引文格式:柯日宏,肖亮,赵大郅,等.融合规则方法和机器学习的居民出行目的的识别方法[J].武汉大学学报(信息科学版),2026,51(6):1273-1282.DOI:10.13203/j.whugis20240329

Citation: KE Rihong, XIAO Liang, ZHAO Dazhi, et al. An Integrated Rule-Based and Machine Learning Method for Trip Purpose Identification[J]. Geomatics and Information Science of Wuhan University, 2026, 51(6): 1273-1282. DOI: 10.13203/j.whugis20240329

融合规则方法和机器学习的居民出行目的识别方法

柯日宏[1] 肖 亮[1] 赵大郅[1] 李胜楠[1] 王 璞[1]

1 中南大学交通运输工程学院,湖南 长沙,410075

摘 要:针对手机定位数据缺乏语义信息的问题,探索手机用户出行目的识别方法,有助突破手机定位数据应用的局限性。当前出行目的识别的研究中普遍存在数据量不足、数据类型单一和非通勤出行目的识别效果不理想等问题。基于大规模、长周期的手机定位数据和多源时空数据,提出一种融合规则方法和机器学习的居民出行目的识别框架。首先,基于规则方法提取用户职住信息,进而识别出通勤出行目的;然后,使用兴趣面数据识别部分非通勤出行目的;最后,构建*K*-Prototypes-XGBoost模型识别剩余非通勤出行目的。模型综合建成环境、用户属性、活动属性和用户历史活动等多源时空数据,构建特征变量,能够有效地捕捉个体出行习惯与行为模式。研究结果表明,所提方法能够较为准确地识别手机用户出行目的;与5种基准方法进行对比,*K*-Prototypes-XGBoost模型预测非通勤出行目的效果最好。利用*K*-Prototypes模型提取相似出行行为的人群再进行预测,能够进一步提高模型性能和可解释性。

关键词:手机定位数据;出行目的;出行链;建成环境;机器学习

中图分类号:TP208 **文献标识码**:A **收稿日期**:2025-04-05

DOI:10.13203/j.whugis20240329 **文章编号**:1671-8860(2026)06-1273-10

An Integrated Rule-Based and Machine Learning Method for Trip Purpose Identification

KE Rihong[1] *XIAO Liang*[1] *ZHAO Dazhi*[1] *LI Shengnan*[1] *WANG Pu*[1]

1 School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China

Abstract: **Objectives**: Aiming at the lack of semantic information in mobile phone data, the method for identifying the trip purposes of mobile phone users is explored to break through the limitations of mobile phone data application. **Methods**: This paper proposes a framework for identifying the trip purposes of residents by combining a rule-based approach and machine learning utilizing large-scale and long-term mobile phone data and multi-source spatiotemporal data. The home and work locations of users are first analyzed using the rule-based approach to identify commuting trips. The area of interest data are then used to identify the purposes of some non-commuting trips. Finally, The *K*-Prototypes-XGBoost model is used to identify the purposes of other non-commuting trips. The proposed model constructs feature vectors based on multi-source spatiotemporal data, including built environment, user attributes, activity attributes and historical activity of user, to capture the travel behavior habits and patterns of individuals. **Results**: Comparative experiments on known trip-purpose activity data from May 1st to 31st, 2023 show that, without introducing spatiotemporal features of historical activity of user and clustering information, extreme gradient boosting (XGBoost) outperforms adaptive boosting, random forest, gradient boosting decision tree, and artificial neural network, achieving accuracy of 73.2%, precision of 74.3%, recall of 73.2%, and *F*1-score

基金项目:湖南省教育厅科学研究重点项目(24A0016);中南大学研究生科研创新项目(1053320252421)。

第一作者:柯日宏,博士生,主要研究方向为交通大数据挖掘与可视化。kerihong@csu.edu.cn

通信作者:王璞,博士,教授。wangpu@csu.edu.cn

of 74.3%. After introducing spatiotemporal features of historical activity of user, the prediction performance of all five models is significantly improved, with the average accuracy, precision, recall, and $F1$-score increasing by 24.2%, 21.2%, 24.1%, and 23.0%, respectively. On this basis, the proposed K-Prototypes-XGBoost model further improves the prediction performance by incorporating user clustering information. Compared with the conventional XGBoost model, its accuracy, precision, recall, and $F1$-score increase by 2.7%, 1.5%, 2.7%, and 2.8%, respectively, and all four evaluation metrics reach 91.5%, indicating the best overall performance in identifying non-commuting trip purposes from large-scale, long-term mobile phone data. The confusion matrix further reveals that the proposed model performs best in identifying medical activities, with an accuracy of 94.7%, whereas dining activities exhibit the lowest recognition accuracy of 88.4%. The main misclassifications occur among leisure and entertainment, dining, and shopping activities, because these activities often share highly overlapping spatiotemporal characteristics and are frequently located within the same commercial complexes or adjacent areas. **Conclusions:** By integrating rule-based inference with the K-Prototypes-XGBoost model, this paper provides an effective solution for recovering the semantic information of trip purposes from mobile phone data. The proposed framework enables a more complete reconstruction of trip purpose chains and promotes the transformation of mobile phone data applications from mobility pattern description to semantic understanding of travel behavior, thereby supporting finer-grained urban transport planning and human activity analysis.

