doi:10.3969/j.issn.1672-1292.2022.02.012

隐私保护下的车辆轨迹联邦嵌入学习与聚类

孔秀平[1],陆　林[2]

(1.扬州工业职业技术学院信息中心,江苏 扬州 225127)
(2.中电云数智科技有限公司,湖北 武汉 430056)

[摘要]　智能网联汽车的高维轨迹数据被广泛用于从车辆的行驶轨迹中发现不同运动模式,从而降低交通风险、提高通行效率. 然而,数据利用过程中的隐私问题日益受到关注,如何在隐私保护的前提下进行算法的研究和应用是当前面临的一大挑战. 针对车辆轨迹数据分散在不同持有方且出于隐私保护无法共享数据的背景,利用差分隐私联邦学习框架来构建序列自编码网络提取轨迹序列的低维表示,并进一步利用轨迹的低维空间向量来发现不同时段下车辆的频繁路线. 提出的框架既通过本地训练避免了用户隐私数据的分享,又能通过高斯差分隐私机制防止模型信息的泄露. 该框架在真实的轨迹数据集上进行了验证,利用 LSTM 自编码作为嵌入学习网络,与非联邦、非差分加密的模型进行了对比分析,最后对三种得到的轨迹嵌入通过聚类分析发现该框架下学习的模型在充分尊重了隐私保护的前提下,仍然能够找出有效的频繁轨迹.

[关键词]　序列自编码,联邦学习,差分隐私,轨迹聚类
[中图分类号]TP311.1　[文献标志码]A　[文章编号]1672-1292(2022)02-0080-07

Privacy-preserved Vehicular Trajectory Embedding Federated Learning and Clustering

Kong Xiuping[1],Lu Lin[2]

(1.Department of Information Center,Yangzhou Polytechnic Institute,Yangzhou 225127,China)
(2.China Electronic Cloud Digital Intelligence Technology Co.,Ltd.,Wuhan 430056,China)

Abstract:High dimensional trajectory data of intelligent networked vehicles are widely used to find different motion patterns from vehicle trajectories,so as to reduce traffic risk and improve traffic efficiency. However,more and more attention has been paid to the privacy problem in the process of data utilization. How to research and apply the algorithm under the premise of privacy protection is a big challenge. In view of the background that vehicle trajectory data are scattered among different owners and cannot be shared due to privacy protection,this paper uses differential privacy federated learning framework to construct a sequence autoencoding network to extracting low dimensional representation of trajectory sequence,and such latent representation is further used to find frequent vehicle routes in different periods. The proposed framework not only avoids the sharing of user privacy data through local training,but also prevents the disclosure of model information through Gaussian differential privacy mechanism. The framework is validated on real trajectory data sets,using LSTM autoencoder as embedding learning network,and compared with non-federated and non-differential encryption models. Finally,through clustering analysis,it is found that the learning model under the framework can still find effective frequent trajectories under the premise of fully respecting privacy protection.

Key words:sequential autoencoder,federated learning,differential privacy,trajectory clustering

随着网联汽车的普及,车联网技术可以采集到大量的车辆时空轨迹数据. 对时空轨迹数据的挖掘和语义推理,既可以发现频繁的行驶路线或异常车辆,也可以帮助推理用户的出行意图,无论对物流行业、交管部门还是运营商来说,都具有很大的应用价值.

本文旨在利用某城市电动网约车实际运行数据,发现不同时段下的频繁运营路线,从而辅助智能交通决策与提升车辆运营效率. 发现频繁路线的通常做法是应用空间聚类算法[1],并将聚类的结果进行过滤

收稿日期:2021-08-31.
基金项目:国家自然科学基金面上项目(61902070).
通讯作者:孔秀平,硕士,工程师,研究方向:计算机网络与隐私保护. E-mail:xiu651015722@qq.com

得到满足要求的频繁线路. 但传统方法需要直接利用原始的轨迹经度、纬度数据来计算不同行驶轨迹之间的相似度,从而忽略了信息共享的隐私问题[2-4].

