基于方差优化谱聚类的热点区域挖掘算法*

梁卓灵[1],元昌安[2],覃　晓[3]**
(1.广西大学,广西南宁　530004;2.广西科学院,广西南宁　530007;3.南宁师范大学,广西南宁　530000)

摘要:为改善交通拥堵的情况,本文利用聚类分析方法对移动轨迹数据进行挖掘,识别居民出行的热点区域。传统的 Ng-Jordan-Weiss (NJW)谱聚类算法常使用 K-means 聚类算法来实现最后的聚类操作,然而 K-means 聚类算法存在对初始值敏感、容易陷入局部最优的缺陷,影响对热点区域的挖掘结果。因此,本研究将方差优化初始中心的 K-medoids 聚类算法运用到谱聚类算法最后聚类阶段,提出基于方差优化谱聚类的热点区域挖掘算法(Hot Region Mining algorithm based on improved K-medoids Spectral Clustering,HRM-KSC),然后在真实的轨迹数据集上进行试验。试验结果发现,HRM-KSC 算法聚类结果的轮廓系数更高,表明 HRM-KSC 算法改善了 NJW 谱聚类算法,提高了聚类质量。

关键词:K-medoids 算法　谱聚类　热点区域　停留点　交通拥堵

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2020)06-0616-06
DOI:10.13656/j.cnki.gxkx.20210119.003

0　引言

近年来,城市的交通拥堵问题日益严重,影响居民的出行以及城市的发展。为改善交通拥堵的情况,可以通过聚类分析的方法对城市居民出行轨迹数据进行挖掘,分析总结居民的出行规律及行为习惯,从而合理地规划城市的公交线路并优化交通资源的配置,为居民推荐合理的出行方式[1,2]。

Ng-Jordan-Weiss (NJW)谱聚类算法是由 Ng 等[3]提出的,属于经典的多路谱聚类算法。NJW 谱聚类算法通过样本数据点的相似矩阵和度矩阵计算求得 Laplician 矩阵,并由 Laplician 矩阵的前 k 个最大特征值对应的特征向量建立一个 $n*k$ 阶的矩阵。最后把矩阵中的每一行看作是空间中的点,用 K-means 聚类算法进行聚类。但 K-means 聚类算法[4]利用簇内点的平均值作为簇的中心点,对孤立点敏感,且其受初始值影响,易陷入局部最优解,不利于对热点区域的聚类挖掘。

轨迹数据是在空间无规则分布的且分布不均匀的数据。K-medoids 聚类算法[5]总是选择簇中位置最接近簇中心的对象作为簇的中心点,能消除对孤立点和噪声点的敏感性,比 K-means 聚类算法更适用于轨迹数据的聚类。但 K-medoids 聚类算法对初始中心点的选取仍是随机的,而利用方差优化初始中心

*国家自然科学基金项目(61962006,61802035,61772091),广西科技开发项目(AA18118047,AD18126015)和广西自然科学基金项目(2018GXNSFDA138005)资助。

【作者简介】
梁卓灵(1996—),男,在读硕士研究生,主要从事数据挖掘研究,E-mail:1104589997@qq.com。
【**通信作者】
覃　晓(1973—),女,副教授,主要从事计算机图像处理与数据挖掘研究,E-mail:7670172@qq.com。

【引用本文】
梁卓灵,元昌安,覃晓.基于方差优化谱聚类的热点区域挖掘算法[J].广西科学,2020,27(6):616-621.
LIANG Z L,YUAN C A,QIN X. Hot Region Mining Algorithm based on Variance Optimization Spectrum Clustering [J]. Guangxi Sciences,2020,27(6):616-621.

的K-medoids聚类算法可改善其对中心点的选取,实现对于热点区域的挖掘。

本文把方差优化初始中心的K-medoids聚类算法[6]运用到NJW谱聚类算法的最后聚类阶段,提出基于方差优化谱聚类的热点区域挖掘算法(Hot Region Mining algorithm based on improved K-medoids Spectral Clustering,HRM-KSC),然后在真实的轨迹数据集上进行试验,证明HRM-KSC算法的聚类效果。

1 K-medoids聚类算法的基本原理

K-medoids聚类算法通常用一个代价函数来评估聚类质量的好坏,以重复迭代的方式寻找到最好的聚簇划分及聚簇中心点。这里使用聚类误差平方和来评估聚类结果质量,定义如下:

$$E=\sum_{j=1}^{k}\sum_{x\in C_j}|x-O_j|^2, \tag{1}$$

其中,x 为各个簇类 C_i 中的样本,O_j 为其聚类中心。

K-medoids聚类算法步骤可描述如下:

