文章编号:1009-6744(2020)05-0169-08 中图分类号:U495 文献标志码:A
DOI:10.16097/j.cnki.1009-6744.2020.05.025

基于改进密度聚类算法的交通事故地点聚类研究

黄 钢*,瞿伟斌,许卉莹
(公安部交通管理科学研究所 道路交通安全公安部重点实验室,江苏 无锡 214151)

摘 要: 交通事故特征受地域分布影响显著,本文对交通事故特征进行优化聚类研究.基于2019年无锡市交通事故数据,调用开放地图接口地理编码解算事故地点经纬度,使用密度聚类算法对事故地点与事故原因进行密度聚类.传统的密度聚类算法依赖距离阈值和样本数阈值的准确输入,为解决这一局限,建立一种自适应搜索距离阈值和样本数阈值的密度聚类模型,并与原始聚类模型进行对比.结果表明,优化算法在参数确定上更加智能,对簇的划分更加准确,对噪声点的识别更加合理.通过机器学习中轮廓系数计算方法计算模型得分,证明了该算法在城市道路交通事故地理位置聚类中的适用性.

关键词: 城市交通;交通安全;地理编码;密度聚类;轮廓系数

Traffic Accident Location Clustering Based on Improved DBSCAN Algorithm

HUANG Gang, QU Wei-bin, XU Hui-ying
(Key Laboratory of Ministry of Public Security for Road Traffic Safety, Traffic Management Research Institute of the Ministry of Public Security, Wuxi 214151, Jiangsu, China)

Abstract: Traffic accident characteristics are significantly affected by regional distribution. In this paper, traffic accident characteristics are clustered by the optimized density-based spatial clustering of applications with noise (DBSCAN) clustering method. The 2019 traffic accident data in Wuxi, China is used as a case study. The open map API is used to obtain the longitude and latitude of the accident location as an input for the proposed method. The traditional DBSCAN clustering algorithm normally requires accurate input of the distance threshold and sample number threshold. This paper develops the DBSCAN clustering model with an adaptive search distance threshold and sample number threshold. The comparison results of the proposed algorithm with traditional algorithm show that the optimized algorithm is more intelligent in determining parameters and more accurate in dividing clusters; the recognition of noise points is more reasonable than the traditional algorithm. The applicability of the algorithm in the geographical location clustering for urban road traffic accidents is proved by calculating the model score of the silhouette coefficient in machine learning.

Keywords: urban traffic; traffic safety; geocoding; density clustering; silhouette coefficient

0 引 言

随着我国道路里程、机动车保有量和机动车驾驶人数量的迅速增长,交通事故总量及伤亡人数居高不下.区域交通事故的特征分布与居民生活地域有明显关系[1],在市域范围内,呈现部分区县或某一村镇特定特征的事故多发情况,对交通事故特征地域分布进行研究具有重要的理论意义和实用价值,可为交管部门制定针对性的事故预防对策提供参考.

交通事故地理位置聚类与基于交通事故空间

收稿日期:2020-05-30 修回日期:2020-08-05 录用日期:2020-08-09
基金项目:公安部科技强警基础工作专项项目/Fundamental Research Funds for Basic Work of Strengthening the Police by Science and Technology of the Ministry of Public Security(2019GABJC24);中央级公益性科研院所基本科研业务费专项资金/ Fundamental Research Funds for Central Public Welfare Research Institute(2020SJA01).
作者简介:黄钢(1991-),男,湖北荆州人,助理研究员. *通信作者:hgtmri@126.com

特征的事故多发点段鉴别具有相同的含义.常用的交通事故多发点段识别方法有:累计频率曲线法[2],用频率曲线表示某个地点交通事故发生的次数;临界率法[3],在给定一个最高事故率的情况下,某一路段事故率超过该值则认定为事故多发段;回归分析法[4],应用logistic回归模型将不同交通流特征与交通事故发生的可能性关联起来,预测事故多发点段.这些方法在交通事故多发点段识别上都已相对成熟,但存在一个共同的问题,即没有考虑事故地理位置信息,研究的是具体某条路上的事故多发区域.学者使用神经网络聚类方法进行基于GIS技术的事故多发点段识别[5],赋予不同道路特征参数不同权重,将所有权重加和后进行评价以获取事故多发点段信息,但存在定位信息不准确问题.

