控制与决策
Control and Decision

区别多种出行方式的城市活动轨迹预测

郭戈, 胡峻豪

引用本文:
郭戈,胡峻豪. 区别多种出行方式的城市活动轨迹预测[J]. 控制与决策, 2023, 38(4): 1022−1030.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1558

区别多种出行方式的城市活动轨迹预测

郭 戈[1,2†], 胡峻豪[3]

(1. 东北大学 流程工业综合自动化国家重点实验，沈阳 110004；
2. 东北大学秦皇岛分校 控制工程学院，河北 秦皇岛 066004；
3. 东北大学 信息科学与工程学院，沈阳 110004)

摘 要: 信息社会中,基于用户的历史活动轨迹发掘和预测人类位置轨迹及活动规律至关重要. 已有研究大多采用基于时间和轨迹间相似度分类的马尔可夫模型,忽略了不同出行方式下的移动规律差异. 对此,区别不同出行方式,基于轨迹的速度、加速度和航向变化速度等特征,用XGBoost算法识别轨迹所对应的出行方式,并采用基于优化的轨迹分割算法,将人类出行轨迹按出行方式分解成多个轨迹,采用由不同出行方式轨迹建立的马尔可夫模型实现出行轨迹的精准预测. 实验表明,不同出行方式的轨迹的移动规律存在显著差异,且所提出方法的预测精度和距离偏差明显优于几个基准方法.

关键词: 马尔可夫模型；轨迹分类；轨迹分割；轨迹预测；出行方式

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1558

引用格式: 郭戈,胡峻豪. 区别多种出行方式的城市活动轨迹预测 [J]. 控制与决策, 2023, 38(4): 1022-1030.

Urban activity trajectory prediction with different travel modes

GUO Ge[1,2†], *HU Jun-hao*[3]

(1. State Key Laboratory of Synthetical Automation for Process Industries，Northeastern University，Shenyang 110004，China；2. School of Control Engineering，Northeastern University at Qinhuangdao，Qinhuangdao 066004，China；3. College of Information Science and Engineering，Northeastern University，Shenyang 110004，China)

Abstract: In the information society, it is very important to discover and predict human position trajectories and activity patterns based on users' historical activity trajectories. Most of the existing studies use the Markov model based on the similarity classification between time and trajectory, ignoring the differences of movement laws under different travel modes. This paper distinguishes different travel modes, based on the characteristics of the trajectory's speed, acceleration, and heading change speed, the XGBoost algorithm is used to identify the travel mode of trajectory, and the optimized trajectory segmentation algorithm is used to decompose the human travel trajectory into travel modes. Markov models trained by trajectories of different travel modes are used to accurately predict travel trajectories. Experiments show that there are significant differences in the trajectory movement laws of different travel modes, and the prediction accuracy and distance deviation of the proposed method are obviously better than several benchmark methods.

Keywords: Markov model；trajectory classification；trajectory segmentation；trajectory prediction；travel mode

0 引 言

随着移动通信、卫星定位导航等技术的快速发展,越来越多的智能感知设备(智能手机、智能手表等)正逐渐走进人们的生活,大规模移动位置数据的获取和应用已成为可能. 特别是挖掘和分析车辆行驶轨迹、个人活动轨迹等数据中的内在规律,实现移动轨迹的预测,对优化城市规划、改善交通、提高服务便捷等具有十分重要的意义[1-2].

González等[3]认为:不同于以往所认知的基于列维飞行模型[4]或随机游走模型[5]的随机行为,人类的移动存在高度的时空规律性和可再现模式. 这种规律是轨迹位置预测等研究的基础,许多学者使用轨迹数据挖掘频繁模式,用于分析预测对象的移动行为,并进一步基于已知轨迹记录对对象下一时刻的位置进行预测. 人类移动轨迹是基于时间的坐标序列,对于这类问题,马尔可夫模型是一种较好的处理方法,

收稿日期: 2021-09-06；**录用日期:** 2022-01-28.
基金项目: 国家自然科学基金项目(61573077, U1808205).
责任编委: 龙建成.
†通讯作者. E-mail: geguo@yeah.net.

它能发现轨迹频繁模式与移动规律,即从数据集中发现频繁出现的模式,如轨迹数据中频繁出现A→B,其中就可能隐藏某种移动规律,因而被广泛应用于相关研究之中. Song等[6]提出了具有回退机制的马尔可夫模型,针对解决稀疏历史的问题,如果n个历史位置不曾出现在数据集中,则模型阶数将从n降低到$n-1$; Gambs等[7]提出了v-mobility Markov chain (v-MMC)模型,用于移动对象位置预测,其本质是高阶马尔可夫模型,但计算量较大; Asahara等[8]借助混合马尔可夫链模型预测行人运动,综合考虑行人的个人特征和历史状态进行预测.

对于基于马尔可夫模型的轨迹预测方法,扩大数据集能够有效提高预测准确率. 但随着数据量的增加,其计算量也将因状态数的增加而增加,需要更多的训练和预测时间; 对于这一问题,一种办法是使用效率更高的算法,另一种办法则是对轨迹分类,将具有相似规律的轨迹归为一类,根据类别进行更加细致的研究. 分类的标准是十分重要的,既要能突出每类的特殊性,也要注意部分与整体间的联系,不能陷入无限细分的误区中. Brockmann等[5]已经证明了人类移动在时空上的规律性,因此,很多研究者在这个方向做出了很多努力,大体上可以分为基于空间与时间的分类.

1) 基于空间的分类. 主要是基于轨迹位置点的空间分布关系进行分类,比如将一个地区划分为多个小区域,对每个小区域分别建立模型. Lv等[9]提出了一种基于HMM (hidden Markov model)的方法,在熵的基础上将用户区分为不同类别,比如有的人整天在城市里随意游荡,移动位置的不确定性较大,即熵值较高; 而有的人仅在有限位置移动,保持着有规律的出行,即熵值较低. 其实质是根据用户轨迹位置的空间分布频率进行分类,但需要足够的历史数据以判断用户类型. Rathore等[10]提出了一种基于马尔可夫模型的方法,根据轨迹间相似度对轨迹分类,适用于存在大量重叠轨迹的密集路网. 其关注的也是轨迹间的空间关系,但计算开销较大.

