第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

一种利用不完整数据检测交通异常的方法

王玉玲 任永功

(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 城市化进程的加快带来了严重的交通问题,检测交通异常成为数据挖掘领域的热点之一。传统道路管理主要是应用视频监控,使得处理交通问题的效率受限。鉴于上述原因,提出了一种利用不完整数据检测交通异常的方法(Traffic Anomaly Detection,TAD)。首先,利用相关性聚类从手机数据中获取车辆密度信息,降低处理不完整数据的计算开销;然后,设计一个自适应无参数检测算法,根据手机呼叫量变化率捕捉车辆的分散式动态异常,以解决道路状况不确定性难题;最后,提出异常轨迹算法来追踪异常分布路线并预测影响范围,提高异常检测效率。实验结果表明,TAD方法在不同的实验环境下能够有效地检测交通异常,与现有算法相比,所提算法在有效性和伸缩性上效果更好。

关键词 异常检测,不完整数据,手机数据,异常轨迹

中图法分类号 TP277 文献标识码 A

Method of Traffic Anomaly Detection with Incomplete Data

WANG Yu-ling REN Yong-gong

(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract Development of urbanization process has brought serious traffic problems, and traffic anomaly detection becomes one of hot spots in the field of data mining. The traditional traffic management mainly uses video monitoring which has a limited efficiency for handling traffic problems. A method of traffic anomaly detection with incomplete data (Traffic Anomaly Detection, TAD) was proposed in this paper. Firstly, the correlation clustering obtains vehicle density information from mobile phone data and reduces the computation costs of processing incomplete data. Secondly, an adaptive parameter-free detection algorithm is designed to capture the distributed dynamic anomalies with phone call volume change rate on the roads, solving the uncertainty problem of road condition. Finally, anomaly trajectory algorithm is devised to retrieve anomaly distribution route and forecast influence scope, improving the efficiency of anomaly detection. Experimental results show that TAD methods can effectively detect abnormal traffic in different experimental conditions and our algorithm is better in efficiency and scalability compared with existing algorithms.

Keywords Anomaly detection, Incomplete data, Mobile phone data, Anomaly trajectory

1 引言

交通异常[1]是根据历史数据与正常稳定的车辆分布相比较而异常的车辆分布。城市中机动车的数量快速增加,道路拥堵现象日趋严重,交通问题严重威胁了中国城镇化的健康发展。应用聚类分析方法解决日益突出的城市交通问题,已经获得了广泛的关注和探讨。目前,视频监控技术[2]是应用最广泛的道路监测方法,具有视频捕获、存储和回放等功能。然而,要保证获取实时道路异常并及时采取有效措施,就需要工作人员不停地监看视频。这种情况下,监控人员极易疲劳,尤其面对多路监控视频时,很难及时对异常事件做出应急反应。因此,迫切需要一种新颖的交通异常检测技术来辅助监控人员工作。

基于聚类分析在数据挖掘中起到的重要作用,目前该领域已取得巨大的研究成果[5-7]:Chakrabarti 等人[5]提出了进化聚类,其在低动态环境中可以处理移动集群数据。Aggarwal 和 Yu[6]将大量努力投入到不确定数据流的聚类研究中,大量数据集可以联合聚类,处理复杂事件。Leung 等人[7]利用相关性聚类处理高维数据,识别存在于集群中任意导向的子空间。由于我们的方案中不断出现新的数据特征,上述方法不能保持良好的计算效率,本文提出的相关性聚类方法能够处理不完全数据问题:从不完整数据中获取隐藏的完整数据,基于相关性通过完整数据研究不完整数据。

近年来对异常检测的相关研究引起了学者们的广泛关注[8-16]。Singh 和 Sayal[8]提出在信息流数据中检测异常的方法。Bu 等人[9]提出通过网络异常来监测本地集群轨迹流。Sun 等人[10]提出一个利用网络分析动态概率并且能够识别相关的异常模式的方法。Aktolga 等人[11]提出了一种排列技术,在信息检索中使用该技术确定实验方案中的离群值。Pang 等人[12]利用出租车的 GPS 数据区监视不正常交通行为

本文受国家自然科学基金项目(F020806),辽宁省高等学校优秀人才支持计划项目(LR2015033),辽宁省科技计划项目(2013405003),大连市科技计划项目(2013A16GX116)资助。

王玉玲(1990—),女,硕士生,主要研究方向为数据挖掘,E-mail:1160173644@qq.com;任永功(1972—),男,博士,教授,主要研究方向为数据库技术、数据挖掘、智能信息计算等。

