引用格式:方志祥,倪雅倩,黄守倩.顾及上网行为特征的手机用户停留行为预测方法[J].地球信息科学学报,2020,22(1):136-144. [Fang Z X, Ni Y Q, Huang S Q. Mobile phone user stay behavior prediction method considering mobile APP usage characterization[J]. Journal of Geo-information Science, 2020,22(1):136-144.] **DOI:**10.12082/dqxxkx.2020.190655

顾及上网行为特征的手机用户停留行为预测方法

方志祥[1],倪雅倩[2],黄守倩[1]

1. 武汉大学测绘遥感信息工程国家重点实验室,武汉 430079;2. 高德软件有限公司,北京 102200

Mobile Phone User Stay Behavior Prediction Method Considering Mobile APP Usage Characterization

FANG Zhixiang[1*], NI Yaqian[2], HUANG Shouqian[1]

1. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; 2. Autonavi Holdings Limited, Beijing 102200, China

Abstract: With the development of information and communication technology, mobile phones have become an indispensable part of human daily life. Human activities have gradually extended from real space to cyberspace. The online behavior of cyberspace in the era of mobile Internet is inseparable from the travel behavior of real space. The current individual travel behavior predictive modeling is less concerned with the relationship between online behavior and travel behavior. A mobile phone user stay behavior prediction model based on the characteristics of online app usage behavior is proposed. Firstly, the time-space constraint is used to define the mobile phone user's stay behavior. Then, from multi-source data, the paper extracts the individual travel behavior's space-time preference, the app usage characteristics such as the APP combination, Internet traffic, Internet access times and other Internet behavior characteristics and weather information, etc. Feature engineering is done in a time and space crossing way, and the mobile phone user stay behavior prediction model with high interpretability from feature to model is constructed. We found the following from the experimental results: (1) The prediction accuracy of the model is 80.31%. After the integration of online behavior characteristics, weather and other external factors, the prediction accuracy is improved by 12.08%, compared with the model only using individual travel characteristics. (2) The prediction accuracy of the model is higher than that of Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT). And it is 1.23% lower than that of Random Forest (RF), but the model in this paper runs faster than RF, and the model solving process is easy to understand and interpretable. Besides, the first-order Markov model has a small amount of calculations and a fast running speed, but the accuracy is low. In general, the model in this paper has higher accuracy and fast running speed, which is more suitable for mobile phone user stay behavior prediction. (3) There is a big difference in the prediction accuracy of different users' stay behaviors prediction. The prediction accuracy of most users is concentrated between 70% and 90%. The highest prediction accuracy is 98.2%, and the worst prediction accuracy is 34.5%. (4) Among the app usage characteristics, the APP

收稿日期:2019-11-04;**修回日期:**2019-12-31.

基金项目:国家重点研发计划项目(2017YFB0503802);国家自然科学基金项目(41231171)。 [**Foundatin items:** National Key Research and Development Program of China, No.2017YFB0503802; National Natural Science Foundation of China, No.41231171.]

作者简介:方志祥(1977—),男,湖北咸宁人,教授,主要从事时空行为建模、导航与位置服务研究。E-mail: zxfang@whu.edu.cn

combination, Internet traffic and Internet access times contribute more to the prediction of mobile user stay behavior. The use of navigation, news and office apps has a particularly significant impact on the prediction results. In addition, comparing to the historical travel behavior characteristics, travel distance and activity radius in the current period have a stronger impact on the prediction of mobile phone user stay behavior.

Key words: mobile phone location data; mobile Internet data; mobile behavior patteren; stay behavior prediction; online behavior characteristics; feature fusion; machine learning; logistic regression model

***Corresponding author:** FANG Zhixiang, E-mail: zxfang@whu.edu.cn

摘要：随着信息通讯技术的发展，手机成为人类日常生活不可缺少的一部分，人类活动逐渐从现实空间延伸至网络空间，在移动互联网时代，网络空间的上网行为与现实空间的出行行为密不可分。当前个体出行行为预测建模较少考虑上网行为与出行行为间的关系，本文提出一种融合上网行为特征的手机用户停留行为预测模型，通过时空约束定义手机用户的停留行为，在考虑个体出行行为时空偏好的同时，融合手机用户使用的APP组合、上网流量、上网次数等上网行为特征以及天气信息等外部特征，从时间、空间的角度进行特征交叉，构建从特征到模型均具有高可解释性的手机用户停留行为预测模型。实验证明：本文模型预测准确率为80.31%，且在融合上网行为特征、天气等外部因素后，比仅使用个体出行特征进行手机用户停留行为预测提升了12.08%。

关键词：手机数据；手机上网流量数据；移动行为模式；停留行为预测；上网行为特征；特征融合；机器学习；逻辑回归模型

1 引言

目前，人类移动行为模式、位置预测等受到广泛关注[1-2]。个体出行行为预测是基于城市居民的历史出行数据，对下一时段的行为和位置进行预测，是理解人类移动行为模式、城市交通规划、传染病防控、精准的基于位置的服务（Location Based Services, LBS）的基础[3-4]。个体出行行为包含移动和停留2个部分，当前研究对移动行为关注较多，停留行为的研究相对较少。在停留模式分析方面，学者围绕停留模式[5-6]、停留时间[7]等展开研究，而在停留行为的预测方面，相关研究较少。

