文章编号: 1001-0920(2019)01-0018-07 DOI: 10.13195/j.kzyjc.2017.1540

基于DBSCAN算法的城轨车站乘客聚集特征分析

李晓璐[1], 于昕明[1], 郗艳红[2], 杨晨光[1], 张 溪[3], 张 彭[3], 朱广宇[1†]

(1. 北京交通大学 城市交通复杂系统理论与技术教育部重点实验室，北京 100044；
2. 北京交通大学 土木建筑工程学院，北京 100044；3. 北京交通发展研究院
北京市城市交通运行仿真与决策支持重点实验室，北京 100073)

摘 要: 发掘并掌握站内乘客群体的聚集时空变化规律,对于优化城市轨道交通线网间车辆的调度,特别是优化灾害条件下的客流组织管理等,具有积极的作用. 针对具有密度分布非均匀特征的车站乘客位置数据集,提出一种基于高斯混合模型的DBSCAN聚类算法. 首先,利用高斯混合模型对数据集进行密度的分层处理;然后,面向不同密度层次的数据集进行局部聚类,确定各密度层数据集的参数,并选取恰当的种子以完成局部聚类簇扩展;最后,将各密度层次数据集的聚类结果进行合并. 通过标准和实测数据的计算结果表明,基于高斯混合模型优化后的DBSCAN算法,对于非均匀密度分布的乘客位置分布数据具有更好的聚类效果.

关键词: 城市轨道交通；乘客聚集特征；非均匀分布；高斯混合模型；密度分层；聚类算法

中图分类号: U293.1 **文献标志码:** A

Analysis of passenger aggregation characteristics of urban rail stations based on DBSCAN algorithm

LI Xiao-lu[1], *YU Xin-ming*[1], *XI Yan-hong*[2], *YANG Chen-guang*[1], *ZHANG Xi*[3], *ZHANG Peng*[3], *ZHU Guang-yu*[1†]

(1. MOE Key Laboratory for Transportation Complex Systems Theory and Technology，Beijing Jiaotong University，Beijing 100044，China；2. School of Civil Engineering and Architecture，Beijing Jiaotong University，Beijing 100044，China；3. Beijing Key Laboratory of Urban Traffic Operation Simulation and Decision Support，Beijing Transport Institute，Beijing 100073，China)

Abstract: Exploring and grasping the temporal and spatial variation rules of passenger group's aggregation in the station has a positive effect on optimizing the scheduling of vehicles in the urban rail transit network, especially optimizing the organization and management of passengers under disaster conditions. In this paper, a density based spatial clustering of applications with noise(DBSCAN) clustering algorithm based on the Gaussian mixture model is proposed for the station passenger location data set with non uniform density distribution. Firstly, the Gaussian mixture model is used to process the density of data sets. Then, local clustering is performed on data sets with different density levels to determine the parameters of each density layer data set. The appropriate seeds are selected to expand the local cluster cluster. Finally, the clustering results of each density hierarchical data set are merged. Through the calculation of the standard and measured data, it is illustrated that the DBSCAN algorithm based on the Gaussian mixture model has better clustering effect for the passenger location distribution data with non-uniform density distribution.

Keywords: urban rail transit；passenger aggregation characteristics；non-uniform distribution；Gaussian mixture model；density layering；clustering algorithm

收稿日期: 2017-11-14；**修回日期:** 2018-04-08.

基金项目: 科技部国家重点研发计划项目(2016YFC0802206-2，2016YFB1200203-02)；国家自然科学基金项目(61872037，61572069，61503022，71501011)；中央高校基本科研业务费专项基金项目(2017YJS308，2017JBM301，2017JBM095)；北京市科技计划项目(Z171100004417024)；深圳市交通公用设施建设项目(BYTD-KT-002-2).

责任编委: 阳春华.

作者简介: 李晓璐(1992−), 女, 博士生, 从事群体行为模式挖掘的研究；朱广宇(1972−), 男, 教授, 博士生导师, 从事智能交通系统数据仿真与分析等研究.

†通讯作者. E-mail: gyzhu@bjtu.edu.cn.

0 引　言

城市轨道车站是乘客群体集聚的重要节点. 由于车站内设施设备的布局相对固定,在特定时空范围内,乘客在站内的走行流线会随着乘客出行需求呈现一定的规律性. 发掘并掌握站内乘客群体的聚集时空变化规律,对于优化城市轨道交通线网间车辆的调度、站内相关设施设备的布局,特别是优化灾害条件下的客流组织管理等,都有积极的作用.