2) 基于时间的分类. 主要是基于轨迹位置点的时间关系进行分类,比如按照小时分类,或是按照工作日与周末分类. Mathew等[11]基于HMM的方法进行轨迹预测,将轨迹分为工作日的日间和夜间以及周末三类,并分别建立模型. 而Wang等[12]提出了一种基于马尔可夫模型的方法,进一步给出了分类的数学依据. 以小时为区间训练马尔可夫模型矩阵,计算不同矩阵间的KL (Kullback-Leibler divergence)散度表征矩阵间相似度,根据相似度对时间分类,用一阶马尔可夫链对两段之间的过渡进行建模. 以上都是根据时间对轨迹进行分类,但不同类型的用户,其移动模式及其时间范围可能是不同的.

上述研究或考虑轨迹的相似度,或按时间段对轨迹进行分类,但都存在一些问题: 没有考虑不同类别间的过渡模式; 所提出的分类原则难以满足所有用户. 本文认为可以考虑用户出行方式,根据出行方式对轨迹进行分类. 人们的出行方式与移动行为有一定的关联性,比如很少有人步行前往距离远的地方,而驾车仅在家附近行驶. 不同出行方式下的移动行为也存在差异,城市内的人员出行大多被限制在固定的路网中. 不同出行方式使用的道路并不相同,比如行人很少在高架桥上行走,汽车难以沿着地铁线路行驶,如果使用公交车或地铁等公共交通,其移动线路更是固定的. 因此,考虑出行方式对轨迹进行分类,或许能更好地挖掘移动轨迹中的移动模式与内在规律.

综上所述,本文认为进行轨迹位置预测时,可以考虑用户的出行方式. 对此,提出一种考虑出行方式的基于马尔可夫模型的轨迹预测方法,按照出行方式对轨迹分类,分别建立多个马尔可夫模型来预测目标的下一个位置,以提高预测准确率,减少模型训练时间. 为了区分轨迹的出行方式,采用基于用户GPS (global positioning system)轨迹数据的出行方式分类方法. 通过从轨迹中提取速度、加速度和航向变化速度等统计特征,使用XGBoost算法进行分类训练. 利用轨迹数据区分非机动方式(步行和自行车)、公交车、汽车(出租车和私家车)和地铁列车4种出行方式. 该方法可以离线运行,不需要地理信息系统(geographic information system, GIS)数据,并可达到较高的准确率. 采用基于优化的轨迹分割方法,确保轨迹中仅包含一种出行方式,以准确判断所提取轨迹的出行方式. 实验结果验证了所提出方法的有效性和优越性.

1　问题描述和系统模型

1.1　问题描述

本文所研究问题的目标是基于群体的移动轨迹数据来预测个人用户未来的轨迹位置,并考虑不同出行方式的影响. 具体而言,根据轨迹的出行方式对轨迹进行分类,分别建立马尔可夫模型$\{P^1, P^2, \ldots, P^k\}$来发现移动规律,根据个人用户当前的轨迹来判断采用的出行方式,采用对应的马尔可夫模型预测未来的轨迹位置. 由于轨迹点是离散数据,难以描述不同轨迹间的相似度,本文将地图划分成不相交的

区域,即网格化. 用网格坐标代替轨迹点经纬度坐标,网格精度越高,越贴近真实世界,计算量也越大. 通过网格化,降低了内存使用,提高了计算效率. 在网格化方法的表示下,本文所研究的轨迹预测问题,可以表示为预测用户下一时刻移动轨迹的网格坐标,并且可以将上一次的预测结果作为已知来预测下一位置,从而迭代地实现长期轨迹预测,但这个过程中,误差也会不断累积,难以准确预测长期移动轨迹. 因此,本文主要关注于用户的下一位置的预测,即根据已知轨迹坐标序列来预测用户在下一时刻的网格坐标. 综上所述,本文所研究的轨迹预测问题定义如下.

问题 已知群体移动轨迹记录$R^C=\{R^v|v\in U\}$,个人用户u当前移动轨迹记录$R^u=\{\mathrm{ST}_1,\mathrm{ST}_2,\ldots,\mathrm{ST}_n\}$,移动轨迹记录包含位置$s_j^u$和时间戳$t_j^u$,目标是预测用户$u$的下一个位置$s_{n+1}^u$.

注1 通过对轨迹分类,可以进行更细致的研究,文献[9-10]基于轨迹点的空间分布关系进行分类,但此类方法计算开销较大,也很难构建不同类别间的过渡模式. 文献[11-12]基于轨迹点的时间关系进行分类,但并不是所有人都有相同的时间规律. 本文则考虑出行方式的影响. 采用不同出行方式时,用户的移动时间、距离和道路都是不同的. 但因为很难直接获取轨迹出行方式,所以需要解决一些新问题:如何判断轨迹中用户采用的出行方式;如何将相同出行方式的轨迹点划归为一条轨迹.

