文章编号：1000-8608(2022)04-0403-08

基于改进 *K*-Medoids 的组合聚类算法及异常值检测研究

贺玉海[1,2,3]， 周庆琨[1]， 程琰晟[1]， 王勤鹏*[1,2,3]

（1.武汉理工大学 船海与能源动力工程学院，湖北 武汉 430063；
2.武汉理工大学 船舶动力工程技术交通行业重点实验室，湖北 武汉 430063；
3.武汉理工大学 船舶与海洋工程动力系统国家工程实验室电控分实验室，湖北 武汉 430063）

摘要： 采用聚类算法和异常值检测算法进行车辆轨迹信息的提取与挖掘，在交通控制与管理、道路拥堵时空分析与治理、用户出行线路规划与推荐，以及自动驾驶决策规划等应用中具有重要意义.针对现有聚类算法和异常值检测算法参数难以控制、算法存在随机性的不足，提出基于 *K*-Medoids 与 DBSCAN 组合的聚类算法.通过对模拟十字交叉路口数据集的训练，得到一个交叉路口最佳聚类模型，并用真实轨迹数据集验证、优化该模型.然后，将交叉路口区域内一段时间内的轨迹聚类数据流进行可视化再现，取得了异常轨迹少、聚类速度快的聚类效果，同时比较选择出算法各个参数的最优值.最后，通过参数传递使 DBSCAN 算法能够更精确地识别出异常轨迹，为交通治理与自动驾驶决策提供指导.

关键词： 车辆轨迹；聚类分析；异常值检测；相似性度量；DBSCAN 算法
中图分类号： TP391 **文献标识码：** A **doi：** 10.7511/dllgxb202204009

0 引 言

随着社会经济与科技的快速发展，国民生活质量正在不断提高.汽车保有量的快速增长，虽然给人们带来了出行上的便利，但也给城市带来了日益严重的交通拥堵以及环境污染等问题.近年来，伴随着人工智能（AI）、机器视觉、物联（车辆）网、5G 等大数据采集与传输技术的日益完善，可以对大量车辆行驶自然轨迹数据进行分析[1]，挖掘特征参数，帮助人们理解轨迹数据中隐含的规律与趋势，给城市道路交通管理与规划提供指导，对于缓解城市交通拥堵发挥了重要作用[2].

聚类是挖掘轨迹数据最常用的手段之一[3-5]，轨迹聚类可以较为精确地识别目标轨迹的行为方式，并对于异常的噪声轨迹进行识别与剔除[6].轨迹聚类一方面可以为研究如潮汐交通的内在规律提供帮助，为大车流的规划、控制提供决策性指导[7]；另一方面，对于无人驾驶汽车来说，异常轨迹的识别也可以降低事故发生的概率，提高安全性[8].郝美薇等[9]提出一种改进的基于密度的 *K*-Means 算法.该算法基于轨迹数据增加轨迹数据关键点密度权值和分布密度选取高密度的轨迹数据点作为初始聚类中心进行 *K*-Means 聚类.张玉西等[10]提出一种改进主成分分析和基于密度的改进 *K*-Means 聚类组合方法，消除了 *K*-Means 算法对初始聚类中心的敏感性及噪声点的干扰.何月等[11]采用了一种基于网格的聚类算法，通过对出租车 GPS 轨迹数据进行处理和聚类分析，最后得到了出租车轨迹热力图并确定了出租车的运营规律.郭景华等[12]通过马尔可夫链蒙特卡罗模拟预测未来时刻的车辆状态，有效解决了不能准确表征前车人类驾驶员驾驶车辆随机运动的问题.

轨迹的相似性度量是轨迹聚类方法的基础，冯琦森[13]与 Zheng 等[14]基于最长公共子序列的轨迹相似度测量方法，并结合 DBSCAN（density-based spatial clustering of application with

收稿日期： 2022-03-09； **修回日期：** 2022-05-22.
基金项目： 国家自然科学基金资助项目（51009112）.
作者简介： 贺玉海（1976-），男，博士，教授，E-mail：hyh@whut.edu.cn；周庆琨（2001-），男，本科生，E-mail：1617542253@qq.com；程琰晟（2000-），男，本科生，E-mail：3369610249@qq.com；王勤鹏*（1983-），男，博士，讲师，E-mail：wangqpkevin@163.com.

noise)算法设计了一种可以识别居民出行热点路线的出租车轨迹数据聚类算法.马占宇[15]通过计算轨迹间的双向距离,提高了相似性计算的准确性,使其适用于从不等长的轨迹数据中提取热点.

异常轨迹检测作为轨迹聚类的目标,切实应用于日常的交通运输规划、智能汽车避障等场景.蒋恩源等[16]利用三帧差法、最小二乘法、自适应分段直线拟合算法等通过结合轨迹转向和速度变化率两个参数建立车辆异常行为检测模型.朱宪飞[17]基于FCM轨迹聚类算法并通过卷积神经网络学习,建立了异常轨迹识别模型.浩欢飞[18]在无监督轨迹模型上引入了增量式EM算法,提出一种增量式轨迹模型,并将该模型应用到车辆轨迹的异常识别上.

