文章编号：2095-2163(2021)04-0014-04 中图分类号：U293.13 文献标志码：A

基于 K-means 的上海地铁站点分级研究

赵 源[1,2]

(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)

摘 要：科学合理的地铁站点分级体系能够为车站资源分配及管理模式的选择提供决策依据。通过运用图论知识，结合 Space L 法构建了上海市地铁全网的拓扑模型。并选取图论中度，接近中心性和介数的 3 个重要度指标，结合各站的客流量分级指标及城市中心站点对各站点的辐射影响指标，运用 K-means 聚类算法，建立了涵盖多指标多因素的站点分级体系。

关键词：上海地铁；图论；K-means 聚类；站点分级

Study on the classification of Shanghai metro station based on K-means

ZHAO Yuan[1,2]

(1 Key Laboratory of Road and Traffic Engineering Ministry of Education, Tongji University, Shanghai 201804, China; 2 Shanghai Rail Transit Operation Management Center, Shanghai 200070, China)

【Abstract】 A scientific and reasonable subway station classification system can provide a decision basis for station resource allocation and management mode selection. By using the knowledge of graph theory and Space L method, the topological model of Shanghai metro network is constructed. In addition, three important indexes of degree, proximity to centrality and betweenness are selected, combining with the passenger flow rating index of each station and the radiation influence index of urban center site on each station, the k-means clustering algorithm is used to establish a station rating system covering multiple indicators and factors.

【Key words】 Shanghai Metro; Graph theory; K - means clustering; Site grading

0 引 言

目前，对于轨道交通站点的分级研究主要依据是车站的位置、用地情况以及客流量等因素，结果往往只是单一特征下的几类车站，不能反映一个车站同时属于多种类型的情况[1]。随着上海地铁网络的快速发展，线网网络的拓扑结构越来越复杂，需要更加科学合理的分级体系，来对站点进行分级研究[2]。本文应用图论的方法，将上海地铁复杂的线网用数学模型方法[3]进行描述。通过研究模型准确地发现线网的规律和特点后，引入简单高效的 k-means 聚类方法，综合各指标对上海现有站点进行分类研究。

1 基于图论的地铁车站重要度研究

现实中复杂的地铁线网可以通过图论进行简化描述，将现实中的站点拟化成图论中的节点，站点之间的线路可拟化成边。不同的线路利用不同的颜色进行区分，然后通过复杂线网的 3 个重要指标(节点度值、接近中心性和介数中心性)的值，对线网节点进行描述[4]，从而了解上海各站站点在线网网络中的重要度。

1.1 上海地铁线网模型建立

在图论中，一般设图 $G=(V,E)$，V 和 E 的值分别代表图 G 的顶点数和边数。若图 G 的顶点数和边数都是有限集，则称 G 为有限图，反之为无限图。若图 G 中，节点之间的边有方向，则称图 G 为有向图，否则称为无向图。有向图中节点的度有出度和入度之分。

在研究上海地铁全网的拓扑模型特性时，需要选择合适的网络拓扑抽象方法。常见构建线网拓扑的方法是 Space L 和 Space P 方法。对比两种建模方法：Space L 方法所建的模型能够很好的重现线网的实际拓扑结构，但对一些交通特性如换乘次数、最短路径等体现的不够明显。Space P 方法能够很好的将换乘特性、出行距离等乘客重点关注的特征参数表示出来，但由于模型的过度抽象，无法较好的体现线网的具体拓扑结构。由于本文的研究目的主要为分析站点的拓扑特性，因此选择使用 Space L 方法来进行实际地铁线网的建模。

为了研究图论角度下，上海地铁线网的一般拓扑结构规律及特征，需对线网进行如下简化：

(1)将各站距离都假定为 1。

(2)共线线路合并为一条线路。

作者简介：赵 源(1978-)，男，博士研究生，高级工程师，主要研究方向：轨道交通运营安全。

收稿日期：2020-12-07

(3)线网网络为无向网络,不分上下行。

(4)换乘站当做一个站,忽略换乘通道。

1.2　线网重要度评价指标

运用 Space L 法建立的上海地铁线网拓扑网络是复杂网络。目前,复杂网络中最主要的统计指标有:节点度值、接近中心性以及介数中心值。运用这些指标对线网中某个节点进行描述,从而反映这个节点在网络中的基本特性。

