智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于时空约束密度聚类的停留点识别方法

陆剑锋， 郭茂祖， 张昱， 赵玲玲

引用本文:
陆剑锋, 郭茂祖, 张昱, 等. 基于时空约束密度聚类的停留点识别方法[J]. 智能系统学报, 2020, 15(1): 59–66.
LU Jianfeng, GUO Maozu, ZHANG Yu, et al. Stay point recognition method based on spatio–temporal constraint density clustering[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(1): 59–66.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201910026

您可能感兴趣的其他文章

基于多粒度结构的网络表示学习
Network representation learning based on multi–granularity structure
智能系统学报. 2019, 14(6): 1233–1242 https://dx.doi.org/10.11992/tis.201905045

低秩分块矩阵的核近似
Kernel approximation of a low–rank block matrix
智能系统学报. 2019, 14(6): 1209–1216 https://dx.doi.org/10.11992/tis.201904058

公理化模糊共享近邻自适应谱聚类算法
Shared nearest neighbor adaptive spectral clustering algorithm based on axiomatic fuzzy set theory
智能系统学报. 2019, 14(5): 897–904 https://dx.doi.org/10.11992/tis.201810002

结合稀疏表示与约束传递的半监督谱聚类算法
A semi–supervised spectral clustering algorithm combined with sparse representation and constraint propagation
智能系统学报. 2018, 13(5): 855–862 https://dx.doi.org/10.11992/tis.201703013

认知视角下的舆论观点句情感计算
Research on computation of affect in public opinion sentences from the cognition viewpoint
智能系统学报. 2017, 12(4): 498–503 https://dx.doi.org/10.11992/tis.201607023

一种改进的自适应快速AF–DBSCAN聚类算法
An improved adaptive and fast AF–DBSCAN clustering algorithm
智能系统学报. 2016, 11(1): 93–98 https://dx.doi.org/10.11992.tis.201410021

微信公众平台

关注微信公众号，获取更多资讯信息

DOI: 10.11992/tis.201910026

基于时空约束密度聚类的停留点识别方法

陆剑锋[1,2], 郭茂祖[1,2], 张昱[1,3], 赵玲玲[4]

(1. 北京建筑大学 电气与信息工程学院, 北京 100044; 2. 北京建筑大学 建筑大数据智能处理方法研究北京市重点实验室, 北京 100044; 3. 北京建筑大学 深部岩土力学与地下工程国家重点实验室, 北京 100083; 4. 哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘 要: 轨迹停留点的识别是轨迹分析、出行活动语义挖掘的关键。针对基于密度聚类的停留点识别方法对时空信息的表达缺陷, 提出新的时空约束停留点识别方法, 在密度聚类中引入轨迹的间接时空特征表示, 将具有时空相似性的轨迹点进行聚合; 采用与聚类过程相统一的时空特征约束对轨迹簇进行细粒度识别。算法在进行约束的时候再次利用到聚类时候所用的输入数据特征, 特征的充分利用提高了识别的准确率。实验结果验证了本文方法的有效性。

关键词: 停留点识别; 密度聚类; 时空约束; 间接时空特征; 时空相似性; 聚合; 过程统一; 细粒度

中图分类号: TP301 **文献标志码:** A **文章编号:** 1673−4785(2020)01−0059−08

中文引用格式: 陆剑锋, 郭茂祖, 张昱, 等. 基于时空约束密度聚类的停留点识别方法 [J]. 智能系统学报, 2020, 15(1): 59–66.

英文引用格式: LU Jianfeng, GUO Maozu, ZHANG Yu, et al. Stay point recognition method based on spatio-temporal constraint density clustering[J]. CAAI transactions on intelligent systems, 2020, 15(1): 59–66.

Stay point recognition method based on spatio-temporal constraint density clustering

LU Jianfeng[1,2], GUO Maozu[1,2], ZHANG Yu[1,3], ZHAO Lingling[4]

(1. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 2. Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 3. State Key Laboratory for Geomechanics and Deep Underground Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100083, China; 4. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: The recognition of the track stay point is the key to the trajectory analysis and the semantic mining of travel activities. Aiming at the defect of spatio-temporal information based on density clustering, the new method of space-time constrained stay point recognition is proposed. In the density clustering, the indirect spatio-temporal feature representation of the trajectory is introduced, and the trajectory points with spatio-temporal similarity are aggregated. The spatio-temporal feature constraint unified with the clustering process is used to fine grain the trajectory cluster. Therefore, when the constraints are used, the input data features used in the clustering are reused, and the full utilization of the features improves accuracy of the recognition. The experimental results verify effectiveness of the proposed method.

Keywords: stay point identification; density clustering; space-time constraint; indirect spatio-temporal feature; spatio-temporal similaily; aggregatied; process uniformity; fine-grained

随着信息技术的发展以及移动通信设备等记录 GPS 轨迹数据的工具的普及, 更为细致和丰富的出行轨迹信息被大量记录下来。将人们的出行 GPS 轨迹数据进行采集、处理并根据停留点特征深入挖掘轨迹的语义信息, 是了解个体轨迹特

收稿日期: 2019−10−31.

