

## 基于聚类的出租车异常轨迹检测

朱 燕,李宏伟,樊 超,许栋浩,施方林

(信息工程大学 地理空间信息学院,郑州 450001)

**摘 要:** 出租车全球定位系统数据中蕴含城市交通和移动对象行为的宏观信息,从中可以挖掘出有价值的异常轨迹模式。将位置和几何形状、行驶时间分别作为出租车轨迹的空间与时间特征,根据特征偏离情况划分时间、空间和时空异常轨迹。从轨迹数据中提取相同起终点的轨迹集,将轨迹划分成轨迹片段,计算轨迹间的相似度并进行基于距离和密度的聚类,在空间特征上初步分离出频繁和稀疏轨迹,根据数据异常判定的  $k\sigma$  准则确定时间特征异常的分离阈值,对时间特征进行再次划分,最终实现出租车异常轨迹检测。实验结果表明,该方法能从异常轨迹中挖掘出个性化路线、异常停留位置和交通路段,为智能交通、物流高效规划和执行等提供参考信息。

**关键词:** 异常轨迹检测;全球定位系统数据;轨迹聚类;时空特征;轨迹模式

**中文引用格式:**朱 燕,李宏伟,樊 超,等.基于聚类的出租车异常轨迹检测[J].计算机工程,2017,43(2):16-20.

**英文引用格式:**Zhu Yan, Li Hongwei, Fan Chao, et al. Clustering-based Taxi Trajectory Outlier Detection[J]. Computer Engineering, 2017, 43(2): 16-20.

## Clustering-based Taxi Trajectory Outlier Detection

ZHU Yan, LI Hongwei, FAN Chao, XU Donghao, SHI Fanglin

(Institute of Geographic Space Information, Information Engineering University, Zhengzhou 450001, China)

**[Abstract]** Taxi Global Position System (GPS) data contain macro information about the behavior of urban traffic and moving object behavior, from which valuable anomalous trajectory patterns can be mined. The location, geometry and travel time are taken as the spatial and temporal characteristics of the taxi trajectory respectively. According to the deviation of the feature, the trajectory anomalies are divided into temporal, space and spatio-temporal outliers. The trajectories of the same starting and ending points are extracted from the trajectory data, and are partitioned into segments. The similarity between trajectories is calculated and clustering based on distance and density is carried out. Frequent and the sparse trajectories are preliminary separated by the spatial characteristics. Based on  $k\sigma$  criterion, the separation threshold of temporal anomaly is determined to realize the classification of the temporal characteristic, and finally the trajectory outlier detection of the taxi is realized. The experimental results show that the method can extract personalized route as well as abnormal parking location and traffic section from abnormal trajectories, providing reference information for intelligent transportation as well as efficient logistics planning and execution.

**[Key words]** trajectory outlier detection; Global Position System (GPS) data; trajectory clustering; spatio-temporal characteristics; trajectory pattern

**DOI:**10.3969/j.issn.1000-3428.2017.02.003

### 0 概述

海量、大比例尺的全球定位系统 (Global Position System, GPS) 轨迹能够揭示城市动态以及与人类行为相关的隐含事实<sup>[1]</sup>, 针对城市移动轨迹挖掘问题的研究已经取得了许多成果, 比如从轨迹数据中提

取兴趣区域<sup>[2]</sup>和热点路径<sup>[3]</sup>, 发现移动对象行为模式和社会角色<sup>[4]</sup>, 利用出租车轨迹数据探索区域间的移动交互时空规律<sup>[5]</sup>, 分析城市路况和交通可达性等, 这些研究大多侧重于频繁模式发现, 而在异常模式发现方面的研究较少。出租车是城市居民出行不可缺少的交通工具, 由于 GPS 设备的安装,

