文章编号:1674-2474(2017)08-0145-07 DOI:10.16339/j.cnki.hdxbzkb.2017.08.022

# 面向群智感知车联网的异常数据检测算法*

徐艺文,徐宁彬,庄重文,陈忠辉†
(福州大学 物理与信息工程学院,福建 福州 350116)

**摘 要**:群智感知车联网利用普通用户的手机或平板电脑等智能终端获得交通数据,解决了车联网以低成本获取足够数据的问题,但却凸显了数据"质"的问题.为此,在分析群智感知车联网的数据结构及数据异常特点的基础上,提出一种适用于群智感知车联网的异常数据检测算法,并依此剔除异常数据,提高数据质量.算法利用核密度估计理论对车联网数据的概率密度进行估计,进而构建信任函数计算被检数据的信任度,后根据统计学理论将信任度小于0的数据判定为异常数据.最后对该算法的可行性及性能进行了仿真,结果表明该算法的性能可满足实用需求,且对比传统的统计检测法在检测率和误检率上具有更好的性能.

**关键词**:车联网;群智感知;异常数据检测;核密度估计
**中图分类号**:TP391 **文献标志码**:A

# An Algorithm of Abnormal Data Detection for Internet of Vehicles Based on Crowdsensing

XU Yiwen,XU Ningbin,ZHUANG Zhongwen,CHEN Zhonghui†
(School of Physics and Information Engineering,Fuzhou University,Fuzhou 350116,China)

**Abstract**:Internet of Vehicles (IoV) based on crowdsensing technology,which gets traffic data by smartphone or panel PC from ordinary person,has solved the problem that getting sufficient data at low cost. However,it also makes a new problem that the data quality of the system is deteriorated. To solve this problem,by analyzing the structure of crowdsensing data and the characteristics of abnormal data in crowdsensing IoV,a data detection algorithm is put forward to eliminate the abnormal data in IoV system and consequently improve data quality. In the algorithm,kernel density estimation theory is used to estimate the probability density of traffic data,and a belief function is then constructed to derive the confidence value of every detected data. According to the statistical theory,the data whose confidence value is less than 0 is regarded as abnormal data. Finally,the feasibility and performance of the presented algorithm are simulated. The results show that the proposed algorithm can meet practical demands and achieve better performance than that of traditional statistical detection methods.

**Key words**:internet of vehicles;crowdsensing;abnormal data detection;kernel density estimation

* **收稿日期**:2017-02-21
**基金项目**:国家自然科学基金海峡联合基金重点支持项目(U1405251),Key Project of National Natural Science Foundation of China (U1405251);国家自然科学基金资助项目(61571129,61601126),National Natural Science Foundation of China(61571129,61601126);福建省自然科学基金资助项目(2015J01250,2016J01299),Natural Science Foundation of Fujian Province(2015J01250,2016J01299)
**作者简介**:徐艺文(1976—),男,福建漳州人,福州大学副教授,博士
† 通讯联系人,E-mail:czh@fzu.edu.cn

近年来,随着汽车数量的持续增长,许多城市的道路承载容量已近饱和,道路拥堵问题日益严重.车联网技术在智能交通系统中的成功应用,使其被认为是解决道路拥堵问题的最佳方法,吸引了大批研究人员的兴趣[1-4],本文的研究也是基于该应用场景.传统车联网的数据采集方式分为固定式采集和浮动式采集两种,但它们都存在明显缺陷,前者安装和维护成本偏高,后者存在浮动车数量少而可能出现数据量不足的问题.为解决以上问题,有些学者提出了群智感知车联网的思路,即利用普通用户的手机或平板电脑等智能终端作为基本感知单元获取所需的交通数据,例如,文献[3]通过读取智能手机的加速度传感器信息,并经过坐标旋转获得车辆的三轴加速度,结合从手机内置 GPS 获得的车速信息,判断车辆运行的颠簸状况及刹车状况,进而判断交通状况是否良好;文献[4]通过实测评估了智能手机用于实时交通流预测的性能.以上研究均验证了群智感知技术应用于车联网数据采集的可行性.

