康忆宁, 张登银. 基于 DAGSVM 的居民出行方式分类研究[J]. 智能计算机与应用,2025,15(3):24-32. DOI:10.20169/j.issn.2095-2163.250304

基于 DAGSVM 的居民出行方式分类研究

康忆宁, 张登银
(南京邮电大学 物联网学院, 南京 210003)

摘 要: 采用移动蜂窝信令数据识别居民出行方式对于规划交通方案、制定交通策略具有十分重要的意义,然而目前大多数研究方法未考虑不同输入组合特征对模型性能的影响,从而导致识别精度不佳。本文提出一种基于 DAGSVM(Directed Acyclic Graph Support Vector Machine)的居民出行方式识别方法。首先,采用基于网格的预处理算法对原始数据进行筛选,提取出信令数据的出行特征,并结合 K-means 聚类特征和隶属度函数特征获得信令用户出行的多维特征数据集;其次,基于上述特征建立用户轨迹的多维度特征集,利用 DAGSVM 模型评估不同场景下不同特征组合对出行模式识别的影响。仿真结果表明,在输入最优特征组合时,未区分时段的准确率为 88.4%,本文方法在高峰时段准确率约为 89.85%,在非高峰时段准确率约为 90.45%,识别精度得以提升。

关键词: 城市交通; 蜂窝信令数据; DAGSVM; 出行方式识别; 多分类器
中图分类号: U495 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0024-09

Research on resident travel mode classification based on DAGSVM

KANG Yining, ZHANG Dengyin
(College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The use of mobile cellular signaling data to identify resident travel modes is of great significance for planning traffic plans and formulating traffic strategies. However, most current research methods do not consider the impact of different input combination features on model performance, which limits recognition accuracy. This paper proposes a resident travel mode recognition method based on Direct Acyclic Graph Support Vector Machine (DAGSVM). Firstly, a grid based preprocessing algorithm is used to filter the original data, extract the travel features of the signaling data, and combine K-means clustering features and membership function features to obtain a multidimensional feature dataset of the signaling user's travel; Secondly, a multi-dimensional feature set of user trajectories is established based on the above features, and the DAGSVM model is used to evaluate the impact of different feature combinations on travel pattern recognition in different scenarios. The simulation results show that when the optimal feature combination is input, the accuracy of the undifferentiated period is 88.4%, and the accuracy of the proposed method is about 89.85% in peak period and 90.45% in off-peak period, which improves the recognition accuracy.

Key words: urban transportation; cellular signaling data; DAGSVM; travel mode identification; multiple classifiers

0 引 言

随着大数据的兴起,公交刷卡数据、电警卡口数据以及停车场流水等数据极大地拓展了交通数据的来源。但这些数据跟随性较差,仅能还原部分轨迹,无法构建用户完整的出行链,属于“跟车不跟人”的数据[1]。蜂窝信令数据则相反,具有很强的跟随性,不会随用户出行过程中交通方式的切换而改变。出行交通信息在交通规划与管理等方面有重要的价值[2]。基于手机信令数据挖掘用户交通方式具备获取成本低、抽样率高、全日数据完整度高、数据更新快等优点,是交通方式判别领域研究的热点[3]。识别用户出行方式,能够掌握居民出行规律、分析交通状况、缓解交通拥堵;并且可以针对车流量和人流量进行规划和调度,促进城市系统的健康发展[4]。

基金项目: 国家自然科学基金(61872423); 江苏省研究生科研创新计划(KYCX23_1054)。
作者简介: 康忆宁(1998—),女,硕士研究生,主要研究方向:信令大数据分析,智能交通。
通信作者: 张登银(1964—),男,博士,研究员,主要研究方向:信号与信息处理,网络技术与信息安全。Email:zhangdy@njupt.edu.cn。
收稿日期: 2023-09-09

早期,因技术受限,采集的信令数据精度较低,导致与交通方式识别相关的研究很少,且研究效果不显著。随着智能手机的普及,这方面的研究逐渐增多。例如,Qu 等学者[5]提出了一种融合 Logitech 模型和阈值规则的交通方式识别算法,通过将交通网络数据和移动信令数据相结合,对汽车、公交车和步行三种方式进行识别。Larijani 等学者[6]通过制定一系列阈值规则来识别用户的轨道交通信息,但识别结果尚不显著。

有学者提出,SVM 算法在参数调节和函数选择上过于敏感,过分依赖人为经验,存在一定的主观性,为了提高分类精度和效率,杨飞等学者[7]在支持向量机(Support Vector Machine,SVM)模型的基础上,加入了基于短时傅里叶变换(Short Time Fourier Transform,STFT)的频域属性,并利用遗传算法(Genetic Algorithm,GA)对 SVM 的惩罚系数和核参数进行联合优化,实验准确率高达 90%。另外,钟舒琦等学者[8]针对手机信令数据,设计了一套用户出行特征分析的框架,包括数据清洗、轨迹点分析、出行链提取、兴趣点分析与出行方式识别。基于兴趣点、路网数据和导航数据将用户的出行方式划分为 4 种不同出行模式,识别算法的准确率比仅使用导航数据的算法提高了 10%。王文静等学者[9]设计了一种基于移动端的个体出行链数据自采方案,同时采集个体的 GPS 数据、基站位置数据以及出行链记录数据。并应用主成分分析和决策树方法筛选出可用于交通方式识别的关键特征,如速度最大值、速度均值、加速度标准差、速度标准差以及方位角变化标准差,这些特征可以有效区分交通方式。但由于选取 GPS 数据作为数据源,会存在建筑物遮挡造成数据缺失的现象,最终导致实验结果不准确。

