doi:10.11959/j.issn.1000-436x.2017287

基于改进密度聚类与模式信息挖掘的异常轨迹识别方法

何明[1,2]，仇功达[1]，周波[1]，柳强[1,3]，曹玉婷[1]
（1.陆军工程大学指挥控制工程学院，江苏 南京 210007；2. 解放军第 61 所，北京 100000；3. 海军指挥学院，江苏 南京 210007）

摘 要：针对社会安全事件中异常行为信息识别挖掘难等问题，提出一种基于改进密度聚类与模式信息挖掘的异常轨迹识别方法。首先，针对采样问题，结合 Hausdorff 距离思想重新定义一种改进型 DTW 距离，用于描述轨迹具体行为，而 MBR 距离下的延伸定义，则用于描述轨迹覆盖区域热度。其次，在 CFSFDP 算法的密度关联与决策模型下，基于支持向量机回归（SVR, support vector regression）提出了特定支持向量机回归（SSVR, specific support vector regression），利用针对性改良下的回归差异非线性识别类中心，实现类的智能识别。最后，通过 2 种密度下的类识别，实现更多异常模式信息的挖掘与 3 种异常轨迹识别。结合上海市与北京市出租车轨迹集进行了仿真实验与数据分析，验证了算法在轨迹聚类异常识别方面的有效性。与传统方法相比，类发现能力提高了 10%，异常轨迹信息得以区别与丰富。

关键词：支持向量机回归；密度聚类；异常轨迹识别；模式信息挖掘
中图分类号：TP301 **文献标识码**：A

Abnormal trajectory detection method based on enhanced density clustering and abnormal information mining

HE Ming[1,2], QIU Gong-da[1], ZHOU Bo[1], LIU Qiang[1,3], CAO Yu-ting[1]
(1. College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China;
2. The 61st Research Institute of PLA, Beijing 100000, China; 3. Naval Command College, Nanjing 210007, China)

Abstract: Aiming at problems of low accuracy in the recognition and difficulty in enriching the information of abnormal behavior in the social security incidents, an abnormal trajectory detection method based on enhanced density clustering and abnormal information mining was proposed. Firstly, combined with Hausdorff distance, an enhanced DTW distance aiming at the problem of sampling to describe the behavior in detail was proposed. And based on the MBR distance, some definitions to describe the geographical distribution of trajectory were proposed. Secondly, with the density-distance decision model of CFSFDP algorithm, intelligent recognition of cluster was realized by using the difference of SSVR which was proposed based on SVR. Finally, based on the analysis of distribution under the two kinds of density, more abnormal information could be mined, three kinds of abnormal trajectories would be recognized. And the simulation results on trajectory data of Shanghai and Beijing verify that the algorithm is objective and efficient. Comparing to existing method, accuracy in the clustering is promoted by 10%, and the abnormal trajectories are sorted, abnormal information is enriched.

Key words: SVR, density clustering, abnormal trajectory detection, pattern information mining

收稿日期：2017-05-12；**修回日期**：2017-11-12
基金项目：江苏省自然科学基金资助项目（No.BK20150721, No.BK20161469）；中国博士后基金资助项目（No.2015M582786, No.2016T91017）；江苏省重点研发计划基金资助项目（No.BE2015728, No.BE2016904）；江苏省科技基础设施建设计划基金资助项目（No.BM2014391）；国家重点研发计划基金资助项目（No.2016YFC0800606）
Foundation Items: The Natural Science Foundation of Jiangsu Province (No.BK20150721, No.BK20161469), China Postdoctoral Science Foundation (No.2015M582786, No.2016T91017), The Primary Research & Development Plan of Jiangsu Province (No.BE2015728, No.BE2016904), The Engineering Research Center of Jiangsu Province (No.BM2014391), The National Key Research and Development Program of China (No.2016YFC0800606)

1 引言

随着传感器网络的普及，人们对移动对象实时位置信息的获取越来越多样便捷。与此同时，重大恐怖事件和突发事件的事后处置已经愈加无法满足当前社会的迫切需求，亟待向事前异常行为的预警预测转型，实现社会安全风险感知与防控。基于时空轨迹数据的挖掘实验[1]有助于发现人类行为模式中隐藏的模式信息、行为意图，而轨迹异常检测[2]能对隐藏在众多正常模式中的异常行为进行识别，而偏僻地域轨迹、跨域流动轨迹、热门区域行为异常轨迹等的识别，对社会安全事件的预警预测具有重要意义。

目前，国内外学者在轨迹数据异常检测方面已经开展了一定的研究。例如，Zhu 等[3]将当前轨迹与历史轨迹对比，提出了一个基于轨迹孤立点分析的热门路径（TPRO）发现方法，同时将与相应时段热门路径差异较大的轨迹视为异常；Chawla 等[4]提出了一个两阶段挖掘优化框架，检测偏离历史轨迹的异常，推断导致异常出现的路径，并揭示异常产生的缘由；文献[5]提出了基于分段与分组密度聚类框架的检测算法 TRAOD，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。

广义而言，历史轨迹与聚类后类中轨迹异曲同工。概括为在某一个距离描述下聚类，计算与热门轨迹的相异程度，识别离群轨迹。但在轨迹聚类上聚类精度有限，另外，离群轨迹涵盖范围太宽泛，直接作为异常轨迹说服力不强，缺乏相关异常信息挖掘补充。

而其中的关键技术轨迹聚类识别研究的热点在于相似度计算[6]与聚类算法。研究者根据轨迹数据的特点，综合时空轨迹多维信息，定义了多种轨迹间相似性度量方法。魏龙翔等[7]提出了一种结合 Hausdorff 距离和最长公共子序列（LCSS, longest common subsequence）的轨迹分类算法，提高了轨迹分类的准确性；文献[8]与文献[9]提出了基于动态时间规整（DTW, dynamic time warping）距离的序列比较方法，用以解决时间扭曲问题，但计算量较大；此外，Chen 等[10]提出了一种新的编辑距离度量（edit distance on real sequence）方法，正态化处理解决了空间维缩放的问题和普通算法对噪声敏感的问题；另外，为了解决整体特征与局部特色的矛盾关系，文献[5]提出了一个分段与分组的轨迹聚类的框架；为了简化分段分组处理，Yuan 等[11]提出了根据实现设置的转向阈值实现拐角检测，对轨迹进行分段。

聚类算法上，不少研究者将各类优异的密度聚类算法应用在轨迹聚类中，如文献[5]在分段分组之后便是通过 DBSCAN(density-based spatial clustering of applications with noise)算法对不同相似度的轨迹实现聚类。文献[12]提出了将 OPTICS(ordering points to identify the clustering structure)算法与另一种轨迹距离概念应用到轨迹数据聚类中的 T-OPTICS 算法。其他密度聚类算法还包括 CFSFDP[13](clustering by fast search and find of density peaks)，通过密度与可达距离分离密度类峰值点，实现快速聚类；适用于多密度聚类的 Chameleon 算法[14]，既考虑了互连性，又考虑了簇间的近似度，特别是簇内部的特征，来确定最相似的子簇。

以上方法在相似度上进行了较好的改良，包括了较多属性信息，以求发现真实相似的轨迹，但对于异常轨迹识别研究较少，多属性信息有助于提取真正热点区域或相似轨迹，但对于类中边缘化轨迹的异常信息提取能力有限；以上方法对聚类算法也做了较好的应用，实现了有效热点轨迹聚类识别，但对众多轨迹数据的小类识别能力依旧有限。

2 本文异常识别方法

如图 1 所示，对异常轨迹进行进一步细分，定义了 3 种异常，针对 3 类异常轨迹，改进并延伸定义了 2 种距离描述与轨迹热度；结合 CFSFDP 密度聚类算法，对支持向量机回归算法中的经验损失进行了针对性的改写，得到特定支持向量机回归，利用 SSVR 回归差异实现其类中心识别，该基于数据集自学习的类中心非线性识别方法智能化水平高、准确性好；最后，针对边缘交界集合的跨区域核心异常轨迹、核心热度区的热门区域行为异常轨迹、离群点的偏僻无人地域藏匿异常轨迹，通过 2 种距离密度下的回归分析，实现更多异常模式信息的挖掘，在这 2 种距离基础上实现了上述 3 类异常轨迹识别。

