.2 基于 DBSCAN 聚类的客户端选择

在传统的联邦学习算法中，一个很重要的环节是按轮次从所有节点中抽取一定数目的节点来参与训练以改进全局模型。FedAvg 算法采用的是随机抽取的方式，随机地从全部节点中抽取指定数量的节点。该方法在面对独立同分布的数据时非常有效，在面对非独立同分布的数据时，训练的效率和精度都会受到较大的影响（详细数据见实验部分）。数据的分布倾斜越严重，训练的精度越低，如 Mnist 数据集在使用 MLP 模型训练时，一类非独立同分布数据对比独立同分布数据下降了 4.53% 的模型精度，二类非独立同分布数据对比独立同分布数据下降了 15.98% 的模型精度，可见数据的非独立同分布情况极大地影响着联邦学习的训练质量。

在联邦学习的应用场景中，各个节点上的数据是独立产生的，因此各节点的本地数据都不可能代表总体的分布，传统的联邦学习将数据视为独立同分布，只用一个模型去适用所有节点数据的方法是不可行的。为了降低数据的非独立同分布给模型精度带来的影响，在先期对用户进行聚类，然后各个聚类内部共同训练一个全局模型是一个较好的选择。

在相同的神经网络随机种子下，相似数据训练出来的神经网络参数是相似的，同时将训练出来的神经网络参数可视为高维的向量，对高维向量进行聚类，这就给在不需要获取节点数据的情况下对节点进行聚类提供了可能，只要各节点上传本地训练的模型参数信息即可。

为了在所有的客户端节点中找到合适的聚类，所有的节点客户端进行一次充分的本地学习，然后将学习的参数和梯度信息上传到参数服务器，由参数服务器运用 DBSCAN 聚类方法对这些节点的模型参数进行聚类，将所有的节点划分入不同的簇中，然后再在不同的簇中进行联邦学习。

**算法 1** DCFL 算法

输入：C，K，E，B，$\eta$

输出：$w_{c_1}$，$w_{c_2}$，…/ * $w_{c_n}$ 代表第 n 个簇的全局模型参数 * /；

/ * 参数服务器执行的操作 * /；

1. 初始化全局模型参数 $w_0$；
2. for each client $k \in S_t$ in parallel do
3. $w^k \leftarrow$ ClientUpdata(k，$w_0$)；
4. $\Gamma \leftarrow$ DBSCAN($w^1$，$w^2$，…，$w^k$) / * $\Gamma$ 为簇的集合 * /；
5. for each cluster($c_1$，$c_2$，…) $\in \Gamma$ do
   5.1 for each round t = 1，2 … do
      5.1.1. $m \leftarrow \max(C \cdot N, 1)$；
      5.1.2. $S_t \leftarrow$ (random set of m clients)；
      5.1.3. for each client $k \in S_t$ in parallel do
         5.1.3.1. $w_{t+1}^k \leftarrow$ ClientUpdata(k，$w_t$)；
      5.1.4. $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$；
6. ClientUpdata(k，w) / * 在客户端 k 执行的操作 * /；
7. $\beta \leftarrow$ (split $P_k$ into batches of size B)；
8. for each local epoch i from 1 to E do
   8.1. for batch $b \in \beta$ do
      8.1.1. $w \leftarrow w - \eta \nabla \ell(w; b)$；
9. return w to sever.

步骤 1－步骤 5 是参数服务器执行的操作，在初始化模型参数后，将初始化的模型发放给各节点，由节点执行本地训练的操作，然后参数服务器收集节点上传的参数信息，由参数服务器对收到的模型参数进行聚类，并将其划分成不同的簇（$c_1$，$c_2$，…）。步骤 5 是在不同的簇 $c$ 中进行多轮的迭代学习，然后将每轮被选择节点的模型参数按样本权重加权平均，最后得到全局模型；步骤 6－步骤 9 是客户端节点执行的操作，每轮被选择的节点在接收到参数服务器发来的全局模型后，利用本地数据进行本地的迭代训练，将训练后的模型参数再发回给参数服务器。