目前,关于城轨车站内乘客交通行为的研究主要集中在乘客个体交通行为分析和群体性行为分析两个方面. 文献[1]对站内乘客个体和乘客流的基本交通特性进行建模分析,建立了站台候车及上下车乘客分布特性的相关数学模型;文献[2]建立了一种离散选择模型,研究地铁车站站台的乘客分布特征. 有关群体性交通行为的研究主要涵盖乘客流关系及乘客群体分布特性两个方面,如:乘客群体的滞留和断流、汇聚和分流、干扰和冲突、交替和自组织等相关内容. 文献[3]对站内不同服务设施处的乘客群体交通特性进行分析,建立了乘客流在站台出口瓶颈处的交通特性模型;文献[4]通过量化乘客流流量、密度与速度,分析乘客流的影响因素并揭示了乘客流状态突变的内在机理;文献[5]利用站台乘客速度和密度的参数特性,刻画了乘客群体候车时的分布特性. 上述研究主要是通过现场调查等方式获取乘客的换乘样本数据,基于仿真等技术手段,对城轨车站乘客群体换乘过程中的滞留、断流、汇聚和分流等现象进行统计和分析.

随着信息技术手段的发展,交通行为的数据采集方式也由现场调查等主动采集转变为被动采集,其中,基于GPS数据的交通行为特性研究已较为成熟,但其主要用于探究室外空间的行为特性分析,不涉及占城市居民活动时间超过70%的室内活动行为. 随着以WiFi等为基础的室内定位技术日益成熟,能够准确地获取到人类在室内的位置数据,有关室内空间人群时空行为特性也取得了较多的成果,主要表现为如下两个方面:1)群体移动模式,如文献[6]利用轨道交通网络系统范围内的WiFi信号数据,实现了站内乘客走行时空轨迹的识别,并且能够还原已发生的乘客的站内走行情况;文献[7]根据某酒店展会获取的参观者室内定位数据,借助序列对比分析法得到参观者在不同展台之间的流模式;文献[8]根据获取的博物馆内游客参观序列数据,分析出游客参观的主导序列,并进一步推断参观完所有展品的最优路径. 2)群体分布感知与推断,如文献[9]提出了一种基于贝叶斯网络方法的解决思路,利用室内定位技术获取部分人群的位置数据,从而推断出所有人群在不同空间位置出现的可能性及分布情况. 最后,在人群分布的热点探测方面,文献[10]基于获取用户在室内的时空位置数据,建立了室内人群空间行为特性分析的理论框架,并根据实测得到的某大学学生室内活动位置数据,利用数据挖掘的算法探测出这些学生不同时段在不同楼层的聚集与分布特征.

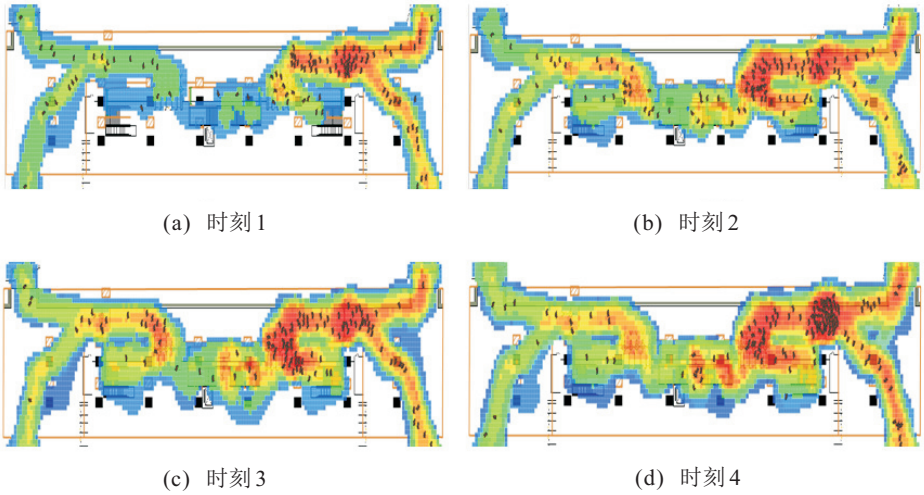
综上所述,基于站内位置数据进行站内乘客群体交通行为特性的研究尚处于探索阶段,其中,基于位置数据进行群体分布感知与推断的相关研究为本文研究城轨车站乘客群体的聚集特征提供了可靠的理论背景支撑. 本文采用DBSCAN算法作为基础算法,通过聚类的方式分析乘客群体在城轨站内的分布情况. 针对站内乘客位置数据集的密度分布非均匀特性,提出一种基于高斯混合模型的DBSCAN聚类算法,通过引入高斯分布的概率密度函数,实现数据集的分层处理,以消除乘客群体密度分布不均匀对聚类效果的影响,并通过簇的快速扩展提升算法的执行效率.

