



研究与开发

基于手机上网记录数据的个体相遇预测

李倩, 江昊, 杨锦涛

(武汉大学电子信息学院, 湖北 武汉 430072)

摘要: 随着反映个体位置信息的数据大量涌现, 人类移动行为的研究引起了业界学者的广泛关注。针对个体相遇预测问题, 使用用户手机终端上网产生的会话数据进行分析。首先基于用户相遇关系进行复杂网络建模, 然后分析了网络拓扑结构特征, 同时引入了用户移动性特征和上网行为特征, 构造基于随机森林的预测模型, 进行个体相遇预测。实验结果表明, 相比于传统的网络拓扑结构特征, 通过引入用户移动性特征和上网行为特征, 能够显著提升预测性能。

关键词: 相遇预测; 移动互联网; 复杂网络; 随机森林

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017249

Individual encounter prediction based on mobile internet record data

LI Qian, JIANG Hao, YANG Jintao

School of Electronics and Information, Wuhan University, Wuhan 430072, China

Abstract: Studies on human movement behavior have drawn much attention with the availability of unprecedented amount of records with high accuracy involving individuals' trajectories. The encounter prediction problem based on the session data generated by users' mobile terminal was studied when users accessed the internet for data usage. Firstly, the network based on the encounter relations between users was constructed. Secondly, the network topology features were analyzed and user mobility characteristics and user internet behavior characteristics were introduced. Finally, the prediction model based on random forest was applied. The experimental results show that compared with the traditional network topology features, the prediction performance can be significantly improved by introducing the user mobility characteristics and user internet behavior characteristics.

Key words: encounter prediction, mobile internet, complex network, random forest

收稿日期: 2017-04-01; 修回日期: 2017-08-18

通讯作者: 江昊, jh@whu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61371126); 国家高技术研究发展计划 (“863” 计划) 基金资助项目 (No.2014AA01A707); 武汉应用基础研究计划资助项目 (No.2014010101010026)

Foundation Items: The National Natural Science Foundation of China(No.61371126), The National High Technology Research and Development Program of China(863 Program)(No.2014AA01A707), The Applied Basic Research Programs of Wuhan (No.2014010101010026)



1 引言

近年来,随着移动互联网技术的飞速发展以及诸如智能手机、平板电脑、笔记本电脑等终端设备的普及,移动互联网用户呈现出爆发式的增长^[1,2]。中国互联网络信息中心(CNNIC)发布的最新报告^[3]表明,截止到2016年12月,网民总数高达7.31亿人,其中手机网民总数为6.95亿人。由于手机的便捷性和移动互联网的高覆盖率,用户能够随时随地使用手机终端上网,与此同时手机用户在接入网络过程产生了大量会话数据,这些数据准确记录了用户的时间和地理位置信息,能够真实反映用户的移动轨迹,为人类移动行为的研究提供了十分丰富的数据源。

随着反映个体位置信息的数据大量涌现,人类时空轨迹信息的挖掘引起了业界学者的广泛关注,其中基于历史轨迹数据的位置预测问题一直是学者们不断研究的热点。Song等人^[4]通过计算个体历史轨迹序列的随机熵、访问量相关熵、时序相关熵,得出了人类移动行为能够达到93%可预测性的结论。Lu等人^[5]分别利用0~7阶的马尔可夫链预测个体轨迹,结果表明对于非静态轨迹数据预测准确率高达95%,其中1阶马尔可夫链的预测效果最佳,这说明人类行为高度依赖于历史行为记录。An等人^[6]采用高斯混合模型来预测手机用户的移动轨迹,通过期望最大方法(EM)来训练参数。Noulas等人^[7]利用历史访问位置、移动距离等特征来构建预测模型,采用了两种监督学习模型(线性回归(linear regression)和M5树(M5 tree))来预测个体的移动轨迹。Monreale等人^[8]使用一种基于关联规则名为T模式树的决策树作为预测模型,通过从历史轨迹数据中挖掘出个体轨迹模式作为该决策树的判别准则。参考文献[9-11]分别提出了基于概率模型、轨迹序列模式挖掘、通信数据统计来预测个体轨迹的方法。参考文献[12,13]引入了社交特征来进行移动轨迹预测,能够很大程度地提升个体位

