

## 轨迹大数据异常检测:研究进展及系统框架\*

毛嘉莉<sup>1,2</sup>, 金澈清<sup>1</sup>, 章志刚<sup>1</sup>, 周傲英<sup>1</sup>



<sup>1</sup>(华东师范大学 数据科学与工程学院, 上海 200062)

<sup>2</sup>(西华师范大学 计算机学院, 四川 南充 637009)

通讯作者: 毛嘉莉, E-mail: jlmiao1231@stu.ecnu.edu.cn

**摘要:** 定位技术与普适计算的蓬勃发展催生了轨迹大数据, 轨迹大数据表现为定位设备所产生的大规模高速数据流。及时、有效地对以数据流形式出现的轨迹大数据进行分析处理, 可以发现隐藏在轨迹数据中的异常现象, 从而服务于城市规划、交通管理、安全管控等应用。受限于轨迹大数据固有的不确定性、无限性、时变进化性、稀疏性和偏态分布性等特征, 传统的异常检测技术不能直接应用于轨迹大数据的异常检测。由于静态轨迹数据集的异常检测方法通常假定数据分布先验已知, 忽视了轨迹数据的时间特征, 也不能评测轨迹大数据中动态演化的异常行为。面对轨迹大数据低劣的数据质量和快速的数据更新, 需要利用有限的系统资源处理因时变带来的概念漂移, 实时地检测多样化的轨迹异常, 分析轨迹异常间的因果联系, 继而识别更大时空区域内进化的、关联的轨迹异常, 这是轨迹大数据异常检测的核心研究内容。此外, 融合与位置服务应用相关的多源异质数据, 剖析异常轨迹的起因以及其隐含的异常事件, 也是轨迹大数据异常检测当下亟待研究的问题。为解决上述问题, 对轨迹异常检测技术的研究成果进行了分类总结, 针对现有轨迹异常检测方法的局限性, 提出了轨迹大数据异常检测的系统架构。最后, 在面向轨迹流的在线异常检测、轨迹异常的演化分析、轨迹异常检测系统的基准评测、异常检测结果语义分析的数据融合以及轨迹异常检测的可视化技术等方面探讨了今后的研究工作。

**关键词:** 异常检测; 轨迹大数据; 概念漂移; 时变进化性

**中图法分类号:** TP311

中文引用格式: 毛嘉莉, 金澈清, 章志刚, 周傲英. 轨迹大数据异常检测: 研究进展及系统框架. 软件学报, 2017, 28(1): 17-34. <http://www.jos.org.cn/1000-9825/5151.htm>

英文引用格式: Mao JL, Jin CQ, Zhang ZG, Zhou AY. Anomaly detection for trajectory big data: Advancements and framework. Ruan Jian Xue Bao/Journal of Software, 2017, 28(1): 17-34 (in Chinese). <http://www.jos.org.cn/1000-9825/5151.htm>

### Anomaly Detection for Trajectory Big Data: Advancements and Framework

MAO Jia-Li<sup>1,2</sup>, JIN Che-Qing<sup>1</sup>, ZHANG Zhi-Gang<sup>1</sup>, ZHOU Ao-Ying<sup>1</sup>

<sup>1</sup>(Institute of Data Science and Engineering, East China Normal University, Shanghai 200062, China)

<sup>2</sup>(Computer School, China West Normal University, Nanchong 637009, China)

**Abstract:** The vigorous development of positioning technology and pervasive computing has given rise to trajectory big data, i.e. the high speed trajectory data stream that originated from positioning devices. Analyzing trajectory big data timely and effectively enables us to discover the abnormal patterns that hide in trajectory data streams, and therefore to provide effective support to applications such as urban planning, traffic management, and security controlling. The traditional anomaly detection algorithms cannot be applied to outlier

\* 基金项目: 国家自然科学基金(61370101, U1501252, U1401256); 上海市教委创新计划(14ZZ045); 西华师范大学国家级项目培育专项(16C005)

Foundation item: National Natural Science Foundation of China (61370101, U1501252, U1401256); Shanghai Municipal Education Commission Innovation Plan (14ZZ045); China West Normal University Special Foundation of National Programme Cultivation (16C005)

收稿时间: 2016-05-25; 修改时间: 2016-08-18; 采用时间: 2016-10-25; jos 在线出版时间: 2016-10-27

CNKI 网络优先出版: 2016-10-27 13:39:47, <http://www.cnki.net/kcms/detail/11.2560.TP.20161027.1339.001.html>

detection in trajectory big data directly due to the characteristics of trajectories such as uncertainty, un-limitedness, time-varying evolvability, sparsity and skewness distribution. In addition, most of trajectory outlier detection methods designed for static trajectory dataset usually assume a priori known data distribution while disregarding the temporal property of trajectory data, and thus are unsuitable for identifying the evolutionary trajectory outlier. When dealing with huge amount of low-quality trajectory big data, a series of issues need to be addressed. Those issues include coping with the concept drifts of time-varying data distribution in limited system resources, online detecting trajectory outliers, analyzing causal interactions among traffic outliers, identifying the evolutionary related trajectory outlier in larger spatial-temporal regions, and analyzing the hidden abnormal events and the root cause in trajectory anomalies by using application related multi-source heterogeneous data. Aiming at solving the problems mentioned above, this paper reviews the existing trajectory outlier detecting techniques from several categories, describes the system architecture of outlier detection in trajectory big data, and discusses the research directions such as outlier detection in trajectory stream, visualization and evolutionary analysis in trajectory outlier detection, benchmark for trajectory outlier detection system, and data fusion in semantic analysis for anomaly detection results.

**Key words:** anomaly detection; trajectory big data; concept drift; time-varying evolutionary

随着传感器网络技术、通信技术和定位技术的发展与日臻成熟,各类定位设备与手机等移动智能终端的广泛应用,使移动对象(人、车辆、轮船、动物等)的位置相关信息得以大规模采集.该类位置数据包含了地理坐标、速度、方向以及时间戳等信息,并以时变进化的形式持续增加且快速更新,被称为轨迹大数据.鉴于轨迹大数据可以准确地记录相当长一段时间之内移动对象的活动情况,可以客观地反映出移动对象个体(或群体)的活动规律,继而引发了数据科学、社会学及地理学等众多领域的学者的普遍关注.相关的研究工作有助于人们更好地理解对象动态演化的移动行为,预测其未来的移动趋势,并为支持基于位置的社交网络、智慧交通管理、城市规划、军事侦察等应用领域提供有效服务.以分析出租车轨迹数据辅助交通规划决策为例,出租车是一种非常重要的城市出行工具,它们通过预装的GPS设备频繁地(每隔0.5~2分钟)向数据中心报告当前的位置信息.因其广泛的长时间分布于城市路网中,海量的出租车轨迹数据集合能够很好地反映道路网络的交通运行状况,所以,出租车也被视为城市路网交通的“流动检测器”.近年来,许多研究工作对出租车的轨迹数据进行离线分析,服务于城市规划<sup>[1-5]</sup>、道路推荐<sup>[6-8]</sup>、交通热点分析<sup>[9]</sup>等领域.随着基于位置服务的应用需求的不断扩展,在线分析需求日益增多,需要在较短时间内实现快速处理和响应.“实时性”俨然成为轨迹大数据除“海量”之外的又一个重要特性.换言之,轨迹大数据以“轨迹流”的形式持续到达,需要被实时处理以便及时发现其蕴含的移动模式.区别于静态轨迹数据库,轨迹流中的数据分布会随着时间的推移而发生演变,各移动对象的移动模式在不同时段表现不一,短时间内难以对动态变化的位置数据建立先验知识,无法使用静态轨迹数据集合的索引技术加以维护.目前,学术界已出现一些针对实时轨迹数据查询处理的研究工作,包括可伸缩的快速轨迹聚类<sup>[10]</sup>及轨迹流聚类<sup>[11-13]</sup>、轨迹流的连续查询<sup>[14]</sup>、热门路径发现<sup>[15]</sup>、汇集模式发现<sup>[16]</sup>、旅伴模式发现<sup>[17,18]</sup>以及实时的个性化拼车<sup>[19]</sup>等.

基于轨迹数据的模式发现旨在从海量轨迹集合中提取诸多移动对象的共性特征,与之相对应的另外一类工作则是面向轨迹大数据发现异常模式,这在许多LBS应用中非常关键.异常,也称为孤立点、新颖点、偏离点、例外点等,异常检测在统计学和数据分析等领域中并不鲜见.Hawkins将孤立点定义为:在数据集中表现与众不同的数据,使人怀疑这些数据并非随机偏差,而是产生于完全不同的机制<sup>[20]</sup>.孤立点的出现常常由于人为误差、仪器误差、源于异类的数据(如欺诈、入侵等)、系统行为改变(如气候变化、顾客的新购买模式、基因突变等)所致,从而表现出与正常行为不一致的现象.异常并非噪声,尽管噪声与异常极为相似,但噪声会降低数据集合的质量,是数据分析的绊脚石;而异常往往预示着有趣事件的发生,从而具有更高的研究价值.因此,在实际应用中,常常需要在数据预处理阶段进行去噪处理.异常检测在数据库、数据挖掘、机器学习和统计学等领域有着广泛应用,包括信用卡或保险业的欺诈检测、网络中的入侵检测及故障诊断、卫星图像分析中的新特征识别、健康医疗监控、公共安全中的突发事件的发生、药物研究中新型分子结构的识别等.以城市交通管理中的异常检测为例,在趋于饱和的城市道路网络中,交通事故、恶劣天气和道路紧急事件等,均会造成道路交通异常由点到面的迅速发展和蔓延,导致整体路网的拥堵或瘫痪.传统的智能交通监控通过定点部署视频检测器或感应线圈等磁性检测器,检测交通流参数的异常.由于此类设备需要昂贵的基础设施投资及维护开销,无法密集安装覆

盖整体路网,导致部分区域缺失或获得不可靠的监控信息.基于城市路网中出租车、公交车等车辆的轨迹数据,使用有效的异常检测方法对其进行实时分析,不仅可以发现交通参与者的路径选择模式的异常情况、识别交通拥堵路段,还能检测路网变化,发现新路以更新交通地图.

目前,对于轨迹数据的异常检测已经开展了深入的研究,但是由于轨迹大数据具有不确定性、稀疏性、偏态分布性、规模大以及更新快等特征,面向轨迹大数据的实时异常检测的研究相对较少<sup>[21,22]</sup>.