群智感知以极低的成本获得海量交通数据,很好地解决了车联网数据"量"的问题,但是另一方面,这些数据来源于大量未经训练或认证的普通用户,恶化了数据"质"的问题,因此如何实现高效、高可靠的异常数据检测在群智感知车联网中显得尤为重要.车联网中,传统的异常数据检测算法主要有物理检测法和统计检测法[5-8],前者主要应用交通流理论针对不同类型的交通数据设置相应阈值,若数据超过阈值则判定为异常,该方法实现简单、检测速度快,但它只适用于交通流理论涉及的数据类型(如速度、流量、时间占有率),不适用群智感知车联网中新的数据类型(如三轴加速度),而且该方法在判定过程中使用大量的经验值作为阈值,导致算法适用范围窄且性能偏低;后者应用统计学理论确定一个置信上限,若数据误差超过该上限则判定为异常,该方法在理论上具有较高的检测性能,但它是在假设数据服从正态分布的前提下进行异常检测的,而实际生活中由于驾驶行为和车辆性能的随机性,车联网数据并不一定服从正态分布,导致该方法在车联网应用中的实际检测性能远不如理论性能.近几年也有学者提出其他一些车联网异常数据检测方法,文献[9]利用支持向量机的回归估计模型,通过计算实际值与预测值之间的残差来判别异常数据;文献[10]以 $k$ 近邻算法为基础,根据数据点与其相邻数据节点之间的距离判断异常数据;文献[11]采用小波分析方法分离交通流数据中的高频与低频分量,进而求得原始信号与重构信号的差值并结合最小二乘法找出异常数据.此外,文献[12]提出一种加入时间关联因子曲线拟合的交通流异常挖掘方法,并运用分箱思想设定正常值动态范围,从而剔除异常数据.这些方法都有各自的优势,但它们均主要针对传统车联网数据进行检测,不适用于群智感知这种具有特殊数据结构及异常特点的应用场景.为解决以上问题,本文分析了群智感知车联网的数据结构及数据异常特点,在此基础上提出一种适用于该场景的基于核密度估计的异常数据检测算法,并通过 Matlab 仿真验证其性能.

## 1 群智感知车联网的数据结构及数据异常的特点

群智感知车联网中,由于群智用户的手机或平板电脑等智能终端内的传感器有限,实际应用中往往不能直接从其内置传感器获得我们所需的数据,而是必须先对读取的传感器数据进行某种变换后才能获得,即群智用户上传的数据(以下简称群智数据)往往体现出如式(1)所示的结构:

$$\boldsymbol{X}_{i_c}=\{x_i^1,x_i^2,\cdots,x_i^M,c\} \tag{1}$$

其中$\{x_i^1,x_i^2,\cdots,x_i^M,c\}$代表第 $i$ 次测量获得的 $M$ 个传感器数据(以下简称测量值),$c$ 代表通过对第 $i$ 次测量值进行变换后获得的系统需要的数据(以下简称变换值),且 $c=1,2,\cdots,N$.例如,在文献[3]的应用场景下,$\{x_i^1,x_i^2,\cdots,x_i^M,c\}$代表直接从手机传感器读取的三轴加速度和速度,$c$ 代表通过三轴加速度和速度判断所得的路面颠簸状况.实际应用中,式(1)所示数据可能出现以下几种异常情况:① 因智能终端传感器自身原因造成测量值出现误差;② 群智数据的提供者未经专业训练,在某些情况下可能使变换值出现偏差,极端情况下可能因误操作而导致变换值严重偏离正常水平;③ 群智数据的提供者未经事先认证,可能出现用户发送假数据甚至恶意数据的情况.以上 3 种情况中,第①种情况是因设备原因产生的固有误差,考虑到大样本情况下该误差可得到有效弥补,而且测量值由智能终端直接测量所得,数据质量相对可靠,所以本文忽略该误差,即假设式(1)所示的群智数据中的每一项测量值都是准确的;第②,③种情况尽管缘由不同,但它们造成的后果最终都体现为群智用户将测量值$\{x_i^1,x_i^2,\cdots,x_i^M,c\}$错变换为 $c'$上报给服务器,这

是群智场景下特有的数据异常,也是本文的研究对象.

根据以上分析,本文将通过核密度估计获得群智数据 $\boldsymbol{X}_{ic}$ 中各元素的概率密度的估计值,进而计算测量值$\{x_i^1,x_i^2,\cdots,x_i^M\}$变换为 $c$ 的概率,并依此构建信任函数,根据该函数值检测异常数据.