基金项目: 国家自然科学基金项目 (61871020, 61502117, 61305013); 北京市教委科技计划重点项目 (KZ201810016019); 北京市属高校高水平创新团队建设计划项目 (IDHT20190506); 国家重点研发计划项目 (2016YFC0600901).

通信作者: 赵玲玲. E-mail: zhaoll@hit.edu.cn.

征的重要方式，能够帮助了解人们的出行需求及行为方式，发现城市居民的出行动态，发掘人的移动性规律，指导以人为本的城市交通建设，因此具有重要的研究意义。但是，GPS 轨迹数据的增多、用户的出行方式多样化以及出行目的复杂性使得对于出行轨迹数据的识别存在困难，已有的识别方法在面对当前轨迹数据的复杂性与多维性的时候越来越难以实现既定的目的。因此，需要更为准确的停留点识别方法，以进一步挖掘轨迹数据所蕴含的深层语义信息，为更为丰富的上层应用提供支撑。

目前，停留点识别的方法分为 3 类：基于聚类策略的方法、基于概率策略的方法和基于区分策略的方法。

基于聚类策略的方法方面，吕志娟[1] 提出的个人轨迹模式挖掘算法能够获得细腻的轨迹数据信息，但是这要花费更多的时间，并且还会引入更高比例的噪声。 Jiang 等[2] 提出的两步法能够有效的聚集并合成接近真实的语义数据，效率高但准确性有待提高。张文元等[3] 的算法简单高效，但不适应于密度分布不均的数据集。杨震等[4] 的方法提高了预测准确率的同时也具有良好的普适性与多步预测性能。石陆魁等[5] 提出的基于时空模式的轨迹数据聚类算法因同时兼顾了轨迹的空间和时间特征，因此在轨迹时空聚类中有更好的描述，但因为加入了时间度量使得聚类的效率有所降低。Fu [6] 等提出的两步聚类算法能够大大降低 GPS 信号丢失和数据漂移的影响，并识别独特的位置。算法的搜索速度快但不是自动的，并且需要一些先验参数。Xiang [7] 等提出的基于序列方式的聚类算法考虑了轨迹的连续性和持续时间，聚类抗噪声能力强，但仅限序列合并数据。此外，Lei Gong[8] 在 2015 年提出的算法 C-DB-SCAN 对停留点的识别准确率达到 90%，但是算法的约束条件方向变化约束存在极大缺陷，即可能适用于识别数据点频率较高的连续 GPS 轨迹数据，然而只要轨迹点停止或变化较小，算法约束条件就无法有效进行识别。Lei Gong 在 2018 年的改进算法利用到熵[10] 的原理。因此预测精度有一定提升，但是该约束也有一定缺陷，即停止的轨迹点或者移速较慢的轨迹点不一定是活动点，它们可能是等车或者是交通道路堵塞的点。

基于概率策略的方法方面，张鹏[11] 提出的数据挖掘算法能够获得一定的用户出行行为特征，但模型的网络资源利用率还需要有效提高。向隆刚等[12] 提出的核密度算法兼顾停留的识别完整性和准确性，可以有效识别复杂多样的轨迹数据，识别准确率高但抗噪声能力不强。Liao 等[13] 提出的条件随机场算法考虑轨迹的前后信息，能够获得重要的出行轨迹数据。该算法考虑的地方较多，因此预测结果会产生很多不可预估的问题。

基于区分策略方法方面，李毓瑞等[14] 提出的基于密度的停留点识别方法能够很好地找到特定的停留点，但是这种方法并不能找到所有的停留点。杜润强[15] 提出的停驻点识别方法有效的避免了常规停留点的识别错误，使得停驻点的识别更合理，但这种仿真式的识别算法在处理轨迹数据噪声上存在一定的问题。HERDER 等[16] 提出的旅游推荐算法能够获得用户的访问信息并且及时给予用户一定的推荐，但该技术并没有发展成熟。

时空语义[17-18] 为停留点的识别提供了新的思路，但是，现有的基于时空语义的轨迹点识别方法以经纬度加上时间戳直接进行聚类分段，聚类后的轨迹段虽然考虑了时间的特征，但是时间戳的差异导致轨迹段的细散化，使得轨迹段的特征不明显，不利于后续的识别。

因此，本文从停留点的识别出发，基于时空约束密度聚类的停留点识别算法对个体出行轨迹进行停留点的识别。针对方法中采用了轨迹的间接时空表示：两点间的距离和平均速度，这既保留了轨迹的时空特征，又减少了轨迹段的分散程度，能够保留停留点和移动点的特征差异。在轨迹段的识别阶段，因为考虑了轨迹点的速度和距离等特征，同时也提出了 3 种约束方法，使得轨迹点的识别更加细腻化，提高了识别性能的同时还能够挖掘更多更深层次的轨迹信息。