当前城市居民的日常活动越来越离不开以智能手机为载体的移动互联网，现实空间与虚拟网络空间活动紧密联系，变革了人类的生活方式和行为习惯。现实空间出行行为与网络空间上网行为联系日益紧密，目前国内外学者从实证分析、行为预测、相关性分析等角度对现实空间与网络空间行为间的关系展开了研究[8-11]，如Qiao等[12]提出了大数据驱动的评级框架，证明了移动行为和上网行为间存在强关联；Zheng等[13]从手机用户轨迹数据中提取用户移动速度作为关键指标，分析其与APP使用模式（数据流量大小、使用频次、APP类别）间的关系，但分析上网行为特征对出行行为预测作用的研究相对较少。Do等[14]分析手机传感器数据和APP使用记录中蕴含的上下文信息，建立个体/群体混合预测模型，预测未来10 min手机用户在现实空间的位置、网络空间使用的APP，位置预测准确率为65.7%；Huang[15]利用Twitter签到数据对现实空间的轨迹进行预测，通过时空密度聚类计算区域间转移概率，建立稀疏移动马尔科夫链（Sparse Mobility Markov Chain, SMMC）模型，在工作日取得了78.94%的准确率；Hong等[16]认为时间和位置是利用手机数据分析与预测人们在网络空间行为的重要特征，能帮助人们理解用户的偏好模式，更准确地预测用户与手机的交互行为。但当前研究主要从上网行为发生的位置入手进行出行预测，对上网行为与出行行为间的关系挖掘相对较少。

综上所述，国内外研究学者从人类移动模式、出行行为预测等对出行行为进行了深入研究，但在探讨个体出行行为与上网行为间的关联时，少有研究探讨手机上网行为特征对个体出行行为预测影响，而在进行出行行为预测时加入手机上网行为特征也许可以提高预测的精度。因此，本文融合手机基站位置更新数据、上网数据、POI数据等多源数据，从行为地理学视角，融合上网行为特征、出行时空行为特征以及天气等外部特征，构建手机用户下一时段的停留行为预测模型。

2 融合上网行为特征的手机用户停留行为预测方法

2.1 停留行为预测框架

本预测方法融合手机用户基站位置更新数据、

上网流量记录数据、POI数据、天气数据等多源数据，从行为地理学、时间地理学角度提取了手机用户出行行为、上网行为的时、空间特征以及体现出行行为与上网行为间的关联的APP组合特征，通过时间、空间维度的特征交叉产生多维特征。由出行特征集、上网特征集、外部特征集等特征集合进行组合，构成停留行为特征集，从而从多个维度刻画手机用户的停留行为，构建从特征到模型求解均具有高可解释性的停留行为预测模型，手机用户停留行为预测方法框架如图1所示。该预测方法主要包括停留行为识别、停留行为预测特征集构建和停留行为预测算法等内容。

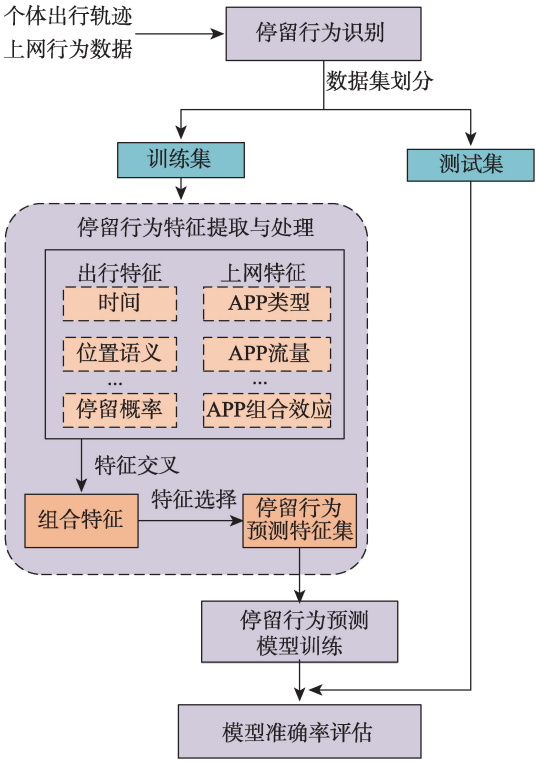

图1 个体停留行为预测的方法流程

Fig. 1 Basic procedures of the individual stay behavior prediction method

2.2 停留行为识别

本研究对个体出行行为的理解是在个体发生停留行为的基础上，将个体在停留行为间的出行视为手机用户的一次移动行为，本预测算法中对个体停留行为的定义为：运动对象在一段时间内保持在某一区域内，则视为该对象在该区域发生了停留行为。其中停留区域、停留时长是定义“停留”这一时空行为的2个必要因素。参考经典的SMoT（Stops and Moves of a Trajectory）停留点识别算法[17]，本文从时空约束的角度给出了停留行为的表达形式如式（1），即时间维度上该个体在该区域内的第一次出现的时间和最后一个轨迹点出现的时间相隔时长不小于时长 T，并且在空间维度上，个体出行轨迹满足任意轨迹点间的欧式距离不超过距离阈值 D。

$$\begin{cases} distance(loc_i, loc_j) \leqslant D \\ timeInterval(t_i, t_j) \geqslant T \end{cases} \tag{1}$$