智能网联汽车的信息共享是富有挑战性的,因为人们日益感觉到数据安全和隐私的重要性[5]. 一般来说,不同渠道收集的不同车辆的状态数据、行驶轨迹都属于个人隐私,信息共享将带来安全隐私隐患. 另外,出租车、网约车包含乘客的敏感信息,例如住所、上班地和通勤路线等,这一事实需要在训练过程中考虑隐私.

针对这一挑战,一种可行的方法通过应用扰乱技术进行隐私保护式分布式聚类[6],但是其隐私保护程度有限. 另一种是通过潜空间表征[7],将原始轨迹压缩到统一维度的特征空间,表示为只有机器可以理解的特征向量. 然后再构建基于潜空间特征的机器学习算法,来解决相应的分类与聚类问题,从而避免对原始数据的直接利用敏感问题. 然而该轨迹表示的训练过程仍然需要数据拥有者分享自身的数据进行集中式学习,仍然存在隐私泄露的风险. 本文提出了一种差分联邦学习的轨迹嵌入学习方法,在充分保护用户隐私的前提下发现个体用户不敏感的频繁路线. 本文的主要贡献包括:

(1)提出了利用横向联邦学习框架在各方不共享原始轨迹数据的前提下,共同学习轨迹自编码模型.

(2)针对解码器可能还原具体用户原始轨迹的问题,利用差分隐私算法对模型进一步加密,避免模型参数泄露.

(3)在真实数据集上对差分后的轨迹嵌入进行了聚类分析,通过对比发现其聚类效果与无隐私保护的方法效果一致.

1 相关工作

1.1 基于轨迹嵌入表示的聚类

Yao 等[8]尝试将每条轨迹转换成固定长度的表示形式,从而很好地编码对象的移动行为. 一旦学习到高质量的轨迹表示,就可以根据实际需要轻松地应用任何经典的聚类算法. Yue 等[9]提出了一种用于移动行为聚类的无监督神经方法-DETECT. 它利用 LSTM 自编码[10]学习了行为潜在空间中轨迹的强大表示,从而使聚类算法(例如 k 均值)得以应用.

给定一个自编码模型 M,它包括编码器 M_{enc}和解码器 M_{dec}. 该模型学习的目标是给定任意变长的轨迹序列 $x \in \mathbf{R}$,学习其定长的嵌入变量 $z \in \mathbf{R}^d$(d 为嵌入空间维度)使得其尽可能包含原有轨迹的信息,表示为:

$$x \approx M_{\text{dec}}(z = M_{\text{enc}}(x)). \tag{1}$$

利用自编码模型将所有车辆的行程轨迹映射到嵌入空间 Z 后,就可以利用传统的聚类算法基于相似度度量进行自动分组. 每一组的中心向量,经过解码器还原后就代表了一条频繁的运行路线.

最新的实验结果表明,序列自编码已经是轨迹序列表征学习的 state-of-art 方法. 然而,上述研究中需要对所有车辆的轨迹进行集中学习,也就是所有车辆都要分享自己的 GPS 数据,这样明显暴露了各用户的相关隐私,比如家庭住址、上班场所以及通勤路线等信息[10].

1.2 基于差分隐私的联邦学习

联邦学习(federated learning,FL)[11-12]模式因其允许数据拥有者在不共享原始数据的前提下多方共同参与进行机器(深度)学习得到了广泛关注. 假设一共有 n 个数据拥有者,对应的数据集为 $D = \{D_1, \cdots, D_n\}$,联邦机器学习的主要思想是每个数据拥有者 k 先接收第三方聚合者的初始化全局模型 M_{FED},接着利用自身的数据进行本地训练得到局部模型 M_k,然后仅通过传递参数来更新全局模型. 聚合者接收数据拥有者的模型参数进行聚合,得到更新后的全局模型. 该过程一直持续直到达到终止条件为止.

