

# 基于注意力机制的用户轨迹识别模型

李昊 曹书瑜 陈亚青 张敏

中国科学院软件研究所可信计算与信息保障实验室 北京 100190

**摘要** 近年来,基于位置服务的应用逐渐开始普及,它在为人们生活提供便利的同时,也对个人隐私造成了巨大威胁。现有研究表明,在具备大量历史轨迹数据的情况下,攻击者能够从匿名化的轨迹数据集中识别出用户身份与轨迹的链接关系。然而,这些相关研究都面临着数据稀疏和数据质量差这两方面的问题。数据稀疏指用户的轨迹往往只分布在局部区域,同时缺乏与自然语言处理领域一样规模庞大的语料库;数据质量差指轨迹中的位置点往往存在采样率低和噪音大的问题。针对上述问题,文中提出了一种基于注意力机制的用户轨迹识别模型,包括位置嵌入模块、基于注意力的位置转移特征编码模块和轨迹用户识别模块。位置嵌入模块用于将原始轨迹中位置点的空间关系嵌入到位置向量中;基于注意力的位置转移特征编码模块用于提取轨迹中位置点间的转移依赖关系,生成轨迹的表征向量;轨迹用户识别模块用于对轨迹表征向量的用户身份进行预测。最后,在 Gowalla 和 Geolife 数据集上进行了实验验证,实验结果表明,所提方案有效缓解了轨迹数据稀疏和数据质量差带来的问题,能够提高轨迹的用户身份识别率。

**关键词:** 轨迹隐私;轨迹用户识别;深度学习;循环神经网络;注意力机制

中图法分类号 TP309

## User Trajectory Identification Model via Attention Mechanism

LI Hao, CAO Shu-yu, CHEN Ya-qing and ZHANG Min

Department of TCA, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

**Abstract** Recently the application of location-based services has gradually become popular. It provides convenience in people's daily life, and also brings a great threat to personal privacy. The existing research shows that, with a large amount of historical trajectory data, attackers can identify the user who generates the trajectory from the anonymous trajectory dataset. In these related studies, both data sparsity and poor data quality are faced. Data sparsity refers to the fact that trajectories are often distributed only in a few local areas, and there is no large corpus contrast to the natural language processing field. The poor data quality refers to the low sampling rate and existing noise of the location points in a trajectory. To address these two problems, this paper proposes a user trajectory identification model based on attention mechanism, including the location embedding module, the attention-based transitional feature encoder and trajectory-user identification module. The location embedding module is used to embed the spatial relation of the trajectory points into the location vector; the attention-based transitional feature encoder is used to extract the sequential dependencies from a single trajectory; and the trajectory-user identification module is used to predict the user identity of the trajectory based on the outputs of the transitional feature encoder. Finally, the experimental verification is carried out on Gowalla and Geolife datasets. The experimental results show that the proposed model in this paper can effectively alleviate the problem of data sparsity and poor data quality, and can achieve better accuracy than existing methods.

**Keywords** Trajectory privacy, Trajectory-user identification, Deep learning, Recurrent neural network, Attention mechanism

## 1 引言

近年来,随着空间定位技术与移动网络的快速发展,基于位置的服务得到了极大的推广应用。截至2018年,我国位置服务市场规模已达859.6亿元<sup>1)</sup>。目前,人们在社交媒体、

购物网站、地图应用等日常软件上的行为都被附加了位置信息,这些位置信息有效地提高了用户在使用软件时的便利性。例如,社交媒体会对用户位置附近的人进行推荐;外卖网站会对用户位置附近的美食餐厅进行推荐;打车软件会显示附近的出租车;地图应用会根据轨迹进行路况显示等。然而,这些

<sup>1)</sup> <https://www.chyxx.com/industry/202001/827088.html>

到稿日期:2021-03-24 返修日期:2021-06-07

基金项目:国家重点研发计划(2018YFC0809300);中国科学院青年创新促进会(2019113)

This work was supported by the National Key R & D Program of China(2018YFC0809300) and Youth Innovation Promotion Association CAS(2019113).