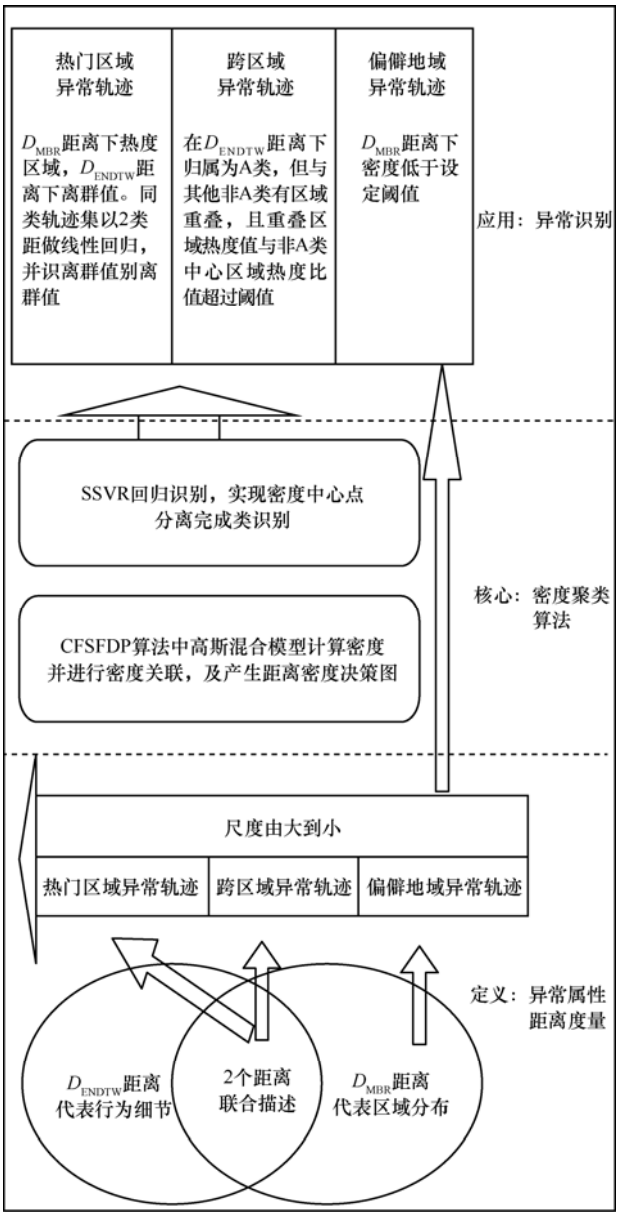


图 1　异常轨迹识别方法框架

3　基于密度类中心 SSVR 差异识别的轨迹聚类

本文在相似性描述上，针对轨迹采样频率多样多变问题，重新定义了 D_{ENDTW} 距离（改进型平均动态时间规划距离），另外在 D_{MBR} 距离（最小外包矩形距离）基础之上，以分段最小外包矩形距离定义了轨迹热度 NED_{M} 与分段轨迹所经过地域最大热度 $DEN_{\mathrm{M_{MAX}}}$。D_{ENDTW} 距离是 D_{DTW} 距离结合 Hausdorff 距离[15]针对采样频率多变进行了改良，表示的是 2 条轨迹间的实际行为具体细节相似度，包括空间位置、速度等属性，后者分段计算最小外包矩形表示轨迹覆盖区域相似度，结合 2 种距离密度下的回归分析，实现更多异常信息的挖掘。

在聚类算法的全局密度计算方面，CFSFDP 算法、DBSCAN 算法等密度聚类算法又有相同的全局密度基础，可以通过可达距计算密度并进行密度关联，也可以以高斯混合模型计算数据点间密度影响情况，得到全局密度函数。在类识别上采用 CFSFDP 密度与距离决策，结合 SSVR 实现峰值点的自学习识别，取代原先的阈值参数输入，完成类识别。

3.1　轨迹距离定义

针对轨迹数据集 $TD=\{T_1,T_2,T_3,\mathrm{L}\ ,T_i,\mathrm{L}\ ,T_{n-1},T_n\}$，其中，单条轨迹表示为 $T_i=\{tr_{i1},tr_{i2},\mathrm{L}\ ,tr_{ip-1},tr_{ip}\}$，而每一个轨迹点又由经纬度、时间以及其他附加信息组成，表示为 $tr_{i1}=(x_{i1},x_{i2},t_{i1},ex_{i1})$。在文献[16,17]中都尽可能以一种距离描述方式包含全部属性信息，本文采用 2 种轨迹相似度描述方式定义 2 条轨迹 $T_i=\{t_{i1},t_{i2},\mathrm{L}\ ,t_{ip-1},t_{ip}\}$ 与 $T_j=\{t_{j1},t_{j2},\mathrm{L}\ ,t_{jq-1},t_{jq}\}$ 间的距离。尽可能保留特征信息，联合描述异常轨迹。

最小外包矩形[18]（MBR, minimun bounding rectangle）：平行于横纵坐标轴的最小外接矩形。Elnekave 等曾采用最小外包矩形的重合部分来度量整条轨迹间的相似性。平滑了细节并缓解了噪声的影响。时间复杂度较小为 $O(n)$。距离 $D_{\mathrm{MBR}}(T_i,T_j)$定义如下。

$$D_{\mathrm{MBR}}(T_i,T_j)=\begin{cases}\infty, & MBR_j \cap MBR_j=\varnothing\\ \dfrac{MBR_j+MBR_j}{2MBR_i \cap MBR_j}, & MBR_j \cap MBR_j\neq\varnothing\end{cases}\tag{1}$$

为了进一步提升精度，以采样点为隔断，将轨迹分段构建平行于横纵坐标轴的最小外接矩形。$SMBR_i$ 表示分段累加下的最小外接矩形。再计算轨迹两两间的重合情况。记分段累加下的最小外接矩形距离为 $D_{\mathrm{SMBR}}(T_i, T_j)$。定义轨迹 T_i 在轨迹集合 cl 中的分段累加最小外接矩形下密度为 $DEN_{\mathrm{M}}(T_i, cl)$，表示为

定义 1

$$DEN_{\mathrm{M}}(T_i,cl)=\frac{1}{SMNR_i}\cdot$$

$$\sum_{j=1}^{cluster}\begin{cases}0, SMBR_i \text{ I } SMNR_j=\varnothing\\ SMBR_i \text{ I } SMNR_j, SMBR_i \text{ I } SMNR_j\neq\varnothing\end{cases} \tag{2}$$

$DEN_{\mathrm{M}}(T_i)$反映出轨迹在空间区域分布上，所处区域位置的轨迹密集程度。在分段累加最小外接矩形距离下，对全局轨迹计算轨迹密度，并对其排序，可获得轨迹相对密度情况。

定义轨迹分段最大局部外包矩形密度 $DEN_{\mathrm{M_{MAX}}}(T_i,cl)$，用于描述轨迹 T_i 所有分段中所经过的区域位置中最核心处的核心程度，如式(3)所示。

$$DEN_{\mathrm{M_{MAX}}}(T_i,cl)=\max(DEN_{\mathrm{M}}(t_{i_{1\_2}},cl,\\ DEN_{\mathrm{M}}(t_{i_{2\_3}},cl),\mathrm{L},DEN_{\mathrm{M}})) \tag{3}$$

其中，t_{i1_2} 表示(t_{i1},t_{i2})这一段轨迹。

$D_{\mathrm{DTW}}(T_i,T_j)$ 考虑了每条轨迹所有采样点信息，反映出轨迹间的整体相似情况。距离 $D_{\mathrm{DTW}}(T_i,T_j)$ 定义如下。

$$D_{\mathrm{DTW}}(T_i,T_j)=\begin{cases}0, p=q=0\\ \infty, p=0,q\neq 0\,\|\,p\neq 0,q=0\\ dist(a_i,b_j)+\min\begin{cases}D_{\mathrm{DTW}}(Rest(T_i),Rest(T_j))\\ D_{\mathrm{DTW}}(Rest(T_i),T_j)\\ D_{\mathrm{DTW}}(T_i,Rest(T_j))\end{cases}, \text{其他}\end{cases} \tag{4}$$

