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17巻 第2号 (1973年3月)

デンドログラムの一つの指標†

古林 隆*

1. まえがき

クラスター分析で用いられる手法のなかには，一つの個体で一つのクラスターを形成するところから出発して，ある規準に関して最も近い二つのクラスターを順々に結合させていく階層的手法が数多くある．階層的手法では，ある時点で同一のクラスターにはいった個体はその後分離されることはなく，距離がいくらのときに，どのクラスターとどのクラスターが結合したかを示すのにデンドログラム（樹形図）が用いられる．

ここでは，デンドログラムの特徴を表わす一つの指標を提案し，その性質をのべる．

また，個体数nのデンドログラムの集合をΔ_n，正整数xの1/2をこえない最大整数を$x|2$で表わすことにする．

2. 指標Kの定義

デンドログラムで，距離は無視して，結合の仕方にだけ注目することにしよう．個体数をnとし，i番目の個体が属するクラスターの結合回数をl_i** とするとき，指標Kを次式で定義する．

(1) $K=\sum_{i=1}^{n} l_i$

とくに，デンドログラムDを強調したときは，$K(D)$とかくことにする．

[例]

図1に示すような$n=10$の二次元のデータに対して，ユークリッド距離に関する最短距離法を適用すると，図2のデンドログラムが得られる．$\boldsymbol{l}=(6654554222)$であるから，$K=6+6+5+4+5+5+4+2+2+2=41$である．

3. 指標Kの性質

(a) 個体数nのデンドログラムDが，最後にデンドログラムD_1とデンドログラムD_2を結合してでき上がったものであれば（図3参照），次式が成りたつ．

---

† 1973年1月18日受理．1972年9月25日，秋季研究発表会講演要旨．

* 埼玉大学教養学部．

** デンドログラムを（距離を無視することによって）ある競技のnチームによるトーナメントの組合せとみるならば，i番目のチームが優勝するために勝たなければいけない試合数がl_iである．

表1　距　離　行　列

	A	B	C	D	E	F	G	H	I	J
A	0.0000	3.6056	5.3852	6.0828	5.0990	8.0623	8.5440	7.2111	5.0000	2.0000
B	3.6056	0.0000	2.0000	3.1623	3.6056	5.8310	7.0711	6.7082	6.0000	5.0000
C	5.3852	2.0000	0.0000	1.4142	3.0000	4.2426	5.8310	6.0828	6.3246	6.4031
D	6.0828	3.1623	1.4142	0.0000	2.2361	2.8284	4.4721	5.0000	5.8310	6.7082
E	5.0990	3.6056	3.0000	2.2361	0.0000	3.0000	3.6056	3.1623	3.6056	5.0990
F	8.0623	5.8310	4.2426	2.8284	3.0000	0.0000	2.0000	3.6056	5.8310	8.0623
G	8.5440	7.0711	5.8310	4.4721	3.6056	2.0000	0.0000	2.2361	5.0990	8.0623
H	7.2111	6.7082	6.0828	5.0000	3.1623	3.6056	2.2361	0.0000	3.0000	6.3246
I	5.0000	6.0000	6.3246	5.8310	3.6056	5.8310	5.0990	3.0000	0.0000	3.6056
J	2.0000	5.0000	6.4031	6.7082	5.0990	8.0623	8.0623	6.3246	3.6056	0.0000