置预测的准确率。

目前个体相遇行为预测仍然是一个比较新颖的问题,大多数学者主要是针对单个个体轨迹进行预测,在个体间相遇问题方面,有学者通过链路预测的方法来预测个体之间的相遇行为。Sun等人^[14]利用公交智能卡数据,研究个体的日常相遇模式,发现了相遇有可复制的周期模式,并且频繁相遇的个体对在不同的社区中有较强的连接。Socievole等人^[15]利用在线社交关系来预测个体间的接触行为,并将该问题看作个体形成的接触图演化问题。Scholz等人^[16]通过构建个体之间形成的社交网络,利用链路预测方法来预测网络的演化趋势和个体对在未来的接触行为。本文基于真实的手机用户上网记录数据,通过复杂网络建模,分析了网络拓扑结构特征、用户移动性特征和用户上网行为特征,构造了基于随机森林的预测模型,并提出了个体相遇预测的算法,针对个体对在将来某个时段的相遇进行预测,从而揭示人与人之间在时空上的关联程度。本文的方法为个体相遇行为预测研究提供了一种新思路。本文的主要贡献如下。

- 提出了一种个体相遇预测算法。首先通过复杂网络建模,利用网络拓扑结构描述用户个体之间的相遇关系,然后在传统网络拓扑结构特征的基础上,分析了用户移动性特征和上网行为特征,构造基于随机森林的预测模型。
- 基于我国某城市真实的手机用户上网记录数据,利用个体相遇预测算法,针对用户个体对在未来某个时段的相遇行为进行预测,实验结果表明本文提出的预测算法能够取得良好的预测效果,并且相比于传统的网络拓扑结构特征,引入用户移动性特征和上网行为特征后,可以显著提升预测性能。

2 个体相遇预测算法

2.1 基于用户相遇关系的复杂网络建模

将用户之间的相遇关系看作图,图中的节点

为用户集合, 边代表用户对之间曾经有过相遇行为, 由此定义无向图 $G = (V, E)$, 其中, V 为节点集, E 为边集, 假设节点 $u, v \in E$, 若 $(u, v) \in E$, 则表示用户 u 和 v 之间有过相遇行为; 若 $(u, v) \notin E$, 则表示用户 u 和 v 没有相遇行为。本文将用户之间的相遇行为定义为用户对在一天中的同一个时段出现在同一个基站下, 使用基站经纬度代表用户个体移动位置。

本文分时段构建用户相遇关系的复杂网络, 通过针对每一个时段构建网络, 将用户个体的相遇预测问题, 转化为对图中连边的预测问题。通过构建有监督的分类学习模型, 利用前若干天某个时段的网络作为训练集, 对分类模型进行训练, 能够预测出用户对在后一天的该时段是否相遇。

2.2 网络拓扑结构特征

基于用户相遇关系的复杂网络, 可以利用传统的拓扑结构特征, 衡量网络中两个用户节点的相似度。本文选取了 10 种经典的网络结构特征^[17], 作为用户节点对的相似性指标, 具体见表 1。

针对网络中的用户节点 u , 定义 u 的邻居节点集为 $\Gamma(u)$, u 的度为 $k(u) = |\Gamma(u)|$ 。

2.3 用户移动性和上网行为特征

人类移动行为受各种各样的因素影响, 仅网络结构特征不能充分反映个体之间的关系, 需要提取更多与用户行为相关的特征, 才能进一步挖掘个体的行为规律, 因此本文引入了用户移动性和上网行为特征。

2.3.1 用户移动性特征

(1) 热点位置距离

用户个体由于人类行为的内在驱动性, 每天经常会访问多个位置, 其中诸如家庭所在地、工作地等热点位置的访问频率显著高于其他位置。同时, 参考文献[18]揭示了个体的移动距离近似服从幂律分布, 用户的空间活动范围是有限的, 用户大部分时间都在一定的活动半径内移动。由此, 本文选取用户轨迹序列中访问频率最高的位置作为用户个体的热点位置, 利用两个用户热点位置的距离, 表征用户对之间的相似度。