的出现。Chawla 等人[13]专注于推理道路交通异常的起因。Ge 等人[14]根据出租车司机的行驶轨迹，提出了一个出租车驾驶诈骗检测系统，其能追踪异常轨迹，进一步提高了检测效率。Chen 等人[15]对移动对象轨迹进行相似性查找。Monreale[16]利用轨迹模型来预测移动对象的移动。上述基于距离的方法不适合本文中的数据特征，本文的实验受到高度动态分散式环境的影响，面临非常有限的车辆行为数据，必须开发新方法处理不完全数据。

传统基于临界值[3]的异常检测方法需要花费大量时间训练阈值。为了解决以上问题和挑战，本文提出一种利用不完整数据检测交通异常的方法(TAD)，可以及时发现道路上的异常事件，追踪动态异常轨迹。城市中几乎每个人都有一部或多部手机，将这些手机作为人身上的"传感器"，通过基站能够获得手机使用者通讯时产生的位置信息。通过分析呼叫量及统计真实车辆数据发现，呼叫量和汽车数量呈近似单调关系，因此可以利用手机呼叫量检测道路异常。

不完整手机数据特性对传统方法提出新的挑战：1)车辆数据动态多变，时间和空间上呈分散式导致手机数据的不确定性；2)每个手机的观察数据极其有限，只有处于通话状态时我们才能记录其位置信息；3)手机的位置有失真性，基站获得的位置信息与实际位置具有一定的偏差[4]。这 3 个挑战降低了传统方法的检测精确度，目前还没有相关文献研究过解决办法。因此，我们给出的决策模式具有以下特点：1)基站的位置固定不变，将每个基站负责的区域作为观察单元；2)手机的通话数量是精确的，可以通过移动电话网络即时收集；3)每天的手机通话量模式图有相似性。

2 问题定义

基于文献[1-17]，给出以下基本概念和定义。

定义 1(呼叫量变化率) 在某段道路 i 上，从时间 t_j 开始的一分钟内所有手机通话量的总数为呼叫量 v_i^j，时间 t_k 对应呼叫量 v_i^k，呼叫量变化率 $\eta_i^{j,k}$ 定义为：

$$\eta_i^{j,k}=\frac{v_i^k-v_i^j}{v_i^j} \tag{1}$$

其中，$v_i^j, v_i^k, i, j, k\in N$。

人们每天的日常行为基本上一致，以天为观察周期的呼叫量模式图具有明显相似的形状，运用呼叫量变化率计算手机使用量更具准确性。正常情况下，在时间 t_j 和 t_k 对应的呼叫量变化率为 $\varphi^{j,k}$，即常规情况下的呼叫量统计结果。

定义 2(完整数据) 给出一个原始数据集 Z 和一个来自于原始数据集的观察数据集 H，对于任意两个元素 $(\delta_1,\delta_2)=Z$，并且假设他们相对应的观察数据是 $(\gamma_1,\gamma_2)\in H$，当且仅当观察数据与原始数据具有相同数值关系时，观察数据叫做完整数据。正式定义如下：

$$\forall[(\delta_1,\delta_2)\in\Delta,\delta_1\geqslant\delta_2],[(\gamma_1,\gamma_2)\in H,\gamma_1\geqslant\gamma_2]$$

定义 3(不完整数据) 给出一个原始数据集 Z 和一个来自于原始数据集的观察数据集 H，对于任意两个元素 $(\delta_1,\delta_2)=Z$，并且假设他们相对应的观察数据 $(\gamma_1,\gamma_2)\in H$，当观察的数据与原始数据不具有相同的数值关系时，观察数据叫做不完整数据。正式定义为：

$$\forall[(\delta_1,\delta_2)\in\Delta,\delta_1\geqslant\delta_2],[(\gamma_1,\gamma_2)\in H,\gamma_1<\gamma_2]$$

车辆密度数据是原始数据，手机数据是被观察的不完整数据，呼叫量是被观察的完整数据。不完整数据[17]需要花费大量的时间消除错误或恢复数据，而且数据在准确性和完整性上的安全系数不高。利用一种基于相关性的聚类方法，从原始数据中检索被观察的完整数据，用完整数据代替原始数据进行实验研究。

定义 4(交通异常) 如果呼叫量改变率大于或小于常规呼叫量改变率，则存在交通异常。至少满足下列一种情况，异常发生：

$$a.\ \eta_i^{j,k}>\varphi^{j,k};b.\ \eta_i^{j,k}<\varphi^{j,k} \tag{2}$$

其中，$\varphi^{j,k}$ 是常规模式下呼叫量的变化率。

不同的时间和地点内临界值各不相同，时间实体 j 也会影响检测结果，时间间隔长，检测结果稳定；时间间隔短，检测结果具有即时性，适合在线方式的异常检测。我们把异常分为两类：积极异常和消极异常。情况 a 代表积极异常，表示汽车数辆在增加；情况 b 代表消极异常，表示汽车数辆在减少。这里我们只考虑积极异常。