## 2　基于核密度估计的异常数据检测算法

### 2.1　算法原理

为了便于理解本文所提算法,现将算法中涉及的各个符号及其定义统一说明,如表 1 所示.

**表 1　符号定义说明**

**Tab.1　Notations′ definition**

| 符号 | 定义 |
|---|---|
| $\boldsymbol{X}_{ic}$ | 群智数据,包含测量值与转换值,即 $\boldsymbol{X}_i c=\{x_i^1,x_i^2,\cdots,x_i^M;c\}$ |
| $x_i^s$ | 第 $i$ 次测量值的第 $s$ 个分量 |
| $s$ | 测量值分量的序号,$s=1,2,\cdots,M$ |
| $c$ | 变换值,$c=1,2,\cdots,N$ |
| $D_c$ | 变换值 $c$ 对应的积累数据集 |
| $f_c(x^s)$ | 群智数据中第 $s$ 个测量值分量的概率密度 |
| $r_{il}$ | 第 $i$ 次测量值转变为变换值 $l$ 的概率密度 |
| $P_{lc}$ | 群智用户将变换值 $l$ 认定为变换值 $c$ 的概率 |
| $r(\boldsymbol{X}_{ic})$ | 被检群智数据 $\boldsymbol{X}_{ic}$ 的信任度 |
| $\alpha$ | 置信概率 |

车联网数据库中的每个变换值分量 $c$ 对应积累数据集 $\boldsymbol{D}_c$,每个 $\boldsymbol{D}_c$ 中各个测量值分量下的所有元素都服从相同的概率分布 $f_c(x^s)$.假设通过核密度估计可足够精确地获得 $f_c(x^s)$,则在此基础上,可将每个群智数据 $\boldsymbol{X}_{ic}$ 中的测量值 $x_i^s$ 代入计算:

$$r_{il}=\prod_{s=1}^{M} f_l(x_i^s);l=1,2,\cdots,N \tag{2}$$

式中 $r_{il}$ 的物理意义是测量值$\{x_i^1,x_i^2,\cdots,x_i^M\}$变换为 $l$ 的概率密度.定义一个先验概率 $P_{ic}$,表示对于某个变换值 $l$,一个未经训练的群智用户将之变换为 $c$ 的概率,则测量值$\{x_i^1,x_i^2,\cdots,x_i^M\}$可变换为 $c$ 上报的最大可能概率密度为:

$$P(\boldsymbol{X}_{ic})=\max_l(r_{il}P_{lc});l=1,2,\cdots,N \tag{3}$$

实际应用中,$P_{lc}$ 一般可通过多次实验获得,对于无法确定该值的应用场景,可以用 Kronecker delta 函数近似,即:

$$P_{lc}=\begin{cases}1 & l=c\\ 0 & l\neq c\end{cases} \tag{4}$$

统计学理论告诉我们:

①概率密度函数 $f(x)$表征的是数据在 $x$ 附近的概率[13];

②具有不寻常低概率的数据对象可定义为异常数据[14].

因此,$P(\boldsymbol{X}_{ic})$相当于$\{x_i^1,x_i^2,\cdots,x_i^M\}$变换为 $c$ 的最大概率,则通过 $P(\boldsymbol{X}_{ic})$可以进行异常数据检测.基于以上分析可构建信任函数:

$$r(\boldsymbol{X}_{ic})=\log\frac{P(\boldsymbol{X}_{ic})}{\alpha} \tag{5}$$

式中 $\alpha$ 为置信概率,在实际应用中一般采用经验值或通过样本训练获得.式(5)采用对数函数是为了使数据具有线性可加性,且具有更紧凑的数量级分布.最后,根据式(5)可计算出所有群智数据的信任度,若满足:

$$r(\boldsymbol{X}_{ic})>0 \tag{6}$$

则认为该数据正常,反之则认为其异常.

以上算法的关键是 $f_c(x^s)$的获得,本文通过核密度估计实现.

