

DOI:10.16452/j.cnki.sdkjzk.2021.01.008

文章编号:1672-3767(2021)01-0061-10

# 基于 GPS 数据的交通出行模式及转换点识别算法

朱少武,孙海春,罗万杰,赵晓凡

(中国人民公安大学 信息技术与网络安全学院,北京 102623)

**摘 要:**GPS 产生的大量数据携带了许多潜在信息。针对一段连续的 GPS 轨迹数据进行研究,识别 GPS 轨迹中的出行模式转换点;以出行模式转换点为分割点,将连续的 GPS 轨迹分割成多个 GPS 轨迹段;利用训练好的随机森林模型识别出不同 GPS 轨迹段的出行模式(包括步行、骑行、公交、地铁、汽车等类型),从而得到一段 GPS 轨迹的完整的出行模式描述。相比以前的研究,区别在于:提出一种基于低速度移动研判的出行模式转换点识别方法;新增加了“稳定速度”和“平均停留间隔”两种区别性特征,对已有基于随机森林的出行模式判定算法进行了改进。实验结果表明:改进之后的出行模式判定算法平均准确率提高了 1.3%;同时,出行模式转换点识别方法大幅降低了算法的时间复杂度,使得 GPS 轨迹数据出行模式研判算法的  $F$ -score 值提高了 4% 左右。

**关键词:**GPS;随机森林模型;转换点识别;移动窗口;交通出行模式

**中图分类号:**TP391

**文献标志码:**B

## Traffic travel mode and conversion point recognition algorithm based on GPS data

ZHU Shaowu, SUN Haichun, LUO Wanjie, ZHAO Xiaofan

(College of Information Technology and Internet Security, People's

Public Security University of China, Beijing 102623, China)

**Abstract:** Big data generated by GPS carries plenty of hidden information. This paper focuses on the travel mode judgment based on GPS data. Initially, a section of continuous GPS trajectory data was selected and divided into different trajectory segments by conversion point recognition. Afterwards the travel modes of different GPS trajectory segments (such as walking, riding, bus, subway, and car) were identified by using the random forest model. Compared with previous studies, the present study proposed a new method of travel mode conversion point recognition based on low-speed motion judgment, added the two distinguishing features of stable velocity and average stop interval, and improved the existing travel mode determination algorithm based on random forest. The experimental results show that the average accuracy of the improved algorithm is improved by 1.3%. The time complexity of the algorithm is greatly reduced compared with the previous methods, and the  $F$ -score value is increased by about 4%.

**Key words:** GPS; random forest model; conversion point recognition; moving window; traffic travel mode

伴随着移动通讯设备普及应用,全球定位系统(global position system, GPS)产生的数据对个人位置的展现越来越充分。依据 GPS 数据识别个人的交通出行模式,可以对居民出行信息准确摸排,有助于城市交通规划和管理,解决当前城市中存在的公交线路规划、交通拥堵等问题;也可分析个人的出行特点,并有针对性地推荐相关服务。利用转换点识别算法识别出 GPS 轨迹数据中的出行模式转换点,可以分段对 GPS 数据

**收稿日期:**2019-05-23

**基金项目:**国家重点研发计划项目(2017YFC0803700);北京市自然科学基金项目(4184099);中国人民公安大学新任教师科研启动基金项目(2018JKF609)

**作者简介:**朱少武(1998—),男,天津人,硕士研究生,主要从事数据挖掘的研究。

孙海春(1985—),女,山东聊城人,讲师,博士,主要从事数据挖掘的研究,本文通信作者。

E-mail: sunhaichun@ppsuc.edu.cn

进行分析研判,给出出行模式标签,最终得到整段GPS数据的出行模式变化过程。

分类算法的不同对最终识别结果的精度有很大的影响。出行模式识别算法主要使用神经网络、随机森林、支持向量机等分类算法。Liang等<sup>[1]</sup>利用卷积神经网络(convolutional neural networks, CNN)进行识别,并加入一些滤波算法对数据进行平滑处理来降低数据的波动,其最高识别精度在94%左右。Xiao等<sup>[2]</sup>利用贝叶斯网络和K2算法,最终得到的步行模式识别率超过97%。Martin等<sup>[3]</sup>开发了一种新的分类算法,并与K近邻分类算法(k-nearest neighbor, KNN)和随机森林算法进行对比,实验表明结合随机森林算法的识别准确率更高,约为94%。Zhu等<sup>[4]</sup>提出一种出行模式选择模型和一种有向图引导的融合Lasso方法,降低了出行模式研判算法的时间复杂度。Guvensan等<sup>[5]</sup>提出基于分段的传输模式检测体系结构,其综合识别准确率在93%左右。此外,有效的识别特征对提升出行模式识别准确率很有帮助。Sun等<sup>[6]</sup>结合谱聚类和隐Markov模型,加入位置信息和信号强度等特征,得到的出行模式算法平均识别准确率超过93%。Dabiri等<sup>[7]</sup>使用CNN算法,利用急动度和方向变化率辅助算法的识别,相比之前研究的准确率提升超过10%。