虽然联邦学习有效避免原始数据的传输和泄露,但是在深度学习模型的联邦训练过程中,仍然存在梯度泄露风险. 针对这一问题,相关的研究利用差分隐私(differential privacy,DP)来解决. 差分隐私最早由 DWORK[13]于 2006 年提出,典型的定义如下:

给定随机化算法 A,对于任意两个相邻数据集 D,D 和 A 可能的输出 O,算法 A 符合(ε,δ)-DP,满足

$$\mathbb{P}[A(D) \in O] \leqslant \mathrm{e}^{\varepsilon}\mathbb{P}[A(D')] + \delta. \tag{2}$$

式中,δ 表示允许 ε-DP 被打破的概率,ε 表示保护级别或隐私预算,ε 越小表示隐私保护级别越高. 此定

义可确保基于算法的输出,对手具有有限的(取决于 ε,δ 的大小)识别任何个人存在或不存在的能力.

深度学习中实现 DP 的主要方式是对模型添加一定的噪声来满足[14]. 常用的噪声机制包括拉普拉斯和高斯噪声. 以高斯差分隐私(Gaussian DP,GDP)为例,高斯机制从服从均值为 μ,标准差为 σ 的分布 $N(\mu,\sigma^2)$ 中,随机采样添加到每一维的模型参数中. 通过本地化差分隐私,每个客户端的梯度上传的第三方可以是不可信的. 该梯度下降算法满足 (ε,δ)-DP,最终模型参数的误差趋近于常数. 相较于同态加密、密钥共享和混淆电路灯隐私保护技术,差分隐私因其计算和传输成本低的优势成为了联邦学习研究的新方向.

2　相关问题与研究方法

2.1　问题定义

定义 1　轨迹数据　考虑一组移动车辆集合,表示为 $V=\{v_1,v_2,\cdots,v_l\}$. 某车辆的轨迹 $\boldsymbol{TR}$ 定义为其在时间维度上的一系列位置点. 位置点为包含时间戳、经度和纬度的三元组,表示为 $\boldsymbol{p}=$<time,lat,lon>. 对于每个对象 v_i,其历史序列可表示为 $\boldsymbol{TR}_{v_i}=(\boldsymbol{p}_{i1},\boldsymbol{p}_{i2},\cdots,\boldsymbol{p}_{im})$.

定义 2　周期时段轨迹　首先将车辆的轨迹点根据时间解析出所在星期,然后给定时间槽(表示为 Δ),将该位置点映射到当天所在的时段. 比如时间槽为 0.5 h,那么一天被划分为 48 个时段. 周期时段轨迹则是将车辆轨迹按星期时段维度进行分组得到的轨迹序列.

定义 3　轨迹嵌入表示　根据定义 1 和定义 2,所有车辆的周期时段轨迹集用 $\boldsymbol{\Psi}=(trip1,trip2,\cdots,tripm)$ 表示. 因为轨迹的变长特性,我们定义轨迹的嵌入映射 $\boldsymbol{\Psi}\rightarrow\mathbf{R}^{m\times d}$,其中 $\mathbf{R}$ 为嵌入空间,d 为嵌入向量维度. 经过该变换每一段轨迹都会被表示一个 d 维的向量表示.

基于以上定义,本文需要解决的问题是,给定任意周期时段内车辆的运行轨迹,找出其中的频繁路线. 该问题的挑战主要是:(1)车辆原始轨迹不能分享以进行集中式的学习. (2)学习到的嵌入需要加密以保证个体用户的轨迹信息不被查询到. (3)最终聚类的效果要保证较高的精确度,即需要隐私保护和模型精度之间的折衷.

2.2　轨迹嵌入的差分联邦学习框架

针对上述问题,本文提出了图 1 所示的差分隐私联邦轨迹嵌入聚类(DP FL of trajectory embedding for clustering,DP-FL-TE4C)框架. 图 1 主要包括了差分隐私联邦轨迹嵌入(DP-FL-TE)学习与轨迹嵌入聚类和解码两部分.

