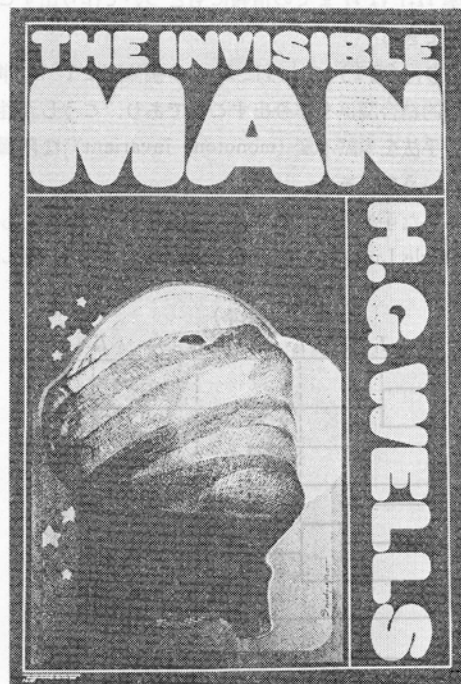


クラスター分析

—SL 法をめぐって—

大 隅 昇



アレックス・グニジェイコ

1. はじめに

最近、盛んにクラスター分析によるデータ解析を望む声が高い。とくに社会調査、意見調査データなどの分析にかかわる人達からの要請には、かなりなものがある。ところでクラスター分析 (Cluster Analysis) とはそもそも何であろうか。ごく簡単にいえば、多次元データ解析のうち、ものを分けるという問題を扱う部分に関連した、かなり総括的な概念であるといつてよい。そして、分類に関する問題は様々の観点から眺めることができる。たとえば、

(I) 事前に、分ける群に関する情報が与えられて、あるいは群の存在がわかっていて、これに観測したデータをわりあてるとき、

(II) 群があるとして、その数を決め、何らかの基準によって群を作り出すことを考える、

(III) 対象をいくつかの群に分けることが適当か、また群が存在するか知りたいとき、
等等、いろいろ考えられよう。

(I) に該当する典型的な方法論として判別分析法が挙げられる。そしてクラスター分析は主として(II)、(III)あたりを扱う方法論であり、与えられたデータをクラスター化する技法を大まかに表わす術語としてクラスター分析が使われているようである。判別分析などがクラスター分析と異なる点は、それが少なくとも群に関する事前情報が何らかの形で与えられている点にある。また状況にもよるが判別分析では少なくともモデルの概念がはっきりとしている(たとえば線形判別である、場合によっては多変量正規分布を仮定するなど)。しかしクラスター分析ではこうしたモデルに関する約束ごとがはっきりしないことが多い。また、現状ではどれがクラスター分析の手法であるといった厳密な枠を用意して説明することが難しい。ある場合は数値分類法 (Numerical Taxonomy) を指し、また狭義のクラスター分析の意味で測定項目の分類として考えられることもある。さらにパターン認識などでは、事前情報を与える予備データのない分類という意味で、教師なしのパターン認識という観点からクラスタリングを捕えている。いずれにせよ、分類に関わる相当広い範囲の概念がクラスター分析という術語で代表されているといつてよい。したがって、共通の問題として、それぞれの分野で使用される用語の混乱、手法の乱立、さらにクラスターをどう定義するかといった点の十分な約束がなされないまま算法だけが計算機ソフトウェアの進歩に並行して発達してきたといつてよい。

現在のようにクラスター分析が注目をあひるようになった理由として、データ解析で扱うデータの性格が多様化したこと、組み合わせ的、反復試行的な技法が要求される計算法を可能にする計算機の支援が十分えられるようになったこと、などが挙げられよう。データの形式だけ考えてみても、(個体)×(変数)の計量データを扱うという従来の多変量解析の枠内に留まらず質的データ(カテゴリーカル・データ)を扱う、あるいは、対象間の関連性を示す類似性の尺度だけが与えられる場合を扱うなど様々である。こうなると、もはや強い制約に縛られた方法論だけでは問題を扱いきれず、別の立場からの解析法としてクラスター分析が注目を集めるようになったことは当然のなりゆきといつてよい。こういう点で、多次元尺度解析や数量化理論に通ずるものがある。しかし、現状では、はなはだ問題が多く、利用経験者の失望を買っていることも事実であろう。だがこうした例の多くは、手法の誤用と過剰な期待、つまりデータ自体の持つ情報以上の何かを期待しすぎること、によるもので、これからは、手法の使い方、運用面の技術といった点が重視されるようになる。