本文旨在通过交通事故信息采集项中记录的事故地点信息寻找交通事故热点区域,密度聚类方法更加适用.基于密度聚类的事故分析方法一般多用于刑侦领域,用于犯罪热点区域研究[6].实际上交通事故的发生与交通参与者的居住地高度关联,区域交通事故往往呈现部分地区集中的情况,特定特征的交通事故(如酒驾醉驾、超速行驶等)尤为明显.应用密度聚类方法将事故地点进行分类划分,可便于交管部门对本区域交通事故进行精细化管理及制定针对性事故预防对策.

1 问题描述与建模

1.1 交通事故地点的经纬度计算

数据来源于2019年无锡市人员伤亡或财产损失事故,部分数据如表1所示.首先需从事故信息中获取事故地点,我国现行的道路交通事故信息采集项中,事故地点记录的是事故发生的地理位置,并未采集经纬度信息.故对事故地理位置进行分析时,需应用地理编码将文字描述的事故地点转换为便于计算的经纬度数据.国内主流的在线地理编码服务由百度地图、高德地图、搜狗地图和腾讯地图等提供[7].

表 1 无锡市2019年原始事故数据(部分)

Table 1 Original accident data of Wuxi in 2019 (partial)

路 名	事故地点	事故原因	职 业	交通方式
港羊线	东港镇港羊路新港路路口(港羊线)	其 他	其 他	驾驶汽车
上伟路	无锡市惠山区上伟路香缇路口	违反交通信号(非)	其 他	驾(驭)非机动车
八文线	锡山区锡北镇八文线变电所路段	醉酒驾驶	职 员	驾驶汽车
硕放街道	硕放街道南星路LD03路灯杆路段	无证驾驶	工 人	驾摩托车
342省道	342省道146 km 300 m	其他操作不当	工 人	驾驶汽车

本文选用百度地图和高德地图地理编码服务,分别调用两者的API,提取百度地图API和高德地图API返回数据,即可获得事故地点的经纬度,两者对同一地点的地理编码如表2所示.参照文献[7]对这两种地图的服务质量进行分析,结果如表3所示.可以看出,高德地图对事故信息中录入的事故地点匹配成功率及精确匹配上服务质量更好,百度地图模糊匹配成功率相对较高.本文使用的地理位置均要求精确匹配,故最终选用高德地图作为本文精确地理编码工具;对剩余19.7%的地点采用百度地图进行模糊地理编码,其中,18.6%的事故地点通过模糊匹配成功,仅1.1%的地点未匹配,定位效果良好.

表 2 地理编码结果

Table 2 Geocoded results

事故地点	百度地图		高德地图	
	经度/(°)	纬度/(°)	经度/(°)	纬度/(°)
东港镇港羊路新港路路口(港羊线)	null	null	31.692 44	120.523 9
无锡市惠山区上伟路香缇路口	31.613 7	120.209 8	31.610 01	120.205 6
锡山区锡北镇八文线变电所路段	31.672	120.429 6	31.666 01	120.445 8
硕放街道南星路LD03路灯杆路段	null	null	31.474 22	120.442 1
342省道146 km 300 m	23.553 03	113.502 5	31.587 05	120.151 3

注:null表明未匹配到数据.