- (1) 轨迹数据的不确定性:由于定位技术的有限精度以及 GPS 定位设备的计算误差、信号衰减及丢失导致采集的位置数据具有空间不确定性;同时,不同采集频率或不同时序长度也会带来位置数据的时序不确定性.空间、时序的不确定性,使得采集到的轨迹数据具有大量的位置偏差,即噪声,从而降低了轨迹异常检测结果的准确性;
- (2) 轨迹数据的稀疏性和偏态分布性:轨迹数据反映了移动对象的活动规律,而移动对象(包括人、动物等)的活动普遍具有周期性,因此,轨迹数据常常表现为不均匀分布.例如,在城市路网中,部分路段在很长时间内仅有少量车辆通过,而少数主干道在短时间内却有大量车辆经过,从而导致轨迹数据呈现稀疏的、偏态分布的特征;
- (3) 轨迹数据的大规模及快速更新:移动对象的位置数据实时地产生且持续增加.只要移动对象处于活动状态,位置信息就会不断产生和累积,因此,轨迹的数据量是无限的,没有固定长度.但是,在线异常检测算法在数据的计算过程中无法保存采集到的全部轨迹数据:一方面,硬件设备没有足够大的存储空间存储无限增长的轨迹数据;另一方面,也缺少合适的方法有效管理轨迹大数据;并且,对于快速到达的轨迹数据,难以精确地定义移动行为的异常特征及提出有效的实时异常检测方法.

异常发现是数据管理与分析的重要内容.迄今为止,国内外研究者相继对异常检测技术进行了综述.Hodge<sup>[23]</sup>,Chandola<sup>[24]</sup>与 Aggarwal<sup>[25]</sup>等人分别对异常检测的方法学和各类典型的异常检测算法进行了分析.Zhang 等人<sup>[26]</sup>回顾了无线传感网络中的异常检测技术.Gupta 等人<sup>[27]</sup>着重分析了面向时序数据的异常检测技术.然而,他们的研究工作并非面向轨迹数据集合.一些轨迹挖掘的综述文献<sup>[28,29]</sup>将轨迹异常检测问题仅作为挖掘技术的一部分,对现有的轨迹异常检测技术缺少详尽的分析与总结.与上述工作不同,本文重点关注面向轨迹大数据的异常发现,回顾了基本的异常检测技术,并按照不同类别对轨迹异常检测技术的研究进展进行了概述,同时指出其存在的局限性及未来的研究方向.

本文第 1 节介绍传统的异常检测技术.第 2 节对轨迹异常检测的概念、模型及轨迹异常检测技术的分类进行概述.第 3 节~第 6 节对基于分类的轨迹异常检测技术、基于历史相似性的异常检测技术、基于距离的异常轨迹检测技术以及基于网格划分的异常轨迹检测技术的基本原理分别进行总结阐述.第 7 节对 13 种具有代表性的轨迹异常检测方法进行对比分析,指出它们存在的局限性,提出轨迹大数据异常检测系统的基本架构,并对未来的研究工作探讨.最后对全文进行总结.

## 1 异常检测概述

异常检测技术由 3 个子问题组成:① 异常定义;② 提出异常检测的方法;③ 解释异常检测的结果.一般来说,可以先定义正常行为的范围,再以此为依据判定与之相异的行为是否异常.但是,受限于不同的应用领域,确定一个通用的涵盖所有异常行为的孤立点定义是非常困难的.现有的异常检测方法主要涵盖基于分布的检测算法<sup>[30]</sup>、基于偏差的检测算法<sup>[31]</sup>、基于深度的检测算法<sup>[32]</sup>、基于距离的检测算法<sup>[33-36]</sup>以及基于密度的检测算法<sup>[37,38]</sup>等.

基于分布的检测方法首先假设给定的数据集符合某种概率分布模型(例如正态分布,也可基于数据集自动产生),再采用不一致性检验来判断是否与该概率分布模型匹配,从而确定孤立点.基于偏差的检测方法主要识别连续序列中明显不同于其他数据的对象,该方法需要事先知道数据的主要特征,不适合处理大规模多属性的现实复杂数据.基于深度的检测方法将数据对象映射到多维空间,并为每个数据点赋予一个深度值.将数据对象按分配的深度值分层组织,处在浅层上的数据对象比处于深层上的更可能成为孤立点.基于深度的方法只适

合处理低维空间数据(三维以下),而对于四维及以上属性的大数据集处理低效.基于距离的检测方法将数据集中远离大多数对象的数据点作为异常,可用于处理高维数据集,但其难点在于距离函数的选择.此外,该方法不适合处理具有不同密度分布的数据.基于密度的局部异常检测算法不再把异常视为二元属性,根据异常与待测对象的局部近邻密度的相关性,采用局部异常因子 LOF 表示对象的异常程度,将具有高 LOF 值的对象视为异常对象.该方法需要执行大量的近邻查询,时间复杂度较高,并且参数选择比较困难.

与传统的孤立点检测技术相比,空间孤立点检测(SOD)需要结合空间持续性以及与近邻对象的自相关性,旨在识别非空间属性异于空间邻域中其他空间相关对象的数据点.主要分为两类检测方法:基于全局的 SOD 方法和基于局部的 SOD 方法:基于全局的 SOD 方法使用地理统计数据的最佳线性无偏估计量识别孤立点,适于中小规模低维数据的全局孤立点检测;基于局部的 SOD 方法首先计算各对象的局部差异,即,该对象在非空间属性部分与其空间近邻对象统计的聚集值之间的差异,随后,根据独立同分布的假设,通过鲁棒性地评价诸如聚集值、平均值以及标准差等模型参数发现孤立点.相对于全局 SOD 方法,局部 SOD 方法具有更高的精确性以及计算效率.Sun 等人<sup>[39]</sup>提出了空间局部异常测量 SLOM,能够捕捉空间数据固有的空间自相关性以及异方差性.结合数据点的局部稳定性,通过比较被检测对象的空间近邻的局部数据行为,为每个数据点计算局部异常程度(SLOM 分值),并将具有较高 SLOM 值的数据点作为孤立点.Chen 等人<sup>[40]</sup>系统地研究了基于局部的 SOD 方法的统计特征,提出了一个广义的局部统计框架(GLS),可有效处理空间数据的全局趋势或局部差异间的依赖关系.

时空数据以时间、空间和非时空属性作为实体固有的基本特征,反映了实体随时间推移而产生的空间状态演变过程.时空数据孤立点(ST-outlier)是指时空对象的行为特征(包含非空间、非时间属性)明显不同于其时空邻域中的其他对象.大多数时空异常检测技术通过先发现空间孤立点再比较其时序近邻的方式,以确认时空孤立点.Birant 等人<sup>[41]</sup>提出了包含聚类、空间近邻检测、时序近邻检测的三阶段检测方法,将非空间属性明显不同于其时空邻域内其他对象的数据作为异常.Cheng 等人<sup>[42]</sup>提出了包含聚类、聚集、比较、确认的四阶段检测方法,通过评价连续时空范围内的变化来检测主题属性异于空间邻域/时间邻域内其他对象的时空异常.Adam 等人<sup>[43]</sup>采用基于距离的异常检测方法,通过寻找微邻域、合并微邻域产生宏邻域,进而识别宏邻域的时空异常.考虑到时空异常不仅出现在一个时期,还可能在形状大小等方面不断进化,Wu 等人<sup>[44]</sup>提出了时空孤立点检测算法 Outstretch,通过检测各时期前  $K$  个高差异区域、发现并存储随时变产生的异常序列、提取高差异区域序列以及子序列等任务实现时空孤立点的检测.

## 2 轨迹异常检测:定义及方法

### 2.1 轨迹与轨迹异常

轨迹是一种重要的时空数据类型,它代表了移动对象持续移动的位置信息历史.轨迹可视为时间到空间的映射,给定某一个时刻  $t(t \in R^+)$ ,通过一个以时间为自变量的连续函数  $F$ ,可以得到该对象在  $t$  时刻所处的  $d$  维空间  $R^d$  中的位置,即  $F: R^+ \rightarrow R^d$ .令  $Tr$  表示某移动对象在二维空间内的轨迹,  $Tr = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_n, y_n, t_n), \dots\}$ , 其中,  $(x_i, y_i, t_i)$  表示移动对象在  $t_i$  时刻的位置信息(经纬度).异常轨迹(或称为轨迹孤立点)既可被看作不遵守某种预期模式的事件,也可被看成是根据相似性准则表现(旅行时间、数据分布等)与其他对象的行为相异.典型案例包括:在同一片海域行为异于其他渔船的船只、风向突变的飓风、出租车的绕路行为、由意外交通事故引发的部分路段堵塞等.因此,在不同应用中,异常被描述为异常的子轨迹<sup>[21,45-50]</sup>、异常的轨迹<sup>[33-35,49,51]</sup>、异常的移动对象<sup>[22]</sup>、异常的路径<sup>[52-55]</sup>、异常的路网区域<sup>[55]</sup>以及异常的事件<sup>[55]</sup>等不同定义.

### 2.2 轨迹异常检测

#### 2.2.1 轨迹的相似性准则

定义轨迹相似性准则是大多数轨迹异常检测技术面临的关键问题,通常以轨迹点或轨迹线段之间的距离测量描述轨迹间的相似程度.研究者针对不同应用提出了几类典型的轨迹相似性度量方法,包括欧氏距离<sup>[56]</sup>、DTW<sup>[57]</sup>、LCSS<sup>[58]</sup>、ERP<sup>[59]</sup>、EDR<sup>[60]</sup>、Hausdorff<sup>[61]</sup>等.欧氏距离用于计算长度相同的两条轨迹间各时间点对应

位置坐标点的距离的聚集值,该方法计算简单,但不适于长度不等、采集频率不同和时间尺度不一致的轨迹数据,且易受噪声数据影响。鉴于位置采集设备的测量误差导致所有轨迹的位置信息采集时刻不能保证一一对应,DTW 方法选择将时间维度拉伸,可用于任意长度轨迹的距离度量。其基本思想是,允许位置序列中的数据元素根据度量的需要重复计算多次。换言之,在保证位置序列中数据元素顺序不变的前提下,通过复制前一个位置点完成时间维的局部缩放,并以计算得到的最小距离作为轨迹间的相似性度量。由于距离计算时两条轨迹包括噪声点在内的所有位置点都要求匹配,所以该方法对噪声数据极为敏感。为消除轨迹中的噪声数据可能导致轨迹间较大距离的影响,LCSS 方法以计算轨迹间最长的相似子轨迹序列来评测轨迹间的相似性,其基本思想是:允许略过一些位置点而不是重排,并通过剔除一些距离较远的位置数据来减少噪声数据的影响。尽管 LCSS 能够处理有噪声的轨迹数据,但是该方法不考虑轨迹间相似子序列的不同间隙大小,降低了相似性测量的准确度。