### 2.2　核密度估计的实现及其在实用中的修正

从数据集中提取独立样本 $X_1,X_2,\cdots,X_n$,假设该样本与数据集具有相同的概率密度函数$f(x)$,则根据核密度估计理论[15],该数据集的核密度估计为:

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \tag{7}$$

式中 $n$ 为样本容量,$h$ 为窗宽,$K(\cdot)$为核函数.理论和大量的实验已证明,在样本容量足够大的情况下(群智感知数据显然符合这一前提),核函数的选取对核密度估计效果的影响不大,只需满足对称和单峰特性即可[15],所以本文选择 Epanechnikov 核(简称依潘核),即:

$$K(u)=\frac{3(1-u^2)}{4},\ |u|\leqslant 1 \tag{8}$$

在核密度估计的实现中,核函数 $K(\cdot)$的支撑集是全体实数.因此,当数据样本的真实密度函数的支撑集(函数 $f(x)$的支撑集指$\{x\mid f(x)\neq 0\}$)有边界时,核密度估计会在其边界处出现偏差,具体体现为边界外的概率密度值仍大于 0,这种现象称为核密度估计的边界效应.例如,车联网数据中的车速是一个非负值,因此在车速为负值时其实际概率密度值应为 0,但是由于边界效应的存在,根据式(7)算得的核密度估计结果在负半轴上却非零,这种情况会影响式(2)的计算结果,造成车速负值情况下 $r_{il}$ 却非零,可能引起后续异常数据检测的误判.为解决该问题,本文采用“边界核”的方法[16]对式(7)进行修正,边界核的表达式为:

$$B(u)=\frac{[a_2(x)-a_1(x)u]K(u)}{a_0(x)a_2(x)-a_1^2(x)} \tag{9}$$

式中：

$$a_i(x)=\int_{\frac{x-b}{h}}^{\frac{x-a}{h}} z^i K(z)\mathrm{d}z$$

其中 $i$ 为 0,1,2；$a$ 为上边界；$b$ 为下边界；$K(z)$为核函数；$h$ 为窗宽.将 $B(u)$替换式(7)中的核函数，可得边界核的核密度估计为：

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}B\left(\frac{x-X_i}{h}\right) \tag{10}$$

核密度估计的边界效应及修正结果将在本文第 3 部分的仿真中给出.

### 2.3 算法流程

综合以上分析，本文所提出的基于核密度估计的异常数据检测算法流程如下.

①划分子数据集：将数据集 $\boldsymbol{D}$ 中的所有数据记录按照变换值 $c(c=1,2,\cdots,N)$划分成 $N$ 个子数据集 $\boldsymbol{D}_c$，并从中分别提取独立样本用于后续核密度估计；

② 核密度估计：利用步骤①提取的数据样本，根据式(10)计算出各子数据集 $\boldsymbol{D}_c$ 中各测量值分量的概率密度函数 $f_c(x^1),f_c(x^2),\cdots,f_c(x^M)$；

③ 计算被检数据的最大可能概率密度：针对每一个被检数据 $\boldsymbol{X}_{ic}=\{x_i^1,x_i^2,\cdots,x_i^M;c\}$，根据式(3)计算 $\boldsymbol{X}_{ic}$中的测量值$\{x_i^1,x_i^2,\cdots,x_i^M\}$变换为 $c$ 上报的最大可能概率密度 $P(\boldsymbol{X}_{ic})$；

④ 计算信任度：根据式(5)计算每一个被检数据的信任度 $r(\boldsymbol{X}_{ic})$；

⑤ 异常数据检测：若被检数据 $\boldsymbol{X}_{ic}$ 的信任度大于 0，则判定其为正常数据，否则判定为异常数据.

## 3 仿真与分析

### 3.1 数据来源

文献[3]给出了通过手机获取汽车三轴加速度和速度并进而判断路面颠簸程度的算法，这是群智感知车联网的一个典型应用，本文将其作为仿真的应用背景.为此，我们设计了一个基于 Android 平台的应用软件，该软件可实时读取手机内置的三轴加速度传感器数据以及 GPS 模块给出的车辆速度数据，并根据文献[3]的算法判断当前车辆行驶路面的颠簸程度.该应用软件的运行界面如图 1 所示，数据记录示例如表 2 所示.


(a) 主界面 (b) 路面颠簸实时检测结果

图 1 应用软件运行界面

Fig.1 Application user interface

利用上述应用软件，我们在福州的多条道路上进行了一个多月的实验，获取了超过 500 万条数据记录，将其作为后续仿真的数据源.