表 3　地理编码服务的匹配率

Table 3　Address matching rates of Geocoding

匹配级别	匹配率/%	
	百度地图	高德地图
精确匹配	63.1	80.3
模糊匹配	31.8	17.9
未匹配	5.1	1.8

将无锡市2019年全部事故地点利用上述方法进行地理编码获取经纬度，绘制成地理信息散点图，结果如图1所示，图中，05、06、11、13、14、81、82分别代表锡山区、惠山区、滨湖区、梁溪区、新吴区、江阴市和宜兴市.从图1可以看出，无锡市梁溪区(市中心)事故较为集中，江阴市(县级市)北部也存在事故聚集的地方，宜兴市(县级市)城区事故较为集中，滨湖区因大面积为太湖水域，事故相对较少，其他区县事故特征无法直接从图1中获取.

图 1　事故地点定位地理分布

Fig. 1　Geographical location

1.2 DBSCAN地理坐标聚类模型

DBSCAN(Density-based Spatial Clustering of Applications with Noise)是由Martin Ester[8]等提出的一种基于密度的空间聚类算法，其将具有足够密度数据的区域划分为 k 个不同的簇，并能在具有噪声数据的空间域内发现任意形状的簇.本文记为 $C_j(j=1,2,\cdots,k)$ 表示第 j 个簇，其中，簇定义为密度相连点的最大集合.聚类过程要满足以下两个条件：最大性，对于空间中任意两点 p、q，如果 p 属于簇 C_j，且 p 密度可达 q，则点 q 也属于簇 C_j；连接性，对于同属于簇的任意两点 p、q，它们彼此是密度相连的.DBSCAN算法具有聚类速度快，有效处理噪声点，发现空间中任意形状簇，无需划分聚类个数等优点；但DBSCAN聚类算法的聚类效果高度依赖输入参数——聚类半径和簇内最少样本点数，在高维数据的聚类中，对距离公式

选取非常敏感，存在“维数灾难”.本文研究的交通事故空间数据不是高维数据，各事故点间距离计算并不复杂，选择距离阈值也相对容易，且该方法能较好地体现事故多发地点的特点，故使用该方法对事故地理位置进行分析是合理且有效的.

为分析特定事故特征的聚类结果，对数据集进行划分，选择“酒驾醉驾”“无证驾驶”“未按规定让行”“超速行驶”“机动车闯红灯”“非机动车闯红灯”这6类事故，使用原始DBSCAN对1.1节解出的经纬度点进行聚类.在无任何先验知识的情况下，统一设定输入参数：EPS(距离阈值)为0.015，MinPts(最少样本点)为6.具体参数及划分结果如表4所示.

表 4 原始DBSCAN算法聚类结果

Table 4 Result of original DBSCAN algorithm

事故原因	数据点数	EPS	MinPts	距离求解模型	簇 数	噪声比/%
酒驾醉驾	626	0.015	6	欧式距离	18	41.0
未按规定让行	311				4	85.2
无证驾驶	126				0	100
超速行驶	37				0	100
机动车闯红灯	97				0	100
非机动车闯红灯	108				0	100

从表4可知，仅酒驾醉驾事故和未按规定让行事故聚类成功，酒驾醉驾事故聚集成18类，其中，41%的数据点标记为噪声点，聚类结果如图2所示，图中，圆点表示噪声点，其他符号标记的点为簇内点(下同).未按规定让行事故聚集成4类，噪声点占比达到85.2%，聚类地理图如图3所示.从图3可以看出，有3个簇聚集在宜兴市区，除滨湖区还存在一个事故集中区域外，其他区县的事故均被标记为噪声点，聚类结果比较粗糙.其他几类事故由于数据点过少及输入参数不准确，将所有的点均标记为噪声点，聚类结果不具参考意义.

