DOI:10.3969/j.issn.1003-5060.2023.09.014

基于集体离群点挖掘的城市交通异常检测研究

黄晓地[1], 朱晓曦[2], 吴淑慧[2], 胡中峰[1]
(1.合肥学院 经济与管理学院,安徽 合肥 230601; 2.合肥工业大学 管理学院,安徽 合肥 230009)

摘 要:针对时空数据环境下的城市交通异常检测问题,文章提出一种基于集体离群点挖掘的"线下拟合-线上检测"一体化检测模型。该模型采用以距离-密度-权重为度量的改进聚类(distance-density-weight *k*-medoids,DDWK-medoids)算法,根据城市交通态势自适应确定交通枢纽点的数量和位置,通过对数据进行不同分辨率拟合,将交通流量异常检测与交通轨迹异常检测相结合。在该检测模型中,数据的时间属性与空间属性均未以数值的形式直接参与计算,有效降低了运算复杂度。实验结果表明,该模型算法对于不同规模数据集的处理速度均显著优于对比算法,尤其是对于样本充足的大规模数据集,检测的准确率更具有明显优势。

关键词:城市交通;异常检测;集体离群点;基于距离-密度-权重度量的改进聚类(DDWK-medoids)算法;交通枢纽点
中图分类号:U491.265 **文献标志码**:A **文章编号**:1003-5060(2023)09-1237-10

Research on urban traffic anomaly detection method based on collective anomaly mining

HUANG Xiaodi[1], ZHU Xiaoxi[2], WU Shuhui[2], HU Zhongfeng[1]
(1. School of Economics and Management, Hefei University, Hefei 230601, China; 2. School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: Aiming at the problem of abnormal state detection of urban traffic in multi-source data environment, an integrated detection model of offline fitting and online detection based on collective anomaly mining is proposed. The model uses distance-density-weight *k*-medoids (DDWK-medoids) algorithm, which is measured by distance-density-weight, to determine the number and location of traffic pivot points according to urban traffic situation adaptively. Based on this, traffic flow anomaly detection and trajectory anomaly detection are combined by fitting multi-source traffic data at different resolutions. In the detection model, neither the temporal nor spatial attributes of the data are directly involved in the computation in the form of numerical values, which effectively reduces the computational complexity. Experimental results show that the runtime of this algorithm is significantly better than that of the comparison algorithm for test datasets of different sizes. Especially for large datasets with sufficient samples, there is clear advantage in terms of detection accuracy.

Key words: urban traffic; anomaly detection; collective anomaly; distance-density-weight *k*-medoids (DDWK-medoids) algorithm; traffic pivot point

城市范围内交通拥堵分为偶发性和常发性拥堵[1],如何尽早识别和精准定位城市路网系统中出现的偶发性拥堵,对于交通安全运营具有重要意义,围绕该主题的研究工作主要分为交通流量

收稿日期:2023-02-23;**修回日期**:2023-05-22
基金项目:国家自然科学基金资助项目(72271075);安徽省高等学校优秀青年人才基金资助项目(2022AH051774)
作者简介:黄晓地(1989—),男,安徽合肥人,博士,合肥学院讲师,硕士生导师;
朱晓曦(1987—),男,安徽合肥人,博士,合肥工业大学副教授,硕士生导师,通信作者,E-mail:zhuxiaoxi@hfut.edu.cn.

异常检测和交通轨迹异常检测2类。交通流量异常检测是通过对环型感应线圈监测器、视频监测器等固定监测器采集的交通流量、平均车速等数据进行分析,挖掘表征交通异常的数据序列[2]。由于监测器位置相对固定,流量异常检测主要以时间序列数据为对象,目前对于交通流量异常检测的研究已相对成熟[3-9],数据处理精度不断提高,但对于交通异常的主动干预能力不足。

交通轨迹异常检测主要通过拟合车辆位置、城市区域流量、道路占有率等含有历史交通数据的样本,构建城市路网轨迹数据库,再从在线数据中挖掘与期望显著偏离的轨迹序列[10]。根据空间属性的标定方法,相关研究主要分为基于区域交通数据的异常检测和基于道路交通数据的异常检测。基于区域交通数据的异常检测需要将城市地图按照一定标准进行划分,如按主干路划分[11];其检测方法[12-14]可以捕获不同区域之间交通流量的相互影响关系,如因果关系、递进关系等,能够根据城市整体交通态势的演变尽早发现异常现象,并定位异常区域。其主要局限在于:划分的路网区域相对固定,区域内的交通信息会受到天气、节假日及区域活动等诸多因素影响;城市不同区域间的交通变化存在时滞,会对异常判定造成显著影响。

基于道路交通数据的异常检测首先将浮动车(浮动车泛指有定位系统和传感设备的车辆)采集的数据按起讫路段、经纬度坐标等地理信息标定到其行驶的道路上,再通过计算道路在分析间隔内的交通运行特征(如路面占有率等),判断道路上是否出现了交通异常[15]。其检测算法[16-18]多以数据融合模型为基础,关联交通数据的时间属性与空间属性,受环境影响较小,能够根据交通实时状况及时识别并定位异常出现位置。其主要局限在于:浮动车数据采集不均衡,部分道路可能由于数据稀疏而导致误判;行驶速度、方向等数据容易受到驾驶习惯、机械故障等偶发性因素干扰,形成噪音。

针对上述问题,本文提出一种将流量异常检测与轨迹异常检测相结合、基于集体离群点挖掘的城市交通异常检测模型。首先,选择路口作为固定监测点,通过对监测点采集的不同交通工具的流量信息进行单一分析和融合分析,利用流量检测的方法检测是否出现异常;同一区域内不同媒介产生的交通流量信息可以相互增强,因此融合后的流量信息能更加全面地反映该路段的交通状态,且能有效避免数据稀疏和噪音的影响。其次,采用以距离-密度-权重为度量的改进聚类(distance-density-weight k-medoids,DDWK-medoids)算法,在连续的时间间隔内,从所有路口监测点中自适应确定交通枢纽点的数量和位置,通过交通枢纽点流量和位置的变化,从轨迹异常检测的角度判定是否出现异常;该算法打破以往区域划分固定的边界限制,可根据城市整体交通态势变化灵活调整交通枢纽点的数量及其覆盖范围。最后,通过构建"集体离群点-城市交通异常"度量规则,提高对处于潜伏期或早期阶段交通异常的甄别能力,延长预警时间,避免或降低可能造成的交通拥堵,增强城市路网系统的抗风险能力。