一些研究利用深度数据挖掘大幅度提高了单种出行模式的识别准确率。Wang等<sup>[8]</sup>提出地铁模式单独识别的方法,并结合地理信息系统(geographic information system, GIS)进行精度识别,使得地铁模式的识别准确率超过98%。Zong等<sup>[9]</sup>提出步行/骑行模式单独识别的方法,在3万余条数据中识别正确率接近100%。然而,一段出行记录中往往涉及多种出行模式。确定出行模式转换点是识别过程的难点,而要精准地识别出一段GPS日志中的不同出行模式更为困难。针对多种模式转换问题的深入研究较少,沈云<sup>[10]</sup>对转换点进行了研究,对比了基于多段窗口识别和基于移动窗口识别,其多段窗口识别率达到78.8%,移动窗口识别率达到76.7%。

目前围绕识别算法的改进大多侧重于提供新的识别算法和新的运动特征,得到了较高的识别精度。已有研究中速度特征选取以平均速度和分位速度为主,但平均速度和分位速度受其他干扰因素影响较大。受到整体速度变化过程的影响,每个静止点的增加会在一定程度上降低平均速度和分位速度。同时,停留点的次数特征受到出行距离的影响,出行距离长的样本的停留次数相对更高。本研究通过“稳定速度”和“平均停留间隔”两个特征来弥补以上不足,并提出了基于低速度移动研判的出行模式转换点识别方法,通过转换点将GPS轨迹分段研判,进一步提高出行模式识别的准确率。

## 1 基于低速度移动研判的转换点识别方法

出行模式转换点识别是出行模式识别过程的重要组成部分,转换点识别的准确率很大程度上影响了出行模式研判的正确率。目前有两种典型的转换点识别方法:基于多段窗口的识别方法和基于移动窗口的识别方法<sup>[10]</sup>。在移动窗口的识别方法中,通常利用遍历的方式,计算每一个数据点窗口两侧的欧氏距离;比较两侧的相似度大小,将低于一定阈值的点认定为转换点。

通过观察个人GPS轨迹数据,对个人出行模式中交通方式的转换过程进行了分析,发现:①每一次交通出行模式转换均包含一定的时间间隙,而这段时间间隙总是被步行模式或者静止模式占用;②由于步行模式和静止模式本身的速度特性,相对其他的出行模式来说可以较为简便地区分。利用以上特征,对基于移动窗口的转换点识别方法进行改进,提出一种基于低速度移动研判的出行模式转换点识别方法。

将一段GPS轨迹数据表示为: $G_0 = \{g | g = (Lat, Lon, H, date, time)\}$ 。其中 $g$ 代表某一个时刻的GPS数据,由5个维度数据组成,分别指某个时刻的纬度值、经度值、海拔高度、日期、时间;转换点集合表示为 $T = \{g | g \in G_0 \wedge g \text{ 是转换点}\}$ 。此外,设定一个速度阈值经验值 $V_0$ 和研判移动窗口的步长 $S$ ,其中, $V_0$ 指步行或者静止时人的最高移动速度, $S$ 决定取样的时间间隔。

首先将出行模式划分为步行模式与非步行模式两类。针对一段GPS轨迹数据中的出行模式转换点候选集的研判方法如下:

- 1) 设定研判窗口大小 $\Delta t$ ,设定研判移动窗口的步长 $S$ ;
- 2) 对于轨迹数据中某一时刻 $t$ 的GPS数据 $g$ ,依据研判窗口大小,完成 $g$ 前后 $\Delta t$ 时间内的GPS出行

轨迹的数据取样;

3) 分别研判取样窗口内轨迹数据的出行模式。如果两段样本中出行模式不同,则  $g$  计入转换点候选集;若果两段样本中出行模式相同, $g$  不计入转换点候选集;

4) 移动步长  $S$ ,选取  $t+S$  时刻的 GPS 数据  $g'$ ;

5) 转入 2),继续研判直到轨迹数据结束。

利用以上方法筛选出一段出行轨迹中的出行模式转换点备选集合。常规情况下,在较短的时间间隔内不会出现多次出行模式的转换。因此,设定筛选阈值  $ST$ ,剔除时间范围内欧氏距离较小、相似度较大的转换点,留下时间范围内唯一的转换点。依据测试数据集情况设定时间  $ST=20\text{ min}$ 。

---

基于低速度移动研判的转换点识别算法:Function TPRA( $G_0, T$ )