Step 1:从数据集中随机选择 k 个对象,作为初始的聚类中心点;

Step 2:根据与中心点距离的远近,将数据集中的其他非中心点对象分配到最近中心点所在的簇类;

Step 3:重新计算每个簇的中心点位置,使其到该簇其他样本的距离总和最小;

Step 4:重复执行Step 2和Step 3,直到聚类误差平方和基本不变,达到指定要求为止。

一般K-means聚类算法是通过计算簇类点的平均值来选取中心点,其对孤立点敏感,选取的中心点可能不存在。与K-means聚类算法不同,K-medoids聚类算法在迭代选取中心点时,总是在中心点的周围选择样本点作为新的中心点,消除了对孤立点的敏感性。

2 基于方差优化初始中心点的K-medoids聚类算法

方差优化初始中心点的K-medoids聚类算法是对K-medoids聚类算法的改进,称之为Standard-Deviation as radius of neighborhood (SD_K-medoids)算法。该算法利用方差是反映样本分布密集或疏散程度的特性[7],以方差度量样本分布的密集程度,采用样本标准差为邻域半径,从不同的样本分布密集区域中选择样本作为K-medoids聚类算法的初始聚类中心。

2.1 基本概念描述

设样本数据集为 $X=\{x_i|x_i\in R^p,i=1,2,\cdots,n\}$,则样本 x_i 和 x_j 间的欧式距离 $\mathrm{dist}(i,j)$ 可定义为

$$\mathrm{dist}(i,j)=\sqrt{\|x_i-x_j\|}。 \tag{2}$$

为消除样本之间的差异性,需对样本进行标准化。本文采用最大最小标准化方法,将样本的属性值 x'_{ij} 映射到[0,1]区间,公式如下所示:

$$x'_{ij}=\frac{x_{i,j}-\min(X_{:,j})}{\max(X_{:,j})-\min(X_{:,j})}。 \tag{3}$$

方差是各个数据与平均数之差的平方和的平均数,而标准差是方差的算术平方根。因此样本 x_i 的方差 V_i 和标准差 std_i 可分别定义如下:

$$V_i=\frac{1}{N-1}\sum_{j=1}^{N}\|x_j-x_i\|^2, \tag{4}$$

$$\mathrm{std}_i=\sqrt{V_i}, \tag{5}$$

其中,N 为样本总数。

SD_K-medoids算法以样本的标准差为邻域半径 r,其值与初始聚类中心相关,即 $r=\mathrm{std}_i=\sqrt{V_i}$,不是固定值。

2.2 SD_K-medoids算法描述

SD_K-medoids算法首先将样本标准化,然后计算数据集中各样本的方差,选择方差最小的样本作为第一个初始聚类中心加入中心点集;然后计算聚类中心的领域半径,从数据集中删除该样本点及该样本领域中的所有数据样本,在剩余数据样本中寻找方差最小的数据样本作为中心点,重复执行,直到选出 k 个初始中心点。剩余步骤则与K-medoids聚类算法相同:将数据集中的样本分配给与其距离最近的中心点,计算聚类误差平方和,计算新的中心点位置;重复执行,当聚类误差平方和不变,结束算法。

其中,SD_K-medoids算法通过方差反映样本分布的特性,每次可以在最密集的区域选取到中心点,使得初始类簇的划分更贴近数据集的真实分布,降低算法收敛的迭代次数,增加其收敛到全局最优解的概率。

使用样本标准差作为领域半径,每次选取一个中心点后,要把其领域内的样本点去除,从而避免初始中心点可能位于同一簇类的缺陷,使初始中心点尽可能地分布在不同的簇类。

所以,面对轨迹数据无规则分布、分布密度不均

匀的特点时,SD_K-medoids 算法可以尽快找到好的热点区域初始中心点,加快收敛速度,增加了收敛到全局最优解的可能。

3 基于方差优化谱聚类的热点区域挖掘算法

对于居民出行热点区域的挖掘,就是寻找居民频繁出行的区域,发现居民出行停留时间较长的聚集地点,因此需对居民出行停留点进行聚类操作。在聚类操作之前,本文主要借鉴文献[8]的方法,从移动对象的轨迹数据中提取居民出行停留点,详细方法本文不再具体说明。