表 1 网络结构特征

网络结构特征	计算式	基本思想
共同邻居 CN 指标	$CN(u, v) = \Gamma(u) \cap \Gamma(v) $	若 u 和 v 的共同邻居数量越多, 则两者有连边的可能性越大
Salton 指标	$Salton(u, v) = \frac{ \Gamma(u) \cap \Gamma(v) }{\sqrt{k(u) \times k(v)}}$	在 CN 的基础上, 考虑了两个节点度的乘积对相似性的作用
Jaccard 指标	$Jaccard(u, v) = \frac{ \Gamma(u) \cap \Gamma(v) }{ \Gamma(u) \cup \Gamma(v) }$	在 u 或 v 的邻居节点集里, 随机选择一个节点, 它是 u 和 v 的共同邻居的概率
Sorenson 指标	$Sorenson(u, v) = \frac{2 \Gamma(u) \cap \Gamma(v) }{k(u) + k(v)}$	从 CN 指标发展而来的, 它加入考虑了 u 和 v 的度之和对相似性的影响
大度节点有利指标 HPI	$HPI(u, v) = \frac{ \Gamma(u) \cap \Gamma(v) }{\min\{k(u), k(v)\}}$	HPI 指标的分母为两个节点度中较小的值
大度节点不利指标 HDI	$HDI(u, v) = \frac{ \Gamma(u) \cap \Gamma(v) }{\max\{k(u), k(v)\}}$	选取两个节点度中的较大值作为分母, 用来抑制大度节点对相似性的影响
LHN-I 指标	$LHN-I(u, v) = \frac{ \Gamma(u) \cap \Gamma(v) }{k(u) \times k(v)}$	若 u 和 v 拥有共同邻居数越多, 两者之间的相似性就越高, 同时纳入考虑了节点度乘积
优先链接指标 PA	$PA(u, v) = k(u) \times k(v)$	只考虑两个节点的度, 节点 u 和 v 之间有连边的可能性与两个节点度的乘积成正比
Adamic-Adar 指标	$AA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\lg k(z)}$	共同邻居节点的度越小, 它对相似度的作用越大, 由此将度的对数取倒数
资源分配指标 RA	$RA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{k(z)}$	从网络资源分配的角度提出的, 假设有若干资源从 u 传送到 v , 两者的共同邻居当作传送的媒介, 并将资源均匀分发给邻居节点, 则 u 和 v 的相似度定义为 v 接收的资源数



设用户个体 u 的移动位置序列为 $Loc(u) = (L_1, L_2, L_3, \dots, L_{n(u)})$, 其中, $n(u)$ 为用户 u 移动位置的个数, 则个体 u 对位置 L 的访问频率为:

$$P(u, L) = \frac{\sum_{i=1}^{n(u)} \delta(L, Loc_i(u))}{n(u)} \quad (1)$$

用户个体 u 访问频率最高的位置为:

$$HotSpot(u) = \arg \max_{L \in Loc} P(u, L) \quad (2)$$

用户 u 和 v 的热点位置之间的空间距离表示为:

$$Dist(u, v) = R \cdot \arccos(\cos(lat_u) \cos(lat_v) \cos(lon_v - lon_u) + \sin(lat_u) \sin(lat_v)) \quad (3)$$

其中, (lon_u, lat_u) 、 (lon_v, lat_v) 分别为用户 u 和 v 的热点位置的经纬度坐标, R 是地球的半径。

(2) 空间相似度

余弦相似度 (cosine similarity)、皮尔逊相似度 (pearson similarity)、KL 散度 (Kullback-Leibler divergence, KLD) 和 JS 散度 (Jensen-Shannon divergence, JSD) 是人类移动行为研究中的常用指标, 用来表征用户位置序列的相似度和差异性。假设用户 u 和 v 访问基站的频次序列分别为 $u = \langle k_1, k_2, \dots, k_i, \dots, k_N \rangle$, $v = \langle l_1, l_2, \dots, l_i, \dots, l_N \rangle$, 其中, N 为用户 u 和 v 访问基站的总数, 则计算式分别如下:

$$cosine = \frac{\sum_i k_i l_i}{\sqrt{\sum_i k_i^2} \sqrt{\sum_i l_i^2}} \quad (4)$$

$$pearson = \frac{\sum_i (k_i - \bar{k})(l_i - \bar{l})}{\sqrt{\sum_i (k_i - \bar{k})^2} \sqrt{\sum_i (l_i - \bar{l})^2}} \quad (5)$$

$$KLD(K \| L) = \sum_i k_i \lg \frac{k_i}{l_i} \quad (6)$$

$$JSD(K \| L) = \frac{1}{2} KLD(K \| S) + \frac{1}{2} KLD(L \| S) \quad (7)$$

其中, $\bar{k}$ 和 $\bar{l}$ 分别为 k_i 和 l_i 的均值, 并且 $S = \frac{1}{2}(K + L)$ 。

(3) 用户相遇次数

人类移动行为通常会呈现群体性活动规律, 比如用户一般都会在早上从家里出发去往工作地上班, 晚上下班后从工作地返回到家里, 用户个体每天会访问多个位置, 可能在不同的位置与多个其他用户个体有相遇行为。一些用户对可能会多次相遇, 比如在同一个人工作单位的同事, 或住在同一个小区的住户, 而另一些用户对可能只是在去往某地的路途中偶然相遇。由此, 本文引入用户相遇次数来表征个体对的相似程度。