虽然研究人员提出并改进了多种聚类算法,并实现了较好效果,但值得注意的是,目前所采用的大部分轨迹数据集是高速公路等直道数据集,轨迹形状较为单一,针对城市十字交叉路口以及环形交叉路口等的轨迹聚类研究相对较少,相关聚类算法与异常值检测方法对于城市交叉路口轨迹数据集的兼容性较差.为了解决这个问题,本文提出一种基于K-Medoids算法与DBSCAN算法的组合聚类算法,通过模拟十字交叉路口数据集的训练,得到一个交叉路口的最佳聚类模型,再用全景相机所采集的真实环形交叉路口的轨迹数据集加以验证,从而为实现车辆监管、轨迹异常识别、缓解交通拥堵等应用提供理论指导.

1 轨迹相似度分析

轨迹相似度函数作为聚类算法的底层函数,可为聚类算法服务,相似度系数与轨迹间的某种距离度量密不可分.通常情况下,采用轨迹距离函数来代替相似度系数函数,也可获得较好的聚类效果.选择合理的轨迹距离函数,对于后面聚类算法的效率与准确性至关重要.一方面,轨迹距离函数的适用性会影响数据的完备性,使数据样本间的差别不够明显,导致聚类簇数不足;另一方面,轨迹距离函数的选择还会影响到聚类算法的索引结构,进而影响聚类效率.

因此,本文从时、空两个维度入手,首先通过"时间戳"将轨迹点集按照时间顺序排列并存储为元胞形式;然后根据Hausdorff距离对形状敏感的特点,求得形状差异较大的城市交叉路口车辆轨迹距离,从而得到轨迹间相似度系数;最后实现轨迹聚类的目标.

Hausdorff距离是描述两组实数点集之间相似程度的一种量度,假设有两组集合$\boldsymbol{A}=(a_1 \quad a_2 \quad \cdots \quad a_p)$,$\boldsymbol{B}=(b_1 \quad b_2 \quad \cdots \quad b_q)$,则这两个点集合之间的Hausdorff距离$H(\boldsymbol{A},\boldsymbol{B})$定义为

$$H(\boldsymbol{A},\boldsymbol{B})=\min_{a\in A}\{\min_{b\in B}\{d(\boldsymbol{a},\boldsymbol{b})\}\} \tag{1}$$

或

$$H(\boldsymbol{A},\boldsymbol{B})=\max\{h(\boldsymbol{A},\boldsymbol{B}),h(\boldsymbol{B},\boldsymbol{A})\} \tag{2}$$

其中

$$h(\boldsymbol{A},\boldsymbol{B})=\max_{a\in A}\min_{b\in B}\|\boldsymbol{a}-\boldsymbol{b}\| \tag{3}$$

$$h(\boldsymbol{B},\boldsymbol{A})=\max_{b\in B}\min_{a\in A}\|\boldsymbol{b}-\boldsymbol{a}\| \tag{4}$$

式中:$\|\cdot\|$是集合$\boldsymbol{A}$和集合$\boldsymbol{B}$间的距离范式.由于地球是一个近似的球体,计算球面上两点间的距离时,使用常用的欧几里得距离(度量)会带来较大的误差,所以本文采用球面距离公式来代替欧几里得距离以减小误差.

由式(1)、(2)可知,Hausdorff距离是一种极大极小函数,具有方向性,或是说存在不对称性,即$h(\boldsymbol{A},\boldsymbol{B})\neq h(\boldsymbol{B},\boldsymbol{A})$,所以将$h(\boldsymbol{A},\boldsymbol{B})$称为单向Hausdorff距离,将$H(\boldsymbol{A},\boldsymbol{B})$称为双向Hausdorff距离.下文若无特殊说明,Hausdorff距离均指的是双向Hausdorff距离$H(\boldsymbol{A},\boldsymbol{B})$.

简而言之,Hausdorff距离是从一个集合到另一个集合中最近点的最大距离,这种距离的度量方式克服了最近距离算法(closest-pair distance,CPD)在相对位置改变时距离也会受到影响的缺点.因此,这种算法对形状敏感而被广泛应用于计算机的视觉识别.考虑到非完全直线道路(十字交叉路口、环形交叉路口等)上车辆轨迹形状差异很大,使用Hausdorff距离可以有效地识别不同方向的车辆并提供相似度系数.该算法的时间复杂度是$O(p,q)$,其中p和q分别为集合$\boldsymbol{A}$和集合$\boldsymbol{B}$中点的个数.所以,如果一条轨迹中存在较多点则会线性增加该算法的时间复杂度,降低运算效率.