图 2 酒驾醉驾事故聚类结果

Fig. 2 Cluster result of drunk driving accidents

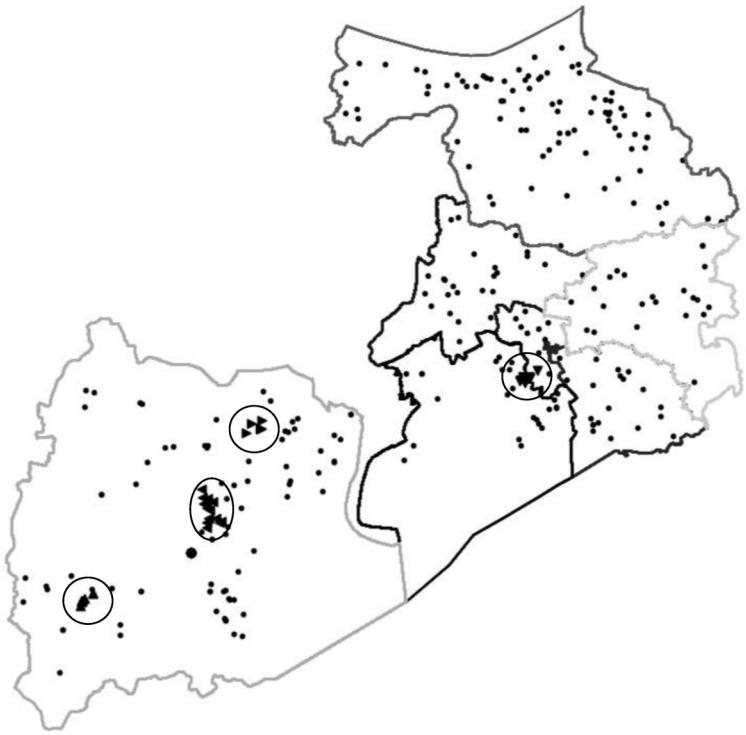


图 3　未按规定让行事故聚类结果

Fig. 3　Cluster result of not to give way accidents

2　自适应参数输入DBSCAN算法

从上述聚类效果来看，DBSCAN聚类算法对输入参数EPS和MinPts非常敏感，尤其是数据点较少的情况，不合适的输入参数可能将所有点标记为噪声点，明显与实际情况相悖. 相关学者对DBSCAN算法进行了改进，Kumar[9]等提出的改进算法加快了高维度下邻域搜索速度，同时指出高维度数据下参数输入存在的问题. 本文继续对DBSCAN算法进行改进，提出一种EPS、MinPts参数自适应选择的A-DBSCAN算法.

2.1　轮廓系数计算方法

在聚类算法中，使用轮廓系数(Silhouette Coefficient)对聚类样本的聚类效果进行评估，轮廓系数的计算模型为

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{1}$$

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{a(i)}{b(i)}-1, & a(i)>b(i)\end{cases} \tag{2}$$

式中：$s(i)$ 为样本 i 的轮廓系数，该值越接近1，说明样本 i 聚类越合理；越接近-1，说明样本 i 更应该分类到另外的簇；越接近0，说明样本 i 在两个簇的边界上. $a(i)$ 为样本 i 到簇内不相似度，为该样本同簇内其他样本的平均距离，该值越小，说明该样本越应被聚类到该簇. $b(i)$ 为样本 i 的簇间不相似度，计算公式为

$$b(i)=\min\left(b_{i1},b_{i2},\cdots,b_{ij}\right) \tag{3}$$

式中：b_{ij} 表示样本 i 到某簇 C_j 所有样本的平均距离.

根据所有样本的轮廓系数计算平均值，即可得到当前聚类模型的总体轮廓系数值，并依据该值确定输入参数.