---

输入:一段 GPS 出行轨迹数据  $G_0$

输出:转换点集合  $T$

1: **Begin**

2: 设置初始变量:研判窗口大小  $\Delta t$ , 步长  $S$ , 速度阈值经验值  $V_0$ , 删选阈值  $ST$ , 欧式距离度量窗口变量  $\Delta t'$ ;

3: while  $i < |G_0|$  do

4: 在  $G_0$  中取出  $g_i$ ;

5: 取  $g_i$  前  $\Delta t$  区间内的 GPS 数据为  $data1$ ;

4: 取  $g_i$  后  $\Delta t$  区间内的 GPS 数据为  $data2$ ;

5: 分别研判  $data1, data2$  两个区间内 GPS 轨迹的出行模式  $Mode1, Mode2$ ;

6: If  $Mode1 \neq Mode2$

7:  $g_i$  计入转换点候选集  $T'$ ;

8: end if

9: end while

10: While  $i < |T'|$  do

11: 在  $T'$  中取出  $g_i$ ;

12:  $g_i$  前  $\Delta t'$  区间内的 GPS 数据为  $data1'$ ;

13: 取  $g_i$  后  $\Delta t'$  区间内的 GPS 数据为  $data2'$ ;

14: 计算  $data1'$  与  $data2'$  的欧式距离  $dis$ ;

15: 保存  $dis$  为  $g_i$  的欧式距离;

16: end while

17: while  $j < |T'|$  do

18: if  $g_{j+1}$  与  $g_j$  的时间距离小于  $ST$

19: if  $g_{j+1}$  的欧式距离小于  $g_j$  的欧式距离

20:  $T' = T' - \{g_{j+1}\}$ ;

21: else

22:  $T' = T' - \{g_j\}$ ;

23: end if

24: end if

25: end while

26:  $T = T'$

27: **End**

---

## 2 借助转换点识别的 GPS 轨迹出行模式生成方法

研判输出一段 GPS 轨迹的出行模式分 5 步进行:①将一段 GPS 数据进行预处理,如:滤波、插值等,剔除脏数据;②根据清洗后的数据计算相关特征;③通过上一章的转换点识别算法识别出 GPS 轨迹中的出行模式转换点,依据转换点将一段 GPS 数据分为不同子段;④利用随机森林分类器,研判不同子段的出行模式;⑤依照时间次序将模式标签连接,得到整段 GPS 数据携带的出行模式信息。方法的总体框架如图 1 所示。

## 2.1 数据预处理

### 2.1.1 高斯滤波

鉴于定位系统的准确性,所有的GPS定位数据都无法绝对精确地定位被采集位置,收集的数据会在一定范围内波动,产生不必要的噪声,因此需要将其进行平滑处理。参考黄仁等<sup>[11]</sup>、孙冰怡等<sup>[6]</sup>、Liang等<sup>[1]</sup>的工作,利用滤波方法将数据进行平滑处理。具体地,利用高斯滤波算法,滤除掉数据中的高斯白噪声。

$$g_i.Lat = (g_{i-1}.Lat + g_i.Lat + g_{i+1}.Lat) / 3. \quad (1)$$

其中: $g_i.Lat$ 为滤波后的第*i*点经纬度值, $g_{i-1}.Lat$ 为原GPS数据中第*i*-1点经纬度值, $g_i.Lat$ 为原GPS数据中第*i*点经纬度值, $g_{i+1}.Lat$ 为原GPS数据中第*i*+1点经纬度值。用一点及其附近点的平均值代替这一点的原值,能够降低数据的突变,使轨迹变得更平滑、连续,更符合实际情况的变化规律。

### 2.1.2 插值法

由于信号传输的不稳定或定位系统采集的问题,在某些时间点上的数据会产生缺失,而在一些信号弱或无信号的隧道等地区甚至会出现大面积的数据缺失,严重影响模式识别。为此,需要将这些数据进行修补,在一定程度上保持数据的合理性。

实验采用了一维线性插值法来补充缺失的数据,使数据频率更稳定。具体地,利用缺失点两侧的数据,将两侧数据的时间、经纬度取均值,按时间以匀速状态插入缺失点。

$$g_i = (g_{i-1} + g_{i+1}) / 2. \quad (2)$$

其中, $g_i$ 为缺失点GPS数据, $g_{i-1}$ 为缺失点上一时刻GPS数据, $g_{i+1}$ 为缺失点下一时刻GPS数据。对缺失点的值进行预测,数据标注为缺失点前后相应数据的均值。

## 2.2 特征的选择

实验选定中位速度、95%分位速度、平均速度、稳定速度、平均停留间隔作为出行模式识别算法的基本特征。特征的选择主要针对速度类进行提取,用95%分位速度代替最高速度,是因为最高速度有可能无法反映真实的速度,是突变的,波动性较强。中位速度、95%分位速度、平均速度的计算很多文献中都有,在此不赘述。

