基于位置信息的移动终端用户异常检测

李 志 马春来 马 涛 单 洪
(国防科技大学电子对抗学院 合肥 230037)

摘 要 针对当前轨迹异常检测中轨迹演化和检测结果类型单一的问题,结合用户历史行为模式、群体结构信息和近邻用户行为,提出一种基于位置信息的移动终端用户异常检测方法。该方法将位置数据转换为时空共现区,进一步挖掘用户行为模式,提取用户群体结构信息。在此基础上,根据历史行为模式异常、伴随行为模式异常、时空共现区行为模式异常、时空共现区流量模式异常和异常用户群体属性5种异常特征,采用随机森林方法构建多分类异常检测模型,识别移动终端用户个体异常、群体异常、时空异常和事件异常现象。在真实数据集上的实验结果表明,所提方法可以有效识别移动终端用户的轨迹演化行为,检测多种类型的异常现象,与同类方法相比具有较高的召回率和较低的误差率。

关键词 移动终端,位置数据,轨迹演化,异常特征,异常分类
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.027

Anomaly Detection Method of Mobile Terminal User Based on Location Information

LI Zhi MA Chun-lai MA Tao SHAN Hong
(College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China)

Abstract Aiming at the problem of trajectory evolution and single-type of detection result in trajectory anomaly detection technology, an anomaly detection method was proposed for mobile terminal user based on location information, which comprehensively utilizes the user historical behavior pattern, group structure information, and behavior of close users. The method converts the location data into the spatio-temporal co-occurrence area (STCOA), and further excavates the user behavior pattern and extractes the user group structure information. On this basis, a multi-class anomaly detection model was constructed by random forest method according to five abnormal characteristics of historical behavior pattern anomaly, accompanying behavior pattern anomaly, STCOA behavior pattern anomaly, STCOA flow pattern anomaly and group attribute of abnormal users. This model can identify individual anomaly, group anomaly, spatio-temporal anomaly and event anomaly of mobile terminal users. Experiments on real data sets show that the proposed method can effectively identify the trajectory evolution behavior and detect various types of anomalies of mobile terminal users. Compared with the similar methods, this method has higher recall rate and lower error rate.

Keywords Mobile terminal, Location data, Trajectory evolution, Abnormal feature, Abnormal classification

异常检测是数据挖掘的重要研究内容[1],被广泛应用于信用卡和保险等应用程序的欺诈检测[2]、网络安全的入侵检测[3]、关键系统的故障检测[4]、公共安全的视频异常监控[5-6]以及军事侦察活动等诸多领域。随着移动通信网络的快速发展和移动终端定位技术的广泛应用,基于位置的服务(Location Based Services, LBS)[7]迅速走进人们生产生活的各方面,产生了大量的位置数据。作为异常检测的一个分支,轨迹异常检测[8]旨在发现移动终端用户的异常行为模式,可在城市计算[9-10]、智能交通监控[11-12]、突发事件预警[13]、特殊人员监控[14]等方面发挥重要作用。

轨迹异常检测主要包括个体[15]或群体[16]行为异常,以及局部[17]或区域[18]流量异常等。个体或群体行为异常是指单个或者多个移动终端用户轨迹与历史行为模式不匹配的现象,如小孩走失、小偷盗窃、组团旅游、犯罪团伙作案等。局部流量异常是指短时间内小范围空间内的轨迹数量发生扰动的现象,如交通事故引起拥堵路段流量减小而周边路段流量增大。区域流量异常是指较大范围空间内轨迹流量较同期有显著波动的现象,如演唱会、体育赛事、节假日等引起的公众集中迁移行为。文献[15-18]对以上现象分别进行了研究,然而轨迹异常通常不是孤立存在的,不同异常现象之间既有区别也有联系。单一的异常检测方法容易因无法甄别不同异常现象间的因果关系而发生误判,导致异常检测的准确率不高。

到稿日期:2018-03-09 返修日期:2018-05-21 本文受国防重点实验室基金项目(9140C130104)资助。
李 志(1990-),男,博士生,主要研究方向为网络安全、位置数据挖掘,E-mail:lizhiwelcome@126.com;马春来(1989-),男,博士,讲师,主要研究方向为信息安全、位置数据挖掘;马 涛(1979-),男,博士,副教授,主要研究方向为信息安全、无线网络;单 洪(1965-),男,教授,博士生导师,主要研究方向为信息安全、位置数据挖掘、无线网络,E-mail:hshan222@163.com(通信作者)。

另一方面,位置数据具有时间演化特性,异常现象在不断变化,这给轨迹异常检测方法提出了新的挑战[8]。

因此,如何全面且准确地检测移动终端用户的轨迹异常现象是亟待解决的问题。移动终端用户行为轨迹不仅与历史行为模式有关,还可能受用户社会关系、轨迹近邻用户的行为和社会事件等因素的影响,因此其异常轨迹检测需要考虑多种因素。文献[19]指出移动社交网络中用户轨迹与周期性的历史行为和好友关系有关;基于此,文献[20]提出一种依据历史位置(H-outlier)和好友圈(F-outlier)的用户异常签到行为检测方法。文献[21]根据近邻用户轨迹的相似度检测轨迹流中的离群点。文献[17]根据路网中的历史流量模式检测异常子图,并结合社交网络信息推断异常事件。上述文献均在特定方面解决了异常轨迹的检测问题,为移动终端用户轨迹的异常检测提供了参考。本文拟针对当前异常轨迹检测方法存在的轨迹演化和检测结果类型单一的问题,综合运用用户历史行为模式、群体结构信息和近邻用户行为,提出一种全面、准确的移动终端用户异常轨迹检测方法。

1　基本思路

本文提出一种基于位置信息的移动终端用户异常检测方法(Anomaly Detection Method of Mobile Terminal User based on Location Information,ADMTUL)。首先,针对移动终端用户行为轨迹的演化特点,在用户历史行为模式异常检测的基础上,结合异常点其他用户的行为和用户所属群体中其他成员的行为,检测可能存在的演化异常现象。其次,分别从用户和地点角度定义多种异常特征,采用随机森林方法构建多分类异常检测模型,识别用户多种类型的异常现象,解决单一类型异常检测方法判定结果不够准确的问题。ADMTUL 方法的基本结构如图 1 所示,主要包括位置数据处理、异常特征检测和异常类型判定 3 个部分。首先,对位置数据进行预处理,从位置数据中提取时空共现区(Spatio-Temporal Co-Occurrence Area,STCOA),在此基础上挖掘用户行为模式,发现用户群体结构。其次,按照历史行为模式异常、伴随行为模式异常、STCOA 行为模式异常、STCOA 流量模式异常和异常用户群体属性 5 种异常特征,采用随机森林方法构建多分类异常检测模型,将异常类型划分为个体异常、群体异常、时空异常、事件异常和无异常 5 类。最后,根据当前待检测位置数据、用户行为模式和群体结构信息计算异常特征值,将其输入异常检测模型并判定用户异常类型。