Key words: mobile phone data; travel purpose; travel chain; built environment; machine learning

探索城市居民出行目的需求的空间分布规律,并研究其与人口结构[1]、土地利用[2]和交通设施分布[3]的关系,是洞察人地交互规律、解构城市复杂系统的基础[4]。手机定位数据具有样本量大、精度高、时效性强和采集成本低等优点,弥补了传统问卷调查和出行日志等数据的不足,为居民出行行为研究提供了新的视角。但是,手机定位数据因缺少语义信息,如出行链、出行目的、活动类型、时空特征等,不能直观反映居民出行目的和活动规律[5]。因此,通过轨迹数据处理技术和出行目的识别后的手机定位数据能够反映居民出行行为和活动规律,为优化城市资源配置和交通管理决策提供参考依据[6]。

目前出行目的识别方法主要分为3类:基于规则的方法、基于概率的方法和机器学习方法。基于规则的方法是基于特定环境和已有数据设置不同的规则来推测出行目的[7-9]。文献[7]基于荷兰1 104名志愿者的出行数据,使用兴趣点(point of interest,POI)和土地使用数据构建规则来推测7种类别的出行目的;文献[10]基于加拿大安大略省23 527人的出行调查数据,使用规则模型与多项Logit模型相结合的方法预测居家、工作和其他3种出行目的,预测精度分别为90%、96%和91%;文献[5]提出职住距离的近邻匹配方法来识别居家/回家和上学/上班活动,准确率分别达到100%和89.8%。由于居民出行模式在不同环境中有所不同,需要花费大量的精力重新校准并设定阈值来制定新的规则。基于概率的方法比基于规则的方法更加灵活,因为其同时考虑到目的之间时空关联,对研究人员的主观知识依赖性小[11]。文献[12-13]提出兼顾时间和空间的规则,并集成到统计算法中,然后根据行程终点与POI集群的距离估计每个POI的概率来判断出行目的;文献[14]引入了一个多项逻辑回归模型,分析纽约高密度地区的49名受访者出行数据,将出行目的分为基于居住地活动和非居住地的活动,最后分类准确率分别为67%和78%;文献[15]通过时间和空间规则筛选出用户停留点附近候选的出行目的,然后使用重力模型计算概率来确定用户最终的出行目的,识别准确率为43%。随着数据可访问性和计算能力的提升,基于机器学习的方法逐渐成为主流,如决策树[16-17]、随机森林(random forest,RF))[18]、极端梯度提升(extreme gradient boosting,XGBoost)[5,19]、神经网络[20-21]等。文献[16-17]使用决策树模型预测出行目的,但是准确率都无法超过80%。相比之下,文献[18]使用RF模型预测156人一周的出行目的,准确率高达84.4%。后来许多研究也证明RF在预测出行目的识别方面表现突出[22-23]。文献[5]提出基于时空密度聚类与规则的出行链提取算法,使用XGBoost模型预测非通勤出行目的,准确率达到87.6%;文献[19]基于社交媒体签到数据,采用XGBoost模型对用户个体的活动进行识别,识别准确率为87%;文献[21]利用人工

神经网络和粒子群优化算法从GPS数据中预测出行目的,准确率达到96.5%;文献[20]利用家庭出行调查数据、历史推特等数据,采用贝叶斯神经网络方法预测多种出行目的,不同出行目的预测的准确率在69.36%~90.52%。

当前,在出行目的识别方面的不足表现为数据量不足、精度低、时间周期短和数据类型单一。当前研究对于非通勤出行目的预测效果不理想,多数研究侧重于使用地理信息数据构建特征变量,较少研究能够综合考虑建成环境、活动属性和用户属性构建特征来预测非通勤出行目的。忽略用户历史活动的时空信息以及相似社会经济属性的用户具有高度相似的出行行为。因此,本文采用湖南省长沙市大规模、高精度的手机定位数据和多源时空数据,弥补当前研究数量不足、数据类型单一等问题,并将基于规则的方法和机器学习方法融合,提出了居民出行目的的识别框架。首先,在出行链提取的基础上,基于规则方法提取用户职住信息,进而识别出通勤出行目的;然后,结合兴趣面(area of interest,AOI)数据识别部分非通勤出行目的,提升了非通勤出行目的的识别的全面性与准确性;最后,综合多源时空数据,如建成环境、用户属性、活动属性和用户历史活动时空信息,构建特征变量,构建*K*-Prototypes-XGBoost模型来预测剩余的非通勤出行目的。模型将用户历史活动时空分布特征纳入特征构建过程,能够更准确地反映用户的出行行为,有效地提升非通勤出行目的识别的准确率。其次,利用*K*-Prototypes聚类算法[24-25]提取相似出行行为的群体,为每个不同群体分别构建XGBoost模型来识别非通勤出行目的,这一过程有效挖掘了群体间的相似性,增强了非通勤出行目的识别方法的准确性和可解释性。

1　出行目的识别方法

本文所提出行目的识别框架如图1所示。首先,提取用户的停留点,以实现完整的出行链构建;其次,开展活动特征分析,挖掘用户的出行特征;然后,基于居民活动特征识别职住信息,在此基础上进一步识别居民通勤出行目的;最后,对非通勤出行目的进行精细化识别。