一段完整的速度数据,必然会有加速-匀速-减速的过程。在运动过程中,取均值会受到加减速时速度变化的影响,使得提取值不准确。而提取稳定速度,可以避免加减速对于提取值的影响,从而得到更为准确的提取值。实验采用以下方法提取一段GPS轨迹数据的稳定速度SV。首先,识别出一段完整的加速-匀速-减速的过程,作为运动趋势;在这段运动趋势中取匀速段的平均速度作为稳定速度。一段GPS轨迹数据 $G_0 = \{g | g = (Lat, Lon, H, date, time)\}$ ,其中 $g$ 代表某一个时刻的GPS数据。用符号 $P = (Lat, Lon, V, A, date, time)$ 表示GPS数据的特征数据,其中6个的维度数据分别指某个时刻的纬度值、经度值、速度、加速度、日期、时间。 $g$ 的特征数据表示为 $P_g$ , $G_0$ 的特征数据表示为 $P_{G_0}$ 。

此外,许多研究在识别中采用停驻点作为研判出行模式的基本特征。如Liang等<sup>[12]</sup>利用公交站停留点占总停留点的比率,提高了bus模式的识别准确率。但在实际中,受到出行段总时间、总距离的影响,停驻次数会产生较大的差异。出行时间长、距离较长的出行段,停驻次数更多。因此,结合出行距离与停驻次数,得到平均停留间隔,在一定程度上消除了出行时间、距离对出行模式研判的干扰影响,对模式识别有较大帮助。

$$Asd = D / SN. \quad (3)$$

其中: $Asd$ 为平均停留间隔, $D$ 为出行段的总路程, $SN$ 为出行段中停驻次数(采集点后10s速度小于0.3m/s时判定为停驻点)。平均停留间隔代表平均情况下两次停驻之间的距离。对于公交、地铁等停驻有规律的模式具有较好的识别效果。

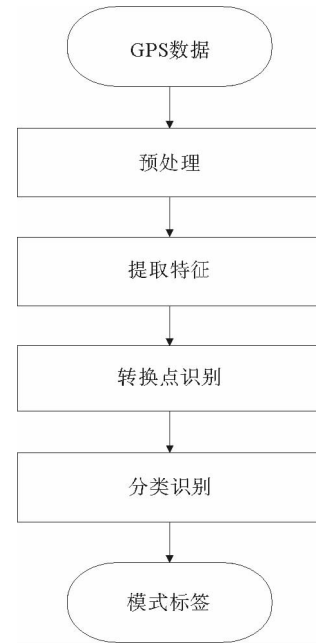


图1 出行模式识别框架图

Fig.1 Framework diagram of travel mode recognition

---

稳定速度提取算法:Function SVE( $P, SV$ )

---

输入:一段 GPS 轨迹的特征数据  $P_{G_0}$

输出:本段 GPS 轨迹的稳定速度特征  $SV$

```

1: Begin
2:   While  $i < |P_{G_0}|$  do
3:     取出  $P_{gi}$ ;
4:     if  $\text{abs}(\text{average}(P_{gi-5}.V, P_{gi-4}.V, \dots, P_{gi-1}.V) - P_{gi}.V) < 0.2 \text{ m/s}$ 
        than
5:        $m = i + 5$ ;
6:       while  $m < |P_{G_0}|$  do
7:         if  $\text{abs}(\text{average}(P_{gi+m+5}.V, P_{gi+m+4}.V, \dots, P_{gi+m+1}.V) - P_{gi+m}.V) < 0.2 \text{ m/s}$ 
            than
8:            $MV \leftarrow i \text{ 至 } i+m \text{ 的中位速度};$ 
9:           else  $m + 1$ ;
10:          end if
11:         end while
12:          $i = i + m$ ;
13:       else  $i = i + 1$ ;
14:       end if
15:        $SV \leftarrow \text{average}(MV)$ ;
        end while
16:   End

```

---

## 2.3 随机森林分类器

随机森林算法的本质是利用阈值进行的二分法,更适合解决分类问题<sup>[12-14]</sup>。因此,实验将一段 GPS 轨迹的出行模式识别问题转换为分类问题,采用随机森林算法完成出行模式识别。

随机森林是由决策树改进而来,由多棵决策树组成。对于单棵树,在训练模型时,通过随机选取数据和特征创建结点,结点值定为将数据最优分割情况时的界值,即分割后 GPS 数据的出行模式某一种标签占全部数据比例最高的情况。当标签种类唯一时,停止生长。用模型识别时,将待测数据与结点的值进行比较,相应落到某一个子结点中,重复比较特征。当数据掉落到某一个叶子结点中时,叶子结点的标签就表示为这段 GPS 轨迹数据的识别结果。将所有决策树的识别结果进行投票,得票最高的模式标签定为这片随机森林的识别结果。