文献[10]通过提取用户的平均速度与加速度特征,结合信令用户的模糊统计特征,将这些多维特征输入 XGBoost 模型进行训练,从而分析出用户的出行方式。Tao 等学者[11]选择在识别过程中引入加速度变量,实验结果表明,识别精确度在 90%以上,进一步验证加速度数据在识别交通方式时有着良好优势。Bolbol 等学者[12]使用支持向量机分类方法进行用户出行方式分类研究,准确率为 88%。Sarker[13]利用决策树构建基于手机信令数据的出行方式分类模型。Bohte 等学者[14]利用速度统计特征量、加速度统计特征量、方向变化率、行程特征量来区分各出行模式。Siła-Nowicka 等学者[15]利用平均速度、平均距离、频率、性别、年龄等特征参数,结合交通路径的背景空间信息,区分出行模式。但由于在同一实验环境下缺少对比实验,无法有效凸显自身算法的优势。

以上各种解决方案并未考虑到以下 2 个方面:

(1)现有文献大多直接聚焦于对用户出行方式的分类,并未事先对用户完整出行轨迹进行刻画,造成链路轨迹的误差较大,影响后续研究。

(2)因交通流存在来回波动,需要评估用户出行的多维特征与分类精度之间的相互影响。

基于此,本文提出了一种 DAGSVM 模型来识别和分类居民出行方式。该模型选择线性核函数、多层感知机(Multiple Layer Perceptron,MLP)核函数和径向基(Radial Basis Function,RBF)核函数构造 3 种 DAGSVM,以测试模型的最佳分选结果。同时,选择了平均速度、加速度、隶属度函数以及聚类等 7 类特征作为输入特征集,在 SVM 模型的基础上结合 DAG 算法构建 DAGSVM 识别模型,并将本文模型与典型的机器学习方法进行性能对比,最后利用 DAGSVM 模型评估不同场景下不同特征组合对出行模式识别的影响。

1　出行特征提取

1.1　速度和加速度特征

本文使用 Haversine 公式对基站间的欧式距离进行计算,计算公式具体如下:

$$harversin(\theta)=\frac{1-\cos\theta}{2} \tag{1}$$

$$h=\cos(\alpha^i_{e\_lat})\times\cos(\alpha^i_{s\_lat})\times haversin(\frac{\alpha^i_{e\_lng}-\alpha^i_{s\_lng}}{2})+haversin(\alpha^i_{e\_lat}-\alpha^i_{s\_lat}) \tag{2}$$

$$l=2\times R\times\arcsin(\sqrt{h}) \tag{3}$$

其中,R 表示地球半径,值为 6 471 km;l 表示基站间的距离;$\alpha^i_{s\_lat}$、$\alpha^i_{e\_lat}$ 分别表示第 i 条数据开始基站和结束基站的纬度;$\alpha^i_{s\_lng}$、$\alpha^i_{e\_lng}$ 分别表示第 i 条数据开始基站和结束基站的经度。信令用户出行时长计算如下所示:

$$T=T^i_{e\_time}-T^i_{s\_time} \tag{4}$$

其中,$T^i_{s\_time}$、$T^i_{e\_time}$ 分别表示第 i 条数据中的开始时间和结束时间。信令用户的平均速度和加速度分别使用如下公式计算:

$$L=2\times l_i\times\Theta \tag{5}$$

$$V^i_v=\frac{L_i}{T_i} \tag{6}$$

$$A_a^i = \frac{|V_v^{i+1} - V_v^i|}{T_i} \tag{7}$$

其中，Θ 表示道路非直线系数，默认值设置为 1.3[16]；L_i 表示第 i 条数据中用户的出行距离；T_i 表示第 i 条数据中用户的出行时长；V_v^i 表示第 i 条数据用户的平均速度；A_a^i 表示第 i 条数据中用户的加速度。

1.2 时空特征

在工作日期间，居民的乘车行为通常具有明确的目的性，主要是通勤需求。乘车行为较固定，时空弹性较小，即乘车时间和乘车地点相对稳定。相反，在休息日期间，居民的出行行为更加随机和多样化，因为没有工作限制，使得居民出行具有较大随机性。

为减少数据噪音并提高研究的可信度，本文选取了工作日期间早晨 06:00 以后的乘车数据和居民轨迹为实验样本。之所以选择这段时间，是因为在凌晨 24:00 至次日 06:00 之间车流量较少，出行活动少，具有较低的研究价值。本文以小时为单位对工作日和休息日的居民出行时间进行划分，并通过乘车数据，分析得出工作日和休息日不同时间段的居民出行量特征曲线如图 1 所示，能够反映出不同时间段居民出行的高峰和低谷。

