

DOI:10.11918/202311028

基于手机信令 HMM 轨迹映射的交通方式识别算法

王江锋, 杨昊, 梁艳平, 张楚瑶

(综合交通运输大数据应用技术交通运输行业重点实验室(北京交通大学), 北京 100044)

摘要:为解决手机信令数据稀疏性限制和重构轨迹特征提取与融合能力不足问题,提出一种出行轨迹重构与多源特征融合的交通方式精准识别算法。构建刻画基站信号传播路径损耗与信号强度的无线信号损耗模型,利用隐马尔可夫模型(hidden markov model, HMM)将手机信令轨迹由基站序列重构为路段节点序列,提出基于无线信号传播隐马尔可夫模型(wireless signal propagation hidden markov model, WP-HMM)的出行轨迹重构方法,用以描述信号强度与距离作用关系。基于出行重构后的轨迹,结合路段类型特征,提出了时空标准化相似性度量算法,以融合导航轨迹特征,并构建了基于随机森林(random forest, RF)的交通方式识别算法。实证分析表明:通过出行轨迹的重构,模型的平均识别精度提高了8%以上,且对新样本具有优异的泛化能力;相较于现有方法,时空标准化相似性度量算法能更准确捕捉轨迹间的移动模式;在不同环境下的轨迹识别中,模型在郊区区域的表现显著高于城区。所提算法在大规模手机信令数据的出行方式识别领域具有重要的应用价值。

关键词: 交通工程; 交通方式识别; 手机信令数据; 路径损耗; 轨迹相似性

中图分类号: U491.1

文献标志码: A

文章编号: 0367-6234(2024)11-0045-10

Transportation mode recognition algorithm based on mobile phone signaling HMM trajectory mapping

WANG Jiangfeng, YANG Hao, LIANG Yanping, ZHANG Chuyao

(Key Laboratory of Transport Industry of Big Data Application Technologies for comprehensive Transport
(Beijing Jiaotong University), Beijing 100044, China)

Abstract: To address the limitations of sparse mobile phone signaling data and the insufficient ability to extract and fuse features in trajectory reconstruction, this paper proposes a precise transportation mode recognition algorithm based on trajectory reconstruction and multi-source feature fusion. A wireless signal loss model is developed to characterize the signal path loss and signal strength of base stations. Using the hidden markov model (HMM), the mobile phone signaling trajectory is reconstructed from a base station sequence to a road segment node sequence. A trajectory reconstruction method based on the wireless signal propagation hidden markov model (WP-HMM) is proposed to describe the relationship between signal strength and distance. Based on the reconstructed travel trajectory and combined with the characteristics of road segment types, a spatiotemporal standardized similarity measurement fusion navigation trajectory feature is proposed, and a transportation mode recognition algorithm based on random forest (RF) is constructed. Empirical analysis shows that, through the reconstruction of travel trajectories, the model's average recognition accuracy improved by over 8%, and it demonstrated excellent generalization ability for new samples. Compared to existing methods, the spatiotemporal normalized similarity measure more accurately captures the movement patterns between trajectories. In trajectory recognition across different environments, the model performs significantly better in suburban areas than in urban areas. The proposed algorithm demonstrates significant application value in the field of transportation mode recognition using large-scale mobile data.

Keywords: traffic engineering; transportation mode recognition; mobile phone signaling data; path loss; trajectory similarity

居民出行调查是开展城市交通规划与管理的基础性工作,传统居民出行调查方法具有成本高、采样率低、周期长等特点,已无法满足大范围、实时性获

取出行起讫点(origin and destination, OD)的需求。随着移动通信技术的发展和智能手机的普及,手机信令数据成为居民出行调查的有效手段之一。手机

收稿日期: 2023-11-08; 录用日期: 2023-11-20; 网络首发日期: 2024-10-15

网络首发地址: <https://link.cnki.net/urlid/23.1235.T.20241015.0849.002>

基金项目: 国家重点研发计划(2022YFB4300404)

作者简介: 王江锋(1976—),男,教授,博士生导师

通信作者: 王江锋, wangjiangfeng@bjtu.edu.cn

信令通过基站记录出行者的出行轨迹信息,可获取大范围出行者的实时出行信息,进而挖掘其出行范围、驻留区域等出行特征,可为交通规划和管控提供良好的出行数据,并已经应用于城市通勤结构^[1]和人口密度估算^[2]等方面,且取得了良好的效果。

早期阶段,关于手机信令数据的研究大多聚焦于宏观维度,多以城市内人口流动、职住空间特征等统计为主。随着技术的不断发展和数据采集手段的不断完善,研究者们逐渐将研究重点从宏观维度转向精细化研究。手机信令数据具有时空不确定性和稀疏性,因此出行轨迹重构对于手机信令数据的精细化研究至关重要。Li 等^[3]考虑稀疏轨迹数据的缺失模式,结合聚类、决策树、随机森林(random forest, RF)等机器学习算法进行稀疏轨迹数据的轨迹重构,为用户行为特征和集体行为模式的分析提供了支持。Cheng 等^[4]提出了两步法以重构时空缺失轨迹,包括基于动态滑动窗口进行细粒度插值和神经网络模型对时空插值结果进行积分两个步骤。针对手机信令数据的轨迹信息稀疏性,Chen 等^[5]从人类流动性特征方面进行缺失位置推断,基于张量分解,结合上下文增强技术进行轨迹重建,填补手机信令轨迹的稀疏轨迹段。Dyrmishi 等^[6]提出了一种基于粒子滤波算法的移动定位和轨迹重建方法,并与卡尔曼滤波方法进行对比评估。李明晓^[7]依据人类移动行为的相似性特征,提出了一种基于多条件数据划分的轨迹重构算法。从轨迹语义角度对移动对象进行分类,尝试解决现有研究中因轨迹稀疏导致的训练数据量不足的问题。郭煜东等^[8]利用线性插值结合 Savitzky-Golay 滤波进行出行路径的拟合与匹配,并提出一系列优化步骤对出行路径进行优化,该方法分别单独考虑轨迹点的最近距离匹配路段再加以优化,并未将轨迹作为连续的状态序列处理,忽略了路径的整体时空连续性。目前,有监督的重构算法需要大规模有标注的轨迹数据,且对数据质量要求较高,实际应用难以提供足够多的完整轨迹,且模型的泛化能力有待验证。此外,多数重构方法依赖于对原始轨迹序列的平滑与插值处理,输出结果仍是基站序列,并未实现基站序列到路网级别的深度重构,且整体轨迹上下文信息考虑不足,重构准确性有待提高。

随着手机信令轨迹交通方式识别领域研究的深入,研究者关注如何充分挖掘信令数据中的潜在信息,探索将其他数据源与信令数据相结合,以提高识别准确率和鲁棒性。Danafar 等^[9]提出了基于信令网络事件的贝叶斯方法。首先分析了信令事件点之间的最短路径,并进行了轨迹重构,通过公共交通网