## 2.4 算法实现过程

将一段 GPS 轨迹数据表示成: $G_0 = \{g | g = (Lat, Lon, H, date, time)\}$ ,  $G_0$  对应的出行模式序列表示为  $M = \langle M_0, M_1, \dots, M_k \rangle$ , 其中,  $M_0$  代表  $G_0$  中第一个子段的出行模式标签。

首先对收集的 GPS 轨迹数据进行预处理和特征计算,处理过程见 2.1 和 2.2 节;然后,将不符合一般规律的数据剔除,如步行速度大于 10 m/s 的;通过转换点识别将待测数据分段,并利用训练好的模型对数据进行模式识别,得到对应的出行模式标签;最后,将得到的模式标签组合形成预测结果。

## 3 实验结果与分析

### 3.1 测试条件与数据集

实验所采用的硬件条件是:主板,华硕 X455LJ;CPU,(英特尔)Intel(R) Core(TM) i5-5200U @ 2.20 GHz;内存,12.00 GB (1 600 MHz)。软件环境是 Microsoft Windows 10 企业版 (64 位)、Pycharm 集成开发环境。

由于研究者所采用的数据集的收集方式有所不同,因此各个数据集中的异常数据量会有很大不同,导致

识别的准确率不同,所以需要将差距较大的数据集分开对比。收集的数据集分为自主收集数据集和工程数据集。自主收集的数据一般是由手机 App 的内置功能实现,采集个体一段时间内的运动情况。工程数据集是由专业的某项工程收集测定,例如大多数研究<sup>[4,7,12,13,17]</sup>采用的工程—Geo Life project<sup>[15-16]</sup>。相对于 App 收集的数据,工程收集的数据更具有专业性、准确性、客观性。实验选用 Geo Life project 数据并进行了筛选,最终采用 982 条单出行模式个人数据、65 条多模式个人数据完成测试。

GPS 数据的交通出行模式序列识别算法:Function TME( $G_0, M$ )

输入:一段 GPS 出行轨迹数据  $G_0$

输出: $G_0$  对应的出行模式序列  $M$

```

1:  Begin
2:    While  $i < |G_0|$  do
3:       $g_i.Lat = (g_{i-1}.Lat + g_i.Lat + g_{i+1}.Lat) / 3;$ 
4:       $i = i + 1;$ 
5:    End while
6:    While  $i < |G_0|$  do
7:       $g_i = (g_{i-1} + g_{i+1}) / 2;$ 
8:       $i = i + 1;$ 
9:    End while
10:    $T = \text{Function TPRA}(G_0, T);$ 
11:   while  $j < |T|$  do
12:      $G_0' = \{g \mid g_j.time \leq g.time < g_{j+1}.time\};$ 
13:      $M_{G_0'} = \text{Random Forest}(G_0');$ 
14:   End while
15:   输出  $M;$ 
16:  End

```

### 3.2 出行模式识别测试对比结果

将“稳定速度”和“平均停留间隔”这两个新增特征加入,对以往基于随机森林的出行模式识别算法进行改进,并将改进后的算法与已有算法的识别准确率进行对比。

测试发现,在每次训练与识别时,算法的准确率略有不同。为降低测试结果的偶然性,重复进行 30 次测试。具体地,进行了 30 次模型训练,产生 30 片随机森林,取 30 片森林的平均准确率作为改进后算法最终的准确率,如表 1 所示。随机森林算法训练过程中,树的数量直接决定了模型的准确度,但当数量到达一定程度后,模型会产生过度拟合,可解释性减弱,导致准确率降低,并且会大大增加模型构建时间。经过多次测试,选定每片森林为 60 棵树。此外,实验计算了 30 次识别准确率的方差,对算法与已有算法的稳定性进行了分析,如表 2 所示。

表 1 添加新特征前/后的 30 次训练识别准确率表

Tab. 1 30 training recognition accuracy before and after adding new features

%

前/后	前/后	前/后	前/后	前/后	前/后
87.3/86.9	69.4/87.7	92.6/88.1	68.6/89.7	81.1/84.2	81.5/83.8
90.6/88.3	90.8/91.4	69.0/90.3	90.3/88.9	84.6/79.9	88.1/86.9
85.4/88.1	87.7/92.8	88.3/86.2	89.5/93.6	88.5/92.0	80.1/87.5
86.4/89.7	85.4/86.2	88.5/85.0	89.7/72.9	86.9/88.7	89.5/86.0
89.9/89.3	88.1/86.9	88.5/87.1	91.0/79.9	89.1/86.9	90.3/91.6