### 3.2 核密度估计结果

表 2 中，数据集每条记录的内容为：

$$\boldsymbol{X}_{ic}=\{x_i,y_i,z_i,v_i;c\}$$

其中 $x_i$，$y_i$，$z_i$和 $v_i$分别代表第 $i$ 次测量所得的三轴加速度和速度，$c$ 代表该次测量变换所得的路面颠簸程度，且 $c\in\{1,2,3,4\}$(应用软件中将颠簸程度按不颠簸、轻度、中度和重度颠簸分为 4 级，且分别对应数值 1～4).

表 2 数据记录示例

Tab.2 Examples of data

| 时间 | 车速/(km·h$^{-1}$) | $x$ 轴加速度/(m·s$^{-2}$) | $y$ 轴加速度/(m·s$^{-2}$) | $z$ 轴加速度/(m·s$^{-2}$) | 颠簸程度 |
|---|---|---|---|---|---|
| 2016－05－06 14:05:01 | 66.6 | －0.569 489 84 | 0.652 773 | 7.022 732 3 | 4 |
| … | … | … | … | … | … |
| 2016－05－14 19:10:21 | 44.1 | 0.526 362 36 | －1.407 769 3 | 7.754 511 | 4 |
| … | … | … | … | … | … |
| 2016－05－22 08:17:27 | 64.08 | －0.752 266 94 | 3.105 414 6 | 5.961 641 | 3 |
| … | … | … | … | … | … |
| 2016－06－06 14:52:51 | 39.6 | 1.493 874 30 | 1.619 646 481 | 3.777 596 5 | 2 |
| … | … | … | … | … | … |
| 2016－06－13 09:26:01 | 29.7 | 0.747 930 60 | －0.401 853 86 | 0.894 661 4 | 1 |
| … | … | … | … | … | … |

图 2 给出了 $c=2$(即轻度颠簸)情况下车辆速度和三轴加速度的核密度估计结果.从图 2 可看出车速明显不服从正态分布,与传统的统计检测法的假设前提不符.车辆发生颠簸时会造成 $z$ 轴加速度在正负值之间变化,因此图 2(c)的曲线显示出双峰结构,也不服从正态分布.

图 2(d)中虚线所示为直接使用依潘核(如式(8))进行核密度估计的结果,可以看出在速度为负值时其概率密度却不为 0,并且区间$[0,h]$($h$ 为窗宽)内的概率密度值明显小于真实值,即出现了边界效应.图 2(d)中实线给出了利用边界核进行修正后的效果,可以看出在速度为负值时的概率密度值被修正为 0,而且该曲线与点划线给出的实际概率密度曲线基本相符.