参考文献[19]将第 2.2 节中 10 种经典的网络结构特征, 在 6 个经典的数据集上进行实验, 并比较其预测精确度, 发现 CN、AA 和 RA 这 3 种指标的表现最好, 因此在分析用户相遇次数和上网行为特征中选取了这 3 种指标。在经典指标 CN、AA 和 RA 的基础上, 参考文献[20]将 3 种指标扩展为含权的形式, 利用网络中边的权重信息, 进一步提升预测准确率。受此启发, 本文基于上第 2.1 节中构建的复杂网络, 以用户节点对之间的相遇次数作为权重, 得到 3 种基于用户相遇次数的相似性指标, 计算式如下:

$$Encounter_WCN(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} e(u, z) + e(z, v) \quad (8)$$

$$Encounter_WAA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{e(u, z) + e(z, v)}{\lg(1 + s(z))} \quad (9)$$

$$Encounter_WRA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{e(u, z) + e(z, v)}{s(z)} \quad (10)$$

其中, $e(u, v)$ 为用户节点 u 和 v 的相遇次数,

$$s(u) = \sum_{z \in \Gamma(u)} e(u, z)。$$

2.3.2 用户上网行为特征

由于手机移动终端的便捷性, 用户每天会在多个位置接入网络进行上网, 在一些诸如家庭、

工作地等用户热点位置, 用户的停留时间显著高于其他位置, 因此用户在这些热点位置的上网时长、消耗的流量也会明显较高。如果用户对在同一个时段出现在同一个位置, 并且在该时段、该位置的上网时长和消耗的流量越大, 说明这个位置很有可能是用户经常有规律性访问的热点位置, 用户对下次在该位置相遇的概率越高。因此, 本文引入了基于上网时长和上网流量特征, 计算式分别如下。

(1) 用户上网时长

$$Dur_WCN(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} d(u, z) + d(z, v) \quad (11)$$

$$Dur_WAA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{d(u, z) + d(z, v)}{\lg(1 + s(z))} \quad (12)$$

$$Dur_WRA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{d(u, z) + d(z, v)}{s(z)} \quad (13)$$

其中, $d(u, v)$ 为用户节点 u 和 v 在同一时段同一位置的上网时长总和, $s(u) = \sum_{z \in \Gamma(u)} d(u, z)$ 。

(2) 用户上网流量

$$Flow_WCN(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} f(u, z) + f(z, v) \quad (14)$$

$$Flow_WAA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{f(u, z) + f(z, v)}{\lg(1 + s(z))} \quad (15)$$

$$Flow_WRA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{f(u, z) + f(z, v)}{s(z)} \quad (16)$$

其中, $f(u, v)$ 为用户节点 u 和 v 在同一时段、

同一位置的上网流量总和, $s(u) = \sum_{z \in \Gamma(u)} f(u, z)$ 。

2.4 基于随机森林的预测模型

本文采用随机森林(random forest, RF)^[21,22]算法进行个体相遇预测, 如图1所示, 模型利用分类回归树(classification and regression tree, CART)作为元分类器, CART算法利用基尼指数最小化准则进行特征选择, 由多个元分类器(CART决策树)组成随机森林, 记为 $\{h(X, \theta_k)\}$, 并且 $\{\theta_k\}$ 为独立同分布的随机向量, 在训练集抽样时利用 Bagging 算法的无放回抽样, 同时通过随机子空间方法抽取部分特征进行训练, 最后由训练出的 k 个决策树投票确定分类结果, 最终分类决策为:

$$H(x) = \arg \max_Y \sum_{i=1}^k I(h_i(x) = Y) \quad (17)$$

其中, $H(x)$ 为组合分类模型, h_i 为单个决策树, Y 为目标变量, I 为示性函数。

在传统的链路预测模型中, 只有网络拓扑结构特征作为输入变量, 通过本文的分析可知, 用户移动性特征和上网行为特征对分类决策也有十分重要的影响, 因此通过引入用户移动性特征和上网行为特征, 旨在提高随机森林模型的预测性能。本文选取了上述的10个网络结构特征、8个用户移动性特征和6个用户上网行为特征, 作为预测模型的输入变量。