通信作者:李昊(lihao@iscas.ac.cn)

位置信息被采集利用的同时,也会对用户的隐私造成威胁<sup>[1]</sup>。据一份调研报告显示,有超过84%的用户对位置隐私服务可能造成的隐私泄露表示担心<sup>1)</sup>。

目前的一些研究工作表明,基于历史轨迹的用户身份识别是有效的。例如,Zang等直接通过提取用户手机频繁访问的top  $k$ 个基站位置,就能够识别出35%的用户<sup>[2]</sup>。Freudiger等提取home/work的位置对作为用户准标识符也能够在一定程度上实现用户身份识别<sup>[3]</sup>。近期,基于深度学习的轨迹用户识别方法得到了广泛关注<sup>[4-6]</sup>。该类方案实现了端到端的轨迹建模,极大地简化了轨迹特征的提取工作,并充分利用了LSTM(Long Short Term Memory)、GRU(Gated Recurrent Unit)等神经网络对长依赖关系的捕获能力,提高了轨迹用户识别的准确率。

然而,这些方案都面临着数据稀疏和数据质量差带来的问题。首先,上述方案将轨迹数据类比为自然语言处理领域的语句,将轨迹点类比为单词,并利用神经网络对轨迹点之间的序列关系进行学习。这种方式忽略了轨迹数据严重的稀疏性。相比自然语言的数据集,轨迹数据集的规模更小,且轨迹点的分布更加集中在局部区域。其次,由于轨迹点的采集受设备性能的影响,不可避免地存在低采样率和误差严重的位置噪音。因此,在对轨迹数据集进行分析时,必须考虑模型在低数据质量条件下的鲁棒性。

针对上述问题,本文提出了一种基于注意力机制的用户轨迹识别模型UTIMAM(User Trajectory Identification Model via Attention Mechanism)。该模型通过注意力机制来提取最能标识用户移动模式独特性的轨迹特征,进而缓解数据稀疏和低数据质量带来的问题,以提高轨迹用户识别的准确率。

本文的主要贡献如下:

(1)本文提出了一种基于注意力机制的轨迹用户识别模型,充分利用了注意力机制来增强网络对用户移动模式独特性的学习。

(2)本文提出了一种针对轨迹中位置点的嵌入表示方法,支持将不同采样率的原始轨迹进行统一压缩表示,并将位置点的空间关系嵌入到位置向量中。

(3)在数据采样频率和数据噪音相差较大的两类真实数据集上进行了实验,结果表明,本文方案对基于深度学习的轨迹用户识别方案的性能提升是有效的,且在高采样率和高噪音数据集上表现最佳。

## 2 相关工作

现有的轨迹用户识别研究工作主要分为3类:基于显式特征匹配的轨迹识别、基于马尔可夫模型的轨迹建模、基于深度学习的轨迹用户链接。

基于显式特征匹配的轨迹识别方法是通过在轨迹中的高频地点进行匹配<sup>[2]</sup>,或提取轨迹的空间距离、序列长度等特征<sup>[7-8]</sup>对轨迹进行相似度计算,并基于来自同一个用户的轨迹之间相似度较高的假设进行用户身份识别。但这些方法没有对序列依赖关系进行建模,因此其准确率较低。

随后,许多研究者开始利用马尔可夫模型对轨迹中的位置转移关系进行建模。Wang等<sup>[9]</sup>提出了一种基于用户移动

行为周期性规律的隐马尔可夫模型,采用一天的不同时间段作为隐含态,将位置点作为观察态,从而学习出用户在不同隐含态下对位置点的转移偏好。Chen等<sup>[10]</sup>进一步改进了该方案的隐含态构建方式,采用时空点的密度聚类结果作为隐含态,以进一步拟合不同用户在轨迹生成偏好上的差异性。这些基于马尔可夫模型的轨迹建模方法能够学习到轨迹中的位置转移关系,但是无法学习到长轨迹的序列依赖关系。同时,马尔可夫模型对用户移动模式进行了很强的假设约束,且其预测效果依赖于特征工程。

