DOI:10.13203/j.whugis20190404 文章编号:1671-8860(2021)06-0799-08

融合Markov与多类机器学习模型的个体出行位置预测模型

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,湖北 武汉,430079
2 高德软件有限公司,北京,102200

摘 要:随着城市化的发展,人们出行的方式逐渐多样化,对人类行为的深入理解以及对个体出行行为的建模预测有助于解释若干复杂的社会经济现象,且在基于位置的服务、交通规划、公共安全等方面具有重要价值。个体出行行为预测建立在深入理解人类活动特性的基础上,而在移动互联网时代,网络空间的上网行为与现实空间的出行行为密不可分。首先基于上网行为特征,融合马尔可夫(Markov)模型和多类机器学习模型,构建了个体出行位置预测模型,该模型使用了基于频率分布图的自适应融合规则,融合了传统的Markov模型和机器学习多分类模型的结果进行个体出行位置预测;然后利用手机数据、上网流量数据、兴趣点数据及天气等多源数据进行个体出行位置预测实验。实验结果表明,该模型的第1个和前3个预测结果中包括正确结果的准确率分别为74.59%、94.19%,均优于基础模型的准确率和利用投票法融合规则融合基础模型的准确率,且预测时间粒度为30 min时,该模型的预测效果较好。

关键词:马尔可夫模型;机器学习模型;出行位置预测;手机数据;特征融合

中图分类号:P208 **文献标志码**:A

随着信息通信技术的发展,数据通信速度和质量不断提升,城市居民的日常活动从现实空间逐渐扩展至网络空间,越发离不开以智能手机为载体的移动互联网。现实空间行为与网络空间上网行为联系日益紧密,探讨个体在现实空间与网络空间的活动差异,建立现实空间活动与虚拟网络空间活动的关联[1-2],有助于个体出行行为预测的研究。

现有研究对移动行为关注较多,在个体移动方面,有关注个体移动行为模式[3-6]、活动空间[7-11]等方面的研究;在出行预测方面,常用的方法包含马尔可夫(Markov)模型、频繁模式挖掘以及神经网络和机器学习方法[12-14];在构建位置预测模型方面,也有学者取得了系列成果[15-17]。随着现实空间出行行为与网络空间上网行为联系日益紧密,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[18-21]。但少有研究探讨手机上网行为特征对个体出行行为预测的影响,应用多模型融合技术预测个体出行位置的研究也较少。Markov预测模型能构建基于停留点语义的出行链,据此进行出行位置的预测,所得结果与实际出行场景更为贴近。机器学习的多分类方法是基于统计的学习方法,准确率较高,但可解释性较差。文献[22]发现融合多个差异较大的分类模型更能提升模型学习的效果,提高准确率。因此,本文利用手机基站位置更新数据、上网数据、兴趣点(point of interest, POI)数据等多源数据,融合上网行为特征、出行时空行为特征及外部因素特征,基于频率分布图的自适应融合规则,融合Markov模型、机器学习多分类模型的预测结果来进行个体出行位置预测。

1 出行位置预测模型的融合方法

1.1 Markov预测模型

Markov预测模型的核心思想是将历史数据中当前状态转移概率最大的状态作为下一状态的预测值。根据Markov理论中转移概率的定义,需要通过条件概率来计算从当前状态转移到

收稿日期:2019-12-06

项目资助:国家自然科学基金(41771473)。

第一作者:方志祥,博士,教授,主要研究时空地理信息系统、人类活动大数据时空建模与分析和行人导航理论与方法。zxfang@whu.edu.cn

下一状态的概率。转移概率在应用于个体出行位置预测时,其定义可参考文献[23]。Markov模型根据对当前状态描述的不同,可以分为一阶Markov和k阶Markov模型。一阶Markov模型仅使用当前时段的位置,对训练数据要求较低;k阶Markov模型则使用更多的历史状态数据,预测的准确率更高,但存在对训练数据要求高、更易冷启动的问题。因此,综合考虑一阶Markov、多阶Markov预测模型的预测结果,有助于提高预测准确率。Markov模型的构建如图1所示。首先根据手机用户的位置更新数据,识别停留点及其语义,构建出行链;然后计算手机用户出行的k阶转移概率$p_{i,j}^t$,构建多个Markov预测模型并进行准确率分析。具体计算公式如下:

$$p_{i,j}^t=\frac{F_{i,j}^t}{\sum_{k=1}^{m}F_{i,k}^t} \tag{1}$$

式中,i、j分别表示用户群体在时段t、$t+1$所在的基站;$F_{i,j}^t$表示手机用户在时段t从基站i移动到基站j的次数;m表示城市区域的手机基站个数。

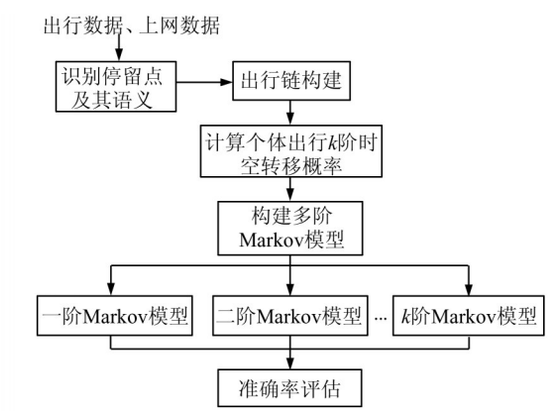

图1 Markov预测模型构建流程图

Fig.1 Construction Procedure of Markov Prediction Model

1.2 机器学习预测方法

本文用到的机器学习预测方法包括决策树(decision tree, DT)、随机森林(random forest, RF)和k近邻(k-nearest neighbor, kNN)算法、支持向量机(support vector machine, SVM)算法等4个经典的多分类机器学习算法。

1)DT算法是一种监督学习算法,该算法构建的DT代表类别属性和属性值间的映射关系,每个内部节点表示某个样本属性,叶子节点表示一个或多个类,分叉路径代表可能的属性值,每个叶子节点为从根节点到该叶子节点所经历的属性路径所表示的类别[24]。回归分类树算法(classification and regression tree, CART)以基尼系数最小化作为决策树样本集属性选择的标准,划分左、右子树。基尼系数的物理意义是随机选择一个样本,该样本在划分后的子集中被错分的可能性计算如下:

$$G(D)=\sum_{i=0}^{n}[\,p_i\cdot(1-p_i)\,]=1-\sum_{i=0}^{n}p_i^2 \tag{2}$$

式中,D表示样本总体;p_i表示第i类样本占样本总体的比例;n表示总类别数。

比较基于不同特征划分DT得到的基尼系数,选取基尼系数最小的特征Y作为DT划分左、右子树的标准。基于Y特征划分的基尼系数的计算方法如下:

$$G(D,Y)=\frac{\|D^{\text{left}}\|}{\|D\|}\cdot G(D^{\text{left}})+\frac{\|D^{\text{right}}\|}{\|D\|}\cdot G(D^{\text{right}}) \tag{3}$$

式中,D^{left}、D^{right}分别表示划分后的左、右子树样本集。

2)RF算法是集成学习引导聚集算法在DT上的改进版,是常用的多分类算法。其核心思想是通过随机采样数据集、随机选择特征,构建多个独立的CART分类器,通过分类结果投票决定最终分类结果。RF算法是一种经典的装袋算法,对训练样本集进行有放回的随机采样,构成多个不同的样本集,分别用于训练多个相对独立的弱分类器,并通过一系列结合策略融合分类结果,形成强分类器[22]。但RF在构建CART基分类器时,是从样本特征中随机选择m个特征(m小于样本特征总数),并从m个特征中选择一个最优特征用于划分DT左、右子树。随机选择特征的个数m能直接影响模型的偏差和方差,m过小可能会导致模型存在较大偏差,因此通常利用交叉验证的方法选择合适的m,以保证预测模型的泛化能力。

3)kNN算法的核心思想是每个样本都可以用它最接近的k个邻居来代表,kNN通过测量不同特征值之间的距离进行分类[22]。在kNN算法中,用于选择的邻居都是已正确分类的对象,该方法依据样本与其最近邻的k个对象的类别来决定样本所属的类别。参数k的选择对算法结果有重要影响,因此从$k=1$起,重复使用检验集估计分类器的误差率,直到确定最合适的k值,通常k不超过20。