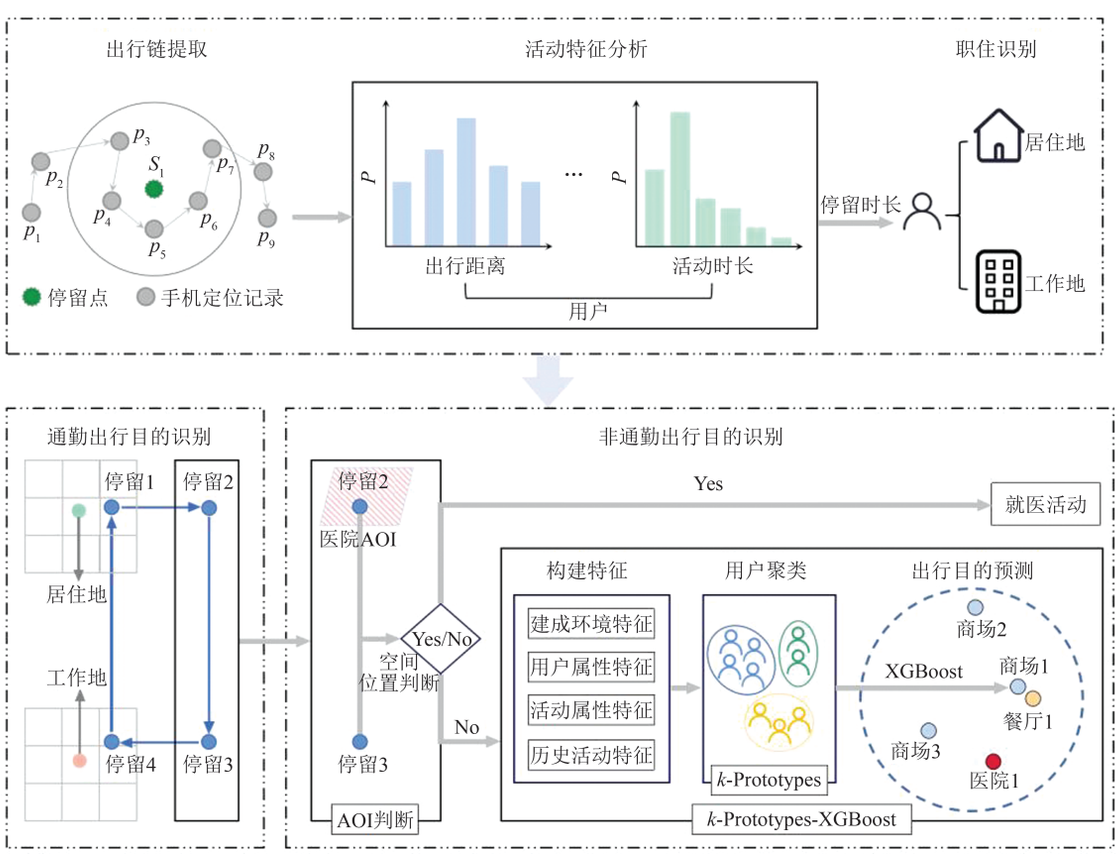


图1　本文模型框架

Fig. 1　Framework of the Proposed Method

1.1 出行链提取

针对时空精度不均匀的手机轨迹数据进行出行链重构,识别出手机用户的停留点和经过点,其中主要通过空间阈值和时间阈值识别停留点。空间阈值是用户在一个位置停留时的游走距离,与定位误差有关[26]。考虑到本文使用的手机定位数据定位误差在10~50 m,将停留点的游走距离设为100 m。时间阈值是用户在一个位置的最小停留时间,通过在一个停留点观察到的第一条和最后一条记录之间的持续时间来衡量,本文的时间阈值设置为10 min。停留点的所有记录的中心点作为停留的空间位置。具体的过程如下:

1)初始化。轨迹点序列为$(p_1,p_2,\cdots,p_n)$,对应的时间序列为$(t_1,t_2,\cdots,t_n)$,其中每个轨迹点p_i包含经纬度坐标(x_i,y_i)和时间戳t_i。

2)计算距离。使用地理距离计算函数计算两个点之间的距离:

$$d(p_i,p_j)=g((x_i,y_i),(x_j,y_j)) \tag{1}$$

式中,$d(p_i,p_j)$为轨迹点p_i和p_j之间的地理距离;$g(\cdot)$表示地理测地距离计算函数;(x_i,y_i)和(x_j,y_j)分别是p_i和p_j的地理坐标。

3)质心计算。设当前停留点的质心为o,其坐标(x_o,y_o)的计算式为:

$$x_o=\frac{1}{|S|}\sum_{p_k\in S}x_k \tag{2}$$

$$y_o=\frac{1}{|S|}\sum_{p_k\in S}y_k \tag{3}$$

式中,S为当前候选停留点包含的轨迹点集合,$|S|$为集合S中轨迹点的数量;p_k为集合S中的第k个轨迹点,其坐标为(x_k,y_k)。

4)判断停留点。初始化停留点集合$S_{\text{start}}=\{p_1\}$,质心$o_{\text{start}}=(x_1,y_1)$,起始时间$t_{\text{start}}=t_1$。对于第$i$个轨迹点$p_i$,计算质心$o$到轨迹点$p_i$的距离:

$$d(o,p_i)=g((x_o,y_o),(x_i,y_i)) \tag{4}$$

若$d(o,p_i)\leqslant 100$ m,则将p_i加入集合S,并更新质心o,否则,计算集合S持续时间$t_{\text{duration}}=t_{i-1}-t_{\text{start}}$,若持续时间$t_{\text{duration}}\geqslant 10$ min,则记录停留点,并将p_i作为新的起点,继续上述过程。

停留点识别示例如图2所示,其中灰色点代表手机用户移动轨迹,点p_3~p_7形成一个停留点S_1,S_1满足时间阈值($t_7-t_3\geqslant 10$ min)和空间阈值(S_1停留半径为100 m)。