络、公共汽车、电车和火车停靠站等信息进行相似度匹配,进一步区分细化交通模式。赖见辉等^[10]根据手机信令事件数据,结合轨道交通路网信息,通过基站与地铁站点的匹配,对地铁乘客出行路径进行推断。杜亚朋等^[11]通过基于密度的轨迹聚类算法融合手机信令轨迹和导航轨迹计算轨迹匹配比值,并进行时间关联识别用户的交通方式。结果表明,结合导航轨迹识别准确率可提升 15%。Liu 等^[12]针对粗粒度的手机信令数据,在动态时间规整(dynamic time warping, DTW)算法基础上提出时空动态时间规整(spatial-temporal dynamic time warping, ST-DTW)算法,整合手机信令轨迹与导航轨迹之间的时空相似性,并对结果进行标准化,以更准确量化轨迹之间的相似性。钟舒琦等^[13]利用导航轨迹聚类得到路径匹配度和时间匹配度,综合匹配度、OD 距离等特征识别用户的出行方式。研究表明,融合兴趣点与导航数据后,用户出行方式的识别准确率显著提高。Chen 等^[14]利用几何覆盖对公共交通导航轨迹和手机信令轨迹进行空间聚类,进一步结合旅行时间识别公共交通出行和私人出行。Lu 等^[15]结合居民出行调查数据、GIS 数据和导航数据,提取出行轨迹时空特征、导航路径特征和出行模式选择特征的识别交通方式,但其导航特征仅采用出行距离和出行时间这类全局特征,未能充分发掘轨迹本身所蕴含的模式信息,导致轨迹的局部细节提取不足。

在手机信令数据交通方式识别任务中融合导航轨迹等其他数据信息可提高模型的识别性能,而轨迹时空相似性度量是轨迹数据融合与特征提取的关键步骤,交通方式识别领域现有文献中的轨迹融合方法主要包括空间覆盖算法^[9-10,13-14]、密度时空聚类算法^[11]和 DTW 类算法^[12]。空间覆盖算法在边界位置误差较大,处理高基站密度区域以及轨迹时空耦合等方面存在一定的局限,且算法较为简单,特征提取能力较差;DTW 类算法和密度时空聚类算法对于参数较为敏感,计算复杂度较高,而且经典的 DTW 算法不适用于不同时空尺度、不同采样频率的非均匀采样轨迹。目前交通方式识别领域内的轨迹相似性算法不适应大规模手机信令数据的处理,对于不同时空尺度、不同采样频率的轨迹特征提取能力有待提高。此外,现有研究大多将基站位置估计为用户实际位置,在轨迹点的稀疏段或高建筑密度地区易产生较大误差。

综上,针对基站位置估计误差与手机信令数据的稀疏性问题,本研究提出考虑无线传播信号的隐马尔可夫模型(wireless signal propagation hidden

markov model, WP-HMM), 将手机信令轨迹映射到路段节点描述的路网, 提高手机定位的准确性和连续性; 针对不同时空尺度、不同采样频率的大规模轨迹特征提取问题, 提出时空标准化轨迹相似性算法 (normalized spatial-temporal trajectory similarity algorithm, NSTSA) 以提取导航轨迹特征, 进一步综合考虑轨迹路网特征、导航轨迹特征和轨迹本身特征, 建立交通方式识别模型。

1 算法构建

基于 WP-HMM 轨迹映射的交通方式识别算法

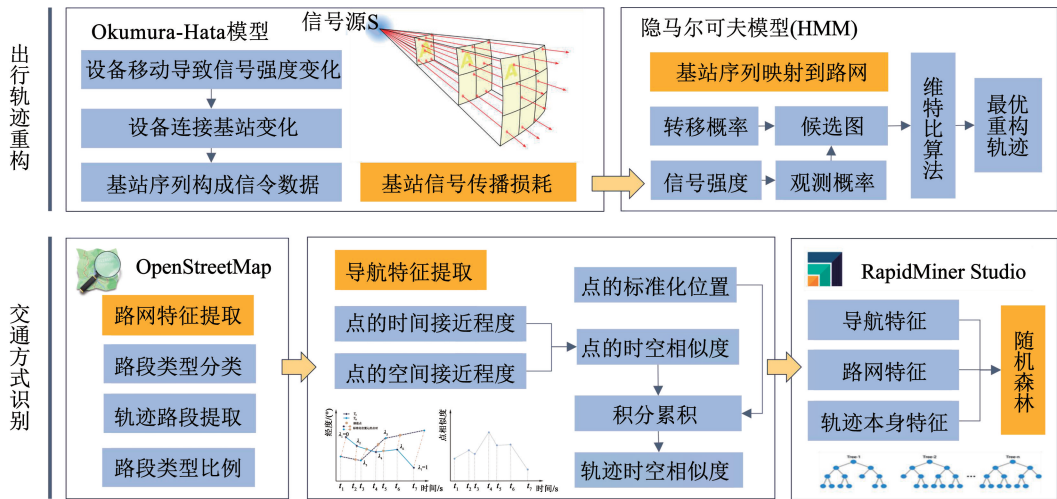


图 1 基于 WP-HMM 轨迹映射的交通方式识别算法框架

Fig. 1 Transportation mode recognition algorithm framework based on WP-HMM trajectory mapping

1.1 基于 WP-HMM 的出行轨迹重构

出行轨迹重构基于 WP-HMM 框架, 包括候选路段的提取、状态转移建模以及路网轨迹解码 3 个阶段^[16-17]。由于信号传播环境的影响、基站位置的变化以及信号强度数据的质量问题等因素, HMM 的定位精度可能受到影响。手机信令的信号传播过程遵循无线信号传播原理, 在观测概率建模时考虑无线信号传播损耗与路段的相对信号强度, 信号传播的路径损耗基于 Okumura-Hata 模型测算^[18]。无线信号传播损耗计算式为

$$L_p = 69.55 + 26.16 \lg f - 13.82 \lg h_b - \alpha(h_m) + (44.9 - 6.55 \lg h_b) \cdot \lg r + C_{\text{cell}} \quad (1)$$

$$\alpha(h_m) = \begin{cases} (1.11 \lg f - 0.7) h_m - (1.56 \lg f - 0.8), & \text{中小型城市} \\ 8.29 (\lg 1.54 h_m)^2 - 1.1, & f \leq 300 \text{ MHz, 大城市} \\ 3.2 (\lg 11.75 h_m)^2 - 4.97, & f > 300 \text{ MHz, 大城市} \end{cases} \quad (2)$$

