

基于 MATLAB 的系统聚类算法分析

李 惠

(东莞城市学院, 广东 东莞 523419)

摘 要 聚类分析法是研究“物以类聚”的一种重要方法, 根据样本或变量间的亲疏程度, 将最相似的对象聚合在一起。本文研究了基于 MATLAB 的系统聚类法中的最短距离法和最长距离法, 依据空气质量方面的 6 个指标将 10 个城市进行了分类。分类结果表明, 最短距离法易将大部分样本聚为一类, 有链接聚合的趋势, 最长距离法能克服最短距离法的这一缺点。

关键词 MATLAB; 系统聚类; 最短距离法; 最长距离法

中图分类号: TP311.13

文献标志码: A

DOI: 10.3969/j.issn.2097-3365.2025.09.001

0 引言

聚类分析法是研究个体分类的一种重要方法, 根据样本或变量之间的亲疏程度, 将最相似的对象结合在一起, 以逐次聚合的方式将样本或变量进行分类^[1]。个体在聚类之前, 能聚成的类的数量与类型均未知, 聚类后, 类中的个体在特征上具有某种相似性, 而不同类之间的个体, 其特征的差异性较大^[2]。随着科学技术的不断进步, 以往仅靠经验进行分类的方法用起来变得越来越困难, 自数学这一有力工具被引入分类学之后, 聚类分析方法变得越来越丰富, 有模糊聚类法、系统聚类法、动态聚类法、密度法等^[3]。本文研究的系统聚类法是当前使用率最高的一种方法, 该方法是通过计算各样本间的距离或相关系数来确定样本间的亲疏程度, 通过亲疏程度的判定完成样本的分类。

MATLAB 软件集成了数值分析、建模和仿真、数据可视化等诸多强大功能^[4], 本文基于 MATLAB 平台, 采用系统聚类法中的最短距离法和最长距离法依据空气质量指标对 10 个城市进行了分类, 并分析了两种方法的优缺点。

1 系统聚类算法

设样品个数为 N , 样本指标项为 M , 在分类过程中, 将每个样品看成是 M 维空间的一个点, 在该空间中定义距离。设原始数据矩阵为:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}$$

其中 x_{ik} ($i=1, 2, \dots, N, k=1, 2, \dots, M$) 为第 i 个样品的第 k 个指标的观测数据, $X_i = (x_{i1} \ x_{i2} \ \cdots \ x_{iM})$ 为矩阵 X 的第 i 行, 代表第 i 个样品 M 个指标的观测值, 任意两个样品 X_i 与 X_j 间的相似性, 可以通过矩阵 X 中的第 i 行与第 j 行的距离来衡量^[5]。

1.1 定义样本距离

令 d_{ij} 表示样品 X_i 与 X_j 的距离, 其值越小表示两个样本接近程度越大, 其值越大表示两个样本接近程度越小。

1. 闵氏距离:

$$d_{ij}(q) = \left(\sum_{a=1}^M |x_{ia} - x_{ja}|^q \right)^{1/q}$$

当 $q=1$ 时为绝对距离:

$$d_{ij} = \sum_{a=1}^M |x_{ia} - x_{ja}|$$

当 $q=2$ 时为欧氏距离:

$$d_{ij} = \left(\sum_{a=1}^M |x_{ia} - x_{ja}|^2 \right)^{1/2}$$

当 $q=\infty$ 时为切比雪夫距离:

$$d_{ij} = \max_{1 \leq a \leq M} |x_{ia} - x_{ja}|$$

当各个样本间的指标观测值相差很大时, 需要对这些指标数据进行标准化处理, 然后用标准化的数据计算距离。

2. 兰氏距离:

$$d_{ij} = \frac{1}{M} \sum_{a=1}^M \frac{|x_{ia} - x_{ja}|}{x_{ia} + x_{ja}}, \quad i, j = 1, 2, \dots, N$$

兰氏距离的优点是能克服各指标间量纲的影响, 缺点是没有考虑各指标间的相关性。

3. 马氏距离:

$d_{ij}^2 = (X_i - X_j)^T \Phi^{-1} (X_i - X_j)$, 其中 Φ 为 M 项指标的协方差矩阵:

$$\Phi = (\sigma_{ij})_{M \times M}$$

$$\sigma_{ij} = \frac{1}{N-1} \sum_{a=1}^N (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j)$$

$$\bar{x}_i = \frac{1}{N} \sum_{a=1}^N x_{ai}, \quad \bar{x}_j = \frac{1}{N} \sum_{a=1}^N x_{aj}$$

$$i, j = 1, 2, \dots, M$$

马氏距离既克服了各指标间相关性的干扰, 又避免了各指标间量纲的影响, 在多元统计分析中有着重要的作用。

任意两个样品 X_i 与 X_j 之间的距离 d_{ij} 计算出来后, 可以得到距离矩阵:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{pmatrix}$$