$dist(a_i,b_j)$表示轨迹 T_i 与 T_j 当前轮次迭代中的最末端匹配点 a_i 与 b_j 之间的点间距离，min()部分表示点 a_i 与 b_j 之前的轨迹的最小匹配代价。从定义可见，在动态规整之后，能获得 2 条轨迹采样点间最小代价下的映射关系集合 $DD=(D_{11},\mathrm{L},D_{ij},\mathrm{L},D_{pq})$，$D_{ij}$ 代表第 i 与第 j 个采样点间的映射代价。而 D_{DTW} 距离正是所有映射代价的总和。显然当采样频率较高时，能获得更多的映射点对，加和代价将会更大，因此，希望获得一个平均值代表相似度；另外，当采样间隔较为随机时，轨迹间会出现多对一的映射情况，稀疏部分缺少相应映射点，在数据采集较为艰巨的情况下，会引入不必要误差。结合 Hausdorff 距离思想：先是找到最准确的映射，再是仅保留最大的映射距离。在一对多的情况下，仅保留最小代价映射。为此在获得最小代价下的映射关系集合 $DD=(D_{11},\mathrm{L},D_{ij},\mathrm{L},D_{pq})$下，重新定义 D_{ENDTW}，表示为

$$D_{\mathrm{ENDTW}}(T_i,T_j)=\frac{1}{sum}\cdot\sum\begin{cases}\sum_{i=1}^{p}\min D_{i*}, D_{i*}\text{的数量大于1}\\ \sum_{j=1}^{q}\min D_{*j}, D_{*j}\text{的数量大于1}\\ \sum_{j=1}^{q}D_{ij}, D_{i*}\text{的数量等于1}\\ \sum_{j=1}^{q}D_{ij}, D_{*j}\text{的数量等于1}\end{cases} \tag{5}$$

其中，*表示 i 或 j 为任意值。

至此，D_{ENDTW} 距离表示的是 2 条轨迹间的实际行为具体细节相似度，包括空间位置细节；D_{SMBR} 距离表示的是轨迹覆盖区域相似度，DEN_{M} 表示轨迹所覆盖区域的轨迹热度、核心程度。$DEN_{\mathrm{M_{MAX}}}$用于描述轨迹 T_i 所有分段中所经过的区域位置中最核心处的核心程度。显然 D_{EBDTW} 距离下的核心是 D_{SMBR} 距离下的核心的充分不必要条件。例如，当一个轨迹与众多轨迹 D_{EBDTW} 距离较大时，并不能推断它是在轨迹稀疏的区域引起的边缘化，还是位于区域活动中心，由于本身行为细节异常引起的 D_{EBDTW} 距离较大。

3.2 基于高斯混合模型的密度计算

采用与 DENCLUE 算法、CFSFDP 算法等密度聚类算法类似的密度空间建立模型。鉴于数据密度的性质，2 个数据点之间密度影响的程度与距离 r 负相关。每一个数据点都有一个密度影响函数，所有数据点影响效应叠加，得到最终密度空间。针对多维数据集 $TD=\{T_1,T_2,T_3,\mathrm{L},T_i,\mathrm{L},T_{n-1},T_n\}$，其中，$T_i=\{tr_{i1},tr_{i2},\mathrm{L},tr_{ip-1},tr_{ip}\}$，计算某一个点 T_i 在其他所有点高斯混合影响下的密度，本文的影响函数采用高斯分布。某一个数据点在所有数据点影响下的密度表示为

$$\rho_{T_i}=\sum_{j=1}^{n}\mathrm{e}^{\frac{-d_{ij}^2}{2\sigma^2}} \tag{6}$$

其中，距离 d_{ij} 为 2 条轨迹间的距离，标准差σ决定了正态分布窗口的大小，文献[7]中讨论了标准差的经验取值，将所有轨迹数据进行从小到大排序，处于前 2%~10%距离作为标准差拥有较好的密度识别效果。

2.3 基于 SSVR 回归差异的密度峰值点识别

CFSFDP 算法为每一个数据点赋予 2 个属性，

分别为数据点密度 ρ_{t_i} 以及最邻近距离 δ_{ti}。最邻近距离 δ_{ti} 指距离所有大于当前点密度的数据点中最近点的距离，定义为

$$\delta_{ti}=\begin{cases}\max_j(d_{ij}),\forall j,\rho_i>\rho_j\\ \min_{j:\rho_j>\rho_i}(d_{ij}),\text{其他}\end{cases} \tag{7}$$

CFSFDP 算法结合密度分布指出，当数据点密度较大时，点间距离越小，指向邻近中密度大于该点的距离也越小。但是，当该点为局部极大值点时，需要跨类去寻找到密度大于该点的数据点，导致最邻近指向距离远大于正常该密度下应该具有的距离。CFSFDP 算法利用该性质通过人为设定阈值定性划分，将密度较大且距离较大的点分离作为类中心。

显然这种阈值线性分割方式会因为几个极显著类的存在导致决策模型上识别尺度压缩，同时密度与距离关系为非线性，这为更多的小类识别带来较大难度。

鉴于密度峰值点显著离群于正常点，且峰值点数目较少，正常点的距离与密度梯度变化关系平稳，就可以得到较为平滑的非线性分离线。通过随机采样部分数据实现正常点的回归模型构建，使所有数据基于该回归模型，直接对密度与距离关系进行判断识别峰值点。本文以 SVR 模型为基础实现非线性拟合，进行判断。当数据量较小时，由于随机采样，回归预测模型易受显著离群峰值点的影响。为此提出了 SSVR 模型。

支持向量机回归[19]可形式化表示为式(8)，第一项为结构风险，由回归线附近的支持向量决定，因此，具有一定的稀疏性；第二项为经验风险，通过对样本数据的测试，结合损失函数进行针对性的调整，能提升回归精度，但是存在过拟合的风险。

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}l_\varepsilon(f(x_i)-y_i) \tag{8}$$

因此，减小正则化常数 C，可以实现回归函数相应的泛化；针对经验风险，ε不敏感损失函数以一个宽为 2ε的容忍间隔带实现对回归线附近波动的平稳回归，因此，可适当增加ε值；此外，容忍间隔带之外的值通过损失惩戒实现回归调整，离群值的影响也同样通过此处引入，但也正因为这段损失函数，可以实现对支持向量得到的结构进行进一步拟合调整。典型的ε-不敏感损失函数 l_ε为

$$l_\varepsilon(z)=\begin{cases}0,|z|\leqslant\varepsilon\\ |z|-\varepsilon,\text{其他}\end{cases} \tag{9}$$

为使损失函数保留对损失的惩戒调整功能，同时减缓对显著离群值的响应，本文对显著离群的惩戒进行改变。改写后的分段损失函数如式(10)所示。

$$l_\varepsilon(z)=\begin{cases}0,|z|\leqslant\varepsilon\\ |z|-\varepsilon,\varepsilon<|z|\leqslant\xi\\ \xi-\varepsilon,\text{其他}\end{cases} \tag{10}$$

对于非线性回归，SVR 先通过非线性映射 $x\to\varphi(x)$，将输入空间映射到高维特征空间（Hillbert 空间），然后在特征空间中进行线性支持向量回归。

假设非线性模型为

$$\hat{f}(x,\omega)=\omega\varphi(x)+b \tag{11}$$

其中，可解得

$$\omega=\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)\varphi(x_i) \tag{12}$$

此时，由于非线性函数 φ 未知，而特征空间的维数很高（甚至 ∞），因此，ω无法显式地表达，引入核函数（kernel function），使函数回归绕过特征空间，直接在输入空间上求取，从而避免了计算非线性映射 φ。最终 SVR 可表示为

$$f(x)=\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)K(x,x_i)+b \tag{13}$$

其中，$K(x,x_i)=\varphi(x_i)\varphi(x)$，即为核函数。常用的核函数有：径向基函数、多项式函数、Sigmoid 函数、线性函数等。由于径向基核函数对应的特征空间是无穷维的，有限的样本在该特征空间中肯定是线性可分的，故本文采用径向基核函数，表示为

$$K(x,x_i)=\exp\left\{-\frac{\|x-x_i\|^2}{p^2}\right\} \tag{14}$$

另外，鉴于密度与最邻近距离决策模型中在高密度与低密度区存在较多非正常密度距离关系点，也可以通过调整采样频率，增加这 2 个区域中的数据点采样以抑制不稳定的过拟合现象。

4　3 种异常轨迹及识别

轨迹的地域分布主要可分为 3 种：核心热度区、边缘交界带以及边缘离群点。针对 3 个区域，提出了 3 种异常轨迹数据。

1) 边缘交界集合的跨区域异常轨迹。正常人的

活动常具有一定的区域性，跨域轨迹往往具有较强的目的性与非常规性，尤其是深入另一区域核心的轨迹，例如，众多犯罪轨迹多具有跨域流窜性，与原有的城市轨迹模式相异。在轨迹聚类中反映为类边缘交界集合中轨迹。