(1)节点度值。节点度的值是指和该节点相关联的边的条数,又称关联度。无向网络中, 节点 i 的度 k_i 定义为与该节点相连接边的数目,网络中所有节点 i 的度 k_i 的平均值,即网络的平均度公式为:

$$\langle \mathrm{k} \rangle = \sum_{i=1}^{n} k_i. \tag{1}$$

依据公式,如果某个节点的度越大,表明与该节点形成连边的数目越多,该节点在网络中的重要性也相对较大。

(2)接近中心性。对于网络中的每一节 v_i, 可以计算该节点到其它节点最短距离的平均长度 L_i。L_i 的倒数即为节点 v_i 的接近中心性,二者的表达式如下:

$$L_i = \frac{1}{n}\sum_{j=1}^{n} d_{ij}.$$

$$CC_i = \frac{1}{L_i} = \frac{n}{\sum_{j=1}^{n} d_{ij}}. \tag{2}$$

依据公式, L_i 值是线网中某个节点至网络中其它节点距离之和。其值越小,说明到网络中其它节点距离较小,节点接近中心性的值就越大,节点 v_i 在网络中空间位置上就相对重要。

(3)介数中心值。介数中心值定义为网络中起始点到终点路径中,所有经过节点 v_i 的最短路径的数目。介数中心性的值,定义为网络中所有节点对之间的最短路径中,经过节点 v_i 的比例之和。其中, g_{st} 为节点 v_s 和节点 v_t 之间的最短路径数目, n_{st}^{v} 为连接节点 s、t 之间最短路径中经过节点 v_i 的最短路数目。如果两节点间不存在路径,此时介数值就为 0。当节点 v_s 和节点 v_t 存在路径时则公式为:

$$BC_i = \sum_{s \neq t} \frac{n_{st}^{v}}{g_{st}}. \tag{3}$$

介数中心性的值从“流量”的角度刻画了该节点在网络中的相对重要程度,值越大说明节点在网络中作为枢纽作用就较大,车站站点在地铁网络中的重要性就越高。

由公式计算出来的部分站点指标数值见表 1。

表 1　部分站点指标数值表

Tab. 1　Numerical table of partial site indicators

编号	站点名称	度	介数	接近中心度
1	枫桥路	1	0	0.127 027
2	大渡河路	1	0	0.131 324
3	娄山关路	1	0	0.141 971
4	宋园路	1	0	0.148 093
5	桂林路	1	0	0.145 203
6	桂林公园	1	0	0.134 084
7	漕宝路	4	0.050 8	0.155 846
8	上海南站	3	0.020 0	0.134 894
9	锦江乐园	1	0	0.118 278
10	曹杨路	3	0.035 8	0.146 393

2　K-means 聚类算法

基于图论建立地铁网络拓扑网络后,可以引入 K-means 算法模型。根据表征地铁线网中车站和区间的局部拓扑重要性指标(度)和全局拓扑重要性指标(节点接近中心性和介数),再结合客流量指标以及上海大型城市中心点站对各站点的辐射影响指标,用 K-means 聚类算法对站点进行聚类分析。

2.1　K-means 算法模型

K-means 聚类是无监督学习的一种聚类方法。聚类算法是针对观测到的数据,根据给定的准则发现它们的共同点,在数据集中寻找“群”[5]。K-means 算法以距离作为数据对象间相似性度量的标准,通常采用欧氏距离来计算数据对象间的距离,欧氏距离的计算公式如下所示:

$$dist(x_i, x_j) = \sqrt{\sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2}. \tag{4}$$

其中, D 表示数据对象的属性个数。

基于 K-means 的车站站点重要度聚类分析基本步骤:

(1)确定分类数 K。从所有样本中,随机选取 K 个对象作为初始的簇中心。

(2)将剩余的点保存到相应的簇中,即计算该点与初始质心间的距离,选取最近的那个质心,并将其存储于该质心所在的簇中。

(3)每个簇的质心进行更新,选择该簇所有点的平均值为新的 k 个质心。

(4)将数据集中所有的点进行新一轮分配。如果所有点的分配结果与上一次一致,即簇的质心不会再发生改变,流程结束。否则,分配完所有的点之

后重新更新每个簇的质心,循环该流程直到所有簇的质心稳定下来为止。

2.2 聚类变量选取

地铁车站在线网中所处区位不同会导致重要度差异较大,聚类指标的选取需结合图论中的指标来反映节点在网络中的重要度。所以,变量可以选用复杂网络的 3 个指标:度、接近中心性以及介数值。除此之外,客流量因素也是影响车站重要度的主要因素[6]。从实际数据量上可知,上海地铁客流量超过 40 万的站点数占比最少,可以归为一类,而日进出站量在 1 万-10 万人/天以下的站点数占比较大,应该在此区间对客流量进行细分(以 1 万人/天为单位划分)。10 万人/天-40 万人/天的车站数随着客流量上升占比变小,以 5 万人/天划分,这样可将全网站点的客流量分为 16 级,客流量分级见表 2。