1 交通特征参数定义

1.1 交通数据流特征参数

城市交通是由人、车、路和环境等多种因素共同形成的复杂系统,具有明显的时空特性。对于各类交通数据流,每个数据实例可抽象为由时间属性、行为属性、空间属性构成的三元组(t,b,l)。

1)时间属性t。该属性包含2个参数,即对数据流进行分段处理的窗口宽度T_{window}和反映各类交通数据采集时点的时间矩阵$\boldsymbol{T}$。相关研究中将城市交通实时预测定义为时间跨度不超过15 min的短时交通预测[19],因此,本文将窗口宽度设置为$T_{\text{window}}=10$ min,如起始点时间标定为5:00,则下一时间点为5:10。对不同城市的交通状态进行预测,窗口宽度可灵活变化。时间矩阵$\boldsymbol{T}$为:

$$\boldsymbol{T}=\begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & & \vdots \\ t_{m1} & \cdots & t_{mn} \end{bmatrix} \tag{1}$$

其中,t_{ij}为第i类交通数据流中第j个分析间隔的时间属性,$i=1,2,\cdots,m$;$j=1,2,\cdots,n$。

2)行为属性b。本文以时间间隔内交通媒介的平均流量表征交通数据的行为属性,即$b_{ij}=q_{ij}/T_{\text{window}}$,$b_{ij}$为第$i$类交通数据流中第$j$个分析间隔的行为属性,$q_{ij}$为第$i$类交通数据流在第$j$个分析间隔内的车流量。行为属性矩阵$\boldsymbol{B}$为:

$$\boldsymbol{B}=\begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix} \tag{2}$$

3)空间属性l。交通数据空间属性标定如图1所示。图1中:x轴为经度坐标;y轴为纬度坐标。将城市地理边界视为数据集边界,设定城

市中心点为坐标原点 O，标定其空间属性为 $l_O=(0,0)$。根据城市范围内所有交通监测点与坐标原点的相对位置，将实际地理位置信息转换为坐标信息，如任一监测点 a 的空间属性为 $l_a=(x_a,y_a)$，不仅可以对实际地理信息进行归一化处理，也显著降低了空间属性的复杂度。

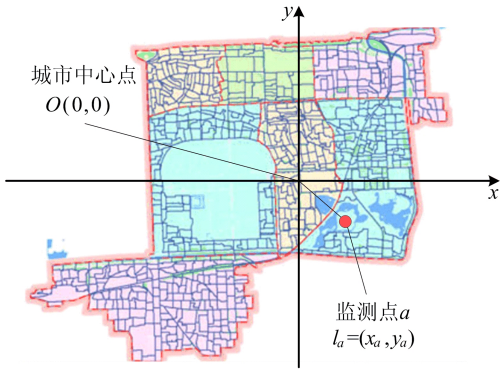


图 1　交通数据空间属性标定

1.2　交通监测点与枢纽点

1）交通监测点。交通监测点是指城市路网系统中客观存在的交通信息采集点（如配备电子设备的交通路口），以同一道路前后路口的平均流量差作为交通监测点采集数据的行为属性。

道路口为城市交通流量相对密集的地点，路口的拥堵或顺畅程度是城市交通态势的有力表征，拟合交通监测点的交通信息能够更加真实地反映城市交通状态的变化趋势，同时可以充分利用现有数据，便于数据预处理。

2）交通枢纽点。城市路网系统中的交通枢纽点如图 2 所示。

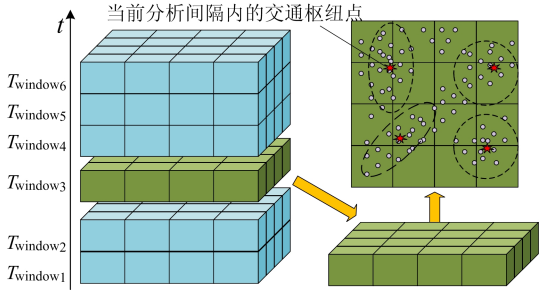


图 2　城市路网系统中的交通枢纽点

图 2 中，正方体的横截面是一个城市实际地理边界的抽象表示，其上每个点代表 1 个交通监测点。在 1 个 T_{window} 内，每个监测点采集的交通数据可表示为三元组 (t,b,l)。

将 1 个 T_{window} 内由所有交通监测点生成的数据视为 1 个数据集，$D_i=(a_1,a_2,\cdots,a_Q)$，Q 为城市边界内交通监测点的数量。对数据集 D_i，以距离度量对交通监测点地理位置信息进行聚类，聚类簇的中心点即交通枢纽点，反映当前分析间隔 i 内城市路网系统中交通流量最密集的位置。本文设计的 DDWK-medoids 算法，可自适应确定分析窗口内满足条件的交通枢纽点数量和位置。

2　检测模型

2.1　集体离群点

根据离群点异常表现方式的特征，可将其分为点离群点、情景离群点和集体离群点[20]。集体离群点通常是由一系列相关数据实例组成，当它们以某种模式共同出现时，明显偏离数据分布的正常期望，但每个数据实例单独分析时不构成离群点[21]。

2.2　基于集体离群点的异常交通状态度量

在样本数据可用性的基础上，本文选择 4 类城市交通数据进行分析，分别是公交车数据集 s_1、出租车数据集 s_2、非机动车数据集 s_3 及其他交通媒介数据集 s_4。在样本数据中，数据集 s_3 占比约 10%，其精确程度对分析精度影响较小，因此，通过共享单车各停放点实时更新的流量数据 f 和其实际地理位置对应的权重 φ 近似得到 s_3。数据集 s_4 包括行人、客货车、私家车等的流量信息，此类数据来源繁杂且完整性不足，基于各类别交通流量变化在同一城市区域内具有一致性的特点，本文在考虑地理因素基础上，通过加权其他 3 类数据集近似表示 s_4。在实际应用中，可以根据城市实际情况，对权重做适当调整，具体信息见表 1 所列。