以 D_{ENDTW} 为准，结合改进 CFSFDP 算法，对全地区轨迹进行聚类，由于全地区轨迹在轨迹间分布位置上的差别远远大于轨迹形状上的差别，因此，识别的是该地区实际轨迹热度分布，并不显式地划分区域。本文将跨越进入另一活动区域的轨迹假想为探针，该探针由连续分段最小外包矩形组成，由每一分段最小外包矩形来感受各自所处位置轨迹采样点密度，保留最大值作为该轨迹跨类到访区域的核心程度，即分段最小外包矩形最大局部密度 $DEN_{\mathrm{M_{MAX}}}$。伪代码如下。

输入 时空轨迹数据集 *TD*, 阈值ω

输出 跨区域异常轨迹

1) 计算所有轨迹间 D_{ENDTW} 距离

2) 计算所有轨迹密度

3) 计算所有轨迹最小指向距离δ_{ti}，完成点间密度关联与密度决策模型

4) 基于 SSVR 识别类中心，完成聚类，$CL=\{cluster(1),\cdots,cluster(k)\}$

5) 计算每一类中心轨迹在自己类中的 $DEN_{\mathrm{M_{MAX}}}$

6) for T_i in *TD*

7) for T_j in *TD*

8) if T_i 与 T_j 不同类且 $D_{\mathrm{MBR}}(T_i, T_j)>0$

9) 计算 $DEN_{\mathrm{M_{MAX}}}(T_i, \text{cluster of } T_j)$

10) 跨域程度=

$$\frac{DEN_{\mathrm{M_{MAX}}}(T_i, \text{cluster of } T_j)}{DEN_{\mathrm{M_{MAX}}}(\text{center trajectory of } T_j, \text{cluster of } T_j)}$$

11) if 跨域程度>ω

12) 输出 T_i

13) end if

14) end if

15) end for

16) end for

2) 核心热度区的热门区域异常轨迹。正常轨迹相似度的降低主要是因为道路选择问题，但存在一类异常轨迹在区域道路上与众多轨迹高度重合，其具体行为信息却有较大差异。这种轨迹对应在人群集中区域异常人员的逗留、徘徊等异常行为。

以 D_{ENDTW} 距离完成聚类，显然 D_{ENDTW} 距离下的类核心轨迹，也一定是 D_{SMBR} 距离下的核心轨迹；而 D_{ENDTW} 距离下的类边缘轨迹，若计算分段累加最小外接矩形下的密度 $DEN_{\mathrm{M}}(T_i)$，发现达到区域核心区域密度值，则说明在轨迹覆盖地域上与频繁类存在较大重合，但是基于具体轨迹行为的聚类结果显示并不属于同一类。即认为在热点区域中的异常轨迹行为。伪代码如下。

输入 时空轨迹数据集 *TD*，阈值φ

输出 热门区域异常轨迹

1) 计算所有轨迹间 D_{ENDTW} 距离

2) 计算所有轨迹密度

3) 计算所有轨迹最小指向距离δ_{ti}，完成点间密度关联与密度决策模型

4) 基于 SSVR 识别类中心，完成聚类，$CL=\{cluster(1),\cdots,cluster(k)\}$

5) 计算每一类中心轨迹在自己类中的 DEN_{M}

6) for *cluster*(*i*) in *CL*

7) for T_i in *cluster*(*i*)

8) 计算 $DEN_{\mathrm{M}}(T_i, \text{cluster of } T_i)$

9) 行为热度= $\dfrac{\rho_{T_i}}{\rho_{\text{center trajectory of } T_i}}$

10) 区域热度=

$$\frac{DEN_{\mathrm{M}}(T_i, \text{cluster of } T_i)}{DEN_{\mathrm{M}}(\text{center trajectory of } T_i, \text{cluster of } T_i)}$$

11) end for

12) for T_i in *TD*

13) if 区域热度/行为热度>φ

14) 输出 T_i

15) end if

16) end for

17) end for

3)离群点的偏僻地域异常轨迹。指的是在及其鲜有轨迹区域，发现的个别轨迹。主要是无人区域嫌疑人员的事先谋划与事后藏匿。

由于 D_{ENDTW} 距离是 D_{MBR} 距离的充分不必要条件。因此，计算 D_{MBR} 距离下的轨迹密度值即可。伪代码如下。

输入 时空轨迹数据集 *TD*, 阈值τ

输出 偏远地区异常轨迹

1) for T_i in *TD*

2) 计算 $DEN_{\mathrm{M}}(T_i, TD)$

3) if $DEN_{\mathrm{M}}(T_i,TD)<\tau$

4) 输出 T_i

5) end if

6) end for

5 实验结果

数据集为上海市 2007 年 2 月 20 日出租车行驶轨迹 GPS 数据集。由编号、时间、经纬度坐标、瞬时行驶速度、瞬时方向、是否载客（“0”表示空车，“1”表示载客）6 个部分组成。

设置窗口时间段 7 点~9 点，以 0/1 载客信息为依据，提取该窗口时间内每一单乘客打车行为，作为一条行为轨迹。如 A 车在窗口时间内，存在 2 段载客，即提取为 2 条行为轨迹。

5.1 ENDTW 距离稳定性测试

从数据集中随机选择某 2 条近似轨迹，如图 2 所示，其中，tra1 轨迹（24 个采样点）保持不变，对 tra2 轨迹（23 个采样点）进行随机下采样，保留起始点与转折点（即第 1、6、14、18、23 个轨迹点），使轨迹形状不变。基于采样点以 ENDTW 距离与 DTW 距离计算轨迹 tra1 与轨迹 tra2 相似度，每一采样点数对应的随机采样实验 50 次。

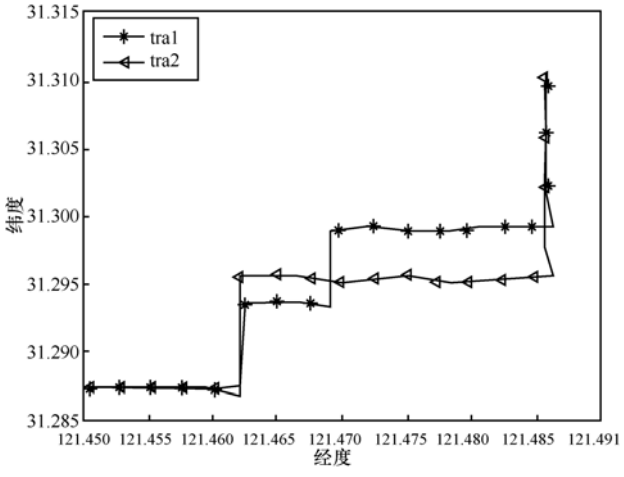


图 2　近似轨迹 tra1 与 tra2 分布示意

实验结果如表 1 所示，以统计学中的变异系数作为稳定性衡量标准，即标准差比方差。ENDTW 距离在不同采样点数下均值的变异系数为 0.173，同时 DTW 距离的变异系数为 0.380。显然，为了衡量不确定轨迹间的相似度时，即不同轨迹随机采样且采样点数不确定情况下，ENDTW 距离描述轨迹相似度比 DTW 距离更具有全局稳定性与统一性。

表 1　ENDTW 距离稳定性测试

采样点数	ENDTW		DTW	
	变异系数	均值	变异系数	均值
100%	—	4.07×10^{-6}	—	1.63×10^{-4}
80%	0.128	5.20×10^{-6}	0.111	1.98×10^{-4}
65%	0.177	5.66×10^{-6}	0.189	2.65×10^{-4}
50%	0.162	5.58×10^{-6}	0.214	3.29×10^{-4}
40%	0.212	6.68×10^{-6}	0.186	4.25×10^{-4}
均值变异系数		0.173		0.380

DTW 距离是最优匹配下的总代价，本文 ENDTW 距离是在最优匹配中删除多对一冗余后的平均值，能在一定程度上避免不同相似程度的片段因为采样点分布问题导致描述不一，由于采样为随机并未设置针对性分布，因此，两者在同一采样数下具有近似的变异系数。采样点数减少，DTW 距离反而增加是由于随机下采样使原先的匹配点缺失，必须在较远处才能找到匹配，导致单个匹配代价上升，ENDTW 距离的相应增加。若采样点数的减少按照匹对关系减少时，DTW 距离又会随着采样数的减少而减少。