编辑距离的基本思想是,计算两条轨迹相同所需的最小开销。基于编辑距离的 ERP 类似于 DTW,可以局部缩放时间维并且适用于不同尺度的轨迹间距离测量。但与 DTW 不同的是:在校正数据遇到不匹配情况时,ERP 使用一个位置常量来评测轨迹不匹配位置点,将位置常量与不匹配点的距离作为测量的距离。该方法支持局部时间位移,却容易受噪声数据的影响。基于编辑距离的 EDR 以正态化处理来解决空间维缩放,它使用匹配阈值通过量化轨迹元素对之间的距离(0 或 1)来降低噪声数据对测量距离的影响。此外,对于两段相似的子轨迹间不相似的部分,EDR 可以根据间隙长度分配不同的惩罚距离,这使其比 LCSS 距离测量更加精确。DTW,LCSS,EDR,ERP 都是基于欧氏空间的相似性度量,它们关注轨迹的全局相似性,不能获得短时间内子轨迹的局部相似性。同时,因其计算复杂度较高,无法直接用于空间受限应用(如道路网络)的轨迹异常检测。Lee 等人<sup>[62]</sup>提出了一种用于子轨迹相似性度量的改进的 Hausdorff 距离计算方法,描述为两条轨迹之间垂直距离、平行距离以及角度距离的加权和。Hausdorff 距离原用于测量一个集合中数据点与另一集合中最近数据点的所有最短距离的最大值,因其独有的方向特性,还被 Roh<sup>[63]</sup>与 Han<sup>[64]</sup>等人用于道路网络中评测不同方向轨迹数据间的相似性,表现为路段间的最长距离。

### 2.2.2 轨迹异常检测方法的分类

现有的轨迹异常检测技术基于不同的应用需求以及不同的实现原理,大致可以划分为 4 类:基于分类的检测方法<sup>[45,46,65]</sup>、基于历史相似性的检测方法<sup>[47,52-55,66,67]</sup>、基于距离的检测方法<sup>[21,22,33-36,48,62]</sup>以及基于网格划分的检测方法<sup>[49-51,68-71]</sup>,如图 1 所示。其中,最直接的轨迹异常检测方法是基于训练标签数据集学习构建一个可区分噪声、异常以及正常轨迹的分类模型。

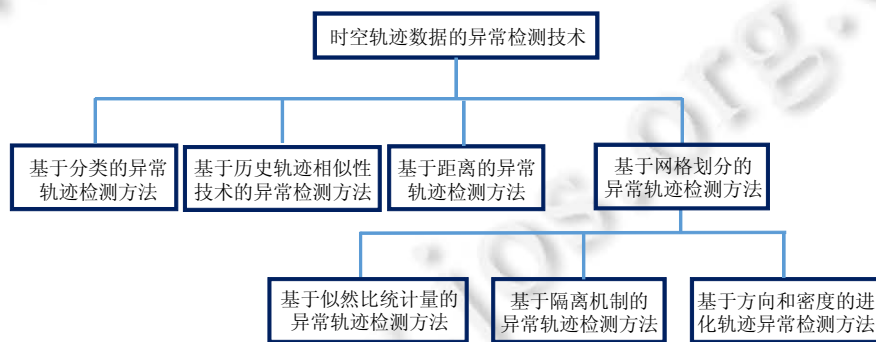


Fig.1 Taxonomy of trajectory outlier detection techniques

图 1 时空轨迹数据的异常检测技术分类

但在实际应用中,很难获得一个覆盖所有异常行为类型的标签训练数据集。在训练标签数据集缺失的情况下,基于历史相似性的检测方法根据采集到的轨迹数据建立用于异常检测的全局特征模型,可以凭借该模型判定其他轨迹数据是否异常。但是此类方法并未考虑随着新增轨迹数据增量更新原有异常检测模型,不能评测轨迹异常的演化行为。基于距离的异常检测方法不需要明确的数据分布,通过距离计算即可识别异常轨迹。但该类

方法涉及轨迹子序列间大量的近邻查询,需要较高的时间开销和存储空间开销.基于网格划分的方法通过将城市地图划分成均等大小的网格单元形式,把异常轨迹的检测转变成异常网格单元符号序列的检测问题.根据所采用的技术的不同,可分为基于似然比统计量的检测方法、基于隔离机制的检测方法以及基于方向和密度的进化轨迹异常检测方法.基于似然比检验统计量的异常检测方法主要识别明显偏离期望行为的邻近网格单元集以及时段区域.基于隔离机制的检测方法通过一定的隔离步骤,识别以少量步骤隔离开的轨迹为异常.基于方向和密度的进化轨迹异常检测方法用于发现方向、密度等不同特征在一段时间内逐渐进化演变而来的异常轨迹.在第3节~第6节,本文将分别阐述这4类轨迹异常检测技术的基本原理.

大部分轨迹异常检测方法只考虑轨迹的空间特征,基于内存处理数据,需要多次扫描数据集,适于静态轨迹数据集的异常发现.由于轨迹是进化时变的位置数据,没有固定的数据分布模式,因而存在概念漂移.轨迹异常通常具有局部性、多样性、进化性、相互关联性等特点,因此亟需高效的检测方法在线识别异常轨迹.

### 3 基于分类的轨迹异常检测

在轨迹数据集中,判定异常轨迹的标准是检测其是否表现与数据集中大多数轨迹的全局特征不一致.基于分类的轨迹异常检测技术通常分为两个阶段.

在训练阶段,使用标签训练集数据学习构建分类器;

在测试阶段,根据分类器将测试实例划分为正常以及异常的两类对象.

基于收集的海量轨迹数据,根据异常模式与时间和位置等属性粒度的相关性,李晓磊等人<sup>[65]</sup>提出了 Motion-Alert 算法,从对象移动路径中提取与时间、位置相关的移动特征 motif.每条路径由位置点序列转变成由一系列的移动 motif 组成,因此,移动路径集被转换成基于 motif 的特征空间.再使用支持向量机学习方法对提取的 motif 进行特征学习,产生用于处理高维特征空间的异常检测分类器.鉴于异常与时空特征多维粒度的相关性,在 Motion-Alert 算法的基础上,李晓磊<sup>[45]</sup>提出了异常轨迹检测算法 ROAM,如图2所示.它先使用 K-means 算法对分段轨迹数据基于各相关属性的 motif 构建了一个多维特征空间,再通过检查轨迹的模式,自动提取特征空间中的多个层次,构建了一个数据的多分辨率视图,有助于分析不同粒度不同维的多个轨迹特征间的复杂关系.最后,建立了一个基于规则的分类器,可分析处理高维特征空间以检测异常轨迹.

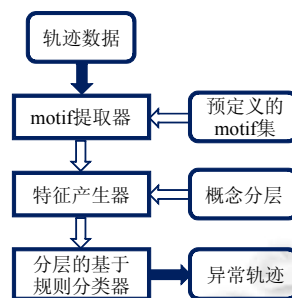


Fig.2 Framework of rule- and motif-based anomaly detection

图2 ROAM 框架

考虑到大多数异常轨迹检测技术以发现整体异常轨迹为目标,容易忽视轨迹子段中可能出现的局部异常行为或异常事件,在视频监控应用中,高阳等人<sup>[46]</sup>提出了一个三阶段的局部异常检测框架 TRASMIL. TRASMIL 先将轨迹划分为独立的子段,并使用连续概率模型对各子段建模表示,再以整条轨迹为包(正常轨迹为负包、异常轨迹作正包)、轨迹子段作为包的示例,运用多示例学习方法构建分类器来检测轨迹的局部异常行为.

在给定好的训练数据集的前提下,基于分类的异常检测技术通常可以获得比无监督技术更好的检测精度.然而,为数据附上正确的标签需要专家人工标注,要获得能够精确代表所有类(包括正常和异常)行为的标签数据涉及较高的计算开销.此外,鉴于进化轨迹流中的异常通常是未知的、时变的,实际应用中,不可能找到涵盖所



有异常实例的训练数据.因此,基于分类的异常检测方法不适用于轨迹流的在线异常检测.

#### 4 基于历史轨迹相似性方法的轨迹异常检测

在静态轨迹数据集中,轨迹的数据分布通常是先验已知的,在缺失标签训练数据集的情况下,可根据历史采集到的大量轨迹数据挖掘所有频繁模式建立全局特征模型,再将异于全局特征模型的数据识别为异常轨迹.在不考虑轨迹数据的时变进化特征时,基于一定数量的历史轨迹数据构建的全局特征模型具有较高的检测精度,因此,基于历史轨迹相似性的异常检测技术在航海、路网交通等领域应用广泛.

针对海上船只移动的随意性以及所产生轨迹的不确定性,Lei 等人提出了 MT-MAD<sup>[47]</sup>方法用于海量航海船只轨迹数据的异常检测.如图 3 所示.

- 在轨迹建模阶段,首先提取所有船只频繁访问的区域,基于频繁区域集将每个轨迹转变成基于区域的移动序列;随后,基于轨迹的空间和序列特征发现频繁序列模式并对相似的行为进行分类,构造模型获取基于区域的轨迹所具有的移动行为;
- 在异常检测阶段,对于每个对象在每个时刻产生的位置数据,使用学习的模型,根据轨迹的 3 种异常特征(如空间、序列、行为特征)对其进行特征化;最后,综合 3 种异常特征分数值使用加权平均的方法,对所有移动对象的子轨迹的异常可疑程度进行排序,将异常可疑程度超过给定阈值的对象视为异常.

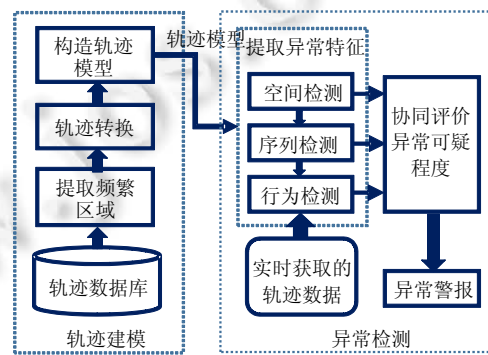


Fig.3 Framework of MT-MAD

图 3 MT-MAD 框架

面向出租车轨迹数据,有大量研究基于轨迹的历史相似性提出了不同的应用.Zhu 等人<sup>[72]</sup>基于轨迹的历史相似性,提出了一个基于轨迹孤立点分析提取时间依赖的受欢迎路径(TPRO)方法.通过对轨迹按源-目的地分组提取特定时段内最热门的  $K$  条路径,同时将与相应时段热门路径差异较大的轨迹视为异常.在发现出租车司机的欺诈绕路行为应用中,Ge 等人<sup>[66]</sup>开发了一个出租车驾驶欺诈检测系统,用于检测基于旅行路线以及基于旅行距离的两类驾驶欺诈行为.Liu 等人<sup>[67]</sup>开发了基于速度的欺诈检测系统(SFDS),采用基于速度信息的聚类方法对出租车的正常行驶行为建模,用于检测出租车的欺诈行为.