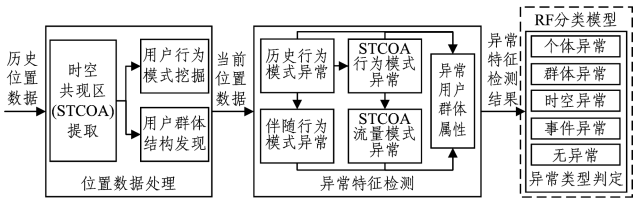


图 1　ADMTUL 框架结构图

Fig. 1　Structure of ADMTUL

2　相关工作

2.1　移动终端用户时空共现区的提取

移动终端的定位技术特性和用户使用习惯决定其位置数据具有以下特点:1)位置传感器不同,空间精度不同,如 GPS、Wifi、IP 地址、GSM 基站数据、蓝牙、射频卡终端等;2)应用目的不同,更新频次不同,如偶发更新的签到数据、连续更新的位置数据等。连续更新的位置数据量巨大,位置冗余性较高,不适合直接用于用户行为模式挖掘。因此,在挖掘用户行为模式之前需要对位置数据进行精度统一和数量压缩。位置数据可表示为 $d=\{u,p,\langle lo,la\rangle\}$,其中 u 为用户,p 为签到时间,$\langle lo,la\rangle$ 为经纬度。时空共现(Spatio-Temporal Co-Occurrences)[22-23]是指用户在时空维度一定区域相遇的事件,发生时空共现事件的时空区域称为时空共现区。设 $z_t^{\rho\tau}$ 表示以时间点 ρ 为起点、以 τ 为时长的时间段,$z_s^{\sigma\lambda}$ 表示以 o 为圆心、以 λ 为半径的空间区域,则 $z_c^{\rho\sigma}=(z_t^{\rho\tau},z_s^{\sigma\lambda})$ 表示不同用户在时间段 $z_t^{\rho\tau}$ 和空间区域 $z_s^{\sigma\lambda}$ 内共现事件的时空共现区(STCOA)。本文使用基于密度的 CFSFDP(Clustering by Fast Search and Find of Density Peaks)算法[24],以自然日为时间周期,对位置数据进行聚类,每一个聚类簇为一个 STCOA,并使用 STCOA 序列挖掘移动终端用户行为模式。

2.2　移动终端用户行为模式的挖掘

人类活动通常具有一定的周期性[19],根据周期性的历史行为模式分析用户异常行为是异常检测的常用方法[8]。STCOA 具有时间和空间双重属性,其频繁序列模式表示用户周期性的行为规律。给定用户 u 的时空共现位置数据集 $Z_c^u=\{z_{c1}^u,z_{c2}^u,z_{ci}^u,\cdots\}$ 和频繁子序列支持度阈值 s,其中 $z_{ci}^u=\{z_{c_1}^{\rho_1\cdot o_1},z_{c_2}^{\rho_2\cdot o_2},\cdots\}$ 为用户 u 单次行为轨迹的 STCOA 序列。移动终端用户行为模式挖掘任务是找到用户 u 的所有频繁 STCOA 子序列集合 $P_c^u=\{p_{c1}^u,p_{c2}^u,p_{ci}^u,\cdots\}$,其中 $p_{ci}^u=\{z_{c_i}^{\rho_i\cdot o_i},\cdots\}$ 为频繁 STCOA 子序列且满足式(1)的约束条件。

$$\begin{cases} Support_{Z_c^u}(p_{ci}^u)\geqslant s \\ Support_{Z_c^u}(p_{ci}^u)=|\{z_{ci}^u\mid p_{ci}^u\subseteq z_{ci}^u,z_{ci}^u\in Z_c^u\}|/|Z_c^u| \end{cases} \tag{1}$$

2.3　移动终端用户群体结构的发现

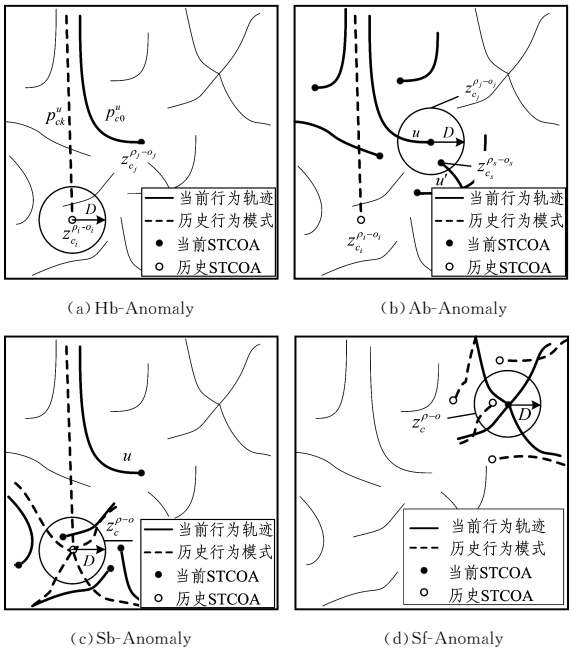
移动终端用户行为受好友影响[19],结合好友行为进行异常检测有助于甄别用户异常行为中的演化现象[20]。移动终端用户群体是具有位置聚集性的好友集合[25-26],通过分析群体成员的行为可以发现异常的用户群体。文献[27]根据位置共现信息估计用户社会关系强度,得到社会关系拓扑图,而后使用社团挖掘方法发现移动用户群体。本文首先根据用户到访 STCOA 的总体属性、用户活跃性、位置多样性和位置特殊性等 4 类特征,使用随机森林算法判断用户社会关系[28];然后使用模块度最大化标签传播(CDMM-LPA)[29]算法发现移动终端用户群体。

3　异常的定义

3.1　异常特征的定义

移动终端用户行为不仅受自身行为规律性的约束,还受外界因素的影响,如临时行程、群体成员行为、道路交通环境、社会事件等。由于引发用户行为改变的因素可能不同,因此在确定异常现象时应区别理解。例如,在学生行为监控中,若某天老师组织郊游,则整个班级学生的行为都会发生改变,此时若仅仅依据用户历史行为模式来判定所有学生的行为异常并向学校或家长发出异常警报,则会引起不必要的恐慌;而若