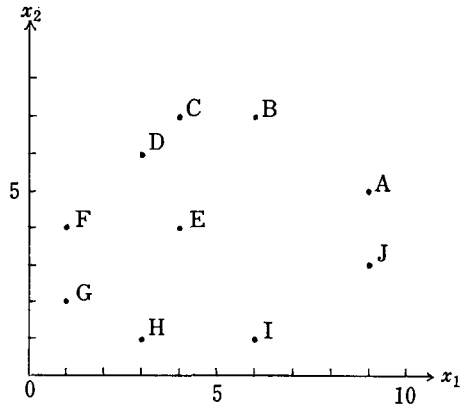


図1　散　布　図

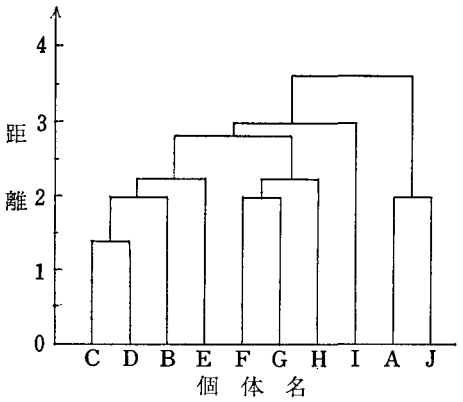


図2　最短距離法によるデンドログラム

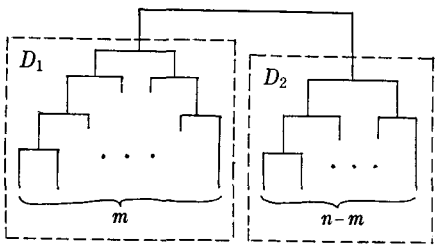


図3　D の分解

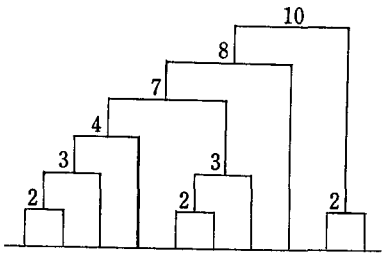


図4　Kの計算
数字は結合してでき上がったクラスターの大きさを示す
$K=2+3+4+2+3+7+8+2+10=41$

(2)　$K(D)=K(D_1)+K(D_2)+n$

ただし，$D_i(i=1$ または$2)$の個体数が1のときは $K(D_i)=0$ とする．

D における $\boldsymbol{l}_i$ の値は，D_1 または D_2 における $\boldsymbol{l}_i$ の値よりも1だけ大きいことから，ただちに証明される．

(b)　K の値は，1個体で一つのクラスターを形成している最初の状態から，n 個体の一つのクラスターになる最後の状態までの間に，結合してでき上がるクラスターの個体数をすべて加えあわせたものに等しい．

これは，(a)の性質より漸化的に証明される．図1の例では，$K=2+3+4+2+3+7+8+2+10=41$ となる（図4参照）．

K の値を計算するには，定義式に従うより，この性質を利用するほうが簡単である．

(c) $K_{\max}(n)=\max_{D\in\Delta n}K(D)$

$K_{\min}(n)=\min_{D\in\Delta n}K(D)$

とすると，次式が成りたつ．

(3) $K_{\max}(n)=n(n+1)/2-1=(n-1)(n+2)/2$

$n=2^p+q$(p,q整数，$0\leqq q<2^p$)とすると

(4) $K_{\min}(n)=p(2^p-q)+(p+1)\times 2q$

$=p\cdot 2^p+(p+2)q$

また，近似的に次式が成りたつ．

(5) $K_{\min}(n)\fallingdotseq n\log_2 n$

(3)，(4)式の証明

$f(n)=(n-1)(n+2)/2$

$g(n)=p\cdot 2^p+(p+2)q$

とおく．

まず，Δ_n の中に $K(D_{\max})=f(n)$ となるデンドログラム $D_{\max}$ が存在することを示そう．

最初に個体1と個体2が結合し，それに個体3が結合し，さらにそれに個体4が結合するというように，一つずつ順々に結合していくときのデンドログラム $D_{\max}$（図5参照)を考えると，$l_1=l_2=n-1, l_3=n-2, \cdots\cdots, l_{n-1}=2$，$l_1=1$ であるから，$K(D_{\max})=n(n+1)/2-1=(n-1)(n+2)/2$ である．

次に，Δ_n の中に $K(D_{\min})=g(n)$ となるデンドログラム $D_{\min}$ が存在することを示そう．

$q>0$ならば，最初の $2q$ 個の個体を2個ずつ組み合わせて，大きさ2のクラスターを q 個つくると，残りの (2^p-q) 個の個体とあわせて，2^p 個のクラスターができ上がる．次に，二つずつクラスターを組み合わせて 2^{p-1} 個のクラスターをつくり，それらをさらに二つずつ組み合わせるという操作をくりかえしてでき上がるデンドログラム $D_{\min}$(図6参照）を考えると