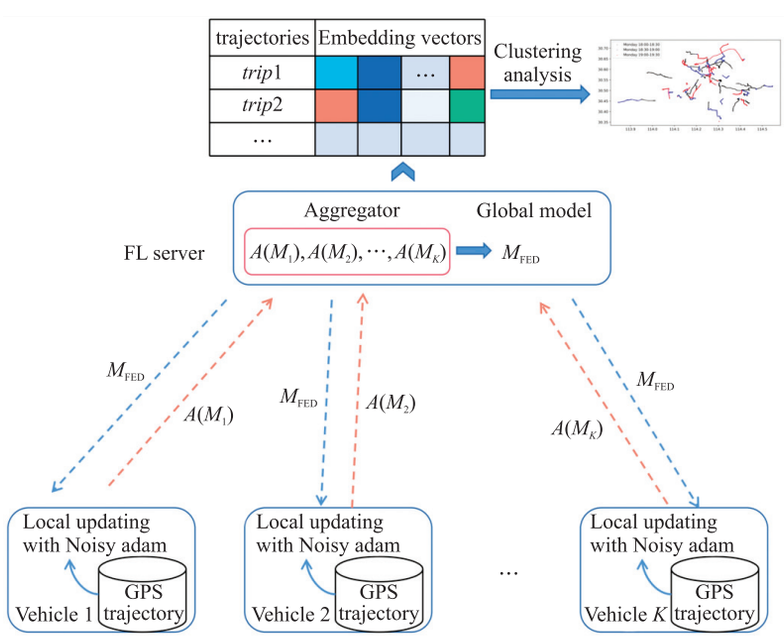

图 1　差分隐私联邦轨迹嵌入聚类框架

Fig. 1　DP FL of trajectory embedding for clustering framework

框架第一部分主要是包括一个第三方服务器(FL Server)和多辆车辆组成的联邦学习流程. 这里假设车辆具有边缘计算能力,他们有选择地参与序列自编码模型的训练,并将添加噪声后的梯度上传给服务器. 第三方服务器在云端进行云计算,它主导模型的联邦训练过程. 它首先将模型及数据处理逻辑发送给图 1 中各个数据持有方,然后接收各本地结点训练好的模型参数,进行聚合得到一个全局最优的模型.

该框架第二部分的工作是 FL Server 接受各方发送的差分加密后的轨迹嵌入向量,应用无监督聚类算法得到频繁的簇,最后对频繁簇的中心嵌入向量应用解码器还原.

2.3 轨迹序列自编码模型

本文提出的框架中使用 LSTM 通过重建轨迹的 GPS 序列来生成固定长度的深度表示的嵌入向量. 在实践中,LSTM 作为循环神经网络(RNN)的变体,通过加入门机制,能更好地捕获时序结构.

假设轨迹自编码模型为 $M(\theta)$,由循环编码器和循环解码器组成,其中 θ 为模型参数. 与通用自动编码器的机制类似,LSTM 编码器的任务是将输入的时段轨迹编码为潜在嵌入,然后使用递归解码器,仅从嵌入中重建轨迹. 对模型进行训练以最大程度地减少重建误差,从而学习具有代表性的嵌入,该嵌入可以完全捕获轨迹中的运动过渡和上下文. 将 $x_i \in \{x_i^{(t)}\}_{t=1}^{T_i}$ 视为长度为 T_i 的(填充)轨迹. x_i 被顺序地馈送到由几个 LSTM 单元组成的循环编码器. 编码器使用每个传递单元 t 更新隐藏状态 $h_{\text{enc}}^{(t)}$ 和其他参数,$h_{\text{enc}}^{(t)} = \sigma(h_{\text{enc}}^{(t-1)}, x^{(t)})$,其中 σ 是神经层激活函数. 最后的隐藏状态 $h_{\text{enc}}^{(t)}$ 称为潜在嵌入 z_i,并被假设为概括表示整个轨迹序列所需的信息. 接下来,解码器尝试以 $h_{\text{enc}}^{(T_i)}$ 为初始状态来重建轨迹,$h_{\text{dec}}^{(1)} = h_{\text{enc}}^{(T_i)}$,借助 $h_{\text{dec}}^{(1)}$ 解码器生成 $\hat{x}^{(1)}$,随后的隐藏状态将以递归方式生成 $h_{\text{dec}}^{(t)} = \sigma(h_{\text{dec}}^{(t-1)}, \hat{x}^{(t-1)})$,$\hat{x}^{(t)} = \sigma(h_{\text{dec}}^{(t)})$,对轨迹数据、编码器和解码器一起训练,训练目标是以最大程度地减少重构误差. 解码器可以从潜在嵌入 z_i 中重建整个增强轨迹,该 z_i 隐式编码了地理环境的过渡模式.