1 基于 DBSCAN 算法的停留点识别方法

本节对密度聚类算法 DBSCAN(density-based spatial clustering of applications with noise) 进行介绍，在 DBSCAN 算法的基础上详细说明改进的 ST_DBSCAN 算法，实验将根据 ST_DBSCAN 算法聚类得到的轨迹段进行约束与识别，最终识别停留点和移动点。

1.1 DBSCAN 算法

DBSCAN 引入了密度可达和密度相连的概念，将密度大于给定阈值的点作为核心点，所有相互可达的点作为一个聚类，不属于任何一个类的点作为噪声数据，因此可以将一个基于密度的簇看作是密度相连的点的最大集合。算法的优点在于可以识别形状复杂的聚类，不受噪声的干扰，而且聚类的结果不受数据输入顺序的影响。缺点是对于定义的参数 Eps 和 MinPts 敏感，而且

通常很难准确的确定该聚类参数, 因此不适于高维数据集。DBSCAN 定义如下:

E 邻域: 给定对象半径为 *E* 内的区域称为该对象的 *E* 邻域。

核心对象: 如果给定对象 *E* 邻域内的样本点数大于等于 MinPts, 则称该对象为核心对象。

直接密度可达: 对于样本集合 *D*, 如果样本点 *q* 在 *p* 的 *E* 邻域内, 并且 *p* 为核心对象, 那么对象 *q* 从对象 *p* 直接密度可达。

密度可达: 对于样本集合 *D*, 给定一串样本点 $p_1,p_2,\cdots,p_n,p=p_1,q=p_n$, 假如对象 p_i 从 p_{i-1} 直接密度可达, 那么对象 *q* 从对象 *p* 密度可达。

密度相连: 存在样本集合 *D* 中的一点 *o*, 如果对象 *o* 到对象 *p* 和对象 *q* 都是密度可达的, 那么 *p* 和 *q* 密度相联。

1.2　ST_DBSCAN 算法

为了表达某些时空序列的时间特征, 在 DBSCAN 算法基础上加入了时间维度, 让一个点仅考虑它设定一定时间长度内的点, 因此能够识别一些来回移动且密度比较大的簇; 输入的数据不再是经纬度转换的空间距离特征, 而是一个点到下一点的距离以及平均速度特征, 这比仅考虑点与点之间的空间距离特征, 有更好的特征选择。因此, 相对于原始的 DBSCAN 算法, 在特征的种类上有更多的选择。时空约束的聚类算法公式如下:

$$C=\sum_1^n(E_1(v,d)\&E_2(t)\&M(p))$$

式中: *C* 为改进的密度聚类算法得到的簇, E_1 为空间邻域, *v* 是输入数据速度, *d* 为输入数据距离, *v* 和 *d* 是经纬度与时间转换后得到的数据; &为逻辑门与, E_2 为时间邻域; *M* 为聚类点的最少点个数。

1.3　基于 ST_DBSCAN 算法的轨迹点聚类

在传统的轨迹点识别方面, 停留点的识别主要以经纬度加上时间戳直接进行聚类分段, 聚类后的簇虽然考虑了时间的特征, 但是时间戳的差异导致轨迹段的细散化, 使得轨迹段的特征不明显, 不利于后续的识别。因此采用了轨迹的间接时空表示: 两点间的距离和平均速度, 这既保留了轨迹的时空特征, 又减少了簇的分散程度, 所以能够保留停留点和移动点的特征差异。图 1 根据 ST_DBSCAN 算法对轨迹点进行聚类分段, 最后利用约束条件识别出停留点和移动点。

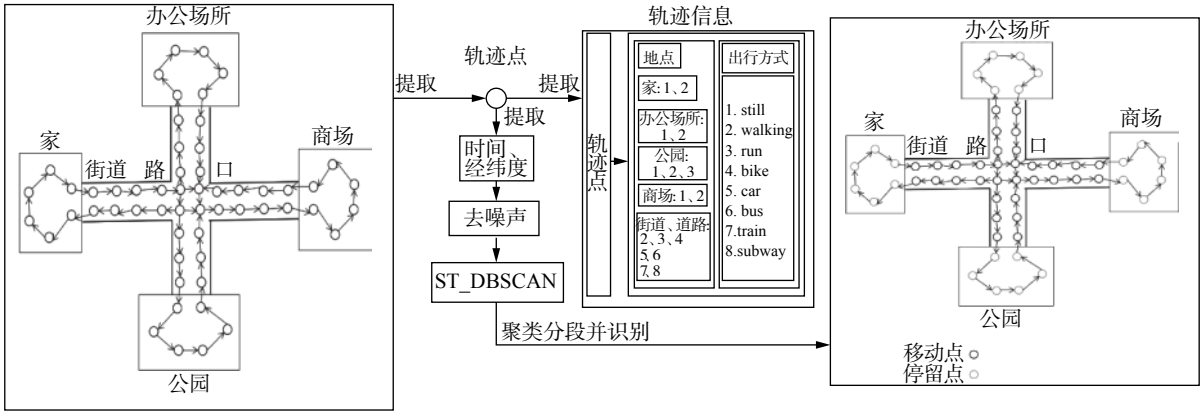

图 1　基于 ST_DBSCAN 算法的停留点识别方法流程

Fig. 1　ST_DBSCAN algorithm clusters and identifies track points

ST_DBSCAN 聚类算法描述

输入　GPS 轨迹数据集 *D*, 空间领域约束 E_1, 时间领域约束 E_2, 聚类点最少点个数 *M*。其中: $D=\{T,v,d\}=\{(t_1,t_2,\cdots,t_n),(v_1,v_2,\cdots,v_n),(d_1,d_2,\cdots,d_n)\}$, *T* 是时间戳, *v* 是速度, *d* 是距离。