定义 5(递归 1-hop 邻域) 给定手机网络中的一个基站，与它相邻的一个基站被定义为递归 1-hop 邻域。例如，给定一个基站，它的 r-hop 邻域是它的 $(r-1)$-hop 邻域的 1-hop 邻域。

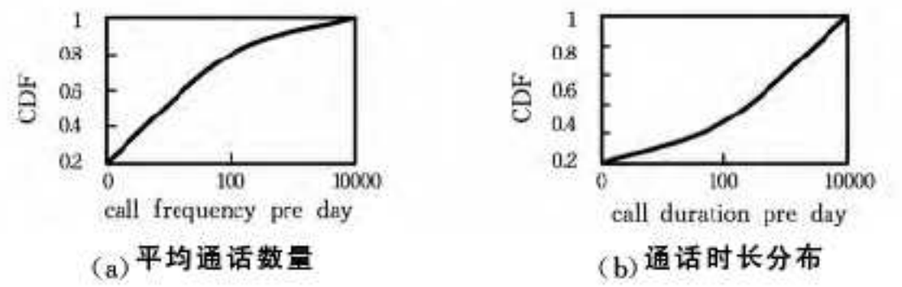
定义 6(异常轨迹) 随着事件的演变，异常位置将会形成一个时空轨迹。事件与交通异常具有空间一致性，也就是说，事件会从一个位置移动到另一个位置，我们把这个异常演变称为异常轨迹。正式定义如下：给定一个初始异常 b_0，发生时间是 t_0，发生位置是 l_0，异常轨迹就是发生在递归 1-hop 邻域内的一个异常序列 $[b_0,t_0,l_0],[b_1,t_1,l_1],\cdots,[b_n,t_n,l_n]$。

获得异常轨迹可以更好地进行异常检测研究，提高检测的准确性和效率。

3 交通异常检测的核心思想

3.1 手机数据与车辆密度相关性分析

正常交通状况下，人们在道路上几乎不使用手机，如果出现交通拥堵现象，使用手机消磨时间的可能性会增大，呼叫量也会随之增加。跟踪一百万个手机用户的使用情况，图 1(a)中 x 轴表示一天内手机的平均通话数量，大约 80% 的用户在一天之内的通话量没有超过 10 个，图 1(b)中 x 轴表示手机的通话持续时间，85% 的通话持续时间不超过 1 分钟，即 85% 的用户跟踪时间少于 1 分钟。用户的跟踪数据极其不完全且不均匀，我们不能简单地使用传统的基于距离的聚类方法收集数据。



(a)平均通话数量　(b)通话时长分布

图 1 手机数据的主观研究

为解决传统方法存在的收集数据的难题，提出了有效的解决思路。记录某段道路上 3 天全部的呼叫量，其中 x 轴是汽车数量，y 轴是相对应的平均呼叫量，统计结果如图 2 所示。3 条线分别表示 3 个在不同时间内进行的实验，在给定的时间和区域内，能够检索出呼叫量和车辆密度之间近似符合单调关系。可以利用手机呼叫量的异常分布来检测交通异常，现在用理论分析来求证这个假设。

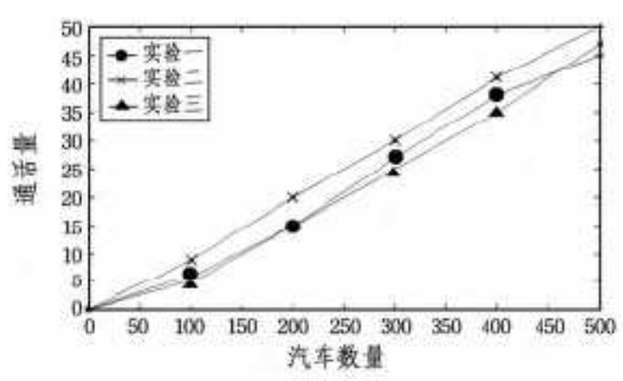


图 2　汽车数量与手机通话量之间的关系

设变量 S 表示在给定的时间和范围内总的通话数量，x_i 是随机变量，表示处于通话过程，x_i 的通话量是 s_i。x_i 存在概率为 p_i 的通话数量 y_i，当通话量 $s_i=0$ 时，概率是 $1-p_i$。我们引用一个指示器变量 v_i，如果 x_i 处于通话过程中并且 $s_i=y_i$ 时，则 $v_i=1$，得到 $S=\sum_{i=1}^{n} v_i x_i$。Y 是道路上呼叫量集合，类似地，可以定义 $Y=\sum_{i=1}^{n} s_i$。

