引文格式:王培晓,张恒才,王海波,等.ST-CFSFDP:快速搜索密度峰值的时空聚类算法[J].测绘学报,2019,48(11):1380-1390. DOI:10.11947/j.AGCS.2019.20180538.
WANG Peixiao, ZHANG Hengcai, WANG Haibo, et al. Spatial-temporal clustering by fast search and find of density peaks[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(11): 1380-1390. DOI: 10.11947/j.AGCS.2019.20180538.

ST-CFSFDP:快速搜索密度峰值的时空聚类算法

王培晓[1,3],张恒才[2,3],王海波[4],吴 升[1,3]
1. 福州大学数字中国研究院(福建),福建 福州 350002; 2. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室,北京 100101; 3. 海西政务大数据应用协同创新中心,福建 福州 350002; 4. 湖北工业大学经济与管理学院,湖北 武汉 430068

Spatial-temporal clustering by fast search and find of density peaks

WANG Peixiao[1,3], ZHANG Hengcai[2,3], WANG Haibo[4], WU Sheng[1,3]
1. The Academy of Digital China, Fuzhou University, Fuzhou 350002, China; 2. State Key Lab of Resources and Environmental Information System, IGSNRR, CAS, Beijing 100101, China; 3. Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350002, China; 4. Economic and management school, Hubei University of Technology, Wuhan 430068, China

Abstract: Spatial-temporal clustering algorithm is the basic research topic of geographic spatial-temporal big data mining. In view of the problem that traditional CFSFDP clustering algorithm cannot be applied in spatio-temporal data mining, this paper proposes a spatio-temporal constraint algorithm called ST-CFSFDP (spatial-temporal clustering by fast search and find of density peaks). ST-CFSFDP adds time constraint on the basis of CFSFDP algorithm, and modifies the calculation strategy of sample attribute value, which not only solves the problem of multi-density peak of single cluster set in the original algorithm, but also can distinguish and identify clusters at the same location and at different times. In this paper, the simulated spatiotemporal data and real indoor location trajectory data were used for the experiment, the results show that the ST-CFSFDP algorithm has a recognition rate of 82.4% at a time threshold of 90 s and a distance threshold of 5 m, which is better than the classic ST-DBCSAN, ST-OPTICS and ST-AGNES algorithm increased by 5.2%, 4.2%, and 7.6%, respectively.

Key words: geospatial-temporal big data mining; CFSFDP algorithm; ST-CFSFDP algorithm; spatio-temporal clustering
Foundation support: The National Key Research and Development Program of China(No. 2017YFB0503500); The Digital Fujian Program(No. 2016-23); The National Natural Science Foundation of China(No. 41771436)

摘 要:时空聚类算法是地理时空大数据挖掘的基础研究命题。针对传统CFSFDP聚类算法无法应用于时空数据挖掘的问题,本文提出一种时空约束的ST-CFSFDP(spatial-temporal clustering by fast search and find of density peaks)算法。在CFSFDP算法基础上加入时间约束,修改了样本属性值的计算策略,不仅解决了原算法单簇集多密度峰值问题,且可以区分并识别相同位置不同时间的簇集。本文利用模拟时空数据与真实的室内定位轨迹数据进行对比试验。结果表明,该算法在时间阈值90 s、距离阈值5 m的识别正确率高达82.4%,较经典ST-DBCSAN、ST-OPTICS及ST-AGNES聚类算法准确率分别提高了5.2%、4.2%和7.6%。

关键词:地理时空大数据挖掘;CFSFDP算法;ST-CFSFDP算法;时空聚类算法
中图分类号:P208 **文献标识码**:A **文章编号**:1001-1595(2019)11-1380-11
基金项目:国家重点研发计划(2017YFB0503500);数字福建建设项目(闽发改网数字函〔2016〕23号);国家自然科学基金(41771436)

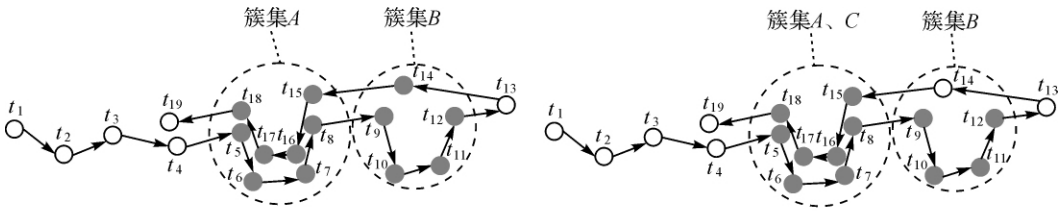
近年来，室内外定位技术迅猛发展，如北斗、GPS、WiFi、蓝牙、UWB(ultra-wideband)等，海量移动终端不断普及，如 PDA、智能手机、平板电脑等，网络技术不断进步，室内外位置服务应用不断增多，如在线导航、基于位置的社交网络、基于位置的广告推送、商业物流调度与管理等，时空轨迹数据爆发式增长[1-2]，为人类出行模式及人类移动性、城市计算、社会计算、交通管理、城市规划、人口流动监测、公共安全保障等研究提供了重要支撑[3-4]。

聚类算法是发现时空模式、探寻移动规律的关键技术之一，旨在将实体划分为一系列具有一定分布模式的簇集，同一簇集中的实体具有较大的相似度，不同簇集中的实体具有较大差别[5-8]，已广泛应用于犯罪热点分析[9]、地震空间分布模式挖掘[10]、制图自动综合[11]、遥感影像处理[12]、公共设施选址[13]、地价评估[14]、用户停留区域识别[15]等研究。

目前，聚类算法大致分为[6,16]：基于划分、基于层次、基于密度、基于格网的聚类算法。基于划分聚类算法使用聚类中心(位于簇集中心附近的一个对象)来表示每个簇集，如经典的 k-means[17]算法和 K-Medoid[18]算法，具有计算逻辑简单、运算效率高等优点；基于层次聚类算法使用逐步合并或分裂的策略来实现聚类，如改进的 CURE[19]算法，不仅能发现球形的空间簇集[20-21]，也能够发现较为复杂结构的空间簇集，但过多的超参数增加了算法的不确定性；基于密度聚类算法，如 DBSCAN[22]将簇定义为密度相连的点的最大集合，但由于采用固定阈值聚类，难以适应空间实体密度的变化[23]，OPTICS[22,24]算法为聚类分析生成一个增广簇排序，解决了 DBSCAN 算法全局密度阈值的缺点；基于格网的聚类算法[25]其聚类效果依赖于网格的大小，如果网格划分过细，则时间复杂度会显著增加，如果网格划分过粗，则聚类质量难以得到保证。