2 聚类算法分析

由于轨迹聚类分析的数据集规模往往十分庞大,选择合适的聚类算法至关重要.DBSCAN算法是一种典型的基于密度的聚类算法,它将簇定义为密度相连的点的最大集合,能够把具有足够

密的区域划分为簇，并可以在有噪声的空间数据集中发现任意形状的簇. 因此该算法聚类速度快，异常值识别准确，但存在参数难以控制的缺点. K-Medoids 算法仅需输入一个参数，但算法存在随机性. 因此将两者以参数传递的方式相结合，提出一种组合聚类算法，可以较好地达到聚类与异常轨迹检测的目的.

2.1　K-Medoids 算法

设在 m 维欧式空间中有 n 个点所构成的数据点集 $\boldsymbol{X}=(\boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \cdots \quad \boldsymbol{X}_n)$，其中 $\boldsymbol{X}_i=(x_{i1} \quad x_{i2} \quad \cdots \quad x_{im})$，$i=1,2,\cdots,n$. 在这个空间中的某个范围内选取 k 个中心位置 $\boldsymbol{V}_i(i=1,2,\cdots,k)$，使这 n 个点到各自所在簇的中心位置的某种距离度量最小，这是一种基于优化思想的目标函数，一般可定义如下：

$$E=\sum_{i=1}^{k}\sum_{\boldsymbol{X}_i\in\boldsymbol{C}_i}\|\boldsymbol{X}_i-\boldsymbol{V}_i\|_m \tag{5}$$

式中：E 是所有点与其归属的簇的中心之间的偏差总和；$\boldsymbol{X}_i$ 是 $\boldsymbol{R}^m$ 空间中的点，表示给定的数据点；$\boldsymbol{V}_i$ 是簇 $\boldsymbol{C}_i$ 的中位数(设 $\boldsymbol{X}_i$ 和 $\boldsymbol{V}_i$ 都是同维度的)；$\|\boldsymbol{X}_i-\boldsymbol{V}_i\|_m$ 表示 $\boldsymbol{X}_i$ 和 $\boldsymbol{V}_i$ 之间的一种 m 阶度量. 通过求 E 的最小值使得生成的簇尽可能地紧凑且相对独立.

K-Medoids 算法用簇的中位数计算使离群噪声点对于聚类的影响减小，避免较小簇的形成，鲁棒性高. 但中心点的寻找采取了轮换的思想，需要遍历所有的点，时间复杂度大，为 $O(n^2kt)$，其中 n 是所有数据点的数目，k 是簇的数目，t 是迭代的次数. K-Medoids 算法通常使用贪婪优化过程来实现：

(1)初始化簇中心点 $\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k$

(2)重复以下步骤直到函数收敛

①确定簇

$$\boldsymbol{C}_k=\{\boldsymbol{X}_i\|\boldsymbol{X}_i-\boldsymbol{V}_k\|^2\leqslant\|\boldsymbol{X}_i-\boldsymbol{V}_\lambda\|^2\} \tag{6}$$

②更新簇的中心点

$$\boldsymbol{m}_k=\boldsymbol{X}_j=\arg\min_{\boldsymbol{X}_l\in\boldsymbol{C}_k}\frac{1}{n_k}\sum_{\boldsymbol{X}_i\in\boldsymbol{C}_k}\|\boldsymbol{X}_l-\boldsymbol{X}_i\|^2 \tag{7}$$

K-Medoids 中的 $\boldsymbol{V}_k$ 一定与数据点重合，因此 K-Medoids 聚类完全依赖于数据点之间的距离. 因为 $\boldsymbol{V}_k$ 与 $\boldsymbol{X}_j$ 重合，那么 $\|\boldsymbol{X}_i-\boldsymbol{V}_k\|^2\equiv\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2$. 在 K-Medoids 聚类过程中评估的所有距离都可以预先计算并存储在一个距离矩阵 $\boldsymbol{D}$ 中，不难看出，这是一个对角线为 0，且 $a_{i,j}=a_{j,i}$ 的大型矩阵，矩阵大小则为轨迹条数. 这样，事先计算距离矩阵可以大幅提高聚类速度，尤其当参数改变或反复聚类以求最佳聚类效果时尤为重要. 所以，式(6)和式(7)可以更新为

$$\boldsymbol{C}_k=\{i\,|\,\boldsymbol{D}_{i,\boldsymbol{v}[k]}\leqslant\boldsymbol{D}_{i,\boldsymbol{v}[\lambda]}\} \tag{8}$$

$$\boldsymbol{m}_k=\arg\min_{l\in\boldsymbol{C}[k]}\frac{1}{n_k}\sum_{i\in\boldsymbol{C}[k]}\boldsymbol{D}_{l,i} \tag{9}$$

2.2　实验分析

本文使用 LISA(Laboratory for Intelligent and Safe Automobiles，UC San Diego Datasets)轨迹数据集为基准轨迹聚类算法提供数据[19].