定理 1　在指定道路区域内，给定车辆集合为 $X(x_i$ 作为一个人，假设一辆车里有一个人持有手机$)$，对应的呼叫量集合为 Y，每辆车内人们打电话的概率是 $P=\{p_1,p_2,\cdots,p_i,\cdots,p_n\}$，其中 $X=x_1+x_2+\cdots+x_n$，$x_1=x_2=\cdots=x_n$，$Y=y_1+y_2+\cdots+y_n$，并且期望值 $E(P)=\mu$。呼叫量与汽车数辆之间具有单调关系。

证明：根据辛钦大数法则，得到

$$\lim_{n\to\infty} P\{\left|\frac{1}{n}\sum_{i=1}^{n} x_i-\mu\right|\geqslant\varepsilon\}=0 \tag{3}$$

其中，$\varepsilon>0$。

方案中约有 100 万个手机，得到 $\frac{1}{n}\sum_{i=1}^{n} x_i=\mu$，注意 $x_1=x_2=\cdots=x_n$，而且 $y_i=p_i x_i$。因此，$Y=y_1+y_2+\cdots+y_n=\sum_{i=1}^{n} p_i x_i$，推出 $Y=(n\mu)\,p_i$。假设 $a=n\mu$，可得到 $Y=a p_i$（单调函数）。

以上证明了呼叫量与车辆密度之间的统计单调关系。事件造成异常使得人们打电话的概率远远高于正常情况，因此我们能够通过呼叫量进行交通异常检测。

3.2　手机数据检测异常

实验中有 100 万个手机产生高度分散式数据流，为了正确有效地生成呼叫量曲线图，我们选择香农定理抽样方法。通过实验可知，呼叫量的曲线图在不同的日子里基本相似，即呼叫量的模式图是稳定的，式(1)描述了在时间间隔内呼叫量的改变值及相应的变化率。例如，给定两个呼叫量序列 $v_i^{j,k}=\{100,190\}$ 和 $v_{i+1}^{j,k}=\{10,90\}$，虽然变化差值相同，但是第二个的改变率是第一个的 10 倍，因此呼叫量改变率可以准确反映出两个呼叫量数据之间的特征。

异常检测需要即时性和精确性，当道路有异常事件发生时应该立即对其检测并且给出警告，所以检测方法必须能够适用于分散式数据流。首先要确定在某段实验道路上的常规呼叫量改变率 $\eta_{normal}^{j,k}$，然后在相应时间指标内计算实时呼叫量改变率并与常规呼叫改变量改变率进行比较，若比较结果大于 0，则表明有异常发生。

$$(\eta_i^{j,k}-\eta_{normal}^{j,k})>0 \tag{4}$$

交通异常发生时，系统将会向数据中心报告，t_s 记录异常开始时间，直到获取的呼叫量变化率等于常规呼叫量变化率，异常结束。

还有一个问题需要解决：历史数据累积超时。系统会收集到不同的数据，因此需要定时更新数据来满足实验的准确性。实验中的更新方法适用于不同的数据：

$$\eta_{update}^{j,k}=\frac{(v_{history}^{k}+v_{new}^{k})-(v_{history}^{j}+v_{new}^{j})}{(v_{history}^{j}+v_{new}^{j})} \tag{5}$$

其中，$v_{history}^{j}$ 和 $v_{history}^{k}$ 是历史数据，不需要重复计算；v_{new}^{j} 和 v_{new}^{k} 是两个更新的呼叫量数据。

如果 $v_{new}\ll v_{history}$，则 $(v_{history}^{j}+v_{new}^{j})\approx v_{history}^{j}$，所以方程式可以简化为：

$$\eta_{update}^{j,k}=\frac{(v_{history}^{k}+v_{new}^{k})-(v_{history}^{j}+v_{new}^{j})}{v_{history}^{j}} \tag{6}$$

现在我们仅需要证明 $v_{new}^{k}-v_{new}^{j}$ 随着时间的变化而变化并且具有自适应性。

3.3　算法描述

本文提出的利用不完整数据检测交通异常主要由两个子算法组成。

3.3.1　异常检测算法描述

根据式(4)利用常规呼叫量改变率检测车辆异常，如果即时呼叫量改变率大于正常情况下的呼叫量改变率，则表示道路上有异常情况发生，输出异常时间和地点；反之，则表示交通情况正常，定时更新数据，具体细节见算法 1。