表 1　4 类交通数据流的具体信息

数据集	一环范围内	二环范围内	二环范围外
s_1	实际采集信息	实际采集信息	实际采集信息
s_2	实际采集信息	实际采集信息	实际采集信息
s_3	$\varphi_1 f$	$\varphi_2 f$	f
s_4	$\varepsilon_1(s_1+s_2+s_3)$	$\varepsilon_2(s_1+s_2+s_3)$	$\varepsilon_3(s_1+s_2+s_3)$

注：φ_1、φ_2 为 s_3 的地理位置权重，$\varphi_1=2.0$，$\varphi_2=1.5$；ε_1、ε_2、ε_3 为 s_4 的地理位置权重，$\varepsilon_1=0.55$，$\varepsilon_2=0.30$，$\varepsilon_3=0.15$。

基于集体离群点的异常交通状态度量主要分为 4 个层次：① 对交通监测点采集的任意一类交通数据流进行异常检测；② 对交通监测点上多类交通数据融合后的累加变化程度进行异常检测；③ 对交通枢纽点的数据变化趋势进行异常检测；④ 对多个交通枢纽点融合的流量变化趋势进行

异常检测。

2.3 模型架构

本文集体离群点检测模型主要分为线下拟合和实时检测 2 个阶段,模型框架如图 3 所示。

1) 线下拟合阶段。首先依据城市地理边界,标定各交通监测点采集数据的时间属性、行为属性和时间属性,拟合各监测点上的单类交通数据和整体交通流量变化趋势,由于监测点物理位置固定,各类交通数据的拟合不涉及空间属性;然后,以 DDWK-medoids 算法自适应确定各分析间隔内的交通枢纽点,拟合交通枢纽点地理位置和交通流量的变化趋势;最后,根据交通枢纽点计算中心点,通过连续分析间隔的中心点变化趋势,拟合城市整体交通状态的变化趋势。

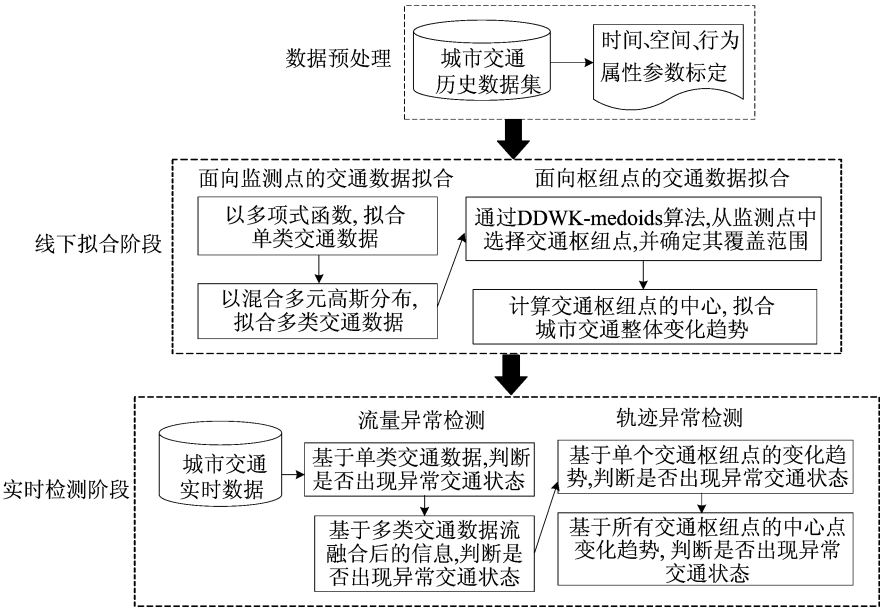


图 3 本文检测模型框架

2) 实时检测阶段。基于线下拟合结果,判断城市路网系统是否出现由集体离群点表征的交通异常。

3 算法框架

交通监测点采集的单类交通数据可表示为二元组(t,b),按精度要求以 MATLAB 函数进行多项式拟合,即

$$p = \text{polyfit}(\boldsymbol{T},\boldsymbol{B},\tau),$$

其中:p 为阶次从高到低的多项式系数;$\boldsymbol{T}=[t_1 \quad t_2 \quad \cdots \quad t_N]$,$N$ 为数据量;$\boldsymbol{B}=[b_1 \quad b_2 \quad \cdots \quad b_N]$;$\tau$ 为多项式阶数。通过改进的混合多元高斯函数,对同一交通监测点采集的多类交通数据进行拟合,以不动点迭代法优化函数参数[22]。

融合后的交通监测点数据可表示为三元组(t,b,l)。本文设计一种 DDWK-medoids 算法,对城市路网系统中的交通数据进行聚类处理,从交通监测点中自适应确定当前时间间隔内的交通枢纽点位置与数量,并据此计算交通中心点,算法框架由参数取值范围自适应确定、预备簇中心点挑选、初始簇中心点筛选、条件重定位迭代、参数对比择优 5 个部分组成,如图 4 所示。

