CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

隐私保护下的车辆轨迹联邦嵌入学习与聚类

Privacy-preserved Vehicular Trajectory Embedding Federated Learning and Clustering

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算法的研究和应用是当前面临的一大挑战。本文针对车辆轨迹数据分进行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。本文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。

问题

定义1 轨迹数据。考虑一组移动车辆集合，表示为$V=\{v_1,v_2,\ldots,v_l\}$。某车辆的轨迹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$p=<time, lat, lon>$。对于每个对象v_i，其历史序列可表示为$\mathrm{TR}_{v_i}=(p_{i1},p_{i2},\ldots,p_{im})$。

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。比如时间槽为半个小时，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2，所有车辆的周期时段轨迹集用$\Psi=(trip_1, trip_2, \ldots, trip_m)$表示。因为轨迹的变长特性，我们定义轨迹的嵌入映射$\Psi\rightarrow R^{m\times d}$，其中$R$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。

方法

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两部分。

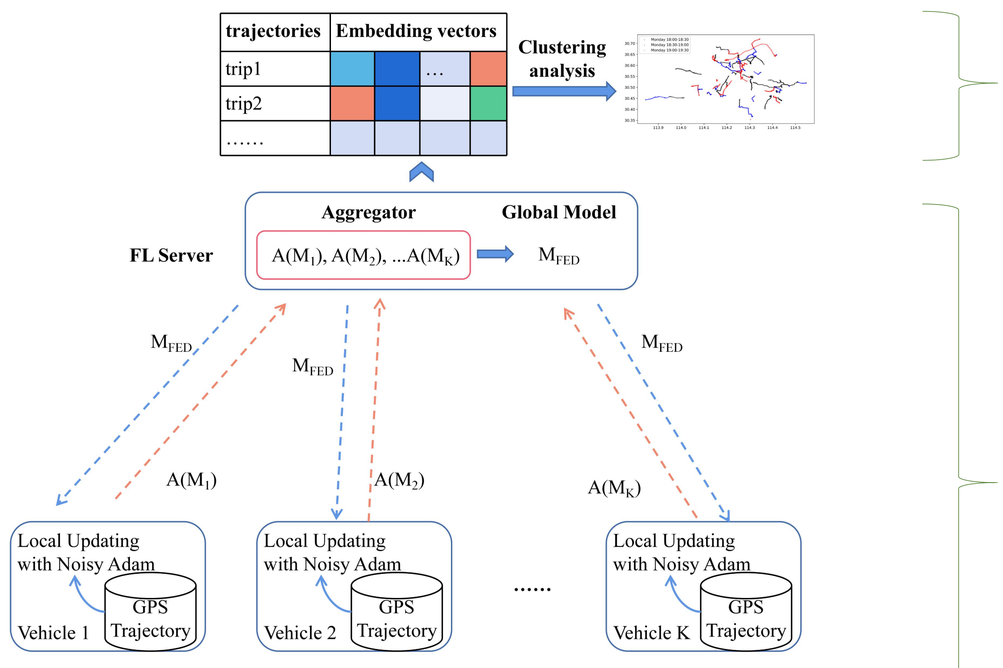

图1 差分隐私联邦轨迹嵌入聚类框架

➢ 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$z_i = A(M_{FED}(x_i))$上传到FL Server。待时段轨迹$x_i, x_i, \ldots, x_n$映射到其相应的定长嵌入$z_i, z_i, \ldots, z_n$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。

➢ 差分隐私联邦轨迹嵌入学习算法

Alg.1 Differential private federated learning of trajectory embedding

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：$M(\theta_{T+E})$	
1	For t=0,…,T-1 do
2	Take a subsampled client from K clients
3	Foreach V_k in this subsample in parallel do
4	Set $M_k(\theta_t)=M(\theta_t)$
5	For j=0,…,E-1 do // Local update E epochs
6	Foreach $x_i \in D_k$ do
7	$v_{t+j}^{(i)} \leftarrow \nabla_\theta \ell(\theta_t, x_i)$
8	$\bar{v}_{t+j}^{(i)} \leftarrow v_{t+j}^{(i)} / \max\{1, \left\|v_{t+j}^{(i)}\right\|_2 / R\}$//梯度裁剪
9	$\tilde{v}_{t+j} \leftarrow \frac{1}{\lvert D_k\rvert}(\sum_{x_i\in D_k} \bar{v}_{t+j}^{(i)} + \gamma R \cdot \mathrm{N}(0,I))$//应用 (ε, δ)-GDP mechanism
10	$m_{t+j} \leftarrow \beta_1 m_{t+j-1} + (1-\beta_1)\tilde{v}_{t+j}$
11	$\mu_{t+j} \leftarrow \beta_2 \mu_{t+j-1} + (1-\beta_2)(\tilde{v}_t \odot \tilde{v}_t)$
12	$w_{t+j} \leftarrow m_{t+j}/(\sqrt{\mu_{t+j}} + \xi)$
13	$\theta_{t+j+1} \leftarrow \theta_{t+j} - \eta w_{t+j}$
14	Return private θ_{t+E}^k to server side
15	Update global model $M(\theta_{t+E})$ with federated average of $\{\theta_{t+E}^k\}_{k=1,\ldots,K}$

实验

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两部分。

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秒获取当前的位置坐标

仿真工具：FedML + Opacus

图2 不同客户端的效果对比

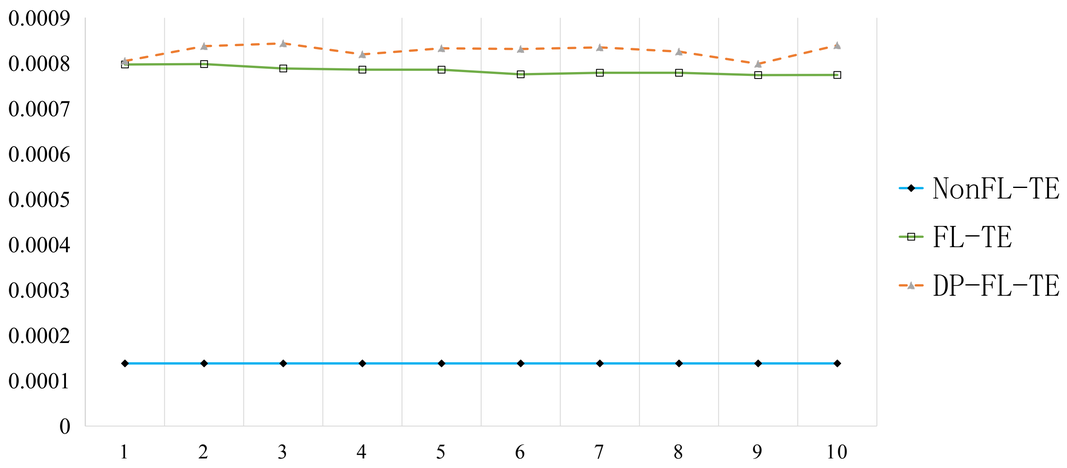

表1 三种模型的聚类效果对比

	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

图2 训练过程对比分析

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，我们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。最后我们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，证明了它可以得到有用的频繁轨迹簇，也就是频繁路线。

投稿人 姓名：孔秀平，陆林
单位：扬州工业职业技术学院