式中: f 为工作频率, MHz; h_b 为基站天线高度; h_m 为移动台天线高度; r 为传输距离; $\alpha(h_m)$ 为有效移动天线修正因子, 是覆盖区大小的函数; C_{cell} 为区域类型校正因子, 郊区取 0, 市区取 3。

框架见图 1。该算法包括基于 WP-HMM 的出行轨迹重构和 NSTSA 增强的多特征融合交通方式识别两部分。出行轨迹重构利用 Okumura-Hata 模型描述道路网实际位置的信号强度, 作为隐马尔可夫模型 (hidden markov model, HMM) 的观测概率, 通过 HMM 框架整合上下文信息, 将基站序列映射到路网重构出行轨迹; 多特征融合交通方式识别部分利用 NSTSA 算法提取导航轨迹特征、融合路网特征和轨迹本身特征, 基于随机森林算法框架构建交通方式识别模型。

空间中点的信号强度或接收功率为

$$P_r = P_t - L_p \quad (3)$$

式中 P_t 为发射功率。

基于道路网某位置的基站信号强度, 定义 WP-HMM 框架下的观测概率服从变量为信号强度均值为零的高斯分布, 基站位置 p_i (原始手机信令轨迹点) 对应候选路段 s_i^j 的观测概率为

$$p(p_i | s_i^j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-0.5 \left(\frac{P_i}{\sigma}\right)^2} \quad (4)$$

式中: σ 为信号强度标准差, P_i 为基站 p_i 在候选路段 s_i^j 处的信号强度。

在 WP-HMM 中, 状态转移概率显著影响路段间的连续性。此概率基于两个关键距离指标计算: 路径规划距离与直线距离。由于距离差异大意味着迂回路径, 因此距离差异小的路段对之间的状态转移概率更大。定义路段对之间的状态转移概率为

$$p(s_i^j \rightarrow s_{i+1}^j | p_{i-1}, p_i) = \frac{\|p_{i-1} - p_i\|_g}{\|s_i^j - s_{i+1}^j\|_r} \quad (5)$$

式中: $p(s_i^j \rightarrow s_{i+1}^j | p_{i-1}, p_i)$ 为 $s_i^j \rightarrow s_{i+1}^j$ 的转移概率, $\|p_{i-1} - p_i\|_g$ 为信令点 p_{i-1} 和 p_i 之间的大圆距离,

$\|s_i^j - s_{i+1}^j\|_1$ 为候选路段对之间的路径规划距离(最短路径距离)。

利用维特比算法求解 WP-HMM 中最有可能状态序列的解码问题。通过动态规划,初始化每个状态的概率,随后遍历整个序列,对每个状态进行迭代更新。完成迭代后,算法回溯构建最优路径,即重构轨迹。

1.2 多特征融合交通方式识别

从重构出行轨迹中提取路网路段类型特征,利用 NSTSA 相似性算法从导航轨迹提取移动模式特征,结合轨迹的基本时空特征作为联合特征集,以

RF 作为分类框架构建交通方式识别模型。

1.2.1 路网特征提取

现有手机信令交通方式识别研究中,尽管多数方法都涉及路网特征的提取,但关于路段类型如主干道、次要道路或住宅区道路等对交通方式的潜在影响则鲜有深入探讨。本文考虑不同交通方式在不同路段类型方面的行驶特性和频率,采用一种路段特征归纳策略量化路段类型对交通方式识别的影响。路网数据从开源地图平台 OpenStreetMap (OSM) 获取,根据 OSM 路段类型的属性标签划分得到路段类型集合,见表 1。

表 1 路段类型集合及对应出行方式

Tab. 1 Set of road segment types and corresponding travel modes

路段集合	OSM 属性标签	对应路段类型	可出行交通方式	不可出行交通方式
R_1	tertiary, residential, secondary, primary, motorway, trunk	城市干道、快速路、城市主要支路等路段	地铁、公交车、小汽车	步行、骑行
R_2	track, path, living street, service	巷道小路、居住区道路、公园车行道路等路段	小汽车、骑行	步行、公交车、地铁
R_3	footway, pedestrian, steps, cycleway	主要包括人行道和自行车道	步行、骑行	公交车、小汽车、地铁

重构轨迹的路段集合 $\{e_i^{\text{road}}(c_{\text{road}}, l_{\text{road}}) | e_i^{\text{road}}, i = 1, 2, \dots, n\}$, 其中 e_i^{road} 为路段, c_{road} 为路段类型标签, l_{road} 为路段长度。定义重构轨迹的路段类型覆盖度为

$$\mu_k = \frac{\sum_{e_i^{\text{road}}, c_{\text{road}}=k} l_{\text{road}}^i}{\sum_{e_i^{\text{road}}} l_{\text{road}}^i}, k = 1, 2, 3 \quad (6)$$

路段类型覆盖度表示为归属于路段类型集合 k 的路段长度之和占总路段长度的比例。根据式(6)可将路段特征归纳为干道路段覆盖度(μ_1), 生活道路路段覆盖度(μ_2)和慢行道路段覆盖度(μ_3)。

1.2.2 基于 NSTSA 的导航轨迹特征提取

本节提出一种时空标准化轨迹相似性算法(NSTSA), 算法能够衡量不同定位精度、采样频率和长度的轨迹。基于 NSTSA 融合手机信令轨迹和导航轨迹, 提取不同交通方式的导航轨迹特征。算法的主要思路是将两条轨迹序列按照轨迹标准化位置进行对齐并度量点对的时空接近程度, 沿着时间轴不断匹配得到整条轨迹的时空相似度。

1) 点的标准化位置

轨迹 $T([p_i(t_i), i = 1, 2, \dots, n])$ 总行程长 $d_{\text{od}} = \sum_{j=0}^{n-1} h(p_j, p_{j+1})$, 其中 $h(p_j, p_{j+1})$ 是两点间的 Haversine 距离。对于轨迹 T 上任意一点, 其在轨迹上的位置是唯一确定的。点 p_i 到轨迹起点的行程距离 $d_i^{\text{start}} = \sum_{j=0}^{i-1} h(p_j, p_{j+1})$, 定义有向轨迹上点的标准化位置 q_i 为

$$q_i = \frac{d_i^{\text{start}}}{d_{\text{od}}} = \frac{\sum_{j=0}^{i-1} h(p_j, p_{j+1})}{\sum_{j=0}^{n-1} h(p_j, p_{j+1})} \quad (7)$$