式中：loc_i、loc_j 分别表示第 i 和第 j 个轨迹点对应的基站位置；$distance(,)$ 是计算两个基站位置间的欧氏距离函数；t_i、t_j 分别表示第 i 和第 j 个轨迹点对应的时刻；$timeInterval(,)$ 是计算两个轨迹点之间时间间隔的函数；D 是空间约束距离阈值；T 是时间约束的间隔时长阈值。

由此，可以从个体出行基站位置更新数据中提取出个体的停留、移动行为，如图2所示。因此，个体的停留区域对应于个体每天经过的基站位置中存在停留行为的区域，提取停留行为即寻找个体每天经过的众多基站位置中产生停留行为的基站。

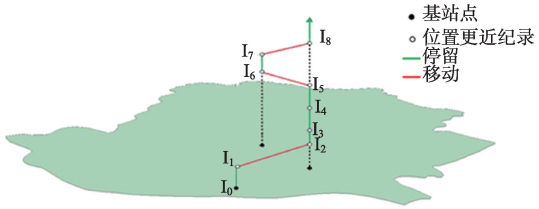

图2 个体出行的停留行为示意

Fig. 2 The example of individual stay behavior

2.3 停留行为预测特征集构建

2.3.1 出行行为、上网行为及外部因素特征提取

手机用户出行行为特征的提取围绕时间、空间2个方面展开，具体特征包括：出行距离、活动半径、轨迹熵、访问位置个数，及下一时段手机用户的历史停留时长平均值、最大值、最小值、中位数、标准差，此外，还通过该基站区域的POI数据定量计算该区域功能的多样性、当前区域POI与“家”和“工作地”所在区域的相似性。其中，区域功能多样性 E_{poi} 的计算参考了熵的概念，通过计算不同类别POI在同一区域出现的混乱程度，分析区域的功能特性，计算方法如式（2）所示。

$$E_{poi} = -\sum p(poi)\log_2 p(poi) \tag{2}$$

式中：E_{poi} 是某基站区域的区域功能多样性；$p(poi)$ 表示某类型POI数量占该区域POI总数的比例；

poi 表示某类别的POI。

手机用户的上网行为的特征提取主要从APP使用类别数、使用APP个数、使用的数据流量，单一类型APP使用频次、产生组合效应的APP使用组合矩阵5个方面提取用户的当前、历史上网行为，具体地包括：上网次数、APP类别数、各类别APP和组合效应APP组合的使用次数、平均使用次数、各类别和组合效应APP组合使用数据流量大小、数据流量大小的平均值、最大值、最小值、中位数、标准差。此外，还用发生上网行为的位置个数来表征上网行为的空间偏好。

外部因素特征则主要包含天气信息、是否为工作日等特征，用于刻画手机用户出行行为的其他影响因素。具体包含：工作日/周末类别、天气、温度、体感温度，以及历史数据中下一个时段的气温、体感温度的最大、最小值、中位数、平均值等。

2.3.2 特征交叉和特征选择

由于出行行为本质是一种时空行为，因此在特征交叉时主要从时间、空间特征交叉的角度通过笛卡尔积的方式构建其时空特征。手机用户所在的基站通过经纬度坐标去重后对区域重新编号，得到 n 个区域组成的区域集合 S 。一天中的时段可以离散化后划分为 m 个时段，通过式(3)所示笛卡尔积公式，可以获取 $m\times n$ 手机用户在特定时段 t 、特定位置 p 的特征。

$$\begin{cases} T=\{t_1, t_2, \cdots, t_m\} \\ S=\{p_1, p_2, \cdots, p_n\} \\ T\times S=\begin{bmatrix} t_1p_1 & \cdots & t_mp_1 \\ \vdots & & \vdots \\ t_1p_n & \cdots & t_mp_n \end{bmatrix} \end{cases} \tag{3}$$

式中：T 表示时段的向量；S 表示位置组成的区域集合；t_m 表示特定时段；p_n 表示特定位置；t_mp_n 是在特定时段、特定位置的特征；$T\times S$ 是手机用户所有特征构成的矩阵。

经过特征交叉后，产生了丰富的特征集，但并非所有特征都会对预测产生帮助，因此需要进行特征选择。相较于其他方法，Filter算法选出的特征通用性强，不需构建分类器就可以快速去除大量不相关的特征，因此，研究采用Filter方法中常用的卡方检验的方法进行特征选择。Filter方法的基本思想是对每一维特征进行打分并赋予权重，然后按照权重排序，选出权重高的特征作为预测特征集。而卡方检验进行特征选择的原理是利用构建列联表，按式(4)计算 χ^2 值分析特征与预测类别间的关联，如果特征与类别间的偏离程度过大，则从特征集中剔除该特征。