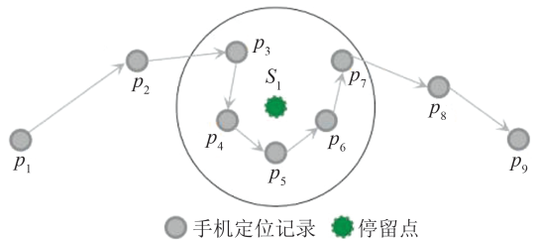


图2 停留点识别示例图

Fig. 2 Stay Point Identification of Mobile Phone User

1.2 通勤出行目的的识别

通勤出行是指从居住地往返工作地的出行[27],出行目的包含回家和上班,因此需要先识别用户的居住地和工作地位置。手机用户的居住地/工作地位置通常由手机用户在夜间/白天停留时长最大[28]和停留次数[29]最多的位置确定。因此,本文将研究区域划分为300 m×300 m的网格[26],该尺度的网格既能覆盖大部分居住区,又能适应较小尺度的工作区,同时能够在误差控制和计算效率之间取得平衡。计算手机用户从19时至次日8时在每个网格及其8个相邻网格中的总停留时长,将该时段内停留时长最大的网格中心标记为用户居住地。将工作日8时至19时之间的活动位置定义为候选工作地,计算手机用户在每个网格及其8个相邻网格中的总停留时长,限制用户在每周内访问其工作停留区域的次数不得少于3次,对于有明确居住地位置的用户,候选工作停留点与居住地的距离要大于500 m[26]。当候选工作停留区域满足上述条件时,将网格中心标记为工作地。进一步将居住地/工作地所在的300 m×300 m网格及其8个相邻网格中的活动视为基于居住地/工作地的活动,如在用户的居住地附近发生的一些购物、娱乐、医疗活动被视为基于居住地的活动,这些活动通常在用户居住社区进行,相应的移动一般不被视为通勤出行[7]。

1.3 非通勤出行目的的识别

1.3.1 基于AOI识别非通勤出行目的

非通勤出行目的通常包含休闲娱乐、购物、医疗、餐饮和生活服务,推断难度较大。许多研究发现,活动位置选择与周边的土地利用及设施类型具有高度相关性[7]。本文采集大型休闲娱乐、购物、医院、餐饮设施、生活服务的AOI数据,将活动的经纬度坐标与大型设施多边形进行空间位置匹配来识别非通勤出行目的。

1.3.2 基于*K*-Prototypes-XGBoost算法识别用户非通勤出行目的

许多未知目的活动无法通过AOI数据确定,

这些活动既没有Wi-Fi信息记录,也不位于AOI内。为了获得用户完整的出行目的活动链,本文提出基于*K*-Prototypes-XGBoost模型来预测剩余非通勤出行目的,模型主要分为3个部分:

1)构建特征变量

考虑到出行目的预测难度较大,当前许多研究提出的预测模型大多考虑了多源时空数据和机器学习模型来提高预测准确率[5-7,19,22-23]。诸多研究发现,个体在某一地点所进行的活动与该地点的建成环境密切相关,因此,本文采用活动周围各类POI的数量来表征建成环境特征。此外,个体的社会经济属性也对其活动有较大的影响[23]。年龄、性别、教育水平、收入水平、是否拥有汽车、是否有子女以及居住地的网格坐标等信息用于表征手机用户的社会经济属性特征。研究发现,不同的活动存在时空异质性[6,11,30],如不同活动的发生时间、持续时长、出行距离、与居住距离和活动所在的位置都存在一定的差异。值得一提的是,用户的历史出行特征对出行目的识别、需求预测有很大的帮助,因此本文将个体历史活动时空信息作为模型的输入特征。综上所述,本文基于建成环境、用户属性、活动属性和用户历史活动时空特征构建特征变量。

2)基于*K*-Prototypes的用户聚类

考虑到用户信息包括多种类型特征,而多数聚类方法如*K*-Means、*K*-Modes仅能处理单一类型的特征[24,31],本文使用*K*-Prototypes算法[24-25]对用户信息进行聚类,提取相似出行行为群体。具体流程如下:

步骤1:建立用户特征集合。用户集合$X=\{X_1,X_2,\cdots,X_n\}$,属性集合$A=\{A_1^r,A_2^r,\cdots,A_p^r,A_{p+1}^c,A_{p+2}^c,\cdots,A_m^c\}$。每个用户含有$m$个属性,前$p$种为数值型,其余为分类型。

步骤2:初始化聚类中心。在初始阶段会随机选择k个数据点作为初始聚类中心,满足簇$C=\{C_1,C_2,\cdots,C_k\}$且$C_i\cap C_j=\varnothing,\mathrm{U}_{I=1}^{k}C_I=X(i,j=1,2,\cdots,k;\ i\neq j)$。簇$C_I$聚类中心为$z_I=(z_I^r,z_I^c)$,数值型和分类型属性的聚类中心分别为$z_I^r=(z_{I,1}^r,z_{I,2}^r,\cdots,z_{I,p}^r)$,$z_I^c=(z_{I,p+1}^c,z_{I,P+2}^c,\cdots,z_{I,m}^c)$。

步骤3:计算数据点x_i至聚类中心z_I距离并重新划分样本点。混合类型的数据距离定义为:

$$d(x_i,z_I)=\sqrt{\sum_{j=1}^{p}(x_{i,j}{}^r-z_{I,j}{}^r)^2}+\gamma\sum_{t=p+1}^{m}\delta(x_{i,t}^c,z_{I,t}^c) \tag{5}$$

$$\delta(x_{i,t}^c,z_{I,t}^c)=\begin{cases}1,x_{i,t}^c\neq z_{I,t}^c\\0,x_{i,t}^c=z_{I,t}^c\end{cases} \tag{6}$$

式中,参数γ用来调整分类型属性的权重。得到所有$d(x_i,z_I)$后,将x_i划分到距离最近的簇中。分别计算C_I中数据点的各数值属性均值以及各分类属性众数,更新聚类中心z_I。

步骤4:重复步骤3,直到满足终止条件。*K*-Prototypes算法最小化目标函数为:

$$F=\sum_{I=1}^{k}\sum_{i=1}^{n}u_{Ii}d(x_i,z_I) \tag{7}$$

式中,判断系数$u_{Ii}\in\{0,1\}$,且$\sum_{I=1}^{k}u_{Ii}=1(1\leqslant i\leqslant n)$,$0<\sum_{i=1}^{n}u_{Ii}<n(1\leqslant I\leqslant k)$。当$u_{Ii}=1$时,表示第$i$个对象属于第$I$个类;当$u_{Ii}=0$时,则不属于第$I$个类。若满足收敛的条件或超过最大的迭代次数,流程终止。