1 面向乘客群体聚集行为的改进型DBSCAN算法

1.1 算法的设计需求及核心思想

乘客走行流线之间的交叉干扰,会导致站内乘客群体走行流线的密度分布不均匀. 图1所示为不同时刻城轨车站乘客群体的位置分布,可以看出其中存在较为明显的密度分布不均匀特性.

对于上述非均匀数据集,传统的DBSCAN算法很难实现聚类;另外,针对具有该类特征的数据集,算法的处理速率通常较低. 为此,人们提出了一系列改进方法,如文献[11-12]从提升簇扩展效率等角度对算法进行了改进,提出应在充分考虑数据集特征的基础上提高算法的执行效率. 因此,本文设计如下解决方案,即:首先采用高斯混合模型,获取站内乘客位置数据集的密度、分布等特征,实现非均匀密度数据集的分层处理;然后,设计一种数据簇的快速扩展处理方法以改进DBSCAN算法,实现对分层后数据集的聚类分析. 算法的核心思想有以下两点:1)面向密度分布不均匀数据集的分层处理;2)簇的快速扩展处理.



(a) 时刻 1 (b) 时刻 2

(c) 时刻 3 (d) 时刻 4

图 1 不同时刻城轨车站乘客群体的位置分布

1.2 基于高斯混合模型的密度分层

用一定数量的高斯分布的线性组合拟合数据集中样本点的分布特征[13],通过EM期望最大化算法(Expectation maximization algorithm)实现模型的参数估计,得到各高斯分布的概率密度函数,针对不同的样本点计算其概率密度函数值,以此度量该样本点的密度分布. 将概率密度函数值相似的样本点划分为同一密度层,以此实现密度分布不均匀数据集的分层处理.

在上述过程中,对高斯混合模型中的未知参数进行估计是数据集分层的核心. 本文利用EM算法进行模型参数的求解.

本文使用的高斯混合模型[14]的概率密度函数形式如下:

$$P(X|\theta)=\sum_{i=1}^{m}\alpha_i p_i(x|\theta_i), \tag{1}$$

$$p(x|\theta_i)=(2\pi)^{-d/2}|\Sigma_i|^{-1/2},$$

$$\exp\Big(-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i)\Big). \tag{2}$$

其中:$\theta=(\theta_i,\alpha_i)^{\mathrm{T}}$,$\theta_i=(\mu_i^{\mathrm{T}},\Sigma_i)^{\mathrm{T}}$,$\mu_i$为第$i$个高斯分布的期望,$\Sigma_i$为第$i$个高斯分布的协方差矩阵,$m$为高斯混合模型的阶数,$d$为向量的维数.

利用EM算法[15]估计高斯混合模型参数的具体步骤如下(通过循环E-Step、M-Step直至最终结果收敛).

E-Step: 由上一M-Step估计得到的参数值来计算隐含变量的值

$$\omega_{ij}^{(t+1)}=p(j|x_i,\theta^{(t)})=\frac{\alpha_j^{(t)}p(x_i|\theta_j^{(t)})}{\sum_{k=1}^{m}\alpha_k^{(t)}p(x_i|\theta_k^{(t)})}=$$

$$\frac{|\Sigma_j^{(t)}|^{-1/2}\exp\Big(-\frac{1}{2}(x_i-\mu_j^{(t)})^{\mathrm{T}}(\Sigma_j^{(t)})^{-1}(x_i-\mu_j^{(t)})\Big)}{\sum_{k=1}^{m}|\Sigma_j^{(t)}|^{-1/2}\exp\Big(-\frac{1}{2}(x_i-\mu_k^{(t)})^{\mathrm{T}}(\Sigma_j^{(t)})^{-1}(x_i-\mu_k^{(t)})\Big)}. \tag{3}$$

M-Step: 根据隐含变量的值来估计未知参数值,即

$$\hat{\alpha}_j^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}\omega_{ij}^{(t+1)}, \tag{4}$$

$$\hat{\mu}_j^{(t+1)}=\frac{\sum_{i=1}^{n}\omega_{ij}^{(t+1)}x_i}{n\hat{\alpha}_j^{(t+1)}}. \tag{5}$$

利用EM算法求解高斯混合模型参数的过程中,E-Step首先利用初始化M-Step得到的$\alpha_j^{(0)}$、$\mu_j^{(0)}$和$\Sigma_j^{(0)}$参数值来估计隐含类别变量$\omega_{ij}^{(t+1)}$的值;然后,将$\omega_{ij}^{(t+1)}$的值代入M-Step中的各个公式重新计算$\alpha_j^{(t+1)}$、$\mu_j^{(t+1)}$、$\Sigma_j^{(t+1)}$各参数值;当取得最大化最大似然估计值时,$\omega_{ij}^{(t+1)}$值需要重新计算. 如此迭代,直至连续两次得到的参数估计值满足算法收敛条件,算法终止.