此外，许多研究学者在经典聚类算法基础之上，提出了时空聚类算法，文献[26—27]在基于密度聚类的基础上提出了时空密度聚类 ST-DBSCAN 和 ST-OPTICS 算法，在时空密度聚类中，使用时间距离将空间邻域扩展到时空邻域，从而寻找时空邻域下密度相连的区域。文献[10，28]采用时空距离(时间距离与空间距离的函数)度量任意两个时空事件的差异，后将时空距离应用到传统的聚类算法中挖掘时空事件的模式；文献[29—30]在空间扫描统计的基础上提出了时空扫描统计算法，通过滑动扫描窗口(空间半径和时间间隔定义的圆柱体)揭示时空事件随时间聚类模式；文献[31]通过窗口距离与时空最近邻的概念重新定义了时空密度，从而提出了 STSNN 算法；文献[32]从统计学的视角提出了 WNN 算法，将时空邻域中的实体分为特征与噪声两个时空泊松点过程，使用密度相连(density-connective)的概念将特征分为多个簇集；文献[33—34]分别在基于划分和基于层次聚类基础上提出了 WKM 和 ST-AGNES 算法，将其分别应用于人类行为和用户停留区域识别中，并取得了较好的聚类效果。

经典的 CFSFDP[35](clustering by fast search and find of density peaks)算法结合了基于划分和基于密度聚类算法的优点，不仅可以发现多密度任意形状的簇集，还可以通过决策图辅助识别聚类中心的个数。但该算法无法很好地应用于时空数据聚类研究之中，究其原因在于 CFSFDP 算法未考虑时间约束，如图 1 所示，并没有正确识别簇集。如错误地将簇集 $A(t_5,t_6,t_7,t_8)$、$C(t_{15},t_{16},t_{17},t_{18})$合并为一个簇集，且并不能发现簇集之间的顺序关系。此外，在单簇集中存在多个密度峰值点时，该算法将会产生错误的时空聚类结果。



(a) 不考虑时间约束的聚类结果　　(b) 考虑时间约束的聚类结果

图 1　是否考虑时间约束的两种聚类结果

Fig.1　Comparison of clustering results with and without time constraints

鉴于此,本文提出一种快速搜索密度峰值的时空聚类算法(spatial-temporal clustering by fast search and find of density peaks, ST-CFSFDP),在CFSFDP算法的基础上引入了时间约束,修改了样本属性值的计算策略。采用模拟数据和真实数据案例进行算法有效性的验证。模拟数据试验结果表明,与CFSFDP算法相比,ST-CFSFDP算法不仅可以克服单簇集中可能存在多密度峰值的不足,且可以区分并识别相同位置不同时间的簇集。以移动对象轨迹中停留点识别为例,ST-CFSFDP算法较经典的ST-DBCSAN、ST-OPTICS和ST-AGNES算法识别正确率分别可提高5.2%、4.2%和7.6%。

1 ST-CFSFDP算法

1.1 CFSFDP算法

CFSFDP算法的核心思想在于聚类中心的计算。聚类中心同时具有两个特点:①局部密度最大,被不超过自身密度的邻居包围;②与局部密度更大的样本点之间的距离相对较大。为识别数据集中$X=\{x_i\}_{i=1}^N$的聚类中心,针对每一个样本点x_i定义了局部密度ρ_i和距离δ_i两个量。其计算方法如式(1)和式(2)所示

$$\rho_i=\sum\nolimits_j \chi(d_{ij}-d_c) \tag{1}$$

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij}) \tag{2}$$

式中,d_{ij}为样本点i和样本点j之间的空间距离;d_c是人为指定的超参数,称为截断距离;$\chi(x)$表示0~1函数,当$x<0$时,$\chi(x)=1$,当$x\geqslant 0$时,$\chi(x)=0$,即局部密度ρ_i表示落在以样本点x_i为圆心,截断距离d_c为半径的圆形区域中样本点个数;δ_i表示局部密度大于x_i的所有样本点中,距离x_i最近的样本点与x_i之间的距离,由于式(2)无法计算密度最大的样本点,因此密度最大的样本点$\delta_i=\max(d_{ij})$。

CFSFDP算法通过各样本点的局部密度ρ_i和距离δ_i确定聚类中心(ρ_i和δ_i均相对较大的样本点)。如图2(a)所示,对于数据集X中每一个样本点x_i,计算其局部密度ρ_i和距离δ_i,生成相应的三元组$(\rho_i,\delta_i,\lambda_i)$,$\lambda_i$表示$\delta_i$对应的节点编号,用于非聚类中心样本点的类别分配,使用ρ_i和δ_i构造数据集X的决策图(decision graph),如图2(b)所示。容易发现,标号为①和⑩的样本点同时具有较大的ρ值和δ值,因此被确定为数据集X的两个聚类中心。为了更加方便地确定数据集X中的聚类中心,文献[35]提出了一个综合考虑ρ和δ的量γ。

$$\gamma_i=\rho_i\delta_i \tag{3}$$

式中,ρ_i表示样本点x_i的局部密度;δ_i表示样本点x_i的距离。显然γ值越大,越有可能是聚类中心,因此将$\{\gamma_i\}_{i=1}^N$进行降序排列,如图3所示。可见非聚类中心的γ值比较平滑,并且非聚类中心过渡到聚类中心存在一个明显的跳跃,此跳跃可以辅助识别聚类中心的个数。聚类中心找到后,剩余样本点根据局部密度降序依次归属到距离其最近邻的且拥有高密度值样本点所在的簇集中,比如非聚类中心样本点中密度最大的是③号样本点,因此③号首先被分配类别,首先获取③号样本点的三元组$(\rho_i,\delta_i,\lambda_i)$,将$x_{\lambda_i}$的类别分配给③号。

非聚类中心点经分配后都将具有类别,为识别数据集X中的离群点,CFSFDP算法为每一个簇集C_k引入了平均密度上界ρ_k^b的概念。ρ_k^b的定义如式(4)所示

$$\rho_k^b=\max_{\substack{d_{ij}<d_c\\ x_i\in C_k \text{ or } x_j\in C_k}}\left(\frac{\rho_i+\rho_j}{2}\right) \tag{4}$$

式中,C_k表示聚类结果后的第k个簇集;x_i表示数据集X中的第i个样本点,x_i与x_j有且仅有一个样本点属于簇集C_k;ρ_i表示样本点x_i的局部密度;d_c表示截断距离;d_{ij}表示样本点x_i与样本点x_j的空间距离。依据平均密度上界ρ_k^b将C_k内部的样本点分为核心部分(cluster core)和边缘部分(cluster halo),C_k内部样本点的局部密度大于ρ_k^b称为核心部分,小于ρ_k^b称为边缘部分,边缘部分即为数据集X中的离群点。

1.2 基于时空约束的ST-CFSFDP聚类算法

ST-CFSFDP算法首先在CFSFDP算法的基础上引入了第2个超参数t_c,针对时间序列数据集X,修改了ρ与δ的计算策略。其中时间约束下的局部密度ρ计算方法如式(5)所示

$$\rho_i=\sum\nolimits_j \chi(d_{ij}-d_c,t_{ij}-t_c) \tag{5}$$

式中,d_{ij}为样本点i与样本点j之间的空间距离;t_{ij}为样本点i与样本点j之间的时间距离;d_c和t_c是人为指定的超参数;d_c为空间截断距离;t_c为时间截断距离;$\chi(x,y)$表示0~1函数,当x和y同时小于0时,$\chi(x,y)=1$,否则,$\chi(x,y)=0$,即时间约束下的局部密度ρ_i表示同时小于空