2.2　基于轮廓系数确定输入参数

针对不同输入参数，模型的轮廓系值越接近1，聚类效果越好. 根据此原理，提出A-DBSCAN聚类算法，流程如图4所示. 首先，根据数据特征确定EPS的步长 L_1 及最大值 $M_{\max,1}$，$M_{\max,1}$ 由任意两个最邻近点距离的最大值确定；确定最少聚类点的步长 L_2 和最大值 $M_{\max,2}$，$M_{\max,2}$ 确定原则为，当

MinPts大于 $M_{\max,2}$ 时，所有点聚集为一个类；构建初始DBSCAN模型，初始距离为步长 L_1，初始最少点数为1，依据2.1节方法计算模型的轮廓系数 S；按照距离步长和点数步长迭代，将计算得到的轮廓系数全部入栈，直到距离参数和点参数达到设定的最大值；根据计算得到轮廓系数的最大值确定最佳EPS值和MinPts值，即为本文提出的A-DBSCAN算法.

3 算法应用与结果分析

应用A-DBSCAN算法对2019年无锡市事故地点经纬度进行聚类(距离求解模型为欧式距离)，计算所有样本轮廓系数平均值作为得分进行评价，聚类结果如表5所示.

从表5可以看出，使用A-DBSCAN算法聚类效果比原始聚类算法(表4)有很大提高，最少聚类簇为6个(无证驾驶事故)，且未出现将大量数据点标记为噪声点的情况.除超速行驶事故外(超速行驶事故是所有特征中样本点最少的)，模型得分均在0.5以上，这表明该模型应用于交通事故地理位置聚类时，除受输入参数EPS和MinPts影响外，还受数据量大小的影响.另外，表5表明，当样本点数据相对较多时，MinPts取值应适当增大；当样本数据点较少时，模型得分相对低一些.

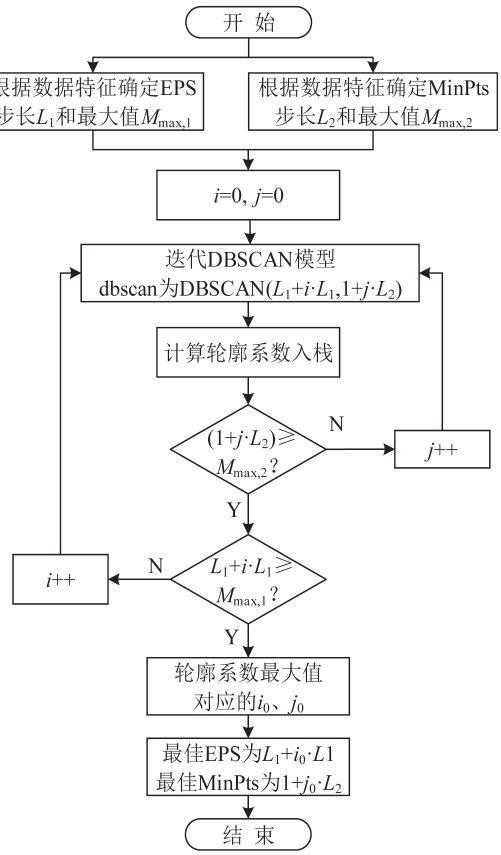


图4 A-DBSCAN聚类算法流程

Fig. 4 Flow chart of A-DBSCAN algorithm

表5 A-DBSCAN算法参数及求解结果

Table 5 A-DBSCAN algorithm results

事故原因	数据点数	EPS	MinPts	簇 数	噪声占比/%	轮廓系数平均值
酒驾醉驾	626	0.015	6	18	41.0	0.696
未按规定让行	311	0.043	6	9	18.6	0.636
无证驾驶	126	0.018	3	8	51.9	0.667
超速行驶	37	0.038	3	6	45.7	0.423
机动车闯红灯	97	0.044	3	11	27.8	0.515
非机动车闯红灯	108	0.034	3	12	27.8	0.627

为验证本文聚类算法先进性，对比其他自适应调参方法对未按规定让行事故进行聚类，结果如表6所示；对比本文A-DBSCAN算法与其他常见聚类算法的聚类结果，如表7所示.