如图 1 所示，该数据集包含 6 个不同场景：3 条模拟公路、1 条真实公路、1 个模拟十字路口和 1 个类似环形交叉路口. 前期使用模拟数据集进行训练，后期使用多个摄像头收集的真实轨迹数据加以验证.

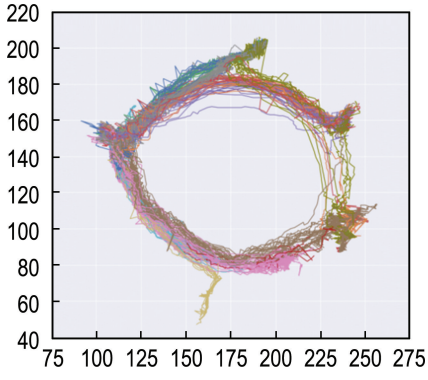


图 1　环形交叉路口 K-Medoids 轨迹聚类

Fig. 1　K-Medoids clustering trajectory at roundabout intersection

该场景有 5 个出入口，中间是一个圆形障碍物，是一个类似环形交叉路口的场景. 考虑到出入量的不同，忽略了行走人数较少的行走方案以及折返回头等规模较小的簇，所以这种场景的聚类是在 $k=12$ 时进行的. 并且通过后续的实验表明：在极端情况下(聚类超过 10 000 次)，$k=12$ 存在一个轮廓系数高达 0.76 的聚类结果，这表明，若想追求最佳聚类效果，可以选择 $k=12$ 并反复训练模型，保存最佳的一个. 结果表明，该方法能较好地对轨迹进行聚类，且各簇间的差别较大，符合人们的认知习惯，效果较好. 但是在图像右侧，算法陷入了局部最优解，这与此部分轨迹质量低、噪声点多以及 k 值的选取有直接联系.

图 2 中，在模拟十字交叉路口的场景下，不考

虑调头(U-turn)的情况发生,理论上存在 $A_4^2=12$ 种可能簇,所以选择 $k=12$ 进行 K-Medoids 聚类.事实证明,簇的聚类标准与十字交叉路口的起点和终点有关.由于存在交通规则的约束,车辆不被允许在十字交叉路口中反复徘徊,这既方便了相似度系数的计算,也为 k 值的选择提供了参考.不同的起点与终点可以被归结为不同的簇中.簇之间差异明显且规模较大,说明聚类效果很好.

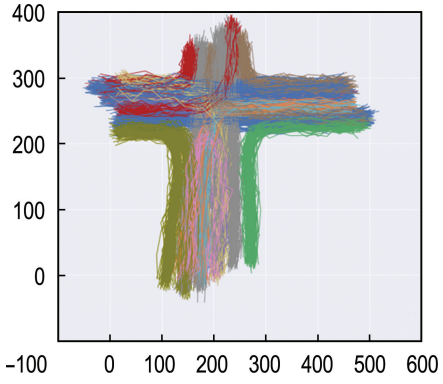


图 2 十字交叉路口 K-Medoids 轨迹聚类

Fig. 2 K-Medoids clustering trajectory at four-exit intersection

3 聚类效果分析

为了能够更科学地评估聚类效果,本文引入轮廓系数 S(silhouette coefficient)作为评价指标.它将聚类效果分为凝聚度与分离度两个方面进行评价,可以用来评价不同聚类算法对同一原始数据的聚类效果,也可以评价运行环境对于聚类效果的影响,是一种评估聚类算法效果的科学客观的手段,在现实生产应用中存在着重要的指导价值.

3.1 单点轮廓系数

首先,对于簇中的每个样本,分别计算它的轮廓系数,对于其中的一个样本 i 来说,其轮廓系数为

$$a(i)=\frac{\sum_{s\in C_n,i\neq s} dis(i,s)}{|C_n|-1} \tag{10}$$

$$b(i)=\min_{C_m:1\leqslant m\leqslant k,m\neq n}\frac{\sum_{s\in C_m} dis(i,s)}{|C_m|} \tag{11}$$

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{12}$$

式中:i 代表簇 C_n 中的对象,凝聚度 $a(i)$代表 i 与 C_n 中剩余对象的均值,分离度 $b(i)$代表 i 与 C_m 中其他点的平均距离,$S(i)$代表轮廓系数.