5.2 跨区域型异常轨迹识别

5.2.1 区域识别

图 3(a)为实验数据集上海全地区 1 500 条轨迹分布。每条轨迹包含数 10 个采样点，每个采样点包含 6 个属性信息。图 3(b)为基于 D_{ENDTW} 距离的峰值决策，是所有轨迹数据点基于密度和最邻近指向距离的分布。人无法从高维空间获悉轨迹间分布情况，图 3(c)是基于轨迹间 D_{ENETW} 距离，将高维轨迹进行降维尺度变换成一个二维数据点后的辅助视图，降维过程会引入一定的偏差，但能大致反应轨迹间的基本分布情况，证明了全局视图下轨迹间重叠交错复杂。

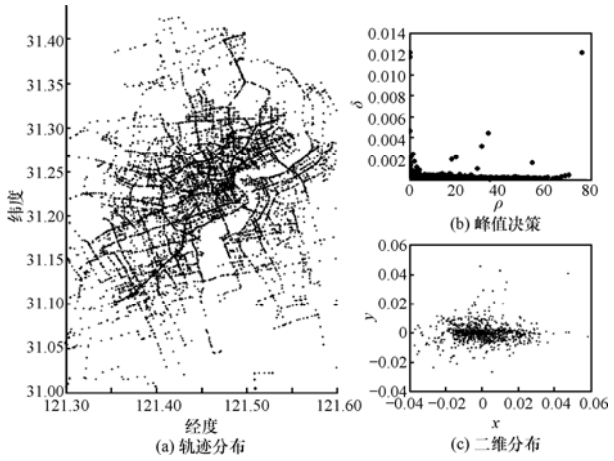


图 3　上海市出租车轨迹数据集分布

对峰值决策图进行 SSVR 回归拟合，在不同数据量随机样本下得到图 4 所示系列结果。图 4(a)和图 4(b)分别在 SVR、SSVR 下得到的回归结果。

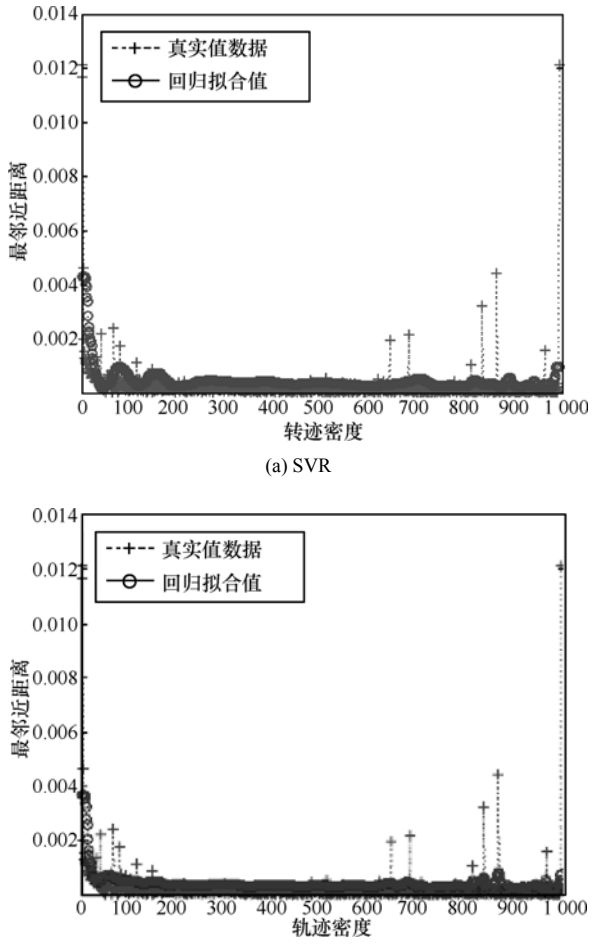


图 4　SVR 与 SSVR 在对应采样量下回归拟合抑制情况

稳定性测如图 5 所示，通过对ε不敏感损失函数的改写，以及对不同密度段数据点区别采样，有效抑制了 SVR 应对数据样本中随机混有不定量离群值的过拟合情况。另外采样间隔从 1 到 5，实则包含着从 20%到 100%数据量的跨度，因此，仅通过部分数据作为训练集进行支持向量机回归预测，具有一定的稳定性。

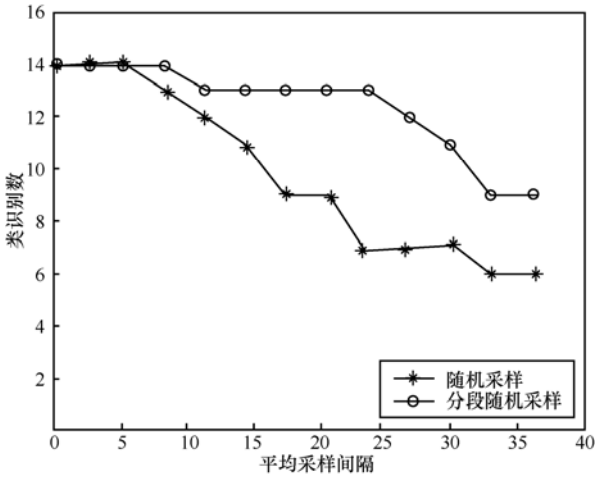


图 5　支持向量机回归稳定性测试

图 6 是本文算法分别应用于上海与北京的轨迹聚类结果，实现了对活动区域类识别。聚类结果和上海市、北京市行政区划分及实际人类活动分中心具有对应关系。也验证了当轨迹分布非常广泛时，轨迹间的地理分布差异将远远大于轨迹间的行为特征差异。但在边界处划分参考性有限，这是由于各类中心轨迹间的主要差异在于区域性，而交界处，轨迹自身行为相