2) 点的相似度

通过时空标准化, 将轨迹点的位置映射到 0 ~ 1 的范围, 确保算法可以适应不同的定位精度、采样频率和轨迹长度。两个点在时间或空间上越接近, 则它们所代表的轨迹元越相似。为了量化描述时空接近程度, 从时间和空间上分别定义点的相似度函数:

$$m_t(p_1, p_2) = \begin{cases} 1, & |t_1 - t_2| \leq 30 \\ 1 - \frac{|t_1 - t_2|}{t^*}, & 30 < |t_1 - t_2| < t^* \\ 0, & \text{其他} \end{cases} \quad (8)$$

$$m_s(p_1, p_2) = \begin{cases} 1, & h(p_1, p_2) \leq 10 \\ 1 - \frac{h(p_1, p_2)}{d^*}, & 10 < h(p_1, p_2) \leq d^* \\ 0, & \text{其他} \end{cases} \quad (9)$$

$$m_{ts}(p_1, p_2) = m_s(p_1, p_2) \times m_t(p_1, p_2) \quad (10)$$

式中: p_1, p_2 分别为轨迹上的点; t_1, t_2 分别为轨迹点对应的的时间; m_{ts} 为定义点的时空相似度; t^*, d^* 分别为时间、空间距离阈值, t^*, d^* 越大, 两点的的时间、空间相似度随时间、空间距离减小得越快, 当两点之间的距离大于 t^*, d^* 时, 相似度为 0。

3) 轨迹时空相似度

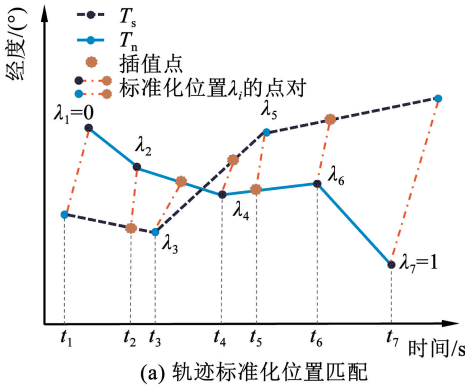
基于点的时空相似度以及点的标准化位置, 对两条轨迹上相同标准化位置的点计算相似度, 并推

广到整条轨迹。手机信令轨迹和导航轨迹的时空相似度为

$$U_{sim}(T_s, T_n) = \frac{\int_{\alpha_1}^{\beta_1} m_{is}(p_1(t_1), p_2(t_2)) v_1(t_1) dt_1}{\int_{\alpha_1}^{\beta_1} v_1(t_1) dt_1} \quad (11)$$

式中

$$\int_{\alpha_1}^{\beta_1} m_{is}(p_1(t_1), p_2(t_2)) v_1(t_1) dt_1 \approx \frac{1}{2} \sum_{i=1}^{n-1} (m_i + m_{i+1}) l_i \quad (12)$$



式中: T_s 和 T_n 分别为手机信令轨迹和导航轨迹; $p_1(t_1)$ ($t_1 \in [\alpha_1, \beta_1]$), $p_2(t_2)$ ($t_2 \in [\alpha_2, \beta_2]$) 分别为手机信令轨迹和导航轨迹上的点, 其中, $[\alpha_1, \beta_1]$ 、 $[\alpha_2, \beta_2]$ 分别为手机信令轨迹和导航轨迹的时间区间; $v_1(t_1)$ 为 T_s 在 t 时刻的速度; m_i 为两条轨迹在相同标准化位置的点相似度; l_i 为 T_s 在区间 $[t_i, t_{i+1}]$ 的长度。

实际计算中, 轨迹由具有不同标准化位置的离散轨迹点组成, 利用线性插值推断轨迹上其他标准化位置的时空坐标, 并利用式 (12) 近似计算积分。算法示意图 2 (纵坐标以经度为例)。

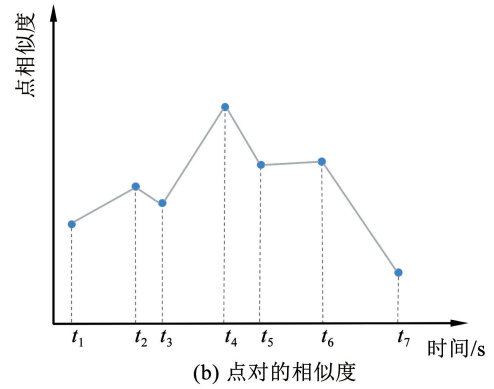


图 2 轨迹相似性计算示意
Fig.2 Trajectory similarity calculation

4) 导航特征提取

从导航地图请求与手机信令轨迹相同 OD 的各交通方式导航轨迹 $T_n^c, c \in M | M = \{ \text{小汽车, 公交车, 地铁, 骑行, 步行} \}$ 。基于 NSTSA 计算手机信令轨迹与各交通方式导航轨迹的时空相似度作为导航轨迹特征, 手机信令轨迹与相应交通方式导航轨迹的相似度为

$$f_n^c = U_{sim}(T_s, T_n^c), c \in M \quad (13)$$

1.3 交通方式识别算法

利用 WP-HMM 算法对手机信令轨迹进行路段 - 节点层面的轨迹重构, 并提取路段类型覆盖度作为路网特征; 基于 NSTSA 算法计算手机信令轨迹与导航轨迹的时空相似度作为导航轨迹特征; 此外, 相比于其他交通方式, 公交车和地铁在运营时段需要在特定站点停靠, 其出行轨迹的速度变化较为频繁, 速度方差较大; 不同交通方式的信令采样频率也有所不同, 计算相邻数据点之间的时间间隔作为信令采样特征。综上, 模型考虑的特征包括路段类型覆盖度, 导航轨迹相似度, 出行速度最大值、均值和方差, 出行距离, 信令时间间隔平均值和最大值。

RF 是一种强大的集成学习算法, 通过构建多个

决策树并结合其预测结果进行分类。其具有适应性强, 处理高维特征数据能力强, 对异常值不敏感, 有效减少过拟合风险的特点。采用 RF 建立手机信令轨迹的交通方式识别算法可表示为

$$\hat{y} = \arg \max_c \sum_{i=1} I(O_i^{tree}(X) = c) \quad (14)$$

式中

$$X = [\mu_1, \mu_2, \mu_3, f_n^c, v_{max}, v_{avg}, v_{var}, d, \Delta t_{avg}, \Delta t_{max}] \quad (15)$$

式中: $O_i^{tree}(X)$ 为第 i 棵树对输入特征向量 X 的预测输出; $I(\cdot)$ 为指示函数, 当括号内的条件成立时取值为 1, 否则为 0; v_{max} 、 v_{avg} 、 v_{var} 分别为出行速度的最大值、均值和方差; d 为出行距离; Δt_{avg} 和 Δt_{max} 分别为信令时间间隔的平均值和最大值。

2 实证分析

2.1 数据准备与预处理

本研究采用 2019 年某地区的手机信令标定轨迹数据, 共计 791 条。从总样本中按各交通方式分层抽取 70% (554 条) 作为训练数据, 其余 30% (237 条) 作为测试数据进行本文提出的方法验证和分析。数据结构见表 2。