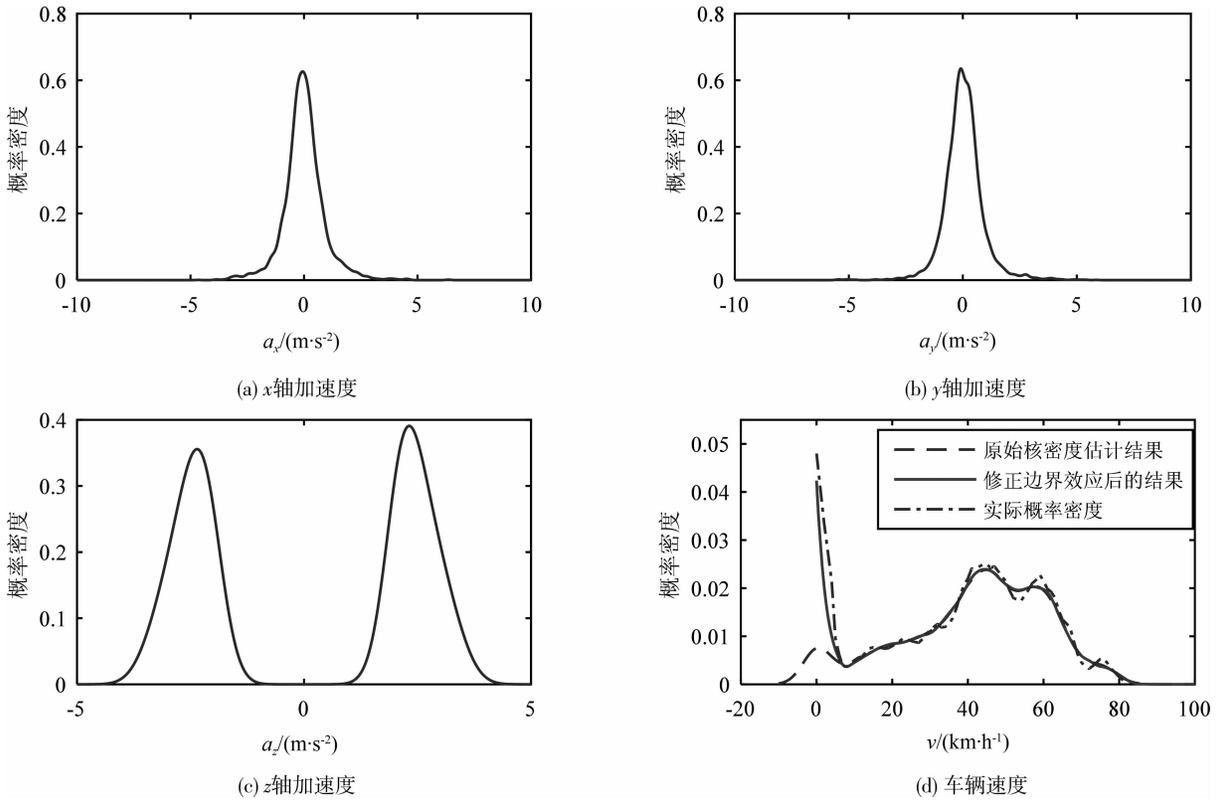



图 2　车辆速度和三轴加速度的核密度估计结果

Fig.2　Results of kernel density estimation for vehicle velocity and three-axis acceleration

### 3.3　与传统统计检测法的性能对比

传统的异常数据检测算法中,物理检测法只能检测交通流理论涉及的数据类型,不适于本文群智感知的应用场景,而传统的统计检测法中,狄克逊准则、格拉布斯准则和肖维勒准则只适用于小样本情况下的异常检测,只有拉依达准则适用于大样本情况,因此我们选择拉依达准则与本文所提算法进行性能对比.在对比实验中,本文所提算法的具体流程可参见 2.3 节,而拉依达准则的处理流程简述如下:

① 将所有数据记录按照颠簸程度 $c$ 的不同划分为 4 个子数据集 $\boldsymbol{D}_c$($c=1,2,3,4$);

② 对每个子数据集中的 4 个测量值分量(速度和三轴加速度)分别抽样,得到各个颠簸程度下,不同测量值分量的数据样本,并通过这些数据样本分别计算它们的均值与标准差;

③ 根据拉依达准则判决依据[13],若被检数据的任一分量与其对应均值的差超过 3 倍标准差,则判定该数据为异常数据.

为模拟实际应用中的恶意数据,在数据集的随机位置人为加入了若干明显异常的数据,然后根据算法检测结果计算检测率 $P_d$ 和误检率 $P_f$,以此来对比两种算法在性能上的优劣.$P_d$ 和 $P_f$ 的表达式如式(11)所示.

$$\begin{cases} P_d = N_r / S_r \\ P_f = N_f / S_d \end{cases} \tag{11}$$

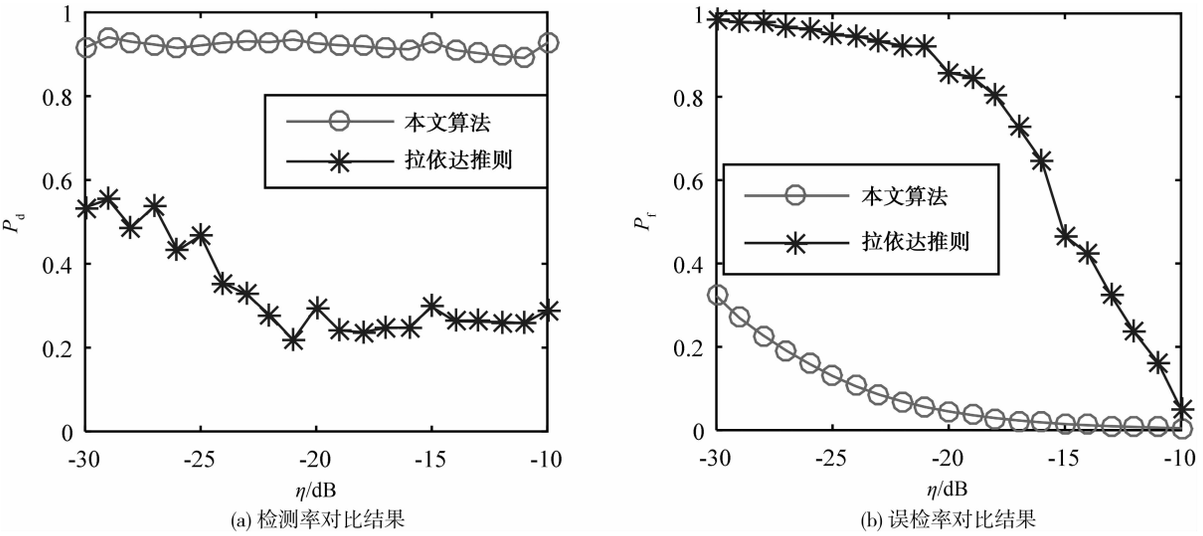
式中:$N_r$ 代表被判为异常的数据中真正异常的数据个数,$S_r$ 代表实际异常数据的个数,$N_f$ 代表被误判为异常数据的个数,$S_d$ 代表被判为异常的数据总数.