表6 自适应调参算法对比

Table 6 Comparison of adaptive parameter adjustment algorithms

调参方法	时间复杂度	EPS	MinPts	簇 数	噪声占比/%	轮廓系数平均值
A-DBSCAN	$O(n\log n)$	0.043	6	9	18.6	0.636
PID自校正	$O(n)$	0.020	8	4	85.2	0.109
SA-DBSCAN	$O(n^2)$	0.036	6	12	31.8	0.135
递归算法	$O(n^2)$	0.016	9	3	90.7	0.011

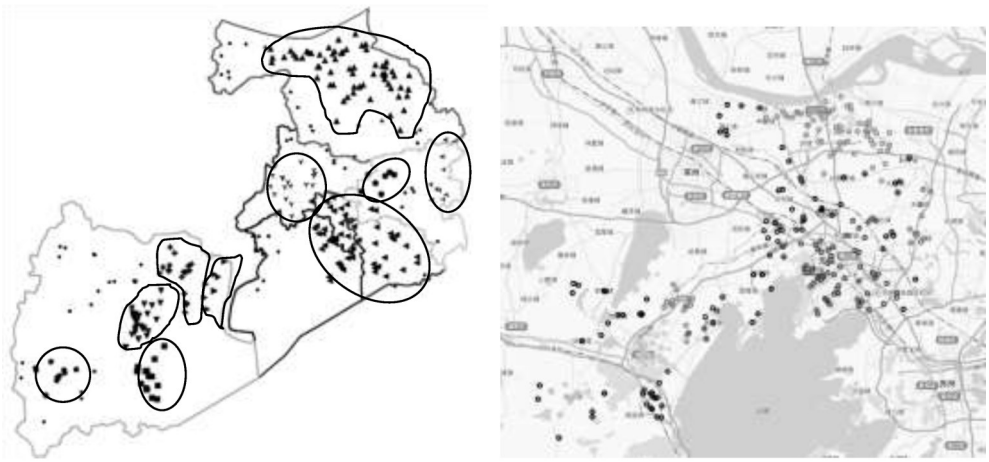
表 7　常见聚类算法对比

Table 7　Comparison of common clustering algorithms

聚类算法	时间复杂度	簇 数	噪声占比/%
A-DBSCAN	$O(n\log n)$	9	18.6
DENCLUE	$O(n^2)$	5	49.7
MEAN SHIFT	$O(n)$	11	21.8
K-MEANS	$O(n)$	2	0.0

从表6可以看出：PID自校正调参算法时间复杂度最低，但噪声占比太高，聚类效果不好；递归调参太复杂，不适用于批量聚类；SA-DBSCAN算法与本文算法有一定的可比性，但本文算法在时间复杂度和模型得分上都优于SA-DBSCAN算法. 从表7可以看出：只有MEAN SHIFT算法对噪声的处理及簇的划分上能与本文算法匹配，但MEAN SHIFT对其他事故类型的聚类存在较大的偏差；K-MEANS算法无法处理含噪声的点集，DENCLUE算法对噪声的处理能力明显偏低；因此，本文算法优势明显.

根据确定参数对6类事故特征进行地理位置聚类. 由于酒驾醉驾事故聚类参数未发生变化，其地理位置聚类结果如图2所示，对照地图可以看出，事故多发区域聚集在市中心全区，太湖国际博览中心附近，镇中公园附近，锡北镇、东港镇、无锡东站附近，申港镇附近，以及桥镇、周铁镇、中心城区、丁蜀镇和张渚镇附近. 未按规定让行事故聚类地理图如图5所示，从图中可以看出，锡山区及滨湖区大部分事故聚集成一个簇，江阴市大部分地区事故亦聚集成簇，宜兴市丁蜀镇、中心城区、万石镇和和桥镇、洋溪镇、张渚镇和西渚镇等几个区域聚集成5个簇，惠山区西北部事故聚集成一个簇，锡山区东港镇和羊尖镇事故聚集成一个簇，共计形成9个事故多发区域. 其他事故特征(无证驾驶、超速行驶、闯红灯)的聚类结果亦有明显区域特征，且事故多发于区域的中心城区和集镇上，囿于篇幅限制，本文不再一一列出进行分析.