近年来,由于深度学习是一种端到端的过程,且神经网络的发展,使其具备了捕获长序列依赖关系的能力,因此基于深度学习的轨迹用户链接研究得到了广泛关注<sup>[11]</sup>。具体地,Luo等<sup>[4]</sup>给出了轨迹用户链接问题的定义,即将轨迹与产生该轨迹的用户身份进行链接,并提出了一种基于循环神经网络的轨迹用户链接方案。接着,Zhou等<sup>[6]</sup>使用两个LSTM模型的分层组合,使其能够对轨迹的层次化复杂依赖关系进行提取,并进一步采用变分自编码器来建模轨迹,有效提高了模型的性能。Wang等<sup>[5]</sup>提出了一个多任务模型,在用RNN进行轨迹用户链接的同时,使轨迹隐向量与产生它的用户隐向量在同一个特征空间中距离接近。多任务模型的训练使得提取轨迹序列特征的共同参数部分能够获得更优解,从而提高整体模型的性能。这些研究工作表明,基于神经网络及其变种的各类模型,能够有效提取轨迹的序列关系,从而更好地实现轨迹用户的链接任务。

然而,这些研究工作都在不同程度上忽略了轨迹数据集与自然语言数据集的差异性,也忽视了不同采样设备和位置服务场景的差异性。在数据稀疏和低数据质量等条件下,它们并不总是有效的,因此,本文提出了一种基于注意力机制的用户轨迹识别方案,旨在增强模型对用户移动模式特异性的提取能力,以解决上述问题。

## 3 问题描述与符号定义

一条轨迹 $T$ 是由一定时间、空间范围内的多个位置点构成的序列。而每个位置点 $p$ 是由一个位置坐标系统的值(如经纬度)和时间戳构成的元组,其形式化定义如下。

**定义1** 轨迹 $T = \{p_1, p_2, \dots, p_m\}$ ,  $p_i$ 表示位置点, $m$ 为轨迹中位置点的个数,也被称为轨迹长度。

**定义2** 位置点 $p_i = (l_i, t_i)$ ,  $l_i$ 表示该位置点的位置坐标值,在本文中指经纬度, $t_i$ 表示该位置点产生的时间(即时间戳形式)。

**定义3**(用户轨迹识别(或称轨迹用户链接)) 可以表示为一个函数 $u_i = f(T_i)$ ,  $T_i \in \Gamma$ ,  $u_i \in U$ ,其中, $T_i$ 表示一条轨迹, $\Gamma = \{T_1, T_2, \dots, T_n\}$ 表示未知用户身份的轨迹集合, $u_i$ 表示函数 $f$ 计算得到的产生 $T_i$ 的用户身份, $U = \{U_1, U_2, \dots, U_k\}$ 为用户身份的集合。

## 4 基于注意力机制的用户轨迹识别模型

基于注意力机制的用户轨迹识别模型如图1所示。它包括3个主要模块:位置嵌入模块、基于注意力的位置转移特征编码模块和轨迹用户识别模块。

<sup>1)</sup> <https://www.wjx.cn/report/4345626.aspx>

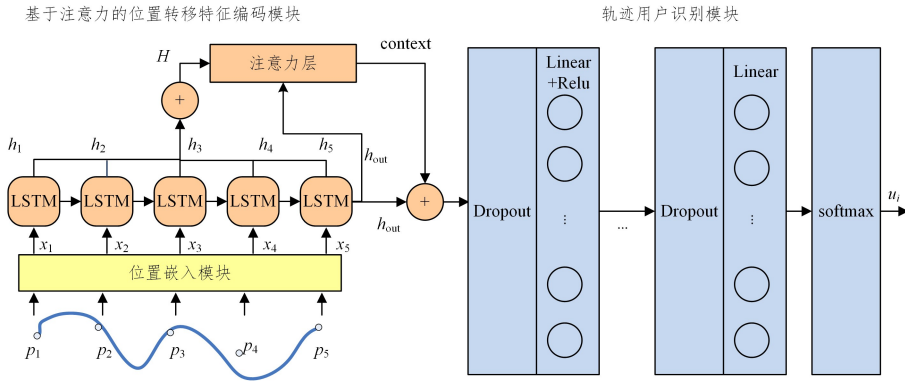


图1 用户轨迹识别模型架构图

Fig. 1 Framework of user trajectory identification model

#### 4.1 位置嵌入模块

同一轨迹数据集中不同用户的采样频率和定位质量存在一定差异,使用户轨迹识别模型陷入对这些差异性的建模,而忽略与问题本质相关的人类移动模式独特性的学习,从而影响模型的泛化性能。为了能够缓解该问题,本文提出了一种新的轨迹泛化压缩和位置嵌入方法。

首先,将轨迹数据集所在区域进行网格化,并分配网格编号。根据网格编号与经纬度范围的对应关系,将位置点  $p_i = (l_i, t_i)$  中的  $l_i$  从经纬度坐标转化为网格编号  $g_i$ 。