4)SVM算法是一类按监督学习方式对数据进行二元分类的广义线性分类器,其基本想法是求解能够正确划分训练数据集、几何间隔最大的

分离超平面[22]。

利用机器学习方法进行个体出行位置预测的流程如图 2 所示。

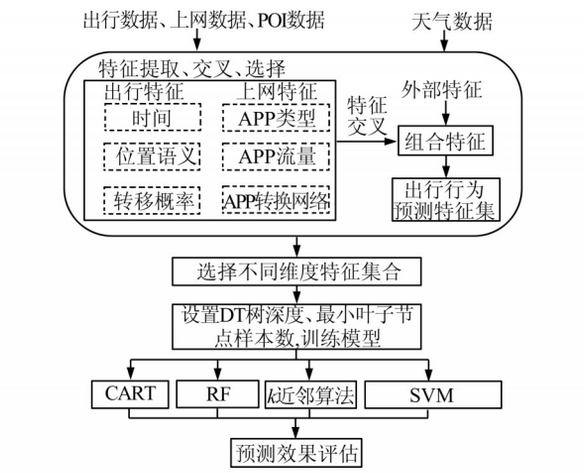

图 2　机器学习预测模型训练流程图

Fig.2　Training Procedure of Machine Learning Prediction Model

1)首先根据手机用户出行数据、上网记录数据,结合 POI 数据、天气数据,提取手机用户的出行特征、上网特征以及外部因素等特征。出行特征包括出行距离、活动半径、轨迹熵、访问位置个数,以及下一时段手机用户历史停留时长的平均值、最大值、最小值、中位数、标准差。然后通过该基站区域的 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,以及手机用户从当前基站转移到其他基站的多个概率中按数值从大到小排序,排在前三的转移概率及其和,用于定量描述手机用户的出行选择。区域功能的多样性 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,分析该区域的功能特性。计算方法如下:

$$E_{\mathrm{POI}}=-\sum(p_{\mathrm{POI}}\cdot\log_2 p_{\mathrm{POI}}) \tag{4}$$

式中,p_{POI} 表示某类型 POI 数量占该区域 POI 总数的比例。

手机用户的上网行为特征包括上网次数、APP 类别数、某类 APP 的使用次数,以及同时使用多个 APP 的次数。其特征组合有 7 个,即不同 APP 同时使用的次数、平均使用次数、使用数据流量,以及数据流量的平均值、最大值、最小值、中位数、标准差。此外,在计算 APP 上网特征时,参考构建 APP 转换关系网络的相关成果[25],构建了相邻时段的 APP 转换关系,以此来作为 APP 上网特征。

外部因素特征包括工作日、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和体感温度的最大、最小值、中位数、平均值等。

2)通过时间、空间维度的特征交叉形成特征集合,利用卡方检验特征选择方法,获取{出行}、{上网行为}、{外部特征}及其特征组合{出行,上网行为}、{外部特征,出行}、{上网行为,外部特征}、{出行,上网行为,外部特征}等 7 个不同的预测特征集。特征交叉是指对手机用户所在的基站进行经纬度坐标去重后,再对区域重新编号,得到 n 个区域组成的区域向量 $\boldsymbol{S}$。1 天中的时段经离散化后划分为 m 个时段,通过笛卡尔积公式可以获取 $m\times n$ 个手机用户在特定时段 t、特定位置 l 的特征,具体计算如下:

$$\begin{cases}T=[t_1,t_2\cdots t_m]\\ \boldsymbol{S}=[l_1,l_2\cdots l_n]\\ T\times\boldsymbol{S}=\begin{bmatrix}t_1l_1 & \cdots & t_ml_1\\ \vdots & \ddots & \vdots\\ t_1l_n & \cdots & t_ml_n\end{bmatrix}\end{cases} \tag{5}$$

式中,T 表示时段向量;$t_i(i=1,2\cdots m)$ 表示特定时段;$\boldsymbol{S}$ 表示位置区域向量;$l_i(i=1,2\cdots n)$ 表示用户所在的基站位置。