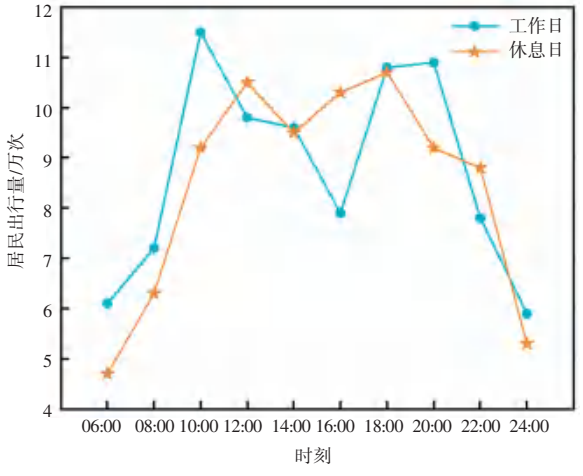


图 1 不同时间段乘客出行量统计

Fig. 1 Passenger travel volume statistics in different time periods

1.3 K-means 聚类特征

K-means 算法是一种基于距离的聚类算法。在本文中，针对提取的出行特征引入 K-means 聚类算法，以计算其聚类特征。由于数据集划分为 5 种不同出行方式，故聚类的簇数量 K 设定为 5；同时，选取平均速度和平均加速度作为 K-means 聚类算法的输入特征。图 2 展示了 K-means 聚类的结果图。图 2 中，标签 Label_1、Label_2、Label_3、Label_4、Label_5 分别代表步行、自行车、公交车、私家车和地铁五种不同的出行方式。通过 K-means 算法的聚类过程，可将样本数据划分到相应的簇中，从而对不同的出行方式进行分类和区分。

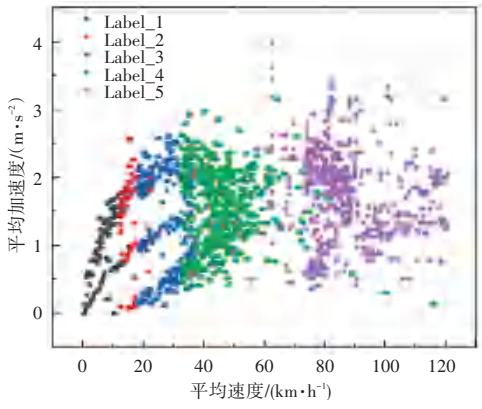


图 2 K-means 聚类结果图

Fig. 2 K-means clustering results diagram

1.4 隶属度函数特征

隶属度函数表达式如下所示：

$$g(x;\mu;z) = e^{\left(-\frac{(x-z)^2}{2\mu^2}\right)} \tag{8}$$

其中，μ 表示高斯函数形状参数，数据分布越分散、μ 越大，反之则越小；z 表示高斯函数位置参数，表示以 $x=z$ 为对称轴，左右完全对称；x 表示每条数据中平均速度或加速度绝对值。5 种不同出行方式的平均速度与加速度对应的高斯隶属度函数值如图 3、图 4 所示。通过隶属度函数，每段信令轨迹中的平均速度与加速度分别会获得一个隶属度值，通过下式来获取每段信令轨迹中平均速度与加速度所对应的联合隶属度值：

$$U_j = s_j \times \alpha_j \qquad (j = 1,2,3,4,5) \tag{9}$$

其中，U_j 表示联合隶属度值；s_j 表示速度对应隶属度值；α_j 表示加速度对应隶属度值。

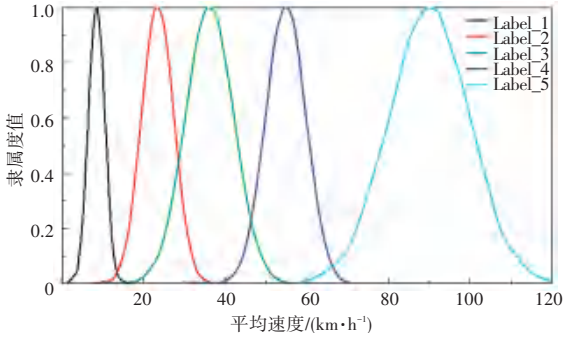


图 3 不同出行方式平均速度对应的隶属度函数值

Fig. 3 Membership function values corresponding to average speed of different travel modes

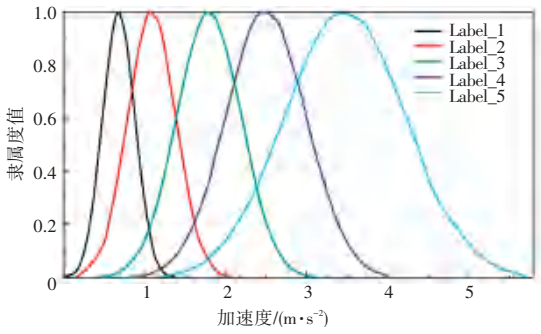


图 4　不同出行方式加速度对应的高斯隶属度函数值

Fig. 4　Gaussian membership function values corresponding to accelerations of different travel modes

因本研究中将居民出行方式分为 5 类,故信令用户的每段信令轨迹会得到 5 个不同的联合隶属度值,最后通过下式来获取最大联合隶属度值所对应的出行方式:

$$U_i = MAX(U_j) \quad (j = 1,2,3,4,5) \quad (10)$$

其中,1,2,3,4,5 分别代表步行、自行车、公交车、私家车和地铁。本文中每条信令轨迹数据都会得到 5 个不同的联合隶属度值。U_i 表示判定出的第 i 条数据所对应的出行方式的标签,当前获得的出行方式标签则是信令用户的模糊统计特征。

2　分类模型设计

2.1　数据预处理

由于采集到的原始蜂窝信令数据中存在大量如“漂移数据”、“乒乓数据”、缺失数据等一系列“脏”数据,需要对其进行预处理。为解决信令数据定位精度不高的问题,本文提出一种基于网格的预处理算法,具体处理流程如下。