1.2 马尔可夫模型

给定一个随机过程$\{x_t,t=0,1,\ldots\}$,其中每个状态都是一个有限的一元状态集$s\equiv\{1,2,\ldots,I\}$中的元素. 对于马尔可夫模型,是指系统过程中的任何一个时刻的状态只依赖于前面M个时刻的状态,被称为M阶马尔可夫模型,即

$$p=p(x_t=s_t|x_{t-1}=s_{t-1},\ldots,x_0=s_0)=p(x_t=s_t|x_{t-1}=s_{t-1},\ldots,x_{t-M}=s_{t-M}).\quad(1)$$

对于最简单、应用也最广泛的一阶马尔可夫模型,有

$$p_{ij}=p(x_t=j|x_{t-1}=i,x_{t-2}=s_{t-2},\ldots,x_0=s_0)=p(x_t=j|x_{t-1}=i).\quad(2)$$

其中:$i,j,s_{t-2},\ldots,s_0\in s$;$p_{ij}$表示该过程前一时刻状态为$i$时,在当前时刻转移到状态$j$的概率. 设$P=(p_{ij})$,则$P$为一阶马尔可夫矩阵的转移概率矩阵,并满足以下性质:

$$p_{ij}\geqslant 0,\quad(3)$$

$$\sum_{i=1}^{I}p_{ij}=1,\ j=1,2,\ldots,I.\quad(4)$$

所以由式(3)和(4),一阶马尔可夫模型的状态概率可以计算为

$$p(x_t=s_t)=\sum_{s_{t-1}}p(x_t=s_t,x_{t-1}=s_{t-1})=\sum_{s_{t-1}}p(x_t=s_t|x_{t-1}=s_{t-1})p(x_{t-1}=s_{t-1}).\quad(5)$$

此时N个随机变量的联合分布可以简化为

$$p(x_1,x_2,\ldots,x_N)=p(x_1)\prod_{n=2}^{N}p(x_n|x_{n-1}).\quad(6)$$

对于本文在轨迹位置预测任务中所建立的马尔可夫模型,状态集s中的元素为轨迹数据中所包含的所有位置点,其目标建立的马尔可夫模型P中的元素p_{ij}即为位置s_i向位置s_j的转移概率,即

$$p_{ij}=\frac{\mathrm{count}((s_i,s_j))}{\mathrm{count}(s_i)}.\quad(7)$$

其中:$\mathrm{count}((s_i,s_j))$是所有轨迹中包含序列$(s_i,s_j)$的轨迹数,$\mathrm{count}(s_i)$是所有轨迹中包含位置$s_i$的轨迹数.

注2 传统的单链马尔可夫模型[6]在面对大规模数据时需要消耗较多的时间,并只考虑了历史位置,若先前的位置是相同的,则将产生相同的预测结果. 为提高预测准确率,进一步发现移动规律,相关学者考虑了轨迹点空间分布关系[9-10]或时间关系[11-12],而本文则考虑出行方式的影响,根据轨迹中用户所使用的出行方式对移动轨迹分类,分别建立马尔可夫模型$P=\{P^1,P^2,\ldots,P^k\}$.

2 轨迹分类和分割方法

本文所使用的数据集是来自微软研究项目的GeoLife数据集,其中部分轨迹有一个关于出行方式的标签,标识出了不同时间段所采用的出行方式. 但大部分轨迹记录中不包含出行方式信息,实际生活中,也难以要求用户上报当前的出行方式. 因此,需要设计合理的轨迹分类方法以判断轨迹的出行方式. 通过提取轨迹中速度、加速度、航向变化速度及其统计特征,拥有标签的轨迹将被用以训练判断出行方式的分类器. 但数据集中没有标识出单次出行的轨迹,还需要将不同出行方式的数据点区分开来,以将同一出行方式的数据点划归在一条轨迹中,准确判断出行方式. 不同出行方式在移动速度、加速度等方面存在差异,因此当其发生显著变化时,就有可能发生了出行方式的改变. 本文将轨迹分割问题转化为一个优化问题[13],从轨迹序列中寻找变化点,使得

拟合代价最小,通过3个步骤对轨迹进行处理,得到仅包含一种出行方式的轨迹,具体内容将在下文中介绍.

2.1 相关定义

定义1 轨迹:用户u的轨迹R^u为一系列带有时间戳的GPS数据点,表示为$R^u = (\mathrm{ST}_1, \mathrm{ST}_2, \ldots, \mathrm{ST}_m)$,其中$\mathrm{ST}_i = (x_i, y_i, t_i)$.

定义2 变化点:表示为CP,被定义为用户在出行中改变其出行方式的点,一次出行轨迹中可能包含零个、一个或多个变化点.

定义3 轨迹段:是用户轨迹的进一步细分,仅包含一种出行方式,表示为$\mathrm{SE} = [\mathrm{ST}_1, \mathrm{ST}_2, \ldots, \mathrm{ST}_n]$.

2.2 轨迹分类算法

为找到最适合轨迹分类任务的算法,本文选择以下分类算法并比较它们的结果,包括极端随机树[14]、XGBoost[15]、LightGBM[16]、Adaboost[17].

极端随机树(extremely randomized trees)[14]是集成学习方法的一种,具有不容易过拟合、抗噪性能良好的优点. 其由多棵决策树[18]组成,每颗决策树都是基于全部数据训练得到的,但在特征的选择上是随机的,多颗决策树组合在一起时往往可以达到较好的预测结果.

XGBoost[15]由许多CART(classification and regression tree, CART)树组成,能够并行处理. 将样本随机分成k份而生成决策树,将多棵决策树并行聚合. 在树形分裂过程中,允许每一轮迭代中使用交叉验证方法获取最优迭代参数,并在代价函数中引入正则化项来控制模型复杂度,以防止过度拟合.

LightGBM[16]是一种分布式的基于决策树算法的梯度提升算法. 通过单边梯度采样(gradient-based one-side sampling, GOSS)来缩小特征分割选择范围,还采用了互斥特征捆绑策略以及带深度限制的叶子生长策略,进一步提高了计算速度.

Adaboost[17],其思想是训练不同的弱分类器,将弱分类器组合构成一个强分类器. 其具有较快的计算速度,不容易过拟合. 与极端随机树相比,Adaboost根据加权投票来决定预测结果,每个弱分类器的权重与错误率有关,而极端随机树则根据所有树的预测结果,以少数服从多数的原则来决定预测结果.