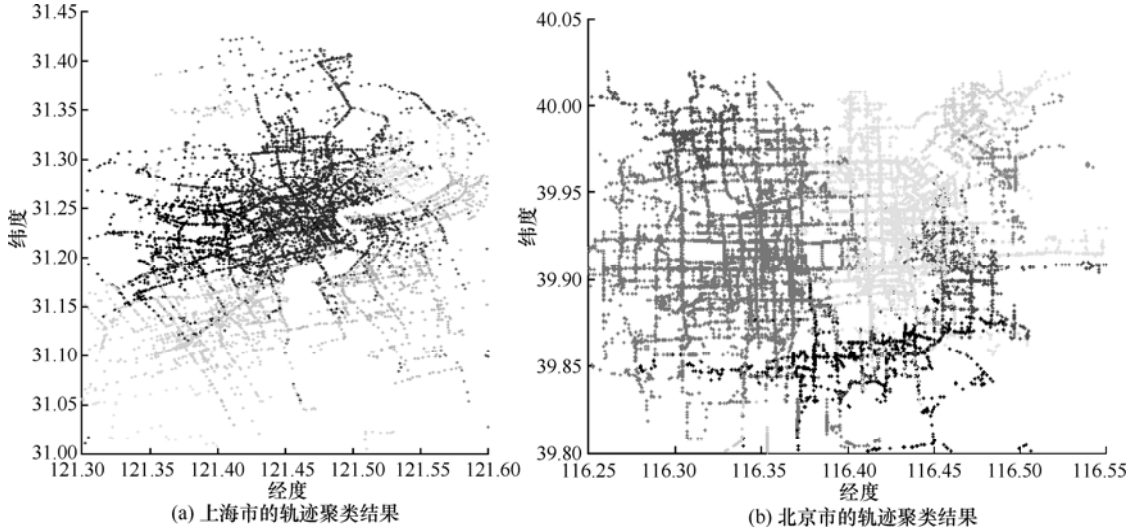


图 6　上海、北京市轨迹数据集聚类结果

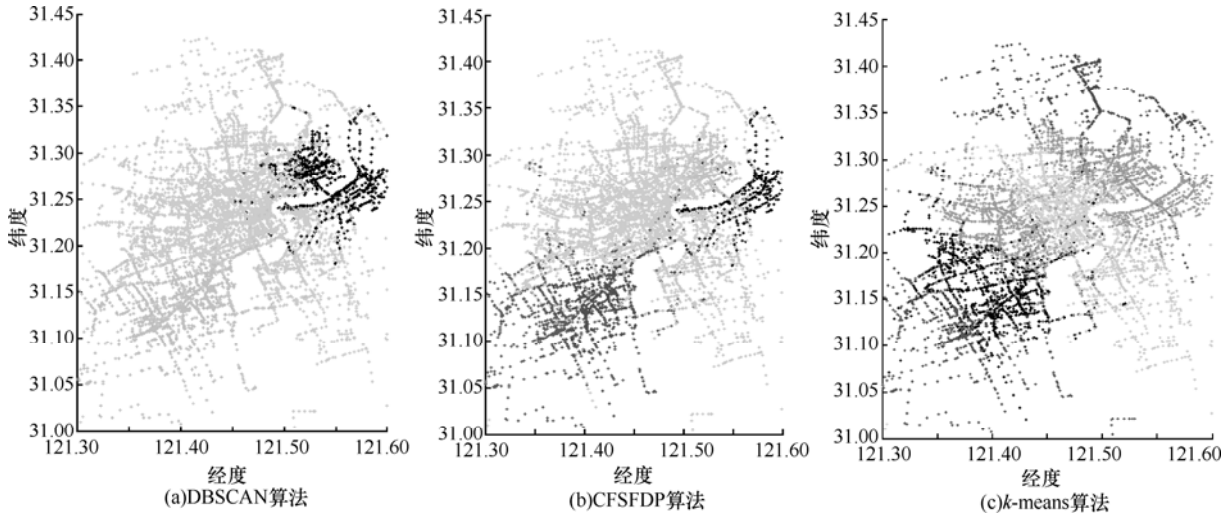


图 7　不同算法下上海市轨迹聚类结果

似性成为划分依据，因此，在交界处轨迹存在延伸。另外，由于离群点的最邻近距离存在个异性，在低密度区回归差异识别时也容易被识别为小类，虽然广义上，离群轨迹也可以被当作一个独自的类。如果仅关注热点区域，也可通过简单的类中心最低密度要求删去离群轨迹。

DBSCAN 算法、CFSFDP 算法、*k*-means 算法下的轨迹聚类结果如图 7(a)~图 7(c)所示。由于全地区轨迹下定量分析比较困难，但结合算法原理及实际聚类效果，能进行定性比较。在本节实验中聚焦到更为具体的街道、轨迹行为，将采用定量分析来对比算法优劣。

结合降维轨迹分布与聚类结果，发现 DBSCAN 算法在数据分布混叠较为厉害情况下，识别能力较弱；CFSFDP 算法在人为阈值监督下经属性关联划分，具有较好的效果，但由于线性划分仍有部分类无法识别；*k*-means 算法的基本上完成了类的识别，类的中心具有一致性，但识别布局方向性，并未体现出轨迹的密度方向延续性，也正是由于其算法限制，仅与距离相关，而距离并不具备方向性。

基于密度的聚类算法都要计算两两间的距离，因此，时间复杂度都是 $O(n^2)$。*k*-means 算法仅计算每个点到种子的距离即可，时间复杂度为 $O(n)$，具有质的优势；DBSCAN 算法通过直接增加一个点邻域 ε 中的所有点要比 CFSFDP 算法及本文算法为每一个点计算最邻近距离复杂度小。但是，DBSCAN 算法在一开始确定完邻域时，完成密度基础时便已决定了之后的聚类结果，需重新输入 ε 才能改变聚类尺度。而 CFSFDP 算法及本文算法可以将两两之间距离得出的密度，以及最邻近距离都作为基础准备，在这个意义上，CFSFDP 算法在聚类这一步复杂度仅是 $O(n)$，本文算法在回归模型建立之后复杂度也仅是 $O(n)$。

5.2.2　异常交界轨迹识别

异常交界轨迹识别的基础与重点在于类的发现能力，实验证明了本文聚类算法在全地图尺度上的类发现能力上的稳定性与优异性。在类识别基础之上，结合跨界进入的区域热度识别异常交界轨迹。在大尺度地域范围下，能识别轨迹呈区域性分布，对应人类活动具有一定的区域性，仅保留所识别 7 个主要热点区域，针对跨区域轨迹，设置跨类深入程度。当跨类深入程度设置为 90% 时，得到如图 8 所示的识别结果，深色轨迹为所识别的跨类轨迹。

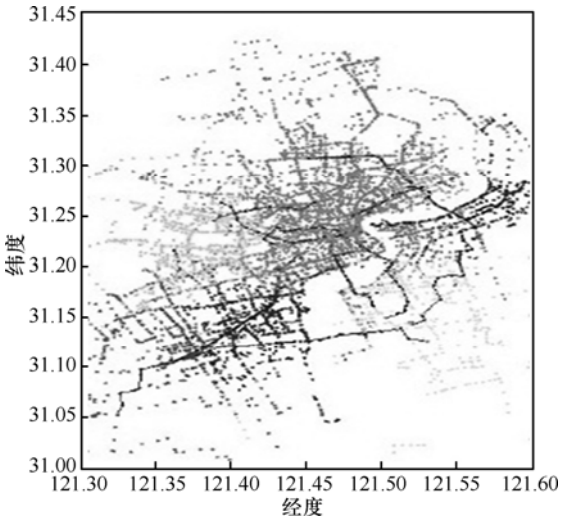


图 8　跨域异常轨迹识别结果

所识别的跨类轨迹多为长时间、大区域跨度轨迹，在 2 个类别中具有一定的跨越性。当跨类深入程度限制放宽之后，不少位于边界处本身在类划分上就具有不确定性的轨迹在这一不确定区域中被作为异常识别，显然并不具有充分的异常参考性。

5.3　高重叠区域中的行为异常轨迹识别

5.3.1　行为模式识别

为进一步挖掘轨迹细节上的属性归属情况，需缩小聚类识别区域，对小范围特定街道区域内的轨迹进行聚类与异常挖掘。在上海市 2007 年 2 月 20 日出租车行驶轨迹 GPS 数据集中通过街道经纬度限制，提取了局部指定区域街道上的 GPS 轨迹数据集，由于 DTW 算法具不可交叉对齐的性质，反向轨迹比邻近街道轨迹的差异显著，为了便于统计误差及展示，仅采用了单向轨迹。

以精度作为性能度量，即分类正确轨迹数比总轨迹数。轨迹标记的标准是以 5 条干道为 5 种中心，轨迹与哪一条干道线型重叠率最高即为该类轨迹。

图 9 是 50 条局部轨迹在本文算法与其他 3 种典型聚类算法下的识别结果。

在小尺度地图范围下，本文算法依旧出色。由于 *k*-means 聚类算法无法实现对任意类形状的识别，在数据点存在某一维度上的延伸时，将被划归为其他类。如图 9(b)所示，以本文算法的类识别数作为参数输入给 *k*-means 算法，所得结果基本上完成了类的识别，类的重心具有一致性，但是类中边缘轨迹存在识别错误。

DBSCAN 算法与 CFSFDP 算法都是基于密度关联的算法，基于相同的密度基础，具有相似的识别效果。CFSFDP 算法是对 DBSCAN 算法的改进，通过阈值设定识别类峰值点，实现更快速的聚类。在人为监督阈值设定之下，CFSFDP 算法可实现比 DBSCAN 算法更加精确丰富的类识别。当峰值点未能从从属点中显著分离时，出现邻近街道轨迹混为一类的识别情况。上海与北京 GPS 数据集局部轨迹类识别精度测试如图 10 所示。