在城市路网的交通异常检测应用中,Li<sup>[52]</sup>,Liu<sup>[53]</sup>,Chawla<sup>[54]</sup>等人基于轨迹的历史相似性提出了不同的异常检测方法.Li 等人<sup>[52]</sup>基于数据点的历史相似近邻,提出了一种针对车辆交通数据的时序孤立点检测算法(TOD).Liu 等人<sup>[53]</sup>通过将城市路网按主要道路划分成不连贯的区域,提出了 STOTree 算法来检测路网中的时空孤立点及孤立点之间的因果关系.Chawla 等人<sup>[54]</sup>提出了一个两阶段的挖掘优化框架,通过对区域之间的交通(不是道路之间)进行建模,检测区域之间偏离历史轨迹的异常链接,推断导致异常链接出现的路径,并揭示异常产生的缘由.

上述方法在采集一定数量的轨迹数据后,建立的特征模型具有较高的异常检测精度,但是随着轨迹数据的不断增加,没有考虑对原有特征模型进行增量更新,若采用阶段性重建特征模型将引致较高的计算开销,因此无

法用于在线评测不断进化的异常轨迹.鉴于此,Pan 等人<sup>[55]</sup>提出了一个交通异常检测框架,用于发现路网中与异常相关的整个子图,如图 4 所示.该方法由离线挖掘、在线异常检测以及异常分析这 3 个阶段组成.与基于交通流量的轨迹异常检测方法不同,该方法旨在识别路网中司机的路径选择行为明显不同于历史路径行为模式的路段子图.对于最新到达的车辆轨迹,在异常检测阶段,不仅可以利用离线挖掘阶段构建的历史路径行为模型实时检测异常路段,还可利用新增轨迹更新路径行为模型.此外,对于异常检测结果,通过引入社交网络平台数据进行术语挖掘,继而发现隐含的异常事件.

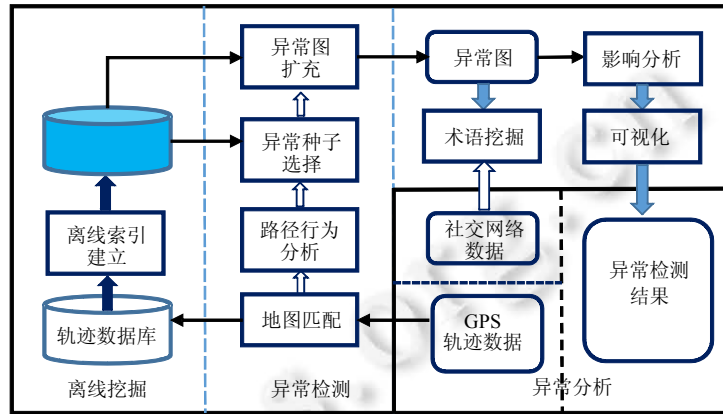


Fig.4 System architecture for traffic abnormality detection

图 4 交通异常检测系统框架

## 5 基于距离的轨迹异常检测

基于距离的轨迹异常检测方法,将轨迹数据集中与大多数轨迹具有较远距离的轨迹视为异常.根据所处理的轨迹数据分布是否先验已知,可分为面向静态轨迹数据集的异常检测技术和面向轨迹数据流的异常检测技术两类.

### 5.1 面向静态轨迹数据集的异常检测

Knorr<sup>[33-35]</sup>首次提出了基于距离的方法检测异常轨迹,轨迹被描述为一个由开始(结束)位置、方向(具有平均值、最小值与最大值)和速度(具有平均值、最小值与最大值的)这 3 个关键特征代表的对象.选择轨迹间特征值的差异加权和作为轨迹相似度准则,通过轨迹间相似性计算检测异常轨迹.此类方法以整条轨迹作为异常检测的基本单位,对于实际应用中长而复杂的轨迹,表现异常的轨迹分段易被整个轨迹平均化,因此,该算法只适用于检测方向、开始(结束)位置、速度完全不同于其他轨迹的孤立点.

为检测如图 5 所示的异常子轨迹, Lee 等人<sup>[48,62]</sup>提出了基于分割检测框架的检测算法 TRAOD.通过二级轨迹划分策略(粗粒度、细粒度)将轨迹划分成线段集,以线段描述轨迹局部特征,结合基于密度的方法,以“与大多数轨迹不存在最小长度的距离”作为异常轨迹划分的判定标准,并根据异常的局部特征在整个轨迹中所占比例进而识别异常轨迹. TRAOD 算法需要预计算各轨迹划分的密度调整系数,存在计算量大、时间开销大、结果准确性不够高等问题.因此,该算法不适用于轨迹流的在线异常检测.

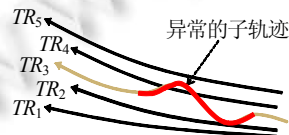


Fig.5 An example of an outlying sub-trajectory

图 5 异常的子轨迹



刘良旭等人<sup>[73]</sup>基于距离的方法提出了一种基于 R-Tree 的异常轨迹检测算法,首先利用 R-Tree 和轨迹间的距离特征矩阵筛选出所有可能匹配的基本比较单元对,再通过计算距离确定其是否局部匹配.为了克服以轨迹分段表示轨迹局部特征所存在的不足,文献[74]进一步提出了以轨迹点表示轨迹局部特征的异常轨迹点检测算法 TraLOD.引入局部异常度表示轨迹点的异常情况,并以相对距离计算轨迹片段之间的不匹配性,使得距离度量更加符合实际意义.基于距离的异常检测方法需要执行轨迹子序列间大量的距离计算,涉及较高的时间开销和存储空间开销,在轨迹流中检测异常轨迹时,必须引入剪枝策略以提高检测效能.

## 5.2 面向轨迹数据流的异常轨迹检测

为持续监控各个时段内明显不同于其空间近邻的异常子轨迹(与正常子轨迹有较大的空间距离偏差),Bu 等人<sup>[21]</sup>结合滑动窗口技术,采用基于距离的异常检测方法,提出了一种用于轨迹流检测异常子轨迹的方法.首先为当前时间窗口内的新到轨迹段在其左右时间窗口(临近时间区域)中寻找同等大小的轨迹段近邻,通过统计近邻,并将近邻数未达到指定阈值的判定为异常轨迹段.为了减少轨迹段之间的距离计算量,基于移动对象在短时间内表现一致的假设,即,轨迹具有局部持续性,采用局部簇对轨迹流进行时序划分.具体来说,通过将限定时段区域内与中心基窗口一定距离阈值范围内的基窗口的轨迹子序列集作为局部簇,将子轨迹间的距离计算转变为局部簇之间的簇集连接操作,减少了大量的距离计算.为进一步提高性能,使用基于分段 VP 树索引结构重新安排簇集连接顺序.

文献[21]仅考虑了检测单个移动对象的异常子轨迹,然而在大规模轨迹流中,多个移动对象的移动模式比单个移动对象的移动路径更为复杂.基于一个移动对象与轨迹流中其他移动对象的近邻关系,Yu 等人<sup>[22]</sup>基于滑动窗口提出了一种检测轨迹流中异常移动对象的增量 INC 算法.首先,结合空间近似性及保持空间近似性时长,提出了轨迹近邻的概念,根据不同同步机制定义了两种近邻概念(点近邻和轨迹近邻),进而提出基于点近邻和轨迹近邻的两种异常定义.基于点近邻的异常检测将一定时间间隔内没有足够数量的点近邻的移动对象作为异常;基于轨迹近邻的异常检测将当前时间窗口中没有足够数量轨迹近邻的移动对象视为异常.为消除 INC 算法优先范围查询策略的影响,减少近邻范围查询导致的 CPU 及内存开销,提出了 MEX 优化框架,引入了 3 种优化原则,包括最小支持检查原则(MSE)、时间意识的检查原则(TAE)以及终身触发检测原则(LTD).

以上两种基于距离方法的轨迹流异常检测技术旨在发现某时刻或某时段内的异常子轨迹、异常移动对象,并未关注轨迹流中进化的异常轨迹.此外,它们只强调轨迹本身基于位置信息的异常行为,忽视了在非位置信息方面明显异于其时空近邻的轨迹.

## 6 基于网格划分的轨迹异常检测

在道路交通的异常研究中,有部分应用采用基于网格划分的轨迹异常检测技术.该类方法旨在从划分成均等大小网格单元的城市路网中识别异常的网格单元序列.根据所采用的技术的不同,可分为基于似然比检验统计量的异常检测、基于隔离机制的异常检测以及基于方向和密度的进化异常检测技术.

### 6.1 基于似然比检验统计量的异常轨迹检测

为挖掘道路交通流中的异常模式,辅助判定路网中发生的非预期事件,Pang 等人<sup>[68]</sup>提出了带参数的基于似然比检验统计量的异常检测方法,识别明显偏离期望行为的邻近网格单元集及时段区域.首先统计一定时间内各网格到达的车辆数,根据用户特定的随机似然函数,对网格中的所有矩形区域进行 LRT 测试并排序,返回与期望行为有最大统计差异,即,最高分值所在的少数矩形区域作为异常.该方法提供了用于发现持续异常以及新兴异常的两类统计模型,并设计了剪枝方法以减少需要检查计算 LRT 的矩形区域.文献[69]使用似然比检验统计量描述交通模式并建立统计模型,进而识别一定时间间隔内具有最大偏离预期行为的异常连续网格区域.

### 6.2 基于隔离机制的异常轨迹检测

根据异常轨迹固有的“稀少而不同”这一特性,即:相对于正常的轨迹,异常轨迹往往表现为数量稀少并异于其他大多数正常轨迹模式.具体来说,对于相同的源-目的地对,异常轨迹表现为与正常轨迹不同的位置序列、或

以不同的顺序通过相同位置等形式.利用异常对隔离机制的敏感性,Zhang 等人结合 iForest 算法与懒惰学习方法提出了 iBat<sup>[51]</sup>检测算法,识别出租车轨迹数据的异常驾驶行为,继而揭示出租车驾驶欺诈或路网改变.通过使用基于隔离的异常轨迹检测方法,iBat 随机选择网格单元,将轨迹集划分成包含该单元以及未包含该单元的数据,直到所有轨迹被隔离完毕.根据隔离异常轨迹的网格单元数通常小于隔离正常轨迹的网格单元数的假设,iBat 方法将以较少步骤迅速隔离的轨迹识别为异常.