**基金项目:**国家自然科学基金“空间数据流的概念漂移问题研究”(41571394)。

**作者简介:**朱 燕(1992—),女,硕士研究生,主研方向为智能交通、空间数据挖掘;李宏伟,教授、博士、博士生导师;樊 超、许栋浩、施方林,硕士研究生。

**收稿日期:**2016-05-24 **修回日期:**2016-07-01 **E-mail:**zy1076645457@163.com

产生了覆盖整个城市范围的 GPS 数据,这些有序 GPS 数据点连接成无数条移动路径,从中可以探测到异常轨迹。异常轨迹表现为属性特征上偏离大多数轨迹的对象,不同偏离方式具有共性行为特征,比如规避交通拥堵、事故的绕道行为,出租车司机欺骗绕道,道路封锁、事故导致的途中长时间停留行为等,因此,从出租车轨迹数据中探测异常轨迹,分析移动轨迹的异常特征和规律,能使人们及时针对异常做出相应处理,以减小该异常带来的损失。

已有异常轨迹检测方法主要包括经典 TROD 算法<sup>[6-7]</sup>、iBAT 算法<sup>[8]</sup>等,由于这些算法本身不考虑时间因素,因此不利于直接发现更多的异常模式。文献[9]提出了独立时间异常轨迹检测方法,先找出时间窗口内历史轨迹的前  $k$  个热点轨迹,再进行同一时间窗口内轨迹与热点轨迹时空编辑距离的比较,得到的结果属于动态异常,但是这种异常的时间和空间特征不明显,不易做出进一步解释。传统异常检测方法中基于聚类的方法把异常检测视为不属于任何簇的对象,能同时得到有意义的聚簇和异常聚簇,实现正常轨迹和异常轨迹的直观对比,而目前轨迹聚类主要用来挖掘轨迹频繁模式,在异常轨迹检测研究中很少使用。目前已有许多轨迹聚类方法被提出,常见方法是对轨迹分段后再进行轨迹形状聚类,与整条轨迹作为聚类单元参与计算的方法相比,更有利于轨迹局部特征的细致比较,如文献[10]提出的基于密度聚类思想的 TRACUS 方法,文献[11]提出的改进 Hausdorff 距离的轨迹聚类算法。

本文研究从出租车 GPS 数据中挖掘有意义的异常轨迹,在提取相同起终点轨迹数据集的基础上,综合考虑轨迹的时空静态特征,介绍时间、空间和时空异常轨迹 3 种异常定义,提出基于聚类的异常轨迹检测(Clustering-based Trajectories Outlier Detection, CTOD)方法,发现易于分析和理解的异常模式。

## 1 相关定义

### 1.1 轨迹

**定义 1(轨迹)** 一条轨迹  $TR$  由一系列点  $\langle p_1, p_2, \dots, p_n \rangle$  组成,其中,  $p_i = (x_i, y_i, t_i)$  且  $t_1 < t_2 < \dots < t_n$ ,  $x$  和  $y$  分别是空间坐标的经度和纬度,  $t$  是点  $p$  运动到此坐标的瞬时时刻。 $TR$  由若干子轨迹  $T$  组成,若子轨迹起点  $S$  所在区域为  $R_1$ 、终点  $D$  所在区域为  $R_2$ ,则表示为  $T = \langle p_i, p_{i+1}, \dots, p_j \rangle \in TR$ ,其中,  $T \cap R_1 = \{p_i\}$ ,  $T \cap R_2 = \{p_j\}$ ,  $1 \leq i < j \leq n$ 。出租车轨迹集中每条轨迹都有各自不同的起终点,为了使轨迹间的比较更加合理,从而增强检测到异常轨迹

的概率,在数据预处理时需要提取相同起终点的子轨迹集。每条子轨迹划分成若干近似轨迹片段,轨迹片段  $L$  为一条线段  $p_i p_j (i < j)$ ,其中  $p_i, p_j$  属于同一条轨迹。

### 1.2 轨迹异常

#### 1.2.1 空间特征异常

轨迹的局部特征中包含位置、方向、转角等形状信息,将其视为轨迹的空间特征,城市中从一个地方移动到另一个地方的轨迹由于不同道路的选择可能在某条道路上发生重合或分离,形成不同的空间特征。在给出空间特征异常定义前,先介绍用于轨迹间空间特征比较的轨迹片段距离和轨迹相似的定义。