在每个学生的异常行为判断中参考群体成员(其他同学)同期的行为,则能避免虚警的发生。同样,在出租车反绕行欺诈监控中,因前方道路发生拥堵、车祸或者临时交通管制,司机会改变行驶路线,此时若仅仅依据历史热点路线判断出租车绕行异常并向监控中心或者乘客推送告警消息,则可能会导致乘客投诉、监控中心处罚,引起司乘纠纷,甚至干扰司机正常的驾驶行为;而如果参考同期其他车辆在绕行点的行为信息,则不会推送告警消息,从而避免不必要的纠纷。因此,本文根据用户历史行为模式和异常点群体成员及其他用户的当前行为,从用户和地点角度定义了 5 种移动终端用户异常特征:历史行为模式异常(Historical Behavior Pattern Anomaly,Hb-Anomaly)、伴随行为模式异常(Accompany Behavior Pattern Anomaly,Ab-Anomaly)、STCOA 行为模式异常(STCOA Behavior Pattern Anomaly,Sb-Anomaly)、STCOA 流量模式异常(STCOA Flow Pattern Anomaly,Sf-Anomaly)和异常用户群体属性(Group Attributes of Anomaly User,Ga-Anomaly)。异常检测算法根据用户的异常特征值判定用户行为的异常类型。图 2 为当前时刻某一区域内用户的行为轨迹,图 2(a)—图 2(d)分别为前 4 种异常特征的示意图。



(a) Hb-Anomaly (b) Ab-Anomaly

(c) Sb-Anomaly (d) Sf-Anomaly

图 2 异常特征情景示意图

Fig. 2 Scenarios of abnormal characteristics

3.1.1 历史行为模式异常

历史行为模式异常是指用户当前行为轨迹与自身周期性的行为规律不相符的现象,如图 2(a)所示。在进行历史行为模式异常检测时,首先查找用户所有频繁 STCOA 子序列集合中与当前轨迹在时间周期或空间位置上接近的序列子集,然后计算当前轨迹与序列子集的时空距离,根据时空距离的差异性确定异常情况。给定用户 u 当前位置数据对应的 STCOA 序列 $p_{c0}^{u}=\{z_{c_j}^{\rho_j-o_j},\cdots,z_{c_r}^{\rho_r-o_r}\}$ 和时空阈值 T,D,在 P_c^u 中查找满足条件的频繁 STCOA 子序列集合 $\mathbb{P}_c^u=\{p_{ck}^u \mid \exists z_{c_i}^{\rho_i-o_i}\in p_{ck}^u, |\rho_r-\rho_i|\leqslant T \text{ or } dist(o_r,o_i)\leqslant D, p_{ck}^u\in P_c^u\}$。若存在 $p_{ck}^u\in\mathbb{P}_c^u$ 满足:

$$\begin{cases}\min\limits_{p_{ck}^u\in\boldsymbol{P}_c^u}\dfrac{1}{|p_{c0}^u|-1}\Big[\sum\limits_{z_{c_j}^{\rho_j-o_j}\in p_{c0}^u}\Big(\min\limits_{z_{c_i}^{\rho_i-o_i}\in p_{ck}^u} d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})\Big)-\\ \max\limits_{z_{c_j}^{\rho_j-o_j}\in p_{c0}^u}\Big(\min\limits_{z_{c_i}^{\rho_i-o_i}\in p_{ck}^u} d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})\Big)\Big]\\ \max\limits_{z_{c_j}^{\rho_j-o_j}\in p_{c0}^u}\Big(\min\limits_{z_{c_i}^{\rho_i-o_i}\in p_{ck}^u} d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})\Big)\geqslant thre(T,D)\end{cases}\tag{2}$$

则称用户 u 在 STCOA $z_{c_i}^{\rho_i-o_i}\in p_{ck}^u$(s.t. $\max\limits_{z_{c_j}^{\rho_j-o_j}\in p_{c0}^u}\Big(\min\limits_{z_{c_i}^{\rho_i-o_i}\in p_{ck}^u} d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})\Big)$)处存在历史行为模式异常,$z_{c_i}^{\rho_i-o_i}$ 为历史异常点,$z_{c_j}^{\rho_j-o_j}$ 为实时异常点。其中,任取 $z_{c_j}^{\rho_j-o_j},z_{c_{j'}}^{\rho_{j'}-o_{j'}}\in p_{c0}^u$,若 $\rho_j\leqslant\rho_{j'}$,则 $\min\limits_{z_{c_i}^{\rho_i-o_i}\in p_{ck}^u} d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})$ 与 $\min\limits_{z_{c_{i'}}^{\rho_{i'}-o_{i'}}\in p_{ck}^u} d(z_{c_{j'}}^{\rho_{j'}-o_{j'}},z_{c_{i'}}^{\rho_{i'}-o_{i'}})$ 满足 $\rho_i\leqslant\rho_{i'}$。$d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})$ 为时空共现区的时空距离,$thre(T,D)$ 为条件时空阈值,$dist()$ 为欧氏空间距离,其定义如式(3)、式(4)所示:

$$d(z_{c_j}^{\rho_j-o_j},z_{c_i}^{\rho_i-o_i})=\begin{cases}dist(o_j,o_i), & |\rho_r-\rho_i|\leqslant T\\ |\rho_j-\rho_i|, & dist(o_r,o_i)\leqslant D\end{cases}\tag{3}$$

$$thre(T,D)=\begin{cases}D, & |\rho_r-\rho_i|\leqslant T\\ T, & dist(o_r,o_i)\leqslant D\end{cases}\tag{4}$$

3.1.2 伴随行为模式异常

伴随行为模式异常是指用户在实时异常点与部分群体成员行为轨迹一致的现象。如图 2(b)所示,当前时刻用户 u 在 STCOA $z_{c_i}^{\rho_i-o_i}$ 处存在历史行为模式异常,且在异常点 $z_{c_j}^{\rho_j-o_j}$ 处与群体成员 u' 相遇。设用户 u 当前的签到位置为 $z_{c_j}^{\rho_j-o_j}\in p_{c0}^u$,所在群体其他成员 $G(u)$ 当前签到的 STCOA 集合为 $p_c^{G(u)}=\{z_{c_s}^{\rho_s-o_s},\cdots\}$,若存在 $z_{c_s}^{\rho_s-o_s}\in p_c^{G(u)}$ 满足:

$$d(z_{c_s}^{\rho_s-o_s},z_{c_j}^{\rho_j-o_j})\leqslant thre(T,D)\tag{5}$$

则称用户 u 在 $z_{c_j}^{\rho_j-o_j}\in p_{c0}^u$ 处存在伴随行为模式异常。

3.1.3 STCOA 行为模式异常

STCOA 行为模式异常是指历史行为模式经过某一 STCOA 的多个用户在同一时期发生历史行为模式异常的现象。如图 2(c)所示,多个历史行为模式经过 $z_c^{\rho-o}$ 的用户在 $z_c^{\rho-o}$ 处发生绕行行为。设历史行为模式经过 $z_c^{\rho-o}$ 的用户集合为 $S(z_c^{\rho-o})$,$\mathcal{S}(z_c^{\rho-o})$ 表示时长为 T 的时间段内 $S(z_c^{\rho-o})$ 中在 $z_c^{\rho-o}$ 处存在历史行为模式异常的用户集合,若异常用户数量 $|\mathcal{S}(z_c^{\rho-o})|$ 满足 $|\mathcal{S}(z_c^{\rho-o})|/S(z_c^{\rho-o})\geqslant\zeta$($\zeta$ 为 STCOA 行为模式异常阈值),则称 $z_c^{\rho-o}$ 在当前时刻存在 STCOA 行为模式异常现象。