3)基于XGBoost的出行目的预测

XGBoost是一种基于梯度提升的机器学习算法,其核心思想是将多个弱学习器的预测结合起来形成一个强学习器[32]。XGBoost具有可扩展性强、运行速度快等优点,尤其适用于处理庞杂的数据集。本文基于上述构建的特征变量作为输入特征,为每一个簇构建XGBoost模型并预测出行目的,其中,将已知目的活动作为训练集,预测剩余未知目的活动。

2　研究区概况及数据说明

2.1　研究区概况

本文研究区选择长沙市,作为湖南省省会城市,长沙市拥有超过1 000万的常住人口,是一个人口密集、人口结构多样且社会经济发展较快的城市,在湖南省及全国范围内均属于重要的交通、经济、文化中心。长沙市在出行需求方面的复杂性和多样性,来源于其庞大且多元化的人口结构,包括不同年龄、性别、职业、收入水平和教育背景的居民群体。此外,长沙市的城市化进程加快,伴随着交通基础设施的发展和变化,出行模式和需求不断演变。因此,长沙市在人口规模、空间分布、交通网络以及社会经济特征等方面为本文提供了丰富的数据基础,成为探索出行目的和行为模式的理想城市案例。

2.2　数据说明

本文采用长沙市2023-05-01—2023-05-31的手机用户轨迹数据,记录了11 319 998个手机用户

的123亿条时空间信息,定位误差约为10～50 m。对于每个移动电话用户,平均每20 min收集一次位置数据,其中约有20%的位置信息由Wi-Fi设备收集生成,记录该位置所在的POI信息。此外,本文收集了手机用户的性别、年龄、收入水平、受教育程度、是否有车等社会经济属性信息,以及POI、人口数据、行政区划数据。手机定位数据和用户属性信息通过软件开发工具包采集,POI数据通过高德地图API开放接口获取,街道尺度人口数据通过红黑人口库获取。

手机定位数据预处理包括:(1)对于同一用户在同一时刻的多条手机定位数据,仅保留一条有效记录。本文通过位移距离异常检测和位移速度异常检测等手段剔除错误位置记录,异常数据的处理阈值被设为:距离阈值3 000 m,速度阈值120 km/h[33]。(2)为满足研究的数据需求,筛选出连续一个月每天12 h有空间位置记录且包含社会经济属性信息的用户轨迹数据,提取出704 591个用户。(3)对手机用户ID脱敏处理,将其空间位置记录按时间排序,构建手机用户轨迹数据记录。

3 实验结果

3.1 特征说明

本文选取5个建成环境特征、9个用户个人信息特征、4个活动属性和2个用户历史活动时空特征构建特征变量。对类别变量采用独热编码[34],剩余变量标准化处理,如表1所示。

表1 特征说明

Table 1 Feature Description

特征	类别	特征描述
建成环境	休闲娱乐	活动点200 m范围内休闲娱乐POI数量
	购物	活动点200 m范围内购物POI数量
	医疗	活动点200 m范围内医疗POI数量
	餐饮	活动点200 m范围内餐饮POI数量
	生活服务	活动点200 m范围内生活服务POI数量
用户个人信息	年龄段	{20～24岁:0,25～29岁:1,30～34岁:2,35～39岁:3,40～44岁:4,45～49岁:5,50～54岁:6,55～59岁:7,60～64岁:8,65～69岁:9,70～74岁:10,75～79岁:11,80～84岁:12,85岁以上:13}
	学历	{低:0,中:1,高:2}
	是否有孩子	{无孩:0,有孩:1}
	性别	{女:0,男:1}
	是否有车	{无车:0,有车:1}
	收入水平	{低:0,中:1,高:2}
	消费水平	{低:0,中:1,高:2}
	是否有工作	{无工作:0,有工作:1}
	居住地经纬度网格坐标	地理坐标系:WGS84
活动属性	活动时长	连续变量,单位为min
	活动开始时间	{凌晨0时—6时:0,早晨6时—8时:1,上午8时—11时:2,中午11时—13时:3,下午13时—17时:4,傍晚17时—18时:5,晚上18时—24时:6}
	至家距离	连续变量,单位为km
	活动所在经纬度网格坐标	地理坐标系:WGS84
历史活动	用户当前活动所在网格各目的历史出行次数	连续变量
	用户当前活动所在时段各目的历史出行次数	连续变量

3.2 活动聚类结果

通过K-Prototypes聚类算法计算在不同K值情况下的误差平方和来选取最佳用户聚类数,如图3所示。由图3可以看出,随着K值的增加,误差平方和逐渐减小;当K=7时,误差平方和的下降幅度变得平缓,表明聚类结果趋于稳定,未再出现明显变化的拐点。因此,本文选择的最佳的用户聚类数为7。

3.3 模型测试结果

采用2023-05-01—2023-05-31一个月已知出

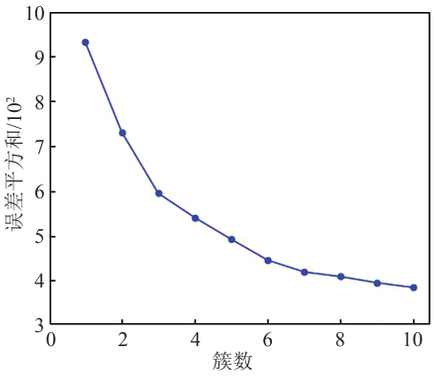


图 3　误差平方和与簇数的关系

Fig. 3　Sum of Squared Errors Under Different Numbers of Clusters

行目的活动数据,验证 *K*-Prototypes-XGBoost 算法预测非通勤出行目的可行性,其中,*K*-Prototypes 聚类的参数 *K* 设置为 7,训练集与测试集比例设为 8∶2。计算不同实验设置下模型预测结果的准确率、精确率、召回率和 *F*1 分数。在对比实验中使用了五折交叉验证,并以各折的评价指标平均值进行对比验证。同时,利用混淆矩阵进一步分析模型在不同活动类别上的识别表现,并对主要的误识别原因进行探讨。