(1) 网格化处理:首先对研究区域进行网格化处理,以基站为网格中心,基站覆盖范围为网格的边长。这样可以将得到的信令用户的运动轨迹映射至对应的网格中。

(2) 剔除 0 记录:在数据集中,剔除数值为 0 的记录。由于用户的轨迹序列是时间连续的,不能直接删除记录为 0 的点。本文解决方案是:判断其上一条时间戳记录和下一条时间戳记录是否相等。如果相等,则直接删除数值为 0 的点;如果不相等,则使用上一条记录和下一条记录的平均值来代替该点。

(3) 过滤漂移数据:采用过滤漂移数据算法对漂移数据进行处理。流程处理步骤见算法 1。通过设定时间窗口(timeWindow),将前后 2 个信令数据轨迹点进行连接,形成一个用于时间窗预测信令数据序列[17]。这样可对漂移数据进行过滤和处理。

(4)过滤乒乓切换数据:采用过滤乒乓切换数据算法对乒乓数据进行处理。流程处理步骤见算法 2。算法 2 的核心思想是将所有信令轨迹点映射至网格内,根据乒乓数据在相邻基站间快速跳变的特性设定时间阈值。如果网格内的轨迹数据在两基站之间快速跳变且跳变时间间隔小于时间阈值,则判定为乒乓数据并将其去除。

通过以上的预处理算法,可以有效地处理原始蜂窝信令中存在的问题,提高数据的质量和可靠性。

算法 1 和 2 中网格大小 n 和时间阈值 T 根据文献[18]设定如下,网格大小 n 设为 500 m,时间阈值 T 为 5 min。算法中各项参数说明见表 1。

表 1　算法 1 和算法 2 中各项参数说明

Table 1　Parameters description of Algorithm 1 and Algorithm 2

参数	含义
$pData[i][lng]$	预处理前数据集中第 i 条数据的经度
$pData[i][lat]$	预处理前数据集中第 i 条数据的纬度
$pData[i][time]$	预处理前数据集中第 i 条数据的时间
$Data[i][time]$	预处理后数据集中第 i 条数据的时间

算法 1　过滤漂移数据算法

输入　预处理前的手机信令数据集 $pData$, 网格中心经度 Lng, 网格中心纬度 Lat, 时间阈值 T, 网格长度 n

输出　预处理后的手机信令数据集 $Data$

```
for i = 1 to n do
    //判断轨迹点是否在网格内
    if InGrid(pData[i],n,Lng,Lat)
    then
        Pass
    else
        j ← i + 1
        timeWindow ← pData[i + 1][time] - pData[i - 1][time]
        //判断时间窗口中是否存在数在网格内
        while pData[j][time] ≤ pData[i][time] + timeWindow do
            if InGrid(Data[i],n,Lng,Lat) then
            //存在则记录该数据
            Data.append(pData[i])
            Data.pop(i, j) //过滤漂移数据
            i ← j
            break
            else
```

17. *Data*. *pop*(*i*)
18. *pData*. *pop*(*i*)
19. *j* ← *i*
20. *i* = *i* − 1
21. end if
22. end while
23. end if
24. end for
25. return *Data*

算法 2 过滤乒乓切换数据算法

输入 预处理前的手机信令数据集 *pData*，网格中心经度 *Lng*，网格中心纬度 *Lat*，时间阈值 *T*，网格长度 *n*

输出 预处理后的手机信令数据集 *Data*

1. for *i* = 1 to *n* do
2. //判断轨迹点是否在网格内
3. if *InGrid*(*pData*[*i*], *n*, *Lng*, *Lat*) then
4. *index* ← *Data*. *indexOf*(*pData*[*i*])
5. //判断网格内数据是否大于时间阈值
6. if *pData*[*i*][*time*] − *Data*[*index*][*time*] > *T* then
7. //存在则记录该数据
8. *Data*. *append*(*pData*[*i*])
9. else
10. *pData*. pop(*i*) //小于时间阈值的判定为乒乓切换数据
11. end if
12. end if
13. end for
14. return *Data*

经过数据预处理后，可得到某一用户出行轨迹对比图。为了更直观地观察到用户的出行轨迹，通过将用户的轨迹点映射至三维空间中展示，具体如图 5 所示。

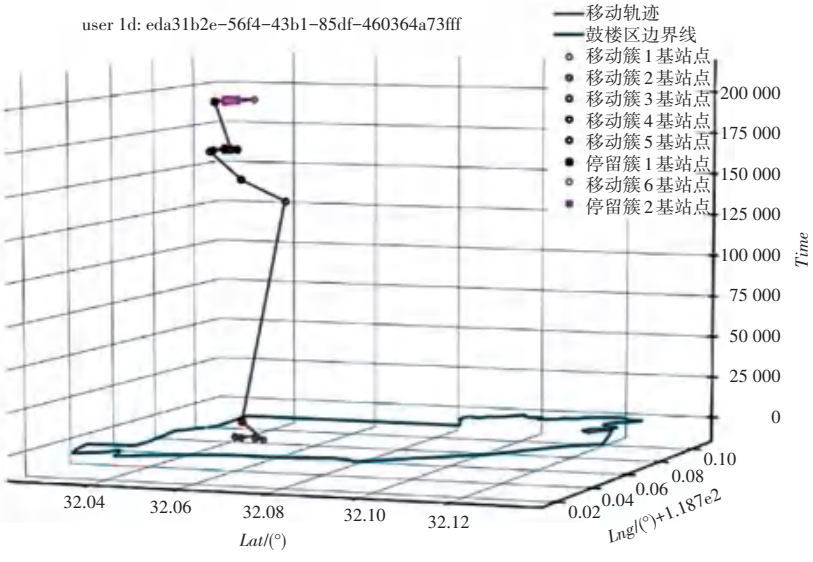


图 5 用户时空轨迹可视化

Fig. 5 Visualization of user space-time trajectory

2.2 DAGSVM 模型优势

SVM 是一种非线性分类和回归分析的方法，其基本原理是通过将数据映射至高维特征空间，并构建一个最优的超平面来实现数据的有效分类。相比较传统的线性分类方法，SVM 具有极高的分类准确性和强大的泛化能力，能够处理高维数据和非线性关系，并对参数变化和异常值具有一定的鲁棒性。这些优点使得 SVM 成为居民出行方式多分类问题的强有力工具，对于解决本文居民出行方式多分类问题有重要价值。