3.1.4 STCOA 流量模式异常

STCOA 流量模式异常是指特定时间段内大量用户在同一区域内的异常点发生历史行为模式异常的现象。如图 2(d)所示,在实时异常点 $z_c^{\rho-o}$ 处多个用户的行为与历史行为的模式不一致。设历史行为模式经过 $z_c^{\rho-o}$ 的用户数量为 $|S(z_c^{\rho-o})|$,实时异常点为 $z^{\rho-o}{}_c$ 的用户数量为 $|\mathcal{S}(z_c^{\rho-o})|$,若 $|\mathcal{S}(z_c^{\rho-o})|/S(z_c^{\rho-o})\geqslant\psi$($\psi$ 为 STCOA 流量模式异常阈值),则称 $z_c^{\rho-o}$ 在当前时刻存在 STCOA 流量模式异常现象。

3.1.5 异常用户群体属性

异常用户群体属性是指异常用户与所在群体中的一定比例(ξ)的其他成员在 Hb-Anomaly,Ab-Anomaly,Sb-Anomaly,Sf-Anomaly 4 种异常特征上保持一致的现象。

3.2 异常类型的定义

异常特征描述了移动终端用户在不同条件下的异常行为表现。异常类型判定根据用户的异常特征判断用户的异常行为类型。综合各异常特征的检测结果进行异常类型判定,可以发现用户的演化异常行为,细化用户的异常类型,提高异常

判断的准确度。本文根据异常现象的影响因素,将用户异常行为分为个体异常(Individual Anomaly,IA)、群体异常(Group Anomaly,GA)、时空异常(Spatio-Temporal Anomaly,SA)、事件异常(Event Anomaly,EA)和无异常(No Anomaly,NA)。

3.2.1　个体异常

个体异常是指用户发生仅与自身因素有关的异常行为的现象。个体异常表现为历史行为模式异常,且不满足其他异常特征。因此,个体异常的表现是:有且只有历史行为模式异常现象。图 2 中的大部分用户都存在历史行为模式异常现象,但从图 2(c)可以看出,由于多个用户在同一 STCOA $z_c^{p,o}$ 处发生历史行为模式异常,因此用户 u 发生主观异常的偶然性较小,存在客观诱因的可能性较大,用户 u 在图 2 中的异常行为不能判定为个体异常。同理,由于图 2(d)中的多个用户在同一个异常点处发生异常现象,因此这些用户均不能被判定为个体异常。

3.2.2　群体异常

群体异常是指群体中的多个成员同时在不同地点发生异常的现象。发生群体异常的用户成员存在历史行为模式异常,但不存在伴随行为模式异常。这是由于如果用户存在伴随行为模式异常关系,说明用户可能处于群体的集体活动状态,其行为可能是正常的。因此,群体异常的表现是:1)同一群体的多个成员同时发生异常;2)异常成员不存在伴随行为模式异常。由于移动终端用户同时存在家人、同事、朋友等多种社会关系[30-31],用户可以属于多个群体,因此在确定群体异常时,需要以历史行为模式异常用户的群体结构为索引,检查群体异常的表现条件。

3.2.3　时空异常

历史行为模式经过某一时空区域的用户集合称为该时空区域的历史用户。时空异常是指一定数量的时空区域历史用户在同一时期远离该时空区域的现象。在时空异常中,用户的异常表现通常与时空区域的局部环境有关,用户间的群体相关性较弱。时空异常的表现是:时空区域内多数历史用户发生历史行为模式异常,存在群体关系的异常用户的比例较小。

3.2.4　事件异常

事件异常是指同一时期大量用户在相同异常区域内聚集的现象。事件异常现象的诱因通常为社会事件,如节假日、体育比赛、演唱会、普通集会等。事件异常中的用户因相同的原因聚集在一起,用户间可以存在群体关系,也可以不存在群体关系。事件异常的表现是:异常区域内存在大量历史行为模式异常的用户,存在群体关系的异常用户伴随不存在群体关系的异常用户而出现。

3.2.5　无异常

无异常是指用户行为具有合理趋向性的现象。无异常包括两种情况:1)用户当前行为符合历史行为模式;2)用户行为发生演化,虽然与历史行为模式不符,但可以查找到导致当前行为与历史行为模式不符合的演化因素,且演化因素无异常行为特征。例如:用户因走访朋友改变了行为路线,团体出差或旅游计划等情况。无异常的表现有 3 种可能的情况:1)用户无历史行为模式异常;2)历史行为模式异常用户与非异常群体成员在空间上满足近邻关系;3)发生历史行为异常的群体所在的时空区域不存在大量个体异常用户。

4　异常检测算法

移动用户异常行为检测算法主要包括异常特征检测、异常类型判定、异常演化检查 3 部分。在进行异常特征检测时,需要对位置数据进行预处理,其主要任务是将用户历史轨迹数据转化为时空共现区,并在此基础上挖掘用户行为模式并发现用户群体结构。

异常特征检测是根据定义给出用户在各异常特征上的二义特征值。Hb-Anomaly 由用户历史行为模式和当前行为轨迹确定;Ab-Anomaly 由用户的 Hb-Anomaly 特征值和群体结构确定;Sb-Anomaly 由用户的 Hb-Anomaly 特征值和经过历史异常点的其他用户行为确定;Sf-Anomaly 由用户的 Hb-Anomaly 特征值和经过实时异常点的其他用户行为确定;Ga-Anomaly 由用户的 Hb-Anomaly,Ab-Anomaly,Sb-Anomaly ,Sf-Anomaly 4 类特征值和群体结构确定。

异常类型判定是根据用户的异常特征值,由异常行为检测模型输出用户的异常类型。异常行为检测模型由统计学习算法在样本数据上训练得到。本文采用随机森林(Random Forest,RF)方法[32]构建异常行为检测模型。随机森林方法不仅实现简单,训练速度快,而且可以有效避免过拟合现象,具有较高的分类准确度。其主要步骤为:

1)依据 Bootstrap 方法,从训练样本集中有放回地抽取 k_{sub} 个样本子集;

2)对于每一个样本子集,按照最大限度生长的原则构建 CART 树;在 CART 树的构建过程中,对于每一个节点,根据 Gini 不纯度选取最具有分类能力的特征进行分裂生长;

3)将生成的 k_{sub} 棵 CART 树组成 RF 分类模型,根据分类模型的投票结果给出分类结果。

演化异常行为检查是根据用户的异常特征值与异常判定结果的一致性来检查用户的演化异常行为。由无异常类型的定义可知,存在正常行为用户因轨迹演化而出现与历史行为模式不相符的现象。此类用户在 Hb-Anomaly 特征上被标记为异常,经过异常行为检测模型后可能被标记为无异常,此时应考虑用户轨迹的演化特性,更新 Hb-Anomaly 的特征值。而特征 Ab-Anomaly,Sb-Anomaly,Sf-Anomaly 和 Ga-Anomaly 均与 Hb-Anomaly 有关联,应该与 Hb-Anomaly 同步更新。因此,异常行为检测算法应根据演化异常行为检查结果在异常特征检测和异常类型判定步骤中循环,直到轨迹演化用户判定完毕。异常检测算法的流程如图 3 所示。

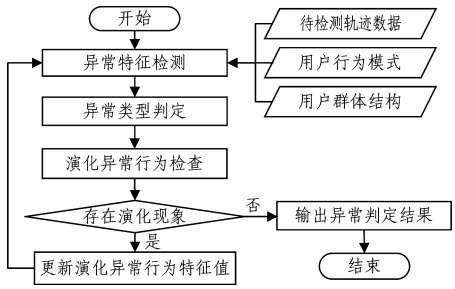


图 3　异常检测算法流程图

Fig. 3　Flowchart of anomaly detection algorithm

异常检测算法的伪代码如算法 1 所示。

算法 1 Anomaly detection

```
Input:Current location data,Historical location data
Output:Abnormal type detection results
1. Extracting Spatiotemporal co-occurrence area;
2. Extracting user group information;
3. For each user do
4.    Mining user historical behavior pattern;
5.    Detection of Hb-Anomaly;
6. loop←TRUE;
7. While(loop) do
8.    loop←FALSE;
9.    For each user do
10.       Detection of Ab-Anomaly;
11.       Detection of Sb-Anomaly;
12.       Detection of Sf-Anomaly;
13.       Detection of Ga-Anomaly;
14.   For each user do
15.     Determination of abnormal type;
16.     If (Existence of evolutionary abnormal behavior)
17.         loop←TRUE;
18.         Updating Hb-Anomaly;
19. Return abnormal type;
```

5 实验与分析

5.1 实验准备

本文将移动终端用户的异常现象分为个体异常(IA)、群体异常(GA)、时空异常(SA)、事件异常(EA)和无异常(NA)5 类。在真实数据集中,用户的大部分行为是正常的,异常位置数据的比例较小,带有标签的异常轨迹更加有限,且异常类型不全。然而,为了取得较好的实验效果,需要数据集同时包含一定数量的各类异常现象的标签样本。因此,本文在真实数据集上根据背景知识进行标注和构造来获取满足实验数量和类型要求的数据集。文献[33]中的数据集由纽约的出租车轨迹(Taxi trip Data,Tt-Data)、单车轨迹(Bike trip Data,Bt-Data)、311 市民服务数据(311 Service Data,311-Data)、路网信息(Road network Data,Rn-Data)、兴趣点数据(POIs,Po-Data)和来自 nycinsiderguide.com 网站的事件报告数据(Events report Data,Er-Data)6 部分组成。此数据集的位置数据量较大,背景知识丰富,能够最大限度地满足本文的实验需求,其统计信息如表 1 所列。

在数据集准备阶段,根据 Er-Data 和 Rn-Data 搜索异常事件所在的时空区域,将 Tt-Data 和 Bt-Data 中在此时空区域内的用户的行为标注为事件异常;根据 311-Data 数据中的"Blocked Driveway"信息和 Rn-Data 标注相同时空区域内 Tt-Data 和 Bt-Data 中的用户行为时空异常;根据 Tt-Data 和 Bt-Data 在数据预处理中获取用户行为模式、用户群体信息和 Po-Data,将同时在旅游景点出现的群体用户标注为无异常类型中的演化部分。对于个体异常和群体异常类数据,由于没有足够的背景信息,因此无法进行标注,实验中采取人工构造的方式。1)个体异常样本构造:修改 Tt-Data 和 Bt-Data 用户的部分位置数据,使其与行为模式不相符。为了避免不同类型的异常事件相互干扰,异常点避开事件异常和时空异常的时空区域。2)群体异常样本构造:根据用户群体结构和行为模式信息,修改 Tt-Data 和 Bt-Data 群体用户的部分位置数据,使群体用户在不同的地点存在与行为模式不相符的签到信息,且修改的签到信息避开其他异常现象的异常点所在的时空区域。

表 1 数据集统计信息

Table 1 Statistical information of datasets

Data	Properties	Values
Tt-Data (2014.01.01—2014.12.31)	Num. of taxicabs	14144
	Num. of trips	165M
	Duration (hour)	36.5M
	Distances (km)	5671M
Bt-Data (2014.01.01—2014.12.31)	Num. of stations	344
	Num. of bikes	6811
	Num. of trips	8M
	Duration (hour)	1.9M
311-Data (2013.05.26—2015.03.10)	Num. of categories	12
	Num. of instances	27M
Rn-Data (2013)	Num. of nodes	79315
	Num. of road segments	83655
	Num. of regions	862
Po-Data (2013)	Num. of categories	14
	Num. of instances	24031
Er-Data (2014.10.31—2014.11.27)	Num. of events	20
	Num. of instances	9083

本文采用随机森林分类方法判定移动终端用户的异常类型,需要一定数量的样本数据训练分类模型,因此实验中将样本数据集划分为训练数据集和测试数据集。由于数据集中各组成部分的收集时间的跨度不同,因此有效标注样本集中在 2014.01.01—2014.12.31 之间。在 ADMTUL 方法的数据预处理阶段,使用 Tt-Data 和 Bt-Data 的数据获取用户行为模式和群体结构信息;模型训练和测试阶段采用 10 折交叉验证方法划分训练集和测试集。