算法 1　交通异常检测(T-scan)算法

输入：呼叫量集合 v(n)；时间序列 T；即时呼叫量 vnew(i)；新事件处理时间 NT

输出：异常点和时间

step1　根据斯皮尔曼相关系数进行聚类，从基站中收集手机呼叫量数据，用呼叫量数据代替车辆密度数据。

step2　获取常规呼叫量改变率。

for(i=0;i++;i<n)

{ 输入常规状况下的呼叫量数据 v(n)；相邻呼叫量时间间隔为一分钟 T;P= polyfit(T, v, 9)，得到常规呼叫量拟合图；n(i)=(v(i+1)−v(i))/v(i)，计算每个时间指标内的呼叫量变化率；}

step3　计算呼叫量改变率。

nnew(i)=(vnew(i+1)−vnew(i))/vnew(i)；计算给定的呼叫量改变率；

step4　实时分布式异常检测。

if

{ nnew(i)−n(NT)>0&&vnew(i)>v(NT);Disp(NT)，输出每个异常时间点；NT=NT+1;}

step5　自适应数据更新。

vdata=((NT+1)+vnew(i+1)−(v(NT)−vnew(i)))/v(NT)+vnew(i)；更新变化率；V(NT)=vdapa

算法 1 的主要优点：它适用于分布式环境，不需要单独维持临界值检测异常；它的即时性收放良好，即使手机的数量增加，当有新的电话进入时，我们也能够计算呼叫量的变化率，并且能够计算出交通异常发生时的相应时间和地点；该算法不需要参数。算法 1 的时间复杂度是 $O(m)$，其中 m 是时间序列长度。

3.3.2　异常轨迹跟踪算法描述

交通异常往往伴随连锁反应，一条路段出现交通异常，很可能导致其他路段也出现异常情况。采用维诺图作为探测单元，一个基站代表一个单元。如图 3 所示，黑圈是没有异常的观察单元，黑点是有异常的观察单元，箭头表示异常轨迹。图 3(a) 显示了随着时间变化产生的异常轨迹，箭头线表示异常发生在不同时刻的实例；在图 3(b) 中显示了异常轨迹随时间变化在空间上形成的序列。

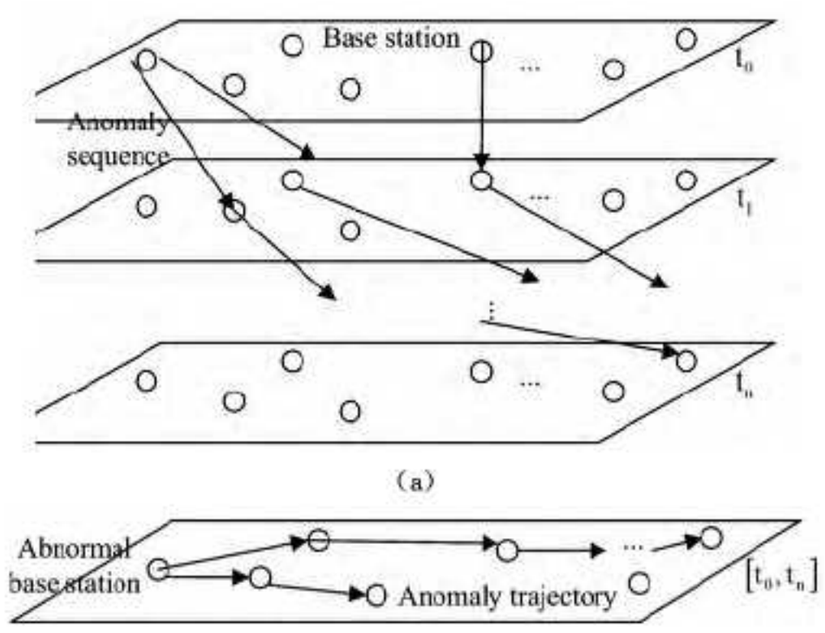


图 3　异常轨迹

交通事件引起的异常在时空上具有一致性，所以道路上出现交通事故时，极有可能引起其他路段的交通堵塞。当在观察单元内发现有一个异常存在时，我们会对它进行标记并且检查临近单元是否出现异常，详细说明见算法 2。

算法 2　捕捉异常轨迹算法(T-miner)

