経営科学（日本オペレーションズ・リサーチ学会邦文機関誌） 第18巻 第1号（1974年1月）

〈総 合 報 告〉

クラスター分析における階層的手法†

——分割問題への適用について——

古 林 隆*

1. はじめに

クラスター分析の目的は，次の二つに大きく分けられる．

(1) 与えられた集合の階層的構造（階層的分類）を求めること．

(2) 与えられた集合の分割の中で，ある規準に従って最適なものを求めること．

クラスター分析の起こりは(1)の場合であるが，最近各方面でクラスター分析が注目されるようになったのは，(2)の場合に使われるようになったからである．この場合は，いわゆる組合せ問題であるから，DPや整数計画法により解くことも考えられている[1][2]が，集合が大きくなれば実際には解けないといってよい．そこで，(1)のために考えだされた多くの階層的手法が，(2)のための簡便法として使われている．

ここでは，おもな階層的手法の手順を説明するとともに，(2)のために階層的手法を適用するときの注意点や考え方を述べてみたい．

なお，分割の対象となる集合の要素を個体と呼び，その総数を n とする．$N=\{1,2,\cdots,n\}$ として，クラスターは N の部分集合と同一視する．また，クラスター A の大きさ（要素の数）を $|A|$ で表わす．

2. 階層的手法

二つのクラスター $A, B(A\cap B=\phi)$ の間の距離 $\delta(A,B)$ の定義が与えられているとして，階層的手法の手順を一般的に記述すると次のようになる．

手順 1. $\Gamma=\{\{1\},\{2\},\cdots,\{n\}\}$ とおく．

$\delta(\{i\},\{j\})(i,j=1,2,\cdots,n;i\neq j)$

を計算する（かまたは読みこむ）．

手順 2. $\min_{A,B\in\Gamma}\delta(A,B)=\delta(G,H)\quad(G,H\in\Gamma)$

となる G,H を求める．

手順 3. G と H を結合する．すなわち，

---

† 1973年10月30日受理．

* 埼玉大学行動科学情報解析センター．

$\Gamma-\{G,H\}+\{G\cup H\}\longrightarrow\Gamma$

とする.

$|\Gamma|=1$ ならば終了.

手順 4. すべての $A(A\epsilon\Gamma, A\neq G\cup H)$ に対して，$\delta(A, G\cup H)$ を計算して，手順 2 にもどる.

ここで，Γ は N の分割であり，手順 3 を実行するごとに．分割数が一つずつへっていく.

階層的手法は，クラスター間の距離 $\delta(A,B)$ の定義によって特徴づけられる．定義の仕方には二通りある.

D 型：個体間の距離行列 (d_{ij}) を用いて定義するもの.

X 型：個体のいくつかの特性値を用いて定義するもの.

D 型には次のようなものがある.

最短距離法 $\delta(A,B)=\min_{i\epsilon A,\, j\epsilon B} d_{ij}$

最長距離法 $\delta(A,B)=\max_{i\epsilon A,\, j\epsilon B} d_{ij}$

群間平均距離法 $\delta(A,B)=\dfrac{1}{|A||B|}\sum_{i\epsilon A,\, j\epsilon B} d_{ij}$

群内平均距離法 $\delta(A,B)=\dfrac{1}{\binom{|A\cup B|}{2}}\sum_{\substack{i,\, j\epsilon A\cup B\\ i<j}} d_{ij}$

X 型には，次のようなものがある．ただし，個体 i の特性 k についての観測値を x_{ik} とし，

$$\bar{x}_k^A=\frac{1}{|A|}\sum_{i\epsilon A} x_{ik}$$

$$S(A)=\sum_k\sum_{i\epsilon A}(x_{ik}-\bar{x}_k^A)^2$$

$$V(A)=\frac{1}{|A|}S(A)$$

とする.

重 心 法 $\delta(A,B)=\sum_k(\bar{x}_k^A-\bar{x}_k^B)^2$

Ward 法 $\delta(A,B)=S(A\cup B)-S(A)-S(B)$

$$=\frac{|A||B|}{|A\cup B|}\sum_k(\bar{x}_k^A-\bar{x}_k^B)^2$$

群内平均距離法 $\delta(A,B)=V(A\cup B)$

群内平均距離法については，文献［3］または［4］を参照されたい．Ward 法については，もとの論文［5］よりも Wishart の論文［6］のほうが計算式がはっきりしていてわかりやすい.

X 型でも，$\delta(\{i\},\{j\})$ は個体 i と個体 j との間の距離と考えられるが，$\delta(A,B)$ の定義式で $A=\{i\}, B=\{j\}$ とおいて計算されなければならない．それに対して D 型では，多くの場合 d_{ij} は個体の特性値 (x_{ik}) を用いて計算されるが，その定義は $\delta(A,B)$ の定義と独立に行なうことができる.