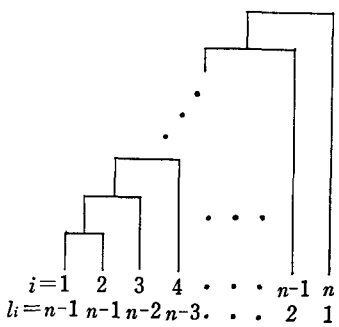


図5 $D_{\max}$

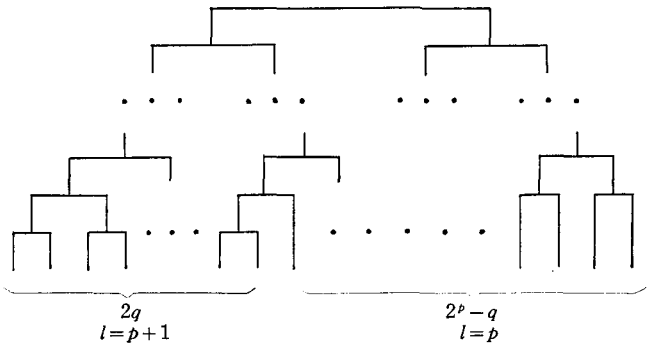


図6 $D_{\min}$

$$l_1=l_2=\cdots\cdots=l_{2q}=p+1, l_{2q+1}=l_{2q+2}=\cdots\cdots=l_n=p$$

よって

$$K(D_{\min})=(p+1)\times 2q+p\times(2^p-q)$$
$$=p\times 2^p+(p+2)q$$

である.

したがって，任意の $D\in\Delta_n$ に対して

(6) $g(n)\leqq K(D)\leqq f(n)$

が成りたつことを証明すればよい.

$$g(1)=f(1)=0$$

であるから，(6)式は $n=1$ のとき成りたつ．$n=1,2,\cdots\cdots,t$ に対して(6)式が成りたつものとする．Δ_{t+1} の任意の要素を D_0 とするとき，D_0 が最後に $D_1\in\Delta_u$ と $D_2\in\Delta_v(u+v=t+1)$ を結合してでき上がったものとすると，(2)式より

$$K(D_0)=K(D_1)+K(D_2)+t+1$$

$n=u, n=v$ に対して(6)式は成りたつから

$$g(u)\leqq K(D_1)\leqq f(u)$$
$$g(v)\leqq K(D_2)\leqq f(v)$$
$$\therefore \min_{\substack{u,v\\ u+v=t+1}}\{g(u)+g(v)+t+1\}\leqq K(D_0)\leqq \max_{\substack{u,v\\ u+v=t+1}}\{f(u)+g(v)+t+1\}$$

f は凸関数であるから，$f(u)+f(v)$ は $u+v=t+1$ という条件のもとでは，$u=1, v=t$ のとき最大となる．このとき

$$f(u)+f(v)+t+1=0+(t-1)(t+2)/2+(t+1)=t(t+3)/2=f(t+1)$$

g も凸関数であるから，$g(u)+g(v)$ は与えられた条件のもとでは，$u=(t+1)/2, v=(t+2)/2$ のとき最小となる.

このとき

$$t+1=2^{p'}+q'(p',\ q'\text{整数},\ 0\leqq q'<2^{p'})$$

とすると

$$u=2^{p'-1}+q'/2$$
$$v=2^{p'-1}+(^{p'}+1)/2$$

である.

$$0\leqq q'|2<2^{p'-1}$$

が成りたつから

$$g(u)=(p'-1)2^{p'-1}+(p'+1)(q'/2)$$

また，$q'<2^{p'}-1$ のときは

$$0\leqq(q'+1)/2<2^{p'-1}$$

が成りたつから

$$g(v)=(p'-1)/2^{p'-1})+(p'+1)\{(q'+1)/2\} \quad (*)$$

$q'=2^{p'}-1$ のときは，$v=2^{p'}$ となり

$$g(v)=p'\cdot 2^{p'}$$

となるが，これは (*) に$q'=2^{p'}-1$を代入したものと一致する．ゆえに

$$g(u)+g(v)+(t+1)$$
$$=(p'-1)2^{p'}+(p'+1)q'+(2^{p'}+q')$$
$$=p'\cdot 2^{p'}+(p'+2)q'=g(t+1)$$