间邻域 d_c 和时间邻域 t_c 的样本点个数。其次，采用样本点的时间距离计算 δ，如图 4(a)所示，簇集 $A(t_3,t_4,t_5,t_6)$、$B(t_{n-4},t_{n-3},t_{n-2},t_{n-1})$虽然在空间距离上相距很近，但是两个簇集的时间距离却很远，因此采用时间距离即可将两个簇集区分，从而产生两个聚类中心。针对单簇集可能存在多密度峰值的不足，本文将样本点的局部密度 $\{\rho_i\}_{i=1}^N$降序排列，获取其降序排列的下标序列 $\{q_i\}_{i=1}^N$，即 $\rho_{q_1}\geqslant\rho_{q_2}\geqslant\cdots\geqslant\rho_{q_N}$，使用下标序列 $\{q_i\}_{i=1}^N$重新对 δ 作了定义，其中 δ 的计算方法如式(6)所示

$$\delta_{q_i}=\begin{cases}\min\limits_{\substack{q_j\\ j<i}}\{t_{q_iq_j}\} & i\geqslant 2\\ \max\limits_{j\geqslant 2}\{\delta_{q_j}\} & i=1\end{cases}\tag{6}$$

式中，$t_{q_iq_j}$ 表示样本点 x_{q_i} 与样本点 x_{q_j} 的时间距离，即修改后的 δ 根据样本点的局部密度降序依次计算，即使单簇集中存在多个密度峰值，依旧仅有一个样本点被选做聚类中心。然而当单簇集的时间跨度过长时，仅使用时间距离计算 δ 同样会将单簇集分裂成多个小簇集。如图 4(b)所示，假设簇集 C 内部的样本点 4、k 具有较高的局部密度 ρ，由于两样本点的时间跨度较大，两点都将具有较大的 δ，依据本文的计算策略 $\gamma_i=\rho_i\delta_i$，簇集 C 将产生两个聚类中心。为解决这个问题，本文添加了聚类中心合并的过程，将识别的聚类中心按时间顺序链接，分别计算相邻两聚类中心的距离，若小于距离邻域 d_c，将相邻两个聚类中心合并，经聚类中心合并后，聚类中心的个数将与簇集的个数一一对应。

在 ST-CFSFDP 算法中，剩余样本点的分配沿时间轴进行。如图 5 所示，针对时间顺序分布的数据集 X，首先识别数据集中的聚类中心（红色样本点为合并后的聚类中心），然后每个聚类中心沿时间轴向前、后搜索本簇集中的剩余样本点，以聚类中心 CP_2 向前搜索为例，寻找 CP_2 前一时刻的样本点 t_{15}，计算样本点与 CP_1 和 CP_2 的距离，如果距离 CP_2 近，则将样本点 t_{15}归属于 CP_2，否则聚类中心 CP_2 向前搜索完毕，以同样的策略向后搜索最终确定该簇集包含的所有样本。所有聚类中心依次搜索完毕后，未被分配类别的样本点为离群点，以图 5 黄色样本点为例，聚类中心 CP_2 延时间轴向后搜索时，由于黄色样本点 OP_1 距离聚类中心 CP_3 较近，因此聚类中心 CP_2 向后搜索结束，当聚类中心 CP_3 延时间轴向前搜索时，黄色样本点 OP_2 距离聚类中心 CP_2 较近，因此聚类中心 CP_3 向前搜索结束，此时黄色样本点 OP_1 与 OP_2 之间的样本点为离群点。ST-CFSFDP 算法整体实现流程如下。

算法　快速搜索密度峰值的时空聚类算法

输入：序列数据 sequence$=\{t_i,p_i\}$，距离阈值 dis_{thr}，时间阈值 t_{thr}

输出：每一个样本点的聚类类别 labels$=\{c_i\}$

```
function ST_CFSFDP (sequence, dis_thr, t_thr)
    //计算样本点的空间距离矩阵
    stDisMat=computeSpatialDisMat(sequence)
    //计算样本点的时间距离矩阵
    timeDisMat=computeTimeDisMat(sequence)
    //计算样本点在时空约束下的局部密度
    densityArr=computeDensity(stDisMat,dis_thr,
    timeDisMat,t_thr)
    //获得局部密度降序排列的下标序列
    densitySortArr=argsort(densityArr)
    //初始化数组，用于存放每个轨迹点的属性δ
    closestDis=[]
    for i=0→len(densitySortArr) do
      //获得当前样本点的索引
      node=densitySortArr[i]
      //获得比当前样本点局部密度更大的索引集合
      nodeIdArr=densitySortArr[i+1:]
      //计算每一个样本点的属性δ
      closestDis[node]=compute(node, nodeIdArr,
      stDisMat,timeDisMat)
    end for
    //计算每一个样本点的属性γ
    gamma=closestDis*densityArr
    //通过决策图得到聚类的数目
    classNum=getNumFromDecisionGraph(gamma)
    //获取样本点的聚类类别
    labels=extract_clustering(gamma,classNum,
    stDisMat,timeDisMat)
    return labels
end function
```

1.3　算法时间复杂度分析

ST-CFSFDP 算法的时间复杂度定性描述了该算法的运行时间。ST-CFSFDP 算法的时间复杂度主要由两部分组成：计算样本点的局部密度 ρ 和距离 δ，其中 ρ 和 δ 的计算都涉及样本点之间的距离计算。在数据集规模为 n 的情况下，①计算任意两个样本点之间的距离，时间复杂度为 $O(n^2)$；②计算 n 个样本点的局部密度 δ，时间复