5.2 评价准则及对比方法

5.2.1 评价准则

现实生活中的异常现象复杂多样,异常用户的属性也各不相同。实验数据集中的背景信息是人工记录的,由于各种可能的原因(如用户不习惯记录异常,异常发生时情况紧急来不及记录,异常现象影响有限而没必要记录等),背景信息记录了实际环境中的部分异常情况。由于背景信息的不全面性,被标注的样本不能涵盖数据集中所有的异常现象,不能使用准确率作为异常检测方法的评价准则。由于 ADMTUL 与现有方法判定异常类型的种类不同,各异常检测方法的输出结果可能存在误分类的情况,因此异常现象分类的召回率和误差率可以作为本文实验中异常检测方法的评价指标。

5.2.2 对比方法

为了验证 ADMTUL 的有效性,实验选取与本文异常检测类型相关的方法进行对比实验。文献[16]抽取群体行为的瞬时状态组成状态转换序列,抽取状态转换序列的特征值生成特征向量,根据各时间窗口状态转换序列的特征向量距离检测集群异常行为。该方法的主要思想是:集群用户行为具有一定的周期性或半周期性,正常行为与异常行为在特征空间的聚集性质不同。文献[17]构建用户在路网中的历史行为模式,将路网子图中的用户行为与历史行为模式不相符的现

象定义为异常。文献[20]使用距离异常计算方法,根据滑动窗口内用户当前位置与邻居位置的差异程度检测历史异常,进一步根据好友圈的签到数据检测历史异常中的好友异常,最终判定有好友异常的用户为异常。文献[21]使用距离检测方法,根据轨迹与时空邻域内邻居轨迹的差异程度来确定异常现象。

5.3　结果分析

ADMTUL 检测移动终端用户的异常并进行分类,文献[16-17,20,21]中的方法仅检测单一类型的异常现象。为了更好地比较异常检测方法的召回率和误差率,需要确定文献[16-17,20,21]中的方法的异常检测结果的类型,实验统计了这些方法异常检测结果中不同类型异常现象的占比,并取各方法比例最高的类型为其异常检测结果类型。进而通过比较 ADMTUL 与以上文献中的方法在各异常类型中检测结果的召回率和误差率,来验证 ADMTUL 的有效性。

在数据集上重复实验 100 次,各异常检测方法的检测结果中不同类型异常现象在样本中的分布情况如图 4 所示,其中,ADCMP,SCTA,HF-outlier 和 DMOTS 分别为文献[16,17,20,21]中的检测方法,Sample 为样本集中各异常类型的占比。

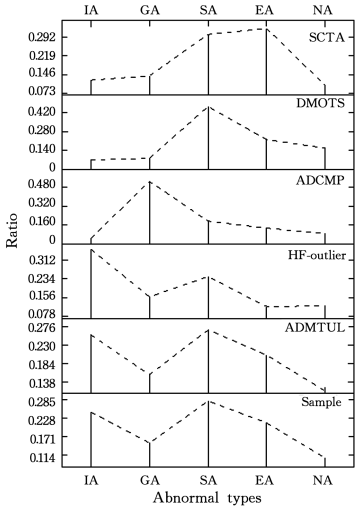


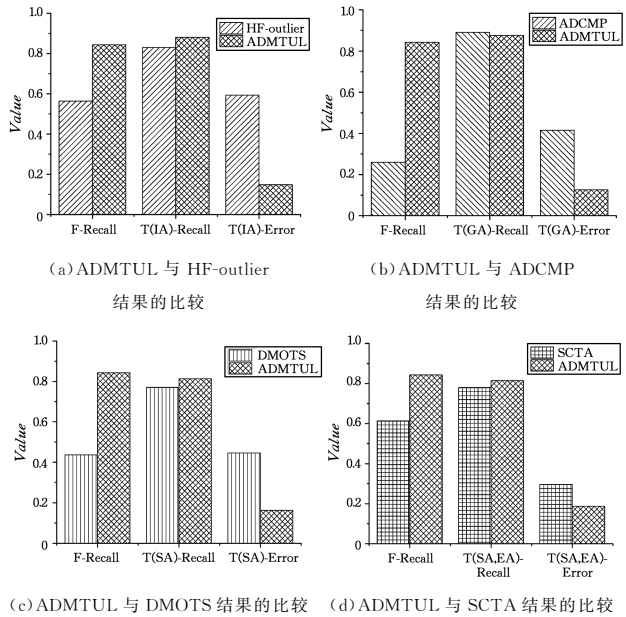
图 4　异常检测结果中异常类型的分布图

Fig. 4　Abnormal type distribution in detection results

从图 4 中可以看出,ADMTUL 检测结果中各异常类型的分布情况与样本集基本相同,而其他方法在异常类型检测上各有侧重。与样本集相比,HF-outlier 方法中个体异常(IA)的占比较大,时空异常(SA)和事件异常(EA)的占比较小,这是由于 HF-outlier 根据滑动窗口内用户位置的距离差异性识别异常现象,因此无法识别 SA 和 EA 中用户不改变行为轨迹的情况;与 HF-outlier 类似,ADCMP 根据状态转换序列中的瞬时状态在特征空间上的差异性甄别集群异常,无法识别不改变行为轨迹的异常用户。两者的区别在于,前者关注个体用户,后者关注集群用户,因此 ADCMP 方法中群体异常(GA)的占比较大。DMOTS 依据相邻轨迹流之间的差异度标识异常现象,无法检测与历史行为不同的异常现象。而轨迹流异常在 SA 中发生得较多,在其他类异常中发生得较少,因此 DMOTS 适用于检测 SA 类异常。SCTA 利用路网中用户的历史行为模式识别异常轨迹,适用于检测道路中的异常现象,因此其异常检测结果中 SA 和 EA 的占比较大。

基于以上分析,可得出如下结论:ADMTUL 使用随机森林方法,通过样本训练构建异常检测多分类模型,因此能够较好地检测各类异常现象;而其他对比方法根据异常检测目的,采用不同的异常度计算方法区分正常现象与异常现象,通常只能检测特定类型的异常,不能对异常现象进行分类。从图 4 的分析中可知,ADMTUL 可以检测多种移动终端用户异常现象;HF-outlier,ADCMP,DMOTS 和 SCTA 分别侧重于检测 IA,GA,SA 和 SAEA 类异常。为了检验 ADMTUL 的异常检测性能,实验比较了 ADMTUL 与 HF-outlier,ADCMP,DMOTS 和 SCTA 检测结果的召回率与误差率。由于异常样本的不完整性,为了使比较结果更加准确,基于已知的样本集计算召回率与误差率。同时,ADMTUL 可以对异常现象分类,而其他方法虽然在异常类别检测上有所侧重,但并没有对异常类型进行分类的义务。