输入：呼叫量集合 v(n)；时间序列 time；邻接矩阵 M

输出：异常轨迹集合 AT

step1　初始化。对每个观察单元初始化相邻矩阵 M。

step2　递增异常轨迹处理。若发现一个异常出现，则检查是否已经对其进行标记和处理。若已经处理，则不进行计算；否则，在新的时间序列内进行计算。

step3　构建异常轨迹堆。如果 i 是一个异常位置，则检测它的相邻节点 i' 是否为异常。若 i' 为异常，则把 i' 放置到堆中；否则检测下一个邻域点。

跃点邻域矩阵方法可以得到具有较高有效性和准确性的实验结果。交通异常轨迹显示事件的影响范围，包括在时间和空间维度范围内异常事件的进展变化情况。算法 2 的时间复杂度是 $O(bn)$，其中 b 是所检测的异常事件的数量，n 是实验中观察单元的数量。

4　实验结果分析

4.1　实验设置

为了验证交通异常检测方法的性能，本文基于人造数据集和真实数据集对所提出的算法进行多角度分析和评价。人造数据集是根据现实生活中真实的数据进行合成而产生的数据集，模拟呼叫量数据，插入 100 个异常，对实验方法进行验证。真实数据集有 3 个：数据集 1 是从 4 月 1 日到 6 月 30 日 3 个月期间的详细手机通话记录，将其作为训练集，可以计算出正常交通状况下道路上的通话情况。数据集 2 是 5 月城市举办马拉松比赛当天的通话情况，这可以帮助我们进行异常研究。数据集 3 是对 500 个人进行主观实验，记录 5 月 1 日到 5 月 31 日期间电话的详细记录及发生的事件。在数据集 3 中要介绍两个主要事件(事件 1 和事件 2)，因为是主观实验，所以我们可以完整地获得事件的真实情况。对比所用指标包括平均的准确率、召回率和 F 值 3 个方面，具体如下所示：

准确率＝正确识别的呼叫量总数/识别出的呼叫量总数

召回率＝正确识别的呼叫量总数/测试集中存在的呼叫量总数

F 值＝正确率 * 召回率 * 2/(正确率＋召回率)

实验环境为 Matlab2012b，电脑配置为 Intel(R) Core(TM) 2 Duo CPU E7500 @ 2.93GHz 2.94GHz 的处理器，2GB 内存，Microsoft Windows7 操作系统。

4.2　检测异常评估

4.2.1　聚类方法对比分析

为了验证文中聚类方法的有效性，选择基于密度的聚类方法(Umicro)[18]作为实验中与基于相关性的聚类方法(CBC)的对比方法，分别在人造数据集和真实数据集上进行实验测试。Umicro 方法和 CBC 方法都是在不确定时间序列中进行，采用 Umicro 作为基线方法合理。

表 1 展示了利用人造数据集对 CBC 和 Umicro 进行比较的结果。可以发现 Umicro 的精确度为 44%，召回率为 36%，F 值为 40%，而 CBC 方法的精确度和召回率都超过了 80%，F 值也高达 79%。相关性聚类方法在人造数据集上具有较高的优越性。

表 1　人造数据集对比结果

数据集	方法	精确度	召回率	F 值
人造	CBC	0.802	0.828	0.780
	Umicro	0.447	0.364	0.401

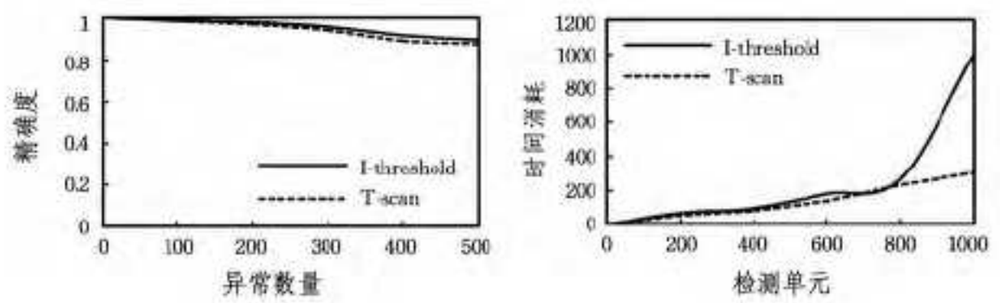
用真实数据集对 CBC 和 Umicro 的性能做进一步验证，表 2 描述了相关的实验结果。数据集 3 中有准确的呼叫量数据和相对应的车辆分布，与真实的车辆分布相比较，Umicro 有明显的误差。Umicro 的精确度小于 44%，我们的方法精确度大余 79%。不完全的数据集和不准确的位置信息使得 Umicro 的实现能力变得更加糟糕，而我们的方法可以从不完全数据中选择完全数据进行异常检测。