杂度为 $O(n^2)$；③计算 n 个样本点的距离 δ，时间复杂度同样为 $O(n^2)$。因此，ST-CFSFDP 算法的时间复杂度为 $O(n^2)$。

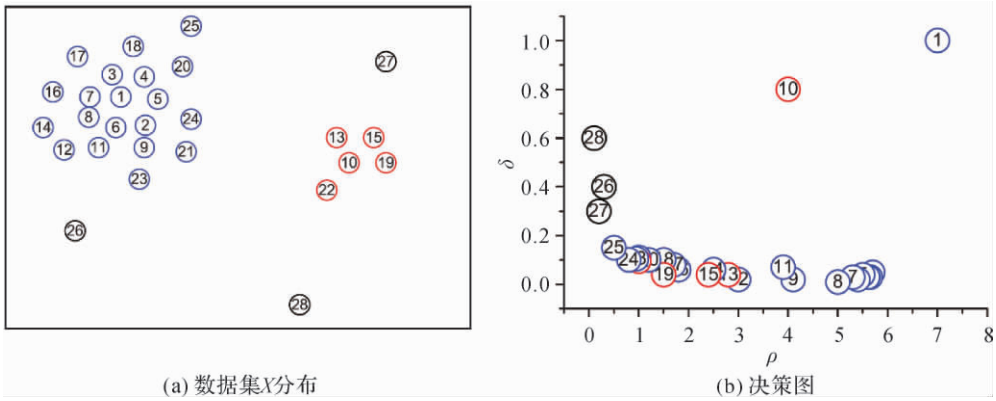


图 2 数据集 X 及其决策图

Fig.2 Dataset X and its decision graph

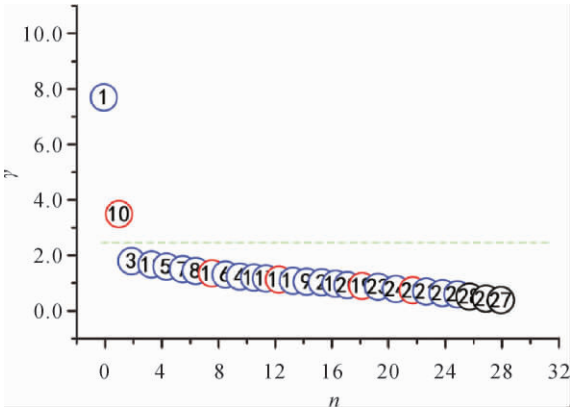
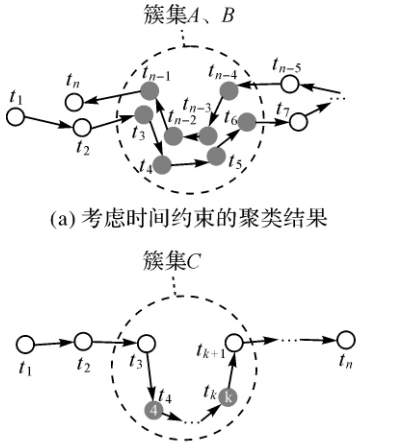


图 3 递减顺序排列的 γ

Fig.3 The value of γ in decreasing order



(a) 考虑时间约束的聚类结果

(b) 考虑单簇集时间距度过长的聚类结果

图 4 CFSFDP 算法改进

Fig.4 Improvement of CFSFDP algorithm

图 5 ST-CFSFDP 算法剩余点分配

Fig.5 Distribution of non-clustered center points of ST-CFSFDP algorithm

2 试验结果与讨论

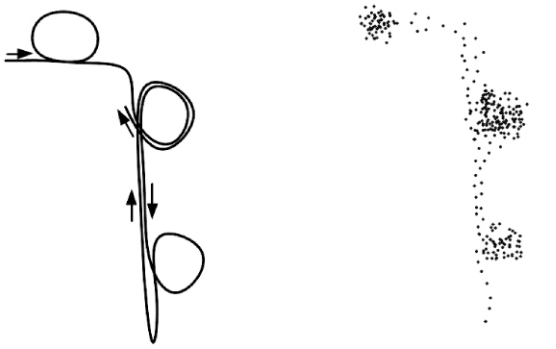
2.1 试验数据

为分析 ST-CFSFDP 算法的性能，分别采用模拟数据集和真实用户轨迹进行试验。采用模拟数据集进行试验的主要目的是以二维图形的方式直观地展示聚类结果，采用真实用户轨迹进行试验的主要目的是使用相关指标定量地评价聚类算法的性能。

本文共模拟 4 组数据量不同的时序数据集 X。每组模拟数据集的生成过程如下：首先生成指定的具有时间约束的几何形状，然后根据指定的几何形状等时间间隔采样而成，其中一组数据如图 6(a)、(b)所示。

真实数据来源于济南市某广场移动用户的 WiFi 定位数据。用户移动定位数据是一种典型的时空数据，在一定程度反映了用户的个人生活习惯和日常行为模式。时空聚类算法常用于从用

户移动定位数据识别用户的停留区域，从而进一步挖掘用户的兴趣爱好，因此本文将 ST-CFSFDP 算法应用于停留区域识别并定量的评价该算法的性能。移动用户定位数据覆盖广场 5 个楼层，平均采样为 1～10 s 不等，定位精度约为 3 m。由于定位数据难以获取用户停留区域的标注数据，因此本文采用爬虫技术从百度地图上获取了该广场的商铺数据，借助商铺数据标注用户轨迹中的停留信息。标注过程如下：首先对商铺数据与蓝牙定位数据求交，然后统计某用户在商铺内部的持续停留时间，若用户在商铺内部的持续停留时间超过一定阈值，则将相应的轨迹点标注为停留点。依据上述规则，本文共标注 472 条用户轨迹（当用户轨迹中停留区域过少时，本文认为该轨迹价值不高，删除该轨迹），轨迹点总记录量为 935 639 个。经标注后的某用户轨迹数据如表 1 所示，每条记录包括用户唯一标识 ID、记录时间、用户位置（X、Y 坐标及所在楼层 ID）、停留的商铺 ID（－1 代表用户未停留）。



(a) 时间约束的数据集X　　(b) 图(a)对应的散点图

图 6　模拟数据集

Fig.6　Simulated data sets

表 1　单用户轨迹数据实例

Tab.1　Samples of user's records　　m

用户 ID	时间	X	Y	所在楼层 ID	停留的商铺 ID
0000CE***	20171220104645	130219***	43904***	1	－1
0000CE***	20171220104657	130219***	43903***	1	1
0000CE***	20171220104705	130219***	43904***	1	1
…	…	…	…	…	…
0000CE***	20171220192033	130219***	43904***	4	－1
0000CE***	20171220192045	130219***	43904***	4	－1

2.2　算法评价指标及对比试验选择

本文以准确率（accuracy）和召回率（recall）作为停留区域识别的定量评价指标。用户停留区域可理解为时间约束下用户轨迹中的簇集，进一步可抽象为[36] S＝（userID，Lng，Lat，arrT，levT），（Lng，Lat）表示停留区域内的某个点坐标，其应最大概率出现在停留区域内部，一般为区域内所有轨迹点的均值坐标，本文为聚类中心坐标；arrT 表示用户抵达停留区域的时间；levT 表示用户离开停留区域的时间。停留区域识别结果是否正确主要依赖 3 个方面：①识别的停留区域和标注的停留区域数量是否一致；②识别的停留区域（Lng，Lat）坐标是否处于标注商铺内部；③识别的停留区域起止时间（arrT，levT）是否与标注数据一致。结合上述 3 个方面，本文 accuracy 和 recall 计算方法如式（7）、式（8）所示。

$$\text{accuracy}=\frac{\text{SC}_{\text{correct}}}{\text{SC}_{\text{algorithm}}} \tag{7}$$

$$\text{recall}=\frac{\text{SC}_{\text{correct}}}{\text{SC}_{\text{label}}} \tag{8}$$