**定义 2(轨迹片段距离)** 在模式识别领域内将线段之间距离定义为垂直距离( $d_{\perp}$ )、平行距离( $d_{\parallel}$ )和角度距离( $d_{\theta}$ )3 个部分的权重和,即  $dist(L_i, L_j) = w_{\perp} \cdot d_{\perp}(L_i, L_j) + w_{\parallel} \cdot d_{\parallel}(L_i, L_j) + w_{\theta} \cdot d_{\theta}(L_i, L_j)$ ,其中  $w_{\perp}, w_{\parallel}, w_{\theta}$  在具体应用中采用适当大小的值,得到两线段的形状匹配距离。各距离的计算示意图如图 1 所示,具体计算公式为:

$$d_{\perp} = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}}$$

$$d_{\parallel} = \text{Min}(l_{\parallel 1}, l_{\parallel 2})$$

$$d_{\theta} = \|L_j\| \times \sin(\theta)$$

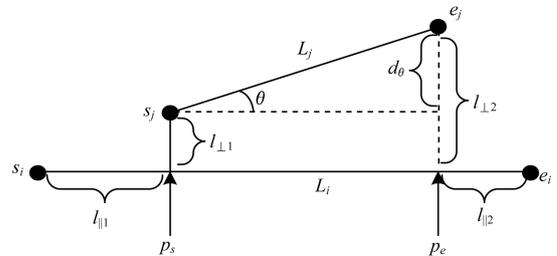


图 1 垂直距离、平行距离和角度距离示意图

给定距离阈值  $\varepsilon$ ,若轨迹片段  $L_i$  和  $L_j$  满足  $dist(L_i, L_j) < \varepsilon$ ,则  $L_i$  和  $L_j$  互为轨迹片段近邻。轨迹片段距离度量的是轨迹片段间的相似程度,由此进一步比较轨迹间的相似性。

**定义 3(轨迹相似)** 对于轨迹  $T_i$  和  $T_j$ ,若互为近邻的所有轨迹片段占各自所属轨迹长度比例的加权平均大于  $f$ ,则认为这 2 条轨迹是空间特征相似,表示为设轨迹  $T_j$  和  $T_i$  的轨迹片段近邻集合为  $C$ ,若下式成立,则  $T_i \sim T_j$ :

$$\left\{ \frac{\sum_{L_i \in \text{Cand}_{L_i \in T_i}} \text{len}(L_i)}{\text{len}(T_i)} + \frac{\sum_{L_j \in \text{Cand}_{L_j \in T_j}} \text{len}(L_j)}{\text{len}(T_j)} \right\} > f$$

每条轨迹  $T$  的相似轨迹集合  $SC(T)$  为轨迹邻域,邻域密度  $|SC(T)|$  表示集合内的轨迹数目。邻域内至少包含  $MinTrs$  个对象的轨迹为核心轨迹。根据 DBSCAN 中密度可达、密度相连的概念将核心

轨迹及其邻域扩充成轨迹簇。

**定义4(空间特征异常)** 通过轨迹片段距离、轨迹相似的密度聚类可以得到空间特征上类似的频繁轨迹聚簇,而不属于任何轨迹簇的轨迹为空间特征异常,即轨迹  $T$  满足  $|SC(T)| < MinTrs$ ,认为  $T$  为空间特征异常。

### 1.2.2 时间特征异常

**定义5(时间特征异常)** 将整条轨迹看成一个对象,其行驶时间耗费视为专题属性特征,行驶时间数值表现异常的轨迹为时间特征异常。

统计判别法中检测异常数据常使用  $k\sigma$  准则。 $k\sigma$  准则不需要指定参数,计算方法更加客观,基本原理为:对于一组数据  $T = \{t_1, t_2, \dots, t_n\}$ ,若满足表达式  $|t_i - \bar{t}| > k\sigma$ ,则为异常数据,其中,  $\bar{t} = \sum_{i=1}^n t_i / n$  为