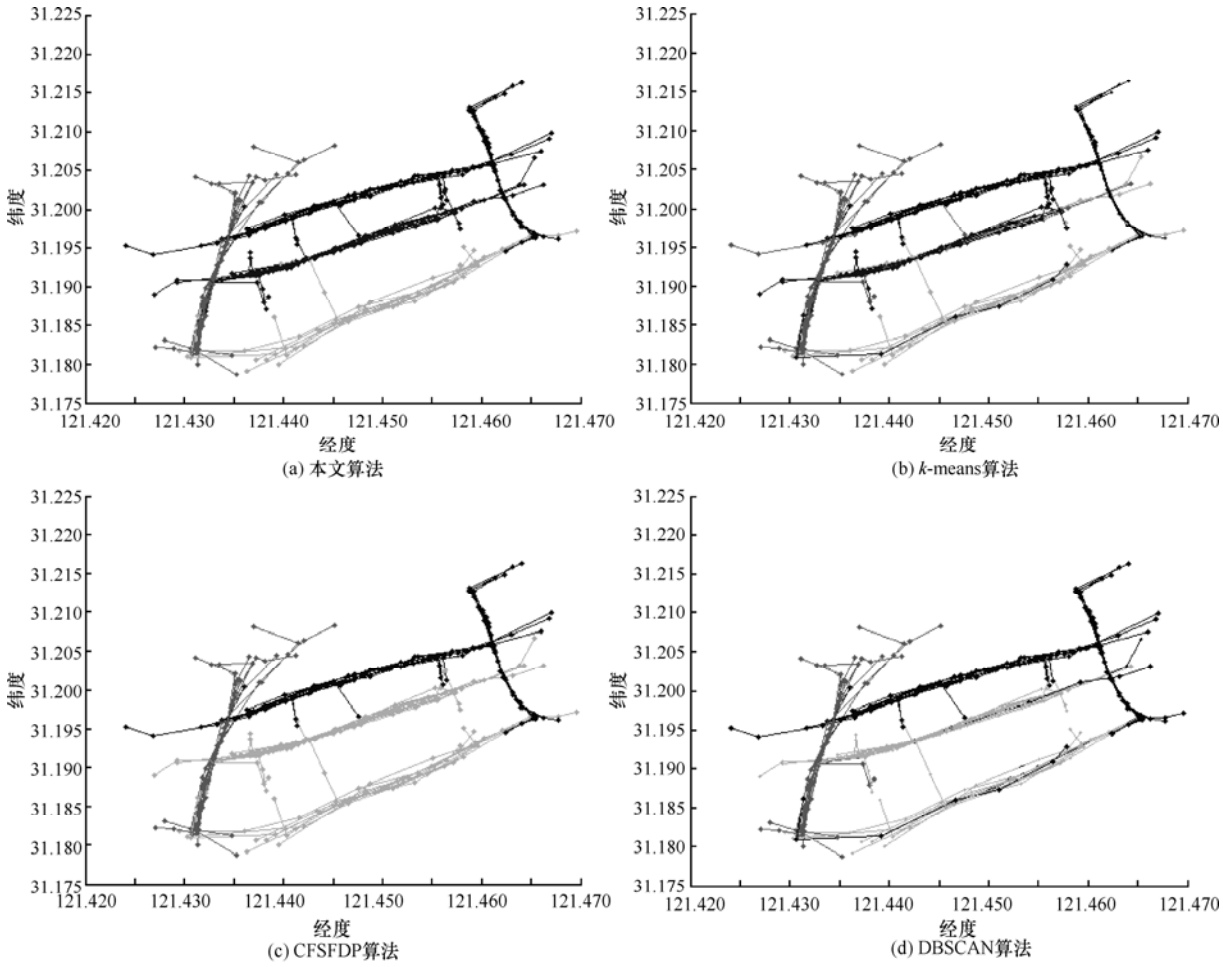
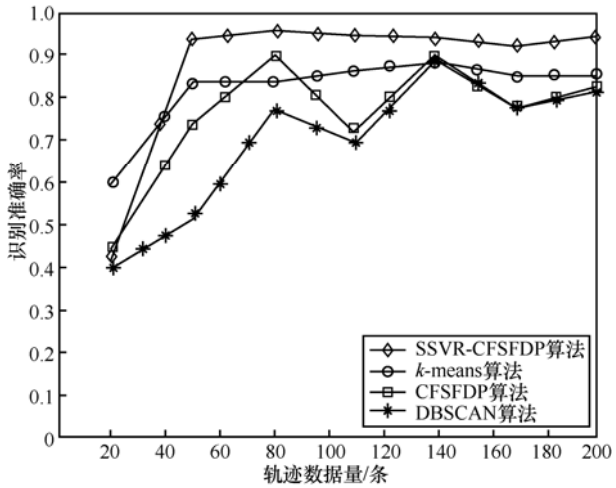
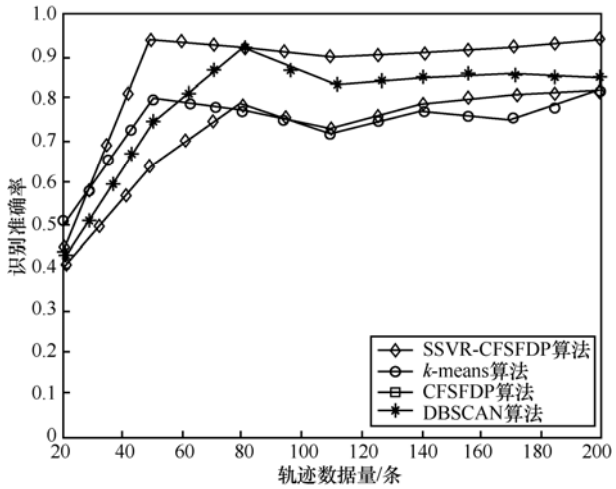


图 9　不同算法聚类结果



(a) 上海局部轨迹类识别精度测试



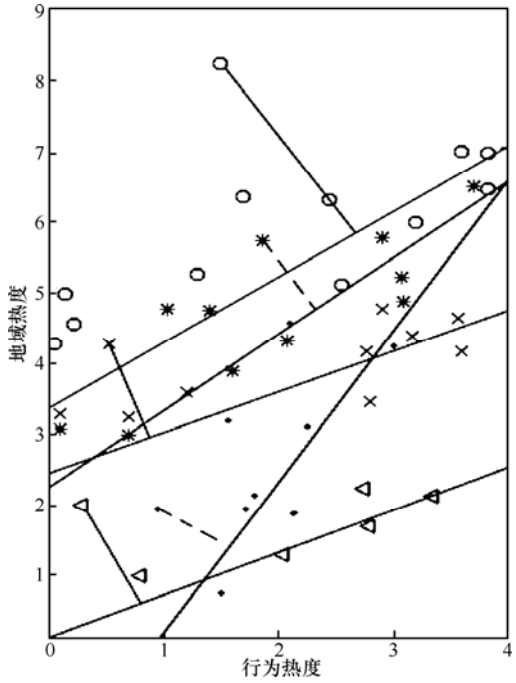
(b) 北京局部轨迹类识别精度测试

图 10　局部轨迹类识别精度测试

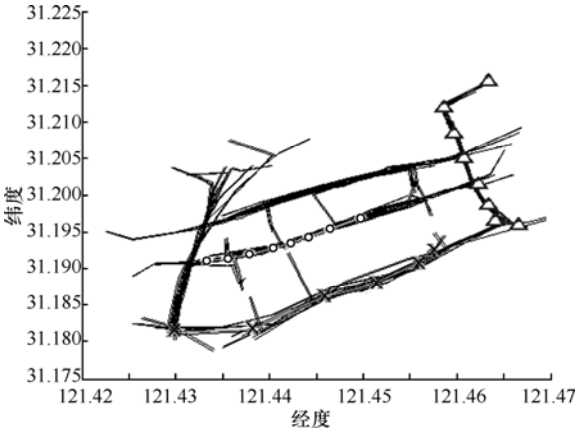
综上所述，*k*-means 聚类对于类中心附近聚类准确度较高，但是边缘处存在较大误差，DBSCAN 算法与 CFSFDP 算法能实现任意形状的聚类，但是受识别精度限制在类识别上会出现合并未能识别的情况，即图 10 中的精度突变。本文算法通过 SSVR 非线性识别 CFSFDP 算法中的密度距离决策图，智能识别了密度类中心点，相比于传统算法有 10%以上的类发现能力。

5.3.2　模式信息与异常识别

小尺度下具体异常轨迹识别同样依赖精准的类识别能力，对比实验部分证明了在应对局部轨迹时的稳定性与优异性。轨迹模式类的识别结果之上，计算每一轨迹 D_{ENDTW} 距离下密度与 DEN_{M} 密度，获得双热度下轨迹分布如图 11(a)所示，并进行回归分析以获得特殊类信息描述。“△”样点轨迹的回归线整体平均位置最低，对应该类轨迹量最少。“•”样点轨迹回归线斜率最大，说明随着行为相似度的逐渐减小，轨迹的区域重合度下降更为厉害，即对应轨迹的行为异常是由于区域异常引起。“○”样点轨迹平均位置最高，表示轨迹数据量较大，而斜率较小，说明异常来源主要是因为行为细节，而并非由轨迹涉及区域多样引起，且存在一个明显偏离回归线的轨迹。