ここにあげたX型手法は，平均，平方和，分散などを用いていることからわかるように，計量値に対する手法であるが，D型は，d_{ij} さえ適当に定めれば，計数値に対しても適用することができる．

3. $\delta(A, G\cup H)$ の計算

階層的手法の手順の中で，計算量に関して問題になるのは，手順4でGとHを結合したあとで$\delta(A, G\cup H)$を計算するところである．これは，もちろん，定義式に従えば常に計算できるが，それでは“うまみ”がない．2節であげた手法のうち，群内平均距離法以外は，GとHを結合する前のクラスター間の距離$\delta(A, G)$，$\delta(A, H)$，$\delta(G, H)$ やクラスターの大きさ$|A|$，$|G|$，$|H|$を用いて計算することができる．このような手法を組合せ的という［7］．

最短距離法 $\delta(A, G\cup H)=\min\{\delta(A, G),\ \delta(A, H)\}$

最長距離法 $\delta(A, G\cup H)=\max\{\delta(A, G),\ \delta(A, H)\}$

群間平均距離法

$$\delta(A, G\cup H)=\frac{1}{|G\cup H|}\{|G|\delta(A, G)+|H|\delta(A, H)\}$$

重心法
$$\delta(A, G\cup H)=\frac{1}{|G\cup H|^2}\{|G||G\cup H|\delta(A, G)+|H||G\cup H|\delta(A, H)-|G||H|\delta(G, H)\}$$

Ward法
$$\delta(A, G\cup H)=\frac{1}{|A\cup G\cup H|}\{|A\cup G|\delta(A, G)+|A\cup H|\delta(A, H)-|A|\delta(G, H)\}$$

群内平均距離法は，組合せ的ではないが，次式で$\delta(A, G\cup H)$を計算する．

$$\delta(A, G\cup H)=\frac{1}{w(A\cup G\cup H)}\{w(A\cup G)\delta(A, G)+w(A\cup H)\delta(A, H)+w(G\cup H)\delta(G, H)-w(A)Q(A)-w(G)Q(G)-w(H)Q(H)\}$$

ここで，$w(A)$，$Q(A)$は次のように定める．

D型では

$$w(A)=\binom{|A|}{2},\quad Q(A)=\frac{1}{w(A)}\sum_{\substack{i,j\in A\\ i<j}} d_{ij},$$

X型では

$$w(A)=|A|^2,\quad Q(A)=\frac{1}{w(A)}\sum_{\substack{i,j\in A\\ i<j}}\sum_{k}(x_{ik}-x_{jk})^2$$

とする．

X型では，$Q(A)=V(A)$ となる．

従来，$\delta(A, G\cup H)$ の求めやすさが強調されすぎたきらいがある．その結果中点法，可変法のようにクラスター間の距離に意味がつかない手法も提唱されている．これはまったく本末転倒であ

って，むしろ$\delta(A, G\cup H)$を計算するのにいちいち定義式にもどらなければならないとしても，距離の定義が目的に適している手法のほうが存在価値は高いといえる．

4. 目 的 関 数

1節で，クラスター分析の目的を，ある規準に従って最適なNの分割を求めることと述べたが，ある規準としてどのようなものを考えればよいであろうか．

規準の定め方の一つとして，Nの分割Γの関数$f(\Gamma)$を考えて，その大小によって分割のよしあしを決める方法がある．すなわち，ある条件のもとで，たとえば，クラスター数（分割数）を指定して，$f(\Gamma)$を最小または最大にするものを最適な分割とするのである．もしこのように目的に対応して目的関数が定められるならば，最適なものが求められないにしても，二つの分割Γ_1, Γ_2のうちどちらがよいかは$f(\Gamma_1)$と$f(\Gamma_2)$の大小によって決められる．

2節にあげた手法に対する目的関数を，三つの型に分けて示すと次のようになる．ただし，$\Gamma=\{C_1, C_2, \cdots, C_\nu\}$とする．

(1) min-max 型

最長距離法 $f(\Gamma)=\max\limits_{\alpha}\max\limits_{i,j\in C_\alpha} d_{ij}\longrightarrow\min$

群内平均距離法 $f(\Gamma)=\max\limits_{\alpha} Q(C_\alpha)\longrightarrow\min$

(2) max-min 型

最短距離法 $f(\Gamma)=\min\limits_{\alpha,\beta}\min\limits_{i\in C_\alpha, j\in C_\beta} d_{ij}\longrightarrow\max$