为实现在线检测异常子轨迹,Chen 等人<sup>[49]</sup>将检测异常所需的网格单元数定义为窗口大小,并基于固定窗口大小及自适应窗口大小提出了 iBOAT 方法.通过对不同轨迹给定一个即时异常分值阈值,以实时识别异常的子轨迹(根据窗口大小内遍历的网格单元符号序列出现的频率与给定的阈值相比,或者说给定的子轨迹在所属的轨迹中是否有足够的频率发生),同时,计算一个进化的异常分值以进一步判定异常路线偏离正常路线的程度.文献[70]进一步分析了 iBOAT 方法中异常阈值及历史轨迹数据集的大小对检测性能的影响,此外,还提供了检测精度及检测成本(计算时间和内存开销)之间的折中办法.为了减少 iBOAT 方法的实时响应时间,Sun 等人<sup>[71]</sup>提出使用倒排索引机制,设计实现了一个可实时检测异常的乘客交付行为的出租车系统.通过提取异常行为的共同特征,揭示欺诈行为背后的动机.

### 6.3 基于方向和密度的进化异常轨迹检测

为识别轨迹数据中基于方向和密度的两类进化的轨迹异常,Ge 等人<sup>[50]</sup>提出了一种提取前  $K$  个进化的异常轨迹的检测算法 TOP-EYE.对于方向异常,首先使用概率模型为各网格中轨迹提取一个描述轨迹移动方向的八值向量,再根据历史轨迹获得该网格轨迹的方向趋势,从而将与所在网格区域方向趋势不同的对象视为异常.而对于密度异常,先将通过每个网格的轨迹数作为各网格的轨迹密度,再基于轨迹实际通过的网格密度计算得到其异常分数.如果该轨迹通过的网格区域密度低于指定的阈值,则该轨迹为密度异常.除了可实时检测基于方向和密度的两类异常轨迹之外, TOP-EYE 还实现了进化异常轨迹的检测.考虑到衰减函数可以降低较早到达数据对进化异常分值的影响以及增加最近到达数据的权重,将最新时刻计算得到的异常分值与经过指数衰减函数处理的历史异常分值进行累加,得到其进化的异常分值.最后, TOP-EYE 方法提取进化异常分值最高的前  $K$  个轨迹作为异常轨迹.该方法借助于进化的异常分值可以有效地辨识噪声和异常.

## 7 未来工作探讨

### 7.1 现有工作的局限性

近 10 年来,轨迹异常检测技术在诸多应用领域提出了大量的检测算法,但是它们针对不同的应用背景提出了不同的异常定义及检测方法.我们从轨迹数据时空相关性、检测的异常类型、异常检测输出结果方式、异常检测处理方式这 4 个方面对 13 种具有代表性的轨迹异常检测方法进行对比分析,比较结果见表 1.它们分别以发现异常子轨迹、异常轨迹、异常的移动对象、异常路段、异常区域、异常事件为目标,结合轨迹数据的空间(或时序)特征,对轨迹数据进行离线(或在线)处理,并以标量(或异常分数)等形式输出检测结果.尽管它们能够有效地发现各类应用中定义的轨迹异常,但仍然存在以下局限性.

(1) 基于历史轨迹数据提取的异常检测模型不能有效地评测新增轨迹数据蕴含的新型异常模式.

部分检测方法依赖于历史搜集的轨迹数据构建用于预测的异常轨迹检测模型.面对无法预知整体数据分布的时变进化的轨迹数据流,此类方法不能结合新增轨迹数据的时空特征动态更新异常轨迹检测模型,而是采用代价较高的周期重建模型的方法.此外,针对历史轨迹数据集的异常检测方法多以离线检测方式,检测结果不能及早警示.而实际应用中,更需要在线实时检测异常轨迹.因此,轨迹数据的历史相似性只能作为异常轨迹检测的辅助评测标准.通过离线分析历史轨迹数据学习产生的异常检测模型,必须结合新增轨迹数据对异常检测模型进行增量更新,才能获得较好的在线异常检测效能.

(2) 以恒定阈值评测异常的检测方法不能有效区分噪声和异常,不适于轨迹大数据的进化异常检测.

大多数异常轨迹检测方法依赖领域知识通过设置特定的阈值来识别异常轨迹,但是通常很难确定一个恰

当的阈值,需要对特定数据集经过多次实验调优计算得到.在阈值设置不当的情况下,极有可能将与异常具有相似行为的噪声数据误判为异常轨迹,或者漏检异常轨迹,从而降低轨迹异常检测的准确性.此外,轨迹数据随时间变化不断进化,不能使用恒定的阈值准确识别更多未知数据分布的轨迹异常.因此,异常检测方法不仅需要设置阈值关注检测某时刻的异常轨迹,还要结合新到轨迹的异常影响因子和轨迹的历史异常分值计算获得轨迹的进化异常分值,进而检测一段时间内的进化轨迹异常,以有效降低噪声数据对轨迹异常检测结果的影响.

(3) 忽视异常轨迹之间的因果联系以及对异常检测结果的语义分析.

大部分异常轨迹检测方法只针对应用数据中的单一定义的异常,如异常的子轨迹、异常的轨迹、异常的移动对象、异常的路段等,从而忽视了轨迹异常之间可能存在的因果联系,以至于不能检测更大范围区域内的异常轨迹.同时,对于异常轨迹检测结果,不重视分析其蕴含的语义信息,不能发现隐含在检测结果背后的异常事件及异常的起因.因此,可结合关联规则方法挖掘异常子轨迹之间的因果联系,发现异常涉及的更大区域,并结合与轨迹数据相关的多源数据对异常检测结果进一步进行语义分析,获知异常轨迹预示的异常事件及异常发生的缘由.

**Table 1** Classification and comparison of trajectory outlier detection techniques

表 1 轨迹异常检测技术的分类及比较

轨迹异常检测技术	轨迹数据相关性		异常类型					异常检测输出		处理方式		
	空间	时序	子轨迹	轨迹	移动对象	路段	区域	事件	标量	异常分数	离线	在线
Knorr <sup>[33-35]</sup>	✓			✓					✓		✓	
Lj <sup>[54]</sup>	✓		✓						✓		✓	
Lee <sup>[65]</sup>	✓		✓						✓		✓	
Lj <sup>[60]</sup>		✓				✓			✓		✓	
Bu <sup>[21]</sup>	✓	✓	✓						✓			✓
Yu <sup>[22]</sup>	✓	✓			✓				✓			✓
Ge <sup>[72]</sup>	✓	✓	✓							✓		✓
Zhang <sup>[68]</sup>	✓	✓			✓				✓		✓	
Chen <sup>[69]</sup>	✓	✓	✓	✓						✓		✓
Lei <sup>[56]</sup>	✓	✓	✓							✓		✓
Chawla <sup>[62]</sup>	✓	✓				✓			✓		✓	
Liu <sup>[61]</sup>	✓	✓				✓			✓		✓	
Pan <sup>[63]</sup>	✓	✓				✓	✓	✓	✓			✓

## 7.2 轨迹大数据的异常检测框架

针对轨迹大数据的时空不确定性、无限性、稀疏性、偏态分布性以及时变进化性等特征,亟需设计鲁棒性强的异常轨迹检测方法,实时检测局部的异常子轨迹,评测一定时段内进化的异常轨迹,并根据异常间的因果联系发现更大范围的异常区域.

图 7 描述了一个轨迹大数据的异常检测系统的基础架构,包括数据预处理、数据存储管理、异常检测、检测结果的可视化语义分析这 4 个阶段.面对持续到达且快速更新的轨迹大数据,首先需要使用预处理技术去除噪声、纠正位置偏差、减少数据处理的规模;随后,为确保以流的形式出现的轨迹大数据能被实时处理,需要对轨迹流数据进行高效的存储管理;更为重要的是,为提高异常检测的准确性,在使用历史轨迹数据提取的异常检测模型辅助判定数据异常的同时,必须有效地通过采用在线异常轨迹检测方法以进一步发现历史特征模型未能揭示的新型异常;最后,对于异常检测结果,不仅需要以可视化的方式展现给终端用户,还应结合与应用相关的多源数据对检测结果进行语义分析,给出关于异常结果合理的语义解释,进而揭示出更有价值的异常事件.

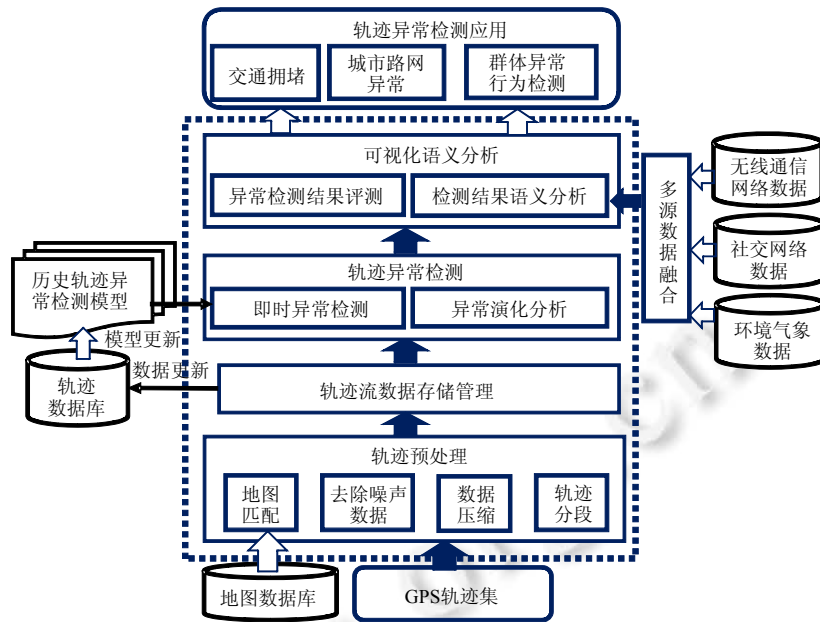


Fig.7 System architecture of outlier detection in trajectory big data

图7 轨迹大数据的异常检测系统架构

### (1) 轨迹预处理

GPS 定位设备容易受建筑物遮挡、外界电磁干扰以及设备故障等因素的影响,产生与真值具有较大偏差的位置数据.另外,一些人为因素(比如司机关闭车载 GPS 设备)也会使得采集设备在部分时段缺失位置信息.由此导致采集的轨迹数据存在经纬度出界、数据丢失、数值异常等问题,严重影响了轨迹异常检测结果的准确度.为了降低噪声数据对异常检测性能的影响,首先,使用数据异常过滤/阈值过滤等方法清洗数据;再对数据进行校验,判断数据是否丢失以及是否包含错误,并根据历史趋势数据等对轨迹数据进行修正;随后,结合均值/中值滤波器、卡尔曼滤波器、粒子滤波器等技术实现数据的平滑去噪.与此同时,为了缩小轨迹大数据实时处理的规模,可运用在线数据压缩技术、采样技术或基于时间间隔/轨迹形状的轨迹分段技术,在保证轨迹的时空特征的前提下,对去除噪声后的轨迹数据实现近似化表示.此外,在交通领域等具体应用中,可采用地图匹配技术修正轨迹数据的偏差,提高异常检测精度.