有向无环图(Directed Acyclic Graph，DAG)由于分类精度高的缘故，常用于处理分类问题并提高其准确率，研究分类过程如图 6 所示。

DAGSVM 是一种以 DAG 作为拓扑结构，SVM 作为分类器组合而成的多类分类器，通过层次结构实现了分类的快速化；通过从全体类别集合中不断删除不可能的类别，从而得到最终结果，这种独特的排除方式保证了分类的准确率。此外，DAGSVM 还具有不易发生因训练样本不均衡、进而对模型准确率造成较大影响的优势，本文采用 DAGSVM 分类模型对居民出行方式分类问题进行处理。图 7 是本文信令用户出行模式识别方法框架图。

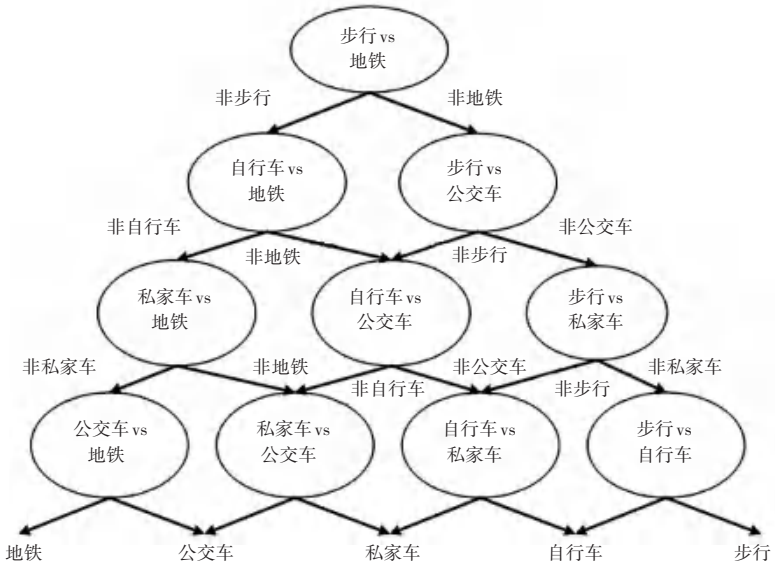


图 6　DAG 多分类算法

Fig. 6　DAG multi-classification algorithm

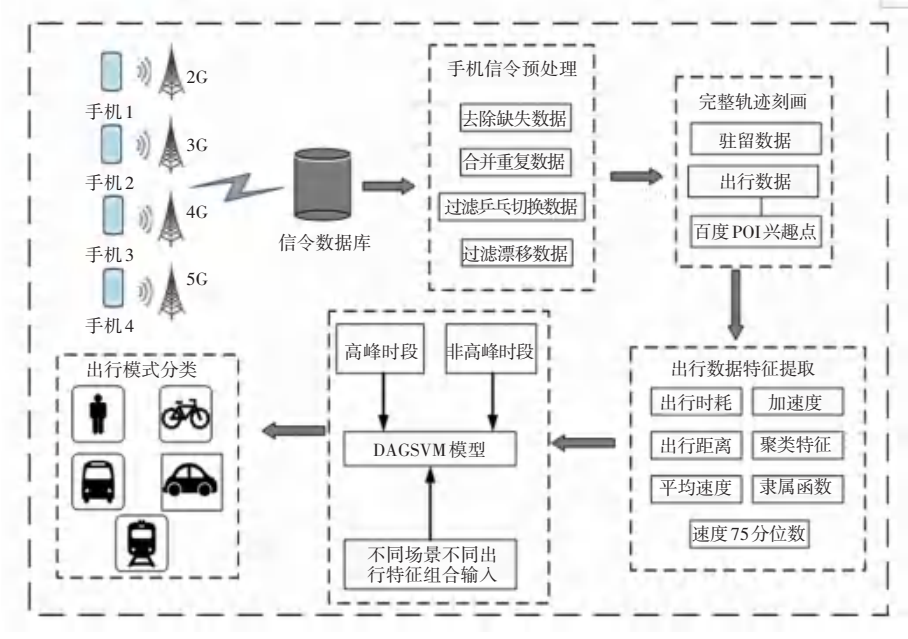


图 7　出行方式识别模型框架图

Fig. 7　Framework diagram of travel mode identification model

2.3　基于 DAGSVM 的出行方式分类

为提高分类精度,在采用 DAGSVM 模型识别出行方式前,对分类器进行训练。首先采用不同交通方式的出行特征作为分类器的训练样本集,即将不同交通速度、加速度、出行时长、K-menas 聚类特征及隶属度函数特征作为输入样本集,5 种交通方式为训练目标,对 DAGSVM 模型进行训练。出行模式对应的出行特征落入其中某个类别对应的特征空间范围,就可以确定此交通方式所属的类别。

DAGSVM 出行方式分类具体实现过程如下:

步骤 1　对采集数据进行预处理。先将所有数据进行归一化处理,以便训练与计算,采用公式 $x_i' = x_i/x_{i\max}$, 其中 x_i 为特征参数, $x_{i\max}$ 为对应特征参数最大值。

步骤 2　构造 $K(K-1)/2$ 个分类器,这里 K 为分选类别数,本文的 K 为 5。第 i 类分类器中,属于该类的样本输出为 1,否则为-1。

步骤 3　将本文多分类问题引入更高维度的空间进行处理,采用径向基函数作为转化函数,最终样本的输出响应区间由 σ 决定,这里 σ 为核函数宽度。σ 越小,响应区间越窄,但结构风险越大; σ 越大,响应区间越宽,函数曲线越光顺,但是经验风险

就越大。选择合适的宽度参数需要在这两者之间进行权衡，σ 值可以在训练过程中根据样本确定。

步骤 4 训练阶段，使用顺序最小优化算法对问题进行求解。

步骤 5 测试阶段，多分类器模型的构造选用 DAGSVM 方法。

步骤 6 最后，将待测试样本数据输入到由步骤 5 得到的出行方式分类模型，模型的输出结果即为出行方式类别。

3 实验结果分析

3.1 实验配置与数据集

本文实验所用的操作系统为 Windows10，内存容量为 16 G，CPU 为 Intel(R) Core(TM) i5-8250U CPU @1.60 GHz。使用 Python 语言并基于 Pycharm 编辑器进行编码实验，调用的 API 工具库有 pandas、numpy、matplotlib、scikit-learn。

本文所使用的移动蜂窝信令数据来源于南京某电信运营商，选取某天在固定区域内的数据进行研究，数据总量 672 367 条，涉及 153 611 位用户和 5 109 座基站。通过对 14 000 条居民轨迹信息的分段分析和参数的提取，共获得出行方式数据 5 580 个。

3.2 性能对比分析

为验证不同核函数在 SVM 中的有效性，本文通过使用不同核函数对实验结果进行对比说明，过程如下。

(1) DAGSVM_线性核函数：使用线性核函数构造 DAGSVM 模型验证信令用户出行方式分选结果。

(2) DAGSVM_RBF：使用 RBF 径向基核函数构造 DAGSVM 模型验证信令用户出行方式分选结果。

(3) DAGSVM_MLP：使用 MLP 多层感知机核函数构造 DAGSVM 模型验证信令用户出行方式分选结果。

实验结果见表 2。使用线性核函数、MLP 核函数和 RBF 核函数构造 3 种 DAGSVM 模型会产生不同的分选结果，在训练样本和测试样本中的分类精确度也不同。

表 2 不同核函数分类效果

Table 2 Classification effects of different kernel functions %

核函数	训练样本分类精度	测试样本分类精度
DAGSVM_线性核函数	78.24	74.10
DAGSVM_RBF	84.75	80.92
DAGSVM_MLP	80.45	78.65

在训练样本中，当使用 DAGSVM_线性核函数模型时，得到的分选准确率仅有 78.24%；当选择 DAGSVM_MLP 模型时，分选准确率相比线性核函数提高 2.21%；当选择 DAGSVM_RBF 模型时，其分选结果相比 DAGSVM_MLP 核函数提高 4.30%。在测试样本中，当使用 DAGSVM_线性核函数模型时，分选准确率仅有 74.10%；当选择 DAGSVM_MLP 模型时，分选准确率相比线性核函数提高 4.55%；当选择 DAGSVM_RBF 模型时，其分选结果相比 MLP 核函数提高 2.27%。

2 组实验结果说明，使用 DAGSVM_RBF 核函数构造的模型效果达到最优，分类结果最为精准。为提高后续实验精度，本文选取 RBF 径向基核函数构造 DAGSVM 模型来对居民出行方式进行分类研究。

为验证所提取特征的有效性，将用户出行特征分别输入至 DAGSVM 模型中进行特征有效性验证。图 8 为不同特征输入至 DAGSVM 模型时各出行模式的准确率。结果表明，聚类特征、隶属度函数特征等 7 类均为有效特征。其中，隶属度函数特征对于步行和自行车的出行模式准确率较高，这是因为步行和自行车的平均速度和加速度更符合高斯分布，故本文选取的是高斯隶属度函数。

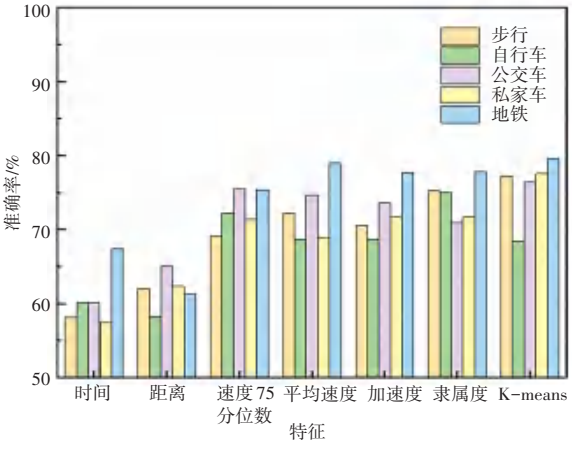


图 8 不同特征输入 DAGSVM 模型的准确率

Fig. 8 Accuracy of DAGSVM model input with different features

此外，本文选取了准确率、平均精确率、平均召回率和 $F1-Score$ 四个指标对不同机器学习算法性能进行评价[19]。计算公式具体如下：

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (14)$$

其中, *TP* 表示真实结果和预测结果均为正的样本数; *FP* 表示真实结果为负、预测结果为正的样本数; *TN* 表示真实结果和预测结果均为负的样本数; *FN* 表示真实结果为正、预测结果为负的样本数; *accuracy* 表示准确率; *precision* 表示精确率; *recall* 表示召回率; *F*1 - *Score* 表示精准率与召回率的调和平均值。