## 4 实验结果

### 4.1 实验环境

本文所有实验在基于 Inter(R) Core(TM) i9-9900KF CPU @3.60 GHz 处理器的 Ubantu18 64 位操作系统，搭配 GeForce RTX2080TI 显卡的计算机上运行。

### 4.2 实验设置

在本地模拟节点和参数服务器，将数据集的数据按照不同的抽样方式分到不同的节点中，节点在本地训练自己的数据，然后由程序对节点训练出来的模型参数信息进行聚合，并分发给下一轮训练的节点。

为了模拟数据的不同分布情况，采用独立不放回的抽样方式每轮抽取定量的数据分配给随机的客户端，这样每个节点的本地数据分布相同，客户端中的数据属于独立同分布的；将数据集中的数据按其标签大小排序，然后按给定数量划分成不同切片的数据集合，将不同切片的数据随机分配给每个节点，这种方法下的客户端本地数据分布并不相同。在 non-IID 划分中，每个节点只会拥有两种标签的数据，在 non-IID2 划分中，每个节点近似只拥有一种标签的数据，因此数据的倾斜程度更高。

实验中共分出了 100 个节点，每个节点的数据占总数据量的 1%，采用了 Mnist 和 Cifar-10 标准数据集进行实验，本地的神经网络模型采用了 MLP 和 CNN 两种神经网络模型。

### 4.3 数据倾斜程度对联邦学习影响实验

采用 Mnist 和 Cifar-10 数据集，在数据分布上采用 3 种分布方式，分别是独立同分布、一类非独立同分布和二类非独立同分布。采用传统的联邦学习算法即 FedAvg 算法分别在不同的数据分布上运行，对比其测试准确率的变化情况，运行结果如图 2 所示。

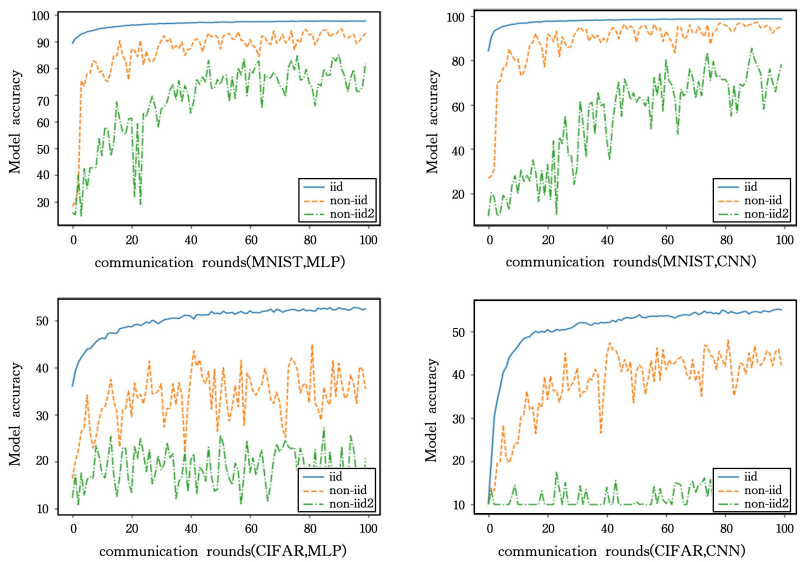



图 2 不同分布的数据集测试精度

Fig. 2 Test accuracy of data sets with different distributions

图 2 给出了 100 轮迭代训练后得到的模型精度，IID 表示数据是独立同分布的，non-IID 和 non-IID2 的数据都是非独立同分布的，但是 non-IID2 的数据倾斜程度比 non-IID 更深。可以看出，随着数据分布的不平衡加深，模型的训练质量也在变低。数据的分布倾斜越严重，训练的精度就越低，如 MNIST 数据集在使用 MLP 训练时，一类非独立同分布数据的精度对比独立同分布数据下降了 4.53%，二类非独立同分布数据的精度对比独立同分布数据下降了 15.98%。可见，数据的非独立同分布情况极大地影响着联邦学习的训练质量。