为比较不同模型在出行目的识别中的性能差异,并验证用户历史活动时空特征和用户聚类信息对模型识别效果的提升作用,本文开展了对比分析。首先,在不添加用户历史活动时空特征和用户聚类信息的情况下,将自适应增强(adaptive boosting,AdaBoost)、RF、梯度提升决策树(gradient boosting decision tree,GDBoost)、人工神经网络(artificial neural network,ANN)和 XGBoost 5 种模型的预测结果进行对比分析,结果如图 4 所示。由图 4(a)可以看出,XGBoost 模型性能最优,准确率为 73.2%、精确率为 74.3%、召回率为 73.2%、*F*1 分数为 74.3%。由图 4(b)可以看出,在添加用户历史活动时空特征和不添加用户聚类信息的情况下,5 种模型的预测结果对比分析显示,添加用户历史活动时空特征能够有效提高预测准确率,5 种模型的平均准确率、精确率、召回率和 *F*1 分数分别提高了 24.2%、21.2%、24.1% 和 23.0%。由图 4(c)可以看出,在添加用户历史活动时空特征和用户聚类信息的情况下提出 *K*-Prototypes-XGBoost,并与 XGBoost 模型的预测结果进行对比分析,发现使用 *K*-Prototypes 算法对用户聚类、提取具有相似出行行为特征的人群后,模型对出行目的的预测性能得到进一步提高,其中准确率、精确率、召回率和 *F*1 分数

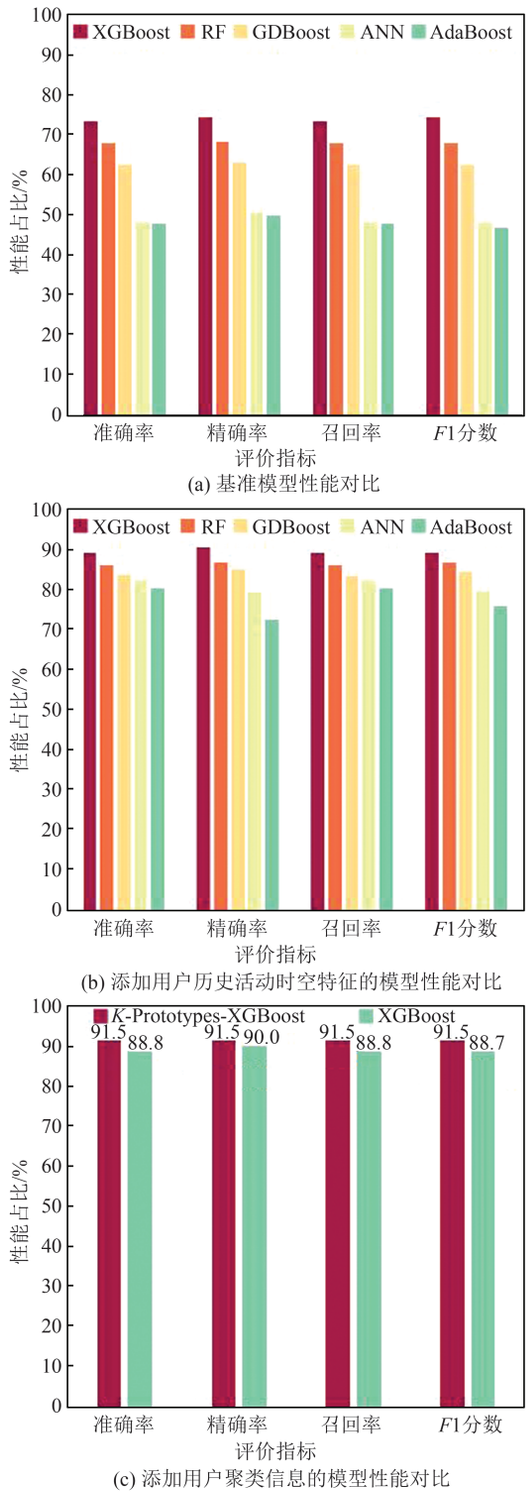


图 4　不同模型预测结果比较

Fig. 4　Prediction Results Comparison of Different Models

分别提高了 2.7%、1.5%、2.7% 和 2.8%。因此,从大规模、长周期的手机定位数据中识别用户的非通勤出行目的时,*K*-Prototypes-XGBoost 模型性能最优,4 个评价指标的值均为 91.5%。

利用混淆矩阵进一步分析 *K*-Prototypes-XGBoost 模型在不同非通勤出行目的类别上的识别表现,如表 2 所示。

表2 *K*-Prototypes-XGBoost的混淆矩阵结果/%

Table 2 Confusion Matrix of *K*-Prototypes-XGBoost/%

真实的出行目的	预测的出行目的				
	休闲娱乐	医疗	生活服务	餐饮	购物
休闲娱乐	92.4	0.3	2.0	2.6	2.7
医疗	0.4	94.7	1.3	1.8	1.8
生活服务	0.8	0.3	93.4	2.5	2.9
餐饮	1.2	0.5	2.8	88.4	7.1
购物	0.9	0.4	2.1	5.1	91.4