(a) 聚类结果　　(b) 聚类地理图

图 5　未按规定让行事故聚类地理分布

Fig. 5　Clustering geographical distribution of not to give way accidents

4　结　论

本文以交通事故空间特征分析为视角，考虑交通事故地点经纬度解算，事故空间特征聚类困难问题，调用在线地图API进行地理编码获取事故地点经纬度，在此基础上，提出一种基于轮廓系数自适应调整原算法输入参数的A-DBSCAN算法，对无锡市不同肇事原因事故数据进行聚类分析，并与原始DBSCAN聚类算法进行比较. 结果表明：事故地点上图率在98%以上，改进算法在参数选择上更加智能，聚类效果更加理想，噪声识别也比原始算法更加合理；A-DBSCAN算法在应用不同数据量进行聚类时，轮廓系数的分值表明聚类效果受数据量大小影响较为显著.

本文分析的数据是二维度的，即仅考虑了事故地点(经纬度)与事故肇事原因的聚类效果，交通事故多维特征的地理聚类分析需进一步研究.

参考文献：

[1] KIM D K. The study of reflecting regional characteristics in car insurance for reduction of traffic accidents[J]. Journal of Korean Society of Transportation, 2015, 33 (3): 223-236.

[2] 方守恩, 郭忠印, 杨轸. 公路交通事故多发位置鉴别新方法[J]. 交通运输工程学报, 2001(1): 90-94. [FANG S E, GUO Z Y, YANG Z. A new identification method for accident prone location on highway[J]. Journal of Traffic and Transportation Engineering, 2001(1): 90-94.]

[3] MILLER TED R. Cost and functional consequences of U.S. roadway crashes[J]. Accidents Analysis and Prevent, 1993, 25(5): 593-607.

[4] XU C C, WANG W, LIU P. Identifying crash- prone traffic conditions under different weather on freeways[J]. Journal of Safety Research, 2013(46): 135-144.

[5] MANDLOI D, GUPTA R. Evaluation of accident black spots on roads using geographical information system (GIS)[C]. Map India Conference, Transportation, 2003.

[6] 颜峻, 袁宏永, 疏学明, 等. 用于犯罪空间聚集态研究的优化聚类算法[J]. 清华大学学报(自然科学版), 2009, 49(2): 176-178. [YAN J, YUAN H Y, SHU X M, et al. Optimal clustering algorithm for crime spatial aggregation states analysis[J]. Tsinghua University (Science & Technology), 2009, 49(2): 176-178.]

[7] 田沁, 巩玥, 亢孟军, 等. 国内主流在线地理编码服务质量评价[J]. 武汉大学学报(信息科学版), 2016, 41 (10): 1351-1358. [TIAN Q, GONG Y, KANG M J, et al. A comparative evaluation of online geocoding services in China[J]. Geomatics and Information Science of Wuhan University, 2016, 41(10): 1351-1358.]

[8] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996: 226-231.

[9] KUMAR K, MSHESH K, REDDY A, MOHAN R. A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method[J]. Pattern Recognition, 2016, 58: 39-48.

上接第155页

[7] 孙鹏, 丁宏飞, 廖勇. 城市轨道交通非高峰期开行方案建模与求解[J]. 计算机工程与应用, 2012, 48(28): 26-30. [SUN P, DING H F, LIAO Y. Modeling and solution for optimization of urban rail transit operation scheme in off-peak period[J]. Computer Engineering and Applications, 2012, 48(28): 26-30.]

[8] 李梦, 黄海军. 基于后悔理论的出行路径选择行为研究[J]. 管理科学学报, 2017, 20(11): 1-9. [LI M, HUANG H J. Modeling route choice behavior based on regret theory[J]. Journal of Management Science, 2017, 20(11): 1-9.]