表 1 不同分布数据 100 轮训练后的精度

Table 1 Accuracy of different distributed data after 100 rounds of training

(单位：%)

| 数据集分布和训练模型 | Mnist | Cifar-10 |
|---|---|---|
| IID(MLP) | 97.64 | 52.52 |
| non-IID(MLP) | 93.11 | 35.44 |
| non-IID2(MLP) | 81.66 | 20.70 |
| IID(CNN) | 98.64 | 55.03 |
| non-IID(CNN) | 95.88 | 42.16 |
| non-IID2(CNN) | 78.23 | 14.60 |

### 4.4 客户端选择实验

首先在 MNIST 数据集上评估 DCFL 算法的性能，MNIST 是一个手写体数据集，一共有 6 万个训练集和 1 万个测试集。

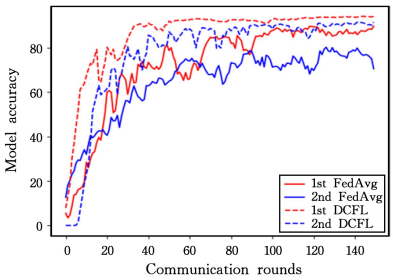
本次实验一共有 100 个节点，为了测试算法在数据分布极不平均的情况下算法的作用，只在 non-IID2 划分的情况下进行实验。联邦学习的输入变量属于优化的范畴，在这里不加以考虑，因此选择默认值。为了测试算法的性能，设置实验以对比 FedAvg 算法和 DFCL 算法训练出来的模型在测试集上的准确率(Test accuracy)。

首先在中央服务器初始化全局模型，然后所有节点接收到初始的全局模型进行充分的本地训练，将本地训练的模型参数上传到中央参数服务器，由参数服务器对节点的模型参数信息进行聚类。

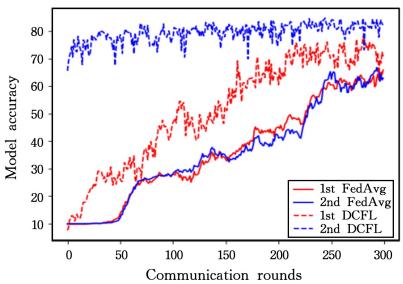
在 MNIST 数据集下，将客户端聚为两个簇时，对比原始 FedAvg 方法使用全部训练集训练的全局模型在各簇测试集上迭代至平稳的准确率，和使用在簇内训练的模型在本簇内的准确率，结果如图 3 所示。

在 non-IID2 型数据上进行联邦学习时，由于各节点本地

数据分布差异过大，学习的曲线容易出现较大的波动，因此在学习时应设置较小的学习率和较大的 $C$（每轮抽样的客户端数占全体客户端的比例）值，由图 3 可以看出，对比传统的联邦学习算法，在 non-IID2 的数据集上，DCFL 具有比 FedAvg 算法更高的准确率，具体数值如表 2 所列。




（a）MLP 模型




（b）CNN 模型

图 3 MNIST 数据集分簇实验

Fig. 3 Mnist data set clustering experiment

表 2 Mnist 数据集上模型的测试准确率

Table 2 Model test accuracy on Mnist

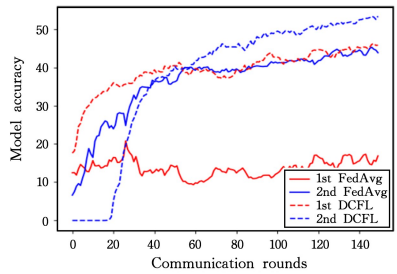
（单位：%）

| 算法和训练模型 | Clust1 | Clust2 |
|---|---|---|
| FedAvg(MLP) | 89.64 | 73.42 |
| DCFL(MLP) | **93.92** | **91.45** |
| FedAvg(CNN) | 65.88 | 62.93 |
| DCFL(CNN) | **71.39** | **82.26** |