针对谱聚类最后聚类阶段 K-means 聚类算法的不足,用 SD_K-medoids 算法来替代,提出基于方差优化谱聚类的热点区域挖掘算法(Hot Region Mining algorithm based on improved K-medoids Spectral Clustering,HRM-KSC)。HRM-KSC 算法的基本思想是把居民出行停留点看作待聚类的空间样本点,在 NJW 谱聚类算法[9]中把停留点集映射到相似度矩阵和拉普拉斯矩阵中,并将拉普拉斯矩阵的特征向量构建为特征矩阵,最后改用 SD_K-medoids 算法对特征矩阵的每行元素进行聚类。

3.1 相似度矩阵的构造

谱聚类把居民出行停留点集看作是一个无向图 $G(V,E,A)$ 的顶点集合 V,由边集 E 把停留点连接起来,而图中权重集 A 表示停留点间的相似性。通过权重来构建停留点集的相似度矩阵,停留点间的相似性常用高斯函数 w_{ij} 来计算:

$$w_{ij}=\begin{cases}\exp(-\mathrm{dist}^2(s_i,s_j)/2\sigma^2), & i\neq j\\ 0, & i=j\end{cases},\quad(6)$$

其中,s_i 和 s_j 分别表示停留点 i 和 j 的特征向量,σ 为尺度参数。

3.2 拉普拉斯矩阵的构造

本文对停留点集的构建采用的是规范的拉普拉斯矩阵 Lsym,其构造公式如下:

$$\mathrm{Lsym}=E-D^{-\frac{1}{2}}WD^{-\frac{1}{2}},\quad(7)$$

其中,W 为相似度矩阵;D 为图的度矩阵,其主对角线上的元素为相似度矩阵 W 的第 i 行元素之和,计算如下:

$$D_{ii}=\sum_{j=1}^{n}W_{ij}。\quad(8)$$

在构建停留点集的相似度矩阵及拉普拉斯矩阵后,将拉普拉斯矩阵前 k 个特征向量构建为特征矩阵,最后用 SD_K-medoids 算法对矩阵的每行元素进行聚类。

HRM-KSC 算法流程如图 1 所示,具体过程描述如下:

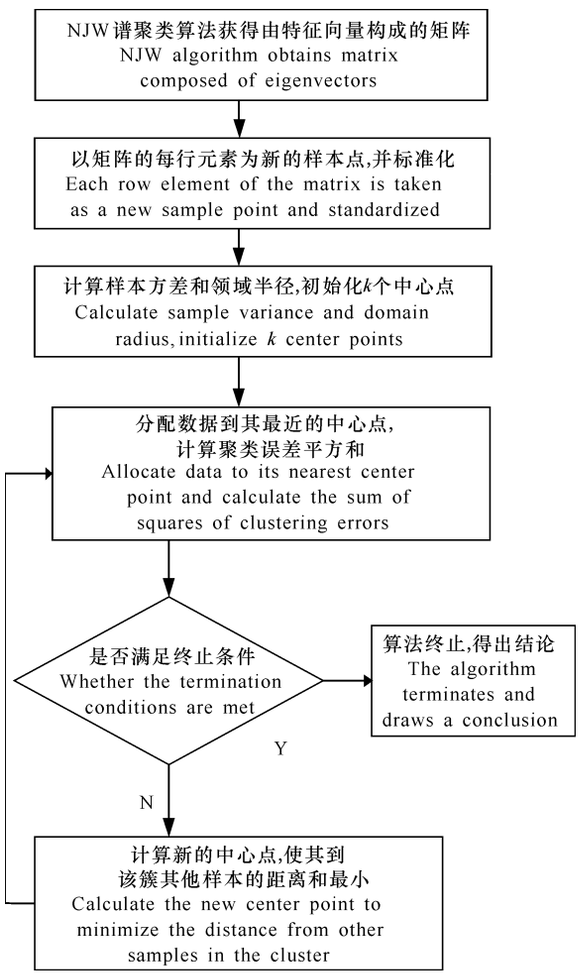

图 1 HRM-KSC 算法流程图

Fig. 1 Flow chart of HRM-KSC

Step 1:利用公式(6)计算停留点集的相似度矩阵;

Step 2:利用公式(7)和(8)计算停留点集的拉普拉斯矩阵;