以上より，$g(t+1)\leqq K(D_0)\leqq f(t+1)$

(6)式は，$n=t+1$のときも成りたつから，すべてのnに対して(6)式は成りたつ（証明終わり）．

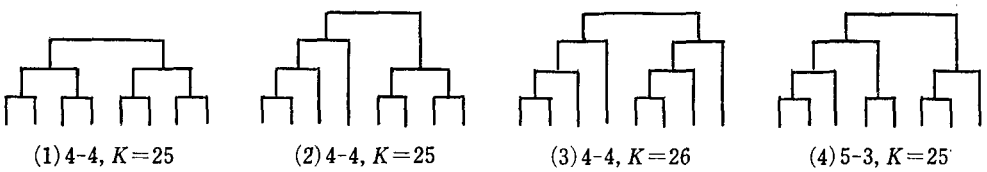
表2に$n=2(1)32$に対する$K_{\max}, K_{\min}, n\log_2 n$の値を示す．

表2 $K_{\max}(n)$, $K_{\min}(n)$, s_nの表

n	$K_{\max}(n)$	$K_{\min}n)$	$n\log_2 n$	s_n	s_n/s_{n-1}
2	2	2	2.00	1	—
3	5	5	4.76	1	1.0000
4	9	8	8.00	2	2.0000
5	14	12	11.61	3	1.5000
6	20	16	15.51	6	2.0000
7	27	20	19.65	11	1.8333
8	35	24	24.00	23	2.0909
9	44	29	28.53	46	2.0000
10	54	34	33.22	98	2.1304
11	65	39	38.05	207	2.1122
12	77	44	43.02	451	2.1787
13	90	49	48.11	983	2.1796
14	104	54	53.30	2179	2.2167
15	119	59	58.60	4850	2.2258
16	135	64	64.00	10905	2.2485
17	152	70	69.49	24631	2.2587
18	170	76	75.06	56011	2.2740
19	189	82	80.71	127912	2.2837
20	209	88	86.44	293547	2.2949
21	230	94	92.24	676157	2.3034
22	252	100	98.11	1563372	2.3121
23	275	106	104.04	3626149	2.3194
24	299	112	110.04	8436379	2.3265
25	324	118	116.10	19680277	2.3328
26	350	124	122.21	46026618	2.3387
27	377	130	128.38	107890609	2.3441
28	405	136	134.61	253450711	2.3491
29	434	142	140.88	596572387	2.3538
30	464	148	147.21	1406818759	2.3582
31	495	154	153.58	3323236238	2.3622
32	527	160	160.00	7862958391	2.3661

4. デンドログラムのパターン

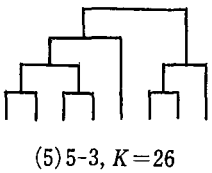
二つのデンドログラムにおいて，距離および個体番号を無視し，しかも並べかえによって重なるならば，パターンが同じであるということにしよう．個体数が8のときのパターンを列挙す