输出　聚类后得到簇 *C* 与噪声 *N*。其中:

$$C=\{C_1,C_2,\cdots,C_n\}=\left\{\begin{array}{c}[(t_1,t_2,\cdots,t_i),(v_1,v_2,\cdots,v_i),(d_1,d_2,\cdots,d_i)],\\ [(t_i,\cdots,t_j),(v_i,\cdots,v_j),(d_i,\cdots,d_j)],\\ \vdots\\ [(t_k,\cdots,t_n),(v_k,\cdots,v_n),(d_k,\cdots,d_n)]\end{array}\right\}$$

并且 $i<j<k<n$。C_n 是聚类得到的具有相似特征的簇。

$$N=\{N_1,N_2,\cdots,N_l,\cdots,N_m,\cdots,N_n\}$$

其中 $l<m<n$, 噪声 *N* 程碎片化, 并且噪声之间的特征不明显, 所以噪声 *N* 不参与下一步识别。

1.4　ST_DBSCAN 算法约束条件

根据上一部分 ST_DBSCAN 聚类算法得到的簇 C, 利用恰当的约束条件来识别停留点和移动点。

约束算法如下:

输入　簇 *C*。其中, $C=\{C_1,C_2,\cdots,C_n\}$。

输出　停留点序列 P_s 和移动点序列 P_m。其中 $P_s=\{P_{s1},P_{s2},\cdots,P_{sn}\}$, $P_m=\{P_{m1},P_{m2},\cdots,P_{mn}\}$。

约束算法:

1)判断各簇 C_1~C_n 是否含有静止的点, 没有

则进行：

2）如果簇 C 的所有速度特征 V_1 的和的平均值 v_{avg}=4 m/s 大于设定阈值则标记为移动点，不符合的进行约束 1 和约束 2：

约束 1 设置一个速度阈值 V_2=0.6 m/s，T_2 为在簇中轨迹点速度特征大于 V_2 所占的百分比。

约束 2 设置一个速度 V_3=1.5 m/s 的阈值，在约束 1 的基础上，并且簇的平均速度小于 V_3 的标记为移动点，其他的标记为停留点。

3）有静止的点，则进行约束 3：

约束 3 计算簇里面速度为 0 的点的百分比 D(若百分比 D 比接近 1 说明它是越有可能是交通堵塞的点) 以及如果簇的 D 符合并且上一个簇是移动点，同时上一个簇的点的移动速度大于 V_4=6 m/s 则标记为移动点。

4）否则标记为停留点。

1.5 时间复杂度

本文算法聚类阶段的时间复杂度是 $O(n\times i)$，n 是点的个数，i 是找出 Eps 领域中的点所需要的时间，最大的时间复杂度是 $O(n^2)$；约束阶段的时间复杂度是 $O(m\times j)$，m 是轨迹段个数，j 是找出停留点和移动点的时间，最大的时间复杂度是 $O(m^2)$。因此本文算法的时间复杂度为 $O(n^2+m^2)$，算法最大时间复杂度为 $O(2n^2)$。

2 实验结果与分析

为了验证本文提出的基于 ST_DBSCAN 聚类算法与约束条件的停留点识别性能，在 Windows7 系统下利用 Python3.6 版本软件编写算法和约束条件。实验共测试 3 种聚类算法：ST_DBSCAN 算法、C_DBSCAN 算法[8] 和 DBSCAN_TE 算法[9]。每种方法的输入数据相同，采用均方根误差 (RMSE) 来度量算法精度，公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(X_{\mathrm{real},i}-X_{\mathrm{pre},i})^2}$$

实验分 3 步进行，数据预处理，参数选择以及评估 ST_DBSCAN 算法的性能。

2.1 数据预处理

实验数据来源于 Ubicomp 2018-SHL 挑战赛的华为数据集 (Sussex-Huawei locomotion dataset)[19]，用户的出行轨迹可视化图如图 2。