由表2可以看出,尽管所提模型在整体上表现良好,其中医疗活动的识别准确率高达94.7%,但休闲娱乐、餐饮和购物活动样本间出现了较多的误识别情况,识别准确率较低,其中餐饮活动识别准确率最低,为88.4%。对这些误识别的样本进行深入分析,发现它们的时空特征具有较高的重叠性,其中,休闲娱乐(电影院、游戏厅等)、购物场所、餐厅通常位于同一商业综合体或邻近区域,不存在明显的界限,导致基于空间特征的区分较为困难。用户可能在相似的时间和空间条件下进行休闲娱乐、餐饮和购物活动,导致这些活动类别的时空信息相似度较高,进而影响了模型的识别准确性。

4 结 语

本文利用长沙市2023-05 704 591个用户的手机定位数据,综合多源时空信息,提出了融合规则方法和机器学习的居民出行目的识别框架,能够较为准确地识别居民出行目的。其中,针对非通勤出行目的识别难度较大的问题,本文将聚类技术与机器学习算法相结合,构建了*K*-Prototypes-XGBoost模型,有效地提升了非通勤出行目的识别的准确性和可解释性。研究结果表明,与常用的5种方法进行对比,*K*-Prototypes-XGBoost模型预测非通勤出行目的效果最好,准确率、精确率、召回率、*F*1分数均高达91.5%。所提模型在不同出行场景下具有较高的泛化能力和适用性,能够有效提升非通勤出行识别的可靠性。此外,发现用户历史活动的时空特征是提高模型性能的关键因素,考虑该特征的5种模型4个评价指标提升均超过20%。与XGBoost相比,*K*-Prototypes-XGBoost模型先对用户聚类,提取相似出行行为的人群,在此基础上预测出行目的能够进一步提高模型性能和可解释性,验证了聚类技术与机器学习算法结合的有效性。

在大数据时代,许多研究从各种数据源调查居民出行行为时,遇到了轨迹数据丰富而活动信息贫乏的困境。本文基于大规模手机定位提出了精度较高的出行目的识别方法,为从个体轨迹中挖掘丰富的活动信息奠定了坚实的基础。但由于本文数据获取难度较大,许多研究无法获取用户历史活动时空信息。另外,本文对于出行目的识别准确率的验证存在一定的局限性,后续将进一步收集用户完整的活动链验证识别结果的准确率。

参 考 文 献

[1] 李锐, 刘朝辉, 吴华意. 城市人口流动感知与建模方法综述[J]. 武汉大学学报(信息科学版), 2025, 50(5): 973-995.
LI Rui, LIU Zhaohui, WU Huayi. A Review of Urban Population Mobility Perception and Modeling Methods[J]. *Geomatics and Information Science of Wuhan University*, 2025, 50(5): 973-995.

[2] MARIA KOCKELMAN K. Travel Behavior as Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from San Francisco Bay Area[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 1997, 1607(1): 116-125.

[3] 苗辰, 王万江. 基于出站旅客出行目的的地铁站口分析及优化:以乌鲁木齐地铁1号线国际机场站为例[J]. 城市建筑, 2020, 17(22): 143-147.
MIAO Chen, WANG Wanjiang. Analysis and Optimization of Subway Station Exit Based on the Travel Purpose of Outbound Passenger—A Case of International Airport Station of Urumqi Metro Line 1[J]. *Urbanism and Architecture*, 2020, 17(22): 143-147.

[4] 黄菲, 肖智雄, 李旭川, 等. 大语言模型智能体驱动的出行日志生成技术[J]. 武汉大学学报(信息科学版), 2026, DOI: 10.13203/j.whugis20250277.
HUANG Fei, XIAO Zhixiong, LI Xuchuan, et al. Travel Diary Generation Based on Large Language Model Agent[J]. *Geomatics and Information Science of Wuhan University*, 2026, DOI: 10.13203/j.whugis20250277.

[5] 周洋, 杨超, 郭唐仪. 基于手机GPS数据的个体出行链提取方法研究[J]. 交通运输系统工程与信息, 2023, 23(5): 45-54.
ZHOU Yang, YANG Chao, GUO Tangyi. Inferring Individual Trip Chains from Smartphone-Based GPS Data[J]. *Journal of Transportation Systems Engineering and Information Technology*, 2023, 23(5): 45-54.

[6] 魁星,关雪峰,杨向阳,等. 顾及时空同步约束的海量轨迹分布式相似查询方法[J]. 测绘地理信息,2025,50(3):103-109.
KUI Xing, GUAN Xuefeng, YANG Xiangyang, et al. Distributed Trajectory Similarity Query with Spatio-temporal Synchronization Constraints[J]. *Journal of Geomatics*, 2025, 50(3): 103-109.

[7] 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*, 2009, 17(3): 285-297.

[8] 王冬根,孙冰夏,宋璟璐. 利用被动式GPS数据的交通行为信息提取方法:发展现状及趋势[J]. 武汉大学学报(信息科学版),2014,39(6):671-681.
WANG Donggen, SUN Bingxia, SONG Jinglu. Methods for Detecting Activity-Travel Behavior Information from Passive GPS Data: State-of-the-Art[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 671-681.

[9] 郭思慧. 职住流的空间分布特征及其影响因素分析:以北京市为例[J]. 北京测绘,2024,38(5):827.
GUO Sihui. Analyses of Spatial Distribution Characteristics of Job-Housing Flows and Their Influencing Factors: Taking Beijing as a Case[J]. *Beijing Surveying and Mapping*, 2024, 38(5): 827.

[10] USYUKOV V. Methodology for Identifying Activities from GPS Data Streams[J]. *Procedia Computer Science*, 2017, 109: 10-17.

[11] NGUYEN M H, ARMOOGUM J, MADRE J L, et al. Reviewing Trip Purpose Imputation in GPS-Based Travel Surveys[J]. *Journal of Traffic and Transportation Engineering* (*English Edition*), 2020, 7(4): 395-412.