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T_i}\sum_{i=1}^{T_i}(x_i^{(t)} - \hat{x}_i^{(t)})^2. \tag{3}$$

2.4 差分隐私联邦训练

本文提出的 DP-FL-TE 实现如算法 1 所示,它主要包括本地局部更新、差分隐私保护机制和云端模型聚合 3 个部分.

算法 1 差分隐私联邦轨迹嵌入学习算法

输入:Dataset $D=\{D_1,\cdots,D_K\}$ corresponding to vehicle set V, loss function $\ell(\theta,x)$

初始化参数:initial weights θ_0, learning rate η, communication rounds T, noise scale γ, gradient norm bound R, local updating epochs E, a small constant ξ.

输出:$M(\theta_{T+E})$

For $t=0,\cdots,T-1$ do
  Take a subsampled client from K clients
  For each V_k in this subsample in parallel do
    Set $M_k(\theta_t)=M(\theta_t)$
    For $j=0,\cdots,E-1$ do //Local update E epochs
      For each $x_i \in D_k$ do
        $v_{t+j}^{(i)} \leftarrow \nabla_\theta \ell(\theta_t, x_i)$
        $\bar{v}_{t+j}^{(i)} \leftarrow v_{t+j}^{(i)} / \max\{1, v_{t+j}^{(i)}{}_2 / R\}$ //梯度裁剪
        $\tilde{v}_{t+j} \leftarrow \frac{1}{|D_k|}\left(\sum_{x_i \in D_k}\bar{v}_{t+j}^{(i)} + \gamma R \cdot \mathrm{N}(0,I)\right)$ //应用(ε,δ)-GDP mechanism
        $m_{t+j} \leftarrow \beta_1 m_{t+j-1} + (1-\beta_1)\tilde{v}_{t+j}$
        $\mu_{t+j} \leftarrow \beta_2 \mu_{t+j-1} + (1-\beta_2)(\tilde{v}_t \odot \tilde{v}_t)$
        $w_{t+j} \leftarrow m_{t+j} / (\sqrt{\mu_{t+j}} + \xi)$
        $\theta_{t+j+1} \leftarrow \theta_{t+j} - \eta w_{t+j}$
    Return private θ_{t+E}^k to server side
    Update global model $M(\theta_{t+E})$ with federated average of $\{\theta_{t+E}^k\}_{k=1,\cdots,K}$

2.4.1　本地局部更新

在 FL 模式下,假设有 K 辆车参与边缘计算,那么每辆车(比如第 k 个)利用自身数据会得到一个局部模型 $M_k(\theta_k)$. 以第 t 次局部更新为例,服务器向所有参与车辆广播最新参数集 θ^t. 接着,每辆车让 $\theta_k^t=\theta^t$. 然后为了节省通信开销,我们采用局部更新策略. 每辆车在本地执行 E 次训练,每次在数据集 D_k 上使用小批量训练样本 D_k'训练. 数据持有方采用局部随机梯度下降的方法来进行本地模型优化,优化算法可以是 SGD 或者 Adam.

2.4.2　差分隐私保护

对模型成功的攻击使攻击者能够与数据集中的具体记录建立链接,从而在记录包含敏感信息时导致隐私泄漏. 通过这些记录链接,可以根据乘客的轨迹分析出工作地点、家庭住址、活动模式和其他敏感信息. 本文与传统局部更新唯一的区别是采用 Noisy Adam[15]算法来实现差分隐私,使得向 FL Server 分享的是隐私保护的模型信息.

算法 1 中的步骤 8 和步骤 9 为差分机制的实现过程. 步骤 8 对更新后的梯度进行裁剪,使得梯度不超过阈值 R. 然后步骤 9 对每一维的梯度参数添加高斯噪声 $\gamma R\cdot \mathrm{N}(0,I)$,其中 γ 为噪声倍数常量. 算法 1 基于严格的差分隐私定义. 根据文献[16],可以证明算法 1 中每个客户端 k 的局部训练符合(ε,δ)-DP. 进一步地,根据序列合成定理,可以推导出 K 个局部模型聚合后的输出满足 $\left(\sum_{k=1}^{K}\varepsilon_k,\delta\right)$ - DP . 最终,可以证明算法 1 满足差分隐私.