图 3 给出了两种算法在不同异常数据规模情况

下的检测率和误检率,为使仿真结果更加直观,图中横坐标 $\eta$ 是异常数据比率的对数值,具体公式为:

$$\eta=10\lg\frac{S_r}{S_n}(\mathrm{dB}) \tag{12}$$

式中 $S_n$ 代表参与异常检测的数据总数.考虑到实际应用中异常数据一般较少,所以仿真中 $\eta$ 的取值范围为−30～−10 dB(即 0.1%～10%).




(a) 检测率对比结果 (b) 误检率对比结果

图 3 与拉依达准则的性能对比

Fig.3 Performance comparison with the Pauta criterion

从图 3 可看出:① 本文所提算法的检测率远高于拉依达准则,这是因为拉依达准则基于数据服从正态分布的假设进行数据检测,而图 2(d)表明车速和 $z$ 轴加速度明显不服从正态分布(尤其 $z$ 轴加速度的分布呈双峰结构,与正态分布的单峰结构差别很大),造成拉依达准则实际检测性能差;② 本文所提算法的误检率远低于拉依达准则,其原因与①相同;③ 随着异常数据的增加,拉依达准则的检测率急剧下降,而本文所提算法的检测率却仅下降少许,这是因为在拉依达准则中,估计的样本方差大于实际样本方差,随着异常数据的增加,方差偏差加大,则一些与真实值偏差较小的异常值会被认定为正常值,从而导致检测率下降,而异常数据的增加对核密度估计几乎没有影响.综上,对比检测率和误检率,本文所提算法具备更佳的性能,且其性能可满足实际应用需求.

## 4 结 论

针对群智感知车联网系统带来的数据质量恶化问题,本文提出了一种基于核密度估计的异常数据检测算法,其区别于传统方法的优点主要在于:

①本文所提算法利用核密度估计直接从数据样本中估计出车联网数据的概率密度,而不依赖对数据分布的任何假设,避免了传统统计检测法因预设分布与实际分布不一致而导致的检测性能下降.

②算法针对群智感知场景下的特殊数据结构而提出,适用性强.

③仿真结果表明,本文所提算法具有较优的检测性能.

值得一提的是,在实际应用中,对筛查出的异常数据进行数据挖掘可能获得正常数据所不能提供的新信息,因此在某些情况下不宜直接丢弃异常数据.此时可根据式(5)将数据按信任度排序,并针对不同信任度等级的数据进行区别化的数据挖掘,从而获得一些隐藏信息.

## 参考文献

[1] 蒋斌,徐骁,杨超,等.路网拥塞控制中的多目标路径决策模型研究[J].湖南大学学报:自然科学版,2015,42(4):121−129.
JIANG Bin, XU Xiao, YANG Chao, *et al*. A multi-objective routing decision model in vehicle transport network congestion control[J]. Journal of Hunan University: Natural Sciences, 2015,42(4):121−129.(In Chinese).

[2] 邱敦国,杨红雨.一种基于双周期时间序列的短时交通流预测算法[J].四川大学学报:工程科学版,2013,45(5):64−68.
QIU Dunguo, YANG Hongyu. A short-term traffic flow forecast algorithm based on double seasonal time series[J]. Journal of Sichuan University: Engineering Science, 2013, 45(5):64−68.(In Chinese).

