文章编号：2095-2163(2022)05-0114-05 中图分类号：U293.13 文献标志码：A

基于 POI-K-means 地铁车站聚类方法研究

赵 源[1,2]，王 越[3]，胡 华[3]

（1 同济大学 道路与交通工程教育部重点实验室，上海 201804；2 上海轨道交通运营管理中心，上海 200070；
3 上海工程技术大学 城市轨道交通学院，上海 201620）

摘 要：为了更好地研究不同类别车站特性及细分问题，地铁车站的精细化分类就显得尤为重要，本文构建了利用 POI(Point of Interest)数据、车站属性、客流特征三类指标基于 K-means 聚类算法进行车站分类的 POI-K-means 车站聚类模型，并对上海 14 条线、共计 416 座车站，分为 6 类，验证了模型的实用性。同时，还对每类车站进行了特征研究，可以为客流研究与预测、地铁车站管理以及周边土地开发提供依据。

关键词：POI；K-means；车站分类

Research on clustering method of metro stations based on POI-K-means

ZHAO Yuan[1,2], WANG Yue[3], HU Hua[3]

(1 Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China;
2 Shanghai Metro Operation Management Center, Shanghai 200070, China;
3 School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In order to better study the characteristics and subdivision of different types of stations, it is particularly important to classify subway stations. The POI-K-means station clustering model for station classification is carried out by the class algorithm, and a total of 416 stations in Shanghai's 14 lines are divided into 6 categories. The practicability of the model is verified, and the characteristics of each type of station are studied, which can provide a basis for passenger flow research and prediction, subway station management and surrounding land development.

【Key words】 POI; K-means; stations classification

0 引 言

随着城市轨道交通的快速发展，车站的数量也在迅速增长，截止 2020 年底，全国轨道交通累计投运车站共计 4 681 座[1]，不同类型的车站客流特征不同，管理方式不同。也有些研究要基于车站的分类，例如在研究客流时间分布特征时需要将车站准确分类才能总结出不同类型车站的客流系数[2]。因此研究车站分类以及建立车站分类模型可以为客流特征研究与预测、地铁车站管理以及周边土地开发提供依据。

马壮林等人[3]采用主成分分析(PCA)方法对轨道交通进出站客流进行特征提取，采用 Hopkins 统计量分析聚类趋势并探讨聚类数量确定方法，采用 CH 系数、轮廓系数和 DB 指标对比分析高斯混合模型(GMM)和 K-means 聚类的优劣，目前大多数分类方法包括：按车站所处的城市位置，分为都市中心站、交通枢纽站等；按场所导向型标准，分为城市外围区、成熟居住区等；按功能导向型标准，分为起点站、换乘站、终点站等；按运营性质，分为中间站、区域站；按车站交通重要性，分为二线换乘、三线换乘等[4]。既有分类方法稍显简单，标准较单一，可能导致一个车站属于多个类别的情况。

为了得到车站的精细化分类，本文总结了影响车站分类的因素：车站自身属性，即是否为起/终点站或者是几线换乘站、客流特征，即早晚高峰时段 5 min 粒度客流占全天客流的比重、POI 特性，即地铁车站 800 m 范围内土地利用情况。构建了 POI-K-means 车站聚类模型并将上海 14 条线、共计 416 座车站，分为 6 类，验证了模型的实用性。

1 车站分类影响因素

1.1 车站属性

车站是轨道交通线网的重要节点，由于在线路

作者简介：赵 源(1978-)，男，博士，高级工程师，主要研究方向：轨道交通运营安全；王 越(1997-)，男，硕士研究生，主要研究方向：轨道交通智慧运营；胡 华(1979-)，女，博士，教授，主要研究方向：轨道交通运营管理。

通讯作者：胡 华 Email：huhua1979@126.com

收稿日期：2020-12-08

中的位置不同，功能不同，所以在确定车站属性类聚类指标时，选取了起点站、终点站、非换乘站、二线换乘站、三线换乘站、四线换乘站 5 个指标，输入数值为 0，1 型，是为 1，否为 0。详见表 1。