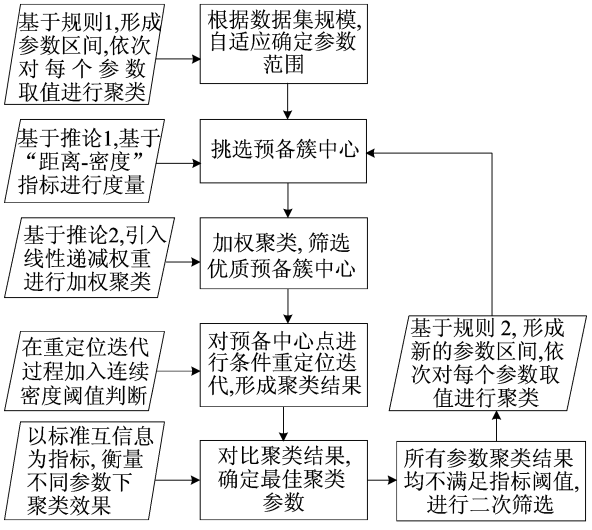


图 4 DDWK-medoids 算法框架

3.1 自适应算法参数取值范围

将期望聚类簇数的取值范围等分,以等分段的上边界值作为区间代表值。引入循环判断准则对参数区间划分的合理性进行判断,提高算法准

确率。具体过程见规则 1 和规则 2。

规则 1　对于包含 Q 个数据的集合，期望聚类簇数 $k\in[2,\sqrt{Q}]$，k 的计算公式为：

$$k=\lfloor k\rfloor,\qquad k=\begin{cases}\sqrt{Q}/p, & \sqrt{Q}/p>2;\\ 2, & \sqrt{Q}/p\leqslant 2\end{cases}\tag{3}$$

其中，p 为[1，10]内正整数。k 的取值集合为 $K=\{k_1,k_2,\cdots,k_m\}$。基于 k 的取值，每个聚类簇内约包含 $\lfloor Q/k\rfloor$ 个数据对象。通过距离度量确定数据集中心点，以中心点范围至少包含 $\lfloor Q/k\rfloor$ 个数据对象的半径 ε 为邻域半径，以 $\lfloor Q/k\rfloor$ 为密度阈值 ρ。

在算法参数对比择优环节，若按照规则 1 确定的算法参数的聚类效果均不满足度量指标的阈值要求，则按照规则 2 进行参数二次选择。

规则 2　选择质量最佳的 2 个期望聚类簇数作为新的参数选择范围区间 $[k_{best1},k_{best2}]$。在新区间内，计算新的期望聚类簇数 k'，即

$$k'=\lfloor k'\rfloor,\quad k'=\sqrt{Q'}/p\tag{4}$$

其中，$Q'\in\mathbf{N}^+$，$Q'\in[\lfloor Q/k_{best2}\rfloor,\lfloor Q/k_{best1}\rfloor]$。

由 k' 取值得到新集合 K'，基于 K'，确定新邻域半径 ε' 与密度阈值 ρ'。在参数对比择优环节，若聚类质量满足度量指标阈值要求，算法终止；否则继续按规则 2 迭代。

3.2　预备簇中心挑选

最佳聚类结果应满足簇内对象高度相似且簇间对象最大程度相异的标准，即簇中心点之间的距离应尽可能远，避免集中在较小范围，陷入局部最优；同时，为避免噪声、离群数据的影响，簇中心点应具备相对较高的邻域密度。

推论 1　数据集最佳聚类结果应使得各个簇中心点的邻域密度相对较高且不同簇中心点之间的距离相对较远。

根据推论 1，以“距离-邻域密度”为度量标准选取预备簇中心，通过删除已形成簇缩小数据集规模，减少数据间的重复计算，具体步骤如下。

1）在数据集中，挑选邻域密度最大的数据 c_1，作为第 1 个预备簇中心，将其邻域半径内的 S_1 个数据聚到一起，形成第 1 个簇 C_1。

2）将 C_1 从原数据集 D 中删除，形成新数据集 D_1，数据规模为 $|D|-|S_1|$。

3）在数据集 D_1 中，以欧式距离为度量，搜寻距离 c_1 最远的数据，若该数据的邻域密度满足密度阈值要求，则将其作为第 2 个预备簇中心；否则选择距离 c_1 次远的数据进行邻域密度度量，依此类推，选择满足密度阈值要求且距离 c_1 足够远的 c_2 作为第 2 个预备簇中心。将 c_2 邻域半径内的 S_2 个数据聚到一起，形成第 2 个簇 C_2。

4）将 C_2 从数据集 D_1 中删除，形成新数据集 D_2，数据规模为 $|D_1|-|S_2|$。

5）计算数据集 D_2 中任意数据 α 到 c_1、c_2 的相对距离 R_d，计算公式为：

$$R_d=\min\{\mathrm{dist}(\alpha,c_1),\mathrm{dist}(\alpha,c_2)\}\tag{5}$$

搜寻满足密度要求且相对 c_1、c_2 距离均足够远的数据作为第 3 个预备簇中心 c_3。聚集 c_3 邻域半径内的 S_3 个数据作为第 3 个簇 C_3。

6）将 C_3 从数据集 D_2 中删除，形成新数据集 D_3，数据规模为 $|D_2|-|S_3|$。

7）重复上述步骤，直到数据集 D 中任意数据都不满足密度阈值要求，终止计算，输出所有的预备簇中心集合 $\{c_1,c_2,\cdots,c_k\}$。

3.3　初始簇中心筛选

在挑选预备簇中心过程中，由于对比空间缩小，后续选出的特征样本相对质量弱于先选出的。

推论 2　首个筛选出的预备簇中心点质量最好，随着数据集规模逐渐缩小，后续筛选出的中心点质量可能呈现递减趋势。

因此，引入惯性权重思想，进一步筛除预备簇中心点集中的劣质簇中心，具体步骤如下。

1）按预备簇中心产生顺序赋予线性递减权重。对先选出的初始簇中心赋予更大的权重，加快局部收敛，控制求解精度与迭代次数；对排序靠后的预备簇中心赋予较小权重，增强全局搜索能力，避免收敛于劣质簇中心形成的局部最优解。权重 w 计算公式为：

$$w=w_{\max}-\frac{n(w_{\max}-w_{\min})}{n_{\max}}\tag{6}$$

其中：$n_{\max}$ 为预备簇中心点总数；n 为当前预备簇中心点产生顺位编号；$w_{\max}=0.9$；$w_{\min}=0.4$。

2）进行权重聚类，对迭代后新形成簇的中心点进行邻域密度判断，若存在劣质簇，即新中心点邻域密度不满足要求，转入步骤 4）；若不存在劣质簇，转入步骤 3）。任意数据 α 与当前簇中心点的加权距离度量计算公式为：

$$\mathrm{dist}(\alpha,c_\eta)=\frac{1}{w}\sqrt{(x_\alpha-x_{c_\eta})^2+(y_\alpha-y_{c_\eta})^2}\tag{7}$$