1.3 不同密度层次数据集的参数确定

经由高斯混合模型分层后的数据集,消除了数据集密度分布不均匀对聚类效果的影响. 可利用传统DBSCAN算法对数据集进行局部聚类,并获取合适的MinPts和Eps参数取值.

在传统DBSCAN算法中,通常根据经验设定阈值参数MinPts = 4;通过遍历各样本点与数据集中所有对象的距离来判断邻域半径Eps参数的取值. 即:在某密度层的数据集中,对任一样本点计算其第k个最近邻距离Dist_k,并将距离值按升序排列,当阈值参数MinPts取值对应多种k值时,可得到如图2所示

的Dist_k分布效果;图3所示则为当阈值参数MinPts = 4时各样本点的Dist_k分布图.

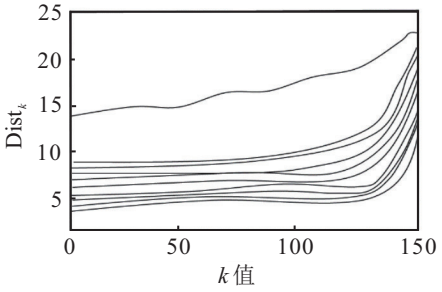


图2　阈值参数MinPts取值未固定时Dist_k分布效果

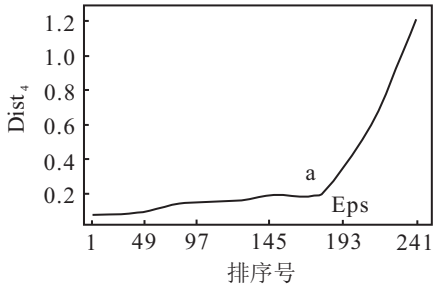


图3　阈值参数$\mathrm{MinPts} = 4$时各样本点的Dist_k分布

由图2和图3可以看出:划分层次后数据集的密度分布曲线并行特征明显;另外,大部分曲线会存在一个拐点. 图2和图3从一定程度上揭示了数据集的密度分布特性. 通常,数据集中的噪音点与类簇应有较大的密度差异,即在Dist_k图中存在某个阈值点,可将数据集的点分为密度差异较大的两部分,该点在图3中表现为Dist_k曲线由缓到陡的突变点a. 该突变点的特点是:突变点之前各样本点的k最近邻距离Dist_k变化较小,而突变点之后各样本点的k最近邻距离Dist_k变化率较大. 即在a点右边的数据点是低密度分布的噪音点,a点左边的数据点是高密度分布的数据点,a点的Dist_4值即为聚类的密度参数Eps,$k = 4$为密度参数的取值[16]. 根据此特点,可将该点对应的Dist_k值作为Eps取值,即确保得到基于密度可达性的最大密度相连对象的集合. 通过此种方法对于密度均匀数据集进行Eps参数确定,具有较好的可行性和有效性,且处理过程相对简单快速.

1.4　类簇的快速扩展

针对传统DBSCAN算法在执行效率方面的缺点,文献[17]提出了一种Fast-DBSCAN算法,在对分层数据集进行聚类的过程中,将邻域内的所有密度可达对象归为同一个簇,然后选取一定数量的代表点作为种子对象进行类簇扩展. 该算法在选取的种子对象数目较少时,会在两轮以上簇扩展过程中出现对象丢失的情况. 本文算法提出,两轮及以上簇扩展时以8个位置点为参照,选取代表点作为种子对象,对于n维空间通常有2^n个象限和$3^n - 1$个参考点,所以种子对象个数最多为$3^n - 1$个. 本文主要针对二维空间进行簇的扩展,因此,任意核心对象的种子对象最多选取为8个是合理的. 其选取方式为:以核心对象p为中心,以其Eps为半径画圆,首先在圆周的最上、最下、最左、最右选取4个点,再依次取上述4个点中每两个点之间圆弧的中点作为另外4个点.

1.5　算法流程

基于高斯混合模型的DBSCAN算法的输入为数据集D、高斯混合模型阶数m及参数初始值、阈值参数MinPts;输出为聚类得到的类或簇. 算法的流程如图4所示,具体步骤如下.