$$\chi^2=\sum_{i=1}^{k}\frac{\left(f_i-np_i\right)^2}{np_i} \tag{4}$$

式中：χ^2 表示卡方检验的检验统计量；np_i 表示特征属性值 i 出现的期望样本数；f_i 表示特征属性值 i 的实际样本数；k 是特征属性值的个数。

2.4 停留行为预测算法

本算法将手机用户停留行为预测理解为二分类问题，分类结果只有停留、移动。逻辑回归是机器学习中解决二分类问题的经典算法[18]，作为一种广义线性模型，训练得到的Logistic模型易于解释，能融合不同来源的特征。Logistic回归模型的基本思想是通过模型训练对不同因素进行权值分配，通过sigmoid函数归一化到区间[0,1]，根据概率值进行分类。Logistic模型的分布函数$f(x)$如式(5)所示，该函数是一条以点(0, 0.5)为中心对称的“S”型曲线。

$$f(x)=\sum_{i=1}^{k}\frac{1}{1+e^{-w^Tx}} \tag{5}$$

$$w^Tx=w_0+w_1x_1+\cdots+w_ix_i+\cdots+w_nx_n \tag{6}$$

式中：x 表示变量；w 表示权值向量；k 表示变量的个数；x_i 表示第 i 个变量；w_i 表示第 i 个变量对应的权重；w^Tx 为线性函数；$f(x)\in\{0, 1\}$ 。

在构建二分类模型时，使用如式(7)和(8)所示的二项 Logistic模型，可由条件概率 $P(Y|x)$ 表示，形成参数化的Logistic分布。

$$P(Y=1|x)=\frac{e^{w^Tx}}{1+e^{w^Tx}} \tag{7}$$

$$P(Y=0|x)=\frac{1}{1+e^{w^Tx}} \tag{8}$$

式中：$Y\in\{0,1\}$ 表示输出的结果，在本实验中0和1依次对应手机用户发生停留行为、不发生停留行为。当线性函数 w^Tx 趋近于无穷大时，$P(Y=1|x)$ 无限趋近于1；当 w^Tx 趋近于负无穷时，$P(Y=1|x)$ 接近0。

本算法利用极大似然估计法对Logistic模型的对数似然函数 L_w 求极大值，即式(9)，估计逻辑回归模型的 w 参数，得到 Logistic模型。在训练预测模型时，为了防止模拟过拟合，本研究采用L2正则化的方法，求解目标函数 $\min\left(L_w+\frac{1}{2}w^Tw\right)$，得到Logistic模型参数 w 。

$$L_w = \sum_{i=1}^{n} y_i \log P(Y=1|x) + (1-y_i)\log P(Y=0|x) \quad (9)$$

式中：L_w 是逻辑回归模型的对数似然函数；y_i 是分类的标签；n 是样本的个数；$P(Y=1|x)$ 是标签为1的概率；$P(Y=0|x)$ 是标签为0的概率。

在构建停留行为预测模型时，本研究考虑到手机上网行为蕴含丰富的个体活动特征，创造性地将手机用户在网络空间的APP使用特征、APP组合效应等上网行为特征作为特征，通过Logistic模型融合上网行为、出行行为特征、外部因素特征等多方面的特征，对下一时段手机用户发生停留行为的概率建模，得到最终的手机用户停留行为预测模型。

3 实验结果与分析

3.1 实验数据

本研究采用的实验数据是由某移动运营商提供的某城市2015年8月10日–2015年8月29日，共计20 d的手机用户基站位置更新数据及手机流量上网的流量收费数据。本研究使用的数据集均经过匿名化处理，在保护用户隐私的前提下，将其用于手机用户个体出行行为预测研究。除此以外，还用到该城市的POI数据以及研究时间段内该城市的天气数据。

手机位置数据以手机基站的经纬度坐标记录手机用户的位置，包括经过脱敏处理后的手机用户ID、记录产生的时间、基站编号、基站经纬度坐标，如表1所示。手机位置数据包括手机用户的主动和被动定位信息，即当用户位置发生基站间变更、接打电话、收发信息时用户的位置信息将被记录（被动记录），当用户长时间（超过1 h时）未发生上述行为时，手机将会主动捕捉其所在的位置（主动记录），因此每个手机用户一天中至少会产生24条定位数据。

手机上网数据主要记录手机用户日常使用手机上网的行为信息，用于运营商的数据流量收费。具体地，手机用户使用GPRS流量访问网页、使用APP应用、接收消息推送等行为会产生流量记录和基于基站的用户位置信息。每条流量收费记录包含匿名手机用户使用流量上网的时间、基站编号、APP类型、流量大小，如表2所示。

该城市的POI数据是通过百度地图Place API接口[19]获取得到的，包含旅游景点、交通设施、公司企业、政府机构、休闲娱乐、购物等17类，共4308条数据。每个POI点的信息包含POI名称、POI类别、经纬度坐标、详细地址等，POI数据将被用于基站区域功能多样性评估。

该城市的天气数据[20]包含每天00:00–24:00每3 h记录一次该城市的天气、温度、体感温度、降雨量等信息。在研究时段内，该城市出现雨天7次，包含2次中雨、1次大雨；气温超过35 ℃的高温天气2次。天气数据将作为外部因素特征用于手机用户

表1 手机位置更新数据实例

Tab. 1 The examples of one user′s mobile phone location update data records

用户ID	日期	时间	事件类型	基站编号	基站经度/°	基站纬度/°
58****	20YY-MM-DD	07:30	1	11**	115.****	29.****
58****	20YY-MM-DD	08:30	5	11**	115.****	29.****
…	…	…	3	…	…	…
58****	20YY-MM-DD	21:30	4	10**	115.****	29.****
58****	20YY-MM-DD	22:30	10	11**	115.****	29.****