2.4.3　云端模型聚合

服务器端接收到这 K 个模型参数后,通过聚合算法合并为一个全局模型 M_{Fed}. 本文采用经典的联邦平均方法来聚合. 以第 t 轮联邦训练为例,云端接收 K 辆车的局部带“噪声”的参数并进行聚合方式为

$$\theta^{t+E}\leftarrow\sum_{k=1}^{K}p_k\bar{\theta}_k^{t+E}. \tag{4}$$

式中,p_k 是第 k 辆车 V_k 的权重,使得 $p_k\geqslant 0$ 且 $\sum_{k=1}^{K}p_k=1$.

2.5　聚类分析

联邦训练结束后,每辆车仅利用编码器将带噪声轨迹嵌入 $z_i=A(M_{\mathrm{FED}}(x_i))$ 上传到 FL Server. 待时段轨迹 $x_i,x_i,\cdots,x_n$ 映射到其相应的定长嵌入 $z_i,z_i,\cdots,z_n$ 后,本文用一个聚类函数去输出 c 个聚类,每个聚类由其质心 $\mu_j\in\{\mu_j\}_{j=1}^{c}$表示. 具体地,我们采用 GMeans 算法,该算法的思想是通过统计检验发现适当数量的聚类,以决定是否将 K 均值中心分割为两个中心. 然后,对每个组统计其中的轨迹计数,当计数大于某阈值时,则认为该分组为一条频繁路线的表示.

3　实验结果与分析

3.1　数据集与参数设定

本文使用一份真实数据集来进行测试,该数据集来自武汉市的 88 辆纯电动网约车,车辆 GPS 传感器每 10 s 获取当前的位置坐标. 数据跨度是从 2018 年 3 月 1 日到 31 日,共抽取到 1 008 条有效轨迹,这些轨迹几乎覆盖了城市所有主干道路,有利于发现用户不同时段的网约出行规律. 本文设置时段 $\Delta=30$ min,那样每个时段最多采集 180 条记录,也是自编码网络输入的最大序列长度. 整个数据集利用固定的随机种子进行二八划分,取 20%放在云端 server 作为测试数据,剩余 80%被划分到 10($K=10$)个客户端来模拟联邦学习.

本文采用了 FedML 框架进行仿真,利用 Facebook 开源的 Opacus 实现差分隐私,具体网络模型使用 Pytorch 来实现. LSTM 编码器的输入为(180,2),输出的嵌入变量 d 为 200,解码器输入输出与编码器刚好相反. 轨迹数据事先经过了归一化,故最后对解码输出加入了 Sigmoid 激活函数. 训练过程中的批大小为 64,非 DP 的方法使用 Adam 优化器,学习率为 0.001,局部迭代次数为 20,最大轮询次数为 1 000,损失函数使用均方误差(mean squared error,MSE).

3.2 不同客户端的效果对比

首先,我们探讨了不同的客户数量对联邦模型在测试集上性能的影响. 仿真结果如图 2 所示,其中显然的 NonFL-TE 在测试集上的损失最低为 0.013 8%. 其次,FL-TE 最低的损失在 0.07% 左右. DP-FL-TE 损失与 FL-TE 相关不大. 另外,可以看到随着客户端数目的增加,FLServer 聚合的全局模型性能逐步提高. FL-TE 和 DP-FL-TE 在客户端数目为 9 时验证损失最小.

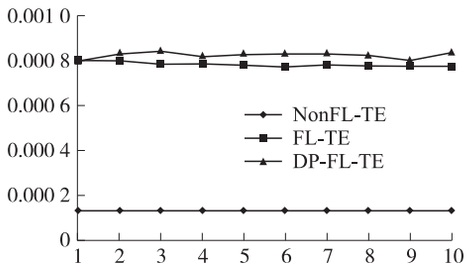

图 2 不同参与客户端数目下的模型在测试集上的 MSE 效果对比

Fig. 2 Comparison of MSE on the test set of models with different number of participating clients

3.3 训练过程对比分析

图 3 显示了 3 种模型在训练集上的平均训练损失情况. 可以看到,NonFL-TE 模型收敛最快且训练损失最低. FL-TE 收敛速度次之,因为其每次训练集的批次受限于本地数据集且利用联邦平均来优化多个局部模型. DP-FL-TE 的联邦平均损失在前两轮急剧下降,最终收敛速度最慢,不过收敛时的训练损失与 FL-TE 相差不大.