表 2　真实数据集对比结果

数据集	方法	精确度	召回率	F 值
数据集 1	CBC	0.792	0.783	0.787
	Umicro	0.425	0.351	0.384
数据集 2	CBC	0.788	0.779	0.783
	Umicro	0.432	0.361	0.393
数据集 3	CBC	0.804	0.783	0.793
	Umicro	0.438	0.357	0.393

4.2.2　异常检测算法评估

支持度相同的情况下，考察本文提出的交通异常检测算法(T-scan)的精确性和有效性，选择基于特定阈值的 I-threshold 算法[19]作为基线算法。这两种算法都属于在时间序列内自适应的异常检测算法，具有可比性。实验数据采用数据集 1、数据集 2 和数据集 3。图 4 所示为 T-scan 和 I-threshold 的精确性和时间消耗实验结果。图 4(a) 中 x 轴表示检测异常车辆的数量，y 轴表示精密度。图 4(b) 中 x 轴表示检测单元的数量，y 轴表示时间消耗。评估显示两种方法的精确度是相似的，但我们的方法在时间消耗上更具有优越性，随着检测单元的增大，T-scan 算法的优势逐渐明显。I-threshold 算法的主要消耗体现在需要不断计算阈值，而我们的方法不需要单独训练阈值，执行效果较好，与 I-threshold 算法相比具有更好的有效性和伸缩性。

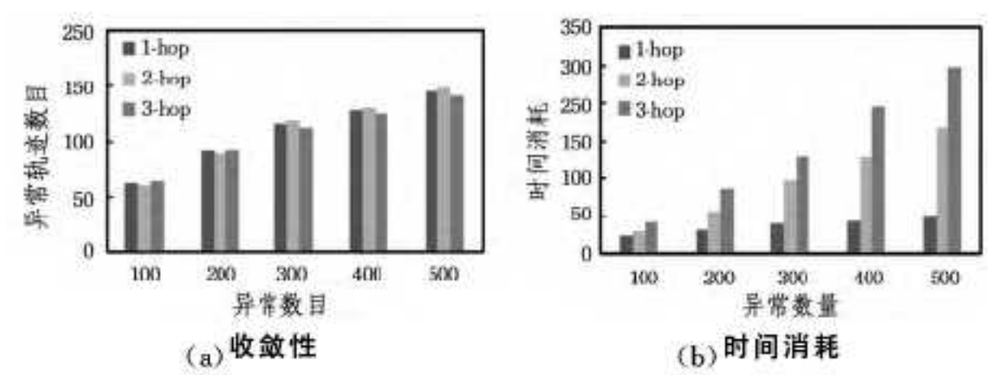


(a) 精确度　(b) 时间消耗

图 4　精确度和时间消耗分析

4.2.3 异常轨迹评估

检索异常轨迹是对异常检测的进一步完善。评估异常轨迹算法(T-miner)采用递归 1-hop 邻域检索方法。图 5(a)中,x 轴表示异常数量,y 轴表示检索异常轨迹的数量,可以发现 1-hop 邻域、2-hop 邻域(包括 1-hop 邻域)、3-hop 邻域(包括 1-hop 邻域,2-hop 邻域)3 个方案都收敛。造成这种结果是因为事件在时间和空间上具有一致性。换句话说,人的移动在时间和空间上是连续的。3 个跃点邻域异常轨迹检索结果相似,但是 1-hop 邻域方案需要更少的存储量,极大降低了时间消耗。图 5(b)中,x 轴是处理的异常数目,y 轴是时间消耗,1-hop 邻域方案的时间消耗比 3-hop 邻域方案少 5 倍。因此,我们的方案可以减少时间和存储成本。



(a)收敛性 (b)时间消耗

图 5 异常轨迹评估

4.2.4 实验方法总结

我们选取城市马拉松比赛事件进行研究总结,图 6 形象地展示了异常检测结果和异常轨迹。白色线代表真实的马拉松比赛路线,黑色的线表示检测到的异常及异常轨迹,箭头表示序列。我们发现异常轨迹不仅显示了事件的影响范围,而且显示了在时间和空间的维度范围内事件的进展变化。