表2 手机上网流量数据实例

Tab. 2 The examples of one user′s Internet traffic data records

用户ID	归属地	日期	时间	基站编号	经度/°	纬度/°	APP	流量/G
58***	HB.WH	20YY-MM-DD	07:32	11**	115.***	29.***	高德地图	0.0016
58***	\N	20YY-MM-DD	08:32	11**	115.***	29.***	微信	0.0092
58***	\N	20YY-MM-DD	09:27	12**	115.***	29.***	QQ	0.0242
…	…	…	…	…	…	…	…	…
58***	\N	20YY-MM-DD	21:06	10**	115.***	29.***	微信	0.0016
58***	\N	20YY-MM-DD	21:34	11**	115.***	29.**	新浪新闻	0.0554

停留行为预测、出行位置预测的建模。

手机位置更新数据和手机流量上网数据存在：字段缺失、记录重复及通信信号漂移导致的用户定位基站跳变等问题，针对上述问题，对手机用户的基站位置更新数据和手机上网数据进行了预处理，具体包含对数据本身缺失值、重复记录的处理和轨迹层面的基站跳变异常数据的消除。

从个体的角度，手机用户的位置更新数据存在很大差异，实验数据中所有手机用户一天内的位置更新记录条数分布图如图3所示，95.54%的手机用户的记录数不超过100条，记录条数只有1~5条的用户占总用户的37.7%。本研究从手机用户的位置更新数据、手机上网数据中筛选出2015年8月10日–2015年8月29日连续20 d每天位置更新记录条数不少于24条、并且每天都有上网数据的手机用户作为实验对象，满足条件的手机用户共计8508人。

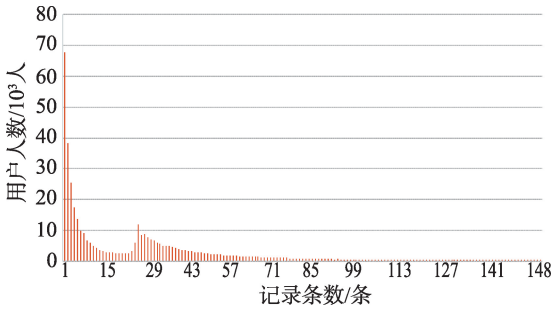

图3　某市2015年8月10日—2015年8月29日手机用户位置更新记录

Fig. 3　Number of location update records for mobile users from Aug.10, 2015 to Aug.29, 2015

3.2 个体停留行为预测实验结果分析

3.2.1 停留行为识别及特征集构建

本实验设置时间阈值T为30 min，距离阈值D为500 m，对06:00–22:00间的停留行为进行识别，手机用户各时段停留行为识别结果如表3所示。约21.42%存在数据记录缺失的情况而无法识别，其中6:00缺失比例最高为33.08%，这可能是手机用户早上关机造成的。在08:00、09:00、17:00、18:00等早晚高峰时段，实验对象的出行更为活跃，识别为移动的比例相对较高。总的来看，选取的实验对象更倾向于停留在某一区域。

对手机用户家和工作地的区域信息，主要根据生活规律来提取，依据人类的生活规律，00:00–06:00为睡眠时间，个体停留在一个区域，因此将一周内个体在00:00–05:00出现概率大于50%的区域作为个体的“家”。相似地，将一周内工作时段（早上9:00–11:00，下午14:00–17:00）出现次数最多的区域作为“工作地”。手机用户的“家”、“工作地”的识别情况如表4所示，利用个体的出行轨迹实现对个体“家”和“工作地”的识别，其中对个体“家”的识别率比“工作地”高31.47%，这可能是由于不同职业工作存在差别造成的，如司机工作时无固定工作地点。

表3　某城市2015年8月10日–2015年8月29日手机用户停留行为识别统计

Tab. 3　The statistics of mobile phone user stay behavior from Aug.10, 2015 to Aug.29, 2015

时段	移动/%	停留/%	无法识别/%
6	14.51	52.41	33.08
7	24.93	49.99	25.07
8	31.66	49.87	18.46
9	33.00	49.57	17.43
10	32.60	49.77	17.63
11	31.79	49.81	18.40
12	30.75	50.16	19.09
13	27.16	51.36	21.48
14	28.41	51.06	20.54
15	28.23	50.96	20.81
16	28.33	50.30	21.37
17	29.59	49.10	21.31
18	30.97	47.30	21.73
19	29.37	50.16	20.47
20	27.33	51.46	21.20

手机用户出行是用户在时间、空间上与城市发生交互的过程，是一个典型的时空行为，特征交叉时，通过对时间、空间特征的交叉，形成了一系列具有现实意义的特征组合。特征提取出的出行行为特征有53维，上网行为特征203维，天气气温、体感温度、工作日类别等外部因素15维。将时段分为7个时段进行时空间特征交叉后，出行行为特征数量大大增加。实验利用卡方检验进行特征选择，筛选出了排名前20%的特征进行后续的预测模型训练。

表4　某城市2015年8月10日—2015年8月29日手机用户地点识别率

Tab. 4　The recognition rate of mobile phone users' home and work location from Aug.10, 2015 to Aug.29, 2015