与已有算法比较,本算法的平均准确率上升了 1.3%,识别准确率波动情况明显改善。综合看来,增加稳定速度与平均停留间隔两个特征对提升出行模式识别算法的准确性有较好效果。

同时,出行模式识别中速度的贡献度较高(贡献度高的特征在识别中的重要性越高,识别时所用到的频率越高,添加后识别准确率也就越大),最大速度、平均速度、中位速度是出行模式中的最主要特征。而添加新的特征后,识别准确率有小幅上升,表明所添加的特征对于不同模式的区分度高于原有特征。

在 30 片森林中选取识别最优的森林模型用作测试出行模式识别混淆矩阵,如表 3 所示。混淆矩阵也称误差矩阵,用于比较分类结果和实际测得值,能够清晰地反映出识别值与真实值的关系。在识别最优的情况下,可以最大程度上降低自然误差,更能反映实际的数据与算法对于识别准确率的影响,便于对特征和算法的分析。

表 2 添加新特征前后准确率对比表

Tab. 2 Contrast of accuracy before and after adding new features

	添加新特征后	未添加新特征
准确率方差	0.001 695 5	0. 003 941 7
平均准确率/%	87.2	85.9

表 3 添加新特征后/前的出行模式识别混淆矩阵

Tab. 3 Travel mode recognition confusion matrix after/before adding new features

识别 实际	后/前					准确率/%	总数
	Walk	Bike	Bus	Subway	Car		
Walk	213/213	11/11	0/0	0/0	0/0	95.1/95.1	224
Bike	9/10	132/130	0/1	0/0	0/0	93.6/92.2	141
Bus	0/2	5/8	20/15	0/0	0/0	80.0/60.0	25
Subway	0/0	0/0	1/3	0/0	3/1	0/0	4
Car	0/0	0/0	2/0	0/0	91/93	97.8/100	93
总数	222/225	148/149	23/19	0/0	94/94		487

由表 3 可知,受到数据限制,在训练集中 bus 和 subway 模式较少,在训练过程中不能很好地拟合这两种模式的特征,从而 bus 和 subway 模式的识别率明显低于其他模式的识别准确率。在添加稳定速度与停留特征之后,bike 模式的识别准确率略有提升。bike 中误识为 walk 和 bus 模式均有减少,表明两种新特征有助于提高识别的稳定性。

在数据限制的情况下,bus 模式与 bike、subway、car 模式均有交织,在添加停留特征后,识别率明显上升。因此对于含有 bus 模式的识别,添加停留特征有利于提升识别准确率。

walk 与 bike 模式识别中,相互均有 5%~10%的误识率(walk 识别为 bike 或 bike 识别为 walk)。实验选取以速度特征为主时,walk 与 bike 模式在速度上有较大的重合,对于识别准确率产生较大影响。因此,对于需要明确区分 walk 与 bike 模式的场景,需要再添加可以明确区分两种模式的特征。

3.3 出行模式转换点测试对比结果

出行模式转换点识别算法利用了步行模式速度低的特性。设置了一个速度阈值经验值  $V_0$  和步长  $S$ 。正常人步行速度约为 1.5 m/s,其他模式的正常速度都大于这个数值,偶尔产生的低速情况持续时间也较短。研判窗口  $\Delta t=300$  s 能够排除偶然情况的干扰。将 300 s 窗口内的平均速度与速度阈值经验值  $V$  对比,能够有效地区分步行模式和其他模式。同时,步长过小会使得移动前后的窗口区别较小,产生无意义的的数据。在保证识别准确率的情况下,尽可能地增大步长能够有效降低识别时间。

对于转换点识别,由于单一的准确率无法正确描述识别准确率,给出以下准确率判别方法:

*precision*:查准率,代表识别出的真实转换点占识别全部转换点数的比例。

$$precision = \frac{N_t}{N_p} \quad (4)$$

$recall$ :查全率,代表识别出的真实转换点占实际全部转换点的比例。

$$recall = \frac{N_t}{N_r} \quad (5)$$

$F-score$ :准确率,代表查准率与查全率的均值,作为识别的准确率。

$$Fscore = \frac{(precision + recall)}{2} \quad (6)$$

其中: $N_t$  是识别正确的转换点个数, $N_p$  是识别的全部转换点个数, $N_r$  是实际出行中应有的转换点个数。

将  $F-score$  作为最终识别的准确率可以有效针对识别转换点冗余和漏识问题,较为准确地描述识别准确率。转换点识别准确率如表 4 所示。

由表 4 可知,所提出方法的查全率较高,而查准率较低。在识别正确个数一定的情况下,查准率较低而查全率较高,表明识别出的转换点数量比真实的全部转换点数量要多,即识别的转换点冗余较大。横向比较发现,由于本方法是基于移动窗口的识别方法改进的,在程序上还保留着移动窗口的框架,因此准确率与移动窗口识别类似,呈现查准率较低、查全率较高和转换点多冗余的特征。