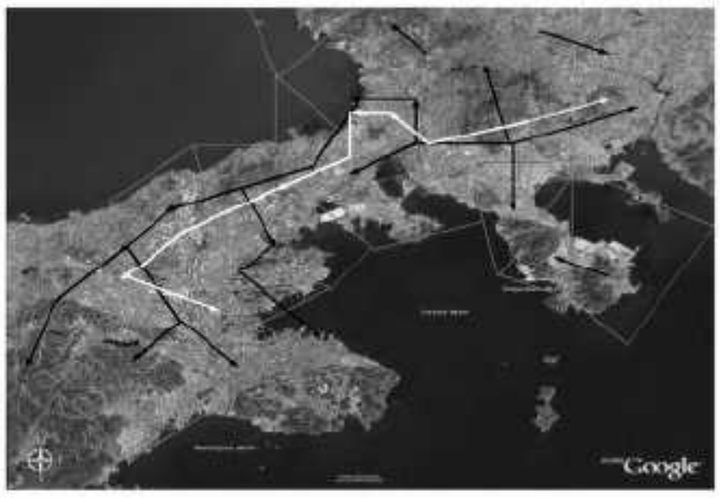


图 6 异常轨迹可视化

结束语 交通问题对城市规划提出了极大的挑战,如何有效解决这一难题已成为数据挖掘领域的研究热点。汽车位置不稳定、收集相关数据的难度大是我们面临的主要问题。本文通过分析大量数据,提出手机呼叫量与汽车密度具有单调性,可以利用不完整手机呼叫量数据检测交通异常。实验中提出的 T-scan 异常检测算法可以有效且实时地检测出道路异常状况;T-miner 异常轨迹追踪算法是对交通异常检测算法的进一步完善。通过实验验证,我们提出的 TAD 方法具有较高的效率,适用于分散式数据集,且优于其他同类方法。

参 考 文 献

[1] Liu Si-yuan, Chen Lei, Ni L M. Anomaly detection from incomeplete data[J]. ACM Transactions on Knowledge Discovery from Data, 2014, 9(2): 1-22

[2] 黄凯奇,陈晓棠. 智能视频监控技术[J]. 计算机学报, 2014, 37(49): 1093-1118

[3] Faloutsos B C, Plant C. Outlier-robust clustering using independent components[C]// Proceedings of the 34th ACM SIGMOD International Conference on Management of Data. Vancouver B C. Canada, 2008: 185-198

[4] DiLorenzo C G, Liu L. Estimating origin destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area[J]. IEEE Pervasive Computing, 2011, 1(10): 109-133

[5] Kumar C R, Tomkins A. Evolutionary clustering[C]// Proceedings of the 12th ACM SIGKDD International Conference on knowledge Discovery and Data. Philadelphia, USA, 2006: 554-560

[6] Aggarwal, Yu P S. A framework for clustering uncertain data streams[C]// Proceedings of the IEEE 24th Internatiional Conference on Data Engineering. New York, USA, 2008: 150-159

[7] Leung W T, Lee D L, Lee W C. Acollaborative location recommendation framework based on co-clustering[C]// Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval. 2011: 305-314

[8] Singh, Sayal M. Privately detecting bursts in streaming, distributed time series data[J]. Data and Knowledge Engineering, 2009, 68(6): 509-530

[9] Chen B L, Fu A W C. Efficient anomaly monitoring over moving object trajectory streams[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 159-168

[10] Zeng S D, Chen H. Burst detection from multiple data streams: A network-based approach[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2010, 40(3): 258-267

[11] Ros A I, Assogba Y. Detecting outlier sections in us congressional legislation[C]// Proceedings of the 34th ACM SIGIR Conference. 2011: 235-244

[12] Pang X, Chawla S, Liu W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data and Knowledge Engineering, 2013, 27: 509-530

[13] Chawla S, Zheng Yu, Hu Jia-feng. Inferring the root cause in road traffic anomalies[C]// Proceedings of the IEEE 12th International Conference on Data Mining. Beijing, China, 2012

[14] Xiong G H, Liu C, Zhou Z H. A taxi driving fraud detection system[C]// Proceedings of the IEEE 11th International Conference on Data Mining. 2011: 181-190

[15] Chen M T, Oria O V. Robust and fast similarity search for moving object trajectories[C]// Proceedings of the 31th ACM SIGMOD International Conference on Management of Data, Baltimore, USA, 2005: 491-502

[16] Monreale F, Tresarti P R, Giannotti F. WhereNext: A location predictor on trajectory pattern mining[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 637-646

[17] Liu Y, Ni L, Fan J. Towards mobility-based clustering[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA, 2010

[18] Aggarwal C, Yu P S. A framework for clustering uncertain data streams[C]// Proceedings of the IEEE 24th International Conference on Data Engineering. New York, USA, 2008: 150-159

[19] Klan K D. Adaptive burst detection in a stream engine[J]. Proceedings of the 24th ACM Symposium on Applied Computing, 2009, 8(12): 1511-1515