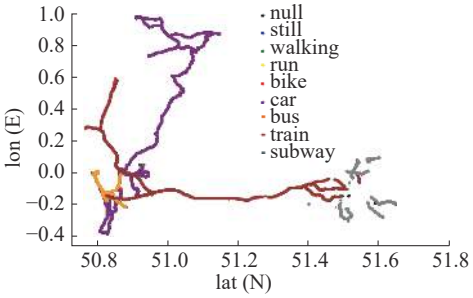

图 2 用户出行轨迹可视化

Fig. 2 Visualization of travel trajectories

表 1 给出了数据的原始标签类型，本文根据原始标签的特点进行数据的再次标注，给出 2 种数据标签：停留点和移动点以及相应的依据。

表 1 数据的原始标签及新标签标注

Table 1 Original label and new label annotation of data

轨迹点类型		数据(停留点定义为0, 移动点为1)	特点
移动点	非活动移动点	Bike:1 Car:1 Bus:1 Train:1 Subway:1	两点之间的距离相差较远，两点之间的速度大，点与点之间的密度稀疏；若有严重交通堵塞状态，则轨迹点的状态可能处于停止状态
	移动停留点	Car:1(Heavy traffic) Bus:1(Heavy traffic)	介于非活动移动点之间，点之间的密度小部分集中
介于移动点和停留点之间的点	移动速度较小且有规律的点	Run:0 Walking;Outside:1 Walking;Inside:0	若接下来是移动停留点或是非活动移动点，那么大概率是移动点；若不是则是停留点
停留点	活动移动点	Still;Stand;Outside：0 Still;Stand;Inside：0	点的密度集中，但无规律
	非移动停留点	Still;Sit;Outside：0 Still;Sit;Inside: 0	点的密度集中且两点之间的速度与距离非常小

此外, 由于数据带有一定噪声, 例如经纬度的漂移, 部分轨迹点的标签有一定的缺失。因此, 对数据进行过滤噪声[20] 处理, 如表 2。

表 2　数据标签

Table 2　Data labels

行动方式	停留点为0; 移动点为1	速度/(m·s^{-1})
Still, walking intside	0	0~1.5
Walking outside	1	0~1.5
run	0	1.5~4
bike	1	4~6
car	1	0或6~25
bus	1	0或6~25
train	1	6~40
subway	1	6~40

此部分的数据标签的特征是按照正常情况下设定的, 例如, car 与 bus 因为交通拥堵而导致的停留将会出现速度为 0 的点, 即用户在交通道路上等待的部分将在轨迹段识别里的约束部分进行识别。

2.2　ST_DBSCAN 算法中的参数选择

基于时空约束的密度聚类算法中, 需要选择合适的参数。空间阈值 eps1 与 MinPts 要与对比算法保持一致, 所以参数选择与对比算法一样。时间阈值 eps2 由于考虑的是连续轨迹点临近的关系, 如图 3, 在 eps2=5 的时候识别率都达到饱和 (user 是用户, user1、user2、user3 是 3 个用户, user1/1 是第 1 个用户第 1 天记录的数据集, user3/3 是第 3 个用户第 3 天记录的数据集), 因而经过调试选取的时间邻域值 eps2 为 5。*D* 是判断一个簇中轨迹点速度为零的百分比, 百分比越高越能够说明这段簇是因为等车或堵车所造成的停留, 如图 4, *D* 在达到 0.2 之后识别率达到饱和, 说明数据集里面因为等车或堵车所造成的停留并没有对整体识别率产生决定性的影响, 此外, 还有可能是数据集里面根本没有等车或堵车所造成的停留。因而经过一段时间的调试之后, 得到比较优秀的结果的参数是空间域 eps1=4, 时间域 eps2=5, MinPts=4, *D*=0.2。因为实验包含 9 组轨迹数据集, 此外, 由于 still、walking outside 和 walking inside 是本文识别轨迹点的算法所关注的重点, 因而对于不同的轨迹数据集来说参数 t_2 是不一样的, 同时精确度变化也是最大的。

如图 5, 每份数据的识别精度在参数 t_2 的调整下变化都特别大而且识别精度变化差异特别的明显, 这说明每个轨迹数据文件因为自身的数据特征不一样, 因而最佳参数也不一样。因此, 本文在识别轨迹数据的时候, 选用的是各自轨迹数据的最佳参数, 而记录数据的时候是记录最优参数所识别的精度。如图 6, 在与其他算法对比的时候将选用最优的参数识别的精度进行对比。