表 2 信令数据结构

Tab.2 Signaling data structure

用户编号	轨迹点编号	时间戳	经度/(°)	纬度/(°)
1000 * * * * * * * * * * 3359	0	7:44:50	115.727 94	40.076 27
1000 * * * * * * * * * * 3359	1	7:44:56	115.728 53	40.075 90
1000 * * * * * * * * * * 3359	2	7:51:59	115.788 72	40.050 41
1000 * * * * * * * * * * 3359	3	7:52:42	115.788 72	40.050 41

由于受环境、传感器误差以及通信干扰等因素影响,获得的用户轨迹往往呈现出明显的抖动和误差,不仅会降低定位的准确性,还会影响后续的轨迹算法分析。为此,引入平滑轨迹预处理以改善定位数据的质量。Rauch-Tung-Striebel (RTS) 平滑是卡尔曼滤波器的扩展,通过利用过去和未来的测量值,结合状态估计和测量更新,对原始定位轨迹进行数据融合和平滑处理,可以有效减少轨迹中的噪声抖动,提升轨迹数据的连续性和一致性。

2.2 参数标定

2.2.1 RTS 平滑参数

初始状态矩阵设为轨迹起点坐标。手机设备或用户方程的移动采用二阶运动方程描述,初始状态转移矩阵和初始观测矩阵设为单位矩阵。初始状态协方差矩阵和转移协方差矩阵采用期望最大化算法 (EM 算法) 进行估计,迭代次数设置为 8。

2.2.2 WP-HMM 算法参数

参考工信部通知以及华为、中兴等企业的移动通信基站设备的参数可知,2019 年中国三大运营商 (中国移动、联通、电信) 主要 4G LTE 基站设备 f 在 1 900 ~ 2 300 MHz 之间, h_b 在 30.0 ~ 50.0 m 之间。因此,本研究 f 取 2 100 MHz, h_b 取 40.0 m, h_m 取 1.7 m。

2.2.3 RF 模型参数

RF 模型参数主要包括要生成的随机树的数量

和树的最大深度。设置随机树数量在 [50, 200] 之间进行实验,大于 110 之后,模型性能趋于稳定且性能较好,因此取随机树数量为 110。树的最大深度控制了树生长的深度,如果深度太大,可能导致过拟合 (overfitting) 问题;如果深度太小,模型可能无法捕捉到数据的复杂关系,导致欠拟合 (underfitting) 问题。结合本文数据集大小,树的最大深度设置为 10。

2.2.4 NSTSA 算法参数

t^* 和 d^* 反映了算法在度量轨迹相似性时对时间偏移和空间偏移的敏感程度,阈值越大,算法对于时空偏移的容忍度越高,反之越小。通过对不同 t^* 和 d^* 取值下模型的预测精度分析,当 $d^* = 2\ 000\ m$ 、 $t^* = 1\ 300\ s$ 时,模型的准确率和结果的稳定性均较好,总体表现最佳。因此, d^* 取 2 000 m、 t^* 取 1 300 s 作为 NSTSA 算法的最终参数。

2.3 结果分析

2.3.1 轨迹实例分析

基于文献 [13, 19] 的方法对冗余数据和乒乓数据进行清洗,并利用 RTS 平滑算法,减少轨迹的跳跃和反复连接;最后将原始手机信令轨迹由基站位置序列映射为由路段结点描述的出行路径,提取更加准确可靠的出行路径信息。某次城区出行所产生的手机信令轨迹、对应的 RTS 平滑轨迹和重构轨迹对比见图 3。

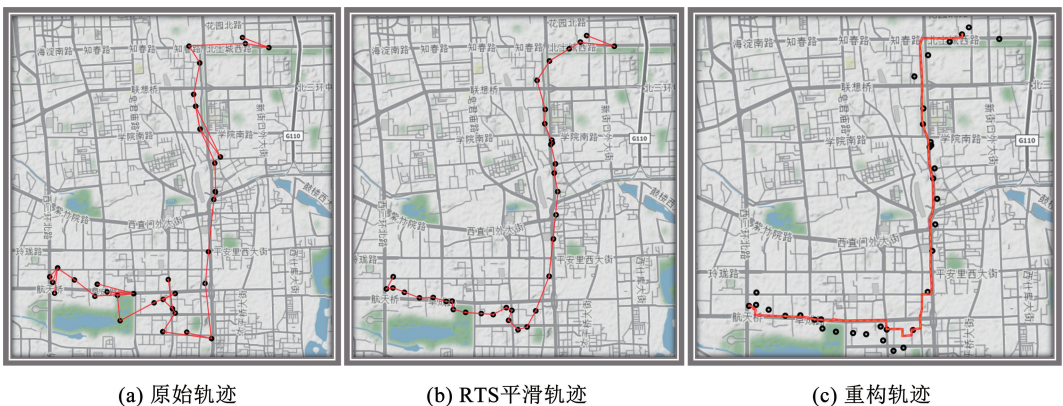
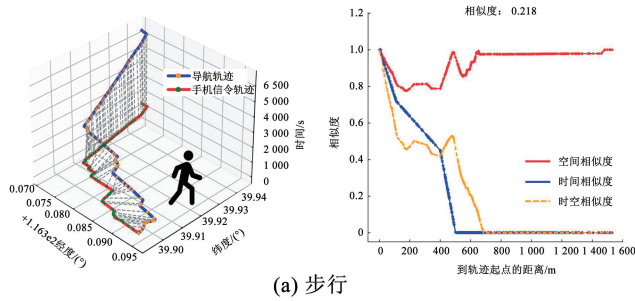


图 3 原始轨迹、RTS 平滑轨迹和重构轨迹对比

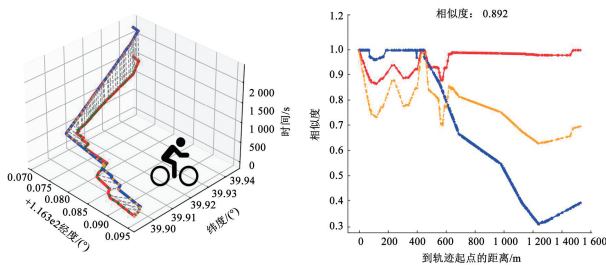
Fig.3 Comparison of original trajectory, RTS smooth trajectory and reconstructed trajectory

导航轨迹与手机信令轨迹的时空关系以及相似度曲线见图 4。图 4(a)~4(e) 中左侧是导航轨迹与手机信令轨迹的时空位置, 虚线的长度表示相同标准化位置的点对的时空距离; 右侧是两条轨迹的时空相似度曲线。图 4(a) 为手机信令轨迹与步行