平均值;  $\sigma = \sqrt{\sum_{i=1}^n (t_i - \bar{t})^2 / (n-1)}$  为标准差。文献[12]通过实验证明,当  $k$  取 1.645 时,此准则对于异常数据的判断较合理可靠,文献[13]以  $k\sigma$  准则为基础,提出一种专题属性双重偏离的时空异常检测方法,因此,本文取此值作为判定依据。在实际情况下,轨迹行驶时间有下限无上限,因此,检测到的异常为行驶时间花费超出上限  $\bar{t} + k\sigma$  的轨迹,时间异常的判断阈值由  $Threshold = \bar{t} + k\sigma$  计算得到。

### 1.3 轨迹模式

根据轨迹的时间、空间特征表现出的正常和异常组合得到 4 种轨迹模式:

**定义6(标准轨迹)** 标准轨迹(Standard)表示空间、时间特征均正常的轨迹。在通常情况下,移动在两地间的轨迹为标准轨迹,即大众频繁选择使用的移动路径,该路径不止一条。

**定义7(时间异常轨迹)** 时间异常轨迹(T-Outlier)表示在空间上未发生偏离、时间特征异常的轨迹。由于恶劣天气、道路中发生严重的交通拥堵、道路封锁等原因都可能导致移动主体在移动过程中产生长时间的滞留行为,使得移动轨迹的时间特征异常。

**定义8(空间异常轨迹)** 空间异常轨迹(S-Outlier)表示在空间上发生偏离、时间特征正常的轨迹。从这种轨迹模式中可以进一步挖掘出规避交通事故、道路施工禁行等产生的绕道行为或者出租车司机选择绕行、缩短行车时间、增加收费的欺骗行为。

**定义9(时空异常轨迹)** 时空异常轨迹(ST-Outlier)表示空间和时间特征均表现异常的轨迹。在移动主体不熟悉区域和行驶路线,途经多处地点或者交通拥堵导致的时间消耗等情况下,易产生与标准轨迹在时空特征上都有较大程度偏离的行驶路径。

## 2 基于聚类的异常轨迹检测方法

为从出租车历史轨迹数据中挖掘上文中定义的轨

迹模式,本文提出基于聚类的异常轨迹检测(CTOD)方法,主要内容包括轨迹预处理、轨迹分割、轨迹聚类和异常轨迹检测,流程如图2所示。

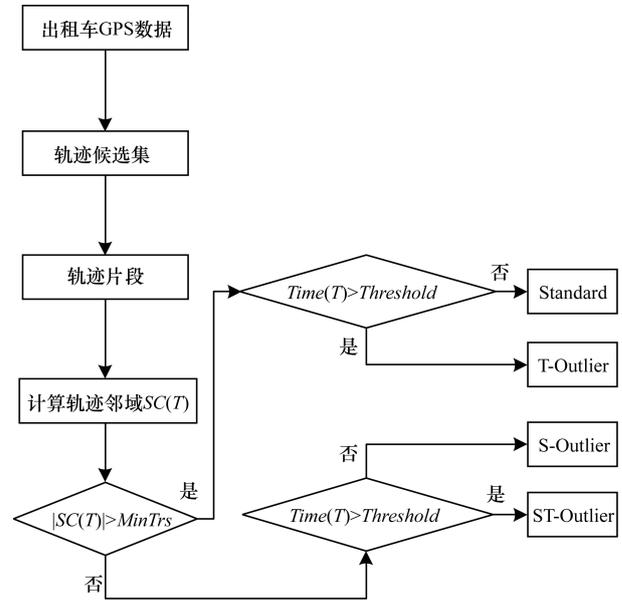


图2 基于聚类的异常轨迹检测流程

### 2.1 轨迹预处理

对于轨迹候选集的提取,设定起终点范围,提取满足从一区域至另一区域移动的轨迹,提取的所有轨迹是本文需要进一步处理的轨迹候选集;再计算轨迹候选集里每条轨迹的行驶时间属性  $Time$ ,获取其统计特征  $\overline{Time}$  和  $\sigma$ ,由  $k\sigma$  准则确定时间阈值  $Threshold$  为  $\overline{Time} + 1.645\sigma$ 。