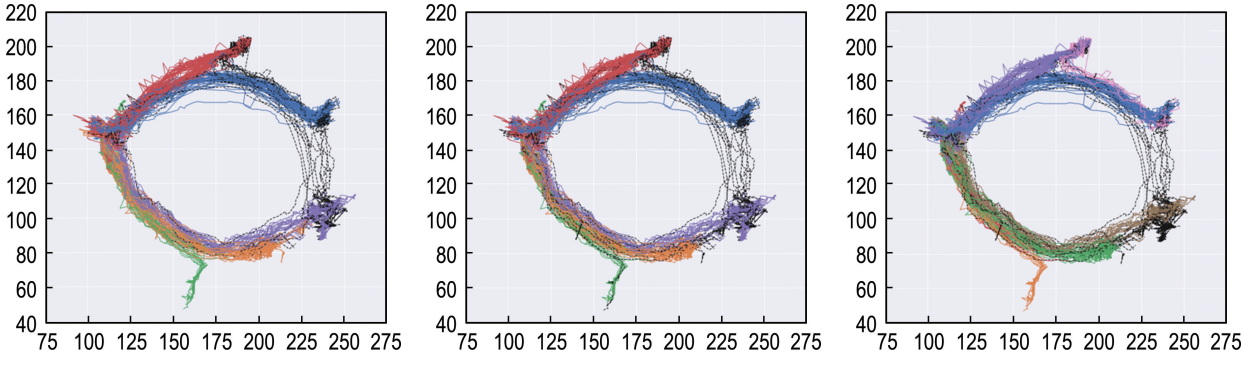
所有样本的轮廓系数的均值称为聚类效果的轮廓系数,定义为 S,是该聚类是否合理、有效的一种度量.聚类效果轮廓系数的取值在[-1,1],值越大,说明同类样本相距越近,不同样本相距越远,则聚类效果越好.

3.2 DBSCAN 算法

DBSCAN 算法对参数很敏感但却不易控制选择,略微改变参数就有可能得出完全不同的聚类效果,而参数的选择目前无规律可循,只能靠反复实验确定.聚类半径愈大,聚类的簇数就越少;最少点越多,识别出的噪声点就越多.

3.3 异常轨迹的分析

在模拟场景中使用 DBSCAN 算法对不同参数聚类效果进行模拟,图 3 为环形交叉路口的聚类效果,其中图3(a)、(b)取簇数为5,图3(c)取簇



(a) 簇数为 5,聚类半径为 150 (b) 簇数为 5,聚类半径为 100 (c) 簇数为 6,聚类半径为 80

图 3 不同参数下的 DBSCAN 异常轨迹检测

Fig. 3 DBSCAN abnormal trajectory detection under different parameters

数为 6. 表 1 为在环形交叉路口场景中使用 DBSCAN算法与 K-Medoids 算法于不同参数条件下聚类效果的比较.

表 1　DBSCAN 与 K-Medoids 参数对比

Tab. 1　Comparison of DBSCAN and K-Medoids parameters

来源	参数	簇数	轮廓系数
图 3(a)	聚类半径=150,最少点=10	5	0.706 9
图 3(b)	聚类半径=100,最少点=10	5	0.726 5
图 3(c)	聚类半径=80,最少点=5	6	0.663 1
	$k=4$	4	0.696 8
K-Medoids	$k=5$	5	0.686 0
	$k=6$	6	0.645 2

由于 K-Medoids 算法每次聚类时会在 n 个样本点中任意选取 k 个点作为中心点,结果具有一定偏差. 因此,具体操作流程如下:

(1)在总体 n 个样本点中任意选取 k 个点作为中心点.

(2)按照与中心点就近原则,将剩余的 $n-k$ 个点分配到当前最佳中心点的类中.

(3)对于第 i 个类中除对应中心点外的所有其他点,按顺序计算当前簇中所有其他点到该中心点的距离之和,选取最小距离时对应的点作为新的中心点.

(4)重复(2)、(3)的过程,直到所有的中心点不再发生变化或已达到设定的最大迭代次数.

(5)最终确定 k 个类.

本场景中,通过 K-Medoids 算法传递的函数可知,簇数应为 4 或 5. 从图 3 与表 1 看出,当簇数为 5(图 3(a)、(b))时,聚类的效果较好,异常检测较为准确,异常轨迹基本聚集在右侧;而当簇数为 6(图 3(c))时,轮廓系数直线下降,使得聚类效果大幅下降.

在十字交叉路口模拟场景下,使用 DBSCAN 算法进行模拟,观察其异常轨迹如图 4 所示.

从图 4 可以看出,轨迹异常的主要原因是异常变道,一方面是因为轨迹数据点较少,导致连接数据点形成折线;另一方面转向异常轨迹较少,说明该模拟数据集对转向轨迹部分模拟较为谨慎,异常值较少.

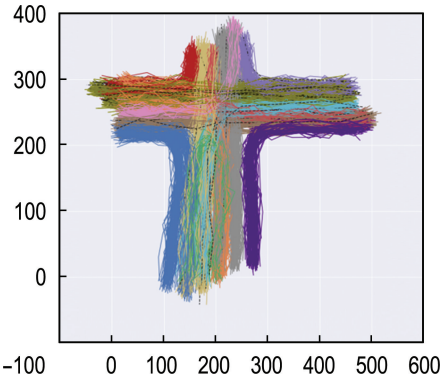


图 4　十字交叉路口异常轨迹检测

Fig. 4　Four-exit intersection abnormal trajectory detection

4　聚类算法的改进

4.1　K-Medoids 的改进

根据 K-Medoids 算法的特点,聚类初始需要对随机种子进行选择,容易导致每次的运行训练结果大不相同. 所以,统一以 1 000 次聚类为基准,测试不同值下的聚类效果. 以轮廓系数为判断依据,分别记录 K-Means 与 K-Medoids 的最大轮廓系数与 1 000 次平均轮廓系数,见表 2、3. 最大轮廓系数代表了这种算法的最好聚类效果,平均轮廓系数则表达了聚类效果的总体水平.