3）计算权重聚类后新形成簇的中心点，将这些中心点作为初始簇中心，进入条件重定位迭代

环节。

4) 删除形成劣质簇的原始预备簇中心点,对其余的原始预备簇中心进行权重迭代,判断是否存在劣质簇,若不存在,转入步骤3);若出现劣质簇,继续迭代计算,直至筛除全部劣质预备簇中心。

3.4 条件重定位迭代

在每次簇中心重定位迭代后加入密度度量,标记不满足阈值要求的劣质簇中心。若被标记的劣质簇中心在后续迭代过程中,聚类质量呈下降趋势(簇密度持续下降),则删除该劣质簇中心,从而尽早筛除可能出现的劣质簇中心,避免冗余迭代,加快算法收敛速度与稳定性。具体步骤如下:

1) 基于当前簇中心点集,进行重定位迭代,对新形成的簇进行密度度量。

2) 若聚类簇的簇内密度均满足密度阈值要求,则清空劣质簇标记,继续按步骤1)处理,直至算法收敛或达到最大迭代次数;若出现不满足密度阈值要求的聚类簇,则标记其为劣质簇,转入步骤3)。

3) 若该簇第1次被标记,则记录当前簇内密度后,继续按步骤1)处理;若该簇已存在标记,则记录当前簇内密度,转入步骤4)。

4) 设定控制精度 ζ,若某簇连续低于密度阈值且聚类质量呈下降趋势,则删除该簇,转入步骤1)。参数 ζ 表示某个簇被允许标记为劣质簇的最大次数。

3.5 算法参数对比择优

选择标准互信息(normalized mutual information,NMI)作为度量指标对不同参数对应的聚类结果进行对比,确定最优聚类参数。

3.6 针对多簇数据集的算法改进

在聚类多簇数据集,即真实聚类簇数接近最大聚类簇数的数据集($k\approx\sqrt{Q}$),聚类簇内的数据与相邻簇的簇内数据之间距离可能小于该数据到簇中心的距离。DDWK-medoids 算法在预备簇中心挑选环节,前几个选出的预备簇中心会将远超过密度阈值的点划分到各自簇中,从而无法挑选出足够的预备簇中心,使预备簇中心数量出现较大偏差,且此偏差无法通过后续步骤得到改善,导致聚类结果的准确性大大降低。

因此,在面向多簇数据集的聚类处理中,在预备簇中心挑选环节,加入密度限定控制,以保证能够找到足够数量的预备簇中心。算法调整如下:对于每个挑选出的符合要求的预备中心点,只从当前数据集中删除离它最近的$\sqrt{Q}$个点,直到数据集中不存在满足密度阈值的点。

4 算法实验

4.1 数据集

选择 UCI(University of California,Irvine)数据库(http://archive.ics.uci.edu/ml/datasets.php)中6个数据集 Iris、Seeds、Survival、Knowledge(即 Knowledge Modeling)、Perfume、Absenteeism 进行测试,所有数据集都经过归一化预处理,数据集具体信息见表2所列。分别用传统 K-medoids 算法与 DDWK-medoids 算法,对5个标准数据集进行对比;采用按3.6节改进的 DDWK-medoids 算法和传统 K-medoids 算法,对 Absenteeism、Perfume 2个多簇数据集进行对比。每个数据集处理50次,计算50次聚类结果评价的最大值、最小值和均值。

表2 数据集具体信息

数据集	数据量	属性数	实际聚类簇数	备注
Iris	150	4	3	标准
Seeds	210	7	3	标准
Survival	306	3	2	标准
Knowledge	403	5	4	标准
Perfume	560	2	20	标准、多簇
Absenteeism	740	21	22	多簇

4.2 聚类质量评价指标

1) 采用轮廓系数(silhouette coefficient,SC),结合内聚度和分离度2种因素,通过计算数据集中所有对象的 SC 平均值对聚类结果进行评价,SC 的取值越接近1,聚类结果越合理;若 SC 取值接近-1,则聚类结果不合理。

2) 采用 NMI,通过对比数据集的实际标签分布和聚类后的分布,对聚类结果进行评价。合理聚类结果的 NMI 取值范围为[0,1],取值越大,聚类结果与真实情况越吻合。

4.3 测试结果分析

最大迭代次数预先设置为100次。对于 DDWK-medoids 算法,设置精度控制参数 NMI 取值为0.55、$\zeta=3$。基于 SC 和 NMI,对2种算法在6个数据集上50次聚类结果的评价结果见表3所列。

由表3可知,在5个标准数据集上,与传统 K-medoids 算法相比,除 Perfume 数据集外,DDWK-medoids 算法的聚类质量有显著优势。这是由于 DDWK-medoids 算法能够消除初始参数选择的随

机性，使得聚类结果总是唯一的，聚类过程更加稳定。对比样本标签，Perfume 数据集为多簇数据集，DDWK-medoids 算法在 SC 与 NMI 2 个评价指标上均弱于传统 K-medoids 算法。

表 3　2 种算法在 6 个数据集上的聚类结果对比

数据集	传统 K-medoids 算法						DDWK-medoids 算法	
	SC			NMI			SC	NMI
	最大值	最小值	均值	最大值	最小值	均值		
Iris	0.689	0.599	0.632	0.770	0.669	0.718	0.716	0.818
Seeds	0.698	0.605	0.645	0.714	0.615	0.676	0.720	0.762
Survival	0.679	0.588	0.629	0.730	0.647	0.689	0.709	0.776
Knowledge	0.668	0.579	0.620	0.750	0.664	0.706	0.704	0.796
Perfume(标准)	0.619	0.589	0.601	0.648	0.618	0.622	0.556	0.590
Perfume(多簇)	0.619	0.589	0.601	0.648	0.618	0.622	0.688	0.734
Absenteeism	0.604	0.536	0.579	0.617	0.564	0.592	0.653	0.717