2.3 轨迹分类特征

不同出行方式所导致的差异不仅体现在出行距离、使用道路等方面,更直接表现在运动特征上,如对速度而言,汽车最快,走路最慢. 但速度也易受交通条件的影响,如交通拥挤时,汽车的平均速度可能与自行车一样慢. 因此,不仅需要选取速度等运动特征的包括平均值、百分位数等统计量,还应考虑航向变化速度[19]、停止率等不易受交通条件影响的特征. 本文从轨迹中提取相关特征,形成特征向量,通过分类算法来判断出行方式. 文中所使用的特征可以分为两类,即轨迹数据点特征和轨迹段特征.

轨迹点特征,即对每个轨迹点运动参数进行计算,包括速度、加速度、航向变化速度,计算方法如下.

1) 速度:对于包含n个数据点的轨迹,能得到$n-1$个时间间隔的速度. 通过计算每两个相邻数据点之间的距离和时间间隔,可以得到距离Δs_i、时间长度Δt_i和平均速度v_i,计算公式如下:

$$v_i = \frac{\Delta s_i}{\Delta t_i}. \tag{8}$$

2) 加速度:对于包含n个数据点的轨迹,能得到$n-2$个时间间隔的加速度,计算公式如下:

$$a_i = \frac{v_i - v_{i-1}}{\Delta t_i}. \tag{9}$$

3) 航向变化速度[19]:对于包含n个数据点的轨迹,能得到$n-2$个时间间隔的航向变化速度,计算公式如下:

$$\Delta\mathrm{long}_i = (\mathrm{long}_i - \mathrm{long}_{i-1}) \cdot \cos(\mathrm{lat}_i), \tag{10}$$

$$\Delta\mathrm{lat}_i = \mathrm{lat}_i - \mathrm{lat}_{i-1}, \tag{11}$$

$$\mathrm{hc}_i = \cos^{-1}\Big(\frac{r_i \cdot r_{i-1}}{\sqrt{r_i \cdot r_i}\sqrt{r_{i-1} \cdot r_{i-1}}}\Big), \tag{12}$$

$$\mathrm{hcs}_i = \frac{\mathrm{hc}_i}{\Delta t_i}. \tag{13}$$

其中:$r_i = (\Delta\mathrm{long}_i, \Delta\mathrm{lat}_i)$,$\mathrm{long}_i$、$\mathrm{lat}_i$代表第$i$个数据点的经度和纬度.

为了更好地发现不同出行方式的差异,还需要提取轨迹段特征,即轨迹点特征在整段轨迹上的统计特征. 所提取的特征如下.

1) 平均值: 反映轨迹段中特征值的集中趋势.

2) 方差: 反映轨迹段中特征值的离散程度.

3) 百分位数: 数据位置的度量,它反映轨迹段中特征值在最小值与最大值之间如何分布,在本文中,选择第25百分位数(下四分位数)和第75百分位数(上四分位数).

4) 四分位范围:上四分位与下四分位之间的差值.

通过计算轨迹段中速度、加速度、航向变化速度的平均值、方差等特征,共能得到共15个特征. 此外再考虑3个全局特征:

5) 总距离: 计算公式为

$$S_{\text{total}} = \sum \Delta s_i. \tag{14}$$

6) 停止率: 速度等于或小于0.6 m/s的间隔在轨迹中的比例,即

$$S_{\text{rate}} = \frac{\text{count}_{v_i \leqslant 0.6\,\text{m/s}}}{n-1}. \tag{15}$$

7) 位移路程比: 位移与路程的比值. 位移,即一段轨迹中起点与终点间的距离;路程,即该轨迹的总距离. 有

$$S_{sd} = \frac{\text{distance}(p_0, p_n)}{\sum \Delta s_i}. \tag{16}$$

文献[20]在利用GPS轨迹数据判断出行方式时还考虑了用户的个人信息;文献[21]用GPS轨迹数据结合大量GIS数据,包括道路网络、地铁网络和实时公交位置,判断出行方式. 上述方法可以达到较高的准确度,但也依赖于足够的个人隐私数据和大量的GIS数据,一方面信息数据难以获取,另一方面计算开销较大. 本文通过提取GPS轨迹特征,无需GIS信息,在保证方法准确度的同时,降低了算法复杂度.

2.4 轨迹分割算法

轨迹分割方法共由3个步骤组成. 第1步,按照基于时间阈值的方法从数据集中提取轨迹,这仅是粗略处理,还需要进一步的细致分割. 第2步,将轨迹拆分为数据点数量小于M的轨迹段,此后,几乎所有轨迹段中将只包含一种或两种出行方式,仍需要下一个步骤来保证轨迹段中只包含一种出行方式. 第3步,将轨迹段表示为$\text{SE} = [\text{ST}_1, \text{ST}_2, \ldots, \text{ST}_n], n \leqslant M$,目标是从中寻找$K$个变化点,即$[\text{CP}_1, \text{CP}_2, \ldots, \text{CP}_K]$,出行方式发生变化时,变化点间的轨迹特征也将发生变化,于是关于轨迹分割的任务就转化为从序列中寻找变化点. 这里计算SE的相关数据特征,用多元时间序列$\{Y_t\}$表示,本文使用速度和加速度作为轨迹特征. 从时间序列信号中寻找变化点的一种常见方法是将其视为一个优化问题,用代价函数来衡量序列的拟合度. 综上,第3步可以被转化为求解以下目标函数[22]:

$$C(\text{CP}) = \sum_{i=1}^{k+1} [c(Y_{\text{CP}_{i-1}+1:\text{CP}_i})] + \gamma f(k). \tag{17}$$

其中:$\text{CP}_0 = 0, \text{CP}_{K+1} = m, Y_{\text{CP}_{i-1}+1:\text{CP}_i}$是连续变化点间的轨迹段,$c$是衡量轨迹段拟合度的代价函数. 本文选择均值漂移模型作为代价函数,对于两个连续变化点间的轨迹段,均值漂移定义为

$$c(Y_{\text{SE}}) = \sum_{i \in \text{SE}} \|Y_i - \bar{Y}\|_2^2, \tag{18}$$

其中$\bar{Y}$是序列$\{Y_{\text{SE}}\}$的平均值.