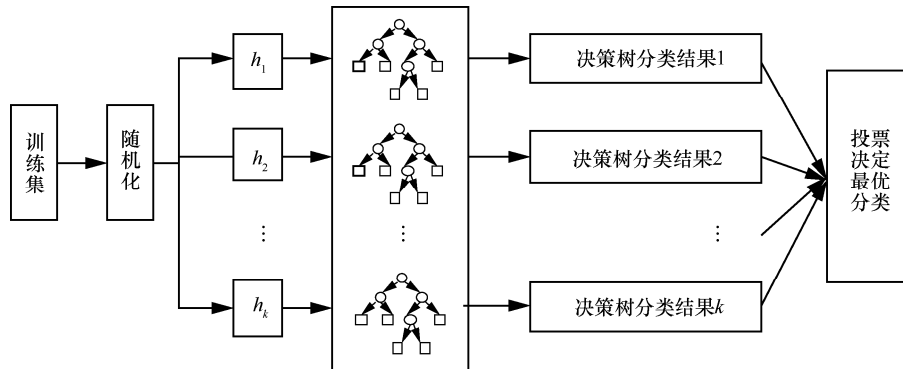


图1 随机森林模型



3 实验结果与分析

本文的手机用户上网记录数据取自我国某省某市,由当地的移动通信运营商提供。对数据进行预处理后,数据规模见表2。

表2 数据集规模

记录数据	规模
记录数/个	42 023 038
用户数/个	1 614 291
基站数/个	10 787
数据集起止时间	2014年11月21日—12月14日

当用户通过手机访问2G或3G网络时,数据业务结束后,用户接入移动互联网产生的相关信息会被记录,主要包括加密后的用户ID、数据流起始时刻及持续时长、数据流的上行和下行流量、用户当前接入的基站经纬度等。如果用户对在一天中的同一个时段出现在同一个基站下,则认为两个用户发生相遇。数据集字段及含义见表3。

表3 数据集字段及含义

字段名	含义
Uid	加密后的用户ID
Longitude	经度
Latitude	纬度
Stime	数据流起始时间
Duration	数据流持续时间/s
Download	下行流量/MB
Upload	上行流量/MB

本文预测算法的总体流程如图2所示,首先基于用户手机上网记录数据,分时段进行复杂网络建模,构建用户相遇关系的复杂网络;然后在传统的网络拓扑特征的基础上,引入了个体移动性和上网行为特征;最后再通过训练集数据,构建随机森林分类模型,基于前若干天某个时段用户关系形成的网络,利用提取出的网络特征和个体行为特征,预测个体对在后一天该时段是否相遇。

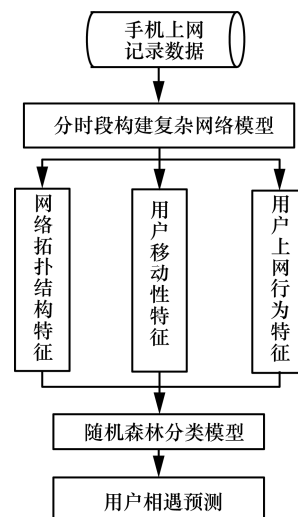


图2 用户相遇预测总体流程

由于数据集规模庞大,本文选取2014年12月2—6日共5天的数据进行分析,利用2014年12月2—5日4天的记录作为训练集,12月6日1天的数据作为测试集,针对0—23时24个时段分别进行用户相遇预测,下面以10时实验为例,进行分析说明。

本文筛选了上网次数大于10次的活跃用户,针对活跃用户进行相遇预测。首先,利用前3天的数据,构建用户相遇关系的复杂网络 G_{train} ,再通过第4天节点之间是否产生连边,进行类标记;然后用第2—4天的数据,构建网络 G_{test} ,对第5天哪些用户节点会产生连边,即发生相遇行为进行预测。训练集和测试集示意如图3所示。

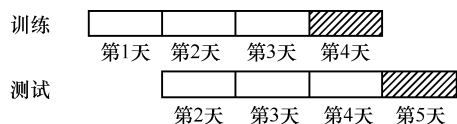


图3 训练集和测试集示意

用户相遇关系形成的复杂网络规模见表4。由于一些特征指标只对全连通的网络有效,因此选取了网络的最大连通子图。由表4可知,测试集网络的最大连通子图中,用户节点数 $N=2\,159$,边数 $M=4\,899$,则潜在可能会产生连边,即可能会在下一时期相遇的用户对有 $C_N^2 = \frac{N \times (N-1)}{2} = 2\,329\,561$,然而实际的新连边数仅有2 427。为了避免盲目地对任意两个