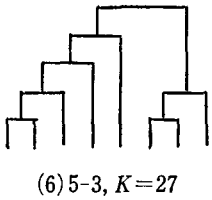

(1) 4-4, $K=25$

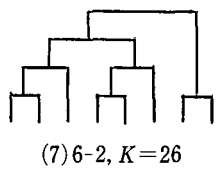
(2) 4-4, $K=25$

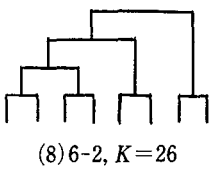
(3) 4-4, $K=26$

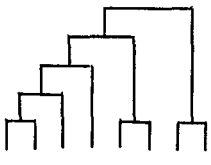
(4) 5-3, $K=25$

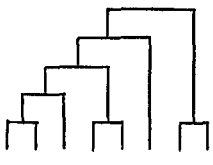

(5) 5-3, $K=26$

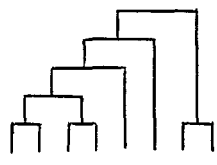

(6) 5-3, $K=27$

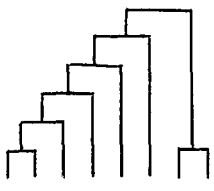

(7) 6-2, $K=26$

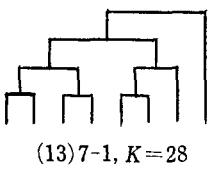

(8) 6-2, $K=26$

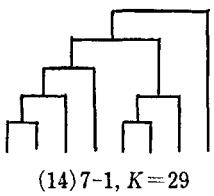

(9) 6-2, $K=27$

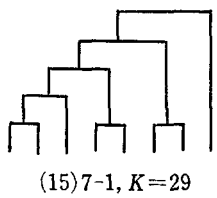

(10) 6-2, $K=28$

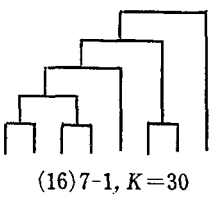

(11) 6-2, $K=29$

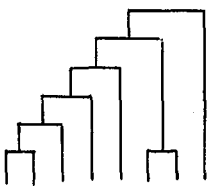

(12) 6-2, $K=30$

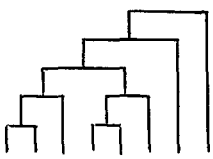

(13) 7-1, $K=28$

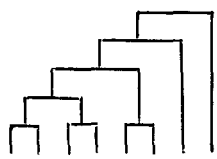

(14) 7-1, $K=29$

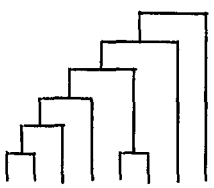

(15) 7-1, $K=29$

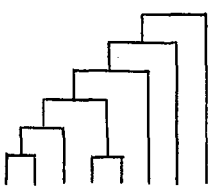

(16) 7-1, $K=30$

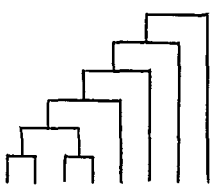

(17) 7-1, $K=31$


(18) 7-1, $K=31$


(19) 7-1, $K=31$


(20) 7-1, $K=32$


(21) 7-1, $K=33$


(22) 7-1, $K=34$

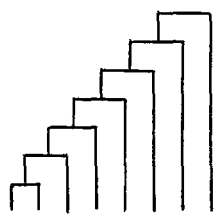

(23) 7-1, $K=35$

図7　$n=8$ のときのパターン
パターンの番号のあとの数字は，最後に結合された二つのクラスターの大きさを示す

表3 K の分布

k $f_n(k)$	k $f_n(k)$	k $f_n(k)$	k $f_n(k)$	k $f_n(k)$	k $f_n(k)$
$n=4$	27 2	$n=10$	$n=11$	$n=12$	73 3
8 1	28 2	34 3	39 3	44 5	74 1
9 1	29 3	35 5	40 7	45 8	75 1
——	30 2	36 7	41 9	46 14	76 1
$n=5$	31 3	37 6	42 12	47 19	77 1
12 1	32 1	38 10	43 13	48 20	
13 1	33 1	39 5	44 10	49 21	
14 1	34 1	40 7	45 13	50 20	
——	35 1	41 8	46 12	51 24	
$n=6$	——	42 7	47 17	52 29	
16 2	$n=9$	43 6	48 12	53 20	
17 1	29 1	44 7	49 14	54 24	
18 1	30 4	45 6	50 9	55 27	
19 1	31 4	46 4	51 12	56 20	
20 1	32 4	47 4	52 10	57 22	
——	33 3	48 4	53 10	58 20	
$n=7$	34 6	49 2	54 7	59 24	
20 1	35 5	50 3	55 8	60 17	
21 2	36 3	51 1	56 7	61 19	
22 1	37 3	52 1	57 5	62 13	
23 3	38 4	53 1	58 4	63 14	
24 1	39 2	54 1	59 4	64 13	
25 1	40 3		60 2	65 11	
26 1	41 1		61 3	66 8	
27 1	42 1		62 1	67 9	
——	43 1		63 1	68 8	
$n=8$	44 1		64 1	69 5	
24 1			65 1	70 4	
25 2				71 4	
26 4				72 2	