群間平均距離法 $f(\Gamma)=\min\limits_{\alpha,\beta}\dfrac{1}{|C_\alpha||C_\beta|}\sum\limits_{i\in C_\alpha, j\in C_\beta} d_{ij}\longrightarrow\max$

重 心 法 $f(\Gamma)=\min\limits_{\alpha,\beta}\sum\limits_{k}(\bar{x}_k^{C_\alpha}-\bar{x}_k^{C_\beta})^2\longrightarrow\max$

(3) min-sum 型

Ward 法 $f(\Gamma)=\sum S(C_\alpha)$

Ward 法と群内平均距離法は，目的関数が先にあってそれに対応してクラスター間の距離を定めたものである．分割の目的が，C_αに含まれるすべての点をそれらの重心$(\bar{x}_k^{C_\alpha})$で近似することと考えるならば，Ward 法は最小二乗近似であり，X型群内平均距離法はmin-max 近似に近いと考えられる．また，等分散の仮定が成りたつように分割したいのであれば，X型群内平均距離法の目的関数は，まさにその要求にあったものといえる．他の手法については，クラスター間の距離の定義が先にあって，あとから目的関数を対応させたものである．最短距離法以外は，対応させた目的関数に関して常に最適解を与えるとは限らない．

2節にあげた手法について，$\delta(A, B)$，$\delta(A, G\cup H)$の計算式，対応する目的関数を表1にまとめて示しておく．

表1 階層的手法一覧表

D型 (d_{ij}は個体iと個体jとの間の距離を表わす)

名　　称	$\delta(A,B)$	$\delta(A,G\cup H)$の計算	目的関数($\Gamma=\{C_1,C_2,\cdots C_v\}$)
最短距離法	$\min_{i\in A,\,j\in B} d_{ij}$	$\min\{\delta(A,G),\delta(A,H)\}$	$\min_{\alpha,\beta}\min_{i\in C_\alpha,\,j\in C_\beta} d_{ij}\to\max$
最長距離法	$\max_{i\in A,\,j\in B} d_{ij}$	$\max\{\delta(A,G),\delta(A,H)\}$	$\max_{\alpha}\max_{i,\,j\in C_\alpha} d_{ij}\to\min$
群間平均距離法	$P(A,B)$	$\frac{1}{\lvert G\cup H\rvert}\{\lvert G\rvert\delta(A,G)+\lvert H\rvert\delta(A,H)\}$	$\min_{\alpha,\beta} P(C_\alpha,C_\beta)\to\max$
群内平均距離法	$Q(A\cup B)$	$\frac{1}{w(A\cup G\cup H)}\{w(A\cup G)\delta(A,G)+w(A\cup H)\delta(A,H)+w(G\cup H)\delta(G,H)-w(A)Q(A)-w(G)Q(G)-w(H)Q(H)\}$	$\max_{\alpha} Q(C_\alpha)\to\min$

ただし，$P(A,B)=\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{i\in A,\,j\in B} d_{ij}$，$Q(A)=\frac{1}{\binom{\lvert A\rvert}{2}}\sum_{i,\,j\in A,\,i<j} d_{ij}$，$w(A)=\binom{\lvert A\rvert}{2}$とする．

X型 (x_{ik}は個体iの特性kについての観測値を表わす)

名　　称	$\delta(\{i\},\{j\})$	$\delta(A,B)$	$\delta(A,G\cup H)$の計算	目的関数($\Gamma=\{C_1,C_2,\cdots C_v\}$)
重心法	$\sum_k(x_{ik}-x_{jk})^2$	$\sum_k(\bar{x}_k^A-\bar{x}_k^B)^2$	$\frac{1}{\lvert G\cup H\rvert^2}\{\lvert G\rvert\lvert G\cup H\rvert\delta(A,G)+\lvert H\rvert\lvert G\cup H\rvert\delta(A,H)-\lvert G\rvert\lvert H\rvert\delta(G,H)\}$	$\min_{\alpha,\beta}\sum_k(\bar{x}_k^{C_\alpha}-\bar{x}_k^{C_\beta})^2\to\max$
Ward法	$\frac{1}{2}\sum_k(x_{ik}-x_{jk})^2$	$S(A\cup B)-S(A)-S(B)$	$\frac{1}{\lvert A\cup G\cup H\rvert}\{\lvert A\cup G\rvert\delta(A,G)+\lvert A\cup H\rvert\delta(A,H)-\lvert A\rvert\delta(G,H)\}$	$\sum_\alpha S(C_\alpha)\to\min$
群内平均距離法	$\frac{1}{4}\sum_k(x_{ik}-x_{jk})^2$	$V(A\cup B)$	$\frac{1}{w(A\cup G\cup H)}\{w(A\cup G)\delta(A,G)+w(A\cup H)\delta(A,H)+w(G\cup H)\delta(G,H)-w(A)V(A)-w(G)V(G)-w(H)V(H)\}$	$\max_{\alpha} V(C_\alpha)\to\min$