经过特征交叉后,就产生了丰富的特征集,但并非所有特征都会对预测有帮助,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,因此本文采用 Filter 方法中常用的卡方检验进行特征选择。其原理是利用构建列联表,计算卡方检验统计值 χ^2,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。具体公式如下:

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-n_i)^2}{n_i} \tag{6}$$

式中,f_i 表示特征属性值 i 的实际样本数;n_i 表示特征属性值 i 出现的期望样本数;k 表示特征的个数。

3)选择不同维度的特征集,设置最大深度、最小叶子节点样本数进行 CART、RF、梯度提升迭代决策树(gradient boosting decision tree,GBDT)分类算法模型进行训练。

4)评估各模型的预测准确率,输出每个测试样本的类别以及各个类别对应的分类概率。

1.3　融合方法

常用的模型结果融合规则主要有投票法、加权融合法和学习法[24]。为了避免多个模型得出的小分类概率结果占多数,忽略分类概率结果的现象,本文提出一种基于频率分布直方图的自适

应模型融合规则，融合不同模型的分类概率，得出最终的预测结果。

对同一个测试样本，不同模型预测结果构成预测结果类别集，对预测结果集中的每个类别的预测概率进行直方图分析，基于直方图加权融合得到该类别的预测概率。图3为某个预测结果在不同模型预测时的预测概率分布图。自适应加权融合规则为：首先根据基础模型对该类别的预测概率进行计算，得到$\{p_j|j=1,2\cdots m\}$，其中，m为基础模型的个数；然后按照该预测概率的最大值、最小值划分k个概率区间，统计预测概率位于区间$[a_i,a_{i+1}]$的频数n_1、$n_2\cdots n_k$，a_i、a_{i+1}分别表示第i个区间下、上限的取值；最后把频数作为权值对区间分类概率均值进行加权，计算融合后模型对该类别的分类概率p。比较预测结果集中各类别的分类概率，选取融合后分类概率最大的类别作为预测结果。a_i、p的计算公式如下：

$$a_i=a_0+\frac{a_k-a_i}{k}i \tag{7}$$

$$p=\frac{\sum_{i=1}^{k}\left(\frac{a_i+a_{i+1}}{2}n_i\right)}{m} \tag{8}$$

式中，$a_0=\min\{p_j|j=1,2\cdots m\}$；$a_k=\max\{p_j|j=1,2\cdots m\}$；$n_i$为第$i$个区间对应的频数，即权值；$m$为基础模型的个数。

基于直方图的多模型融合规则根据分类概率进行自适应的权值设置，计算得到的权重取决于基础模型对样本的分类概率的分布情况，通过这种形式的加权能够一定程度提升高分类概率模型的权重，得到相对准确的模型融合结果。

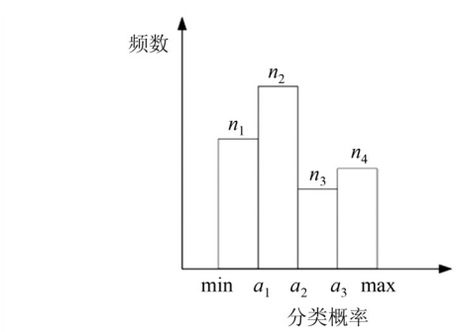

图3 预测类别的分类概率分布

Fig.3 Classification Probability Distribution of Prediction Categories

本文提出的多模型融合的基础框架如图4所示。首先利用训练好的基础预测模型对测试样本集进行预测，获取样本预测结果和预测结果中各类别的分类概率，并统计各模型的预测准确率，如表1所示；然后利用直方图分析基础模型的预测结果和分类概率，应用本文提出的自适应加权融合规则融合基础模型的预测结果并输出。