ると，図7のようになる．

個体数 n のデンドログラムのパターンの総数を s_n とすると，次の漸化式が成りたつ．

$$(7)\quad s_n=\begin{cases} s_1s_{n-1}+s_2s_{n-2}+\cdots\cdots+s_{(n-1)/2}s_{(n+1)/2} & (n\text{ 奇数},\ n\geqq 3)\\ s_1s_{n-1}+s_2s_{n-2}+\cdots\cdots+s_{n/2-1}s_{n/2+1}+\dfrac{1}{2}s_{n/2}(s_{n/2}+1) & (n\text{ 偶数}) \end{cases}$$

ただし $s_1=1$ とする．

(7)式は，最後に $D_1\epsilon\Delta_m$ と $D_2\epsilon\Delta_{n-m}(1\leqq m\leqq n\rfloor 2)$ が結合してでき上がるデンドログラムのパターンの数は，$m<n/2$ ならば，s_ms_{n-m}

n が偶数で $m=n/2$ ならば $\dfrac{1}{2}s_m(s_m+1)$ であることから導かれる．

表2に，$n=2(1)32$ に対する s_n の値を示す．

パターンが同じであるデンドログラムでは K の値は同じである．個体数 n のデンドログラム

で$K=k$となるもののパターンの数を$f_n(k)$とすると，次の漸化式が成りたつ．

$$(8)\quad f_n(k)=\sum_{m=1}^{(n-1)\mid 2}\sum_{i=0}^{k-n} f_m(i)f_{n-m}(k-n-i)+g_n(k)$$

ここで

$$g_n(k)=\begin{cases} 0 & (n\text{奇数},\ n\geqq 3)\\ \displaystyle\sum_{i=0}^{(k-n-1)/2} f_{n/2}(i)f_{n/2}(k-n-i) & (n\text{偶数},\ k\text{奇数})\\ \displaystyle\sum_{i=0}^{(k-n)/2-1} f_{n/2}(i)f_{n/2}(k-n-i)+\frac{1}{2}f_{n/2}\left(\frac{k-n}{2}\right)\left\{f_{n/2}\left(\frac{k-n}{2}\right)+1\right\} & (n,k\text{偶数}) \end{cases}$$

であり，また

$$f_1(k)=\begin{cases} 1 & (k=0)\\ 0 & (k\neq 0) \end{cases}$$

とする．

表3に，$n=4(1)12$に対するKの分布を示す．

5. 考　　察

図8，図9は，図1の例に，最長距離法，平均距離法を適用したときのデンドログラムである．

Kの値は次のようになる．

最短距離法　41

最長距離法　35

平均距離法　36

三つの手法のなかでは，Kの値が一番大きくなるのは，最短距離法であり，最長距離法と平

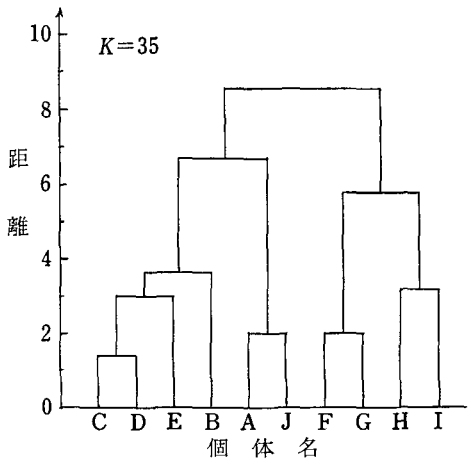


図8　最長距離法によるデンドログラム

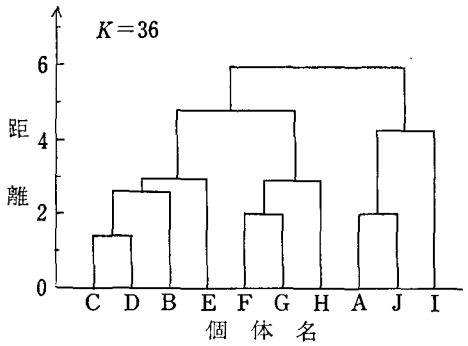


図9　平均距離法によるデンドログラム

均距離法ではあまり差がなく，平均距離法のほうがKの値が小さくなることもありえることが，いくつかの例で示されている．

次に，三つの手法による結果で，クラスター数が2,3,4のときのクラスターの大きさを示すと，表4のようになる．

表4 クラスターの大きさの分布

クラスター数	最短距離法 ($K=41$)	最長距離法 ($K=35$)	平均距離法 ($K=36$)
2	8,2	6,4	7,3
3	7,2,1	4,4,2	4,3,3
4	4,3,2,1	4,2,2,2	4,3,2,1