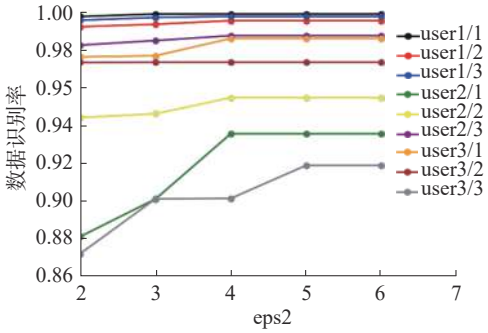

图 3　数据识别率与时间邻域 eps2 的关系曲线

Fig. 3　Relationship between data recognition rate and time neighborhood eps2

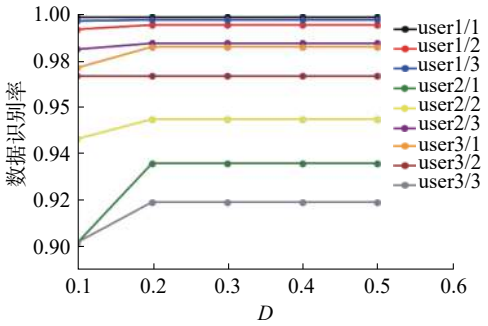

图 4　数据识别率与参数 *D* 的关系曲线

Fig. 4　Relationship between data recognition rate and parameter *D*

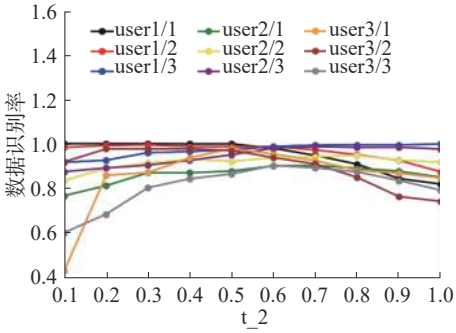

图 5　数据识别率与参数 t_2 的关系曲线

Fig. 5　Relationship between data recognition rate and parameter t_2

2.3　ST-DBSCAN 算法性能评估

为了评估本文方法的性能, 选取了同样基于聚类策略的 C_DBSCAN 算法[8] 以及 DBSCAN_TE

算法[9]。此外，本文算法的创新点在于聚类输入数据的不同，为了验证这一策略的有效性，在ST_DBSCAN 算法的基础上，将在聚类阶段输入的数据改为经度和纬度，并记为 ST_DBSCAN_GPS；在聚类阶段输入的数据为速度和距离的记为 ST_DBSCAN。本文将 ST_DBSCAN _ GPS 也作为对比方法，以验证本文算法的性能。为了达到相同的测试条件，本文在进行识别之前对 3 种算法进行了相同的预处理，在最优参数与最优准确率选择方面，3 个算法的思路是一样的，实验结果均选取了 3 种方法识别停留点的最优准确率。实验结果如表 3、图 7 和图 8 所示。

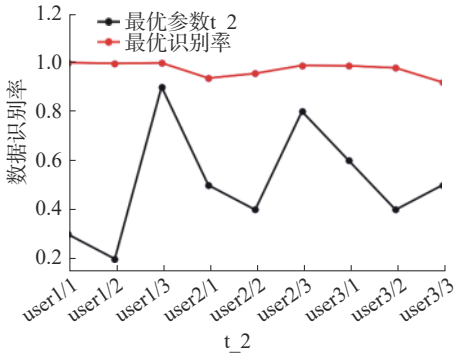

图 6　数据的最优识别参数 t_2 与最优识别率

Fig. 6　optimal identification parameter t_2 and optimal identification rate of data

表 3　3 种方法的识别准确率对比

Table 3　Identification of the three methods for optimal accuracy

用户	收集时间	轨道点个数	识别精度/%				运行时间/s			
			方法1	方法2	方法3	方法4	方法1	方法2	方法3	方法4
User1	Day1	57 988	86.08	86.21	98.63	99.87	105.71	87.82	84.11	81.68
	Day2	69 277	94.97	95.49	94.3	99.52	193.08	110.99	106.08	102.44
	Day3	44 147	86.76	92.68	99.70	99.74	48.00	46.72	48.27	46.82
User2	Day1	90 615	77.99	84.59	90.92	93.56	231.59	179.35	179.02	174.33
	Day2	50 456	76.46	86.40	93.03	95.46	63.81	59.16	59.75	59.85
	Day3	83 689	81.12	86.76	93.16	98.73	179.23	165.58	166.9	156.81
User3	Day1	23 133	85.71	96.60	98.58	98.58	27.54	23.06	21.63	21.96
	Day2	82 284	70.59	85.33	96.09	97.33	214.43	167.94	160.77	149.98
	Day3	55 913	73.48	85.57	91.88	91.89	106.64	78.29	70.86	72.35

注：方法1代表C_DBSCAN；方法2代表DBSCAN_TE；方法3代表ST_DBSCAN_GPS；方法4代表ST_DBSCAN。

图 7 是 3 种算法的识别精度对比。由图 7 可知，在本文所列举的数据集上，ST_DBSCAN 算法基本上是最优的，ST_DBSCAN 算法识别精度明显优于 DBSCAN_TE 算法和 C_DBSCAN。与 DBSCAN_TE 和 C_DBSCAN 算法相比在所有用户的每条轨迹的停留点识别精度上均有大幅提升。其原因在于本文的 ST_DBSCAN 算法的约束针对数据集进行了设计，因而该算法有一定的针对性，所以识别精度很高。因此，在新的数据集上的测试任务需要进行一定的参数调整，以适应新的轨迹数据，但 3 种约束的设计思想保持不变，仅需对约束的参数进行一定的调整。此外，由于每个数据的特征不一样，因而识别误差也有所不同。C_DBSCAN 算法的均方根误差为 19.79%，DBSCAN_TE 算法的均方根误差为 12.00%，ST_DBSCAN_GPS 的均方根误差为 5.73%，ST_DBSCAN 的均方根误差为 3.93%。因此，ST_DBSCAN 算法对于停留点的识别是有效的。