然后,将轨迹  $T = \{p_1, p_2, \dots, p_m\}$  中的位置点数量进行压缩。压缩方式为  $\Phi = C(T)$ , 其中,函数  $C$  能够将轨迹  $T$  中具有相同网格编号的相邻位置点放入一个集合  $\Phi_i$ , 形成输出集合  $\Phi = \{\Phi_1, \Phi_2, \dots, \Phi_k\}$ ,  $|\Phi_i| \geq 1$ 。

最后,将  $\Phi$  中的每个  $\Phi_i$  用  $p_i' = (g_i', tc_i')$  代替,其中,  $g_i'$  为  $\Phi_i$  中的相同网格编号,  $tc_i'$  为  $\Phi_i$  中所有位置点之间的最大时间间隔。

此时,  $T = \{p_1, p_2, \dots, p_m\}$  被转化为  $T' = \{p_1', p_2', \dots, p_n'\}$ 。这种轨迹泛化压缩方法,使得噪音位置点的权重  $tc$  变得很小,同时,对高采样率的轨迹中的位置点数量进行了极大地缩减。

由于位置点的分布与自然语言处理领域的单词分布都符合幂律分布<sup>[5]</sup>,因此本文采用了 Word2Vec<sup>[12]</sup> 的 skip-gram 方式进行位置点嵌入,将位置点  $p_i$  转化为位置向量  $v_i$ 。由于 skip-gram 模型能够在嵌入过程中保持一条轨迹中的位置点的有序性,因此嵌入后的两个位置向量越相似,两个位置点的空间可达性就越高。

#### 4.2 基于注意力的位置转移特征编码模块

经过位置嵌入模块后的轨迹可以表示为  $T = \{x_1, x_2, \dots, x_n\}$ , 其中,  $x_i = v_i tc_i$  为位置点  $i$  对应的嵌入向量,即位置点  $p_i$  和权重  $tc_i$  的乘积。

一条轨迹  $T$  将被输入循环神经网络 RNN 进行序列关系的学习, RNN 的第  $t$  步输出记为  $h_t$ , 最后一步的输出记为  $h_{out}$ , 拼接所有步骤的输出得到矩阵  $H$ 。本文采用了 LSTM 或多层 LSTM 来实现 RNN。

在注意力层,将根据轨迹  $T$  经过 RNN 后的输出计算注意力向量 context, 其计算方式如下:

$$Score(h_{out}, H) = h_{out}^T H \quad (1)$$

$$\alpha = softmax(Score(h_{out}, H)) \quad (2)$$

$$context = H\alpha^T \quad (3)$$

注意力层输出的向量 context 得到了用户轨迹识别任务需要关注的序列信息,再将其与  $h_{out}$  进行拼接,作为轨迹用户识别模块的输入。

#### 4.3 轨迹用户识别模块

在构建轨迹用户识别模块时,使用了 MLP(Multi-Layer Perceptron),并在最终输出结果上采用 softmax 进行非线性变换,输出一条轨迹  $T$  属于用户的概率分布,而用户身份则被编码为 one-hot 向量。

由于本文将用户轨迹的识别任务归类为分类任务,因此采用经典的交叉熵损失函数进行模型训练。

$$L = - \sum_{i=1}^{|T|} \sum_{j=1}^{|u|} \mathbf{u}_j \log(\mathbf{T}_i') \quad (4)$$

其中,  $\mathbf{u}_j$  是输入轨迹对应的用户 one-hot 向量,  $\mathbf{T}_i'$  是轨迹用户识别模块对一条轨迹序列的预测输出,而交叉熵损失函数  $L$  描述了两个分布之间的差异。因此,模型训练的目标是最小化交叉熵损失函数,使模型的预测结果更加准确。

此外,本文在 MLP 的每层神经元之间加入了 dropout 层,它将随机舍弃一些神经元的信息传递,避免某些预测结果只在固定特征组合下才生效,从而抑制过拟合的产生。

### 5 实验结果与分析

#### 5.1 实验数据集

本文选择了采样频率和噪音大小截然不同的两个数据集进行实验。

(1) Gowalla 签到数据集<sup>[13]</sup>。本文选取了签到点密集的北纬 40 度左右的美国东海岸 3 个城市区域的签到点作为实验数据集。

(2) Geolife 轨迹数据集<sup>[14]</sup>。该数据集来源于微软亚洲研究院 Geolife 项目,记录了用户居家、工作、购物、远足、旅游等活动的轨迹。