因为一段轨迹中的出行方式是不确定的,所以变化点的数量K是未知的. 在式(16)中使用了惩罚函数$f(K)$来约束变化点的数量,避免过度地分割,并通过改变惩罚系数γ来调整得到的变化点数量,而不是预先定义变化点数量.

通过以上3个步骤,就得到了仅包含一种出行方式的轨迹段. 使用前文介绍的出行方式分类器来判断所采用的出行方式,如果连续轨迹段的出行方式相同,则将其合并为一个轨迹段. 轨迹分割算法的伪代码如下.

算法1 轨迹分割算法.

输入: 按时间阈值分割的轨迹集合S,阈值M;

输出: 仅包含一种交通方式的轨迹traj.

1) initialize $\text{traj} = [\,]$

2) for traj in S

3) $\text{seg} = \text{trajToSeg}(M)$ // 将轨迹traj划分为轨迹点数量小于M的轨迹段

4) for SE in seg

5) $\text{CP} = \text{Optimization}(\text{SE})$ // 找出代价最小的变化点CP

6) for i in range(len(CP)):

7) if $\text{traj}[-1].\text{TravelMode} == \text{SE}_{\text{CP}_{i-1}:\text{CP}_i}.\text{TravelMode}$ then

8) $\text{traj}[-1] = \text{traj}[-1] + \text{SE}_{\text{CP}_{i-1}:\text{CP}_i}$ // 合并出行方式相同的轨迹段

9) else

10) $\text{traj.append}(\text{SE}_{\text{CP}_{i-1}:\text{CP}_i})$

11) end if

12) end for

13) end for

14) end for

注3 对轨迹进行分类,很重要的是根据预设的定义计算轨迹相关特征并进行区分. 文献[9]通过用户出现的位置及其概率计算熵向量,使用聚类算法对轨迹进行分类,其需要足够的历史数据来判断用户类型,可能存在冷启动问题. 文献[10]在不考虑轨迹方向的前提下计算轨迹间DTW (dynamic time warping)距离,根据轨迹间距离进行分类,并计算每类的代表性轨迹,该轨迹是“虚构”的,可能不属于该类中的任何轨迹,仅描述该类轨迹的主要运动模式,根据待预测轨迹与代表性轨迹之间的距离判断所属的类别. 文献[11-12]则根据轨迹位置点的时间对轨迹进行分类,计算量小,但不同类型的用户,其关于时间的模式可能是不同的. 而本文通过提取轨迹特征判断

出行方式,使用基于优化的轨迹分割方法,将同一出行方式的位置点划归为一条轨迹,在保证方法准确性的同时,降低了复杂度和计算量.

3 实 验

本文所使用的GeoLife数据集来自微软研究项目,从2007年4月到2012年8月共收集了182个用户的GPS轨迹数据,该数据集内包括一系列含时间戳的位置点、经纬度、海拔等信息. 在本节中,首先应用轨迹分类算法来训练判断轨迹出行方式的分类器,然后将数据点分割为仅包含一种出行方式的轨迹,再根据出行方式类别分别建立马尔可夫模型用于轨迹位置预测,通过与其他方法进行比较来验证本文所提出方法的有效性和优越性.

3.1 分类器的训练

数据集中除了用户的GPS轨迹数据以外,还包括部分用户的出行方式标签,每个标签包含用户ID、开始时间、结束时间、出行方式等信息. 将标签与轨迹数据记录进行匹配,就得到了与出行方式相匹配的轨迹. 出行方式标签包括walk(步行)、bike(自行车)、bus(公交车)、car(私家车)、taxi(出租车)、train(火车)、airplane(飞机)等. 因本文更关注于城市内居民的交通出行,故剔除airplane(飞机)和train(火车). 考虑移动规律的相似性且简化模型训练过程,将car(私家车)、taxi(出租车)合并为car(汽车). 根据轨迹的出行方式分别建立马尔可夫模型矩阵,计算JS散度(jensen－shannon divergence)以表征各矩阵间的差异性. JS散度值越接近0,表示矩阵间差异性越小;越接近1,表示矩阵间差异性越大. 计算结果见表1. JS散度的计算方法如下:

$$\mathrm{JS}(P||Q)=\frac{1}{2}\mathrm{KL}(P||M)+\frac{1}{2}\mathrm{KL}(Q||M), \tag{19}$$