Step 3:把拉普拉斯矩阵前 k 个特征向量组成的特征矩阵归一化为矩阵 $Y=[y_1,y_2,\cdots,y_k]$,把矩阵 Y 的每一行向量作为一个数据点 y_i';

Step 4:把所有数据点 $y_i'(i=1,2,\cdots,k)$ 作为新的样本点,根据式(3)对样本进行标准化。

Step 5:根据式(4)计算数据集中各样本的方差,如 V_i 表示第 i 个样本的方差值;其次初始化中心点集 M 为空,即 $M=\{\}$。

Step 6:从数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 中寻找方差最小的样本 $x_{\min}$,将其作为第一个类簇的初始聚类中心 C_1 加入到中心点集中,即 $M=M\cup\{C_1\}$;根据

式(5)计算邻域半径 r_1,从数据集中删除该样本以及该样本领域中的所有数据样本。

Step 7:重复执行 Step 6,直到选出 k 个初始中心点,即$|M|=k$。

Step 8:将数据集中的样本分配给与其距离最近的中心点,由公式(1)计算聚类误差平方和。

Step 9:计算每个簇的新中心点位置,使其到该簇其他样本的距离总和最小。

Step 10:重复执行 Step 8—Step 9,若聚类误差平方和变化不大,结束算法;否则继续迭代。

4 实验分析

为测试 HRM-KSC 算法的性能,本文使用微软亚洲研究院的 geolife 数据集,从 2008 年 8 月到 10 月的轨迹数据中提取出居民出行停留点集。其中,居民停留点的提取参照文献[8]中的方法。本次实验在 Win10 64 bit 操作系统,8 GB 内存,CPU 2.60 GHz 的环境下进行,用 Python 语言实现。

为度量 2 种算法在实验中的表现,采用 SC 轮廓系数作为评价指标。SC 轮廓系数常用于度量未知类别的聚类数据集,表示聚类结果中各簇类的稠密程度及簇间的离散程度。

SC 轮廓系数计算公式如下:

$$SC(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{9}$$

其中,$a(i)$表示计算样本 i 到同簇类其他样本的平均距离,$b(i)$表示计算样本 i 到其他样本的平均距离。SC 在[−1,1]区间内取值。当 SC 越接近 1 时,表示聚类效果越好。

本次实验测试了 2 种算法在用户停留点集上的聚类效果,在一定区间内选取 3 种较好的结果,如表 1 所示。

表 1 算法在停留点数据集上聚类结果

Table 1 Algorithm of clustering results on stay points dataset

算法 Algorithm	参数 Parameter	SC 轮廓系数 Silhouette coefficient	运行时间 Running time (s)
NJW	$k=3,\sigma=10$	0.616	3.609 3
	$k=4,\sigma=15$	0.603	3.562 5
	$k=5,\sigma=20$	0.628	3.625 0
HRM-KSC	$k=3,\sigma=10$	0.682	5.521 2
	$k=4,\sigma=15$	0.627	5.562 8
	$k=5,\sigma=20$	0.630	5.672 6

观察表 1 中 2 种算法在停留点数据集上的表现,发现 HRM-KSC 算法在选取相同参数的情况下,轮廓系数指标均比 NJW 谱聚类算法高,表明 HRM-KSC 算法的聚类结果中同簇类点更紧密,不同簇类点更离散,聚类结果更合理。

为更进一步展示 2 种算法的聚类结果,采用 Python 的 matplotlib 库,以经纬度为坐标画出不同参数条件下的聚类结果(图 2—4)。从图中可以看出,在相同参数条件下,HRM-KSC 聚类划分结果更合理,尤其在图 2 和图 3 中表现更明显,其原因是 NJW 谱聚类算法在最后阶段所使用的 K-means 算法,对于初始中心点的选取不够理想,影响了聚簇的划分效果。