由图 8 可知，ST_DBSCAN 的运行时间相对其他两种算法和 ST_DBSCAN_GPS 运行时间最少，同时，ST_DBSCAN 的识别率显著优于 ST_DBSCAN_GPS，说明了本文所提供的以速度与距离为聚类条件的可行性。

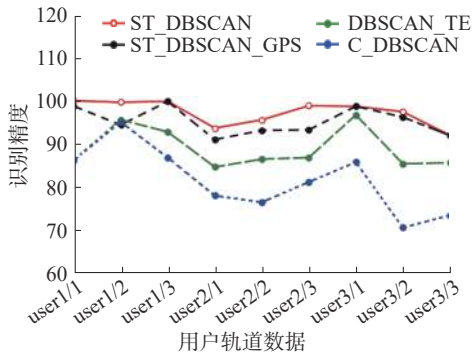

图 7　3 种算法的识别精度对比

Fig. 7　Comparison of recognition accuracy of three algorithms

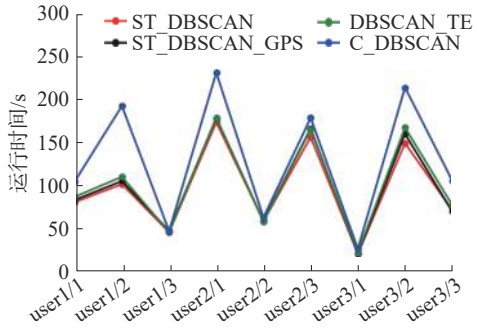

图 8　3 种算法的运行时间对比

Fig. 8　Comparison of the running time of the three algorithms

但是, ST_DBSCAN 方法仍存在一定的识别错误, 其原因在于: 部分停留点和移动点的簇的特征特别相似。如人在特定状态时的不移动 (属于停留点) 与等车、等红绿灯和堵车 (属于移动点) 的特征非常相近, 因而在建立约束条件的时候无法有效识别。

3　结束语

本文采用基于时空约束的密度聚类算法对 GPS 轨迹数据进行识别, 首先对数据轨迹进行聚类分段, 然后利用约束条件来获得停留点和移动点。实验结果表明本文的算法是有效的。本文还有许多需要提高的地方, 比如在识别非常相似的簇时需要获得更有识别率的特征; 在提高算法的泛化能力方面, 可进一步优化识别方法的约束条件, 减少约束参数。同时, 实验结果也表明轨迹数据特征的不足, 以后需要加入更多维度的特征, 此外, 约束条件还会考虑其他的特征, 如加速度、方向角和方向变化率等, 这些为本文的未来研究方向。

参考文献:

[1] 吕志娟. 基于 Lifelog 数据的个人轨迹模式挖掘算法的研究与应用 [D]. 沈阳: 东北大学, 2015: 20–30.
LYU Zhijuan. Research and application of personal trajectory pattern mining algorithm based on Lifelog data[D]. Shenyang: Northeastern University, 2015: 20–30.

[2] JIANG Renhe, ZHAO Jing, DONG Tingting, et al. A density-based approach for mining movement patterns from semantic trajectories[C]//TENCON 2015-2015 IEEE Region 10 Conference. Macao, China, 2015.

[3] 张文元, 谈国新, 朱相舟. 停留点空间聚类在景区热点分析中的应用 [J]. 计算机工程与应用, 2018, 54(4): 263–270.
ZHANG Wenyuan, TAN Guoxin, ZHU Xiangzhou. Application of stay points spatial clustering in hot scenic spots analysis[J]. Computer engineering and applications, 2018, 54(4): 263–270.

[4] 杨震, 王红军. 基于 Adaboost-Markov 模型的移动用户位置预测方法 [J]. 计算机应用, 2019, 39(3): 675–680.
YANG Zhen, WANG Hongjun. Location prediction method of mobile user based on Adaboost-Markov model[J]. Journal of computer applications, 2019, 39(3): 675–680.

[5] 石陆魁, 张延茹, 张欣. 基于时空模式的轨迹数据聚类算法 [J]. 计算机应用, 2017, 37(3): 854–859, 895.
SHI Lukui, ZHANG Yanru, ZHANG Xin. Trajectory data clustering algorithm based on spatiotemporal pattern[J]. Journal of computer applications, 2017, 37(3): 854–859, 895.

[6] FU Zhongliang, TIAN Zongshun, XU Yanqing, et al. A two-step clustering approach to extract locations from individual GPS trajectory data[J]. ISPRS international journal of geo-information, 2016, 5(10): 166.

[7] XIANG Longgang, GAO Meng, WU Tao. Extracting stops from noisy trajectories: a sequence oriented clustering approach[J]. ISPRS international journal of geo-information, 2016, 5(3): 29.

[8] GONG Lei, SATO H, YAMAMOTO T, et al. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines[J]. Journal of modern transportation, 2015, 23(3): 202–213.