由表 3 可知，按 3.6 节改进的 DDWK-medoids 算法在 SC 和 NMI 2 个评价指标上均明显优于传统 K-medoids 算法，且聚类结果依然稳定，由此可见 DDWK-medoids 算法处理多簇数据集的有效性。

对于多簇数据集，改进的 DDWK-medoids 算法聚类效果更佳，但挑选出的预备簇中心数量较多，在迭代次数和收敛速度等指标上弱于原 DDWK-medoids 算法。因此，对于有先验知识的数据集，可选择有针对性的参数设置策略；对于没有先验知识的数据集，可通过对比 2 种参数设置策略下的聚类质量，择优确定最终的聚类结果。

5　实例测试

基于本文检测模型对合肥市 2019—2020 年的历史交通数据进行拟合，并据此对 2021 年 5 月交通数据进行预测，识别其中可能存在的由集体离群点表征的异常交通状态。城市路网交通模型地理边界由合肥市区与肥西县、肥东县、长丰县的边界及绕城高速公路圈定，在边界内有 16 786 个数据可用的交通监测点。

5.1　测试过程可视化示例

1）监测点单类交通数据异常状态检测。基于单类交通数据的异常交通状态预测结果如图 5 所示。对比样本数据拟合出的历史变化趋势，时刻 t_1、t_2 的实时交通流量明显超出历史流量最高线，可判定该分析时刻内监测点采集的交通数据构成集体离群点，区域交通可能出现异常。

2）监测点多类交通数据异常状态检测。基于历史数据拟合出多类交通数据变化趋势如图 6 所示。对比实时采集的交通数据，$[t_1, t_2]$时间段内的各类交通流量融合后的变化幅度明显超过该时段的流量历史最高线，因此可判定在该时段内监测点采集的各类交通数据构成集体离群点，当前区域的交通状态可能出现异常。

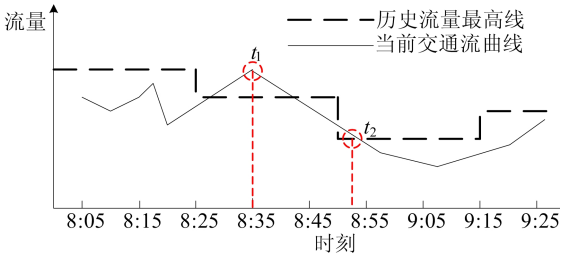


图 5　基于单类交通数据的异常交通状态预测结果

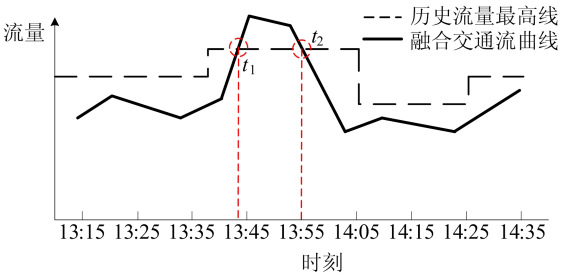
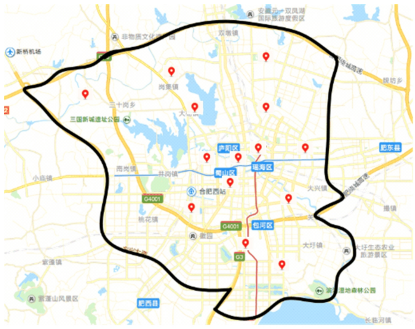


图 6　基于多类交通数据的异常交通状态预测结果

3）交通枢纽点异常交通状态检测。基于 DDWK-medoids 算法确定各时间间隔内的交通枢纽点，拟合枢纽点覆盖范围内交通状态的实时变化趋势。通过对比相同时间内交通枢纽点的历史变化趋势，检测该交通枢纽点覆盖范围内的交通数据中是否存在由多个监测点共同构成的集体离群点，并据此判断是否出现异常交通状态。

对实验结果进行分析发现，交通枢纽点的数量和位置会随着不同的时间发生变化，主要分为工作日和节假日期间。标定的合肥市路网交通模

型如图 7 所示(https://map.baidu.com/search/合肥市)。从图 7a 可以看出,工作日期间的交通枢纽点有 14 个;从图 7b 可以看出,节假日期间的交通枢纽点有 9 个。



(a) 工作日

(b) 节假日

图 7 城市路网交通模型边界内不同时间下的交通枢纽点分布

基于历史数据拟合出某交通枢纽点的变化趋势如图 8 所示。

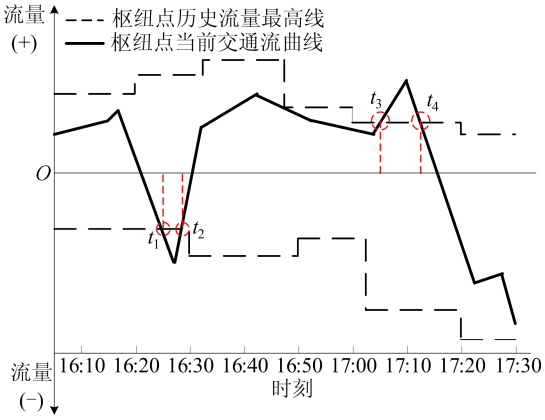


图 8 基于交通枢纽点的异常交通状态预测结果

在纵坐标上,交通枢纽点往城市中心方向的流量变化标记为“+”,背离城市中心方向的流量变化标记为“−”。对比实时采集的交通数据,该交通枢纽点在$[t_1,t_2]$和$[t_3,t_4]$时间段内对应的交通流量明显超过历史流量最高线。该枢纽点覆盖范围内的各监测点,在该时间段内采集的交通数据共同构成集体离群点,该交通枢纽点覆盖范围内的交通状态可能出现异常。