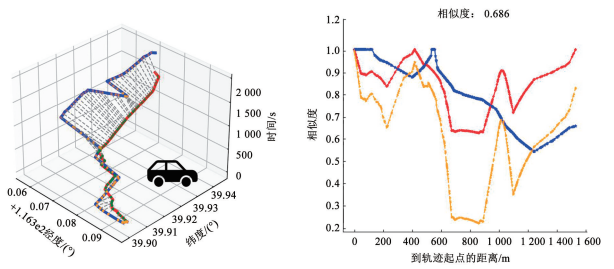
轨迹的时空关系。由图 4(a) 可知, 尽管步行轨迹在空间上与手机信令轨迹相互接近, 但对应的标准化位置点在时间维度上明显滞后, 导致两者在时空维度上的距离不断增大。与其他交通方式轨迹相比, 骑行轨迹与手机信令轨迹在时空维度上最为接近。



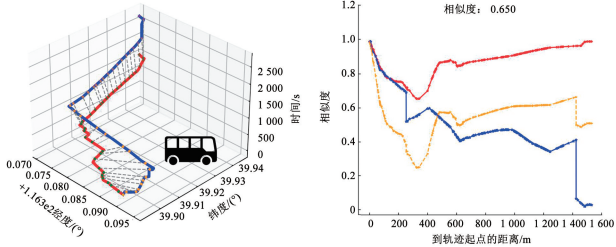
(a) 步行



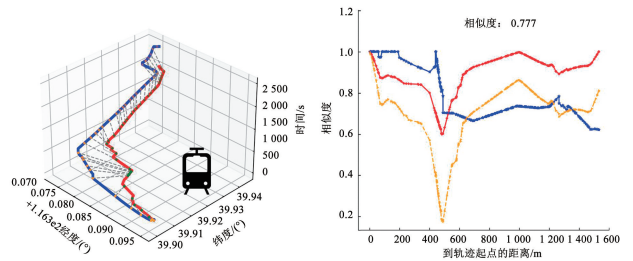
(b) 骑行



(c) 小汽车



(d) 公交车



(e) 地铁

图 4 导航轨迹与手机信令轨迹时空关系及相似度曲线

Fig. 4 Spatial-temporal relationship and similarity curves between navigation trajectory and mobile phone signaling trajectory

通过时空相似度曲线与真实时空位置关系的对比检验可知, 时空相似度曲线能够准确地评估两条轨迹在时空维度上的接近程度, 验证了时空标准化轨迹相似度算法的有效性。

2.3.2 模型分析

本节分析轨迹重构阶段、路网特征以及不同相

似度算法对模型的影响, 并进行敏感性分析。评估指标为准确率。精确率指在被所有预测为正的样本中实际为正样本的概率, 召回率是指在实际为正的样本中被预测为正样本的概率, 准确率为两者的调和分数。模型识别结果见表 3。

表 3 模型识别结果

Tab. 3 Model recognition results

识别的交通方式	真实的交通方式					总计	评价指标		
	步行	骑行	公交车	小汽车	地铁		精确率/%	召回率/%	准确率/%
步行	16	1	0	0	0	17	94.1	84.2	88.9
骑行	1	29	1	2	0	33	87.9	90.6	89.2
公交车	2	2	48	5	2	59	81.4	80.0	80.7
小汽车	0	0	10	86	2	98	87.8	91.5	89.6
地铁	0	0	1	2	27	30	90.0	87.1	88.5
总计	19	32	60	95	31	237	87.3	87.3	87.3

1) 轨迹重构分析

为了系统研究轨迹重构以及路段类型特征对模型精度的影响,重点考虑了轨迹是否经过重构阶段和是否融入路段类型特征。在轨迹未重构条件下,手机信令轨迹通过基站序列描述,不包括路段类型属性。因此设计 3 种场景进行对比实验。

图 5 为轨迹重构对比实验结果。由图 5 可以看出:通过 WP-HMM 轨迹重构,5 种交通方式的平均准确率相对于无轨迹重构情况提高 8% 以上。其中,步行、骑行、公交车、小汽车提升幅度较大;尽管地铁轨迹运行环境大多位于地下,与地面交通方式相比信号传播环境更为封闭且复杂,但准确率仍提升 3%。说明轨迹重构增强了轨迹数据的可解释性和操作性,使得模型能够更为高效地捕捉数据中的关键特征与潜在模式。此外,当引入路段类型特征后,模型对于各种交通方式的平均预测准确率进一步提高,达到了 87.3%,证明路段类型特征对于交通模式识别任务的重要性,能为分类任务提供更多的上下文信息。

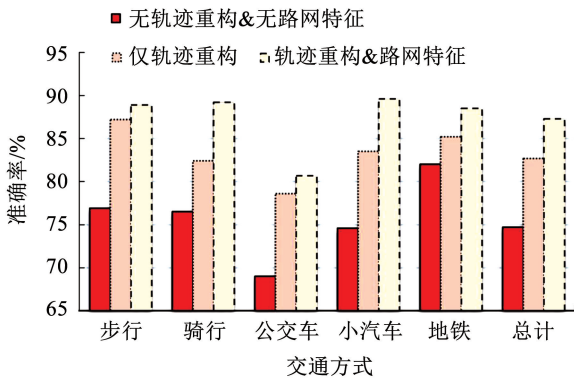


图 5 轨迹重构对比实验结果

Fig. 5 Comparison of results of trajectory reconstruction

2) 相似度算法分析

为了验证本文提出的相似性算法的有效性,将

本文算法与密度时空聚类和 ST-DTW 算法对比分析。为便于比较,3 种算法均采用 Havesine 公式计算距离。此外,为评估 3 种算法在大规模轨迹数据应用的潜力,采用 60 对测试轨迹(轨迹点个数均匀分布)进行相似性计算,并记录算法运行时间。

不同相似度算法实验结果见表 4。由表 4 可知:本文算法的准确率略高于 ST-DTW 算法,密度聚类算法准确率最差。本文算法在步行、骑行和公交车的识别上准确率较高,尤其在步行和骑行的识别上,超过了其他两种算法;步行和骑行的轨迹通常为曲折,反映了行人或骑行者在城市环境中的微观行为和选择,如避开障碍物或选择短路径等,尤其是当轨迹数据的采样频率和定位精度不一致时,现有方法往往难以捕捉这些细微的差异。本文算法通过对轨迹点进行标准化,能够捕捉轨迹路径的局部特征。在处理小汽车和公交车这类高速动态且时间敏感的轨迹时,略逊于 ST-DTW 算法。ST-DTW 算法特别适用于处理具有时间伸缩性和非线性特征的轨迹序列数据,能有效地适应小汽车和公交车这类交通模式的快速变化和复杂路线。本文算法虽在捕捉低速轨迹,如步行和骑行的局部特征方面表现优异,但由于其核心建立在轨迹行程距离的标准化基础上,对于高速移动的轨迹,如小汽车和公交车,其时间敏感性处理能力较弱,仍有优化和提升的空间。