### 2.2 轨迹分割

轨迹分割的合理与否直接影响轨迹片段间距离比较的准确程度,进一步决定了轨迹聚类的有效性。TRACCLUS 方法采用最小描述长度原理对轨迹进行分段,但并不能找到最优划分结果。轨迹划分的前提是找到能保持和反映轨迹整体以及局部变化的特征点,因此,本文以轨迹转角大于角度阈值  $w$  的采样点作为轨迹的拐点或突变点,分割方法参考文献[14],实现对轨迹候选集的分割,得到近似轨迹片段。

### 2.3 轨迹聚类与异常轨迹检测

本文方法与 TRACCLUS 轨迹聚类方法不同的是聚类对象仅为轨迹,而不对轨迹片段进行聚类,最终得到标准轨迹簇和 3 种异常轨迹,具体包括以下步骤:

1) 设定轨迹片段距离阈值  $\varepsilon$ ,密度阈值  $MinTrs$ ,轨迹片段距离权值  $w_{\perp}, w_{\parallel}, w_{\theta}$ ,比例参数  $f$ 。

2) 初始化标准轨迹分类 ID,选择一条轨迹  $T_i$  与候选轨迹  $T$  中其余轨迹两两间进行轨迹片段距离计

算,根据轨迹相似的定义,得到轨迹邻域 $SC(T_i)$ 。

3) 若 $|SC(T_i)| > MinTrs$ ,且 $Time(T_i) > Threshold$ ,将 $T_i$ 标记为 T-Outlier; 若 $|SC(T_i)| > MinTrs$ ,且 $Time(T_i) < Threshold$ ,新建一个轨迹簇 $N$ ,将其设置标准轨迹的簇 ID,并把邻域中所有密度可达轨迹都加入 $N$ 中,同时扩展聚类。

4) 若 $|SC(T_i)| < MinTrs$ ,且 $Time(T_i) > Threshold$ ,则标记为 ST-Outlier; 若 $|SC(T_i)| < MinTrs$ ,且 $Time(T_i) < Threshold$ ,则标记为 S-Outlier。

### 3 实验结果与分析

#### 3.1 实验数据预处理

实验使用由 cabspotting 项目提供的真实出租车 GPS 轨迹数据<sup>[15]</sup>,数据集包括在美国旧金山 2008 年 5 月 19 日—2008 年 6 月 10 日期间采集的约 500 辆出租车行驶轨迹,大约 110 万个轨迹点。每个轨迹点包括 4 个属性:纬度,经度,计费状态及时间戳,采样时间间隔约为 1 min。计费状态值为 1 时表示载客运营,为 0 时表示空载,于是利用 0 和 1 之间的间歇能得到乘客上车或下车的时空位置。此外,只有由连续计费状态均为 1 的轨迹点组成的行驶轨迹能被认定为车辆完成了一次有意义的路径,才能真正揭示移动对象的移动规律。本文对数据进行了上下车位置点的提取和可视化,并得到热点图,如图 3 所示,图 3(a)为上下车点分布图,图 3(b)为下车点热度分布,颜色越深的区域表明此区域越热门,乘客越多。

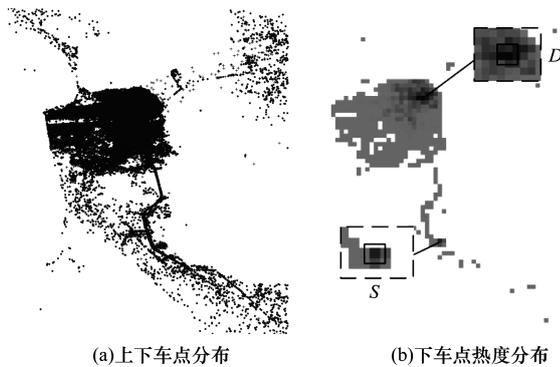


图 3 上下车点可视化结果

本次实验中选择提取最热门的 2 个区域(图 3(b)中实线黑框)之间的行驶轨迹作为候选轨迹集,实地为机场 $S$ 去往一中心住宅区 $D$ 的轨迹,并将时间范围限制在 2008 年 6 月 1 日—2008 年 6 月 10 日每天 05:00—23:00,将数据经过去噪处理后,得到 12 732 个轨迹点,提取 587 条候选轨迹,并将轨迹按时间顺序依次编号为 1~587。对所有候选轨迹行驶时间值进行统计分析,确定时间特征异常阈值 $Threshold$ 为 32 min。