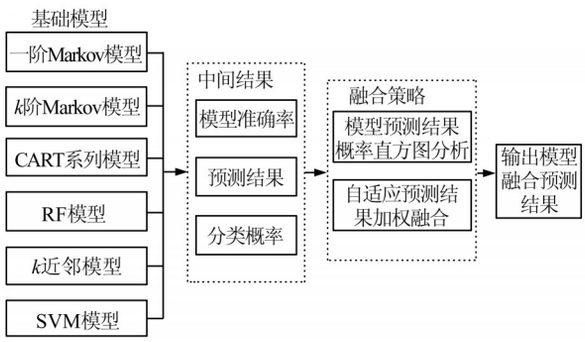

图4 出行预测多模型融合流程

Fig.4 Flowchart of Multiple Travel Prediction Models Fusion

表1 多模型的预测结果

Tab.1 Prediction Results of Multiple Models

模型	类别1	类别2	…	类别n	预测结果
模型1	0.33	0.21	…	0	1
模型2	0.05	0.90	…	0.02	2
模型3	0.25	0.24	…	0	1

2 实验结果与分析

2.1 实验数据

本文采用的实验数据是某城市2015-08-10—2015-08-29共计20 d的手机用户基站位置更新数据和手机上网的流量收费数据，以及研究时间段内该城市的POI数据及天气数据。

手机位置数据是以手机基站的经纬度坐标记录用户的位置，包含的数据字段有用户ID、日期、时间、事件类型、基站编号以及基站经度、纬度等。该数据中包括用户的主动和被动定位信息，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）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1天中至少会产生24条定位数据。

手机上网数据主要记录了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。手机用户使用通用无线分组业务（general packet radio service，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以及基于基站的用户位置信息，因此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、流量大小等。

城市的POI数据是通过百度地图 Place API 接口得到的,包含旅游景点、交通设施、政府机构、休闲娱乐、购物等 17 类,共 4 308 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。天气数据包含每天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。天气数据将作为外部因素特征用于出行位置预测的建模。

由于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,因此本文对这两种数据进行了预处理,包括对数据本身缺失值、重复记录的处理,以及对轨迹层面的基站跳变异常数据的消除。

本文从手机用户的位置更新数据、手机上网数据中筛选出连续 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 8 508 人。

2.2　个体出行位置预测实验结果分析

将每个手机用户前 80% 的数据记录作为训练数据,后 20% 的数据作为测试数据。实验采用预测准确率 C 作为评价标准,对本文构建的手机用户出行位置预测模型进行评估。计算公式如下:

$$C=\frac{N_R}{N} \tag{9}$$

式中,N_R 表示预测结果正确的样本数;N 表示测试集样本数。

个体出行位置预测为多分类问题,在下一时段手机用户出行位置具有多种选择,因此本文统计了 top1、top3 预测准确率。其中,top1 表示预测的第一个结果即为正确结果的概率,top3 表示在模型预测给出的前 3 个结果中包含正确结果的概率。

2.2.1　不同算法预测准确率对比

在提取手机用户的出行特征、上网特征、外部因素的基础上,计算不同时段手机用户在不同基站间的出行转移概率,构建针对手机用户出行位置预测的出行特征集、上网行为特征集。使用 §1.2 中的 7 个特征组合,分别训练 DT、RF、GBDT 算法、kNN 算法、SVM 算法、一阶 Markov 模型、最常访问位置预测(most frequented location model, Most Value)模型等基础模型,同时对基础模型进行多种组合,应用本文提出的多模型融合预测方法进行融合分析。

个体出行位置基础模型预测准确率对比如表 2 所示,其中 CART、RF、GBDT、kNN、SVM 算法给出的是使用不同特征集合所构建的模型的准确率最大值。表 2 结果表明,本文提出的模型预测准确率最高。部分基础模型组合的预测准确率如表 3 所示。

表 2　个体出行位置基础模型预测准确率对比/%

Tab.2　Comparison of Prediction Accuracy of Different Prediction Algorithms/%

基础模型	top1准确率	top3准确率
CART算法	70.56	94.01
RF算法	69.82	87.69
kNN算法	63.30	87.84
SVM算法	57.52	86.54
一阶Markov模型	56.84	91.49
GBDT算法	72.80	92.77
Most Value模型	51.29	55.63

表 3　个体出行位置模型组合融合预测准确率对比/%

Tab.3　Comparison of Prediction Accuracy of Different Combined Prediction Algorithms/%