ただし，$\bar{x}_k^A=\frac{1}{\lvert A\rvert}\sum_{i\in A}x_{ik}$，$S(A)=\sum_k\sum_{i\in A}(x_{ik}-\bar{x}_k^A)^2$，$V(A)=\frac{1}{\lvert A\rvert}S(A)$，$w(A)=\lvert A\rvert^2$とする．

5. むすび

クラスター間の距離が持つ意味は，階層的構造あるいはそれを表現したデンドログラム（樹形図）を求める際には重要であるが，最適な分割を求めたいときには，むしろ対応する目的関数が目的にあっているかどうかのほうが重要である．したがって，まず目的にあった目的関数を定め，それに対応する手法を選んでいくのが当を得ている．ただ，前節にもふれたように，最短距離法以外は，一般に近似解しか与えないから，いくつかの手法による結果の中からよいものは採用することも考えられる（その意味では，クラスター間の距離に意味がつかない手法も，それによる解が結果的にある目的関数を最小にしたり，最大にしたりすることもありえるから，使って

みる価値がないわけではない).

目的関数にあった手法がない場合，現在ある手法の中から最善のものを選ぶという受動的態度にとどまらず，それにあった手法すなわちクラスター間の距離を考えるという能動的態度が望まれる．その際，クラスター間の距離が意味を持たなくても（多くの場合，目的関数値の変化分としての意味を持っている)．また組合せ的でなくても，最適なものあるいはそれに近いものを作りだす手法が“よい手法”であるといえる．

ここで一つの例を紹介してしめくくりとしよう．

ある工場では，n種類の道具を格納する倉庫を建設することを考えている．道具iと道具jを同時に使う作業の頻度f_{ij}はわかっている．できるだけ倉庫の開閉回数を少なくするには，どのように組み合わせて倉庫に格納すればよいであろうか．

与えられた条件のもとでは，目的関数を

$$f(\Gamma)=\sum_{\alpha}\sum_{\substack{i,j\in C_\alpha\\ i<j}} f_{ij}$$

とし，倉庫の数すなわちクラスター数を指定して，$f(\Gamma)$ を最大にするようなクラスターを求めることにすればよい．これは，(最小問題に変えることによって) 4節の分類でいえば(3)min-sum型にはいるから，Ward法と同じように考えて，クラスター間の距離は次のように定めるとよい．

$$\delta(A,B)=-\{T(A\cup B)-T(A)-T(B)\}$$

ここで$T(A)=\sum_{\substack{i,j\in A\\ i<j}} f_{ij}$ とする．

このとき，

$$\delta(A,G\cup H)=\delta(A,G)+\delta(A,H)$$

が成りたつから，手順4の計算の手間は非常に少ない．

この問題では，実際には道具の大きさと倉庫の容量をあわせて考える必要があるから，得られた解がそのまま実行可能とは限らないが，手法の選択あるいは開発における考え方を理解してもらえれば幸いである．

参 考 文 献

[1] Jensen, R. E., “A Dynamic Programming Algorithm for Cluster Analysis,” *Operations Research*, **17** (1969), 1034-1057.

[2] Rao, M. R., “Cluster Analysis and Mathematical Programming,” *Journal of the American Statistical Association*, **66**(1971), 622-626.

[3] 古林 隆，“クラスター分析における階層的手法と目的関数”，埼玉大学教養学部紀要，**8**(1972)，17-22.

[4] 古林 隆，“新しい階層的手法——群内平均距離法”，日本OR学会1973年度春季研究発表会アブストラクト集．

[5] Ward, J. R., “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association*, **58**(1963), 236-244.

[6] Wishart, D., “An Algorithm for Hierarchical Classifications,” *Biometrics*, **25**(1969), 165-170.

[7] Lance, G. N. and W. T. Williams, "A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems," *Computer Journal*, **9**(1967), 373–380.

総合報告として，次の文献がある．

Cormack, R. M., "A Review of Classification," *Journal of the Royal Statistical Society*, **A134**(1971), Series A, 321–367.