#### 3.2 结果分析

实验参数 $w, w_{\perp}, w_{\parallel}, w_{\theta}$ 参考已有研究设定相应值,参数 $\varepsilon, f$ 和 $MinTrs$ 对检测方法的影响比较大,经过多次调整设置,当 $\varepsilon = 0.005, MinTrs = 20, f = 0.8, w = 20, w_{\perp} = 0.8, w_{\parallel} = 0.05, w_{\theta} = 0.15$ 时,实验得到了较为稳定合理的结果,如图 4 所示。图 4(a)是从起点 $S$ 到终点 $D$ 的所有轨迹。2 类标准轨迹如图 4(b)所示,Standard1 有 138 条,Standard2 有 429 条。3 种异常轨迹的检测结果如图 4(c)所示。

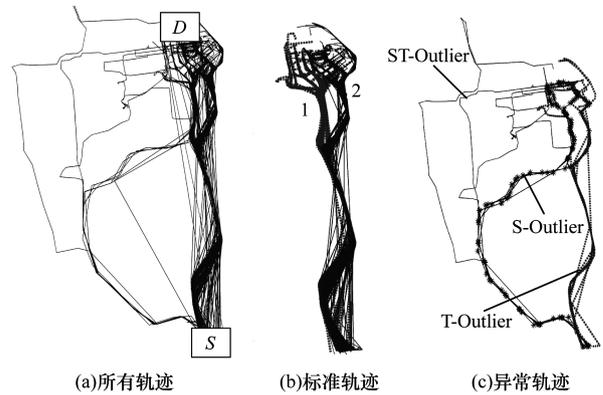


图 4 真实数据集实验结果可视化

下面对实验结果进行分析:

1) 从图 4(b)中可以了解到人们从机场移动到住宅区频繁使用的路径,这些路径在靠近住宅区的路段出现了道路分流选择,而且并不是视觉上路程最短的选择,更多偏向于 Standard2,这一定程度上反映了移动对象的路径选择习惯。

2) 偏离了标准轨迹的空间异常轨迹在行驶距离上有所增加,在同等速度条件下行驶时间花费应该更多,但将其行驶时间与同时间(1 h)窗口内所有标准轨迹的平均行驶时间进行比较,结果相差不多,甚至更少。轨迹行驶时间分布图(图 5)也体现这一点,并且能从中观察到少数空间异常和时间异常伴随发生的现象。说明出租车司机可能为了减少行驶时间、避免拥堵或其他事故而选择绕道而行,根据驾驶经验产生的行为,这样的异常轨迹也许会是用户在紧急情况下所需要的快捷路径。

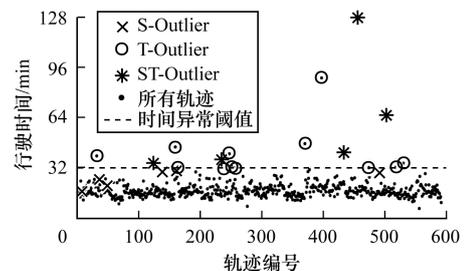


图 5 轨迹行驶时间分布

3)有些时空异常轨迹有较长路段和标准轨迹重合,部分轨迹片段过度偏离了标准路径选择,并且行驶时间花费更多,可以怀疑出租车司机的这种绕道行为隐含欺骗性。考虑到出租车载客的实际情况,产生这种偏离也有可能是乘客们不是直达目的地、而是沿途有多个下车地点的原因。

4)图6(a)是其中一条时空异常轨迹点的可视化结果,图6(b)对应的是一条时间异常轨迹,可以发现,这些轨迹的局部线段在可视化后呈现出互相紧挨着甚至重叠的一串串点,事实上出租车轨迹数据采样间隔是相同的,说明轨迹在途中某地产生了长时间停留或者行驶速度特别缓慢,其中图6(a)表示的轨迹行使用时最长长达2h以上,在途中2个路段花费时间均超过0.5h。从这些异常轨迹中可以挖掘出异常的停留位置和交通异常路段,为智能交通、物流高效规划和执行提供有利的参考信息。