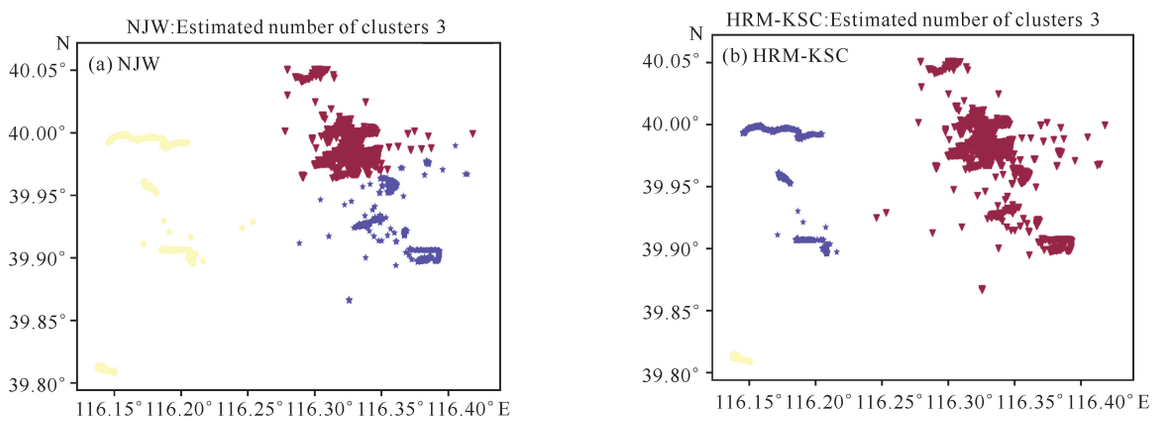

图 2 $k=3,\sigma=10$ 时的实验结果

Fig. 2 Experimental results when $k=3,\sigma=10$

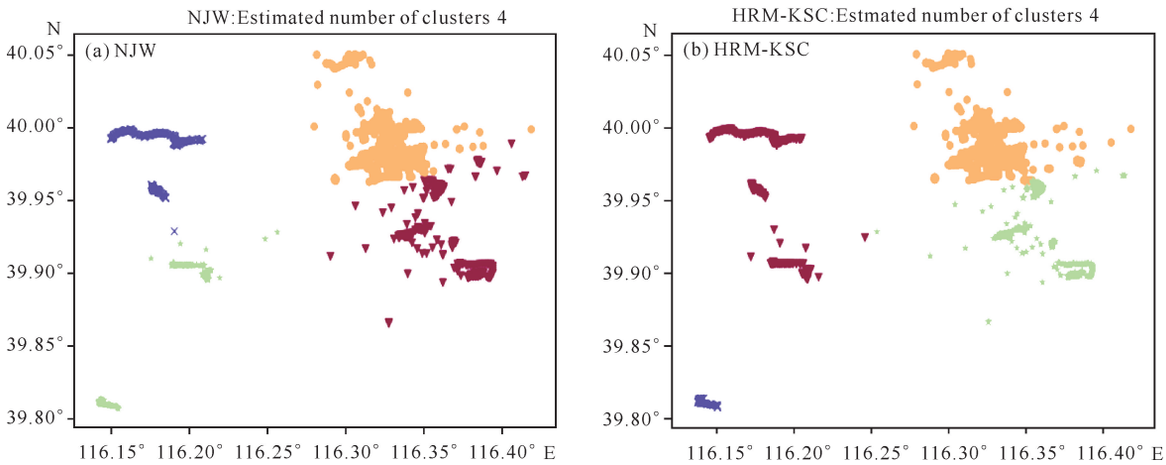

图 3　$k=4,\sigma=15$ 时的实验结果

Fig. 3　Experimental results when $k=4,\sigma=15$

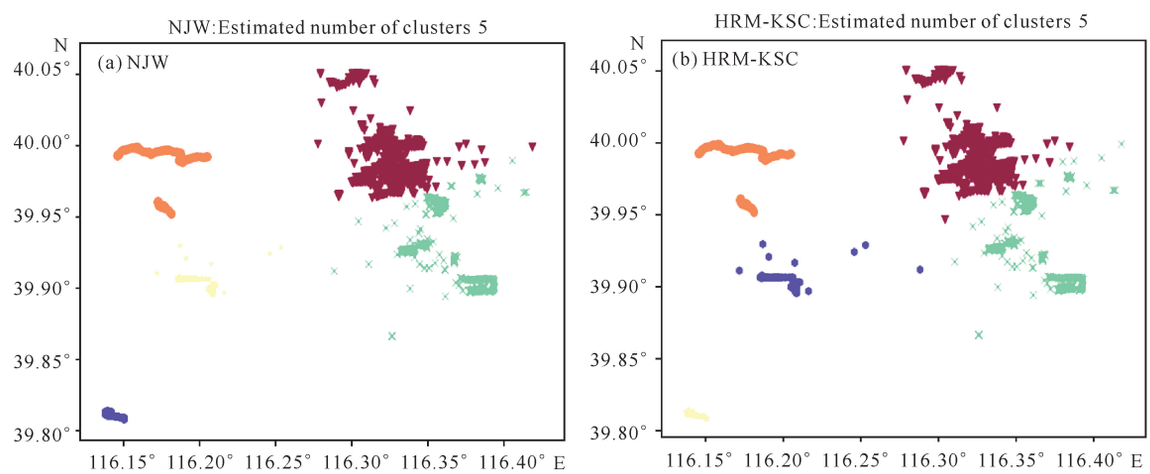

图 4　$k=5,\sigma=20$ 时的实验结果

Fig. 4　Experimental results when $k=5,\sigma=20$

5　结论

本文针对 NJW 谱聚类算法最后阶段的 K-means 聚类算法对初始点敏感的缺陷,利用 SD_K-medoids 算法计算样本方差和领域半径,优化对初始中心点的选取,提出基于方差优化谱聚类的热点区域挖掘算法(HRM-KSC)。在居民停留点数据集上进行 HRM-KSC 算法和 NJW 谱聚类算法的对比实验,结果表明 HRM-KSC 算法的聚类质量更高,聚类效果更好。后续期望改善谱聚类算法中高斯函数尺度参数的选取,以及研究如何确定聚类数目,以进一步提高谱聚类算法的聚类质量。

参考文献

[1] WEI W,QI Y. Information potential fields navigation in wireless ad-hoc sensor networks [J]. Sensors, 2011, 11(5):4794-4807.

[2] WEI W,XU Q,WANG L,et al. GI/Geom/1 queue based on communication model for mesh networks [J]. International Journal of Communication Systems, 2014, 27(11):3013-3029.

[3] NG A Y,JORDAN M I,WEISS Y,et al. On spectral clustering:Analysis and an algorithm [C]. Neural Information Processing Systems, 2001:849-856.

[4] ASHBROOK D,STARNER T. Using GPS to learn significant locations and predict movement across multiple users [J]. Personal and Ubiquitous Computing, 2003, 7(5):275-286.

[5] 谢娟英,郭文娟,谢维信. 基于邻域的 K 中心点聚类算法[J]. 陕西师范大学学报:自然科学版,2012,40(4):16-22.

[6] 谢娟英,高瑞. 方差优化初始中心的 K-medoids 聚类算法[J]. 计算机科学与探索,2015,9(8):973-984.

[7] 谢娟英,高瑞. Num-近邻方差优化的 K-medoids 聚类算法[J]. 计算机应用研究,2015,32(1):30-34.

[8] 张伟玲.基于轨迹数据的城市交通需求热点区域推荐研究[D].兰州:兰州交通大学,2017.