用户都进行相似特征计算,减小程序计算量,提升算法效率,需要缩小候选集范围,针对相遇概率更大的用户对进行相遇预测。

表4 用户相遇关系形成的复杂网络规模

网络	节点数/个	边数/条	群聚系数
G_train	6 039	9 086	0.419
G_train 最大连通子图	2 368	5 269	0.533
G_test	6 233	9 399	0.423
G_test 最大连通子图	2 159	4 899	0.547

最短路径长度概率密度分布如图4所示。图4计算了下一时期中存在连边的用户节点对,在上一时期网络中的最短路径长度的概率密度分布。由图4可知,下一时期产生连边的节点对中,在上一时期网络中的最短路径长度为1和2的占了55.827%的大部分比例,并且随着路径长度的增大,所占比例越小。因此,本文只针对最短路径长度为1和2的用户节点对,进行下一时期的相遇预测。

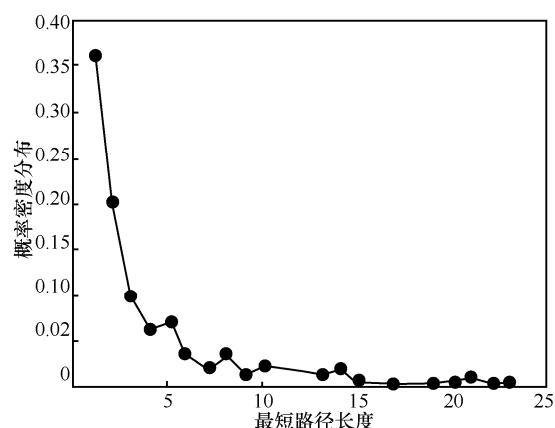


图4 最短路径长度概率密度分布

然后,基于构建的复杂网络,不仅提取了传统的网络结构特征,另外还引入了用户移动性特征和上网行为特征,提取的特征总数为24个,特征集合见表5。

表5 特征集合

特征名称	特征数/个
网络结构特征	10
用户移动性特征	8
用户上网行为特征	6

本文用来衡量预测算法的评价指标为AUC,它能够从整体上衡量算法的精确度^[17],计算式如下:

$$AUC = \frac{n' + 0.5n''}{n} \quad (18)$$

每次在测试集里面随机取出一条边与随机取出的不存在的边进行比较,如果测试集里的边的分数值大于不存在的边的分数值,就得1分;如果两者相等,就得0.5分。独立地比较 n 次,其中上述第一种情况出现了 n' 次,两分数值相等的情况出现了 n'' 次。

利用基于随机森林的预测模型,进行用户相遇预测的AUC结果如图5所示,横坐标为0—23时段,纵坐标为该时段对应的预测算法AUC值。图中对只用网络结构特征和利用所有特征的算法结果进行了对比,由图5可知,通过引入用户移动性特征和上网行为特征后,预测算法的AUC值明显提高,各个时段的AUC值稳定在0.75附近变化波动。实验结果证明了本文所提算法的有效性,说明该算法能够较准确地预测用户相遇行为。此外,选用了参考文献[23]中的常用非监督算法进行对比实验,该算法基于局部信息即利用共同邻居指标进行预测,由于只考虑了单一的基于局部结构信息的相似性指标,算法的精确度较低。

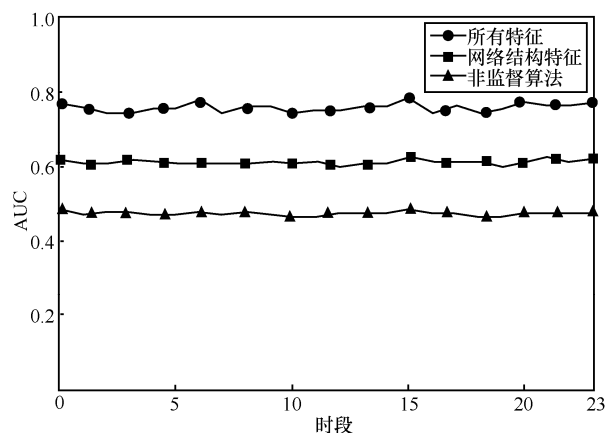


图5 预测算法精确度AUC



随机森林模型中各个特征的重要性程度如图 6 所示。基于用户上网时长和上网流量指标 Flow_WRA、Flow_WCN、Dur_WAA、Flow_WAA、Dur_WRA、Dur_WCN 和用户空间相似度 KLD 等指标对个体对是否有相遇行为的影响较大, 本文引入的用户移动性特征和上网行为特征的重要性程度都排在前列, 明显高于传统的网络结构性特征, 说明这些特征对用户的相遇行为都有较大的影响力, 因此通过引入这些特征, 可以在一定程度上帮助提升预测性能。