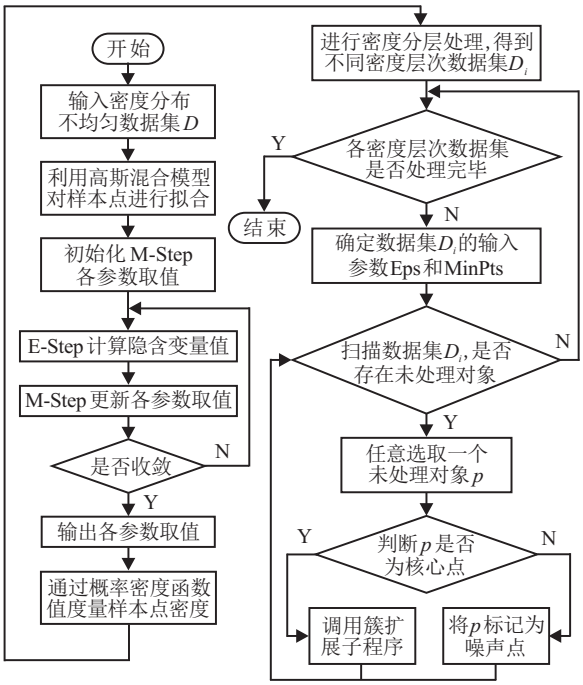


图4　改进算法的流程

Step 1: 确定数据集D,对其中对象进行初始化,将样本点标记为未归类状态.

Step 2: 利用高斯混合模型对数据集D中的样本点进行拟合,借助EM算法求解模型参数.

Step 3: 将高斯混合模型表示为多个高斯分布的线性组合关系式,以各样本点概率密度函数值为依据,将密度分布不均匀数据集进行分层处理,得到不同密度层次的数据集$\{D_1, D_2, \cdots, D_n\}$.

Step 4: 针对不同密度层次的数据集$D_i(1 \leqslant i \leqslant n)$,设定阈值参数$\mathrm{MinPts} = 4$,确定$\mathrm{Eps}_i(1 \leqslant i \leqslant n)$参数取值,依次利用传统DBSCAN算法对数据集$D_i$进行局部聚类.

Step 5: 任意选取数据集D_i内一个未被处理的对象p,查找其关于Eps_i和MinPts密度可达的对象总数,判断该数目与MinPts的大小关系. 若大于等于

MinPts的值,则认为p是核心对象,此时将p的Eps邻域内所有密度可达的对象归为一个簇,使用Step6进行簇的扩展,直到没有点可以加入该簇,于是核心对象p的此次聚类完成;若小于MinPts的值,则p被标记为噪声点. 对象p处理完成后,重复Step 5,直到数据集D_i内没有未被处理的点为止,此时该密度层次的数据集已完成聚类,转入下一密度层次数据集进行分析.

Step 6: 在簇内选取k个代表对象作为种子加入Seeds中,判断Seeds中种子对象s_i是否为核心点. 若不是核心点,则直接将其从Seeds中移除;若是核心点,则将s_i邻域内未归类对象加入簇,并在s_i邻域内继续选取k'个代表对象作为种子加入Seeds′中. 之后对Seeds′中的种子对象进行遍历,将其中未归类种子加入Seeds中作为新的种子对象,此时s_i处理完毕并将其从Seeds中移除. 重复上述过程,直到Seeds中的所有种子对象均处理完,此时核心点p的簇扩展结束,转至下一核心点的簇扩展过程.

Step 7: 当所有密度层次的数据集D_i均完成聚类后,将各个数据集的聚类结果进行合并,即可输出整个数据集D的最终聚类结果.

2 数值实验与结果分析

在本节中,选取UCI数据库中具有较强代表性的3种数据集: Iris、Wine和Poker Hand,分别针对传统DBSCAN算法和本文算法进行对比测试. 3组数据集的情况如表1所示.

表 1 测试数据集

数据集	数据规模	数据维度	类别	数据特征
Iris	150	4	3	小规模低维度
Wine	178	13	3	小规模高维度
Poker Hand	1 025 010	11	10	大规模高维度

引入聚类准确率(Accuracy)[18]这一指标衡量两种算法的聚类效果. 为了保证聚类准确率指标的可靠性,采用多次试验准确率的平均值作为最终的指标值,计算结果如表2所示. 显然,本文算法在不同类型数据集上均具有较好的聚类准确性.

表 2 两种算法的聚类准确率

算法	数据集		
	Iris	Wine	Poker Hand
DBSCAN算法	0.646 7	0.623 8	0.663 2
本文算法	0.873 3	0.791 7	0.810 2

3 实际算例

为进一步验证本文算法的实用性,采用北京市某城轨换乘站的实际数据,分别采用传统DBSCAN算法与本文算法进行聚类分析.