4) 交通中心点异常交通状态检测。计算交通枢纽点的中心点,根据历史交通数据拟合交通中心点位置变化趋势,如图 9 所示。图 9 中:虚线为基于历史数据拟合出的交通中心点位置的变化趋势;在纵坐标上,将中心点往城市中心方向的距离变化标记为“+”,背离城市中心方向的距离变化标记为“−”。

对比实时采集的交通中心点位置变化数据,中心点在$[t_1,t_2]$和$[t_3,t_4]$时间段内的位置偏移量明显超过历史最高值,即表明存在由多个交通枢纽点共同构成的集体离群点。通过进一步判断中心点位置异常变化的方向,可识别出现交通状态异常的具体枢纽点及其覆盖范围。

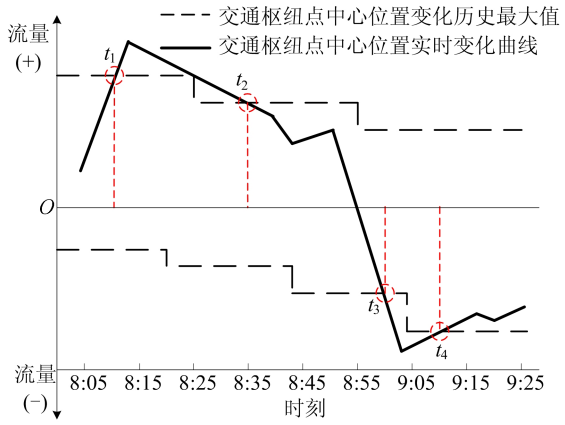


图 9 基于交通中心点的异常交通状态预测结果

5.2 测试结果

通过拟合 2019—2020 年的历史交通数据样本,对 2021 年 5 月交通数据进行预测。样本标注的异常交通状态指对城市交通流畅性产生显著影响的交通故障,即在城市路网中出现明显的交通能力下降、交通流失效、磁滞等现象,并非实际交通事故报警统计。5 月单日内不同时间段交通异常度量的箱线图分布如图 10 所示。

从图 10 可以看出,工作日出现的交通异常远多于周末或节假日,尤其是时间段[7,9)、[13,15)、[19,21)。其中[7,9)为早高峰时间段,[13,15)为午饭和上班、上学时间段,[19,21)为晚高峰时间段,均与实际交通状况相符。

5 月逐 5 d 预测结果准确率与误报率见表 6 所列,预测的平均准确率为 91.46%,平均误报率为 4.46%。预测准确率是指与实际故障标签对

比，正确预测的交通故障概率，为预测的交通故障标签数与标定的实际交通故障标签数的比值。误报率是指将正常交通状态误识别为异常状态的概率，为误识别次数与正确预测次数的比值。

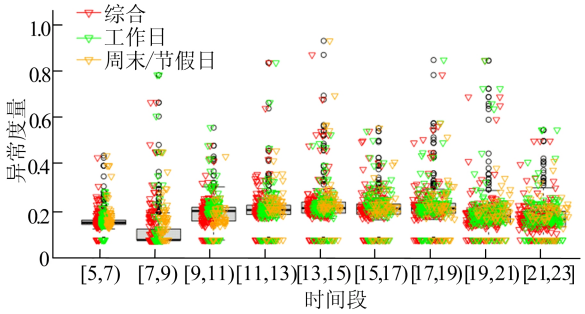


图 10　5 月单日内不同时间段交通异常度量的箱线图分布

表 6　5 月逐 5 d 检测结果与实际标签对比结果

日期	实际交通故障标签数	预测交通故障标签数	预测准确率/%	误报率/%
1—5	11 432	11 161	93.67	4.23
6—10	9 563	9 229	91.34	5.66
11—15	10 268	9 344	88.21	3.16
16—20	7 792	7 372	89.73	5.44
21—25	11 369	10 870	91.30	4.72
26—30	9 431	9 228	94.51	3.53

5.3　对比实验

对比实验选择北京城市交通数据（https://www.beijingcitylab.com/），实验主要对比线上检测结果的效率，对比算法选择狄利克雷过程混合模型（Dirichlet process mixture model，DPMM）[23]、主成分分析（principal component analysis，PCA）法[4]及流量矩阵法 SETMADA（Simultaneously Estimate Traffic Matrix and Detect Anomaly）[24]。

本文算法与对比算法在测试数据集上的运行时间如图 11 所示。

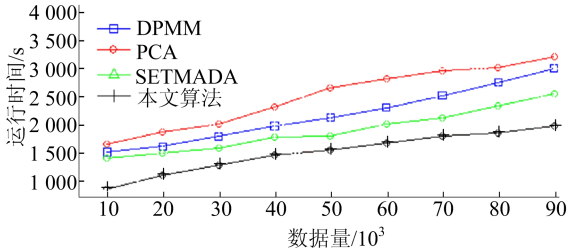


图 11　本文算法与对比算法在测试数据集上的运行时间曲线

在本文检测模型中，数据的时间属性，如某个交通数据产生的具体时刻信息，未直接参与到模型计算中，其地理位置的空间属性也通过坐标转换的方式降低了复杂度，因此，在不同数据量上的运行速度明显优于 3 种对比算法。

4 种算法不同数据量下检测的准确率如图 12 所示。从图 12 可以看出，本文检测模型在大规模数据处理上具有明显优势，能够从大量样本数据集中不断完善数据拟合程度。

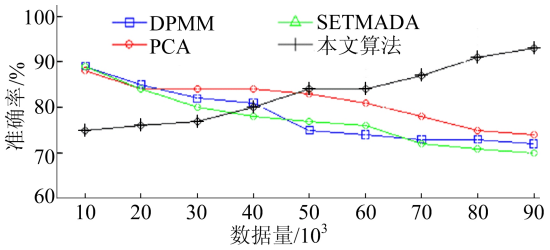


图 12　本文算法与对比算法在测试数据集上的准确率曲线

6　结　　论

本文提出基于集体离群点挖掘的城市交通异常检测方法，其核心思想是将城市交通异常模式检测问题转换为面向多源交通数据流的集体离群点挖掘问题，实例测试和对比实验结果表明，本文算法具有以下优势：