为充分验证本文模型的有效性,研究中将所提取的平均速度、加速度、隶属度函数、聚类等 7 类特征作为输入特征集,利用 DAGSVM 模型搭建出行模式识别模型,同时采用准确率、平均准确率、平均召回率和 *F*1 - *Score* 四个指标与传统的机器学习模型进行性能对比实验。图 9 为不同机器学习模型在出行方式识别的性能结果。结果表明,利用 DAGSVM 搭建的出行模式识别模型识别效果最佳,准确率约为 90.2%,优于传统的 KNN、DT、RF、BP 和 SVM 模型。

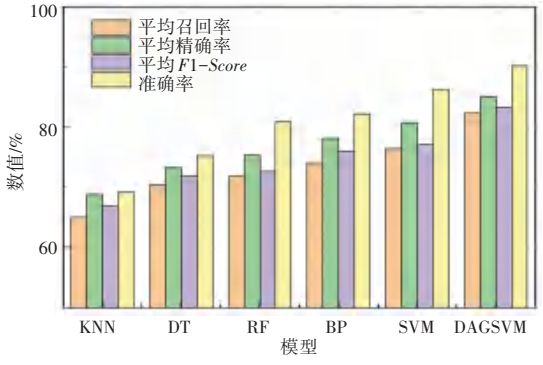


图 9　不同机器学习模型在信令数据中性能比较

Fig. 9　Performance comparison of different machine learning models in signaling data

道路交通状况因易受各种外在因素影响从而发生变化,会在一定程度上影响到对出行方式的选取,本文考虑将提取的多维出行特征进行特征融合处理。图 10 表示未区分高峰非高峰时段下不同特征数量的最优特征组合,当组合特征选取平均速度、速度 75 分位数、加速度、隶属度函数特征和聚类特征五个输入特征时,模型识别精确率最高,为 88.4%。

为了验证区分高峰时段(早上 6:30-9:30 与下午 17:30-19:30)和非高峰时段对预测结果的影响,本文选取在未区分时段下的最优特征组合(即平均速度、速度 75 分位数、加速度、隶属度函数特征和聚类特征)作为模型输入,利用 DAGSVM 模型与其他机器学习模型进行对比实验,实验结果如图 11 所示。结果表明,本模型在高峰时段和非高峰时段的分类效果最优,准确率高达 89.85%和 90.45%,分别比未区分高峰非高峰时段的准确率提高了 1.45%和 2.05%。

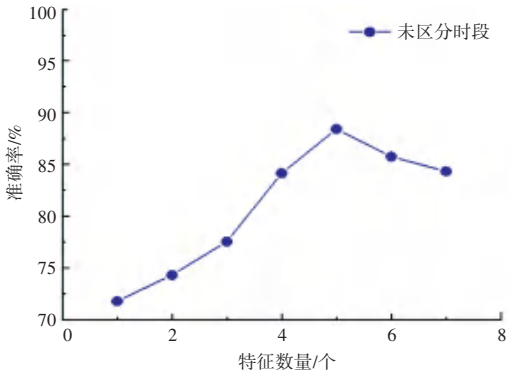
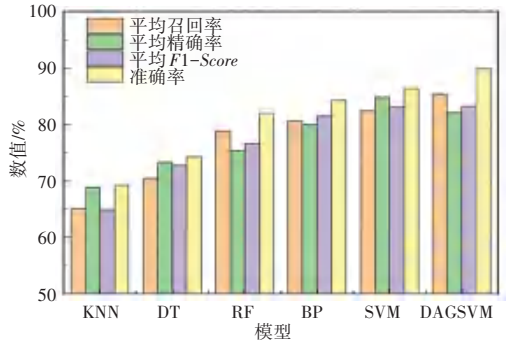
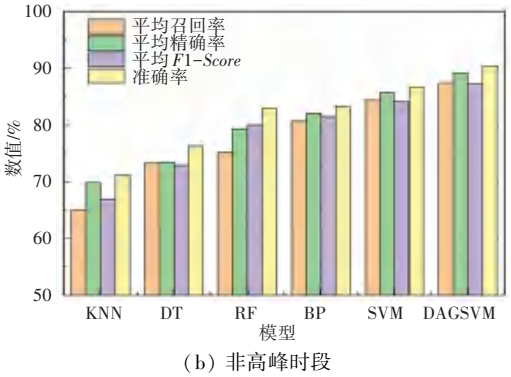


图 10　不同特征数量的最优特征组合结果

Fig. 10　The results of optimal feature combination with different number of features



(a) 高峰时段



(b) 非高峰时段

图 11　不同时段下的出行方式识别性能结果图

Fig. 11　Travel mode recognition performance results in different time periods

为了更直观地表达本文出行模式识别的结果,选取 5 月 25 日南京市北京西路附近一名 ID 为 a5ec39da-6bc3-524c-70d5-c403b261841 的信令用户部分出行轨迹数据进行出行模式识别。通过调用高德地图 API 进行数据可视化,结果如图 12 所示,该用户先后采用步行、自行车和公交车的出行模式。