为了将这些签到点和 GPS 定位点构建为轨迹形式,本文采用了 Liu 等<sup>[15]</sup> 提出的方案。先将一名用户的所有签到点或 GPS 定位点连接为一条轨迹,然后以时间间隔阈值  $\Delta t$  将轨迹  $T$  切分为多条子轨迹  $T = \{T_1, T_2, \dots, T_m\}$ , 即如果两个签到点或位置点的时间间隔超过阈值  $\Delta t$ , 则进行轨迹的切分。

本文在实验中去掉了一些长度小于 2 的轨迹,以及轨迹数量过少的用户。其原因在于轨迹长度过短或者轨迹数量太少,都难以让模型学习到用户的移动模式。最后,将每名用户轨迹数量的约 20% 作为测试集,将剩下的数据作为训练集对

轨迹集合进行划分。实验采用的最终数据集的信息如表 1 所列。

表 1 数据集统计信息

Table 1 Statistics of dataset

数据集	用户数	轨迹数	训练集	测试集
Gowalla	156	21 730	19 858	1 872
Geolife	85	11 846	8 895	2 268

## 5.2 对比方法

本文选取了以下 4 种用户轨迹识别方法与本文方案进行对比。

(1)LCSS<sup>[16]</sup> (Longest Common Sub-Sequence):最长公共子串算法是一种采用动态规划的方式对两个序列进行相似性比较的经典方法。

(2)LDA(Linear Discriminant Analysis):LDA 算法是一种经典的线性学习方法,被很多分类任务采用。本文采用了 Mikolov 等<sup>[12]</sup>的词袋模型 BoW 对一条轨迹进行嵌入表示,并采用了奇异值分解 SVD 作为其优化求解方法。

(3)SVM<sup>[17]</sup> (Support Vector Machine):本文同样采用词袋模型 BoW 对轨迹进行嵌入表示,然后采用线性核的 SVM 进行分类学习。

(4)TULER<sup>[4]</sup> (TUL via Embedding and RNN):将循环神经网络 RNN 用于用户轨迹识别任务的前沿工作,具有很好的识别效果。

## 5.3 评价指标

为了更好地与其他方案进行对比,本文采用了文献[4]中的评价指标:准确率(Accuracy)、Top K 准确率(Accuracy@K)和 macro-F1。

其中,Top K 准确率指模型对一条轨迹有多个预测结果,按照概率大小进行排序。若 top  $k$  个预测结果中包含了该轨迹真实的用户身份,则认为预测结果正确。

而 macro-F1 则是衡量多分类任务性能的另一项重要指标,其计算式为:

$$macro-F1 = \frac{2 * macroP * macroR}{macroP + macroR} \quad (5)$$

其中,macroP 和 macroR 分别为多分类任务中所有分类的平均准确率和平均召回率。

## 5.4 参数说明

实验中的主要参数如表 2 所列。

表 2 实验参数设置

Table 2 Parameters used in experiment

参数	数值
Grid size	500 m 或 1 km
Embedding size	250
hidden size	400
learning rate	0.002~0.0001
dropout rate	每 5 个 epoch 减小 0.8
Batch size	64

## 5.5 结果分析

实验结果如表 3 和表 4 所列,黑体表示同一个指标中的最优结果,下划线表示的是次优结果。本文提出的 UTIMAM 模型在 GPS 类型的数据集 Geolife 上的各项评估指标

均优于其他方案。在签到数据类型的数据集 Gowalla 上的各项指标均优于同样采用了深度学习的 TULER,却未能优于 SVM。

表 3 Gowalla 数据集上不同用户轨迹识别方案的结果对比

Table 3 Results comparison among different user trajectory identification methods on Gowalla dataset

模型	Accuracy	Accuracy@3	Accuracy@5	Accuracy@10	macro-F1
LCSS	0.4663	0.6143	0.6667	0.7244	0.3379
LDA	<u>0.5769</u>	0.6864	0.7115	0.7309	<u>0.6284</u>
SVM	<b>0.5879</b>	<u>0.6875</u>	<b>0.7203</b>	0.7504	<b>0.6315</b>
TULER	0.4850	0.6303	0.6746	<u>0.7574</u>	0.4992
UTIMAM	0.5005	<b>0.6898</b>	<u>0.7153</u>	<b>0.7842</b>	0.5211