组合模型	top1准确率	top3准确率
Markov模型、DT、SVM	74.14	93.65
Markov模型、DT、kNN	73.97	92.92
Markov模型、kNN、SVM	68.16	91.59
Markov模型、SVM、RF	73.35	93.66
Markov模型、kNN、SVM、RF	72.22	93.79
Markov模型、DT、kNN、SVM	73.54	94.13
本文模型	74.59	94.19

表 4 给出了本文提出的融合模型与投票法融合策略对比结果。从表 4 可以看出,本文基于直方图的融合策略的 top1、top3 准确率分别为 74.59%、94.19%,相比投票法融合策略的 top1、top3 准确率分别提升 1.69%、3.61%,可见手机用户下一时段位置预测的 top3 准确率相比 top1 有大幅提升,平均预测准确率达到 94.19%,为提供更好的基于位置的服务打下基础。将表 4 与表 1 对比可知,基于直方图的多模型预测方法比准确率最高的基础模型的 top1、top3 准确率分别提升了 1.79%、0.18%。

表 4　本文融合模型与投票法融合策略预测准确率对比/%

Tab.4　Comparison of Prediction Accuracy Between Our Proposed Method and the Vote Strategy/%

融合方法	top1准确率	top3准确率
投票法融合	72.90	90.58
本文模型	74.59	94.19

2.2.2 不同时间粒度预测准确率对比

实验分别在未来10 min、20 min、30 min等3个时间粒度下对手机用户下一位置进行预测，实验结果对比如表5所示。由表5可知，随着时间粒度的增大，手机用户个体出行位置预测准确率逐步升高，预测时间粒度达到30 min后，top1准确率达到74.59%。随着预测时间粒度的增大，轨迹数据量大大增加，综合考虑预测效果和数据处理效率，以30 min为时间粒度的预测效果最佳。

表5 不同时间粒度预测准确率对比/%

Tab.5 Comparison of Prediction Accuracy Under Different Temporal Granularities/%

时间粒度/min	top1准确率	top3准确率
10	69.80	92.84
20	71.50	94.27
30	74.59	94.19

此外，本文还对不同时段手机用户个体位置的预测结果进行分析，以07:00手机用户的位置预测为例，即以10 min为时间粒度进行位置预测时，就使用06:50的个体位置结合历史数据对手机用户07:00的位置进行预测，其他类推。图5给出了对个体07:00—21:00的位置预测的top1准确率。从图5可以看出：(1)不同时间粒度下，不同时段个体出行位置预测准确率的变化存在相似之处，早上的预测准确率比下午高；(2)以30 min为时间粒度时，除了12:00、15:00和17:00以外，其他时段预测准确率均高于其他时间粒度；(3)在17:00，10 min和20 min的时间粒度较30 min的时间粒度预测效果更好，可能在出行频繁时段需要进行更加精细的出行行为刻画。整体上，本文提出的基于直方图的多模型融合个体位置预测模型对未来30 min手机用户位置的预测最佳，预测top1准确率达到74.59%，取得了不错的预测效果。

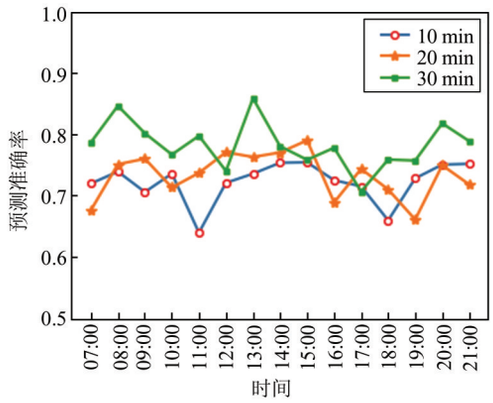

图5 个体位置预测top1准确率对比

Fig.5 Comparison of Individual Position Prediction top1 Accuracy

3 结 语

本文融合了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，对手机用户个体出行构建了基于Markov与多类机器学习模型融合的个体出行位置预测模型，旨在为手机用户提供更好的基于位置的服务。实验结果表明，本文模型的top1准确率、top3准确率分别为74.59%、94.19%，相比准确率最高的基础模型分别提高了1.79%、0.18%，相比投票法融合规则准确率分别提升1.69%、3.61%。此外，本文还在多个时间粒度下，对个体位置预测的准确率进行了对比，发现以30 min为预测时间粒度时预测效果较好。