实际数据来源为该换乘站内WiFi定位系统测试过程中,以固定周期(10 s)不断地向数据库服务器上传的乘客位置数据. 经数据预处理后的数据集如表3所示.

表 3 WiFi定位数据预处理后格式

编号	User_ID	TimeStamp	X	Y
1	0017	20160717230110	117.78	8.76
2	0059	20160717230130	86.40	5.78
3	0068	20160717230150	23.20	8.48
4	0094	20160717230210	102.84	8.84
⋮	⋮	⋮	⋮	⋮
495	0397	20160717230120	72.90	3.68
496	0415	20160717230200	113.09	7.89

表3中第1列为数据编号,第2列为移动设备(或乘客)编号,第3列表示数据获取的时间,第4、第5列记录了该时刻乘客的地理位置坐标.

3.1 基于传统DBSCAN算法的站内乘客位置数据的聚类分析

利用传统DBSCAN算法对上述乘客位置数据进行聚类分析,其中阈值参数MinPts设定为4. 分别选取不同的Eps值可得到不同的聚类效果.

当$\mathrm{MinPts}=4, \mathrm{Eps}=1$时,算法的聚类效果如图5所示. 图中不同颜色的“×”代表不同类别,“○”则代表噪声点. 由图5可知,此时只能实现高密度分布的乘客位置数据的聚类分析,而对于低密度分布的乘客位置数据,则基本无法完成聚类,导致大量乘客位置数据被处理为噪声点,无法呈现乘客群体性换乘行为的聚集与分布特性.

当$\mathrm{MinPts}=4, \mathrm{Eps}=1.5$时,算法的聚类效果如图6所示. 此时,能够满足低密度分布的乘客位置数据的聚类分析,但由于邻域参数Eps取值过大,无法正确识别相近区域内乘客簇的聚集与分布特征,导致高密度分布的乘客位置数据被聚类为同一类.

由此可见,利用传统DBSCAN算法对密度分布非均匀的乘客位置数据进行聚类分析,若参照高密度分布的乘客位置数据选取全局Eps参数值,则会导致低密度分布的乘客位置数据无法实现聚类,并产生大量噪声点;若参照低密度分布的乘客位置数据选取全局Eps参数值,则会导致高密度分布的乘客位置数据被合并为同一类簇,影响聚类结果的准确性.

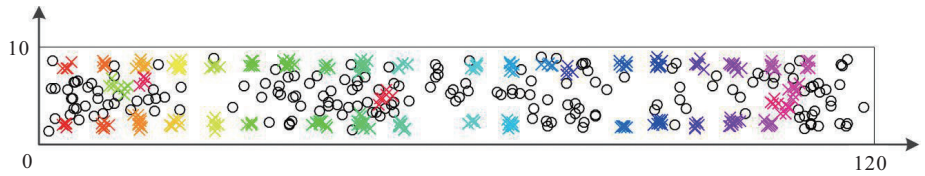


图 5　(DBSCAN算法对乘客位置数据聚类效果(MinPts = 4, Eps = 1.0)

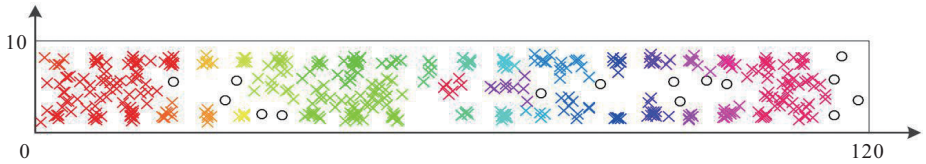


图 6　(DBSCAN算法对乘客位置数据聚类效果(MinPts = 4, Eps = 1.5)

3.2　基于改进型DBSCAN算法的站内乘客位置数据的聚类分析

利用本文提出的基于高斯混合模型的DBSCAN算法,面向上述数据集进行聚类测试.

采用高斯混合模型对测试乘客位置数据集进行拟合,设定模型阶数为$m = 4$,由式(1)和(2)得到概率密度函数关系式,拟合此刻站台上乘客位置数据集的分布特征. 拟合过程中采用EM算法参数求解流程,对该高斯混合模型中的未知参数$\theta_i = \{\alpha_i, \mu_i, \Sigma_i\}$ $(1 \leqslant i \leqslant 4)$进行估计. 各参数的估计结果如表4所示.