式中，SC_{label} 表示标注的停留区域个数；$\text{SC}_{\text{algorithm}}$ 表示算法识别的停留区域个数；$\text{SC}_{\text{correct}}$ 表示算法判断正确的停留区域个数，本文采用 SO 和 TO[15] 两个函数共同判断某个停留区域是否识别正确（SO 函数判断识别结果和标注数据在空间上是否邻近，TO 函数判断识别结果和标注数据在时间上是否重合）。本文将 ST-DBSCAN、ST-OPTICS、ST-AGNES、ST-CFSFDP 时空聚类算法进行对比，探讨 4 种时空聚类算法在不同阈值下准确率和召回率的变化情况。

2.3　模拟数据集上的测试结果分析

CFSFDP 与 ST-CFSFDP 算法在模拟数据集上的聚类结果如图 7（e）、7（f）所示：①对于区域 A，由于存在两个聚类中心，使用 CFSFDP 算法，分裂成两个小簇集，产生了错误的聚类结果，而使用 ST-CFSFDP 算法，依据改进的属性计算策略，

只聚类成一个簇集,从而得到更合理的聚类结果;②对于区域 B,由于 CFSFDP 算法未考虑数据的时间约束,因此只聚类出一个簇集,而 ST-CFSFDP 算法考虑数据的时间约束,所以可以聚类出相同位置但不同时间的两个簇集;③对于区域 C,使用 ST-CFSFDP 算法识别出两个聚类中心,但仅产生一个簇集,原因是其中一个小簇集时间跨度过长,在该簇集中识别出两个聚类中心,但 ST-CFSFDP 算法最终会将这两个聚类中心合并(相邻聚类中心的距离小于距离阈值 d_c),因此最终只产生一个簇集。综上所述,与 CFSFDP 算法相比,ST-CFSFDP 算法不仅克服了单簇集中可能存在多密度峰值的不足,且实现了考虑时间约束的时空聚类。ST-CFSFDP 算法与 ST-DBSCAN、ST-OPTICS、ST-AGNES 算法的聚类结果如图 8 所示,进一步说明 ST-CFSFDP 算法具有发现时空簇集的能力。

由上文分析可知,ST-CFSFDP 算法具有良好的聚类性能,不仅解决了单簇集存在多密度峰值的问题,还能正确发现时空数据集的簇分布特征。进一步分析算法的运行时间,表 2 为两种聚类算法在 4 组模拟数据集上运行时间的对比。试验结果可以看出,ST-CFSFDP 算法的时间开销要略大于 CFSFDP 算法,这主要是因为 ST-CFSFDP 算法在计算样本点局部密度时增加了时间开销。但是这两种算法的时间复杂度的均为 $O(n^2)$。与 ST-DBSCAN、ST-OPTICS、ST-AGNES 算法的运行时间相比,ST-CFSFDP 的运行时间较小,能够较快地得到聚类的运行结果。

表 2 时空聚类算法运行时间在模拟数据集对比分析

Tab.2 The running time of spatio-temporal clustering on simulated data set s

数据集 X 记录数	CFSFDP	ST-CFSFDP	ST-DBSCAN	ST-OPTICS	ST-AGNES
314	0.004 5	0.004 8	0.005 2	0.031 2	0.003 3
3454	0.215 9	0.298 2	0.805 8	1.972 3	0.232 4
6363	0.683 9	0.842 0	3.122 1	10.671	0.791 8
15 623	4.009 4	5.879 1	15.627 6	26.515	4.891 7

2.4 真实数据集上的测试结果分析

为评价 ST-CFSFDP 算法性能,本文参考室内房间宽度将初始距离阈值设置为 1 m,并以 0.5 m步长递增;初始时间阈值设置为 50 s,以20 s 步长递增;以启发式的方法[22]设置 ST-DBSCAN、ST-OPTICS 算法中 minPts 参数,minPts=ln(N),N 为某用户轨迹点数量。算法识别结果的准确率对比如图 9 所示。4 种算法的识别准确率变化趋势在整体上较为相似,在时间阈值固定的情况下,准确率均随着距离阈值的增加呈现先增加后下降的趋势。综合 4 种算法在不同阈值下的表现可以发现:①ST-CFSFDP 算法对参数敏感度低,ST-DBSCAN 与 ST-OPTICS 算法对参数敏感度较高(准确率随着超参数的改变波动较大);②相较 ST-DBSCAN、ST-OPTICS、ST-AGNES 算法,ST-CFSFDP 算法准确率较高,在时间阈值 90 s 时最为明显,ST-CFSFDP 算法准确率最高可达 82.4%,与现有算法相比,最高可提升 7.6%;③ST-DBSCAN、ST-OPTICS 算法本质上是一种算法,所以两种算法的准确率高度一致;④ST-AGNES 算法隐含时间约束,仅需要距离阈值即可完成聚类,因此算法准确率不受时间阈值的影响。

算法识别结果的召回率对比如图 10 所示。4 种算法的召回率均随距离阈值增加呈现先增加后下降的趋势,与算法准确率一致。当时间阈值为 90 s 时,4 种算法的召回率较高,原因在于人工标注时用户的停留时间阈值与 90 s 较接近。在此阈值下 4 种算法将会得到较精确的识别结果。与此同时 ST-CFSFDP 算法的召回率略高于其余 3 种算法且准确率与召回率波动最小,原因为 ST-CFSFDP 算法采用聚类中心作为用户最可能的停留位置,而聚类中心作为局部范围内密度最大的点,受阈值影响较小。本文将 4 种算法最高准确率相应的运行时间进行对比,结果如表 3 所示。ST-CFSFDP 算法与 ST-AGNES 算法相比,ST-CFSFDP 算法的运行时略高,但算法的识别正确率却可提升 7.6%;与 ST-DBSCAN 算法相比,ST-CFSFDP 算法的运行时间略有降低且识别准确率提升了 5.2%;与 ST-OPTICS 算法相比,ST-CFSFDP 算法的运行时间得到了大幅度提升且识别准确率也有一定程度的提高。

表 3 4 种算法运行时间对比分析

Tab.3 The running time of four algorithms

算法名称	最高准确率/(%)	运行时间/s
ST-CFSFDP	82.4	84.658
ST-DBSCAN	77.2	129.464
ST-OPTICS	78.2	285.152
ST-AGNES	74.8	65.326