$$M=\frac{1}{2}(P+Q), \tag{20}$$

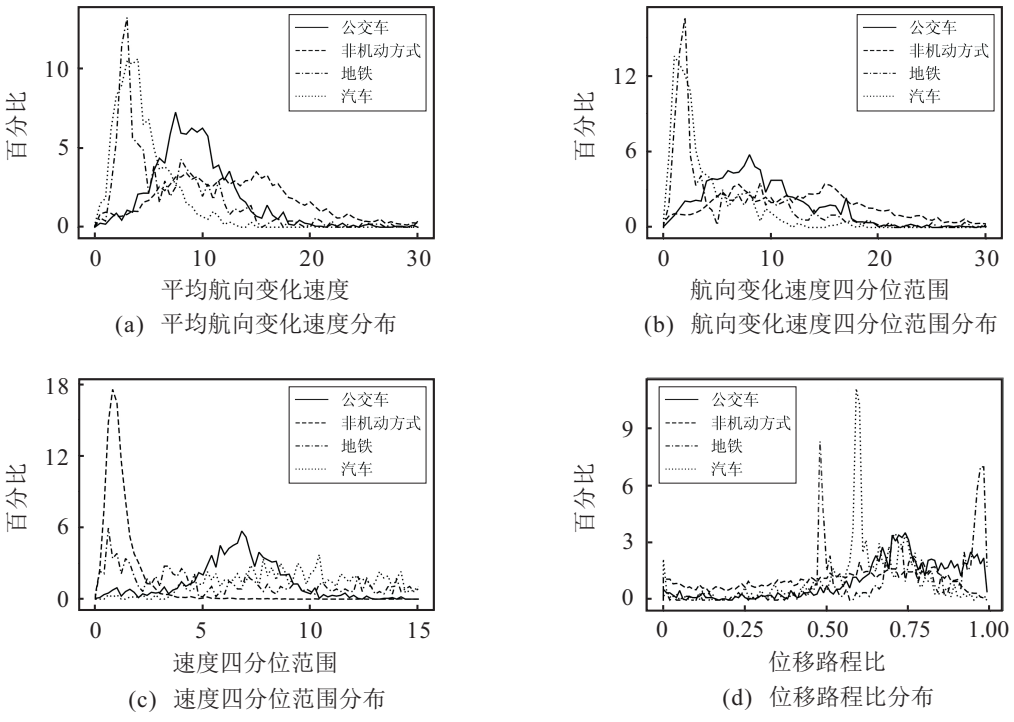
$$\mathrm{KL}(P||M)=\sum P(x)\log\left(\frac{P(x)}{M(x)}\right)\cdot\pi_p, \tag{21}$$

其中π_p表示马尔可夫矩阵P的稳态分布.

表1　各出行方式马尔可夫矩阵间JS散度

出行方式	walk	bus	subway	car	bike
walk	0	0.619 4	0.481 7	0.995 4	0.124 1
bus	0.619 4	0	0.604 7	0.530 4	0.668 8
subway	0.481 7	0.604 7	0	0.682 2	0.693 0
car	0.995 4	0.530 4	0.682 2	0	0.452 6
bike	0.124 1	0.668 8	0.693 0	0.452 6	0

比较各出行方式的马尔可夫矩阵间的JS散度,发现不同出行方式的移动模式确实存在差异,所以根据出行方式对轨迹分类并进行轨迹预测是有依据的且可行的. 此外,发现walk(步行)、bike(自行车)两种出行方式的差距较小,其JS散度值为0.124 1,为简化模型训练,将这两类合并为non-motorized(非机动方式). 所以最终用于分类器的出行方式标签为non-motorized(非机动方式)、bus(公交车)、subway(地铁)、car(汽车). 采用前文中介绍的方法提取轨迹数据特征. 图1展示了一些特征数据的分布情况.



(a) 平均航向变化速度分布　(b) 航向变化速度四分位范围分布

(c) 速度四分位范围分布　(d) 位移路程比分布

图1　部分特征数据分布情况

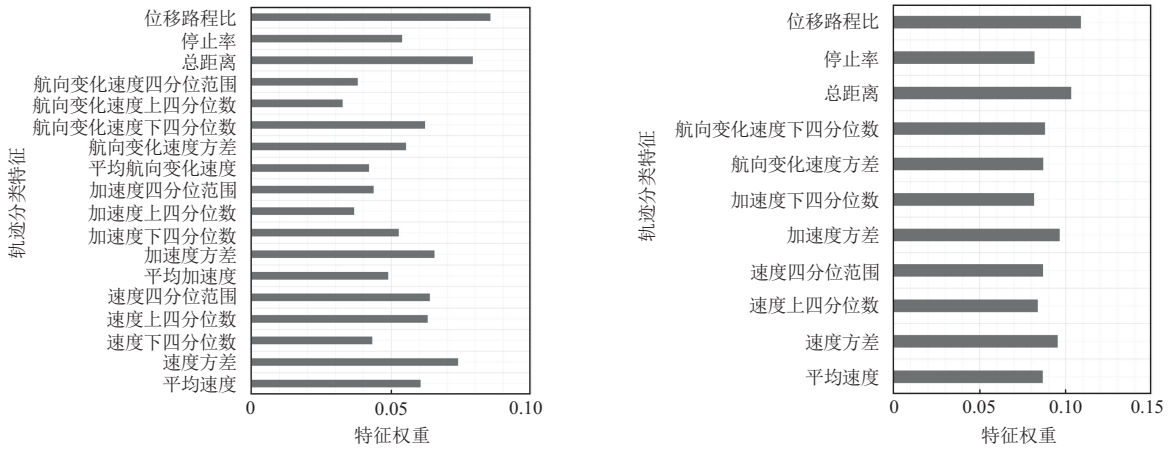
比较图1中不同出行方式的特征数据的分布可以看出,不同出行方式在移动特征方面存在较大差异,可以据此判断轨迹所使用的出行方式. 根据提取到的全部18个特征,使用前文介绍的极端随机树、XGBoost、LightGBM、Adaboost分类算法分别构建分类器,训练数据集情况见表2,预测结果见表3. 其中XGBoost算法的分类正确率最高,为93.89 %. 为进一步提高分类正确率,对各特征的重要性进行了评估,如图2所示,并进行特征筛选,在剔除了权重较低的特征后,分类准确率提高至94.24 %. 利用该分类器可对数据集中更多的没有出行方式标签的轨迹进行分类,但由于数据集中没有标识出单次出行的轨迹,还需要采用前文中介绍的轨迹分割方法,从数据集中提取、分割轨迹,使得到的轨迹中仅包含一种出行方式.

