第3卷 第2期 无锡轻工业学院学报 JOURNAL OF THE WUXI INSTITUTE OF LIGHT INDUSTRY Vol. 3 No. 2 1984

关于最短距离法的合理性

费荣昌 李中奇 吴大伟* 吴有炜

（基础课部）

摘 要

本文证明最短距离法在分类函数

$$E(G_1, G_2, \cdots, G_k) = \sum_{i=1}^{k} SMST(G_i)$$

之下是最合理的。

把 n 个样品 $X_1, X_2, \cdots, X_n$ 分成 k 类 $G_1, G_2, \cdots, G_k$，$R(G_i)$ 表示类 G_i 的直径，方开泰、马逢时(1979)定义 $R(G_i)$ 为 G_i 的最小支撑树 $MST(G_i)$ 的最大边长，并证明了系统聚类的最短距离法在分类函数

$$E(G_1, G_2, \cdots, G_k) = \max_{1 \leqslant i \leqslant k} R(G_i)$$

之下是最合理的[1][2]。这时，两类的距离与直径的关系是不明显的。

文[4]中曾指出也可定义 $R(G_i)$ 为 G_i 的最小支撑树的重量，即 $MST(G_i)$ 的全部边长的和，记作

$$R(G_i) = SMST(G_i)。$$

这时，最短距离法中类 G_i 与类 G_j 的距离为

$$D_{ij} = R(G_i \cup G_j) - R(G_i) - R(G_j),$$

它与直径之间具有明显的关系。

本文将证明最短距离法在分类函数

$$E(G_1, G_2, \cdots, G_k) = \sum_{i=1}^{k} R(G_i) = \sum_{i=1}^{k} SMST(G_i) \qquad (*)$$

之下也是最合理的。

在最短距离法的系统聚类过程中，开始时各样品自成一类，记作 $G_1^{(0)}, G_2^{(0)}, \cdots, G_n^{(0)}$；然后第一步合并其中最近的两类，得 n−1 个类，记作 $G_1^{(1)}, G_2^{(1)}, \cdots, G_{n-1}^{(1)}$；类似地记第 n−k 步的 k 个类为 $G_1^{(n-k)}, G_2^{(n-k)}, \cdots, G_k^{(n-k)}$。用 G 表示 n 个样品 $X_1, X_2,$

本文收到日期 1984 年 1 月 24 日

*无锡教育学院

…，X_n 组成的集合，如果G中任两样品间的距离都不相等，那未 MST(G) 是唯一的，它的边长各不相等，记为 $l_1<l_2\cdots<l_{n-1}$ (为了方便，l_i 既表示边长，又表示相应的边)。第 n－k 步合并的两类正是 l_{n-k} 所连接的两类(指 l_{n-k} 两端分别所属的类)，并且 $MST(G_i^{(n-k)})$ $(i=1,2,\cdots,k)$ 必为 MST(G) 的子集[1][2]，于是，由聚类过程可知

$$E(G_1^{(n-k)},G_2^{(n-k)},\cdots,G_k^{(n-k)})=\sum_{i=1}^{k}R(G_i^{(n-k)})$$

$$=\sum_{i=1}^{k}SMST(G_i^{(n-k)})=l_1+l_2+\cdots+l_{n-k}$$

把G任意分成K类 $G_1,G_2,\cdots,G_k$，设 $\bigcup_{i=1}^{k}MST(G_i)$ 的边为 $l'_1,l'_2,\cdots,l'_{n-k}$ (同时表示相应的边长)，则

$$E(G_1,G_2,\cdots,G_k)=\sum_{i=1}^{n-k}l'_i$$

如有

$$\sum_{i=1}^{n-k}l'_i\geqslant\sum_{i=1}^{n-k}l_i,$$

则就证明了 $G_1^{(n-k)}$，$G_2^{(n-k)},\cdots,G_k^{(n-k)}$ 是使分类函数(*)达到极小的精确最优解。

记

$$L=\{l_1,l_2,\cdots,l_{n-1}\},\quad L'=\{l'_1,l'_2,\cdots,l'_{n-k}\}$$

当 L′ 是L的子集时，因为 $l_1,l_2,\cdots,l_{n-k}$ 是L中最短的 n－k 条边，所以

$$\sum_{i=1}^{n-k}l'_i\geqslant\sum_{i=1}^{n-k}l_i$$

当 L′ 不是L的子集时，用反证法来证明，假设

$$\sum_{i=1}^{n-k}l'_i<\sum_{i=1}^{n-k}l_i$$

设 L′ 中不属于L的边有 r ($\leqslant$n－k) 条，不妨记作 $l'_1,l'_2,\cdots,l'_r$。先把 l_1' 添到 MST (G) 上，必形成回路。回路中必有属于 L－L′ 的边 (否则 $\bigcup_{i=1}^{k}MST(G_i)$ 将形成回路)，记其中之一为 l_{i_1}，把 l_{i_1} 去掉；然后添上 l'_2，又形成回路，回路中必有属于 L－L′ 的边，记其中之一为 l_{i_2}，把 l_{i_2} 去掉；依此做下去，最后添上 l_r'，形成回路，回路中必有属于 L-L′ 的边，记其中之一为 l_{i_r}，把 l_{i_r} 去掉。所得联结图必为联结 $X_1,X_2,\cdots,X_n$ 的一棵树，由上述添边、去边过程可知，它的重量

$$W=(l_1+\cdots+l_{n-1})+(l'_1+\cdots+l'_r)-(l_{i_1}+\cdots+l_{i_r})$$

$$=(l_1+\cdots+l_{n-1})+(l'_1+\cdots+l'_{n-k})-(l_{i_1}+\cdots+l_{i_r}+l'_{r+1}+\cdots+l'_{n-k})。$$

注意到 $l_{i1},\cdots,l_{ir},l'_{r+1},\cdots,l'_{n-k}$ 是L中不同的边，有

$$l_{i1}+\cdots+l_{ir}+l'_{r+1}+\cdots+l'_{n-k}\geqslant l_1+\cdots+l_{n-k},$$

并由反证法假设，得

$$W\leqslant(l_1+\cdots+l_{n-1})+(l'_1+\cdots+l'_{n-k})-(l_1+\cdots+l_{n-k})<l_1+\cdots+l_{n-1},$$

这与 MST(G) 是最小支撑树矛盾。

参　考　文　献

[1] 张尧庭，方开泰，多元统计分析引论，科学出版社，1982。

[2] 方开泰，潘恩沛，聚类分析，地质出版社，1982。

[3] 方开泰，聚类分析(I)，数学的实践与认识，1，1978。

[4] 方开泰，有序样品的一些聚类方法，应用数学学报，第5卷，第1期，1982。

The Rationality of Shortest Distance Method

By

Fei Rongchang, Li Zhongji, Wu Dawei and Wu Youwei

ABSTRACT

Proof of the rationality of shortest distance method is given when classification function

$$E(G_1,G_2,\cdots,G_k)=\sum_{i=1}^{k}SMST(G_i).$$