### (2) 轨迹流数据的存储管理

轨迹大数据的一个关键特性就是大规模的实时轨迹数据以高速数据流的形式到达,由于轨迹数据时变进化,对于新到达的轨迹数据,难以迅速建立数据的先验知识,也不可能对其永久化存储.受限于流处理在响应时间方面的要求,处理过程主要依赖于内存完成,可通过设计恰当的概要数据结构实现其处理.对轨迹数据流的管理,除了采用单机流数据管理系统以外,可以考虑分布式数据流管理系统.尽管单机流数据管理系统能够处理一定量的数据流,但在海量数据环境下,连续到达的轨迹大数据远远超出了单台物理机器的计算能力.为了实时地获得轨迹大数据的处理结果,可选择分布式的计算架构.代表性的分布式流数据处理系统如 Twitter 的 Storm 系统、Yahoo 的 S4 系统、LinkedIn 的 Samza 以及加州大学伯克利分校推出的 Spark 系统等.在对持续到达的流式轨迹数据进行存储管理的同时,将新到达的轨迹数据流同步写入磁盘的轨迹数据库,进而更新提取的轨迹异常检测模型.

### (3) 轨迹的异常检测

基于大规模历史轨迹数据挖掘建立的全局特征模型,在不考虑轨迹数据的时变进化特征时,具有较高的异常检测精度.但以历史轨迹数据提取的异常检测模型为参照标准,不能确保准确检测流中最新到达的轨迹数据

的新型异常模式,需要设计适合有限系统资源的轻量级在线异常检测方法.因此,在轨迹大数据的即时异常检测阶段,采用历史轨迹提取的异常检测模型与在线异常检测算法评测相结合的方式.对于最新到达的轨迹数据,先使用基于历史轨迹提取的特征模型检测异常行为,同时采用在线异常检测方法识别新型的轨迹异常(比如,将异于其局部时空近邻的行为表现作为评测依据),综合二者的检测结果作为最终的即时异常检测结果.鉴于轨迹异常固有的进化特性,在异常演化分析阶段,结合移动对象的即时异常检测结果与其历史异常检测结果,尽早发现随时间推移不断进化的异常轨迹/移动对象.此外,根据新到轨迹数据的特征更新已有的异常检测模型.对于地理上分布的定位设备产生的分布式轨迹数据,还要考虑资源受限环境下最小化计算开销及通信开销,设计高效的分布式轨迹流在线异常检测算法.

#### (4) 异常检测结果的可视化语义分析

为提升轨迹数据的解释能力、及时展示并反馈轨迹的异常检测结果,可根据具体的应用需要选择合适的可视化交互分析技术.一方面,可借助人机交互技术,提供参数/阈值的设置以及检测方法的选择等功能界面,引导用户逐步地进行轨迹数据分析,参与具体的异常检测交互式分析过程;另一方面,借助于图表、动画等形式清晰地展示异常检测的结果,帮助用户理解轨迹异常检测结果的由来.在此基础上,可结合其他相关数据源(如微博、博客等社交网络平台数据、电信数据等)对所识别的轨迹异常进行语义分析,发现异常之间的语义联系以及异常产生的缘由,从而识别其隐含的异常事件.对于语义分析结果,可以使用特定的颜色、不同层次的亮度凸显异常及异常之间的联系,有助于判定异常的发展趋势及影响范围,辅助决策者尽早制定异常控制策略.

### 7.3 未来展望

轨迹异常检测技术在近 10 年得到了广泛而深入的研究,取得了明显的进展,但对于轨迹大数据的异常检测,仍然需要多种技术的协同支持.在未来的工作中,对于轨迹流的在线异常检测、轨迹异常的演化分析、异常检测系统的基准评测、异常检测结果语义分析的数据融合、异常检测的可视化技术等方面,仍亟待进一步加以研究.

#### 7.3.1 轨迹流的在线异常检测

海量、快速、时变的位置序列以数据流的形式持续到达,没有固定的数据分布模式,存在概念漂移.传统的异常检测技术因数据分析涉及较高的计算和内存开销,不能直接用于轨迹流的异常检测.轨迹流的异常检测正面临数据的质量低劣性、流式数据的处理高效性、异常的语义多样性、轨迹异常的时变进化性等技术难题,亟需设计高效的检测方法在线识别异常轨迹.这要求轨迹流异常检测算法在有限的内存空间中建立和维护高效的概要数据结构,存放每个时间间隔内到达的轨迹数据集的时空特征(如中心点位置、角度等值的线性/平方和、平均速度、轨迹分段数、最晚到达的轨迹时间等).使用概要数据结构在保持轨迹分布特征的前提下,可以有效降低存储空间开销以及计算时间开销,有助于提高在线异常检测的效率.对于轨迹流中的概念漂移,可以结合滑动窗口、时间衰减窗口模型等技术,使用高效的分析方法对“无限”的位置数据流进行一遍扫描,及时识别异常子轨迹.例如,Bu<sup>[21]</sup>和 Yu<sup>[22]</sup>等人采用滑动窗口技术,基于距离的思想分别实现了在线检测异常子轨迹以及异常的移动对象.文献[49]基于隔离的思想提出了在线识别异常子轨迹的检测方法.考虑到轨迹流中的异常通常不是孤立存在的,可将同一时段内出现异常的子轨迹联系起来,推断它们之间的因果关系和相互作用,进而识别更大范围内的异常区域.文献[55]实现了在线检测路网中的异常路段及异常所影响的路网区域.此外,在设计分布式轨迹流在线异常检测算法时,可以改进现有单机算法,使之并行化;也可以设计全新的分布式异常检测算法,以满足分布式计算的需求.值得注意的是:在提取分布式环境中的全局异常轨迹时,还要考虑分布式环境下设备间传送信息时的隐私保护问题.

#### 7.3.2 轨迹异常的演化分析

轨迹大数据随时间的推移而不断进化,轨迹异常也非恒定不变,即使当前时刻轨迹数据与其空间近邻相似,但在一段时间以后仍有可能演变成异常.预先训练好的异常检测模型不能准确地评测轨迹流中不断进化的异常,而阶段性重建分类模型因其计算开销太大也不能有效解决这一问题.因此,如何评测轨迹异常随时间出现、成长、转换甚至消失的趋势,是轨迹大数据异常检测技术的难点.这要求轨迹异常检测方法不仅能够区分噪声、



实时检测异于相邻位置序列的异常,还要追踪一段时间内进化演变而来的异常轨迹.可以采用统计移动对象的进化轨迹异常分值的方法来检测进化的异常轨迹.比如,文献[50]提出的 TOP-EYE 方法结合指数衰减函数可以有效区分噪声和异常数据,识别基于方向与基于密度的两类进化的轨迹异常.在统计各移动对象实时的局部轨迹异常分值时,可结合前向/后向衰减方法(指数衰减函数、多项式衰减函数、界标窗口模型等)以降低历史轨迹异常分值的权重,并与最新时刻的轨迹异常分值累加得到进化的轨迹异常分值.通过观测连续时间间隔内各轨迹的进化异常分值,并与预设的轨迹异常阈值进行比较,将某时刻高于异常阈值的轨迹作为进化的异常轨迹.

### 7.3.3 轨迹异常检测系统的基准评测

轨迹异常检测技术已有 10 多年的研究发展历史,但是对于具体应用而言,仍然缺少能客观比较、评测不同异常检测系统性能的规范.轨迹大数据异常检测的评测基准依赖于实际应用,每类典型应用都需要对应的评测基准,同时,评测基准需要具备可迁移性、可扩展性以及可理解性.以设计面向路网轨迹大数据的异常检测基准为例,在数据生成方面,如何在分布式轨迹流环境下生成反映应用需求的实时性的数据流是一大挑战.应使用仿真数据生成器产生具有不同数据密度分布的车辆移动轨迹数据集,再以概要数据结构等形式提取轨迹的时空特征.随后,在此基础上定义路网中可能存在的多种异常,并为它们提供检测方法的缺省参数配置.在功能负载方面,包括发现异常子轨迹、异常轨迹、异常移动对象、异常路段、异常区域和异常事件等典型的处理任务.评测基准的性能度量则选择系统的异常检测执行时间、查全率、查准率、检测真正率、误检率、漏检率以及可伸缩性等,同时提供检测性能测试工具,用于全面评测不同轨迹异常检测方法的性能差距.总之,轨迹大数据异常检测系统的基准评测面临诸多挑战,需要在异常检测评测基准的仿真数据生成、功能负载、评测架构、度量选择等多个方面展开积极而深入的探索,符合轨迹大数据异常检测的实时处理需求,进而推动轨迹异常检测技术的健康发展.

### 7.3.4 异常检测结果语义分析的数据融合

随着位置信息数据源的日益增多,采集到的位置数据类型、数据结构呈现多样化的趋势.在轨迹异常检测过程中,不再只限于分析移动对象产生的位置信息,还可以结合非时空属性数据以及与位置服务应用紧密联系的其他领域数据(比如无线网络、社交媒体、交通管理、环境气象等领域数据)共同检测异常轨迹.采用数据融合技术将多源异质数据匹配融合,辅助异常检测,有助于获知异常轨迹中隐含的语义信息,进一步分析异常的成因以及预测未来的变化趋势.例如, Pan 等人<sup>[55]</sup>基于车辆轨迹数据检测异常轨迹,并根据所检测异常发生的时段及位置,利用在线社交媒体(如微博)数据挖掘与异常相关的术语以揭示其隐含的异常事件.需要注意的是:不同领域的数据具有不同的表示形式、数据分布、数值范围和密度,数据融合需要针对不同特点的数据形式研究有效处理差异的方法.类似于轨迹数据的预处理工作,数据分析之前需要数据清洗工作,降低噪声对检测结果分析的影响.需要开发数据解析工具和数据转换工具,实现将异构数据统一转变成一致的结构化形态.此外,随着各类数据源产生的大规模异构数据的快速增长,不仅需要基于知识间的内在机理,对跨领域的多类型数据增量融合统一存储,还应处理好结构化数据(表格等)、半结构化数据(微博等社交媒体数据)以及非结构化数据(图片等)之间的关联,消除模式、标示符以及数据冲突.这为数据存储、解析、转换及关联处理提出了新的要求.