1）提高异常检测准确性。从多个粒度对城市交通数据进行融合分析时，变化程度的叠加效应有助于更加准确地发现潜在交通异常。基于集体离群点挖掘，不仅可识别单一监测点上由多数据流共同反映的集体异常，还可识别交通枢纽点覆盖范围内多个监测点共同构成的集体异常。结果表明本文算法对交通异常状态预测的平均准确率为 91.46%，平均误报率为 4.46%；在对比实验中，随样本数据量增大，检测准确率呈上升趋势。

2）降低模型计算复杂度。在 DDWK-medoids 算法模型中，数据的时间属性与空间属性未以数值的形式直接参与计算，有效降低了运算复杂度，检测效率显著提高。实验表明，对于不同数据量的样本数据集，本文算法处理速度均明显高于对比算法。

3）线下与线上相结合提高检测效率。拟合阶段属于线下处理，既保留了大数据实证分析带来的高精度优势，也不影响基于集体离群点挖掘的交通状态实时预测效率。实时检测阶段可以将流量异常检测与轨迹异常检测结合起来，进一步提高检测的有效性。实验表明，样本数据量越大，本文算法的线下拟合越准确，检测准确率越高。

[参 考 文 献]

[1] BACHECHI C, ROLLO F, PO L. Detection and classification of sensor anomalies for simulating urban traffic scenarios[J]. Cluster Computing: the Journal of Networks, Software Tools and Applications, 2022, 25: 2793-2817.

[2] DJENOURI Y, BELHADI A, CHEN H C, et al. Intelligent deep fusion network for urban traffic flow anomaly identification[J]. Computer Communications, 2022, 189: 175-181.

[3] XIAO F, CHEN L, ZHU H, et al. Anomaly-tolerant network traffic estimation via noise-immune temporal matrix completion model[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1192-1204.

[4] LI M, HAN D Z, LI D, et al. MFVT: an anomaly traffic detection method merging feature fusion network and vision transformer architecture[J]. EURASIP Journal on Wireless Communications and Networking, 2022, 2022(1): 39-60.

[5] 吴中明,李敏,徐红利. 解动态交通配流问题的快速投影梯度算法[J]. 系统工程理论与实践, 2021, 41(10): 2696-2709.

[6] 祁朵,毛政元. 基于自适应时序剖分与KNN的短时交通流量预测[J]. 地球信息科学学报, 2022, 24(2): 339-351.

[7] HUSSAIN F, LI Y F, ARUN A, et al. A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts[J]. Analytic Methods in Accident Research, 2022, 36: 100248.

[8] 胡松,林鹏飞,翁剑成,等. 基于改进Apriori算法的乘客公共交通依赖性层级转移分析[J]. 东南大学学报(自然科学版), 2022, 52(2): 344-351.

[9] YANG Y, TIAN N, WANG Y P, et al. A parallel FP-growth mining algorithm with load balancing constraints for traffic crash data[J]. International Journal of Computers, Communications & Control, 2022, 17(4): 4806.

[10] 李超能,冯冠文,刘如意,等. 一种基于重构误差的交通轨迹异常检测方法[J]. 计算机科学, 2022, 49(2): 149-155.

[11] 高志波,吴志周,郝威,等. 智能网联车环境下交叉口车流轨迹优化模型[J]. 交通运输系统工程与信息, 2021, 21(2): 91-97.

[12] HUANG G L, DENG K, HE J. Cognitive traffic anomaly prediction from GPS trajectories using visible outlier indexes and meshed spatiotemporal neighborhoods[J]. Cognitive Computation, 2020, 12: 967-978.

[13] 黄艳国,刘红军,金超. 基于数据挖掘的路网交通拥堵特征分析[J]. 科学技术与工程, 2022, 22(29): 13083-13089.

[14] ZHANG X C, ZHENG Y, ZHAO Z X, et al. Deep learning detection of anomalous patterns from bus trajectories for traffic insight analysis[J]. Knowledge-Based Systems, 2021, 217: 106833.

[15] 周启帆,董志鹏,徐银,等. 基于轨迹数据的大规模路网交通拥挤时空关联规则挖掘[J/OL]. 系统仿真学报, 2022: 1-11. (2022-12-09)[2023-05-20]. https://doi.org/10.16182/j.issn1004731x.joss.22-0898.

[16] YU W H, HUANG Q H. A deep encoder-decoder network for anomaly detection in driving trajectory behavior under spatio-temporal context[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 115: 103115.

[17] WANG X D, SUN L J. Diagnosing spatiotemporal traffic anomalies with low-rank tensor autoregression[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(12): 7904-7913.

[18] CRUZ M, BARBOSA L. Learning GPS point representations to detect anomalous bus trajectories[J]. IEEE Access, 2020, 8: 229006-229017.

[19] 《中国公路学报》编辑部. 中国交通工程学术研究综述·2016[J]. 中国公路学报, 2016, 29(6): 1-161.

[20] CHATTERJEE A, AHMED B S. IoT anomaly detection methods and applications: a survey[J]. Internet of Things, 2022, 19: 100568.

[21] WANG C H, ZHOU H, HAO Z Q, et al. Network traffic analysis over clustering-based collective anomaly detection[J]. Computer Networks, 2022, 205: 108760.

[22] 黄晓地,朱晓曦,胡中峰. 一种基于集体离群点检测变速箱故障的方法[J]. 安徽理工大学学报(自然科学版), 2022, 42(6): 29-36.

[23] BAGHDADI A, MANOUCHEHRI N, PATTERSON Z, et al. Hierarchical Dirichlet and Pitman-Yor process mixtures of shifted-scaled Dirichlet distributions for proportional data modeling[J]. Computational Intelligence, 2022, 38(6): 2095-2115.

[24] GUO K, HU Y L, QIAN Z, et al. Dynamic graph convolution network for traffic forecasting based on latent network of Laplace matrix estimation[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(2): 1009-1018.

(责任编辑 张淑艳)