表 2　K-Means 轮廓系数

Tab. 2　K-Means silhouette coefficient

k	最大轮廓系数	平均轮廓系数
9	0.549 789 192	0.381 925 524
10	0.569 332 637	0.372 777 093
11	0.591 295 699	0.361 867 228
12	0.627 745 135	0.341 413 586
13	0.608 983 982	0.320 250 119
14	0.581 951 705	0.306 494 351
15	0.578 180 944	0.270 147 162

表 3　K-Medoids 轮廓系数

Tab. 3　K-Medoids silhouette coefficient

k	最大轮廓系数	平均轮廓系数
9	0.691 851 210	0.467 300 725
10	0.707 256 006	0.466 050 869
11	0.700 612 598	0.454 678 437
12	0.705 240 965	0.440 911 568
13	0.691 733 109	0.438 290 216
14	0.693 499 715	0.398 094 902
15	0.690 114 559	0.388 829 766

通过对比表 2 与表 3 的轮廓系数可知，*K*-Medoids 算法的聚类效果普遍优于 *K*-Means 算法. 同时由表 3 可知，随着 k 值的提高，最大轮廓系数基本保持在 0.7 左右，但是，平均轮廓系数却在降低，说明初始随机种子越多，聚类效果越不稳定，容易陷入局部最优解，造成聚类效果下降. 综合考虑时间复杂度以及运算成本，当 k 在 9～11 时，既节约了计算成本，同时也可以取得较好较稳定的聚类效果. 值得一提的是，在极端情况下（聚类超过 10 000 次），$k=12$ 存在一个轮廓系数高达 0.76 的聚类结果，这表明若想追求最佳聚类效果，可以选择 $k=12$ 并反复训练模型，保存最佳的一个.

4.2 DBSCAN 算法的改进

与 *K*-Means 及 *K*-Medoids 算法不同，DBSCAN算法具有较好的稳定性，实验的可重复性高，所以在获取轮廓系数的时候，并没有进行过多次的聚类，其轮廓系数如表 4 所示.

表 4 DBSCAN 轮廓系数

Tab. 4 DBSCAN silhouette coefficient

簇数	聚类半径	最少点	轮廓系数
16	250	10	0.656 902 085
16	300	10	0.665 892 783
15	350	10	0.696 143 983
14	400	10	0.715 396 944
12	450	10	0.732 157 376
13	500	10	0.735 170 657
12	550	10	0.716 214 347

关于 DBSCAN 轮廓系数各种参数的调整需要注意的是：簇数与聚类半径参数直接相关，但并不意味着改变聚类半径参数一定会改变簇数，同时两者之间大致呈负相关，但绝非标准负相关；基于预处理后的轨迹数据，通过 Alteryx 软件计算轨迹的标准差，筛选出比标准差大的轨迹点记为噪声点，最少点参数的修改只影响最终噪声点的识别与剔除，通过软件计算得出噪声点所占比例小于 1%，所以对于轮廓系数的影响可以忽略不计，故本数据未更改最少点参数，但为了数据的完整性记录下来；同一参数与原始数据多次实验，轮廓系数差别在 1×10^{-5} 量级上，故不记录平均轮廓系数，以最大轮廓系数代替.

结果表明，DBSCAN 算法的聚类效果普遍优于 *K*-Means 与 *K*-Medoids 算法，且聚类速度更快，轮廓系数随着聚类半径的增大先增后降，最佳聚类效果应当出现在聚类半径为 450～500 时，这时聚类的簇数为 12 或 13，与 *K*-Means 和 *K*-Medoids最佳聚类效果簇数相符. 两者相互验证，证明了前期假定的簇数为 12 是正确的.

综上所述，由于 *K*-Medoids 算法参数少，使用 *K*-Medoids 算法可以确定最佳聚类簇数，再将最佳聚类簇数传递给 DBSCAN 算法，使其可以确定自身的相关参数，从而达到精确聚类并识别异常轨迹的目的.

5 结 论

（1）针对交叉路口轨迹形状差异大、交叉点类型多、轨迹异常等特点，提出了一种基于 *K*-Medoids 和 DBSCAN 算法相结合的聚类算法. 分别使用模拟数据集和真实交通数据集进行训练. 最后，根据轮廓系数评价聚类效果. 结果表明，*K*-Medoids 算法由于自身的局限性，在聚类速度上比 DBSCAN 算法慢，但它仍然是目前可用的最快的聚类算法之一，最佳聚类效果更好. *K*-Medoids 算法具有随机性，需要大量重复训练才能得到最佳模型，而 DBSCAN 算法则更稳定. *K*-Medoids 只需要为集群指定一个 k 值，该方法为 DBSCAN 算法提供了参数参考，进而得到高轮廓系数的模型参数. 实验结果表明，两者的有机结合可以获得更好的聚类效果和异常轨迹识别效果.