補遺：D型群内平均距離法の適用例

図1に示すような2次元の20個のデータに対して

$$d_{ij}=\{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2\}^{1/2}$$

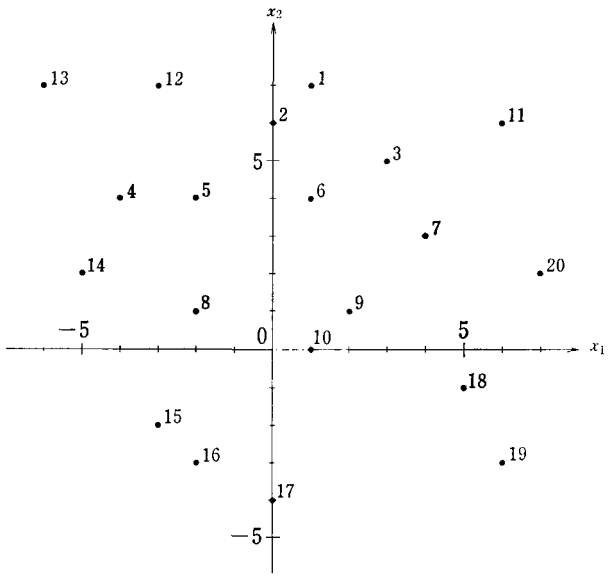

図1 散 布 図

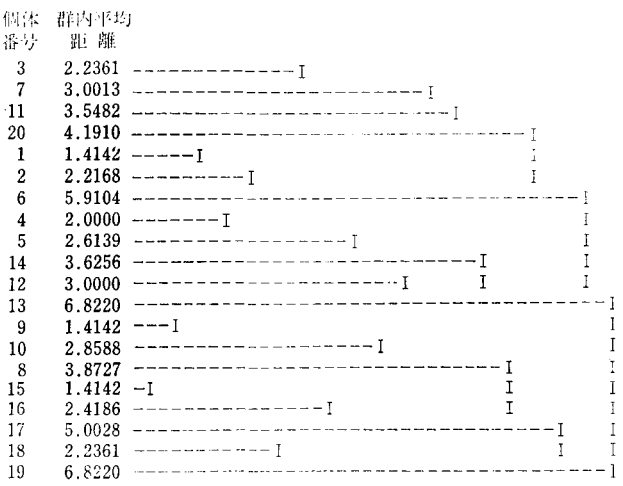

図2 デンドログラム

とおいて，D型群内平均距離法を適用した結果，得られたデンドログラムを図2に示す（このデータは，手順の説明用に作ったものであるから，実際的な意味はない）．

デンドログラムで，個体番号の次の数値は二つのクラスターが結合するときの距離，すなわち結合されてでき上がったクラスターの群内平均距離を表わす．クラスターはその中で一番下側に並んでいる個体で代表させ，クラスターAとクラスターBが結合したときの距離$\delta(A,B)=Q(A\cup B)$は，上側のクラスターの中で一番下側に並んでいる個体のところに示されている．たとえば，{3}と{7}が結合してでき上がるクラスター{3,7}の群内平均距離は$d_{37}=2.2361$であり，それと{11}が結合してでき上がるクラスター{3,7,11}の群内平均距離は

$$\frac{1}{3}(d_{37}+d_{3,11}+d_{7,11})$$

$$=\frac{1}{3}(2.2361+3.1623+3.6056)=3.0013$$

である．また，{3, 7, 11, 20} と {1, 2, 6} が結合してでき上がるクラスター {3, 7, 11, 20, 1, 2, 6} の群内平均距離は，個体番号20のところに示されていて，4.1910であることがわかる．

ここで，クラスター数を5に指定すると，表2に示すようなクラスターが得られる．このときの各クラスター内の個体間距離を表3にあげておく．

表2　個数5のときのクラスター

クラスター	群内平均距離
3, 7, 11, 20	3.5482
1, 2, 6	2.2168
4, 5, 14, 12, 13	3.6256
9, 10, 8, 15, 16, 17	3.8727←max
18, 19	2.2361

表3　クラスター内の個体間距離

3						
7	2.2361					
11	3.1623	3.6056				
20	5.0000	3.1623	4.1231			平均 3.5482
	3	7	11	20		
1						
2	1.4142					
6	3.0000	2.2361				平均 2.2168
	1	2	6			
4						
5	2.0000					
14	2.2361	3.6056				
12	3.1623	3.1623	5.3852			
13	3.6056	5.0000	5.0990	3.0000		平均 3.6256
	4	5	14	12	13	
9						
10	1.4142					
8	4.0000	3.1623				
15	5.8310	4.4721	3.1623			
16	5.6569	4.2426	4.0000	1.4142		
17	5.3852	4.1231	5.3852	3.6056	2.2361	平均 3.8727
	9	10	8	15	16	17
18						
19	2.2361					平均 2.2361
	18	19				