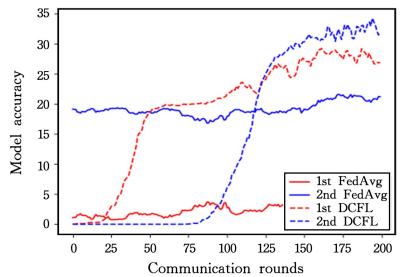
综上，MNIST 数据集在分簇的情况下，每个簇内使用 DCFL 训练出来的模型都比使用传统 FedAvg 算法有较大提升。使用 MLP 模型进行训练时，第一个簇和第二个簇分别提升了4.28%，18.03%；使用 CNN 模型进行训练时，第一个簇和第二个簇分别提升了 5.51%，19.33%。

Cifar-10 数据集是一个用于识别普适物体的小型数据集，一共包含了如飞机、汽车等 10 个类别的 RGB 彩色图片，对比 MNIST 数据集来说，Cifar-10 是 3 通道的彩色 RGB 图像，MNIST 是灰度图像，相比手写字符，Cifar-10 含有的是现实世界中真实的物体，不仅噪声很大，而且物体的比例、特征都不尽相同，这也为识别带来了很大的困难，从图 1 可以看出，面对非独立同分布数据时，采用 FedAvg 算法训练出来的模型精度受到了极大的影响。迭代至平稳后模型的具体测试精度如表 3 所列。

Cifar-10 数据集受数据的非独立同分布影响较大，当数据处于极端的非独立同分布时，模型的精度会急剧下降到 10%～20%，但对模型聚类后，因为簇内的数据差异性得到了降低，在 MLP 模型中训练出来的模型精度分布提高了 25.13%和 10.11%，近似达到了数据独立同分布的水平；在 CNN 模型中，应用 FedAvg 算法时，第一个分组的测试准确率只有 3.56%，这是当节点数据分布太大时可能出现的情况，全局模型并不适用于所有节点。在应用 DCFL 算法进行实验时，测试准确率分别提高了 24.56%和 13.6%。




（a）MLP 模型




（b）CNN 模型

图 4 Cifar-10 数据集分簇实验

Fig. 4 Cifar-10 data set clustering experiment

表 3 Cifar-10 数据集上模型的测试准确率

Table 3 Model test accuracy on Cifar-10

（单位：%）

| 算法和训练模型 | Clust1 | Clust2 |
|---|---|---|
| FedAvg(MLP) | 18.60 | 42.32 |
| DCFL(MLP) | **43.73** | **52.43** |
| FedAvg(CNN) | 3.56 | 20.12 |
| DCFL(CNN) | **28.12** | **33.72** |

**结束语** 本文针对联邦学习中，客户端节点数据非独立同分布对联邦学习模型精度产生影响的问题，提出了一种基于 DBSCAN 聚类的集群联邦学习方法（DCFL）。在联邦学习的客户端节点数据属于非独立同分布的情况下，通过对各客户端节点本地训练的模型参数聚类的方式将各节点划分到不同的簇中，提升了簇内数据的相似度，减轻了数据的不同分布带来的影响，从而提高了模型的训练精度。

然而，本文的工作也存在很多不足，从本质上来说对数据集的分簇是一种需要平衡的事情，簇分得越多，簇内的模型精度就越高，但模型的泛化性就越差，簇分得少，簇内的差异过大，模型的精度也会降低。在使用 DBSCAN 方法聚类时，聚类的选择是由两个参数（扫描半径（eps）和最小包含点数（minPts））决定的，如何获取到一个好的分组与这两个参数的选择有很大的关系，在后期的研究中，会研究参数的自适应选择，以获取最好的分组。

## 参 考 文 献

[1] SHULTZ D. When your voice betrays you[J]. Science, 2015, 347(6221): 494-494.

[2] YANG Q, LIU Y, CHEN T, et al. Federated Machine Learning [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[3] BONAWITZ K. Practical Secure Aggregation for Privacy-Pre-

serving Machine Learning[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.

[4] AI and Data Privacy Protection: The Way to Federated Learning [J]. Journal of Information Security Research, 2019, 5(11): 961-965.

[5] SATTLER F, WIEDEMANN S, MULLER K R, et al. Robust and Communication-Efficient Federated Learning From Non-i. i. d. Data[J]. IEEE Trans Neural Netw Learn Syst, 2020, 31(9): 3400-3413.