[9] GONG Lei, YAMAMOTO T, MORIKAWA T. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines[J]. Transportation research procedia, 2018, 32: 146–154.

[10] GINGERICH K, MAOH H, ANDERSON W. Classifying the purpose of stopped truck events: an application of entropy to GPS data[J]. Transportation research part C: emerging technologies, 2016, 64: 17–27.

[11] 张鹏. 基于蜂窝网络数据的用户移动性分析和兴趣区挖掘 [D]. 北京: 北京邮电大学, 2018: 31–52.
ZHANG Peng. User mobility research and interest region mining based on cellular networks traffic[D]. Beijing: Beijing University of Posts and Telecommunications, 2018: 31–52.

[12] 向隆刚, 邵晓天. 载体轨迹停留信息提取的核密度法及其可视化 [J]. 测绘学报, 2016, 45(9): 1122–1131.
XIANG Longgang, SHAO Xiaotian. Visualization and extraction of trajectory stops based on kernel-density[J]. Acta geodaetica et cartographica sinica, 2016, 45(9): 1122–1131.

[13] LIAO Lin, FOX D, KAUTZ H. Extracting places and activities from GPS traces using hierarchical conditional random fields[J]. International journal of robotics research, 2007, 26(1): 119–134.

[14] 李毓瑞, 陈红梅, 王丽珍, 等. 基于密度的停留点识别方

法 [J]. 大数据, 2018, 4(5): 80–93.
LI Yurui, CHEN Hongmei, WANG Lizhen, et al. Stay point identification based on density[J]. Big data research, 2018, 4(5): 80–93.
[15] 杜润强. 基于手机轨迹数据的用户出行及停驻点识别系统研究 [D]. 北京: 北京工业大学, 2014: 35–45.
DU Runqiang. The research on users' movements and stay point identification based on mobile data[D]. Beijing: Beijing University of Technology, 2014: 35–45.
[16] HERDER E, SIEHNDEL P, KAWASE R. Predicting user locations and trajectories[M]//DIMITROVA V, KUFLIK T, CHIN D, et al. User Modeling, Adaptation, and Personalization. Cham: Springer, 2014: 86–97.
[17] DAMIANI M L, GÜTING R H. Semantic trajectories and beyond[C]//Proceedings of 2014 IEEE 15th International Conference on Mobile Data Management. Brisbane, Australia, 2014.
[18] KHOSHAHVAL S, FARNAGHI M, TALEAI M. Spatio-temporal pattern mining on trajectory data using ARM[J]. International archives of the photogrammetry, remote sensing and spatial information sciences, 2017, XL Ⅱ-4/W4: 395–399.
[19] Sussex-Huawei locomotion dataset[EB/OL]. http://www.shl-dataset.org/download/.
[20] ZHENG Yu. Trajectory data mining: an overview[J]. ACM transactions on intelligent systems and technology, 2015, 6(3): 29–36.

作者简介：

陆剑锋，硕士研究生，主要研究方向为机器学习、智慧城市。

郭茂祖，教授，博士生导师，主要研究方向为机器学习、智慧城市、生物信息学。主持和参与国家自然科学基金面上项目、北京市属高校高水平创新团队建设计划项目和北京市教委科技计划重点项目等。曾获得教育部高等学校科学研究优秀成果自然科学二等奖、省科技进步二等奖等。发表学术论文 200 余篇。

赵玲玲，讲师，博士，主要研究方向为城市计算、生物信息学。主持和参与多项国家自然科学基金项目。发展学术论文 50 余篇。

新书介绍：Mobile Information Service for Networks

本书从信息技术角度，按照网络分层的架构，针对如信息服务组网节点发现的不确定问题、传输的不稳定问题、设备间潜藏关系隐蔽性问题等，梳理当前网络移动信息服务的关键支撑方法，给出典型网络移动信息服务应用的场景及相关技术。

网络移动信息服务是指基于通信网络平台，通过各种移动设备，以无线接入网络的方式，可以通过网络发布的，为人类提供各种各样服务的平台无关的功能实体。其具体表现形式为人们所使用的各种基于移动端应用提供的便捷服务，譬如路线规划服务，移动支付服务等。

人类生活的方方面面已离不开这些信息服务，在享受这些信息服务带来的便捷、高效、智能等好处的背后，相对于有线网络环境，当前无线移动网络环境的动态性、不确定性增强，这给整个的移动信息服务方法及技术理论体系带来了从网络链路层到应用层的诸多挑战。本书从信息技术角度，按照网络分层的架构，针对如信息服务组网节点发现的不确定问题、传输的不稳定问题、设备间潜藏关系隐蔽性问题等，梳理当前网络移动信息服务的关键支撑方法，给出典型网络移动信息服务应用的场景及相关技术。

本书作者为同济大学蒋昌俊教授和东华大学李重副教授，由 Springer 出版社进行出版，全球发行，共 8 章。本书不仅适合信息技术领域的研究生和相关研究人员参考，而且适合网络移动信息服务领域的相关研究人员阅读。希望本书的出版，能够进一步推动现有网络移动信息服务技术的发展。