很明显, D 是一个实对称矩阵, 且 $d_{11} = d_{22} = \cdots = d_{NN} = 0$, 根据 D 可对 N 个样品进行分类。

1.2 聚类方法

计算出距离矩阵 D 后, 将每个样品当成一类, 则 N 个样品共有 N 类, 根据聚类方法将相关的类合并, 再计算新类与其他类间的距离, 根据新的距离矩阵和聚类方法再进行类的合并, 其流程图如图 1 所示。常用的聚类方法有最短距离法、最长距离法、中间距离法、质心距离法、类平均法等, 本文选取最短距离法与最长距离法进行分类比较。

(1) 最短距离法: 设有 N 个样品, d_{ij} 表示第 i 个样品与第 j 个样品的距离, 用 $G_1, G_2, \dots, G_N$ 表示初始类, 并类的原则为: 类与类之间距离最近的两类合并。用 G_{pq} 表示 G_p, G_q 间的距离, 规定 $D_{pq} = \min_{\substack{i \in G_p \\ j \in G_q}} \{d_{ij}\}, p \neq q$, 当 $p=q$ 时, $D_{pq}=0$ 。将样品 p 与样品 q 并为同类, 记 $G_r = \{G_p, G_q\}$, 再计算新类 G_r 与其他类 $G_k (k \neq p, q)$ 的距离, 重复以上步骤至完成所有样品的并类。

(2) 最长距离法: 该方法与最短距离法相反, 合并类的准则是类与类之间距离最长的两类合并, 即:

$$D_{pq} = \max_{\substack{i \in G_p \\ j \in G_q}} \{d_{ij}\}, p \neq q$$

设在某一步将 G_p, G_q 合并为一类, 记为 G_r , 则 G_r 与其他类 $G_k (k \neq p, q)$ 的距离公式:

$$D_{rk} = \max_{\substack{i \in G_r \\ j \in G_k}} \{d_{ij}\} = \max \{ \max_{\substack{i \in G_p \\ j \in G_k}} \{d_{ij}\}, \max_{\substack{i \in G_q \\ j \in G_k}} \{d_{ij}\} \}$$

再找距离最大的合并为一类, 直至所有样品归为一类。

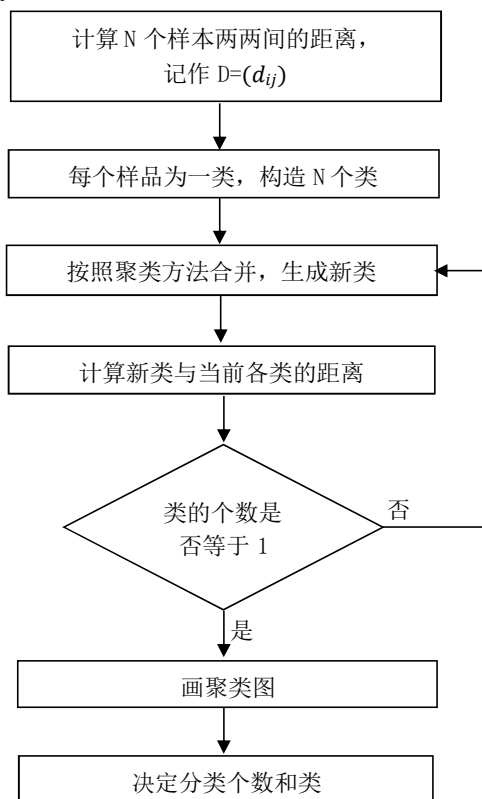


图1 系统聚类流程图

2 实例

根据国家统计局网站的数据, 有表 1 所示的 10 个城市 6 种空气质量指标: (1) x_1 为二氧化硫; (2) x_2 为二氧化氮; (3) x_3 为可吸入颗粒物 (PM_{10}); (4) x_4 为一氧化碳日均值第 95 百分位浓度 (毫克/立方米); (5) x_5 为臭氧 (O_3) 最大 8 小时第 90 百分位浓度 (微克/立方米); (6) x_6 为细颗粒物 ($PM_{2.5}$)。以上指标在未特殊说明情况下, 均指年平均浓度 (微克/立方米)。

将该 10 个城市看成 10 个样本, 根据系统聚类法, 依据 6 个空气质量指标将这 10 个城市进行分类。

在 MATLAB 上将数据做标准化处理后, 得到样本的欧氏距离矩阵:

表 1 城市空气指标观测值

城市						
1 温州	17	50	75	1.7	134	46
2 湖州	22	48	87	1.4	166	64
3 绍兴	36	50	96	1.4	167	63
4 合肥	23	31	113	1.6	69	83
5 芜湖	27	28	96	1.9	72	67
6 湘潭	34	45	108	1.7	131	73
7 岳阳	30	29	130	2.2	74	58
8 常德	36	25	101	2.9	98	71
9 广州	17	48	67	1.5	165	49
10 韶关	33	31	66	2.7	152	49