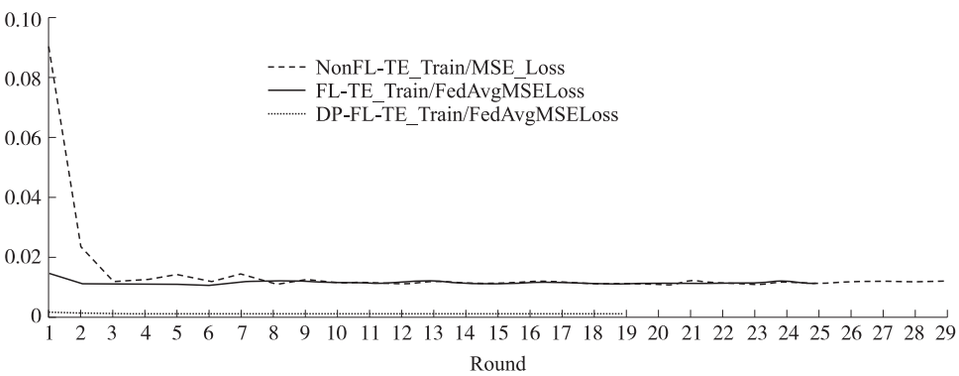

图 3 不同模型的训练损失对比

Fig. 3 Training loss comparison of different models

我们观察到:(1)在差分隐私联邦模式下,聚合模型虽然精度有所损失,但最终依然能够有效地收敛. (2)模型精度的提高会导致隐私预算上升,也就是模型信息泄露的风险增加. 通过隐私分析结合训练损失,可以辅助确定两者的折衷点.

3.4 频繁轨迹聚类挖掘

本文的最终目的是对隐私轨迹进行聚类,对学习到的轨迹嵌入进行聚类,对比非联邦模式下的结果来验证 DP-FL-TE 的精度损失是否影响到最终聚类结果的有效性. 探索使用该框架来聚类将所有车辆轨迹数据,将相似的轨迹组合在一起,从而发现频繁行驶路线.

具体验证手段为,对 3 种模型输出的嵌入表示分别利用 Gmeans 算法进行自动聚类,最大簇数目为 100,然后将聚类簇大小不小于 10 条轨迹的结果,作为频繁轨迹集合. 令 3 种模型输出的聚类结果分别为 $\boldsymbol{C}$,$\boldsymbol{C}_{\mathrm{Fed}}$,$\boldsymbol{C}_{\mathrm{DPFed}}$,对应的轨迹嵌入矩阵为 $\boldsymbol{Z}$,$\boldsymbol{Z}_{\mathrm{Fed}}$,$\boldsymbol{Z}_{\mathrm{DPFed}}$,然后将($\boldsymbol{Z}$,$\boldsymbol{C}$)、($\boldsymbol{Z}$,$\boldsymbol{C}_{\mathrm{Fed}}$)、($\boldsymbol{Z}$,$\boldsymbol{C}_{\mathrm{DPFed}}$)作为输入,使用 Silhouette 系数、Calinski Harabasz score(CH_score)和 Davies-Bouldin score(DB_score)进行聚类效果验证. 这 3 种度量方法中,前两个方法的值越高表示聚类效果越好,第三个则相反.

表 1 中列出了这三种模型进行 10 次聚类后的结果. 可见,3 种模型输出的频繁簇数目大致相近,且非联邦模型聚类效果最好. 以非联邦模型聚类结果为基准,可以发现联邦模型输出聚类结果在非联邦嵌入上的划分性能虽然有所下降,但下降幅度不大,且有趣的是 DP-FL-TE4C 稍微好于 FL-TE4C. 这是因为一方面通过图 3 可以看到 DP-FL-TE 收敛时的训练损失水平与 FL-TE 模型的几乎相同,甚至略优. 另一方面,DP-FL-TE4C 通过添加高斯噪声,使得轨迹自编码网络相对 FL-TE4C 具有更好的泛化能力,使得来自不同车辆相似的轨迹在潜空间上的生成表示更为接近.