参 考 文 献

[1] Xiao Y, Wang B, Liu Y, et al. Analyzing, Modeling, and Simulation for Human Dynamics in Social Network [J]. *Abstract and Applied Analysis*, 2012, (6 684):552-582

[2] Croitoru A, Wayant N, Crooks A, et al. Linking Cyber and Physical Spaces Through Community Detection and Clustering in Social Media Feeds [J]. *Computers, Environment and Urban Systems*, 2015, 53:47-64

[3] Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding Individual Human Mobility Patterns [J]. *Nature*, 2018, 453(7 196):779-782

[4] Ahas R, Aasa A, Silm S, et al. Daily Rhythms of Suburban Commuters' Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data [J]. *Transportation Research Part C*, 2010, 18(1):45-54

[5] Zhou Tao, Han Xiaopu, Yan Xiaoyong, et al. Statistical Mechanics on Temporal and Spatial Activities of Human [J]. *Journal of University of Electronic Science and Technology of China*, 2013, 42(2): 481-540(周涛，韩筱璞，闫小勇，等. 人类行为时空特性的统计力学[J]. 电子科技大学学报，2013, 42(2):481-540)

[6] Shaw Shihlun, Fang Zhixiang. Rethinking Human Behavior Research from the Perspective of Space-time GIS[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6):667-670(萧世伦，方志祥. 从时空GIS视野来定量分析人类行为的思考[J]. 武汉大学学报·信息科学版，2014, 39(6):

667-670)

[7] Fan Y, Khattak A J. Urban Form, Individual Spatial Footprints, and Travel: Examination of Space-Use Behavior[J]. *Transportation Research Record Journal of the Transportation Research Board*, 2008, 2 082:98-106

[8] Xu Y, Shaw S L, Zhao Z, et al. Another Tale of Two Cities-Understanding Human Activity Space Using Actively Tracked Cellphone Location Data [J]. *Annals of the Association of American Geographers*, 2016, 106(2):489-502

[9] Chen B Y, Wang Y, Wang D, et al. Understanding the Impacts of Human Mobility on Accessibility Using Massive Mobile Phone Tracking Data [J]. *Annals of the American Association of Geographers*, 2018, 108(4):1-19

[10] Kang Chaogui, Liu Yu, Wu Lun. An Analysis of Entropy of Human Mobility from Mobile Phone Data [J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(1):63-69(康朝贵, 刘瑜, 邬伦. 城市手机用户移动轨迹时空熵特征分析[J]. 武汉大学学报·信息科学版, 2017, 42(1): 63-69)

[11] Yang Xiping, Fang Zhixiang, Zhao Zhiyuan, et al. Analyzing Space-Time Variation of Urban Human Stay Using Kernel Density Estimation by Considering Spatial Distribution of Mobile Phone Towers [J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(1):49-55(杨喜平, 方志祥, 赵志远,等. 顾及手机基站分布的核密度估计城市人群时空停留分布[J]. 武汉大学学报·信息科学版, 2017, 42(1): 49-55)

[12] Zhang C, Han J, Shou L, et al. Splitter: Mining Fine-grained Sequential Patterns in Semantic Trajectories[J]. *Proceedings of the VLDB Endowment*, 2014, 7(9):769-780

[13] Hou J, Zhao H, Zhao X, et al. Predicting Mobile Users' Behaviors and Locations Using Dynamic Bayesian Networks [J]. *Journal of Management Analytics*, 2016, 3(3):191-205

[14] 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

[15] Song C, Qu Z, Blumm N, et al. Limits of Predictability in Human Mobility [J]. *Science*, 2010, 327 (5 968):1 018-1 021

[16] Yan X Y, Wang W X, Gao Z Y, et al. Universal Model of Individual and Population Mobility on Diverse Spatial Scales[J]. *Nature Communications*, 2017, 8(1): 1 639-1 648

[17] Ozer M, Keles I, Toroslu H, et al. Predicting the Location and Time of Mobile Phone Users by Using Sequential Pattern Mining Techniques [J]. *The Computer Journal*, 2016, 59(6): 908-922