表 2 客流分级

Tab. 2 Passenger flow classification

等级	客流量(万人/天)	等级	客流量(万人/天)
1	≤1	9	9-10
2	1-2	10	10-15
3	2-3	11	15-20
4	3-4	12	20-25
5	5-6	13	25-30
6	6-7	14	30-35
7	7-8	15	35-40
8	8-9	16	>40

以此分级标准对各站点划分客流等级后,作为一项指标放入模型中。此外,上海是国内的一线大都市,城市中心站点本身的影响力很大,所以在进行站点重要度分级时应考虑城市中心站点是否会对市区段的其它站点产生客流的辐射影响,从而造成同类型站点间重要度的差异。因此,本研究选取上海市的 4 个城市中心站点。其中包括:一个城市主中心(人民广场站),三个城市副中心(世纪大道站、徐家汇站、中央公园站)。利用 Dijkstra 算法求解出所有车站至 4 个城市中心站的最短距离(相邻站间距为 1),由于主中心影响比副中心的影响稍大,设受主中心影响的站点距离权重为 0.4,受副中心影响的站点距离权重为 0.2。例如,由 Dijkstra 算法求出大渡河站到人民广场的距离为 8,到世纪大道、徐家汇的距离分别是 12 和 5。将该值乘以各自的权重后可以作为一项指标放入模型中。

由于各指标数据具有不同量纲及单位,为了使得分类的结果更加合理有效,应先将各项数据进行标准化处理。Z-score 标准化方法是一个分数与平均数的差再除以标准差的过程,可以用式(5)表示:

$$Z = (x - \mu) / \sigma. \tag{5}$$

其中,Z 值即为变量标准化后的数值;x 为实际变量值;μ 为同一类变量的均值;σ 为变量的标准差。

2.3 站点重要度分级

由于地铁实际线网与纯图论下的网络会存在较大的差异,单纯利用求得的图论指标进行计算时,会存在一定的异常点,需要结合实际情况进行处理。如,对一些对外接驳其它交通方式的终点枢纽站指标参数进行适当提高,从而保证站点分级的合理性。为了更加明确区分各类型车站的特征,需要经过多次迭代才能得到最优的车站等级聚类结果。将上海全网地铁站点进行聚类分析后,可将现有站点重要度分为 12 个等级,即 $k = 12$,各等级代表站点见表 3。

表 3 部分车站聚类结果

Tab. 3 Cluster results of some stations

车站	重要度等级	车站	重要度等级
人民广场站	12	打浦桥	6
世纪大道站	11	金科路	5
南京西路	10	上海西站	4
上海火车站	9	广兰路	3
新天地	8	佘山	2
宜山路	7	临港大道	1

对聚类结果进一步分析,并按照各站点实际情况对各类别进行特征归纳排序,可得如下结论:

(1)第一类站点:重要度等级为 1。站点是各线的郊区站,位置比较偏僻,该类站点皆为非换乘站和终点站,吸引的客流量较低,因此重要度最低。

(2)第二类站点:重要度等级为 2 的站点。虽然是各线的郊区段站点且不是换乘站,但位置相对不太偏僻,图论指标中度较小,客流量也偏低,因此重要度相对较低。

(3)第三类站点:重要等级为 3 级,站点为各条线路的终点站及郊区段接入市区段的前段站点。终点站吸引客流范围较广,且可能接驳公交或者其它交通方式,图论指标较小,有少量的通勤客流,重要度比郊区段稍高。

(4)第四类站点:重要度等级是 4 级的站点,为各线郊区段接入市区段的枢纽站点及郊区的换乘站点。由于距市区较近所以受市区段的影响,有大量的通勤客流,因此重要度也略微有所提高。

(下转第 20 页)