为了使比较结果更加客观,实验定义了全异常样本(不包括 NA 类型)召回率(Full Sample Recall,F-Recall)、类型样本召回率(Type Sample Recall,T-Recall)和类型样本误差率(Type Sample Error,T-Error) 3 个指标。F-Recall 是指检测结果落在全部异常类型中的样本数量与异常样本总体数量的比值。T-Recall 是指检测结果落在特定异常类型中的样本数量与该类样本总体数量的比值。T-Error 是指检测结果落在全部异常类型中的样本数量与落在特定异常类型中的样本数量之差同落在全部异常类型中的样本数量的比值。ADMTUL 与 HF-outlier,ADCMP,DMOTS 和 SCTA 在全部样本上的 F-Recall,以及在 IA,GA,SA 和 SAEA 类上的 T-Recall 与 T-Error 如图 5 所示。



(a)ADMTUL 与 HF-outlier 结果的比较　(b)ADMTUL 与 ADCMP 结果的比较

(c)ADMTUL 与 DMOTS 结果的比较　(d)ADMTUL 与 SCTA 结果的比较

图 5　不同方法异常检测结果的召回率与误差率的比较

Fig. 5　Comparison of recall and error rate of abnormal detection results between different methods

从图 5(a)中可以看出,ADMTUL 在 F-Recall,T(IA)-Recall,T(IA)-Error 3 个指标上较 HF-outlier 均有领先优势。

与 ADMTUL 相比，HF-outlier 在召回率上较弱的原因是：HF-outlier 主要依据滑动窗口内历史行为的差异性检测异常用户，无法发现具有平滑但与历史行为模式不一致的异常用户。上述情况在 HF-outlier 的检测过程中普遍存在，且 SA 和 EA 类型的比重较 IA 高，因此 HF-outlier 的 T(IA)-Recall 较 ADMTUL 低，且在 F-Recall 上的差异性更加明显；同时，由于 HF-outlier 的异常检测方法无法区分异常类别，导致 T(IA)-Error 较差。图 5(b)显示，ADCMP 在指标 F-Recall 上较 ADMTUL 差，在 T(GA)-Recall 上较 ADMTUL 好。这是由于 ADCMP 主要检测集群用户的异常行为，无法发现其他类别的异常现象，因此 F-Recall 不高；同时，ADCMP 认为凡是瞬时状态发生变化的集群都存在异常，其群体异常检测条件较 ADMTUL 宽松，因此 T(GA)-Recall 较 ADMTUL 好。正是由于 ADCMP 检测条件的宽松性，没有考虑到用户的群体信息和轨迹演化特性，将 SA 中的非群体异常、EA 中的群体异常及 NA 中的群体演化行为标记为异常，才导致了其 T(GA)-Error 较差。在图 5(c)中，一方面，DMOTS 采用距离误差计算方法，只能检测轨迹流中空间维度上的差异性，忽略了时间维度(例如：发生交通拥堵时，部分车辆不改变行驶路线，而是降低行驶速度，在时空共现区上表现为到达目的地的时间比历史行为模式迟)，因此其 T(SA)-Recall 较 ADMTUL 低；另一方面，DMOTS 不考虑移动对象的历史行为模式，存在将符合自身历史行为轨迹而与邻居轨迹不同的对象标记为轨迹流异常的现象(例如：公共交通的行驶路线和停靠站行为通常与道路上的其他车辆不同，但符合其历史行驶模式)，因此 T(SA)- Error 较 ADMTUL 差。此外，DMOTS 能且只能检测其他类型中的轨迹流异常现象，无法区分异常类型，导致 T-Recall 较低，且增加了检测结果在 SA 上的误差率。图 5(d)表明，SCTA 在 F-Recall，T(SA，EA)-Recall，T(SA，EA)-Error 3 个指标上均较 ADMTUL 差，其原因在于：SCTA 根据路网历史流量模式判别异常，可以检测 SA 和 EA 两种类别的异常现象。但 SA 和 EA 可能存在不引起道路流量明显变化的异常现象，例如流量不饱和道路上的小型交通事故等 SA 类异常和在特定时段流量时常饱和的公共场所举行集会等 EA 类常。SCTA 因对上述异常现象的漏检导致其 T(SA，EA)-Recall 较 ADMTUL 低。而 SCTA 由于无法检测其他类型异常，且无法区分其他异常类型中的路网流量模式异常现象，因此 F-Recall 和 T(SA，EA)-Error 较 ADMTUL 差。

从召回率与误差率的比较分析中可以得出以下结论：1)ADMTUL 采用时空共现区表示用户位置信息，可以同时识别用户在空间和时间维度的异常现象；2)结合历史行为模式和历史异常点其他用户行为，有效提高了 ADMTUL 对地点和事件异常的识别率；3)以群体结构信息为辅助，通过分析实时异常点其他用户行为，提高了个体和群体中演化异常的检测准确度和事件异常的识别率；4)使用分类模型有效识别用户不同类型的异常行为，提高了 ADMTUL 检测结果的整体召回率，降低了各类型检测结果的误差率。

综上所述，与同类方法相比，ADMTUL 根据 Hb-Anomaly，Ab-Anomaly，Sb-Anomaly，Sf-Anomaly 和 Ga-Anomaly 5 种特征构建分类模型，能够有效识别移动终端用户不同类型的异常现象，检测结果具有较高的召回率和较低的误差率。

结束语 移动终端用户异常检测具有广泛的应用前景，现有方法存在无法识别演化轨迹和检测结果类型单一的问题。针对该问题，本文结合用户历史行为模式、群体结构信息和近邻用户行为，提出一种基于位置信息的移动终端用户异常检测方法(ADMTUL)。ADMTUL 首先针对移动终端用户行为轨迹的演化特点，在用户历史行为模式异常检测的基础上，结合异常点其他用户和群体成员的行为来检测可能存在的演化异常现象；其次，分别从用户和异常地点角度定义历史行为模式异常、伴随行为模式异常、STCOA 行为模式异常、STCOA 流量模式异常和异常用户群体属性 5 种特征，采用随机森林方法构建多分类异常检测模型，识别用户多种类型的异常行为。在真实数据集上进行的实验表明，ADMTUL 可以有效识别用户的演化轨迹，检测多种类型的异常现象，与同类方法相比具有较高的召回率和较低的误差率。

参 考 文 献

[1] SHIKHA A，JITENDRA A. Survey on Anomaly Detection using Data Mining Techniques[J]. Procedia Computer Science，2015，60(1)：708-713.

[2] LIU T B，LIU S P. Fraud detection model & application for credit card acquiring business based on data mining technology [C]// 4th International Conference on Electrical & Electronics Engineering and Computer Science. Atlantis Press，2016.