[12] GONG L, LIU X, WU L, et al. Inferring Trip Purposes and Uncovering Travel Patterns from Taxi Trajectory Data[J]. *Cartography and Geographic Information Science*, 2016, 43(2): 103-114.

[13] WANG P F, LIU G N, FU Y J, et al. Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model[J]. *ACM Transactions on Intelligent Systems and Technology*, 2018, 9(3): 1-26.

[14] CHEN C, GONG H M, LAWSON C, et al. Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned from the New York City Case Study[J]. *Transportation Research Part A: Policy and Practice*, 2010, 44(10): 830-840.

[15] FURLETTI B, CINTIA P, RENSO C, et al. Inferring Human Activities from GPS Tracks[C]// The 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, USA, 2013.

[16] OLIVEIRA M G S, VOVSHA P, WOLF J, et al. Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, 2405(1): 33-41.

[17] LU Y M, LIU Y. Pervasive Location Acquisition Technologies: Opportunities and Challenges for Geospatial Studies[J]. *Computers, Environment and Urban Systems*, 2012, 36(2): 105-108.

[18] MONTINI L, RIESER-SCHÜSSLER N, HORNI A, et al. Trip Purpose Identification from GPS Tracks[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, 2405(1): 16-23.

[19] 郭茂祖,陈加栋,张彬,等. 融合空间偏好和语义的个体活动识别方法[J]. 国防科技大学学报,2022,44(3):57-66.
GUO Maozu, CHEN Jiadong, ZHANG Bin, et al. Method for Individual Activities Recognition Incorporating Spatial Preference and Semantics[J]. *Journal of National University of Defense Technology*, 2022, 44(3): 57-66.

[20] CUI Y, MENG C S, HE Q, et al. Forecasting Current and Next Trip Purpose with Social Media Data and Google Places[J]. *Transportation Research Part C: Emerging Technologies*, 2018, 97: 159-174.

[21] XIAO G N, JUAN Z C, ZHANG C Q. Detecting Trip Purposes from Smartphone-Based Travel Surveys with Artificial Neural Networks and Particle Swarm Optimization[J]. *Transportation Research Part C: Emerging Technologies*, 2016, 71: 447-463.

[22] 丁韦娜,陈碧宇,张津萌. 可达性与非工作出行需求的非线性关系分析:基于手机大数据的实证分析[J]. 测绘地理信息,2025,DOI:10.14188/j.2095-6045.20240755.
DING Weina, CHEN Binyu, ZHANG Jinmeng. Research on Nonlinear Associations Between Accessibility and Non-work Travel Demand: An Empirical Analysis Based on Mobile Big Data[J]. *Journal of Geomatics*, 2025, DOI: 10.14188/j. 2095-6045.20240755.

[23] ERMAGUN A, FAN Y L, WOLFSON J, et al.

Real-Time Trip Purpose Prediction Using Online Location-Based Search and Discovery Services[J]. *Transportation Research Part C: Emerging Technologies*, 2017, 77: 96-112.

[24] 翟鸣宇, 程建, 王苏桐, 等. 基于*K*-Prototype聚类的学生教育画像分析[J]. 大连理工大学学报(社会科学版), 2021, 42(6): 22-31.
ZHAI Mingyu, CHENG Jian, WANG Sutong, et al. Portraying Student Education Based on the *K*-Prototype Clustering Method[J]. *Journal of Dalian University of Technology* (*Social Sciences*), 2021, 42(6): 22-31.

[25] RANTI K S, SALIM K, YULIONO A D, et al. Clustering Behavioral Data for Advertising Purposes Using *K*-Prototypes Algorithm [J]. *International Journal of Innovative Technology and Exploring Engineering*, 2019, 9(1): 2329-2334.

[26] JIANG S, YANG Y X, GUPTA S, et al. The TimeGeo Modeling Framework for Urban Motility Without Travel Surveys[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, 113(37): E5370-E5378.

[27] CHEN Y Y, YIN C, SUN B D. Nonlinear Associations of Built Environments Around Residences and Workplaces with Commuting Satisfaction[J]. *Transportation Research Part D: Transport and Environment*, 2024, 133: 104315.

[28] 许宁, 尹凌, 胡金星. 从大规模短期规则采样的手机定位数据中识别居民职住地[J]. 武汉大学学报(信息科学版), 2014, 39(6): 750-756.
XU Ning, YIN Ling, HU Jinxing. Identifying Home-Work Locations from Short-Term, Large-Scale, and Regularly Sampled Mobile Phone Tracking Data [J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 750-756.

[29] TOOLE J L, COLAK S, STURT B, et al. The Path most Traveled: Travel Demand Estimation Using Big Data Resources[J]. *Transportation Research Part C: Emerging Technologies*, 2015, 58: 162-177.

[30] 柯玮文, 吴升, 柯日宏. 基于OD流向语义和时空语义聚类的居民出行特征分析方法[J]. 地球信息科学学报, 2023, 25(11): 2150-2163.
KE Weiwen, WU Sheng, KE Rihong. A Method for Analyzing Residents' Travel Characteristics Based on OD Flow Semantics and Spatio-temporal Semantic Clustering[J]. *Journal of Geo-Information Science*, 2023, 25(11): 2150-2163.

[31] HUANG Z X, NG M K. A Fuzzy *K*-Modes Algorithm for Clustering Categorical Data [J]. *IEEE Transactions on Fuzzy Systems*, 1999, 7 (4): 446-452.

[32] CHEN T Q, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016.

[33] CAI M, ZHANG Z X, XIONG C, et al. An Adaptive Staying Point Recognition Algorithm Based on Spatiotemporal Characteristics Using Cellular Signaling Data[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23 (8): 10458-10468.

[34] OKADA S, OHZEKI M, TAGUCHI S. Efficient Partition of Integer Optimization Problems with One-Hot Encoding [J]. *Scientific Reports*, 2019, 9: 13036.