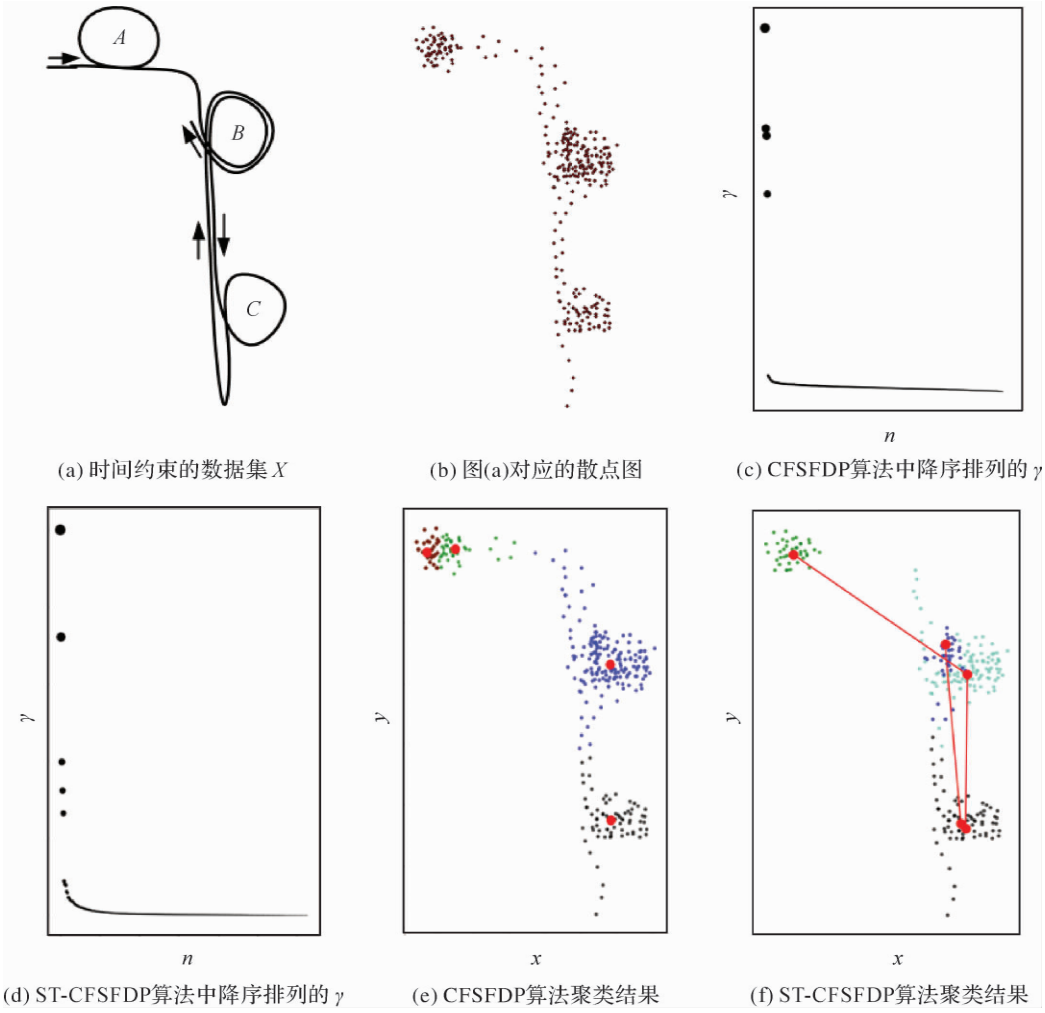


图 7　ST-CFSFDP 算法与 CFSFDP 算法的对比结果

Fig.7　Comparison between ST-CFSFDP algorithm and CFSFDP algorithm

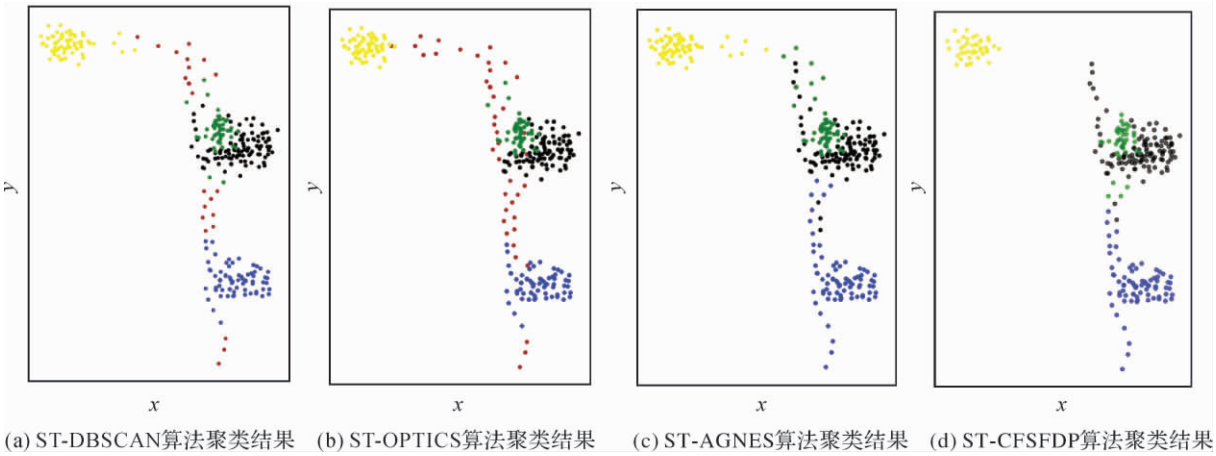


图 8　时空聚类算法聚类结果在模拟数据集对比分析

Fig.8　Spatio-temporal clustering results on simulated data sets

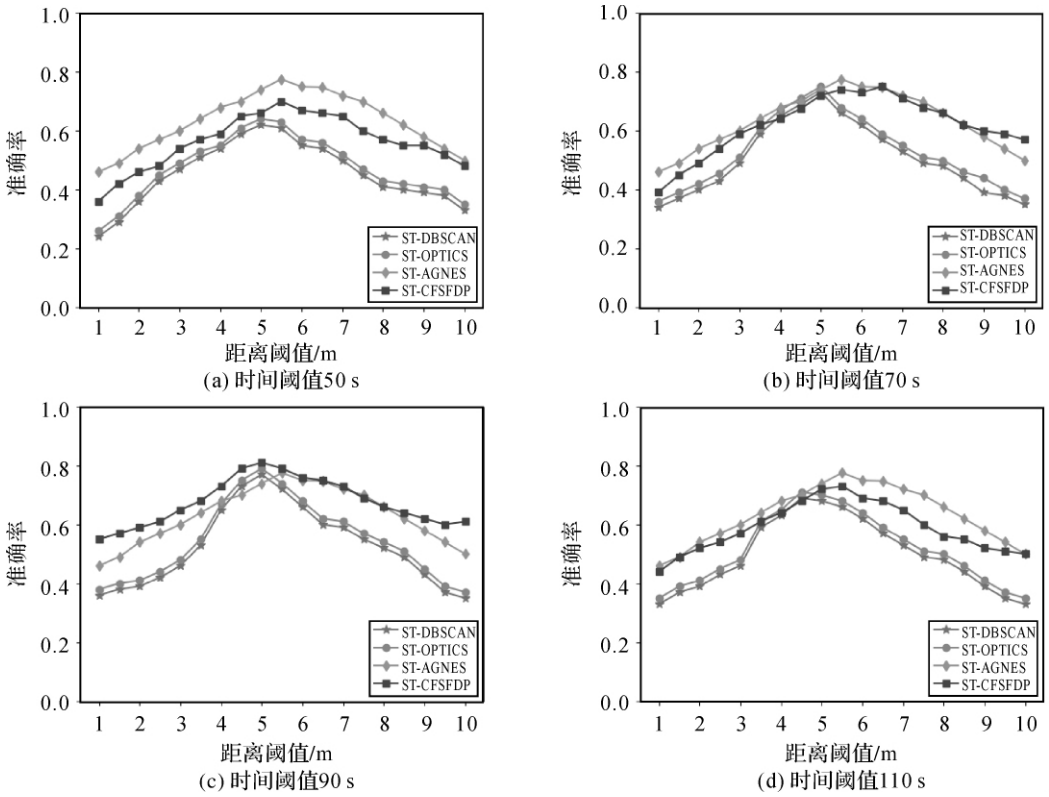


图 9 不同算法识别结果的准确率对比

Fig.9 Comparison of recognition accuracy of different algorithms

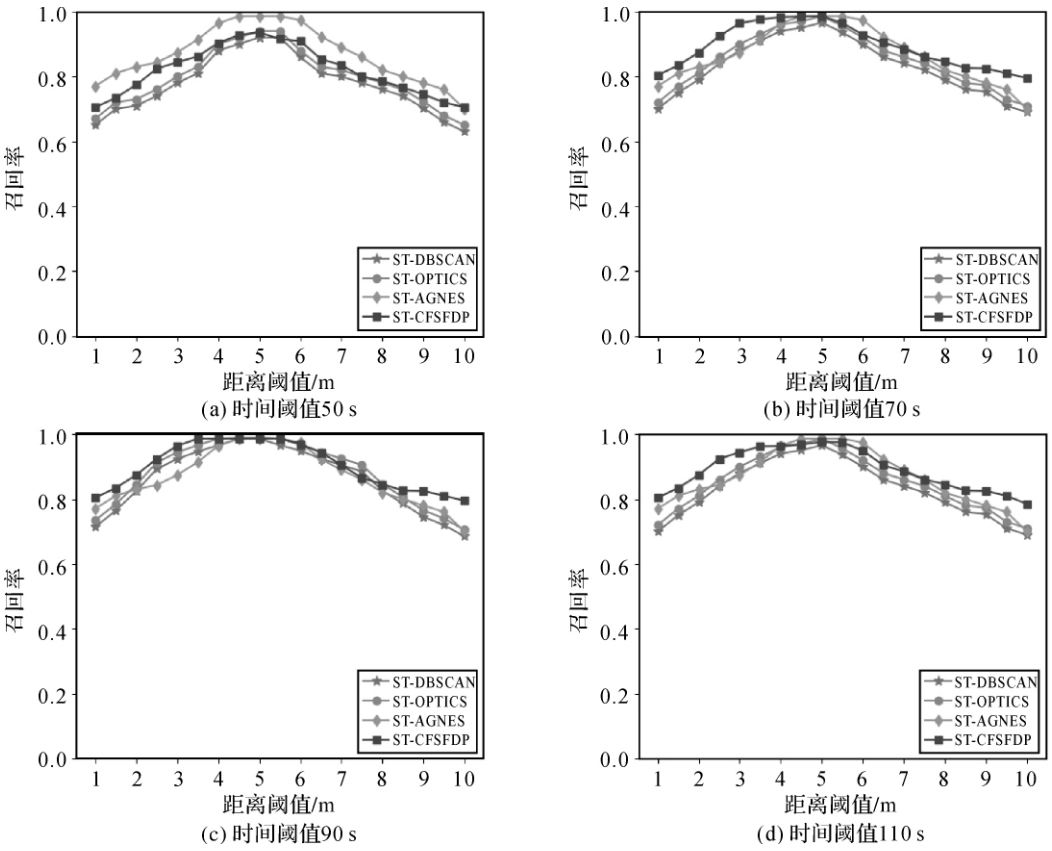


图 10 不同算法识别结果的召回率对比

Fig.10 Comparison of recognition recall of different algorithms

3　结论与展望

CFSFDP 算法是一种新颖的空间聚类算法，通过计算样本点的属性值 γ，能够快速发现数据集中的密度峰值点。然而当数据集的某簇集存在多密度峰值点时，该算法易产生错误的聚类结果，且 CFSFDP 算法无法实现考虑时间约束的时空聚类。针对以上两点不足，本文提出了时空聚类算法 ST-CFSFDP。ST-CFSFDP 算法在 CFSFDP 算法基础上加入时间约束，并修改了样本属性值的计算策略。为验证算法的有效性，首先采用模拟数据进行定性试验。结果表明，与原有的 CFSFDP 算法相比，ST-CFSFDP 算法不仅可以克服单簇集中可能存在多密度峰值的不足，且可以区分并识别相同位置不同时间的簇集。最后本文将 ST-CFSFDP 算法应用于用户的停留区域识别，结果表明，ST-CFSFDP 算法在时间阈值90 s、距离阈值 5 m 的识别正确率高达 82.4%，较经典的 ST-DBCSAN、ST-OPTICS 和 ST-AGNES 算法分别提高了 5.2%、4.2%和 7.6%。

本文尚在以下方面存在不足，需在后续工作中进一步研究：受试验数据采样间隔的限制，现有算法仅采用秒级时间粒度的数据进行验证，当定位数据的采样间隔增大时，算法识别准确率及可靠性需要进一步探究。

致谢：感谢上海图聚智能科技股份有限公司为本研究提供室内定位轨迹试验数据！

参考文献：