4 结束语

本文主要研究了个体相遇预测问题, 并利用现网实测的用户手机上网产生的会话数据, 提出了一种用户相遇预测算法, 首先基于用户相遇关系, 进行复杂网络建模, 然后提取网络结构特征, 同时针对用户的移动行为和上网行为, 引入了用户移动性特征和上网行为特征, 最后构建基于随机森林的预测模型。本文将该算法应用在我国某市真实的用户手机上网数据集上, 基于构建的用户相遇关系复杂网络, 利用提取出的网络结构特征和用户行为特征, 对用户在未来某时段的相遇

行为进行预测, 取得了较好的预测效果。并且与只用网络结构特征和利用所有特征算法做对比, 实验结果说明, 引入的用户移动性特征和上网行为特征能够在一定程度上提高算法的预测性能。本文对用户相遇预测算法的分析, 对人类移动规律和行为模式的研究有着重要的意义。

未来的研究方向是基于用户移动行为的相似度, 利用聚类算法对用户进行聚类, 通过只针对同一个类内相似度更高的用户对进行相遇预测, 进一步提升预测的准确率, 并针对用户上网记录中的内容属性, 构造新的属性特征, 分析上网内容对用户相遇行为的影响。

参考文献:

- [1] 卢卫, 陆希玉. 4G 时代移动互联网的发展趋势[J]. 电信科学, 2014, 30(5): 51-54.
LU W, LU X Y. Mobile internet trends in 4G era[J]. Telecommunications Science, 2014, 30(5): 51-54.
- [2] 林金桐, 许晓东. 第五代移动互联网[J]. 电信科学, 2015, 31(5): 7-14.
LIN J T, XU X D. The 5th generation of mobile internet[J]. Telecommunications Science, 2015, 31(5): 7-14.
- [3] 中国互联网络信息中心. 第 39 次中国互联网络发展状况统计报告[R]. 2017.
China Internet Network Information Center. The 39th China