表 1　车站属性聚类指标

Tab. 1　Stations attribute clustering index

指标	标号	数值类型
起/终点站	A_1	0，1
非换乘站	A_2	0，1
二线换乘站	A_3	0，1
三线换乘站	A_4	0，1
四线换乘站	A_5	0，1

1.2　车站客流特征

相比道路流量、公交客流量，城市轨道交通客流量有很大的不同，由于城市轨道交通有着固定的发车间隔与营业时间，使得其统计的客流量在不同时间粒度（如 5 min、15 min、30 min、60 min）都可以显示出客流本质特征，要使车站做到精细化的分类，所以选择 5 min 时间粒度，而在全天客流中早晚高峰最具代表性，为了使指标更能代表客流趋势，这里将 5 min 客流与当天进站或者出站总客流的比值作为聚类客流特征指标，其中包括早晚高峰各 2 h 进出站客流各 48 个、共 96 个指标，见表 2。

表 2　客流特征指标

Tab. 2　Passenger flow characteristic index

车站	进站客流标号	出站客流标号	数值类型（客流量/总进站或出站）
07:00 ~ 07:04	B_1	B_{49}	比值
07:05 ~ 07:09	B_2	B_{50}	比值
07:10 ~ 07:14	B_3	B_{51}	比值
…	…	…	…
08:50 ~ 08:54	B_{23}	B_{71}	比值
08:55 ~ 08:59	B_{24}	B_{72}	比值
17:00 ~ 17:04	B_{25}	B_{73}	比值
17:05 ~ 17:09	B_{26}	B_{74}	比值
17:10 ~ 17:14	B_{27}	B_{75}	比值
…	…	…	…
18:45 ~ 18:49	B_{46}	B_{94}	比值
18:50 ~ 18:54	B_{47}	B_{95}	比值
18:55 ~ 18:59	B_{48}	B_{96}	比值

1.3　车站客流吸引范围内 POI

POI（一般作为 Point of Interest 的缩写，也有 Point of Information 的说法），通常称作兴趣点，泛指互联网电子地图中的点类数据，POI 数据目前可通过高德地图或者百度地图等方式获取，主要包含名称、地址、坐标、类别四个属性；源于基础测绘成果、即数字线划地图（Digital Line Graphic，DLG）产品中点类地图要素矢量数据集；在地理信息系统（Geographic Information System，GIS）中指可以抽象成点进行管理、分析和计算的对象。通常情况下，POI 分类一共有 3 级，但是对于分类的个数大同小异。高德地图针对全上海的 POI 分类中，一级分类有 23 个，二级分类有 267 个，三级分类有 869 个。研究中给出部分 POI 分类见表 3。具体的餐饮类别 POI 数据见表 4。表 4 中包含了经纬度等重要信息。

表 3　POI 分类

Tab. 3　POI classification

一级分类	二级分类	三级分类
餐饮服务	餐饮相关场所、中餐厅、外国餐厅、快餐厅等	四川菜、韩国料理、肯德基、甜品店、糕饼店等
购物服务	商场、超级市场、家居建材市场、体育用品店等	便民商店/便利店、家电电子卖场、品牌鞋店、农副产品市场、小商品市场等
…	…	…
公司企业	公司、农林牧渔基地、工厂等	广告装饰、建筑公司、机械电子、商业贸易、渔场、林场等

表 4　车站 POI 指标

Tab. 4　Stations POI indicators

指标	标号	数值类型（个数）
餐饮服务	C_1	数值
风景名胜	C_2	数值
公共设施	C_3	数值
公司企业	C_4	数值
购物服务	C_5	数值
交通设施服务	C_6	数值
金融保险服务	C_7	数值
科教文化服务	C_8	数值
商务住宅	C_9	数值
生活服务	C_{10}	数值
体育休闲服务	C_{11}	数值
通行设施	C_{12}	数值
医疗保健服务	C_{13}	数值
政府机构及社会团体服务	C_{14}	数值
住宿服务	C_{15}	数值
汽车服务	C_{16}	数值

为了更好地统计地铁车站附近 POI 数量，故划分一定范围，对于站点吸引范围，学者认为根据实际

情况取 400 m 到 800 m 之间[5],目前应用较为广泛的是 800 m,以 800 m 为半径画圆为地铁车站的缓冲区域,统计缓冲区内各类 POI 数据的个数作为车站分类的 POI 指标。

在确定 POI 分类指标时,选取对地铁车站影响较大的兴趣点作为车站分类的指标,并且将个别 POI 分类进行了整合或拆分,例如将汽车服务、汽车维修、汽车销售、摩托车服务统一为汽车服务,将“事件活动”、“地名地址信息”、“室内设施”、“道路附属设施”对车站无影响的类别不纳入指标选取中。由表 4 可知,车站附近 POI 数据指标共 16 个。