[1] LIU Hongzhi, WU Zhonghai, ZHANG Xing. CPLR: Collaborative pairwise learning to rank for personalized recommendation[J]. Knowledge-Based Systems, 2018, 148(5): 31-40.

[2] ZHOU Yuwem, HUANG Changqin, HU Qintai, et al. Personalized learning full-path recommendation model based on LSTM neural networks[J]. Information Sciences, 2018, 444(5): 135-152.

[3] XUE Hao, HUYNH D Q, REYNOLDS M. SS-LSTM: ahierarchical LSTM model for pedestrian trajectory prediction[C]// Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, NV: IEEE, 2018.

[4] MENG Fanrong, YUAN Guan, LÜ Shaoqian, et al. An overview on trajectory outlier detection[J]. Artificial Intelligence Review, 2018(10): 1-20.

[5] 李志林，刘启亮，唐建波. 尺度驱动的空间聚类理论[J]. 测绘学报，2017，46(10)：1534-1548. DOI：10.11947/j.AGCS.2017.20170275.
LI Zhilin, LIU Qiliang, TANG Jianbo. Towardsa scale-driven theoryfor spatialclustering[J]. Acta Geodaetica et Cartographica Sinica, 2017, 46(10): 1534-1548. DOI: 10.11947/j.AGCS.2017.20170275.

[6] 刘启亮，邓敏，石岩，等. 一种基于多约束的空间聚类方法[J]. 测绘学报，2011，40(4)：509-516.
LIU Qiliang, DENG Min, SHI Yan, et al. A novel spatial clustering method based on multi-constraints[J]. Acta Geodaetica et Cartographica Sinica, 2011, 40(4): 509-516.