地点	识别率/%
家	78.53
工作地	47.06
同时识别率	40.09

3.2.2 不同预测算法对比

在本实验中，对手机用户停留行为的预测是对某一时段，手机用户在基站区域发生停留行为的预测，预测结果为手机用户是否发生停留。结合对于用户停留行为的定义，算法中对用户停留行为的预测粒度为30 min，因此算法对比选取的时间尺度同样也为30 min，空间精度为基站范围（300~500 m）。停留行为预测准确率PRE的计算公式如下：

$$PRE=\frac{N_{PR}}{N} \tag{10}$$

式中：N 表示测试集样本数；N_{PR} 表示预测结果正确的样本数；PRE 是停留行为预测的准确率。

实验选取支持向量机（Support Vector Machine, SVM）、梯度下降决策树算法（Gradient Boosting Decision Tree，GBDT）、随机森林（Random Forest，RF）等常用的机器学习分类算法[21-23]、一阶 Markov 模型[15,24]，以及最常访问位置预测（Most Frequented Location Model, Most Value）模型[14]，从预测准确率、运行时间两方面与本文模型进行了对比，实验对比结果如表5所示。由表5可知：本文模型准确率高于支持向量机、GBDT方法，比随机森林预测准确率低1.23%，但在运行速度上，本文模型比随机森林模型运行更快，并且模型求解过程易于理解，可解释性强。一阶Markov模型[25]计算量小，运行速度快，但由于仅考虑了各时段的出行状态转移概率，准确率较低。MostValue模型基于统计的思想将该时段历史状态中出现频率最高的结果作为预测结果，是一种常用的算法准确率对比的基础模型[14]。总的来说本文模型准确率较高，且运行速度快，更适用于手机用户停留行为预测。

表5 不同预测算法用于手机用户停留行为预测的效果比较

Tab. 5 Comparison of mobile phone user stay behavior prediction effects between different prediction algorithms

预测模型	准确率/%	运行时间/s	参数设置
本文模型	80.31	0.141	
SVM	77.35	3.996	kernel='linear' probability=True
GBDT	76.13	1.940	n_estimators=200
RF	81.54	0.960	n_estimators=200
一阶Markov	71.75	–	
MostValue	69.45	–	

3.2.3 预测算法适应性分析

实验分析了不同手机用户个体停留行为预测准确率，平均准确率为 80.31%，标准差为 10.1%。如图4所示。不同用户停留行为预测准确率存在较大差异，大部分用户的预测准确率集中在70%~90%之间，预测准确率最高的可达98.2%，最差的预测准确率为34.5%。

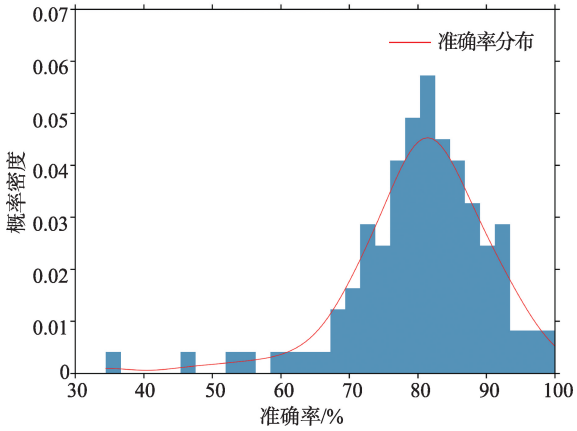

图4 手机用户停留行为的预测准确率分布

Fig. 4 The prediction accuracy distribution of mobile phone user stay behavior

接着实验比较了使用不同维度的特征集对停留行为预测结果的准确率、召回率。根据特征提取的维度，将特征分为出行行为特征集、上网行为特征集、外部因素特征集，组合起来可以得到{出行行为，上网行为}，{出行行为，外部因素}，{上网行为，外部因素}，{出行行为，上网行为，外部因素}等特征集合。分别使用单一维度特征、两两组合特征、所有特征进行停留行为预测模型的构建，模型预测实验结果如图5所示。分析可知：只使用单一维度的特征时，出行行为特征集合预测准确率为比外部

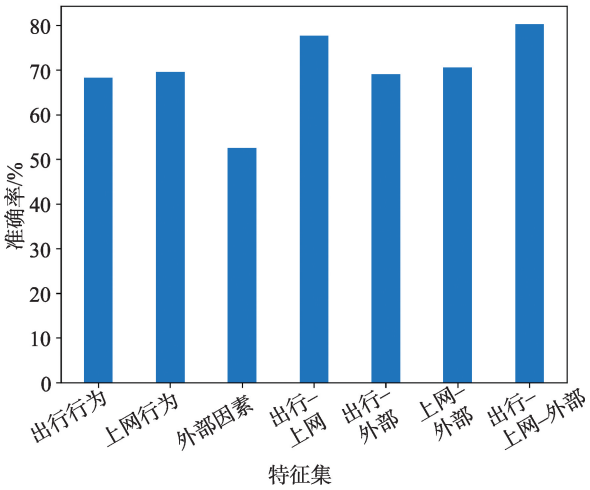

图5 不同维度的特征集用于个体停留行为预测的实验结果对比

Fig. 5 Comparison of mobile phone user stay behavior prediction results with different feature sets