2 K-means 算法

2.1 算法原理

K-means 聚类算法是由 Steinhaus(1955 年)、Lloyd(1957 年)、Ball & Hall(1965 年)、McQueen(1967 年)分别在各自不同的科学研究领域独立地探讨提出的[6]。K-means 算法、也称作快速聚类法,是一种非监督的聚类算法。对于给定的样本集,按照样本之间的距离大小,将样本集划分为 K 个簇。让簇内的点尽量紧密地连在一起,而让簇间的距离尽量地大。如果用数据表达式表示,假设簇划分为$(C_1,C_2,\cdots,C_k)$,那么最小化平方误差 E 可用如下公式计算求出:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2 \tag{1}$$

其中,μ_i 是簇 C_i 均值向量,有时也称为质心,表达式如下所示:

$$\mu_i=\frac{1}{|c_i|}\sum_{x\in C_i}x \tag{2}$$

聚类过程示例如图 1 所示。图 1(a)表达了初始的数据集,假设 $k=2$。图 1(b)中,随机选择了 2 个 k 类所对应的类别质心,即图中的红色质心和蓝色质心,并分别求取样本中所有点到这 2 个质心的距离,再标记每个样本的类别为和该样本距离最小的质心的类别,见图 1(c),经过计算样本和红色质心与蓝色质心的距离,得到了所有样本点的第一轮迭代后的类别。此时标记为红色和蓝色的点分别求其新的质心,见图 1(d),新的红色质心和蓝色质心的位置已经发生了变动。图 1(e)和图 1(f)重复了图 1(c)和图 1(d)的过程,即将所有点的类别标记为距离最近的质心的类别并求得新的质心。最终得到的 2 个类别见图 1(f)。

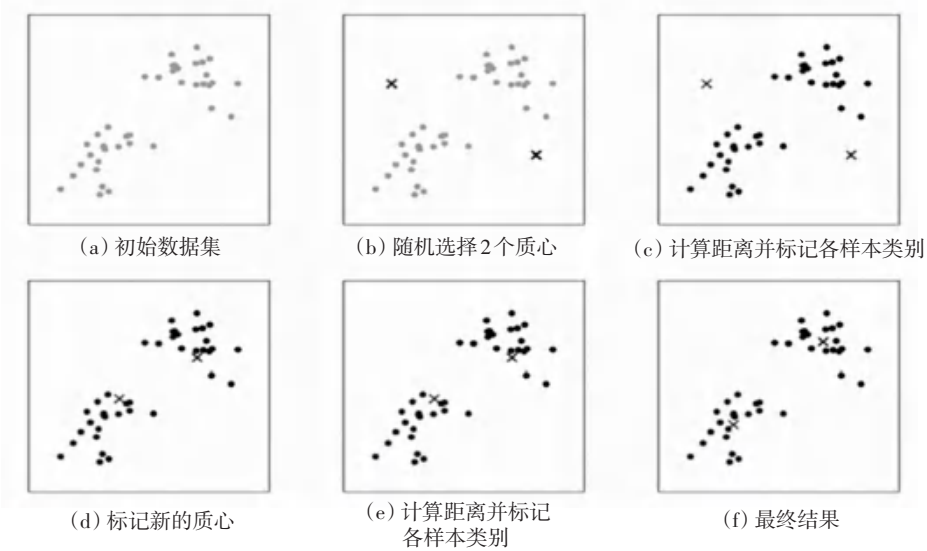

图 1 聚类过程示例

Fig. 1 An example of clustering process

总地来说,K-means 算法步骤为:

Step 1 选择 K 个聚类的初始中心。

Step 2 对任意一个样本点,求其到 K 个聚类中心的距离,将样本点归类到距离最小的中心的聚类,如此迭代 n 次。

Step 3 每次迭代过程中,利用均值等方法更新各个聚类的中心点(质心)。

Step 4 对 K 个聚类中心,利用 Step2、Step3 迭

代更新后，如果位置点变化很小（可以设置阈值），可判定为达到了稳定状态，迭代结束。对不同的聚类块和聚类中心可选择不同的颜色标注。

2.2　基于 POI-K-means 地铁车站聚类模型

在分类过程中，最主要的是对分类指标的选取，本研究分类指标共包含 3 个部分，分别是：车站属性指标、车站客流特征指标以及车站附近 POI 数据指标。

在选取完车站聚类指标后，形成的初始矩阵见表 5。由于指标数值的类型和单位不同，而且数值差距过大，故将矩阵归一化，归一化方法对 K-means 聚类的有效性也通过各种数值实验证明，基本上是 Z-Score、Min-Max 和小数缩放方法。实验分析表明，Z-Score 在 3 个归一化过程中表现良好，准确度更高，因此该方法减少了迭代次数[7]。所以本模型使用 Z-Score 标准化，将变量统一转化为同一个量级，可以将数据有效地转换为统一的标准，Z-Score 的数学公式可写为：