实验随机选取一段 GPS 数据,将移动窗口方式与本方式进行对比。为避免偶然因素对实验结果的影响,对同一段数据进行 10 次识别,识别时间如图 2 所示。

表 4 转换点识别准确率对比

Tab. 4 Comparison of conversion point recognition accuracy %			
类型	本研究	多段窗口 <sup>[10]</sup>	移动窗口 <sup>[10]</sup>
$precision$	77.8	81.3	72.3
$recall$	87.7	76.5	81.5
$F-score$	82.8	78.8	76.7

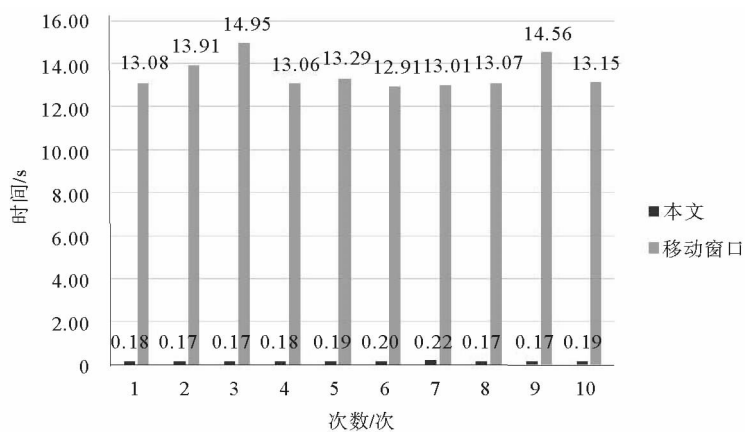


图 2 转换点识别时间对比图

Fig. 2 Comparison of conversion point recognition time

在移动窗口的研究中,需对数据进行遍历,计算特征的欧氏距离时间复杂度较高。已有工作<sup>[10]</sup>对数据集中每一点前后的欧氏距离做了计算,先提取了前后数据集的特征,然后通过特征计算得到前后的欧氏距离,提取欧氏距离的极值点,最后进行筛选。算法用移动步长代替遍历,并且只判断窗口内的平均速度,减少了计算量。判定为初步转换点的窗口后,才进行模式识别和欧氏距离计算。由实验结果可知,本算法的时间复杂度有明显降低。

### 3.4 实验结论

研究针对 GPS 数据进行挖掘,从一段 GPS 数据中研判出这段数据的出行模式及其转换点。通过“稳定

速度”和“平均停留间隔”两个特征的加入,交通出行模式识别的整体精度和 bus 模式识别准确度有较大提高,但对于 walk 和 bike 两种模式的识别效果改善不明显。此外,由于数据集限制,用于训练的数据较少,对于 bus 模式和 subway 模式的识别尚不能做出精确判断。

选用训练数据集中转换间隔较长的数据较多,使本算法对长间隔转换数据的识别精确度较高,对于长时间 GPS 数据的转换点识别有较好的应用价值。

## 4 结束语

本研究提出了一种基于低速度移动研判的出行模式转换点识别方法,增加了两种用于出行模式研判的新特征。通过步行模式的低速特征和转换点两端的欧氏距离判断出 GPS 轨迹数据中的出行模式转换点,并利用“稳定速度”和“平均停留间隔”提高了出行模式识别的精度。通过对比实验发现,本研究提出的方法能够较好地提高长时间 GPS 数据的模式识别的准确率和转换点识别的准确率、降低转换点识别的时间复杂度。但是,研究的转换点研判算法不适用于短时间内多次模式转换的情况,模式识别也仅局限在设定的 5 种模式中。下一步将针对短时间内多次转换的情况进行研究。