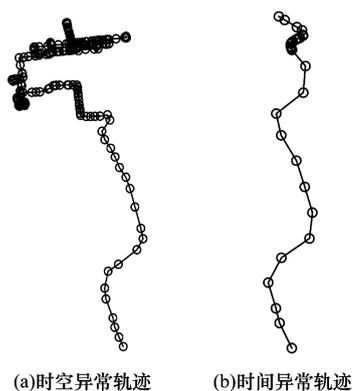


图6 时空与时间异常轨迹可视化结果

#### 4 结束语

本文通过分析城市中相同起终点间移动的出租车轨迹的时空特征,对异常轨迹进行定义,提出基于聚类的异常轨迹检测方法。实验结果表明,该方法能从复杂的出租车轨迹数据中将易于解释的时间异常、空间异常以及时空异常轨迹分离出来。下一步将研究出租车异常轨迹实时检测方法并提高检测准确率。

#### 参考文献

- [1] 郑宇. 城市计算概述[J]. 武汉大学学报(信息科学版), 2015, 40(1): 1-13.
- [2] Giannotti F, Nanni M, Pinelli F, et al. Trajectory Pattern Mining [C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2007: 330-339.
- [3] Chen Zaiben, Shen Hengtao, Zhou Xiaofang. Discovering Popular Routes from Trajectories [C]//Proceedings of International Conference on Data Engineering. Washington D. C., USA: IEEE Computer Society, 2011: 900-911.
- [4] 马宇驰, 杨宁, 谢琳. 基于轨迹时空关联语义和时态熵的移动对象社会角色发现[J]. 计算机研究与发展, 2012, 49(10): 2153-2160.
- [5] Yue Yang, Wang Handong, Hu Bo, et al. Exploratory Calibration of a Spatial Interaction Model Using Taxi GPS Trajectories [J]. Computers, Environment and Urban Systems, 2012, 36(2): 140-153.
- [6] Lee J, Han Jiawei, Li Xiaolei. Trajectory Outlier Detection: A Partition-and-detect Framework [C]//Proceedings of ICDE'08. Washington D. C., USA: IEEE Computer Society, 2008: 140-149.
- [7] 黄添强, 余养强, 郭躬德, 等. 半监督的移动对象离群轨迹检测算法[J]. 计算机研究与发展, 2011, 48(11): 2074-2082.
- [8] Zhang Daqing, Li Nan, Zhou Zhihua, et al. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces [C]//Proceedings of ACM International Conference on Ubiquitous Computing. New York, USA: ACM Press, 2011: 99-108.
- [9] Zhu Jie, Jiang Wei, Liu An, et al. Time-dependent Popular Routes Based Trajectory Outlier Detection [C]//Proceedings of WISE'15. Berlin, Germany: Springer, 2015: 16-30.
- [10] Lee J G, Han Jiawei, Whang K Y. Trajectory Clustering: A Partition-and-group Framework [C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2007: 593-604.
- [11] 陈锦阳, 宋加涛, 刘良旭, 等. 基于改进 Hausdorff 距离的轨迹聚类算法[J]. 计算机工程, 2012, 38(17): 157-161.
- [12] Jiang Shengyi, Li Qinghua. GLOF: A New Approach for Mining Local Outlier [C]//Proceedings of the 2nd International Conference on Machine Learning and Cybernetics. Washington D. C., USA: IEEE Computer Society, 2003.
- [13] 李光强, 郑茂仪, 邓敏. 时空数据异常探测方法[J]. 计算机工程, 2010, 36(5): 35-36.
- [14] 袁冠, 夏士雄, 张磊. 基于结构相似度的轨迹聚类算法[J]. 通信学报, 2011, 32(9): 103-110.
- [15] Piorkowski M, Sarafijanovic-Djukic N, Grossglauser M. CRAWDAD Data Set Epfl/Mobility/Cab [EB/OL]. (2009-02-24). <http://crawdad.cs.dartmouth.edu/epfl/mobility>.

编辑 陆燕菲