[3] MOHAN P, PADMANABHAN V N, RAMJEE R. Nericell: rich monitoring of road and traffic conditions using mobile sm-

artphones[C]//Proceedings of the 6th ACM conference on Embedded network sensor systems. New York: ACM Press, 2008:323－336.

[4] ANSAR1 R, SARAMPAKHUL1 P, GHOSH1s, *et al*. Evaluation of smart-phone performance for real-time traffic prediction [C]//Proceedings of 17th International IEEE Conference on Intelligent Transportation Systems. New York: IEEE, 2014: 3010－3015.

[5] 李颖宏,张永忠,王力.道路交通信息检测技术及应用[M].北京:机械工业出版社,2014:239－243.
LI Yinghong, ZHANG Yongzhong, WANG Li. Road traffic information detection technology and application[M]. Beijing: China Machine Press, 2014:239－243.(In Chinese)

[6] 徐程,曲昭伟,陶鹏飞,等.动态交通数据异常值的实时筛选与恢复方法[J].哈尔滨工程大学学报,2016,37(2):211－217.
XU Cheng, QU Zhaowei, TAO Pengfei, *et al*. Methods of real-time screening and reconstruction for dynamic traffic abnormal data[J]. Journal of Harbin Engineering University, 2016, 37(2):211－217.(In Chinese)

[7] 刘喜梅,刘义芳,高林.小样本道路旅行时间数据中的异常点剔除算法[J].青岛科技大学学报:自然科学版,2015,36(3):346－354.
LIU Ximei, LIU Yifang, GAO Lin. Algorithm outlier filtering for small simple data of travel time[J]. Journal of Qingdao University of Science and Technology: Natural Science, 2015, 36(3):346－354.(In Chinese)

[8] 陆化普,孙智源,屈闻聪.基于动态阈值的交通流故障数据实时识别方法[J].土木工程学报,2015,48(11):126－132.
LU Huapu, SUN Zhiyuan, QU Wencong. Real-time identification of traffic erroneous data based on dynamic threshold[J]. China Civil Engineering Journal, 2015, 48(11): 126－132.(In Chinese)

[9] 李成兵,姚琛.交通流异常数据检测研究及实证[J].计算机工程与应用,2013,49(20):244－246.
LI Chengbing, YAO Chen. Study of recognizing discrepant traffic data and its validation[J]. Computer Engineering and Applications, 2013, 49(20):244－246.(In Chinese)

[10] DANG T T, NGAN H Y T, LIU W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data [C]// Proceedings of IEEE International Conference on Digital Signal Processing. New York: IEEE, 2015:507－510.

[11] 李志敏,易良友,薛平,等.基于小波分析的交通流量异常数据检测[J].计算机应用研究,2011,28(5):1677－1678.
LI Zhiming, YI Liangyou, XUE Ping, *et al*. Short-term traffic flow detection based on wavelet[J]. Application Research of Computers, 2011, 28(5):1677－1678.(In Chinese)

[12] 陈珂,邹权.融入时间关联因子曲线拟合的交通流异常挖掘方法[J].计算机工程与设计,2013,34(7):2561－2565.
CHEN Ke, ZOU Quan. Traffic flow anomaly mining method of curve fitting of adding time correlation factor[J]. Computer Engineering and Design, 2013, 34(7): 2561－2565.(In Chinese)

[13] 邓泽清,陈海英.概率论与数理统计[M].北京:科学出版社,2014:26－32.
DENG Zeqing, CHEN Haiying. Probability and statistics[M]. Beijing: Science Press, 2014:26－32.(In Chinese)

[14] 王星.大数据分析:方法与应用[M].北京:清华大学出版社,2013:31.
WANG Xing. Big data analyze: methods and applications[M]. Beijing: Tsinghua University Press, 2013:31.(In Chinese)

[15] 张夏菲.非参数核密度估计负荷模型在电网可靠性评估中的应用[D].重庆:重庆大学电气工程学院,2010:31－35.
ZHANG Xiafei. The application of non-parametric kernel density estimation load model in power system reliability evaluation [D]. Chongqing: School of Electrical Engineering, Chongqing University, 2010:31－35.(In Chinese)

[16] SIMONOFFJ S. Smoothing methods in statistics[M]. Germany: Springer, 1996:49－54.