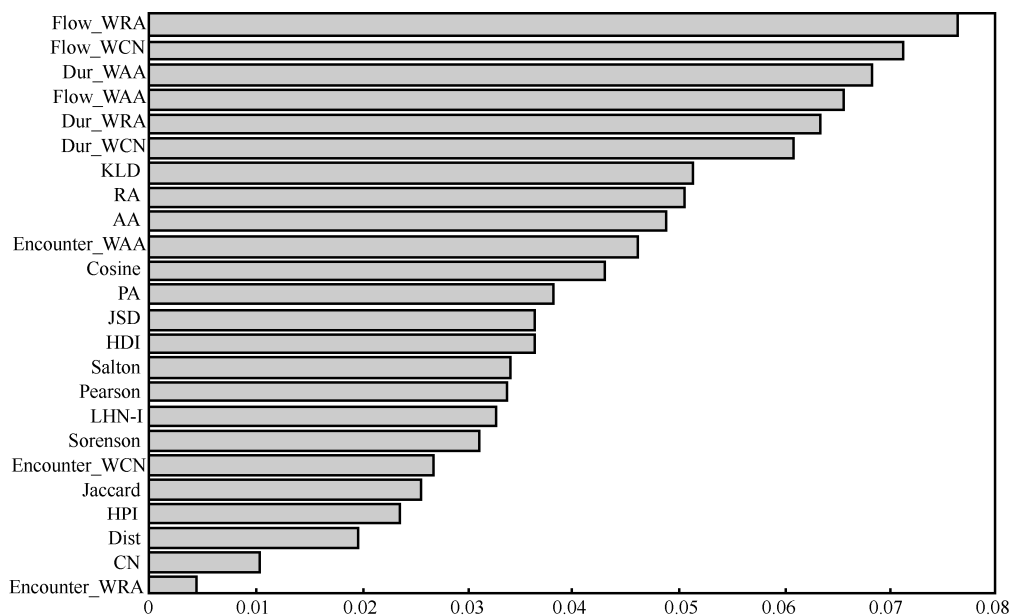


图 6 特征重要性分布

- statistical report on internet development[R]. 2017.
- [4] SONG C, BARABÁSI A L. Limits of predictability in human mobility[J]. *Science*, 2010, 327(5968): 1018-1021.
 - [5] LU X, WETTER E, BHARTI N, et al. Approaching the limit of predictability in human mobility[J]. *Scientific Reports*, 2013, 3(10): 2923.
 - [6] AN N T, TU M P. A Gaussian mixture model for mobile location prediction[C]//International Conference on Research, Innovation and Vision for the Future (RIVF), February 12-14, 2007, Okamoto, Kobe, Japan. New Jersey: IEEE Press, 2007: 914-919.
 - [7] NOULAS A, SCELLATO S, LATHIA N, et al. Mining user mobility features for next place prediction in location-based services[C]//International Conference on Data Mining, December 10-13, 2012, Brussels, Belgium. New Jersey: IEEE Press, 2012: 1038-1043.
 - [8] MONREALE A, PINELLI F, TRASARTI R, et al. Wherenext: a location predictor on trajectory pattern mining[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1, 2009, Paris, France. New York: ACM Press, 2009: 637-646.
 - [9] 苏孝强, 倪宏. 基于概率的移动用户移动模型[J]. *微计算机信息*, 2011, 27(6): 1-3.
SU X Q, NI H. Probability-based mobility model for mobile users[J]. *Microcomputer Information*, 2011, 27(6): 1-3.
 - [10] 程贤亮, 徐小良. 基于序列挖掘的用户移动位置预测[J]. *工业控制计算机*, 2013, 26(3): 70-72.
CHENG X L, XU X L. Location prediction based on sequential mining[J]. *Industrial Control Computer*, 2013, 26(3): 70-72.
 - [11] 刘震, 付俊辉, 赵楠. 基于移动通信数据的用户移动轨迹预测方法[J]. *计算机应用和软件*, 2013, 30(2): 10-17.
LIU Z, FU J H, ZHAO N. Users mobile track prediction method based on mobile communication data[J]. *Computer Applications and Software*, 2013, 30(2): 10-17.
 - [12] GONG Y, LI Y, JIN D, et al. A location prediction scheme based on social correlation[C]//Vehicular Technology Conference, May 15-18, 2011, Yokohama, Japan. New Jersey: IEEE Press, 2011: 1-5.
 - [13] ZHANG D, ZHANG D, XIONG H, et al. Nextcell: predicting location using social interplay from cell phone traces[J]. *IEEE Transactions on Computers*, 2015, 64(2): 452-463.
 - [14] SUN L, AXHAUSEN K W, LEE D H, et al. Understanding metropolitan patterns of daily encounters[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, 110(34): 13774-13779.
 - [15] SOCIEVOLE A, DE RANGO F, MARANO S. Link prediction in human contact networks using online social ties[C]//Third International Conference on Cloud and Green Computing, September 30-October 2, 2013, Karlsruhe, Germany. New Jersey: IEEE Press, 2013: 305-312.
 - [16] SCHOLZ C, ATZMUELLER M, STUMME G. On the predictability of human contacts: influence factors and the strength of stronger ties[C]//International Conference on Privacy, Security, Risk & Trust, September 3-5, 2012, Amsterdam, Netherlands. New Jersey: IEEE Press, 2012: 312-321.
 - [17] 吕琳媛. 复杂网络链路预测[J]. *电子科技大学学报*. 2010, 39(5): 651-660.
LV L Y. Link prediction on complex networks[J]. *Journal of University of Electronic Science and Technology of China*, 2010, 39(5): 651-660.
 - [18] GONZÁLEZ M C, HIDALGO C A, BARABÁSI A L. Understanding individual human mobility patterns[J]. *Nature*, 2008, 453(7196): 779-782.
 - [19] ZHOU T, LV L, ZHANG Y C. Predicting missing links via local information[J]. *European Physical Journal B*, 2009, 71(4): 623-630.
 - [20] LV L, ZHOU T. Link prediction in weighted networks: the role of weak ties[J]. *Europhys Lett*, 2010, 89(1): 18001.
 - [21] BREIMAN L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5-32.
 - [22] 王铮, 任华, 方燕萍. 随机森林在运营商大数据补全中的应用[J]. *电信科学*, 2016, 32(12): 7-12.
WANG Z, REN H, FANG Y P. Application of random forest in big data completion[J]. *Telecommunications Science*, 2016, 32(12): 7-12.
 - [23] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. *Journal of the Association for Information Science & Technology*, 2007, 58(7): 1019-1031.

[作者简介]



李倩 (1992-), 女, 武汉大学电子信息学院硕士生, 主要研究方向为大数据分析、挖掘、移动互联网等。



江昊 (1976-), 男, 武汉大学电子信息学院教授、博士生导师, 主要研究方向为大数据分析、挖掘、移动互联网、无线网络、空间综合信息网络等。



杨锦涛 (1993-), 男, 武汉大学电子信息学院硕士生, 主要研究方向为大数据分析、挖掘、移动互联网、物联网等。