$$z = \frac{x - \mu}{\delta} \tag{3}$$

其中，μ 为总体数据的均值；δ 为总体数据的标准差；x 为个体的观测值。

Z-Score 最突出的优点就是简单，容易计算，能够应用于数值型的数据，并且不受数据量级的影响，因为其作用就是消除量级给分析带来的不便。但是需要指出的是，Z-Score 本身没有实际意义，具体的现实意义需要在比较中得以实现，这也是 Z-Score 的缺点之一。

表 5　车站初始矩阵

Tab. 5　Stations initial matrix

车站	A_1	A_2	…	A_5	B_1	B_2	B_3	…	B_{96}	C_1	C_2	C_3	C_4	…	C_{16}
徐泾东	1	1	…	0	45	52	12	…	22	68	10	84	451	…	35
陆家嘴	0	1	…	0	34	73	8	…	115	49	8	108	761	…	43
世纪大道	0	0	…	1	44	44	9	…	46	67	11	61	543	…	42
浦东国际机场	1	0	…	0	57	25	11	…	330	80	4	83	596	…	62
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

手肘法[8]是一种利用 *SSE* 和 *K* 值的关系图确认最优 *K* 值的方式，*SSE* 还可以替换为样本点到聚类中心欧式距离平均值，本文选用 *SSE* 利用手肘法确定最佳 k 值。在 K-means 算法中，最主要的步骤就是确定 k 值，每一步都可以计算出 *loss* 值、又称为 *SSE*。*loss* 值的计算方式就是每个聚类的点到其质心的距离的平方，如式(4) 所示：

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} | x - \mu_i |^2 \tag{4}$$

指定一个 *Max* 值，即可能的最大类簇数。然后将类簇数 *K* 从 1 开始递增，一直到 *Max*，计算出 *Max* 个 *SSE*。根据数据的潜在模式，当设定的类簇数不断逼近真实类簇数时，*SSE* 呈现快速下降态势，而当设定类簇数超过真实类簇数时，*SSE* 也会继续下降，但下降会迅速趋于缓慢。通过画出 $k-SSE$ 曲线，找出下降途中的拐点，即可较好地确定 k 值。

利用 Python 编程实现确定 k 值与分类的部分，总的分类模型如图 2 所示。

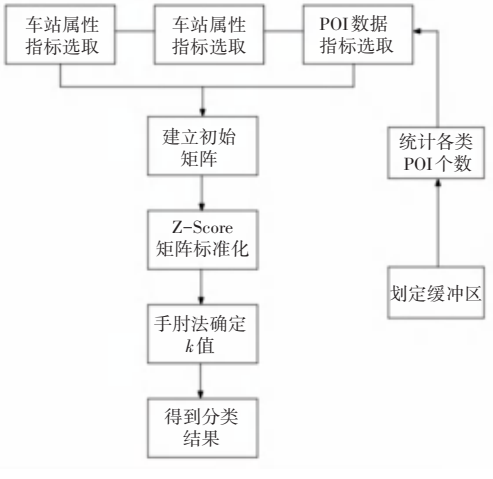

图 2　分类流程图

Fig. 2　Classification flow chart

3　实例验证

3.1　上海地铁车站分类

上海城市轨道交通线网截止 2020 年底共有运

营车站 430 座,本次研究选取运营时间较长的 416 座,其中包括 1 号线、2 号线、3 号线、4 号线、5 号线、6 号线、7 号线、8 号线、9 号线、10 号线、11 号线、12 号线、16 号线、17 号线。

基于 AFC 数据、上海 POI 数据,分别确定车站属性、车站附近 POI 数据、早晚高峰客流特征三类指标进行聚类,如图 3 所示,利用手肘法得到最佳 k 值,在 $k=6$ 时出现明显的拐点,所以将上海地铁车站分为 6 类,聚类结果见表 6。