(a) 双热度下轨迹分布



(b) 热门区域异常轨迹

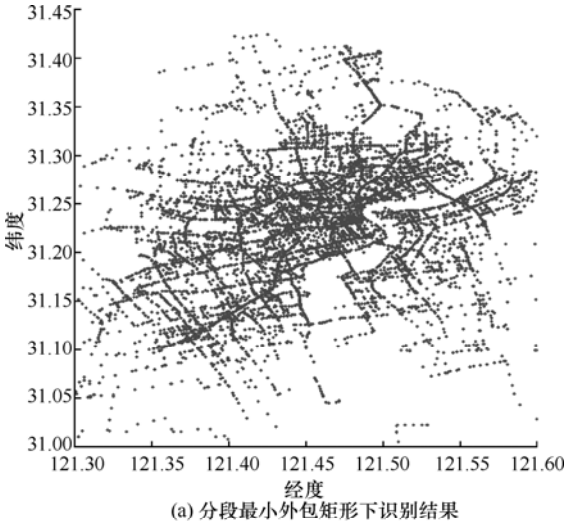
图 11　双热度下轨迹识别

结合热门活动区域中的异常行为轨迹识别可得结果如图 11(b)所示。出租车轨迹总体来说异常性较小，因此，仅保留了 2 种密度下核心度比值差异最大的 3 条轨迹。“△”样点轨迹则由于存在主干道路折返导致区域重叠率不降，而相似行为密度显著

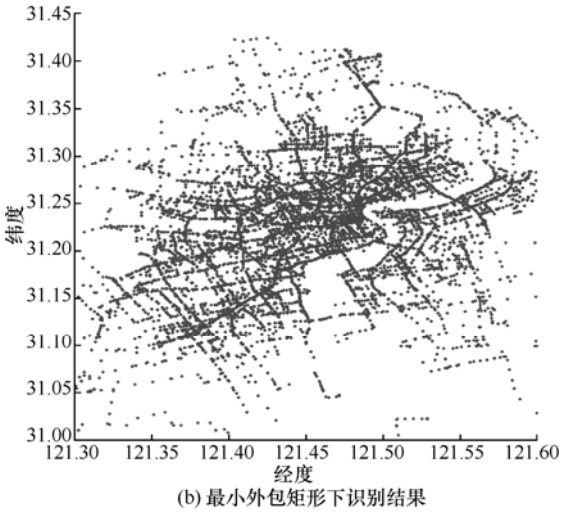
降低；“○”样点异常轨迹为在核心，“×”样点标识轨迹则是街道上缓慢行进的短轨迹，因此，区域重叠率最高。因为速度异常，导致即使沿着同一路段，在未改变区域重合度的情况下，时间距离永远存在，相似行为密度适当降低。

5.4　偏僻地域异常人员的活动轨迹识别

直接在大尺度地域范围下，仅保留重叠率 2%以下的轨迹，图 12(a)所示为分段最小外包矩形下识别结果，图 12(b)所示为最小外包矩形下识别结果。分段的好处在于能避免因轨迹转向导致的最小外包矩形代表性下降问题，有利于发现穿插在某些区域中偏僻街道上的轨迹，但分段将导致时间复杂度由 $O(n)$变为 $O(mn)$。



(a) 分段最小外包矩形下识别结果



(b) 最小外包矩形下识别结果

图 12　相同阈值下 2 种 MBR 距离下偏远地区轨迹检测

6　结束语

本文主要对 DTW 距离结合 Hausdorff 距离进行了采样问题上的改进用于描述行为细节相似度，并延伸定义最小外包矩形距离及区域热度，共同识别异常轨迹并丰富了模式信息，对于异常轨迹进一步细分为 3 种异常；将 SSVR 与 CFSFDP 相结合，以数据自身学习下的非线性阈值代替了线性阈值设定，使峰值识别更智能精确。结合上海出租车轨迹数据集验证了对 3 种异常轨迹的识别能力。并且分别对比验证了在大尺度与局部街道上，热门区域、轨迹近似模式具有有效稳定的发现能力。

今后的研究，可以继续丰富对属性的距离描述，实现对轨迹的理解，构建更加高层次的语义理解层，在识别轨迹异常的同时生成更加详细与准确的语义描述。

参考文献：

[1] ZHENG Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(3).

[2] 毛嘉莉, 金澈清, 章志刚, 等. 轨迹大数据异常检测:研究进展及系统框架[J]. 软件学报, 2017, 28 (1):17-34.
MAO J L, JIN C Q, ZHANG Z G, et al. Anomaly detection for trajectory big data: advancements and framework[J]. Journal of Software, 2017, 28(1): 17-34.

[3] ZHU J, JIANG W, LIU A, et al. Time-dependent popular routes based trajectory outlier detection[C]//WISE. 2015: 16-30.

[4] CHAWLA S, ZHENG Y, HU J. Inferring the root cause in road traffic anomalies[C]//IEEE International Conference on Data Mining. 2012: 141-150.

[5] BIRANT D, KUT A. ST-DBSCAN: an algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[6] TOOHEY K, DUCKHAM M. Trajectory similarity measures[J]. Sigspatial Special, 2015, 7(1): 43-50.

[7] 魏龙翔, 何小海, 滕奇志,等. 结合 Hausdorff 距离和最长公共子序列的轨迹分类[J]. 电子与信息学报, 2013, 35(4):784-790.
WEI L X, HE X H, TENG Q Z, et al. Trajectory classification based on Hausdorff distance and longest common subsequence[J]. Journal of Electronics and Information Technology, 2013, 35(4):784-790.

[8] YUAN G, SUN P, ZHAO J, et al. A review of moving object trajectory clustering algorithms[J]. Artificial Intelligence Review, 2016:1-22.

[9] CHEN L, ÖZSU M, ORIA V. Robust and fast similarity search for moving object trajectories[C]//The 2005ACMSIGMOD International Conference on Management of Data.2005: 491-502

[10] CHEN L, NG R T. On the marriage of Lp-norms and edit distance[C]//VLDB. 2004: 792-803.

[11] YUAN G, XIA S, ZHANG L, et al. An efficient trajectory-clustering algorithm based on an index tree[J]. Transactions of the Institute of

Measurement & Control, 2012, 34(7):850-861.

[12] NANNI M, PEDRESCHI D. Time-focused clustering of trajectories of moving objects[J]. Journal of Intelligent Information Systems, 2006, 27(3): 267-289.

[13] RODRIGUEZ A, LAIO A. Machine learning clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492-1496.

[14] SHANG F, JIAO L C, SHI J, et al. Fast density-weighted low-rank approximation spectral clustering[J]. Data Mining & Knowledge Discovery, 2011, 23(2):345-378.

[15] WANG J, WU J, WANG J, et al. Multi-criteria decision-making methods based on the Hausdorff distance of hesitant fuzzy linguistic numbers[J]. Soft Computing, 2016, 20(4): 1621-1633.

[16] WEI L X, HE X H, TENG Q Z, et al. Trajectory classification based on Hausdorff distance and longest common subsequence[J]. J Electron Inf Technol 2013,35(4):784-790

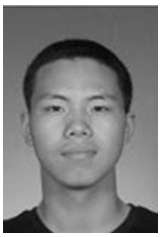
[17] KHOSHAEIN V. Trajectory clustering using a variation of Fréchet distance[D]. Ottawa, Canada: University of Ottawa, 2014.

[18] JEUNG H, YIUM L, JENSEN C S. Trajectory pattern mining[M]//Computing with Spatial Trajectories. Springer, 2011:143-177.

[19] 郭虎升, 王文剑. 动态粒度支持向量回归机[J]. 软件学报, 2013(11): 2535-2547.
GUO H S, WANG W J. Dynamical granular support vector regression machine[J]. Journal of Software, 2013(11): 2535-2547.

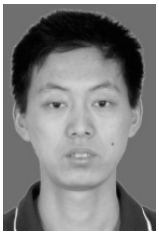
作者简介：

何明（1978-），男，新疆石河子人，博士，陆军工程大学教授，主要研究方向为传感器网络。



仇功达（1992-），男，浙江余姚人，陆军工程大学硕士生，主要研究方向为机器学习。

周波（1982-），男，江苏淮安人，陆军工程大学博士生，主要研究方向智能信息处理。



柳强（1983-），男，辽宁锦州人，博士，海军指挥学院讲师，主要研究方向指挥控制系统工程。

曹玉婷（1989-），女，江苏南京人，陆军工程大学硕士生，主要研究方向为智能信息处理。