表2 出行方式分类器训练集

出行方式	轨迹数量
非机动方式	4 487
公交车	1 593
地铁	518
汽车	961
合计	7 559

表3 分类算法的正确率

分类算法	分类正确率/%
极端随机树	92.64
XGBoost	93.89
LightGBM	93.54
Adaboost	89.72
XGBoost(特征筛选后)	94.24



(a) 筛选前特征权重 (b) 筛选后特征权重

图2 各个特征权重

3.2 马尔可夫模型预测

通过前文介绍的基于优化的轨迹分割方法,从数据集中提取分割得到仅包含一种出行方式的轨迹,并通过分类器将轨迹按出行方式分类,共得到了25 375条轨迹作为训练集,见表4. 分别建立马尔可夫模型,一共得到4个矩阵.

表4 训练集轨迹组成

出行方式	轨迹数量	所占比例/%
非机动方式	8 599	33.89
公交车	4 817	18.98
地铁	5 980	23.57
汽车	5 979	23.56

3.3 实验结果

为了评估算法的表现,本文采用2个评价指标:预测准确率和预测平均偏差.

1) 预测准确率. 已知轨迹$\{\mathrm{ST}_1,\mathrm{ST}_2,\ldots,\mathrm{ST}_n,\mathrm{ST}_{n+1}\}$,预测轨迹$\{\mathrm{ST}_1,\mathrm{ST}_2,\ldots,\mathrm{ST}_n,T_{n+1}\}$,$\mathrm{dist}(p,q)$表示$p$和$q$两点间的欧氏距离,$H(\mathrm{ST}_{n+1},T_{n+1})$表示是否预测正确. 当$\mathrm{dist}(p_{n+1},q_{n+1})\leqslant\varepsilon$时,表示预测正确,此时有$H(\mathrm{ST}_{n+1},T_{n+1})=1$,所以预测准确率为

$$\mathrm{accuracy}=\frac{\sum H(\mathrm{ST}_{n+1},T_{n+1})}{|\mathrm{traj}_p|}. \tag{22}$$

其中:$|\mathrm{traj}_p|$表示轨迹的数量;$H(\mathrm{ST}_{n+1},T_{n+1})$定义如下:

$$H(\mathrm{ST}_{n+1},T_{n+1})=\begin{cases}1,\ \mathrm{dist}(\mathrm{ST}_{n+1},T_{n+1})\leqslant\varepsilon;\\ 0,\ \mathrm{dist}(\mathrm{ST}_{n+1},T_{n+1})>\varepsilon.\end{cases} \tag{23}$$

2) 预测平均偏差. 已知轨迹$\{\mathrm{ST}_1,\mathrm{ST}_2,\ldots,\mathrm{ST}_n,\mathrm{ST}_{n+1}\}$,预测轨迹$\{\mathrm{ST}_1,\mathrm{ST}_2,\ldots,\mathrm{ST}_n,T_{n+1}\}$,则平均偏差距离定义为

$$\mathrm{deviation}_{\mathrm{avg}}=\frac{\sum\mathrm{dist}(\mathrm{ST}_{n+1},T_{n+1})}{|\mathrm{traj}_p|}. \tag{24}$$

本文中ε取值为400 m,预测的时间间隔为5 min.

为了表明本文所提出方法的优势,将其与以下几种基准预测方法进行比较:

1) 具有回退机制的单链马尔可夫模型[6]:为了提高马尔可夫模型在稀疏历史上的性能,如果当前位置的n个历史位置不曾出现在训练集中时,则马尔可夫模型的阶数将从n降低到$n-1$;

2) 基于空间分类的马尔可夫模型[9]:根据用户轨迹位置的空间分布频率,即熵,对用户进行分类,分别建立模型;

3) 基于时间分类的隐马尔可夫模型[11]:这是马尔可夫模型的扩展,假设观测状态是由不可观测的隐状态间的马尔可夫过程生成的,其根据工作日白天、工作日夜间、周末对轨迹分类,分别建立模型;

4) 基于时间分类的马尔可夫模型[12]:根据不同时间段的马尔可夫模型矩阵间的KL散度,将数据集所按照的时间分为3类,即7 : 00 am～ 1 : 00 pm, 1 : 00 pm～ 7 : 00 pm, 7 : 00 pm～7 : 00 am (+1)分别建立马尔可夫模型矩阵;

5) XGBoost轨迹预测器:该方法基于XGBoost算法[15],根据历史轨迹位置作为特征,对下一位置进行预测;

6) 神经网络轨迹预测器:该方法基于RNN (recurrent neural network)分类器,由一个输入层、两个隐藏层和一个输出层组成,以历史轨迹作为特征,对下一位置进行预测.

表5展示了不同方法的预测效果,可以发现本文所提出的方法在预测准确度和预测平均偏差上都优于其他方法,说明对于轨迹预测,考虑出行方式是可行且有效的. 基于时间分类的马尔可夫模型考虑了不同时间段用户的移动模式可能存在差异,为每个时间段分别建立马尔可夫模型,在6种对比方法中表现最好,但没有考虑不符合该时间规律的用户. 具有回退机制的马尔可夫模型仅建立了一组转移内核,当用户改变移动模式时,其预测结果就会受到影响. 而对于基于机器学习模型XGBoost以及神经网络RNN的预测方法,神经网络轨迹预测器表现更好,但由于数据集中包含坐标点数量众多,而包含某一具体坐标点的轨迹数据却数量较少,难以实现准确预测. 综上所述,本文所提出的方法在大多数情况下都比其他算法表现更好,从而验证了本文方法的优越性.

表 5　不同方法预测效果对比

预测方法	预测准确率/%	预测平均偏差/m
基于出行方式分类	75.02	511.47
单链模型	71.85	778.82
基于空间分类	66.57	683.77
HMM	68.98	688.87
基于时间分类	73.23	561
XGBoost	66.81	627.73
RNN	69.52	665.78

对于不同的出行方式,预测结果存在一定差异. 表6展示了不同出行方式的预测效果. 其中:对于公交车的预测准确率最高,预测平均偏差距离最小,可能是公交车行驶路线更为固定,使得能实现更加准确的预测;对于与公交车类似的地铁表现较差,可能是由于位于地面下,其GPS数据存在更大的扰动和误差;而对于非机动方式,其预测准确率较低,可能是其移动具有更多的不确定性和偶然性,缺乏目的性. 对于未来位置的预测主要是基于用户的历史出行规律,因此对于偶然性、随机性较强的行为表现较差,难以实现更准确的预测. 此外,对于不同出行方式的预测与前文介绍的基准方法相比,仅基于时间分类的马尔可夫模型效果优于非机动方式和地铁,而公交车、汽车的预测结果均优于基准方法,这也说明了不同出行方式下移动规律的差异和本文方法的优越性.