表 4　高斯混合模型参数估计结果

高斯分支	混合比	均值(x, y)	协方差矩阵[2×2阶]
1	0.300 3	(13.429 9, 5.469 0)	[25.753 8　−10.762 9; −10.762 9　14.645 3]
2	0.186 4	(46.904 6, 3.790 5)	[25.753 8　0.162 3; 0.162 3　24.645 3]
3	0.232 2	(73.201 7, 6.406 8)	[16.855 6　−0.897 4; −0.897 4　17.021 6]
4	0.281 1	(106.489 7, 4.730 9)	[30.253 9　15.456 3; 15.456 3　16.102 6]

较多测试乘客所在的位置一定程度上决定了各单高斯分布均值的范围,同时乘客位置数据点被拟合到相应的单高斯分布中,而剩余的测试乘客位置数据则被拟合到最合适的某个单高斯分布中.

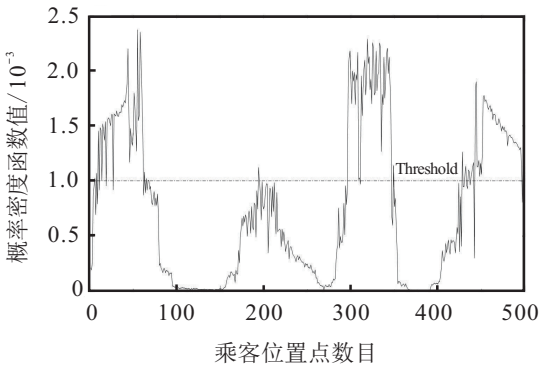


图 7　乘客位置数据密度层次划分示意

设定合适的阈值[19]对数据集进行层次划分,可得到针对该数据集的分层结果,如图7所示.

依据单一密度层次数据集的参数确定方法,得到数据集中第一密度层次数据的聚类参数MinPts = 4, Eps = 1,第二密度层次数据的聚类参数MinPts = 4, Eps = 1.5. 按照Eps_i从小到大的取值顺序依次进行聚类,即:优先聚类较高密度层次的乘客位置数据,再对较低密度层次的数据进行聚类,将完成聚类的簇标记为C_{ij}(密度层次为D_i的第j个簇),以保证完成聚类的样本点不会再被重复聚类. 当各个密度层次的数据都完成聚类后,没有被聚类的数据点被标记为噪声点. 合并不同密度层次数据的局部聚类结果,即可得到该时刻换乘站站台上乘客群体行为的聚集与分布特性. 如图8所示.

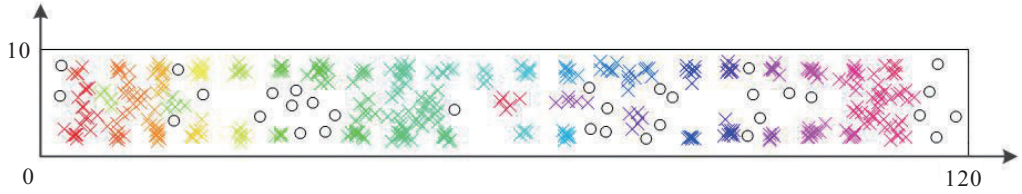


图 8　改进型DBSCAN算法对乘客位置数据聚类效果(MinPts = 4; $\mathrm{Eps}_1 = 1.0$, $\mathrm{Eps}_2 = 1.5$)

相比于图5、图6所示的聚类效果,改进型算法对高密度分布、低密度分布的乘客位置数据集都有更好的聚类效果,同时避免了大量乘客位置数据被误处理为噪声点,从而克服了传统算法对相近区域内乘客

簇合并失效的缺陷.

4 结 论

本文研究了城轨车站的乘客聚集行为,针对具有密度分布非均匀特征的换乘站乘客位置数据集,设计了一种基于高斯混合模型的DBSCAN算法,通过引入高斯分布的概率密度函数实现数据集的分层处理,消除了密度分布不均匀对聚类效果的影响,并采用簇的快速扩展改善了算法执行效率低的缺点. 通过对标准和实测数据的计算,验证了本文算法的有效性和实用性.

关于算法中如何自动获取不同密度层次数据集的参数,进一步提升算法的聚类效率等问题,是本文下一步的研究方向.

参考文献(References)

[1] 曹守华. 城市轨道交通乘客交通特性分析及建模[D]. 北京:北京交通大学交通运输学院, 2009.
(Cao S H. Analysis and mdeling on pssengers traffic characteristics for urban rail transit[D]. Beijing: School of Traffic and Transportation, Beijing Jiaotong University, 2009.)

[2] 宋庆梅, 吴非, 袁振洲. 城市轨道交通站台的乘客分布特性分析及建模[J]. 城市轨道交通研究, 2011, 14(9): 43-47.
(Song Q M, Wu F, Yuan Z Z. Urban rail transit station platform passenger distribution character[J]. Urban Mass Transit, 2011, 14(9): 43-47.)