[6] LI T, SAHU A K, TALWALKAR A, et al. Federated Learning: Challenges, Methods, and Future Directions [J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[7] MCMAHAN H B, MOORE E, D RAMAGE, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[J]. arXiv: 1602. 05629, 2016.

[8] KAIROUZ P, BRENDAN H. McMahan. Advances and Open Problems in Federated Learning[J]. arXiv: 1912. 04977, 2021.

[9] LI X, HUANG K, YANG W, et al. On the Convergence of FedAvg on Non-IID Data[J]. arXiv: 1907. 02189, 2019.

[10] ZHAO Y, LI M, LAI L, et al. Federated Learning with Non-IID Data[J]. arXiv: 1806. 00582, 2018.

[11] JIANG Y, KONEN J, RUSH K, et al. Improving Federated Learning Personalization via Model Agnostic Meta Learning[J]. arXiv: 1909. 12488, 2019.

[12] MUHAMMAD K. FedFast: Going Beyond Average for Faster Training of Federated Recommender Systems[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.

[13] GHOSH A, HONG J, YIN D, et al. Robust Federated Learning in a Heterogeneous Environment[J]. arXiv: 1906. 06629, 2019.

[14] SATTLER F, MULLER K R, SAMEK W. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints [J]. IEEE Transactions on Neural Network Learning and Systems, 2021, 32(8): 3710-3722.

[15] ESTER M, KRIEGEL H P, SANDER J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// AAAI Press. 1996.

[16] KONEN J, MCMAHAN H B, YU F X, et al. Federated Learning: Strategies for Improving Communication Efficiency[J]. arXiv: 1610. 05492 2016.

[17] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards Federated Learning at Scale: System Design [J]. arXiv: 1902. 01046, 2019.

[18] MCMAHAN H B, RAMAGE D, TALWAR K, et al. Learning Differentially Private Recurrent Language Models [J]. arXiv: 1710. 06963, 2017.

**LU Chen-yang**, born in 1997, postgraduate. His main research interests include federated learning and machine lear-ning.

**MA Wu-bin**, born in 1986, Ph. D, lecturer. His main research interests include data engineering and cyber-physical systems.

(上接第 231 页)

[39] PĂUN A, PĂUN R. Membrane computing and economics: Numerical P systems [J]. Fundamenta Informaticae, 2016, 73(1/2): 213-227.

[40] ZHANG Z, SU Y, PAN L Q. The computational power of enzymatic numerical P systems working in the sequential mode[J]. Theoretical Computer Science, 2018, 724: 3-12.

[41] PAN L Q, ZHANG Z, WU T F. Numerical P systems with production thresholds [J]. Theoretical Computer Science, 2017, 673: 30-41.

[42] LIU L, YI W, YANG Q. Numerical P systems with Boolean condition [J]. Theoretical Computer Science, 2019, 785: 140-149.

[43] WU T F, PAN L Q, YU Q, et al. Numerical Spiking Neural P Systems. IEEE transaction on neural networks and learning systems [J]. IEEE Transaction on Neural Networks and Learning Systems, 2020(99): 1-15.

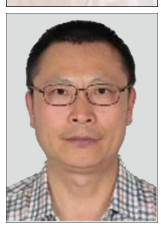
[44] KOREC I. Small universal register machines [J]. Theoretical Computer Science, 1996, 168: 267-301.

[45] YIN X, LIU X. Dynamic Threshold Neural P Systems with Multiple Channels and Inhibitory Rules [J]. Processes, 2020, 8(10): 1281.

[46] PENG H, LI B, WANG J. Spiking neural P systems with inhibitory rules [J]. Knowledge-Based Systems, 2020, 188: 105064.

[47] WU T, ZHANG T, XU F. Simplified and yet Turing universal spiking neural P systems with polarizations optimized by anti-spikes [J]. Neurocomputing, 2020, 414: 255-266.

**YIN Xiu**, born in 1996, postgraduate. Her main research include membrane computing, data mining and machine learning.



**LIU Xi-yu**, born in 1964, Ph. D, professor, doctoral supervisor. His main research include membrane computing, data mining and machine learning.