此外,本文算法在时间性能上表现较优,仅用时 0.086 1 s,而 ST-DTW 和密度时空聚类算法分别需要 0.113 4 s 和 0.145 7 s(见表 4)。根据时间复杂度理论分析可知,对于两个长度分别为 l_1 和 l_2 的待比较序列,ST-DTW 和密度时空聚类算法的时间复杂度均为 $O(l_1 \cdot l_2)$,本文算法的时间复杂度为 $O(l_1 + l_2)$ 。综上所述,本文算法在保证准确率的同时,在时间性能上具有明显的优势,更适用于大规模手机信令数据与其他轨迹数据的相似性计算。

表 4 不同相似度算法实验结果

Tab. 4 Experimental results of different similarity algorithms

相似度算法	准确率/%						执行时间/s
	步行	骑行	小汽车	公交车	地铁	总计	
密度聚类	0.821	0.794	0.786	0.794	0.820	0.797	0.145 7
ST-DTW	0.821	0.870	0.854	0.908	0.871	0.869	0.113 4
本文算法	0.889	0.892	0.807	0.896	0.885	0.873	0.086 1

3) 样本量敏感性分析

根据对本文提出算法的敏感性分析,考察在总数据量固定的条件下,随机分层抽取数据集作为测试样本,测试样本占数据总量分别设为 30%、35%、40%、45% 和 50%。为了保证可复现性,随机种子

设为 0。

模型准确率随测试样本占比变化见图 6。由图 6 可以看出,当测试样本占比从 30% 逐步增加到 50% 时,不同交通模式的识别准确率呈现出不同的变化趋势。其中,由于步行和骑行的低速性,所以识

别准确率较为稳定, 显示出对样本量变化的鲁棒性。公交车和小汽车作为更加复杂的交通模式, 轨迹受多种因素影响, 如路网条件、交通规则和行驶行为, 识别准确率逐渐降低。尤其是小汽车, 在不同路段类型上行驶的自由度更高, 导致其轨迹多样性增加。地铁由于轨迹固定和速度较快, 识别准确率虽然也受测试样本占比的影响, 但相较于小汽车和公交车变化幅度相对较小。

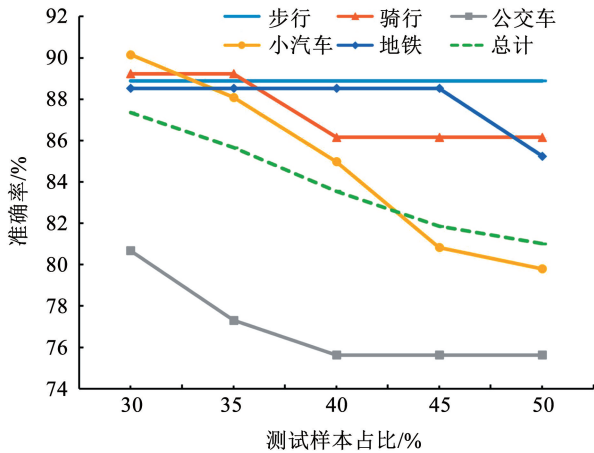


图 6 模型准确率随测试样本占比变化

Fig. 6 Accuracy of the model varies with the proportion of test samples

显然, 随着测试样本占比的增加, 训练样本量随之减少, 模型所提取的特征无法充分泛化以覆盖新的特征模式, 从而导致模型准确率有一定程度的下降。在训练样本与测试样本数量相同的情况下, 模型总体识别准确率仍能达到 81%。可见训练样本数量更多地影响模型对于出行速度、信令时间间隔等轨迹特征的提取表达能力, 对于路段类型特征、导航轨迹特征影响相对较小, 模型对于新样本有一定的泛化能力。

4) 环境变量敏感性分析

在城市交通模式识别中, 基站布局和城市结构等环境变量对基于手机信令的模型性能可能造成影响。由于城区与郊区在基站密度、城市布局及交通模式等方面存在显著差异, 城区通常具有高密度的基站和复杂的道路网络, 而郊区则相对基站稀疏且道路结构简单。因此, 环境变量敏感性从城区与郊区两种环境类型考虑。以北京市为例, 北京市城区包括海淀、朝阳、东城、西城、丰台和石景山, 其他区域为郊区。利用地理信息系统工具确定轨迹点坐标与城区地理边界的空间关系。对于跨越城区和郊区的轨迹, 根据轨迹在各区的停留时间或经过的点数比例判断归属。例如, 如果轨迹在城区的停留时间占总时间比例更大, 则视为城区轨迹, 否则归类为郊区轨迹。

区轨迹。

基于上述划分标准, 得到城区轨迹和郊区轨迹的比例为 482:309。为了深入分析模型对环境因素的敏感性, 根据轨迹出行方式及其归属区域对数据进行分层抽样, 城区和郊区模型准确率对比见图 7。由图 7 可知, 模型在城区的识别准确率低于郊区。原因是城区内密集的建筑群和复杂的道路网结构导致信号衰减和多路径反射, 降低了信令数据的质量。另外, 城区居民的移动行为更为多样和复杂, 包括短距离步行、多种交通工具的换乘等, 增大了从信令数据中识别出准确出行方式的难度。相比之下, 郊区由于基站覆盖较为广泛且分布简单, 交通模式也相对单一, 信令数据在这些区域能够更加准确地反映用户的实际移动轨迹。因此, 模型在郊区的轨迹识别准确率相对较高。

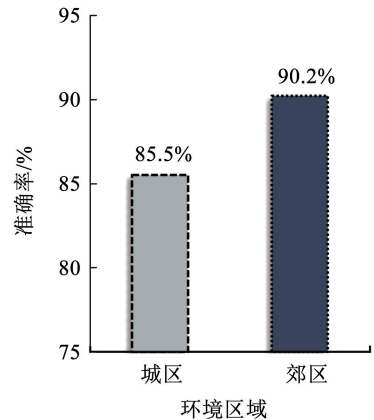


图 7 城区和郊区模型准确率对比

Fig. 7 Comparison of accuracy between urban and suburban models

3 结 论

本研究利用 Okumura-Hata 模型对基站信号强度传播进行建模, 考虑无线信号传播原理构建 WP-HMM 模型对手机信令轨迹进行路网层重构; 融合多源特征数据, 利用 NSTSA 相似性算法从导航轨迹中提取移动模式特征, 并结合轨迹的基本时空特征构建交通方式识别模型, 得到如下结论:

1) 本文模型对于各交通方式的识别准确率可达 87.3%。相较于无轨迹重构, 通过基于 WP-HMM 的轨迹重构, 模型的识别准确率显著提高。

2) 与密度时空聚类、ST-DTW 等相似性算法相比, 本研究提出的相似性算法能准确刻画两种轨迹之间的相似关联, 且时间复杂度较优, 更适合大规模手机信令数据与其他轨迹的特征提取。