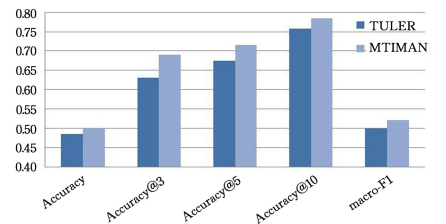
表 4 Geolife 数据集上不同用户轨迹识别方案的结果对比

Table 4 Results comparison among different user trajectory identification on Geolife dataset

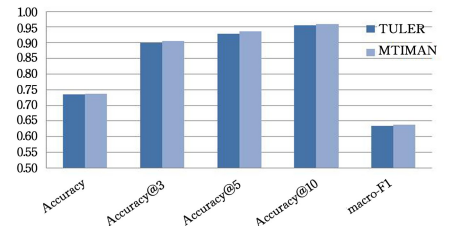
模型	Accuracy	Accuracy@3	Accuracy@5	Accuracy@10	macro-F1
LCSS	0.5599	0.7310	0.7852	0.8417	0.3869
LDA	0.6216	0.7681	0.8064	0.8430	0.5073
SVM	0.7301	0.8955	<b>0.9400</b>	<u>0.9634</u>	0.6145
TULER	<u>0.7345</u>	0.8994	0.9294	0.9559	<u>0.6343</u>
UTIMAM	<b>0.7363</b>	<b>0.9047</b>	<u>0.9369</u>	<b>0.9594</b>	<b>0.6377</b>

具体地,LCSS 是通过轨迹中位置点的匹配进行相似度计算,进而识别用户身份,因此其在两个数据集集中的表现都是最差的。而 LDA 和 SVM 都是对采用词袋模型表示的轨迹特征进行分类的模型,SVM 在 Gowalla 数据集上的表现非常优异,甚至高于深度学习模型 UTIMAM 与 TULER。这表明签到数据集集中的位置点本身的序列关系是不明显的,其轨迹完全可以由无序列性的词袋模型进行表示。而 SVM 与 LDA 对于这种无序的特征分类是非常擅长的,因此效果很好。在 Geolife 这种具有高采样率,但是存在一定噪音的 GPS 轨迹数据集上,深度模型的优势得到了体现,UTIMAM 与 TULER 的性能比其他非深度模型都更好。

更进一步地,将本文方案 UTIMAM 与现有方案 TULER 进行了详细对比,结果如图 2 所示。



(a) Gowalla 数据集



(b) Geolife 数据集

图 2 UTIMAM 与 TULER 在 Gowalla 和 Geolife 上的结果对比

Fig. 2 Results comparison between UTIMAM and TULER on Gowalla and Geolife

可以看出,在 Gowalla 数据集和 Geolife 数据集上,UTIMAM 的表现均优于 TULER,并且在 Gowalla 上的优势更

显著。这是由于 UTIMAM 采用了注意力机制,对循环神经网络的每一步输出都计算了注意力分值,使得每一步输出的隐向量对最终分类任务的贡献有所不同,增强了模型对用户移动模式特异性的提取能力。因此,UTIMAM 具有很好的鲁棒性,能够同时适用于签到数据集和 GPS 轨迹数据集的用户轨迹识别任务。

最后,通过对比不同采样频率和噪音程度的数据集对本文方案 UTIMAM 的影响,进而分析 UTIMAM 方案的性能影响因素,结果如图 3 所示。

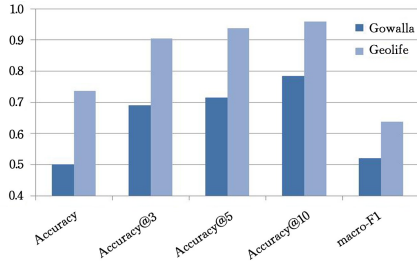


图 3 UTIMAM 在不同采样率与噪音程度的数据集上的结果对比

Fig. 3 Results comparison for UTIMAM on datasets with different sampling frequency and noise level