$$D = \begin{pmatrix} 0.0 & 2.0 & 3.3 & 4.4 & 3.6 & 3.6 & 4.3 & 4.9 & 1.0 & 3.5 \\ 2.0 & 0.0 & 1.9 & 3.5 & 3.2 & 2.3 & 4.0 & 4.5 & 1.7 & 3.7 \\ 3.3 & 1.9 & 0.0 & 4.0 & 3.5 & 1.6 & 3.9 & 4.1 & 3.2 & 3.6 \\ 4.4 & 3.5 & 4.0 & 0.0 & 1.8 & 2.7 & 2.7 & 3.3 & 4.7 & 4.8 \\ 3.6 & 3.2 & 3.5 & 1.8 & 0.0 & 2.5 & 1.9 & 2.4 & 3.9 & 3.3 \\ 3.6 & 2.3 & 1.6 & 2.7 & 2.5 & 0.0 & 2.9 & 3.1 & 3.8 & 3.7 \\ 4.3 & 4.0 & 3.9 & 2.7 & 1.9 & 2.9 & 0.0 & 2.5 & 4.8 & 3.9 \\ 4.9 & 4.5 & 4.1 & 3.3 & 2.4 & 3.1 & 2.5 & 0.0 & 5.2 & 2.9 \\ 1.0 & 1.7 & 3.2 & 4.7 & 3.9 & 3.8 & 4.8 & 5.2 & 0.0 & 3.5 \\ 3.5 & 3.7 & 3.6 & 4.8 & 3.3 & 3.7 & 3.9 & 2.9 & 3.5 & 0.0 \end{pmatrix}$$

采用最短距离法进行聚类,结果如图 2 所示,最终将这 10 个城市分成了两大类:

第 1 类:温州、广州、湖州、绍兴、湘潭、合肥、芜湖、岳阳、常德。

第 2 类:韶关。

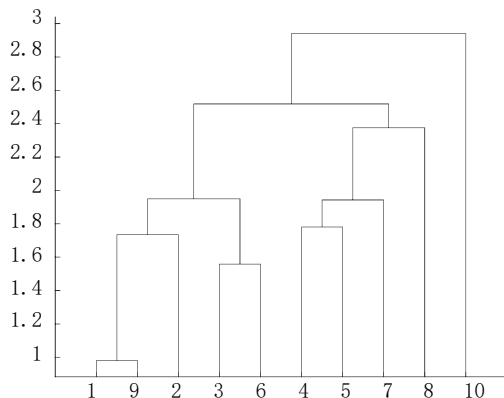


图 2 最短距离法聚类结果

采用最长距离法进行聚类,结果如图 3 所示。该方法将这 10 个城市分成了两大类:

第 1 类:温州、广州、湖州、韶关、绍兴、湘潭。

第 2 类:合肥、芜湖、岳阳、常德。

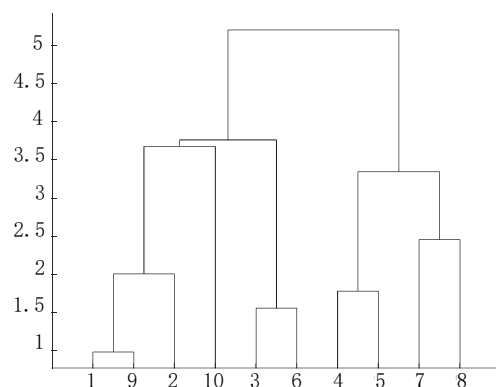


图 3 最长距离法聚类结果

比较两种算法的分类结果可以看出,最短距离法简单好用,且能直观地体现聚类的含义,但在上面的实例中,10 个城市中有 9 个城市被归成了一类,可见该算法有链接聚合的特点,容易将大部分样本聚为一类,原因在于该算法首先查找的是类与类间的距离最小者,合并该距离最小的两类以后,新类与其他类的距离也缩小了,因此容易聚合成一个比较大的类。最长距离法的分类结果中第一类城市有 6 个,第二类城市有 4 个,该方法因其两类合并以后,新类与其他类的距离为原来类中的距离最大者,因此加大了合并后的类与其他类的距离,故能克服最短距离法链接聚合的不足。

3 结束语

本文采用系统聚类法中的最短距离法和最大距离法,依据空气质量指标对部分城市进行了分类,由分类结果可知,不同的分类方法产生的结果会不同。在实际应用时,应根据分类问题的实际情况和各方法的优缺点选取合适的聚类方法。

参考文献:

- [1] 周博文. 基于聚类分析与优化算法的短期电力负荷预测研究 [D]. 哈尔滨: 哈尔滨理工大学, 2024.
- [2] 刘茜, 张静怡, 李玉静. 唐山市老年住院患者共病模式系统聚类分析 [J]. 中国预防医学杂志, 2023, 24(12): 1333-1338.
- [3] 田永明, 武士锋, 赵晰. 2 型糖尿病患者中医证型系统聚类分析及与疾病分期的关系 [J]. 实用医学杂志, 2023, 39(22): 2994-2999.
- [4] 李文天. Matlab 软件在职业教育课程教学中的应用 [J]. 电子技术, 2024, 53(12): 252-253.
- [5] 赵静, 但琦. 数学建模与数学实验: 第 5 版 [M]. 北京: 高等教育出版社, 2022.