因素进行预测高15.73%，上网行为用于停留行为预测比使用出行特征预测准确率高出1.30%，说明上网行为特征能较好地描述个体出行的停留行为，上网行为与出行行为具有很强的关联。组合使用出行行为、上网行为、外部特征多个维度的特征集后模型预测准确率达到80.31%，出行行为与上网行为特征集组合使用时，模型预测准确率比仅使用出行行为特征提高了9.47%，在进一步加入天气等外部因素后，模型预测准确率又提升了2.61%。而天气、工作日类型等外部因素在停留行为的预测上作用有限。

此外，本研究还利用卡方检验分析了单个特征与停留行为相关的可能性，如图6所示，给出了按卡方值从大到小排序的前20个特征。显然相对历史出行行为特征，当前时段的出行距离、活动半径等特征对停留行为预测的影响更为强烈。同时外部因素与停留行为最为相关的特征为下一时段的历史最低温度，体现了气温对出行行为的影响。在描述手机用户上网行为的特征中，手机用户在当前时段上网的记录条数重要性最高，一定程度上说明了手机上网频繁程度与手机用户出行行为间的联系。APP类型数、APP个数、APP上网数据流量也对手机用户停留行为的预测有一定帮助。另外有3种类别的APP使用特征重要性较高，分别是导航出行类APP、新闻类APP、办公类APP。对停留行为预测的特征卡方值的分析进一步验证了上网行为与出行行为的联系，融合上网行为特征，提升了手机用户出行行为预测的效果。

4　结论与展望

当前个体出行位置预测较少考虑个体上网行为特征对出行的影响，因此本文融合手机用户基站位置更新数据、上网流量记录数据、POI数据、天气数据等多源数据，利用手机用户日常出行行为与上网行为间的关联，对个体出行构建下一时段停留行为的预测模型，通过提取手机用户的出行特征、上网行为特征，并结合外部因素特征，进行特征交叉和选择后，将其用于个体下一时段停留行为的预测，创造性地将手机用户上网行为特征用于用户停留行为预测，且与其他特征结合实现了较好的预测效果。实验结果表明：融合上网行为特征的手机用户停留行为预测模型预测准确率为80.31%，相较于仅使用出行行为特征进行停留行为预测的准确率提升了12.08%。而在手机上网行为特征中，手机用户使用APP上网次数、APP种类数、流量等上网行为特征对手机用户的停留行为预测有较大贡献，特别是导航出行类、新闻类、办公类APP的使用情况。

对于下一时刻停留行为的预测，本文仅针对下一时段手机用户是否发生停留行为展开研究，在未来还可充分发挥POI、土地利用数据等提取区域功能特性，从空间角度对手机用户的下一停留位置进行预测，构建一个更为通用的个体停留行为位置预测模型，更好地发挥LBS服务的高商业价值。若是可获取到手机用户的个体社会经济数据，可将其加入本文模型中来分析不同特征人群的上网行为和

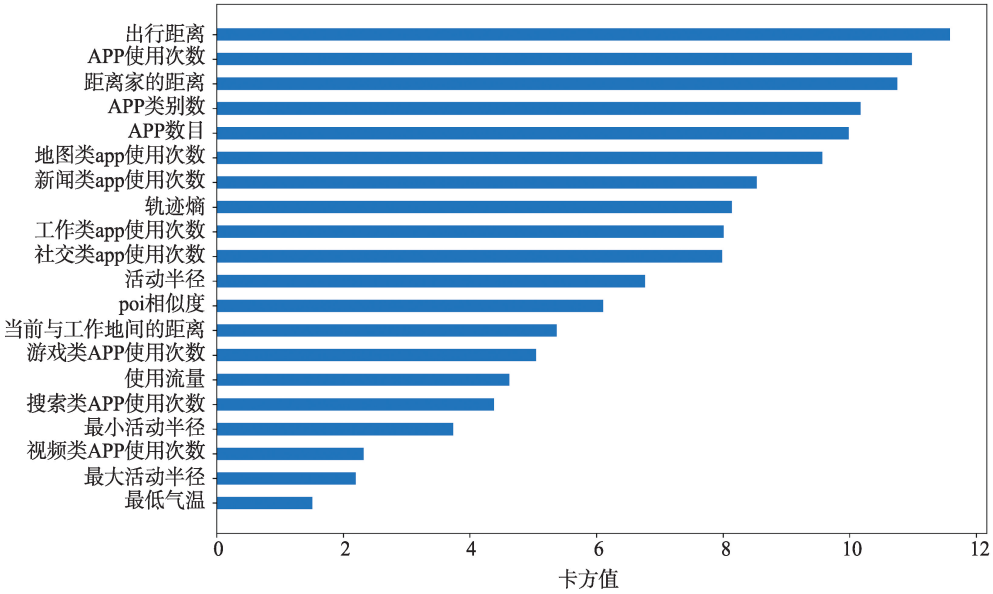

图6　用于停留行为预测的top20特征的卡方值

Fig. 6　The chi-square value of the top20 stay behavior prediction features

出行行为。本文研究主要针对手机用户个体,后续可对手机用户群体停留行为进行研究,此外,地理空间停留与网络空间行为之间的关系机理也是一个有理论研究意义的问题,可在今后的研究工作中展开。

参考文献(References):

[1] Hofman J M, Sharma A, Watts D J. Prediction and explanation in social systems[J]. Science, 2017,355(6324):486-488.