3) 通过样本量敏感性分析, 训练样本数量更多地影响模型对于出行速度、信令时间间隔等轨迹特

征的提取表达能力,对于路段类型特征、导航轨迹特征影响相对较小,模型对于新样本有较优的泛化能力。

4)通过环境变量敏感性分析,模型在郊区的轨迹识别准确率高于城区的轨迹识别准确率。

参考文献

- [1] LIU Yaolin, FANG Feigu, JING Ying. How urban land use influences commuting flows in Wuhan, Central China: a mobile phone signaling data perspective [J]. *Sustainable Cities and Society*, 2020, 53: 101914. DOI: 10.1016/j.scs.2019.101914
- [2] YANG Junyan, SHI Yi, YU C, et al. Challenges of using mobile phone signaling data to estimate urban population density: towards smart cities and sustainable urban development [J]. *Indoor and Built Environment*, 2019, 29 (2): 147. DOI: 10.1177/1420326X19893145
- [3] LI Mingxiao, GAO Song, LU Feng, et al. Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data[J]. *Computers, Environment and Urban Systems*, 2019, 77: 101346. DOI: 10.1016/j.compenvurbysys.2019.101346
- [4] CHENG Shifen, LU Feng. A two-step method for missing spatio-temporal data reconstruction [J]. *ISPRS International Journal of Geo-Information*, 2017, 6(7): 187. DOI: 10.3390/ijgi6070187
- [5] CHEN Guangshuo, VIANA A C, FIORE M, et al. Complete trajectory reconstruction from sparse mobile phone data [J]. *EPJ Data Science*, 2019, 8(1): 30. DOI: 10.1140/epjds/s13688-019-0206-8
- [6] DYRMISHI S, HADACHI A. Mobile positioning and trajectory reconstruction based on mobile phone network data: a tentative using particle filter [C]//7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). Heraklion; IEEE, 2021: 1. DOI: 10.1109/MT-ITS49943.2021.9529277
- [7] 李明晓. 手机信令轨迹重构与预测方法研究 [J]. *测绘学报*, 2023, 52(2): 341
LI Mingxiao. Research on the reconstruction and prediction of cell phone signaling derived trajectories [J]. *Acta Geodaetica et Cartographica Sinica*, 2023, 52 (2): 341. DOI: 10.11947/j. AGCS.2023.20210141
- [8] 郭煜东, 杨飞, 周涛, 等. 基于手机信令的城市机动化方式细分双层模型研究 [J]. *交通运输系统工程与信息*, 2023, 23(3): 101
GUO Yudong, YANG Fei, ZHOU Tao, et al. Two-layer model to distinguish urban motorized travel mode based on mobile phone signaling data [J]. *Journal of Transportation Systems Engineering and Information Technology*, 2023, 23 (3): 101. DOI: 10.16097/j.cnki.1009-6744.2023.03.012
- [9] DANAFAR S, PIORKOWSKI M, KRYSCZCZUK K. Bayesian framework for mobility pattern discovery using mobile network events [C]//25th European Signal Processing Conference (EUSIPCO). Kos; IEEE, 2017: 1070. DOI: 10.23919/EUSIPCO.2017.8081372
- [10] 赖见辉, 陈艳艳, 钟园, 等. 基于手机定位信息的地铁乘客出行路径辨识方法 [J]. *计算机应用*, 2013, 33(2): 583
LAI Jianhui, CHEN Yanyan, ZHONG Yuan, et al. Travel route identification method of subway passengers based on mobile phone location data [J]. *Journal of Computer Applications*, 2013, 33(2): 583. DOI:10.3724/SP.J.1087.2013.00583
- [11] 杜亚朋, 雒江涛, 程克非, 等. 基于手机信令和导航数据的出行方式识别方法 [J]. *计算机应用研究*, 2018, 35(8): 2311
DU Yapeng, LUO Jiangtao, CHEN Kefei, et al. Recognition of urban travel method based on cell phone signaling and navigation map data [J]. *Application Research of Computers*, 2018, 35(8): 2311. DOI: 10.3969/j.issn.1001-3695.2018.08.018
- [12] LIU Yong, ZHAO Na, HAN Wei, et al. User behavior patterns and activity recognition based on mobile big data [C]//2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA). Huaihua; IEEE, 2021: 203. DOI: 10.1109/ICEITSA54226.2021.00048
- [13] 钟舒琦, 邓如丰, 邓红平, 等. 基于兴趣点与导航数据的手机信令数据出行方式识别 [J]. *中山大学学报(自然科学版)*, 2020, 59(3): 87
ZHONG Shuqi, DENG Rufeng, DENG Hongping, et al. Recognition of traffic mode of mobile phone data based on the combination of point of interest data and navigation data [J]. *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 2020, 59(3): 87. DOI: 10.13471/j.cnki.acta.snus.2020.03.010
- [14] CHEN Jiatao, XIONG Chen, CAI Ming, et al. A travel mode identification framework based on cellular signaling data [J]. *Mobile Information Systems*, 2022, 2022: 2113213. DOI: 10.1155/2022/2113213
- [15] LU Zhenbo, LONG Zhen, XIA Jingxin, et al. A random forest model for travel mode identification based on mobile phone signaling data [J]. *Sustainability*, 2019, 11(21): 5950. DOI: 10.3390/su11215950
- [16] NEWSON P, KRUMM J. Hidden Markov map matching through noise and sparseness [C]//Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle; Association for Computing Machinery, 2009: 336. DOI: 10.1145/1653771.1653818
- [17] ZENG Zhe, ZHANG Tong, LI Qingquan, et al. Curvedness feature constrained map matching for low-frequency probe vehicle data [J]. *International Journal of Geographical Information Science*, 2016, 30(4): 660. DOI: 10.1080/13658816.2015.1086922
- [18] HATA M. Empirical formula for propagation loss in land mobile radio services [J]. *IEEE Transactions on Vehicular Technology*, 1980, 29(3): 317. DOI: 10.1109/T-VT.1980.23859
- [19] 王彦琛, 杨飞, 李荣玲, 等. 手机信令定位频率对交通方式识别的影响 [J/OL]. *西南交通大学学报*. (2022-10-14) [2023-10-28]. <http://kns.cnki.net/kcms/detail/51.1277.U.20221014.1129.010.html>
WANG Yanchen, YANG Fei, LI Rongling, et al. Impact evaluation of location frequency on travel mode extraction using cellular phone data [J/OL]. *Journal of Southwest Jiaotong University*. (2022-10-14) [2023-10-28]. <http://kns.cnki.net/kcms/detail/51.1277.U.20221014.1129.010.html>