Gowalla 数据集由用户主动签到产生,其定位可以被用户修正,因此数据噪音较小。然而,用户产生签到点的频率远低于 Geolife 这种 GPS 数据集的位置点采集频率,因此其签到点之间的序列性更差。从图 3 可以看出,UTIMAM 在 Geolife 上的表现更优,这表明 UTIMAM 对轨迹特征的提取主要受位置点序列性的影响,而较少受位置点定位不准带来的噪音影响。由于 UTIMAM 采用了带注意力机制的循环神经网络,因此能够更好地提取轨迹中的序列特征,进而提高用户轨迹识别的准确率。

**结束语** 随着基于位置服务技术的快速发展,轨迹隐私问题得到了广泛关注。近年来,许多针对用户轨迹识别(或称轨迹用户链接)问题的研究越来越多。然而,轨迹数据集存在的数据稀疏和数据质量差这两方面的问题导致这些用户轨迹识别方法在不同数据集上并不总是有效的。针对该问题,本文提出了一种基于注意力机制的用户轨迹识别模型 UTIMAM,旨在提取用户移动模式的特异性特征,以提高其在不同数据集上的性能表现。在两类差异明显的轨迹数据集上进行实验,根据实验结果可以看出,UTIMAM 的性能比之前同样基于循环神经网络的 TULER 均更优,且在两类数据集上有着良好的综合表现。下一步工作将考虑在多个异构数据集上进行用户轨迹识别任务,以使研究成果具有更广泛的适用性。

## 参考文献

- [1] FENG D G. Big data security and privacy preservation[M]. Tsinghua University Press, 2018.
- [2] ZANG H, BOLOT J. Anonymization of location data does not work: A large-scale measurement study[C]// Proceedings of the 17th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2011: 145-156.
- [3] FREUDIGER J, SHOKRI R, HUBAUX J P. Evaluating the privacy risk of location-based services[C]// Proceedings of the 2011 International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2011: 31-46.
- [4] LUO X, GAO Q, ZHOU F, et al. Identifying Human Mobility via Trajectory Embeddings[C]// IJCAI. 2017: 1689-1695.
- [5] WANG G, LIAO D, LI J. Complete User Mobility via User and Trajectory Embeddings[J]. IEEE Access, 2018, 6: 72125-72136.
- [6] ZHOU F, GAO Q, TRAJCEVSKI G, et al. Trajectory-User Linking via Variational AutoEncoder[C]// IJCAI. 2018: 3212-3218.
- [7] YI B K, JAGADISH H V, FALOUTSOS C. Efficient Retrieval of Similar Time Sequences under Time Warping[C]// Fourteenth International Conference on Data Engineering. IEEE Computer Society, 1998.
- [8] CHEN L, NG R. On the Marriage of Edit Distance and Lp Norms[C]// Very Large Data Bases. 2004.
- [9] WANG R, ZHANG M, FENG D, et al. A De-anonymization Attack on Geo-Located Data Considering Spatio-temporal Influences[C]// Proceedings of the 2015 International Conference on Information and Communications Security. Cham: Springer, 2015: 478-484.
- [10] CHEN Z, FU Y, ZHANG M, et al. The De-anonymization Method Based on User Spatio-Temporal Mobility Trace[C]// Proceedings of the 2017 International Conference on Information and Communications Security. Cham: Springer, 2017: 459-471.
- [11] QUAN B, YANG B C, HU K Q, et al. Prediction Model of Ship Trajectory Based on LSTM [J]. Computer Science, 2018, 45(z2): 126-131.
- [12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[C]// International Conference on Learning Representations: Workshops Track. Computer Science, 2013.
- [13] GowallaData[EB/OL]. [2020-01-29]. <http://snap.stanford.edu/data/loc-gowalla.html>.
- [14] GeoLifeData[EB/OL]. [2020-01-29]. <http://research.microsoft.com/enus/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.asp>.
- [15] LIU Q, WU S, WANG L, et al. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts[C]// Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016.
- [16] DAS G, GUNOPULOS D, MANNILA H. Finding Similar Time Series[C]// Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery. 1997: 88-100.
- [17] CORTES C, VAPNIK V. Support-Vector Networks [J]. Machine Learning, 1995, 20(3): 273-297.



**LI Hao**, born in 1983, Ph. D, associate professor, master's supervisor, is a member of China Computer Federation and Youth Innovation Promotion Association. His main research interests include data privacy and access control.