（2）面对海量数据，*K*-Medoids 算法的运算时间随着数据量的增加而线性增加. 因此，在处理大量数据时，聚类效率不容忽视；如果样本集的密度不均匀，且聚类间距相差很大，DBSCAN 算法的聚类质量较差，这就需要对数据进行一定的规范化. 同时，该算法对高维数据的聚类效果不是很好. 对于三维数据，应进行降维变换或使用相应的高维相似度系数函数.

（3）对于轨迹异常检测，需要提高 DBSCAN 算法检测局部异常的准确性. 这些异常状态表现为由于设备测量误差导致的轨道曲率异常，但整个轨道仍在正常车道上.

（4）对于动态轨迹聚类，需要在轨迹完全驱动之前预测聚类结果，这就需要对静态轨迹进行分段聚类，并对每个类别进行可能性分析. 动态预测在实际无人驾驶中具有较高的实用价值和研究价值，本

实验可为后续的动态预测提供一些基础数据.

参考文献：

[1] 田钧方，朱陈强，贾　宁，等. 基于轨迹数据的车辆跟驰行为分析与建模综述 [J]. 交通运输系统工程与信息，2021，**21**(5)：148-159.
TIAN Junfang，ZHU Chenqiang，JIA Ning，*et al*. Review of car-following behavior analysis and modeling based on trajectory data [J]. **Journal of Transportation Systems Engineering and Information Technology**，2021，**21**(5)：148-159.（in Chinese）

[2] 张小芳，冯慧芳. 基于轨迹大数据的动态最优路径规划 [J]. 计算机与现代化，2021(11)：82-88.
ZHANG Xiaofang，FENG Huifang. Dynamic optimal path planning based on trajectory big data [J]. **Computer and Modernization**，2021(11)：82-88.（in Chinese）

[3] 牟乃夏，徐玉静，张恒才，等. 移动轨迹聚类方法研究综述 [J]. 测绘通报，2018(1)：1-7.
MOU Naixia，XU Yujing，ZHANG Hengcai，*et al*. A review of the mobile trajectory clustering methods [J]. **Bulletin of Surveying and Mapping**，2018(1)：1-7.（in Chinese）

[4] 韩　晨. 面向轨迹大数据的高效聚类算法设计与实现 [D]. 呼和浩特：内蒙古大学，2021.
HAN Chen. Design and implementation of efficient clustering algorithm for large amounts of trajectory data [D]. Hohhot：Inner Mongolia University，2021.（in Chinese）

[5] 翟俐民. 基于轨迹聚类和 LSTM 的航迹预测方法研究 [D]. 成都：四川大学，2021.
ZHAI Limin. Research on track prediction method based on trajectory clustering and LSTM [D]. Chengdu：Sichuan University，2021.（in Chinese）

[6] 张　雷. 面向 GPS 数据的轨迹聚类与异常检测算法研究 [D]. 沈阳：辽宁大学，2019.
ZHANG Lei. Trajectory clustering and anomaly detection algorithm for GPS data [D]. Shenyang：Liaoning University，2019.（in Chinese）

[7] 饶　磊，刘艳芳，罗园园，等. 基于通勤轨迹的潮汐交通拥堵路段识别与分析 [J]. 地理空间信息，2021，**19**(4)：89-96.
RAO Lei，LIU Yanfang，LUO Yuanyuan，*et al*. Identification and analysis of tidal traffic congested road sections based on commuter trajectory [J]. **Geospatial Information**，2021，**19**(4)：89-96.（in Chinese）

[8] 程云爻. 基于循环神经网络的异常轨迹检测系统的研究与实现 [D]. 北京：北京邮电大学，2021.
CHENG Yunyao. The research and implementation of anomalous trajectory detection system based on RNN [D]. Beijing：Beijing University of Posts and Telecommunications，2021.（in Chinese）

[9] 郝美薇，戴华林，郝　琨. 基于密度的 K-Means 算法在轨迹数据聚类中的优化 [J]. 计算机应用，2017，**37**(10)：2946-2951.
HAO Meiwei，DAI Hualin，HAO Kun. Optimization of density-based K-Means algorithm in trajectory data clustering [J]. **Journal of Computer Applications**，2017，**37**(10)：2946-2951.（in Chinese）

[10] 张玉西，苏小会，高广棵，等. 改进主成分和 K-均值聚类算法的行驶工况 [J]. 科学技术与工程，2021，**21**(8)：3199-3205.
ZHANG Yuxi，SU Xiaohui，GAO Guangke，*et al*. Driving conditions of a car based on improved principal component and K-Means clustering algorithm [J]. **Science Technology and Engineering**，2021，**21**(8)：3199-3205.（in Chinese）