一般に，クラスターの大きさがそろっていないとき（大きさの分散が大きいとき）は，K値は大きく，大きさがそろっているときは，K値は小さくなる．クラスター分析ではクラスターの大きさがそろっていることが望まれることが多いが，そのような要求をK値の上限を与えることによって表わすことができる．すなわち，Kによって一つの必要条件を表わすことができる．

また，少数の大きなクラスターができて，残りの小さいクラスターと結合していく現象を“鎖効果”とよんでいるが，K値によって鎖効果をある程度定量的に表わすことができよう．したがって，どの手法が鎖効果を起こしやすいかを，K値によって示すことができるし，同種類の集団がいくつかあるとき，内部でのクラスターのでき方のちがいも示すことができる．

$(l_1, l_2, \cdots\cdots, l_n)$ の関数としては，Kのほかに次のようなものが考えられる．

$$V=\sum l_i^2/n-(k/n)^2$$

$$H=\sum l_i\, 2^{-li}$$

Vは分散であり，Hは

表5 K, V, Hの関係

パターンの番号	$l_i=a$ となる l_i の数							K	V	$\sqrt{V}$	H
	$a=1$	2	3	4	5	6	7				
1			8					24	0	0	3.000
2,4		1	5	2				25	0.109	0.331	2.875
3,5,7,8		2	2	4				26	0.688	0.829	2.750
6,9		2	3	1	2			27	1.234	1.111	2.688
13	1		1	6				28	1.000	1.000	2.375
10		3		3	2			28	1.500	1.225	2.563
14,15	1		2	3	2			29	1.484	1.218	2.313
11		3	1		4			29	1.984	1.409	2.500
16	1		3		4			30	1.938	1.392	2.250
12		3	1	1	1	2		30	2.688	1.639	2.469
18,19	1	1		2	4			31	2.109	1.452	2.125
17	1		3	1	1	2		31	2.609	1.615	2.219
20	1	1		3	1	2		32	2.750	1.658	2.094
21	1	1	1		3	2		33	3.109	1.763	2.031
22	1	1	1	1		4		34	3.688	1.920	2.000
23	1	1	1	1	1	1	2	35	4.484	2.118	1.984

$$\sum 2^{-l_i}=1$$

に注目すれば，エントロピーと考えられる．図7に示した$n=8$のときのパターンに対して，K, $V, \sqrt{V}, H$を計算したものを表5に示す．$K, \sqrt{V}, H$間の相関係数は次のとおりである．

	K	$\sqrt{V}$	H
K	1	0.949	−0.965
$\sqrt{V}$	0.949	1	−0.903
H	−0.965	−0.903	1

パターン10と13のように，Kの値は同じであるが，V, Hが異なるものもあるし，パターン12と18のように，KとVの大小関係が逆のものもあるが，計算が簡単であることを考慮すれば，KはV, Hに劣らない指標であるといえよう．

補　遺

本文でとりあげた三つの手法におけるクラスター間の距離の定義を示しておく．

d_{ij}をi番目の個体とj番目の個体の間の距離とし，クラスターは個体番号の集合で表わすことにすると，クラスターAとクラスターBの間の距離$\delta(A,B)$は，三つの手法ではそれぞれ次のように定義される．

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

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

平均距離法　$$\delta(A,B)=\frac{1}{|A||B|}\sum_{i\in A,\, j\in B} d_{ij}$$

参考文献

[1] 奥野忠一ほか，多変量解析法，日科技連出版社，1971.

[2] 矢島敬二ほか，“クラスター・アナリシス”，オペレーションズ・リサーチ，**16** (1971)，8－11.

[3] Lance, G. N. and W. T. Williams, “A general theory of classificatory sorting strategies. **I**. Hierarchical systems,” *Comp. J.*, **9** (1967), 373－380.

[4] Sokal, R. R. and P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freedman and Comp., 1963.