[2] Subrahmanian V S, Kumar S. Predicting human behavior: The next frontiers[J]. Science, 2017,355(6324):486-488.

[3] 钮心毅,丁亮,宋小冬.基于手机数据识别上海中心城的城市空间结构[J].城市规划学刊,2014,6:61-67. [Niu X Y, Ding L, Song X D. Understanding urban spatial structure of Shanghai central city based on mobile phone data [J]. Urban Planning Forum, 2014,6:61-67.]

[4] 萧世伦,方志祥,陈碧宇,等.城市人群活动时空GIS分析[M].北京:科学出版社, 2017. [Shaw S L, Fang Z X, Chen B Y, et al. Spatiotemporal GIS analysis of urban crowd activities[M]. Beijing: Science press, 2017.]

[5] Schneider C M, Belik V, Couronné T, et al. Unravelling daily human mobility motifs[J]. Journal of the Royal Society Interface, 2013,10(84):20130246.

[6] 徐金垒,方志祥,萧世伦,等.城市海量手机用户停留时空分异分析—以深圳市为例[J].地球信息科学学报,2015,17(2):197-205. [Xu J L, Fang Z X, Shaw S L, et al. The spatio-temporal heterogeneity analysis of massive urban mobile phone users' stay behavior: A case study of Shenzhen city[J]. Journal of Geo-information Science, 2015,17(2):197-205.]

[7] 吕遒健.基于无线网络流量的用户移动性分析与应用[D].北京:北京邮电大学,2017. [Lv Q J. Analysis and application of user mobility based on cellular network traffic [D]. Beijing: Beijing University of Posts and Telecommunications, 2017,76-90.]

[8] Wu Z, Wu H, Tong Z. Predict user in-world activity via integration of map query and mobility trace[C]. Proceedings of the 4th International Workshop on Urban Computing, Sydney, Australia, 2015:1-9.

[9] Cao H, Chen Z, Xu F, et al. Revisitation in urban space vs. online: A comparison across POIs, websites, and smartphone apps[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018,2(4):156.

[10] 卢扬,赵志丹,蔡世民.基于移动终端上网数据的移动模式分析及轨迹预测[J].复杂系统与复杂性科学,2015,12(2):53-59. [Lu Y, Zhao Z D, Cai S M. Mobile pattern analysis and trajectory prediction using mobile GPRS data[J]. Complex Systems and Complexity Science, 2015,12(2):53-59.]

[11] Li Q, Qiao Y, Lin W, et al. Understanding correlation between offline mobility and online browsing tendency in mobile network[C]. 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, Quebec, Canada, IEEE, 2017:1-7.

[12] Qiao Y, Zhao X, Yang J, et al. Mobile big-data-driven rating framework: Measuring the relationship between human mobility and app uage behavior[J]. IEEE Network, 2016,30(3):14-21.

[13] Zheng L, Feng Y, Zhou W, et al. Inferring correlation between user mobility and app usage in massive coarse-grained data traces[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018,1(4):153.

[14] Do T M T, Gatica-Perez D. Where and what: Using smartphones to predict next locations and applications in daily life[J]. Pervasive & Mobile Computing, 2014,12:79-91.

[15] Huang Q. Mining online footprints to predict user's next location[J]. International Journal of Geographical Information Systems, 2017,31(3):523-541.

[16] Hong C, Miao L. Mining smartphone data for app usage prediction and recommendations: A survey[J]. Pervasive & Mobile Computing, 2017,37:1-22.

[17] Alvares L O, Bogorny V, Kuijpers B, et al. A model for enriching trajectories with semantic geographical information[C]. Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems. Seattle, Washington, USA, ACM, 2007, 22.

[18] Peng C Y J, Lee K L, Ingersoll G M. An introduction to logistic regression analysis and reporting[J]. Journal of Educational Research, 2002,96(1):3-14.

[19] 百度开放平台API, http://lbsyun.baidu.com/[2017,12].

[20] 中国天气网, http://www.weather.com.cn/[2018,08].

[21] Fernandes R, D Souza, R G L. A new approach to predict user mobility using semantic analysis and machine learning[J]. Journal of Medical Systems, 2017,41(12):188-200.

[22] Xie J, Song Z, Li Y, et al. A survey on machine learning-based mobile big data analysis: Challenges and applications[J]. Wireless Communications and Mobile Computing, 2018:1-19.

[23] Yuan Z, Umek A, Obinikpo A A, et al. A time-fependent multi-class SVM algorithm for crowdsourced mobility prediction[J]. IPSI BGD Transactions on Internet Research, 2018,14(1):1-7.

[24] Menz L, Herberth R, Luo C, et al. An improved method for mobility prediction using a Markov model and density estimation[C]. 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain. IEEE, 2018:1-6.

[25] Castro P S, Zhang D, Li S. Urban traffic modelling and prediction using large scale taxi GPS traces[C]. International Conference on Pervasive Computing. Berlin: Springer, 2012:57-72.