[11] 何　月，王崇倡. 基于时空聚类的出租车载客热点区域挖掘研究 [J]. 测绘与空间地理信息，2020，**43**(1)：99-102.
HE Yue，WANG Chongchang. Research on taxi pick-up hotspots based on time and space cluster [J]. **Geomatics and Spatial Information Technology**，2020，**43**(1)：99-102.（in Chinese）

[12] 郭景华，李克强，王　进，等. 基于危险场景聚类分析的前车随机运动状态预测研究 [J]. 汽车工程，2020，**42**(7)：847-853，859.
GUO Jinghua，LI Keqiang，WANG Jin，*et al*. Study on prediction of preceding vehicle′s stochastic motion based on risk scenarios clustering analysis [J]. **Automotive Engineering**，2020，**42**(7)：847-853，859.（in Chinese）

[13] 冯琦森. 基于出租车轨迹的居民出行热点路径和区域挖掘 [D]. 重庆：重庆大学，2016.
FENG Qisen. Research on residents′ trip hot routes and attractive areas based on taxi trajectory data [D]. Chongqing：Chongqing University，2016.（in Chinese）

[14] ZHENG Linjiang，FENG Qisen，LIU Weining，*et al*. Discovering trip hot routes using large scale taxi trajectory data [C]// **Lecture Notes in Computer Science（Including Subseries Lecture Notes in**

Artificial Intelligence and Lecture Notes in Bioinformatics). Australia: Springer Verlag, 2016.

[15] 马占宇. 基于出租车轨迹数据的居民出行热点区域与路径挖掘 [D]. 郑州: 郑州大学, 2020.
MA Zhanyu. Research on mining residents' trip attractive areas and popular routes based on taxi trajectory data [D]. Zhengzhou: Zhengzhou University, 2020. (in Chinese)

[16] 蒋恩源, 王学军. 基于跟踪轨迹的车辆异常行为检测 [J]. 吉林大学学报(信息科学版), 2016, **34**(1): 98-103.
JIANG Enyuan, WANG Xuejun. Vehicle abnormal behavior detection based on trajectory tracking [J]. **Journal of Jilin University (Information Science Edition)**, 2016, **34**(1): 98-103. (in Chinese)

[17] 朱宪飞. 交通系统监控环境下车辆异常行为识别算法研究 [D]. 济南: 山东大学, 2018.
ZHU Xianfei. Study on vehicle abnormal behavior recognition algorithm under traffic supervising condition [D]. Jinan: Shandong University, 2018. (in Chinese)

[18] 浩欢飞. 增量式建模下的车辆轨迹识别与在线异常检测研究 [D]. 哈尔滨: 哈尔滨工程大学, 2014.
HAO Huanfei. The study of vehicle trajectory recognition and online anomaly detection with incremental modeling [D]. Harbin: Harbin Engineering University, 2014. (in Chinese)

[19] MORRIS B, TRIVEDI M. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation [C]// **2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009-06-20**. Piscataway: IEEE, 2009: 312-319.

Research on combinatorial clustering algorithm and anomaly detection based on improved *K*-Medoids

HE Yuhai[1,2,3], ZHOU Qingkun[1], CHENG Tansheng[1], WANG Qinpeng*[1,2,3]

(1. School of Naval Architecture, Ocean and Energy Power Engineering, Wuhan University of Technology, Wuhan 430063, China;
2. Key Laboratory of Ship Power Engineering Technology Transportation Industry, Wuhan University of Technology, Wuhan 430063, China;
3. Electronic Control Sub Laboratory of National Engineering Laboratory of Naval Architecture and Ocean Engineering Power Systems, Wuhan University of Technology, Wuhan 430063, China)

Abstract: The extraction and mining of vehicle trajectory information using clustering algorithm and anomaly detection algorithm are of great significance in applications such as traffic control and management, spatial and temporal analysis and management of road congestion, user travel route planning and recommendation, and autonomous driving decision planning. A clustering algorithm based on a combination of *K*-Medoids and DBSCAN is proposed to address the shortcomings of existing clustering algorithms and anomaly detection algorithms, which are difficult to control the parameters and have randomness. Through training on simulated four-exit intersection datasets, an optimal clustering model for intersections is obtained, and the model is validated and optimized with real trajectory datasets. Then, the trajectory clustering data flow in the intersection area over some time is reproduced visually, and the clustering effect of fewer abnormal trajectories and faster clustering is achieved, while the optimal values of each parameter of the algorithm are selected by comparison. Finally, the parameter transfer enables the DBSCAN algorithm to identify the abnormal trajectories more accurately and provide guidance for traffic management and autonomous driving decisions.

Key words: vehicle trajectory; cluster analysis; anomaly detection; similarity measure; DBSCAN algorithm