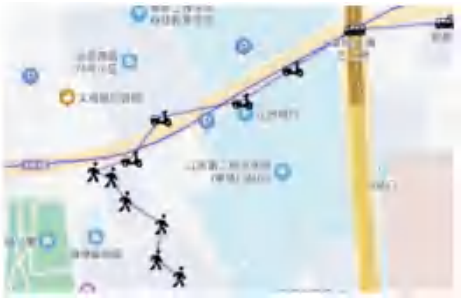


图 12 某手机信令用户出行模式识别结果图

Fig. 12 User travel pattern recognition results of a mobile signaling

4 结束语

本文以南京市内居民出行为例,提出了基于 DAGSVM 的用户出行模式识别方法。首先,采用基于阈值的预处理算法对原始信令数据集中的“漂移数据”和“乒乓数据”进行去除,并在此基础上对用户轨迹进行刻画;其次,引入手机信令数据的隶属度函数特征、聚类特征等多维时空轨迹特征,并对这些特征进行特征有效性验证;最后,利用 DAGSVM 搭建基于手机信令数据的出行模式识别模型,通过与传统的机器学习模型进行性能比较来验证模型的有效性。仿真结果表明,利用 DAGSVM 搭建的出行模式识别模型效果最佳,准确率约为 90.2%,优于传统的 KNN、DT、RF、BP 和 SVM 模型。

此外,还通过评估不同特征的最优特征组合对出行模式识别精度的影响,在未区分高峰非高峰时段找出最优特征组合,并在该特征组合基础上验证高峰时段和非高峰时段的模型性能。利用 DAGSVM 模型分别对 3 种场景搭建出行模式识别模型,并与传统的机器学习模型进行对比实验。仿真结果表明,本模型在高峰时段和非高峰时段的分类效果最优,准确率分别为 89.85% 和 90.45%,分别比未区分高峰非高峰时段的准确率提高了 1.45% 和 2.05%。

经过以上研究,本文在居民出行方式的识别和分类方面取得了令人满意的结果。但选取的输入变量有待丰富,下面研究将进一步丰富信令用户的出行特征,并加强与信令用户的出行轨迹的结合,从而使结果更加精确,能够为交通规划和出行方式优化等领域提供重要的参考价值。

参考文献

[1] 段征宇,赵浩然. 基于手机信令数据的居民活动空间及影响因素分析[J]. 交通与运输, 2023, 39(2):10-14.

[2] 师展,安艾芝,樊重俊,等. 基于 GA-DBN 模型的分类方法研究[J]. 智能计算机与应用,2023,13(5):23-31.

[3] 姚雅慧,张戎. 基于轨迹数据的货车停车目的识别方法[J]. 交通运输系统工程与信息, 2023,23(2):1-10.

[4] 廖阳,文义凡,李迎峰. 车联网环境下司机路径选择行为研究[J]. 智能计算机与应用,2023,13(3):58-63.

[5] QU Yingchun, GONG Hang, WANG Pu. Transportation mode split with mobile phone data[C]//Proceeding of the IEEE 18th International Conference on Intelligent Transportation Systems (ITSC 2022). Piscataway, NJ:IEEE, 2022:285-289.

[6] LARIJANI A N, OLTEANU-RAIMOND A M, PERRET J, et al. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region[J]. Transportation Research Procedia, 2015,6(3):64-78.

[7] 杨飞,姜海航,刘好德,等. 基于 GPS 轨迹数据的不同交通状态下交通方式识别流程优化方法[J]. 交通运输系统工程与信息, 2020, 20(4):83-89.

[8] 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):87-96.

[9] 王文静,陈艳艳,刘冬梅,等. 移动端个体出行链数据自采设计及出行特征选择[J]. 城市交通,2020, 18(6):110-117.

[10] YAO L, BAO J, DING F, et al. Research ontraffic flow forecast based on cellular signaling data[C]//Proceedings of the 5th IEEE International Conference on Smart Internet of Things (IEEE SmartIoT). Piscataway, NJ: IEEE, 2021: 1-7.

[11] TAO Feng, TIMMERMANS H J P. Transportation mode recognition using GPS and accelerometer data[J]. Transportation Research Part C Emerging Technologies, 2022, 37(3): 118-130.

[12] BOLBOL A, CHENG Tao, TSAPAKIS I, et al. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification[J]. Computers, Environment and Urban Systems, 2022, 36(6): 526-537.

[13] SARKER I H. A machine learning based robust prediction model for real-life mobile phone data[J]. Internet of Things, 2019, 5: 180-193.

[14] BOHTE W, MAAT K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands[J]. Transportation Research Part C: Emerging Technologies, 2019, 17(3): 285-297.

[15] SIŁA-NOWICKA K, VANDROL J, OSHAN T, et al. Analysis of human mobility patterns from GPS trajectories and contextual information[J]. International Journal of Geographical Information Science, 2020, 30(5): 881-906.

[16] LIU Huabin, SHAO Chunfu. Recognition method of urban residents′ travel mode based on GPS data[C]//Proceedings of the 3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT2019). Dalian, China: Atlantis Press, 2019: 101-105.

[17] DABIRI S, MARKOVIĆ N, HEASLIP K, et al. A deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data[J]. Transportation Research Part C: Emerging Technologies, 2020, 116: 102644.

[18] 陈略,熊宸,蔡铭. 手机信令的时空密度轨迹点识别算法[J]. 计算机工程, 2021,47(3):83-93.

[19] GUO Maozu, LIANG Shutong, ZHAO Lingling, et al. Transportation mode recognition with deep forest based on GPS data[J]. IEEE Access, 2020, 8: 150891-150901.