表 1　3 种模型聚类效果对比

Table 1　Comparison of clustering effect of three models

	Clusters	Silhouette	CH_score	DB_score
NonFL-TE4C	44±2.46	0.46±0.01	1130.37±66.79	0.69±0.02
FL-TE4C	44±3.09	0.41±0.02	904.00±96.14	0.80±0.04
DP-FL-TE4C	45±2.15	0.42±0.02	1032.38±63.01	0.74±0.03

4　结论

本文提出了一种隐私保护的车辆轨迹深度表征框架. 该框架一方面使用序列自编码网络学习原始 GPS 轨迹的嵌入表示,解决了序列长度不一致(高维度)问题. 另一方面通过联邦学习模式避免用户隐私的直接暴露,并进一步利用差分隐私来避免模型参数的第三方攻击.

为了证明该框架的有效性,利用实际数据集通过联邦学习仿真,比较了该框架下模型的嵌入学习效果. 最后将其应用到实际场景中,通过对真实数据集上的轨迹嵌入学习以及聚类效果对比,可获得有用的频繁轨迹簇,即频繁路线.

[参考文献](References)

[1] ATEV S,MILLER G,PAPANIKOLOPOULOS N P. Clustering of vehicle trajectories[J]. IEEE Transactions on Intelligent Transportation Systems,2010,11(3):647-657.

[2] BIAN J,TIAN D Y,TANG Y Y,et al. A survey on trajectory clustering analysis[J]. arXiv Preprint,2018:1-40.

[3] 柳盛,吉根林. 空间聚类技术研究综述[J]. 南京师范大学学报(工程技术版),2010,10(2):57-62.

[4] 陈铭,吉根林. 一种基于相似维的高维子空间聚类算法[J]. 南京师大学报(自然科学版),2010,33(4):119-122.

[5] BRELL T,BIERMANN H,PHILIPSEN R,et al. Conditional privacy:users' perception of data privacy in autonomous driving[C]//Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems. Anchorage,USA:IEEE,2019:352-359.

[6] 姚瑶,吉根林. 面向垂直划分数据库的隐私保护分布式聚类算法[J]. 南京师范大学学报(工程技术版),2008,8(4):099-102.

[7] KYUNGHY N C,BART V M,CAGLAR G,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv Preprint arXiv:1406.1078,2014.

[8] YAO D,ZHANG C,ZHU Z H,et al. Trajectory clustering via deep representation learning[C]//International Joint Conference on Neural Networks. Anchorage,USA:IEEE,2017.

[9] YUE M X,LI Y G,YANG H Z,et al. DETECT:deep trajectory clustering for mobility-behavior analysis[C]//2019 IEEE Internation Conference on Big Data,Los angeles,USA:IEEE,2019.

[10] 张新峰,闫昆鹏,赵珣. 基于双向 LSTM 的手写文字识别技术研究[J]. 南京师大学报(自然科学版),2019,42(3):58-64.

[11] 杨强,刘洋,陈天健,等. 联邦学习[M]. 北京:电子工业出版社,2020.

[12] YANG Q,LIU Y,CHEN T J,et al. Federated machine learning:concept and applications[J]. ACM Transactions on Intelligent System Technology,2019 10(2):1-19.

[13] DWORK C,ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science,2014,9(3/4):211-407.

[14] ABADI M,CHU A,GOODFELLOW I,et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York,USA:Association for Computing Machinery. 2016:308-318.

[15] BU Z Q,DONG J S,LONG Q,et al. Deep learning with Gaussian differential privacy[J]. arXiv Preprint arXiv:1911.11607,2019.

[16] MIRONOV I,TALWAR K,ZHANG L. Rényi differential privacy of the sampled Gaussian mechanism[J]. arXiv Preprint arXiv:1908.10530,2019.

[责任编辑:陈　庆]