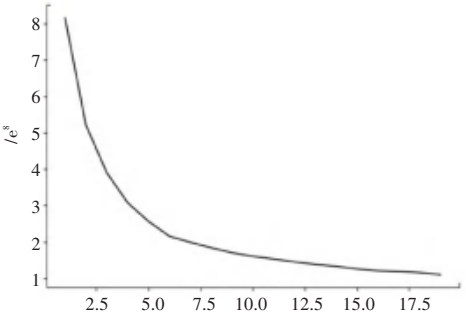

图 3 上海地铁车站分类 k 值手肘图

Fig. 3 The elbow diagram of the k-value of Shanghai subway stations classification

表 6 K-means 聚类结果

Tab. 6 K-means clustering results

类别	车站
1	陆家嘴、张江高科、南京西路等
2	豫园、迪士尼、上海科技馆等
3	松江大学城、上海大学等
4	上海火车站、虹桥火车站、上海南站、浦东国际机场等
5	徐泾东、花木路、虹口足球场等
6	巨峰路、灵岩南路、凌空路等

3.2 各类车站特征分析

根据统计每个类别 POI 个数,分析其土地利用特点以及客流特征,得到以下类型描述。

(1)商务型:地铁车站周边用主要有办公楼、密集的公司、少量的住宅和商户,地面大部分建筑为高层办公楼,土地开发强度高,土地利用率高,高峰时期的交通较为复杂,接驳方式众多,POI 类别中商务写字楼占比最多。

(2)休闲旅游型:地铁车站周边多为景区、音乐厅、体育场、公园等公共场所及建筑,这种类型涉及土地范围稍广,往往换乘线路比较多,配套商业也较多,土地开发率也相对较高,在节假日客流较多,POI 中餐饮服务、购物服务占比较多。

(3)居住型:地铁车站周边多为住宅,商业用地较少且开发程度已经完成,功能比较单一,早晚高峰客流特征明显,接驳方式多以公交、单车为主。

(4)交通枢纽型:地铁车站以大型客运站、火车站、高铁站、机场为主,该类型往往对地上、地下空间利用范围较广,有一些配套的商业,客流量也较大,接驳方式最为全面,从 POI 占比看交通设施服务类占比最大。

(5)活动型:地铁车站周围以大型场馆为主,在活动期间客流骤增,周边场地大,可容纳大量客流,接驳方式主要为地铁、出租。

(6)混合型:地铁车站周边土地利用复杂,多为住宅及学校、办公,商业用地较多且开发程度较高,潮汐客流特征明显,接驳方式众多,POI 类别中生活服务类占比较多。

4 结束语

本文为了得到车站精细化分类,总结了影响车站分类的因素:车站自身属性,即是否为起终点站或者是几线换乘站、客流特征,即早晚高峰时段 5 min 粒度客流占全天客流的比重、POI 特性,即地铁车站 800 m 范围内土地利用情况。构建了 POI-K-means 车站聚类模型,并将上海 14 条线、共计 416 座车站,分为 6 类,验证了模型的实用性。

参考文献

[1] 中国城市轨道交通协会. 城市轨道交通 2020 年度统计和分析报告[EB/OL]. [2021-04-10]. https://www.camet.org.cn/tjxx/7647.

[2] 刘剑锋,罗铭,马毅林,等. 北京轨道交通网络化客流特征分析与启示[J]. 都市快轨交通,2012,25(05):27-32.

[3] 马壮林,杨兴,谭晓伟,等. 基于客流时间序列的城市轨道交通车站分类[J]. 长安大学学报(自然科学版),2021,41(06):113-126.

[4] 金磊,彭建,柳昆,等. 城市地铁车站分类理论及方法研究[J]. 地下空间与工程学报,2010,6(S1):1339-1342,1375.

[5] HSIAO S, LU Jian, STERLING J, et al. Use of geographic information system for analysis of transit pedestrian access[J]. Transportation Research Record Journal of the Transportation Research Board, 1997, 1604(1):50-59.

[6] 王千,王成,冯振元,等. K-means 聚类算法研究综述[J]. 电子设计工程,2012,20(07):21-24.

[7] DAULA U,ISMAIL B M. A study of normalization approach on K-Means clustering algorithm[J]. International Journal of Applied Mathematics and Statistics™,2013,45(15):439-447.

[8] 吴广建,章剑林,袁丁. 基于 K-means 的手肘法自动获取 K 值方法研究[J]. 软件,2019,40(05):167-170.