[9] 覃晓,梁伟,元昌安,等.基于遗传优化谱聚类的图形分割方法[J].计算机科学,2017,44(1):100-102,133.

Hot Region Mining Algorithm based on Variance Optimization Spectrum Clustering

LIANG Zhuoling[1],YUAN Chang'an[2],QIN Xiao[3]

(1. Guangxi University,Nanning,Guangxi,530004,China;2. Guangxi Academy of Sciences,Nanning,Guangxi,530007,China;
3. Nanning Normal Universety,Nanning,Guangxi,530001,China)

Abstract:In order to improve the traffic congestion,this article uses the cluster analysis approach to mine the mobile trajectory data and identify the hot region of residents' travel. The traditional Ng-Jordan-Weiss (NJW) spectral clustering algorithm often uses K-means clustering algorithm to achieve the final clustering operation. However,K-means clustering algorithm has the disadvantages of being sensitive to the initial value and easy to fall into the local optimum,which will affect the mining results of hotspot area. Therefore,the K-medoids clustering algorithm of variance optimization initial center is applied to the final clustering stage of the spectral clustering algorithm,and a Hot Region Mining algorithm based on improved K-medoids Spectral Clustering (HRM-KSC) is proposed,and then experiment on real trajectory data sets. The experiment results find that the HRM-KSC algorithm clustering results have higher silhouette coefficient,which indicates that the HRM-KSC algorithm improves the NJW spectral clustering algorithm and the clustering quality.

Key words:K-medoids algorithm,spectral clustering,hot region,stop point,traffic congestion

责任编辑:米慧芝