[3] CHRISTIANA I，VASOS V，CHARALAMPOS S. An Intrusion Detection System for Wireless Sensor Networks[C]// International Conference on Telecommunications. 2017.

[4] ZHOU J，YANG Y，DING S，et al. A Fault Detection and Health Monitoring Scheme for Ship Propulsion Systems using SVM Technique[J]. IEEE Access，2018，PP(99)：1.

[5] WANG T，HICHEM S. Detection of Abnormal Visual Events via Global Optical Flow Orientation Histogram[J]. IEEE Transactions on Information Forensics & Security，2014，9(6)：988-998.

[6] LI A，MIAO Z J，CEN Y G，et al. Anomaly detection using sparse reconstruction in crowded scenes[J]. Multimedia Tools & Applications，2017，76(24)：26249-26271.

[7] ANIND D，JEFFREY H，EYALDE L，et al. Location-Based Services[J]. IEEE Pervasive Computing，2017，9(1)：11-12.

[8] MAO J L，JIN C Q，ZHANG Z G，et al. Anomaly detection for trajectory big data：Advancements and framework[J]. Journal of Software，2017，28(1)：17-34. (in Chinese)
毛嘉莉，金澈清，章志刚，等. 轨迹大数据异常检测：研究进展及系统框架[J]. 软件学报，2017，28(1)：17-34.

[9] ZHENG Y U，LIU Y C，YUAN J，et al. Urban computing with taxicabs[C]// International Conference on Ubiquitous Computing. ACM，2011.

[10] ZHENG Y. Urban computing and large data[J]. Communication of China Computer Federation，2013，9(8)：8-18. (in Chinese)
郑宇. 城市计算与大数据[J]. 中国计算机学会通讯，2013，9(8)：8-18.

[11] LIU H P，JIN C Q，ZHOU A Y. Popular Route Planning with

Travel Cost Estimation[C]// Proceedings, Part II, of the 21st International Conference on Database Systems for Advanced Applications-Volume 9643. 2016.

[12] DUAN X Y, JIN C Q, WANG X L, et al. Real-Time Personalized Taxi-Sharing[M]. Springer International Publishing, 2016.

[13] PANG L, CHAWLA S, LIU W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data & Knowledge Engineering, 2013, 87(9): 357-373.

[14] REN M Q, SONG R X, WANG M, et al. Detection of Students' Abnormal Behavior in the Intelligent Campus[J]. Natural Science Journal of Harbin Normal University, 2017, 33(3): 20-24. (in Chinese)
任孟其,宋汝鑫,王萌,等. 面向智慧校园的学生异常行为检测[J]. 哈尔滨师范大学自然科学学报,2017,33(3):20-24.

[15] DONGHER S, MINGHUNG S, DAVIDC Y, et al. Personal mobility pattern mining and anomaly detection in the GPS era [J]. American Cartographer, 2016, 43(1): 55-67.

[16] YANG S, ZHOU W B. Anomaly Detection on Collective Moving Patterns: Manifold Learning Based Analysis of Traffic Streams [C]// IEEE Third International Conference on Privacy, Security, Risk and Trust. 2012.

[17] PAN B, ZHENG Y, WILKIE D, et al. Crowd sensing of traffic anomalies based on human mobility and social media[C]// ACM Sigspatial International Conference on Advances in Geographic Information Systems. 2013.

[18] PANG L, CHAWLA S, LIU W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data & Knowledge Engineering, 2013, 87(9): 357-373.

[19] EUNJOON C, SETHA M, JURE L. Friendship and mobility: user movement in location-based social networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA, 2011.

[20] ZHAO G Z, QI J P, YU Y W, et al. Online check-in exception detection in mobile social networking[J]. CAAI Transactions on Intelligent Systems, 2017, 12(5): 752-759. (in Chinese)
赵冠哲,齐建鹏,于彦伟,等. 移动社交网络异常签到在线检测算法[J]. 智能系统学报,2017,12(5):752-759.

[21] YU Y W, CAO L, ELKEA R, et al. Detecting moving object outliers in massive-scale trajectory streams[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 422-431.

[22] CRANDALL D J, BACKSTROM L, COSLEY D, et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(52): 22436-22441.

[23] TAN R, GU J Z, CHEN P, et al. Link Prediction Using Protected Location History[C]// Fifth International Conference on Computational and Information Sciences. IEEE, 2013.

[24] RODRIGUEZ A, LAIO A. Machine learning. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492.

[25] LIM K, CHAN J, LECKIE C, et al. Detecting Location-centric Communities using Social-Spatial Links with Temporal Constraints[C]// European Conference on Information Retrieval. 2015.

[26] CHLOË B, VINCENZO N, SALVATORE S, et al. Social and place-focused communities in location-based online social networks[J]. European Physical Journal B, 2013, 86(6): 1-10.

[27] VISHNU J, KINSHUK B, ANSHU K, et al. Discovering Local Social Groups using Mobility Data[J]. International Journal of Computer Applications, 2015, 120: 15-19.

[28] MA C L, SHAN H, MA T, et al. An Improved Random Forests Algorithm with Application to Social Ties Inferring of LBS Users[J]. Journal of Chinese Computer Systems, 2016, 37(12): 2708-2712. (in Chinese)
马春来,单洪,马涛,等. 随机森林改进算法在 LBS 用户社会关系推断中的应用[J]. 小型微型计算机系统,2016,37(12):2708-2712.

[29] CHEN J, WAN Y. Research on label propagation algorithm based on modularity maximization in the social network[J]. Journal on Communications, 2017, 38(2): 25-33. (in Chinese)
陈晶,万云. 社交网络中基于模块度最大化的标签传播算法的研究[J]. 通信学报,2017,38(2):25-33.

[30] WANG Z, ZHANG D Q, ZHOU X S, et al. Discovering and Profiling Overlapping Communities in Location-Based Social Networks[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2014, 44(4): 499-509.

[31] GUO J, SONG P. Community Discovery with Location-Interaction Disparity in Mobile Social Networks[J]. ZTE Communications, 2015, 13(2): 53-61.

[32] MA C L, SHAN H, MA T, et al. Random Forests Based Method for Inferring Social Ties of LBS Users[J]. Computer Science, 2016, 43(12): 218-222. (in Chinese)
马春来,单洪,马涛,等. 一种基于随机森林的 LBS 用户社会关系判断方法[J]. 计算机科学,2016,43(12):218-222.

[33] ZHENG Y, ZHANG H C, YU Y. Detecting collective anomalies from multiple spatio-temporal datasets across different domains [C]// Sigspatial International Conference on Advances in Geographic Information Systems. Bellevue, WA, USA: ACM, 2015.