### 7.3.5 轨迹异常检测的可视化交互分析

面对纷繁复杂的轨迹大数据,可视化技术通过图表、动画等形式直观地描述轨迹数据及异常检测结果.同时辅以各类交互机制,帮助人们发现轨迹蕴含的移动趋势,了解轨迹数据的不确定性及内在错误,实现轨迹数据的异常检测.例如,文献[55]使用堆积图<sup>[75]</sup>、流向图<sup>[76]</sup>以及路网可视化<sup>[77]</sup>等技术为用户提供了关于异常路段区域的两种视图(导航视图和分析视图).在展示海量轨迹数据的时空特征时,容易出现大量位置信息叠加在一起导致视觉混淆的情况,这是轨迹异常检测可视化需要解决的第 1 个问题.可以考虑按场景展示,使用图模型方式,并以数据聚合的形式来解决这一问题.除了时空特征以外,轨迹数据通常包含速度、方向、高度等非时空特征属性,如何借助可视化技术发现与轨迹异常息息相关的特征属性及属性间的联系,这是轨迹异常检测可视化面临的第 2 个问题.可使用不同的颜色及标签注释的方式对不同的非时空属性加以区分,并用特定颜色加上高亮

的方式凸显异常.此外,对于轨迹中的点、线、面等数据形式,可以借助点边图、时间变化折线图、3D柱状图、热力图、标签云、密度图等方式分别加以展示.然而,可视化技术不仅用于显示轨迹数据及异常检测结果,更重要的是,凭借交互机制帮助用户直接参与具体的异常检测交互式分析过程.通过提供缩放、平移、旋转、标记、聚焦、历史回溯、参数设置等功能,帮助用户交互式地选择感兴趣的区域及特定的对象,对比不同参数获得的不同检测结果,以全面可视分析的形式解决用户异常检测中的不同需求.

## 8 结 论

轨迹大数据的异常检测对交通管理、安全监控、城市规划等诸多领域有着广泛的应用前景与研究价值.轨迹数据因其固有的不确定性、稀疏性、偏态分布性、时变进化性等显著特征,使其与传统的关系数据、空间数据以及时空数据在异常的定义和检测方法方面均有明显的差异,也使得现有的异常轨迹检测技术无法更好地适应轨迹大数据在数据质量低劣性、数据处理高效性、概念漂移的鲁棒性、异常轨迹的时变进化性、异常轨迹间的因果关联性、异常检测结果的可视化语义分析等方面所面临的新技术挑战.本文系统地梳理和分析了现有轨迹异常检测技术的研究和发展现状,并以轨迹数据时空相关性、检测的异常类型、异常检测输出结果方式、异常检测处理方式为依据,对13种具有代表性的轨迹异常检测方法进行了详细的对比分析,指出了现有方法的局限性.在此基础上,提出了轨迹大数据异常检测的基本系统架构,指明了轨迹异常检测未来的研究方向,实现了对轨迹异常检测算法及关键技术问题的深入研究.可以看出:尽管轨迹的异常检测技术已有10多年的研究历史,但是轨迹大数据异常检测的研究与应用仍处于很不成熟的阶段,这与其广泛的市场需求和应用前景很不吻合.为了促进轨迹大数据异常检测技术的稳健发展,亟需设计鲁棒性强的在线轨迹异常检测方法,有待系统、深入地开展相关理论和实践的研究工作.

### References:

- [1] Liao L, Patterson DJ, Fox D, Kautz H. Learning and inferring transportation routines. *Artificial Intelligence*, 2007,171(5-6): 311–331. [doi: 10.1016/j.artint.2007.01.006]
- [2] Liu L, Andris C, Ratti C. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 2010,34(6):541–548. [doi: 10.1016/j.compenvurbysys.2010.07.004]
- [3] Phithakittinukoon S, Veloso M, Bento C, Biderman A, Ratti C. Taxi-Aware map: Identifying and predicting vacant taxis in the city. In: *Proc. of the Ambient Intelligence*. 2010. 86–95. [doi: 10.1007/978-3-642-16917-5\_9]
- [4] Lippi M, Bertini M, Frasconi P. Collective traffic forecasting. In: *Proc. of the ECML-PKDD*. 2010. 259–273. [doi: 10.1007/978-3-642-15883-4\_17]
- [5] Zheng Y, Liu YC, Yuan J, Xie X. Urban computing with taxicabs. In: *Proc. of the UbiComp*. 2011. 89–98. [doi: 10.1145/2030112.2030126]
- [6] Yuan J, Zheng Y, Zhang CY, Xie X, Sun GZ, Huang Y. T-Drive: Driving directions based on taxi trajectories. In: *Proc. of the GIS*. 2010. 99–108. [doi: 10.1145/1869790.1869807]
- [7] Liu HP, Jin CQ, Zhou AY. Popular route planning with travel cost estimation. In: *Proc. of the DASFAA*. 2016. 403–418. [doi: 10.1007/978-3-319-32049-6\_25]
- [8] Zheng VW, Cao B, Zheng Y, Xie X, Yang Q. Collaborative filtering meets mobile recommendation: A user-centered approach. In: *Proc. of the AAAI Conf. on Artificial Intelligence*. 2010. 236–241. <http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1615>
- [9] Zheng Y, Zhang L, Xie X, Ma W. Mining interesting locations and travel sequences from gps trajectories. In: *Proc. of the WWW*. 2009. 791–800. [doi: 10.1145/1526709.1526816]
- [10] Deng Z, Hu YY, Zhu M, Huang XH, Du B. A scalable and fast OPTICS for clustering trajectory big data. *Cluster Computing*, 2015, 18(2):549–562. [doi: 10.1007/s10586-014-0413-9]
- [11] Costa G, Manco G, Masciari E. Dealing with trajectory streams by clustering and mathematical transforms. *Journal of Intelligent Information Systems*, 2014,42(1):155–177. [doi: 10.1007/s10844-013-0267-2]
- [12] Yu YW, Wang Q, Wang XD, Wang H, He J. Online clustering for trajectory data stream of moving objects. *Computer Science and Information Systems*, 2013,10(3):1293–1317. [doi: 10.2298/CSIS120723049Y]

- [13] Mao JL, Song QG, Jin CQ, Zhang ZG, Zhou AY. TScluWin: Trajectory stream clustering over sliding window. In: Proc. of the DASFAA. 2016. 133–148. [doi: 10.1007/978-3-319-32049-6\_9]
- [14] Nehme RV, Rundensteiner EA. SCUBA: Scalable cluster-based algorithm for evaluating continuous spatio-temporal queries on moving objects. In: Proc. of the EDBT. 2006. 1001–1019. [doi: 10.1007/11687238\_58]
- [15] Sacharidis D, Patroumpas K, Terrovitis M, Kantere V, Potamias M, Mouratidis K, Sellis TK. On-Line discovery of hot motion paths. In: Proc. of the EDBT. 2008. 392–403. [doi: 10.1145/1353343.1353392]
- [16] Zheng K, Zheng Y, Yuan NJ, Shang S. On discovery of gathering patterns from trajectories. In: Proc. of the ICDE. 2013. 242–253. [doi: 10.1109/ICDE.2013.6544829]
- [17] Tang LA, Zheng Y, Yuan J, Han JW, Leung A, Hung CC, Peng WC. On discovery of traveling companions from streaming trajectories. In: Proc. of the ICDE. 2012. 186–197. [doi: 10.1109/ICDE.2012.33]
- [18] Li XH, Ceikute V, Jensen CS, Tan KL. Effective online group discovery in trajectory databases. IEEE Trans. on Knowledge and Data Engineering, 2013,25(12):2752–2766. [doi: 10.1109/TKDE.2012.193]
- [19] Duan XY, Jin CQ, Wang XL, Zhou AY, Yue K. Real-Time personalized taxi-sharing. In: Proc. of the DASFAA. 2016. 451–465. [doi: 10.1007/978-3-319-32049-6\_28]
- [20] Hawkins DM. Identification of Outliers. London: Chapman and Hall, 1980.
- [21] Bu YY, Chen L, Fu AWC, Liu DW. Efficient anomaly monitoring over moving object trajectory streams. In: Proc. of the SIGKDD. 2009. 159–168. [doi: 10.1145/1557019.1557043]
- [22] Yu YW, Cao L, Rundensteiner EA, Wang Q. Detecting moving object outliers in massive-scale trajectory streams. In: Proc. of the SIGKDD. 2014. 422–431. [doi: 10.1145/2623330.2623735]
- [23] Hodge VJ, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review, 2004,22(2):85–126. [doi: 10.1007/s10462-004-4304-y]
- [24] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys, 2009,41(3):75–79. [doi: 10.1145/1541880.1541882]
- [25] Aggarwal CC. Outlier Analysis. Springer-Verlag, 2013. [doi: 10.1007/978-1-4614-6396-2]
- [26] Zhang Y, Meratnia N, Havinga PJM. Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 2010,12(2):159–170. [doi: 10.1109/SURV.2010.021510.00088]
- [27] Gupta M, Gao J, Aggarwal CC, Han JW. Outlier detection for temporal data: A survey. IEEE Trans. on Knowledge and Data Engineering, 2014, 26(9):2250–2267. [doi: 10.1109/TKDE.2013.184]
- [28] Zheng Y. Trajectory data mining: An overview. ACM Trans. on Intelligent Systems and Technology, 2015,6(3):29. [doi: 10.1145/2743025]
- [29] Jeung H, Yiu ML, Jensen CS. Trajectory pattern mining. In: Computing with Spatial Trajectories. New York: Springer-Verlag, 2011. 143–177. [doi: 10.1007/978-1-4614-1629-6\_5]
- [30] Barnett V, Lewis T. Outliers in Statistical Data. Hoboken: John Wiley & Sons, 1994.
- [31] Aggarwal CC, Yu PS. Outlier detection for high dimensional data. In: Proc. of the SIGMOD. 2001. 37–46. [doi: 10.1145/375663.375668]
- [32] Johnson T, Kwok I, Ng RT. Fast computation of 2-dimensional depth contours. In: Proc. of the KDD. 1998. 224–228. <https://www.aaai.org/Papers/KDD/1998/KDD98-038.pdf>
- [33] Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large datasets. In: Proc. of the VLDB. 1998. 392–403. <http://www.vldb.org/conf/1998/p392.pdf>
- [34] Knorr EM, Ng RT. Finding intensional knowledge of distance-based outliers. In: Proc. of the VLDB. 1999. 211–222. <http://www.vldb.org/conf/1999/P21.pdf>
- [35] Knorr EM, Ng RT, Tucakov V. Distance-Based outliers: Algorithms and applications. VLDB Journal, 2000,8(3-4):237–253. [doi: 10.1007/s007780050006]
- [36] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proc. of the SIGMOD. 2000. 427–438. [doi: 10.1145/342009.335437]
- [37] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. In: Proc. of the SIGMOD. 2000. 93–104. [doi: 10.1145/342009.335388]
- [38] Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C. LOCI: Fast outlier detection using the local correlation integral. In: Proc. of the ICDE. 2003. 315–326. [doi: 10.1109/ICDE.2003.1260802]