[7] 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. DOI：10.13474/j.cnki.11-2246.2018.0001.
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. DOI: 10.13474/j.cnki.11-2246.2018.0001.

[8] 牟乃夏，张恒才，陈洁，等. 轨迹数据挖掘城市应用研究综述[J]. 地球信息科学学报，2015，17(10)：1136-1142.
MOU Naixia, ZHANG Hengcai, CHEN Jie, et al. A review on the application research of trajectory data mining in urban cities[J]. Journal of Geo-information Science, 2015, 17(10): 1136-1142.

[9] KALANTARI M, YAGHMAEI B, GHEZELBASH S. Spatio-temporal analysis of crime by developing a method to detect critical distances for the Knox test[J]. International Journal of Geographical Information Science, 2016, 30(11): 2302-2320.

[10] ZALIAPIN I, GABRIELOV A, KEILIS-BOROK V, et al. Clustering analysis of seismicity and aftershock identification[J]. Physical Review Letters, 2008, 101(1): 018501.

[11] WANG Jiao, CHENG Weiming, ZHOU Chenghu, et al. Automatic mapping of lunar landforms using DEM-derived geomorphometric parameters[J]. Journal of Geographical Sciences, 2017, 27(11): 1413-1427.

[12] ZHAO Quanhua, LI Xiaoli, LI Yu, et al. A fuzzy clustering image segmentation algorithm based on hidden Markov random field models and Voronoi tessellation[J]. Pattern Recognition Letters, 2017, 85(2): 49-55.

[13] ACEDO-HERNÁNDEZ R, TORIL M, LUNA-RAMÍREZ S, et al. Automatic clustering algorithms for indoor site selection in LTE[J]. EURASIP Journal on Wireless Communications and Networking, 2016(12): 87-98.

[14] 姜波，叶灵耀，潘伟丰，等. 基于需求功能语义的服务聚类方法[J]. 计算机学报，2018，41(6)：1255-1266.
JIANG Bo, YE Lingyao, PAN Weifeng, et al. Service clustering based on the functional semantics of requirements[J]. Chinese Journal of Computers, 2018, 41(6): 1255-1266.

[15] 林楠，尹凌，赵志远. 基于滑动窗口的手机定位数据个体停留区域识别算法[J]. 地球信息科学学报，2018，20(6)：762-771.
LIN Nan, YIN Ling, ZHAO Zhiyuan. Detecting individual stay areas from mobile phone location data based on

moving windows[J]. Journal of Geo-information Science, 2018, 20(6): 762-771.

[16] 周世波, 徐维祥. 密度峰值快速搜索与聚类算法及其在船舶位置数据分析中的应用[J]. 仪器仪表学报, 2018, 39(7): 152-163.
ZHOU Shibo, XU Weixiang. Clustering by fast search and find of density peaks andits application in ship location data analysis[J]. Chinese Journal of Scientific Instrument, 2018, 39(7): 152-163.

[17] MACQUEEN J. Some methods for Classification and analysis of multivariate observations[C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley, Calif.: University of California Press, 1967: 281-297.

[18] PARK H S, JUN C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.

[19] 王寅同,王建东,陈海燕,等.一种代表点的近似折半层次聚类算法[J].小型微型计算机系统,2015, 36(2):215-219.
WANG Yintong, WANG Jiandong, CHEN Haiyan, et al. An algorithm for approximate binary hierarchical clustering using representatives[J]. Journal of Chinese Computer Systems, 2015, 36(2): 215-219.

[20] ESTIVILL-CASTRO V, LEE I. Multi-level clustering and its visualization for exploratory spatial analysis[J]. Geoinformatica, 2002, 6(2): 123-152.

[21] HANRAHANP, SALZMAN D, AUPPERLE L. A rapid hierarchical radiosity algorithm[C]// Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques.New York, NY: ACM, 1991: 197-206.

[22] ESTER M, KRIEGEL H P, SANDERJ, et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, Oregon:AAAI Press, 1996: 226-231.

[23] 唐建波, 邓敏, 刘启亮. 时空事件聚类分析方法研究[J]. 地理信息世界, 2013, 20(1): 38-45.
TANG Jianbo, DENG Min, LIU Qiliang. On spatio-temporal events clustering methods[J]. Geomatics World, 2013, 20(1): 38-45.

[24] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure [C]// Proceedings of ACM-SIGMOD International Conference on Management of Data. Philadelphia PA:ACM, 1999.

[25] WANG Wei, YANG Jiong, MUNTZ R R. STING: a statistical information grid approach to spatial data mining [C]// Proceedings of the 23rd International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1997: 186-195.

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

[27] AGRAWAL K P, GARG S, SHARMA S, et al. Development and validation of OPTICS based spatio-temporal clustering technique[J]. Information Sciences, 2016, 369: 388-401.

[28] BAIESI M, PACZUSKI M. Scale-free networks of earthquakes and aftershocks[J]. Physical Review E, 2004, 69(6): 066106.

[29] KULLDORFF M, HEFFERNAN R, HARTMAN J, et al. A space-time permutation scan statistic for disease outbreak detection[J]. PLoS Medicine, 2005, 2(3): e59.

[30] GAUDART J, POUDIOUGOU B, DICKO A, et al. Space-time clustering of childhood malaria at the household level: a dynamic cohort in a Mali village[J]. BMC Public Health, 2006(6): 286-298.

[31] LIU Qiliang, DENG Min, BI Jiantao, et al. A novel method for discovering spatio-temporal clusters of different sizes, shapes, and densities in the presence of noise[J]. International Journal of Digital Earth, 2014, 7(2): 138-157.

[32] PEI Tao, ZHOU Chenghu, ZHU A'xing, et al. Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise[J]. International Journal of Geographical Information Science, 2010, 24(6): 925-948.

[33] LEIVA L A, VIDAL E. Warped k-means: an algorithm to cluster sequentially-distributed data[J]. Information Sciences, 2013(237): 196-210.

[34] 王培晓, 王海波, 傅梦颖, 等. 室内用户语义位置预测研究[J]. 地球信息科学学报, 2018, 20(12): 1689-1698.
WANG Peixiao, WANG Haibo, FU Mengying, et al. Research on semantic location prediction of indoor users[J]. Journal of Geo-information Science, 2018, 20(12): 1689-1698.

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

[36] LI Quannan, ZHENG Yu, XIE Xing, et al. Mining user similarity based on location history[C]// Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Irvine, California:ACM, 2008: 34.

(责任编辑:丛树平)

收稿日期: 2018-11-23
修回日期: 2019-04-08

第一作者简介: 王培晓(1994—),男,硕士生,研究方向为地理信息服务、时空数据挖掘。
First author: WANG Peixiao(1994—), male, postgraduate, majors in geographic information services and spatio-temporal data mining.
E-mail: peixiao_wang@163.com
通信作者: 吴升
Corresponding author: WU Sheng
E-mail: ws0110@163.com