表 6　不同出行方式预测效果对比

出行方式	预测准确率/%	预测平均偏差/m
非机动方式	71.91	562.97
公交车	79.06	397.41
地铁	72.36	630.56
汽车	77.23	517.65

4　结　论

为了准确地预测用户未来的轨迹位置,本文使用能够有效描述轨迹位置点之间状态转移关系的马尔可夫模型,并考虑了用户采用不同出行方式时移动规律的不同. 为了判断不同的出行方式,本文首先从轨迹中提取速度、加速度、航向变化速度等统计特征,通过XGBoost算法训练判断轨迹出行方式的分类器;然后采用基于优化的轨迹分割方法,使得分割得到的轨迹中仅包含一种出行方式;接着使用训练得到的分类器对轨迹进行分类,将轨迹分为non-motorized (非机动方式)、bus (公交车)、subway (地铁)、car (汽车)四类;进而分别建立马尔可夫模型用于轨迹位置预测,并计算了矩阵间差距,验证了不同出行方式下移动模式的差异;最后通过实验验证了本文所提出的预测算法的效果和优越性. 但仍存在一些问题需要进一步研究和解决:

1) 由于大部分数据集中没有关于出行方式的标签,本文为了使得到的轨迹仅包含一种出行方式,采

用基于优化的轨迹分割方法,但仍可能无法实现理想的分割,即得到的轨迹中包含一种以上的出行方式,从而影响对轨迹出行方式的判断,因此仍需进一步研究.

2) 需要进一步提高轨迹分类的准确率. 本文关于轨迹分类的特征仅包括速度、加速度、航向变化速度等特征,这使得算法具有速度快、可以离线预测等优点. 因此,可以进一步考虑GIS信息,提高分类准确率.

参考文献(References)

[1] Shi H Z, Li Y, Cao H C, et al. Semantics-aware hidden Markov model for human mobility[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(3): 1183-1194.

[2] Yang Z, Sun H L, Huang J B, et al. An efficient destination prediction approach based on future trajectory prediction and transition matrix optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(2): 203-217.

[3] González M C, Hidalgo C A, Barabási A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453 (7196): 779-782.

[4] Edwards A M, Phillips R A, Watkins N W, et al. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer[J]. Nature, 2007, 449(7165): 1044-1048.

[5] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel[J]. Nature, 2006, 439(7075): 462-465.

[6] Song L B, Kotz D, Jain R, et al. Evaluating next-cell predictors with extensive Wi-Fi mobility data[J]. IEEE Transactions on Mobile Computing, 2006, 5(12): 1633-1649.

[7] Gambs S, Killijian M O, del Prado Cortez M N. Next place prediction using mobility Markov chains[C]. Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility — MPM'12. New York: ACM Press, 2012: 1-6.

[8] Asahara A, Maruyama K, Sato A, et al. Pedestrian-movement prediction based on mixed Markov-chain model[C]. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems — GIS'11. New York: ACM Press, 2011: 25-33.

[9] Lv Q J, Qiao Y Y, Ansari N, et al. Big data driven hidden Markov model based individual mobility prediction at points of interest[J]. IEEE Transactions on Vehicular Technology, 2017, 66(6): 5204-5216.

[10] Rathore P, Kumar D, Rajasegarar S, et al. A scalable framework for trajectory prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3860-3874.

[11] Mathew W, Raposo R, Martins B. Predicting future locations with hidden Markov models[C]. Proceedings of the 2012 ACM Conference on Ubiquitous Computing — UbiComp'12. New York: ACM Press, 2012: 911.

[12] Wang H, Zeng S, Li Y, et al. Predictability and prediction of human mobility based on application-collected location data[J]. IEEE Transactions on Mobile Computing, 2021, 20(7): 2457-2472.

[13] Dabiri S, Lu C T, Heaslip K, et al. Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(5): 1010-1023.

[14] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[15] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[16] Meng Q, Ke G. LightGBM: A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017, 30: 3147-3155.

[17] Yu H H, Mouliny P. Regularized Adaboost for content identification[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, 2013: 3078-3082.

[18] Rokach L, Maimon O. Decision trees[J]. IEEE Transactions on Systems Man & Cybernetics: Part C, 2005, 35(4): 476-487.

[19] Zheng Y, Chen Y K, Li Q N, et al. Understanding transportation modes based on GPS data for web applications[J]. ACM Transactions on the Web, 2010, 4(1): 1-36.

[20] Bantis T, Haworth J. Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics[J]. Transportation Research — Part C: Emerging Technologies, 2017, 80: 286-309.

[21] Biljecki F, Ledoux H, Van Oosterom P. Transportation mode-based segmentation and classification of movement trajectories[J]. International Journal of Geographical Information Science, 2013, 27(2): 385-407.

[22] Killick R, Fearnhead P, Eckley I A. Optimal detection of changepoints with a linear computational cost[J]. Journal of the American Statistical Association, 2012, 107(500): 1590-1598.

作者简介

郭戈(1972–), 男, 教授, 博士生导师, 从事网联车辆协同控制、智能交通系统、共享出行系统优化与控制等研究, E-mail: geguo@yeah.net;

胡峻豪(1998–), 男, 硕士生, 从事智能交通系统、轨迹预测的研究, E-mail: 750336437@qq.com.

(责任编辑: 李君玲)