[18] Qiao Y, Zhao X, Yang J, et al. Mobile Big-Data-Driven Rating Framework: Measuring the Relationship Between Human Mobility and APP Usage Behavior[J]. *IEEE Network*, 2016, 30(3):14-21

[19] 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):153-174

[20] Do T M T, Gatica-Perez D. Where and What: Using Smartphones to Predict Next Locations and Applications in Daily Life[J]. *Pervasive and Mobile Computing*, 2014, 12:79-91

[21] Huang Q. Mining Online Footprints to Predict User's Next Location [J]. *International Journal of Geographical Information Systems*, 2017, 31 (3): 523-541

[22] Zhou Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press, 2016(周志华. 机器学习[M]. 北京:清华大学出版社, 2016)

[23] Fang Zhixiang, Ni Yaqian, Zhang Tao, et al. Using Terminal Location Spatio-Temporal Transfer Probability to Predict Subscriber Base Size of Communication Base Station[J]. *Journal of Geo-information Science*, 2017, 19(6):772-781(方志祥,倪雅倩,张韬,等. 利用终端位置时空转移概率预测通讯基站服务用户规模[J]. 地球信息科学学报, 2017, 19(6):772-781)

[24] Sun Juan. Fuzzy Decision Tree Induction Based on Optimization of Parameters [J]. *Computer Engineering and Applications*, 2012, 48(23):148-154(孙娟. 智能参数学习的模糊决策树算法[J]. 计算机工程与应用, 2012, 48(23):148-154)

[25] Fang Zhixiang, Yu Chong, Zhang Tao, et al. A Mixed Markov Method to Predict the Surfing Time Period of Mobile Phone Users[J]. *Journal of Geo-information Science*, 2017, 19(8):1 019-1 025(方志祥,于冲,张韬,等. 手机用户上网时段的混合 Markov 预测方法[J]. 地球信息科学学报, 2017, 19(8):1 019-1 025)

A Multi-model Fusion Model of Individual Travel Location Prediction Using Markov and Machine Learning Methods

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: **Objectives**: With the development of urbanization, people's travel behaviors have diversified. An in-depth understanding of human behavior and the modeling and prediction of individual travel behaviors are helpful in explaining several complex socio-economic phenomena, and are important in offering location-based services, transportation planning, and public safety. Individual travel behavior prediction is based on a deep understanding of human activity characteristics. In the era of mobile Internet, the online behavior of cyberspace is inseparable from the travel behavior of real space. **Methods**: This paper integrates individuals' mobile phone tracking data and Internet traffic data, and constructs a multi-model fusion model of individual travel location prediction on Markov and machine learning methods. Considering the classification probability of prediction results, an adaptive fusion strategy based on frequency distribution graph is proposed. The prediction results of Markov model and machine learning multi-classification model are merged together to obtain the final mobile phone user travel location prediction result. **Results**: This paper performs individual travel location prediction experiments on the basis of multi-source data. And the experiments show that the correct rate of the first result and the top three results of the multi-model fusion location prediction model based on histogram is respectively 74.59% and 94.19%, higher than the prediction accuracy of the basic model with the highest accuracy and the vote strategy. **Conclusions**: Under the prediction time granularity of 30 minutes, the individual travel location prediction is better.

Key words: Markov model; machine learning method; travel location prediction; mobile phone location data; feature fusion

First author: FANG Zhixiang, PhD, professor, specializes in space-time GIS, spatiotemporal modeling of urban big data and pedestrian navigation. E-mail:zxfang@whu.edu.cn

Foundation support: The National Natural Science Foundation of China(41771473).

引文格式: FANG Zhixiang, NI Yaqian, HUANG Shouqian. A Multi-model Fusion Model of Individual Travel Location Prediction Using Markov and Machine Learning Methods[J]. Geomatics and Information Science of Wuhan University, 2021, 46(6):799-806. DOI:10.13203/j.whugis20190404(方志祥, 倪雅倩, 黄守倩. 融合Markov与多类机器学习模型的个体出行位置预测模型[J]. 武汉大学学报·信息科学版, 2021, 46(6):799-806. DOI:10.13203/j.whugis20190404)