- [39] Sun P, Chawla S. On local spatial outliers. In: Proc. of the ICDM. 2004. 209–216. [doi: 10.1109/ICDM.2004.10097]
- [40] Chen F, Lu CT, Boedihardjo AP. GLS-SOD: A generalized local statistical approach for spatial outlier detection. In: Proc. of the KDD. 2010. 1069–1078. <http://dl.acm.org/citation.cfm?doi=1835804.1835939>
- [41] Birant D, Kut A. Spatio-Temporal outlier detection in large databases. *Science*, 2006,14(4):291–297. [doi: 10.2498/cit.2006.04.04]
- [42] Cheng T, Li ZL. A multiscale approach for spatio-temporal outlier detection. *Trans. on GIS*, 2006,10(2):253–263. [doi: 10.1111/j.1467-9671.2006.00256.x]
- [43] Adam NR, Janeja VP, Atluri V. Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. In: Proc. of the SAC. 2004. 576–583. [doi: 10.1145/967900.968020]
- [44] Wu E, Liu W, Chawla S. Spatio-Temporal outlier detection in precipitation data. In: Proc. of the KDD Workshop on Knowledge Discovery from Sensor Data. 2008. 115–133. [doi: 10.1007/978-3-642-12519-5\_7]
- [45] Li XL, Han JW, Kim S, Gonzalez H. ROAM: Rule- and motif-based anomaly detection in massive moving object data sets. In: Proc. of the SDM. 2007. 273–284. [doi: 10.1137/1.9781611972771.25]
- [46] Yang WQ, Gao Y, Cao LB. TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning. *Computer Vision and Image Understanding*, 2013,117:1273–1286. [doi: 10.1016/j.cviu.2012.08.010]
- [47] Lei PR. A framework for anomaly detection in maritime trajectory behavior. *Knowledge and Information Systems*, 2016,47(1): 189–214. [doi: 10.1007/s10115-015-0845-4]
- [48] Lee JG, Han JW, Li XL. Trajectory outlier detection: A partition-and-detect framework. In: Proc. of the ICDE. 2008. 140–149. [doi: 10.1109/ICDE.2008.4497422]
- [49] Chen C, Zhang DQ, Castro PS, Li N, Sun L, Li SJ. Real-Time detection of anomalous taxi trajectories from GPS traces. In: Proc. of the MobiQuitous. 2011. 63–74. [doi: 10.1007/978-3-642-30973-1\_6]
- [50] Ge Y, Xiong H, Zhou ZH, Ozdemir HT, Yu J, Lee KC. Top-Eye: Top- $k$  evolving trajectory outlier detection. In: Proc. of the CIKM. 2010. 1733–1736. [doi: 10.1145/1871437.1871716]
- [51] Zhang DQ, Li N, Zhou ZH, Chen C, Sun L, Li SJ. iBAT: Detecting anomalous taxi trajectories from GPS traces. In: Proc. of the Ubicomp. 2011. 99–108. [doi: 10.1145/2030112.2030127]
- [52] Li XL, Li ZH, Han JW, Lee JG. Temporal outlier detection in vehicle traffic data. In: Proc. of the ICDE. 2009. 1319–1322. [doi: 10.1109/ICDE.2009.230]
- [53] Liu W, Zheng Y, Chawla S, Yuan J, Xie X. Discovering spatio-temporal causal interactions in traffic data streams. In: Proc. of the SIGKDD. 2011. 1010–1018. [doi: 10.1145/2020408.2020571]
- [54] Chawla S, Zheng Y, Hu JF. Inferring the root cause in road traffic anomalies. In: Proc. of the ICDM. 2012. 141–150. [doi: 10.1109/ICDM.2012.104]
- [55] Pan B, Zheng Y, Wilkie D, Shahabi C. Crowd sensing of traffic anomalies based on human mobility and social media. In: Proc. of the SIGSPATIAL/GIS. 2013. 334–343. [doi: 10.1145/2525314.2525343]
- [56] Jonkery R, De LG, Van DV. Rounding symmetric traveling salesman problems with an asymmetric assignment problem. In: Proc. of the Operations Research. 1980. 623–627. [doi: 10.1287/opre.28.3.623]
- [57] Yi BK, Jagadish H, Faloutsos C. Efficient retrieval of similar time sequences under time warping. In: Proc. of the ICDE. 1998. 201–208. [doi: 10.1109/ICDE.1998.655778]
- [58] Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories. In: Proc. of the ICDE. 2002. 673–684. [doi: 10.1109/ICDE.2002.994784]
- [59] Chen L, Ng R. On the marriage of LP-norms and edit distance. In: Proc. of the VLDB. 2004. 792–803. <http://www.vldb.org/conf/2004/RS21P2.PDF>
- [60] Chen L, Ozsu MT, Oria V. Robust and fast similarity search for moving object trajectories. In: Proc. of the SIGMOD. 2005. 491–502. [doi: 10.1145/1066157.1066213]
- [61] Daniel PH, Klara K, Jon MK. On dynamic voronoi diagrams and the minimum hausdorff distance for point sets under euclidean motion in the plane. In: Proc. of the Symp. on Computational Geometry. 1992. 110–119. [doi: 10.1145/142675.142700]
- [62] Lee JG, Han JW, Whang KY. Trajectory clustering: A partition-and-group framework. In: Proc. of the SIGMOD. 2007. 593–604. [doi: 10.1145/1247480.1247546]
- [63] Roh GP, Hwang SW. NNCluster: An efficient clustering algorithm for road network trajectories. In: Proc. of the DASFAA. 2010. 47–61. [doi: 10.1007/978-3-642-12098-5\_4]

- [64] Han B, Liu L, Omiecinski E. Road-Network aware trajectory clustering: Integrating locality, flow, and density. IEEE Trans. on Mobile Computing, 2015,14(2):416–429. [doi: 10.1109/TMC.2013.119]
- [65] Li XL, Han JW, Kim S. Motion-Alert: Automatic anomaly detection in massive moving objects. In: Proc. of the ISI. 2006. 166–177. [doi: 10.1007/11760146\_15]
- [66] Ge Y, Xiong H, Liu C, Zhou ZH. A taxi driving fraud detection system. In: Proc. of the ICDM. 2011. 181–190. [doi: 10.1109/ICDM.2011.18]
- [67] Liu SY, Ni LM, Krishnan R. Fraud detection from taxis' driving behaviors. IEEE Trans. on Vehicular Technology, 2014,63(1): 464–472. [doi: 10.1109/TVT.2013.2272792]
- [68] Pang LXL, Chawla S, Liu W, Zheng Y. On mining anomalous patterns in road traffic streams. In: Proc. of the ADMA. 2011. 237–251. [doi: 10.1007/978-3-642-25856-5\_18]
- [69] Pang LXL, Chawla S, Liu W, Zheng Y. On detection of emerging anomalous traffic patterns using GPS data. Data & Knowledge Engineering, 2013,87:357–373. [doi: 10.1016/j.datak.2013.05.002]
- [70] Chen C, Zhang DQ, Castro PS, Li N, Sun L, Li SJ, Wang ZH. iBOAT: Isolation-Based online anomalous trajectory detection. IEEE Trans. on Intelligent Transportation Systems, 2013,14(2):806–818. [doi: 10.1109/TITS.2013.2238531]
- [71] Sun L, Zhang DQ, Chen C, Castro PS, Li SJ, Wang ZH. Real time anomalous trajectory detection and analysis. Mobile Networks and Applications, 2013,18(3):341–356. [doi: 10.1007/s11036-012-0417-8]
- [72] Zhu J, Jiang W, Liu A, Liu GF, Zhao L. Time-Dependent popular routes based trajectory outlier detection. In: Proc. of the WISE. 2015. 16–30. [doi: 10.1007/978-3-319-26190-4\_2]
- [73] Liu LX, Qiao SJ, Liu B, Le JJ, Tang CJ. Efficient trajectory outlier detection algorithm based on R-tree. Ruan Jian Xue Bao/Journal of Software, 2009,20(9):2426–2435 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/3580.htm> [doi: 10.3724/SP.J.1001.2009.03580]
- [74] Liu LX, Le JJ, Qiao SJ, Song JT. Trajectory outliers detection based on local outlying degree. Chinese Journal of Computers, 2011,34(10):1966–1975 (in Chinese with English abstract).
- [75] Byron L, Wattenberg M. Stacked graphs—Geometry & aesthetics. IEEE Trans. on Visualization & Computer Graphics, 2008,14(6): 1245–1252. [doi: 10.1109/TVCG.2008.166]
- [76] Buchin K, Speckmann B, Verbeek K. Flow map layout via spiral trees. IEEE Trans. on Visualization & Computer Graphics, 2011,17(12):2536–2544. [doi: 10.1109/TVCG.2011.202]
- [77] Wei LY, Zheng Y, Peng WC. Constructing popular routes from uncertain trajectories. In: Proc. of the KDD. 2012. 195–203. [doi: 10.1145/2339530.2339562]

#### 附中文参考文献:

- [73] 刘良旭, 乔少杰, 刘宾, 乐嘉锦, 唐常杰. 基于 R-Tree 的高效异常轨迹检测算法. 软件学报, 2009, 20(9): 2426–2435. <http://www.jos.org.cn/1000-9825/3580.htm> [doi: 10.3724/SP.J.1001.2009.03580]
- [74] 刘良旭, 乐嘉锦, 乔少杰, 宋加涛. 基于轨迹点局部异常度的异常点检测算法. 计算机学报, 2011, 34(10): 1966–1975.



毛嘉莉(1979—),女,四川南充人,博士,副教授,主要研究领域为数据挖掘,基于位置的服务.



章志刚(1988—),男,博士,主要研究领域为分布计算,基于位置的服务.



金澈清(1977—),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为海量数据管理,包括基于位置的服务,数据质量,不确定数据管理.



周傲英(1965—),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为 Web 数据管理,数据密集型计算,内存集群计算,大数据基准测试和性能优化.