[3] 李灿. 城市轨道交通枢纽乘客流交通特性分析及建模[D]. 北京: 北京交通大学交通运输学院, 2008.
(Li C. Study and model of the characteristic of passengers flow in urban railway transit hub[D]. Beijing: School of Traffic and Transportation, Beijing Jiaotong University, 2008.)

[4] 张琦, 韩宝明. 城市轨道交通车站乘客群体行为特征研究[J]. 城市交通, 2010, 8(4): 41-46.
(Zhang Q, Han B M. A study on characteristics of passenger collective behaviors on urban rail transit stations[J]. Urban Transport of China, 2010, 8(4): 41-46.)

[5] 许阳. 城市轨道交通车站站台候车乘客分布规律研究[D]. 北京: 北京交通大学交通运输学院, 2015.
(Xu Y. Research on the spatial distribution of waiting passengers on the platform of a rail transit station[D]. Beijing: School of Traffic and Transportation, Beijing Jiaotong University, 2015.)

[6] 李思杰, 朱炜, 黄兆东. 基于WIFI数据的城市轨道交通乘客出行时空轨迹推定[J]. 华东交通大学学报, 2017, 34(2): 85-92.
(Li S J, Zhu W, Huang Z D. Travel time-space trajectory characterization of urban rail transit network based on WIFI data[J]. J of East China Jiaotong University, 2017, 34(2): 85-92.)

[7] Delafontaine M, Versichele M, Neutens T, et al. Analysing spatiotemporal sequences in Bluetooth tracking data[J]. Applied Geography, 2012, 34(1): 659-668.

[8] Yoshimura Y, Sobolevsky S, Ratti C, et al. An analysis of visitors' behavior in the louvre museum: A study using bluetooth data[J]. Environment & Planning B Planning & Design, 2014, 41(6): 1113-1131.

[9] Liebig T, Andrienko G, Andrienko N. Methods for analysis of spatio-temporal bluetooth tracking data[J]. J of Urban Technology, 2014, 21(2): 81-91.

[10] Petrenko A, Sizo A, Qian W, et al. Exploring mobility indoors: An application of sensorbased and GIS systems[J]. Trans in Gis, 2014, 18(3): 351-369.

[11] Xiong Zhongyang, Chen Ruotian, Zhang Yufang, et al. Multi-density DBSCAN algorithm based on density levels partitioning[J]. J of Information and Computational Science, 2011, 9(10): 2739-2749.

[12] 许芳芳. 基于DBSCAN优化算法的Web文本聚类研究[D]. 上海:华东师范大学信息科学技术学院, 2011.
(Xu F F. The research on web text clustering based on DBSCAN optimized algorithm[D]. Shanghai: School of Information Science Technology, East China Normal University, 2011.)

[13] Liu Zunxiong, Chen Ying, Tian Shanshan, et al. Multivariate Gaussian mixture model based clustering with truncated and censored data[J]. J of Information & Computational Science, 2015, 12(2): 775-785.

[14] Nikos V, Ar Istd Is L. A greedy algorithm for Gaussian mixture learing[J]. Neural Processing Letters, 2002, 15(1): 77-87.

[15] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. J of the Royal Statistical Society, Series B: Methodological, 1977, 39(1): 1-38.

[16] 王实美. 基于DBSCAN的自适应非均匀密度聚类算法研究[D]. 北京: 北京交通大学电子信息工程学院, 2017.
(Wang S M. Research on adaptive varied density clustering algorithm based on DBSCAN[D]. Beijing: School of Electronic and Information Engineering, Beijing Jiaotong University, 2017.)

[17] Zhou Shui-geng, Zhou Ao-ying, Jin Wen, et al. FDBSCAN: A fast DBSCAN algorithm[J]. J of Software, 2000, 11(6): 735-744.

[18] 许合利, 牛丽君. 基于层次与密度的任意形状聚类算法[J]. 计算机工程, 2016, 42(7): 159-164.
(Xu H L, Niu L J. Arbitrary shape clustering algorithm based on hierarchy and density[J]. Computer Engineering, 2016, 42(7): 159-164.)

[19] 谢江. 针对非均匀密度环境的DBSCAN自适应聚类算法的研究[D]. 重庆:重庆大学计算机学院, 2015.
(Xie J. A self-adaptive density-based clustering algorithm for discovering density varied clusters[D]. Chongqing: College of Computer Science, Chongqing University, 2015.)

(责任编辑: 李君玲)