### 参考文献:

- [1] LIANG X, WANG G. A convolutional neural network for transportation mode detection based on smartphone platform[C] // IEEE International Conference on Mobile Ad Hoc and Sensor Systems. IEEE Computer Society, 2017: 338-342.
- [2] XIAO G, JUAN Z, ZHANG C. Travel mode detection based on GPS track data and Bayesian networks[J]. Computers Environment and Urban Systems, 2015, 54(1): 14-22.
- [3] MARTIN B D, ADDONA V, WOLFSON J, et al. Methods for real-time prediction of the mode of travel using smartphone-based GPS and accelerometer data[J]. Sensors, 2017, 17(9): 2058-2078.
- [4] ZHU X, LI J, LIU Z, et al. Learning transportation mode choice for context-aware services with directed-graph-guided fused lasso from GPS trajectory data[C] // IEEE International Conference on Web Services. IEEE Computer Society, 2017: 692-699.
- [5] GUVNRSAN M A, DUSUN B, CAN B, et al. A novel segment-based approach for improving classification performance of transport mode detection[J]. Sensors, 2018, 18(1): 87-105.
- [6] 孙冰怡, 吕巍, 李文洋. 基于智能手机传感器和 SC-HMM 算法的行为识别[J]. 吉林大学学报理学版, 2013, 51(6): 1128-1132.  
SUN Bingyi, LÜ Wei, LI Wenyang. Activity recognition based on smart phone sensors and SC-HMM algorithm[J]. Journal of Jilin University (Science Edition), 2013, 51(6): 1128-1132.
- [7] DABIRI S, HEASLIP K. Inferring transportation modes from GPS trajectories using a convolutional neural network[J]. Transportation Research Part C: Emerging Technologies, 2018, 86(1): 360-371.
- [8] WANG B, GAO L, JUAN Z. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(5): 1547-1558.
- [9] ZONG F, BAI Y, WANG X, et al. Identifying travel mode with GPS data using support vector machines and genetic algorithm[J]. Information, 2015, 6(2): 212-227.
- [10] 沈云. 基于 GPS 轨迹数据的交通出行方式识别研究[D]. 北京: 北京交通大学, 2017.  
SHEN Yun. Research on traffic travel mode identification based on GPS trajectory data[D]. Beijing: Beijing Jiaotong University, 2017.
- [11] 黄仁, 田丰, 田维兴. 基于加速度传感器的运动模式识别[J]. 计算机工程与应用, 2015, 51(6): 235-239.  
HUANG Ren, TIAN Feng, TIAN Weixing. Motion pattern recognition using acceleration transducer[J]. Computer Engineering and Applications, 2015, 51(6): 235-239.
- [12] LIANG J, ZHU Q, ZHU M, et al. An enhanced transportation mode detection method based on GPS data[J]. Communications in Computer and Information Science, 2017, 727: 605-620.
- [13] ZHU Q, ZHU M, LI M, et al. Transportation modes behaviour analysis based on raw GPS dataset[J]. International Journal of Embedded Systems, 2018, 10(2): 126-136.
- [14] 陈俭新, 赵红领, 李润知, 等. 基于移动互联网技术的出行模式识别方法[J]. 计算机工程与设计, 2015, 36(9): 2532-2538.

- CHEN Jianxin, ZHAO Hongling, LI Ruizhi, et al. Transportation mode recognition method based on mobile internet technology[J]. Computer Engineering and Design, 2015, 36(9): 2532-2538.
- [15] ZHENG Y, LI Q N, CHEN Y K, et al. Understanding mobility based on GPS data[J]// Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008). Seoul, Korea. ACM Press: 312-321.
- [16] ZHENG Y, ZHANG L Z, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories[C]// Proceedings of International conference on World Wild Web (WWW 2009). Madrid, Spain. ACM Press: 791-800.
- [17] ZHU Q, ZHU M, LI M, et al. Identifying transportation modes from raw GPS data[C]// International Conference of Young Computer Scientists. Springer Singapore, 2016.
- [18] ZHENG Y, WANG L, LIU L, et al. Learning transportation modes from raw GPS data: US, 0073202[P]. 2013-05-21.
- [19] 李锋, 潘敬奎. 基于三轴加速度传感器的人体运动识别[J]. 计算机研究与发展, 2016, 53(3): 621-631.
- LI Feng, PAN Jingkui. Human motion recognition based on triaxial accelerometer[J]. Journal of Computer Research and Development, 2016, 53(3): 621-631.
- [20] DAS R, WINTER S. Detecting urban transport modes using a hybrid knowledge driven framework from GPS trajectory [J]. ISPRS International Journal of Geo-Information, 2016, 5(11): 207-231.
- [21] 叶帅男. 基于深度卷积网络的交通方式识别[D]. 成都: 西南交通大学, 2018.
- YE Shuainan. Transportation mode detection based on deep convolutional neural network[D]. Chengdu: Southwest Jiaotong University, 2018.

## 符 号 说 明

符号名	符号
速度阈值	$V_0$
研判窗口	$\Delta T$
欧式距离度量窗口变量	$\Delta t'$
路程距离	$D$
移动步长	$S$
删选阈值	$ST$
轨迹数据	$G_0$
轨迹数据点	$g$
转换点集合	$T$
纬度	$Lat$
经度	$Lon$
海拔高度	$H$
日期	$date$
时间	$time$
GPS数据的特征数据	$P$
停驻次数	$SN$
稳定速度特征	$SV$
速度特征	$V$
加速度特征	$A$
平均速度	$MV$
平均停留间隔	$Asd$
出行模式标签序列	$M$
出行模式标签	$Mode$
查准率	$precision$
查全率	$recall$
准确率	$F-score$

(责任编辑: 傅 游)