ここでクラスター分析の現状を全てにわたって述べることは無理なことであるが、最近、「分類法の分類」という観点から多くのレビューが現われている。このうち、代表的なものをいくつか文末に挙げたので多くを知りたい読者はこれらを参考にするとよい。むしろ、ここでは1つの試みとしてある手法に話題を限定して、それを考えるだけでも様々の立場から議論が可能であることを示すに留めよう。

2. いわゆる Single Linkage 法について

扱うデータの様式として一般的なものの1つは対象間の関連の度合が類似度あるいは非類似度(または距離)で与えられる場合であろう。いま、 n 個の分類対象を、

$$X = \{O_1, O_2, \dots, O_i, \dots, O_n\}$$

で表わす。そして対象(ここでは個体と呼ぶことにする) O_i と O_j との差異を表わす数値、つまり非類似度(差異度)を考えることにして、これを $d(O_i, O_j)$ または簡単に d_{ij} と書く。すると全体の個体間の関連は、行列、

$$D = (d_{ij}) \quad (i, j = 1, 2, \dots, n)$$

で表わせる。ここで d_{ij} はつぎの2つの条件を満たすものとしよう。

- (i) X 内の任意の O_i に対して、 $d_{ii} = 0$ (反射律)
- (ii) X 内の任意の個体 O_i, O_j に対して

$$d_{ij} = d_{ji} \quad (i \neq j) \quad (\text{対称律})$$

もちろん推移律として三角不等式が成り立つならばメトリックな(通常の距離の公理を満たす)距離として扱うことができ都合がよい。しかし、現実には、(i)、(ii)しか満足しないときもあるし、場合によっては(ii)の成立も怪しくなることすら考えられる。ある種の手法は、こうしたデータを扱うという点で多次元尺度解析と非常に類似する。とくに、代数的な関係(relation)の概念や、グラフ理論の知識を活用する手法がこうしたデータに適していて、実際、いくつかの(凝集型の)階層的な手法はこうした観点から議論されている。

ところで、階層的(hierarchical)とは、 d_{ij} から作られるつぎのような性質をいう。いま、 X のすべての可能な分割の集合を $S(X)$ と書く。このとき1つの階層構造(あるいは階層的分割)を水準 l_k により $\{l_0, l_1, \dots, l_k, \dots, l_{n-1}\}$ と表わすと、これがつぎの性質を満たすことである。

$$(P1) \quad l_k \in S(X), \quad 0 \leq k \leq n-1$$

(P2) 水準 l_0 では、個体は個々のクラスターを作る、つまり1つ1つの個体が1つのクラスターで、 n 個のクラスターがある。そして、水準 l_{n-1} で全個体が1個のクラスターを作る。

(P3) $l_0 \leq l_1 \leq l_2 \leq \dots \leq l_k \leq \dots \leq l_{n-1}$ であって、ある水準 l_{k+1} は l_k までの情報にもとづいて作り出すことができる。

つまり、(図1)のように X 内の個体間に1つの単調な階層的結合関係を作り出すことであり、こうした性質をもつ手法を単調不変(monotone invariant)な階層的な手法というのである。

こうした手法の中でもっともよく知られているものに、Single Linkage 法(SL 法と略す)と Complete Lin-

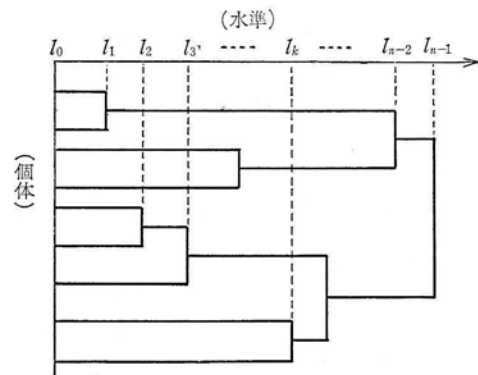


図1 単調不変な階層構造の例

kage 法 (CL 法と略す) がある。これは、あるクラスター C_i と C_j を結合して、水準 l_k におけるクラスターの集合、 $\{C_1, C_2, \dots, C_{n-k}\}$ を作る方式をつぎのように考えるものをいう。

SL 法のと看

$$\begin{aligned} \min \{d_{ij} | O_i \in C_i, O_j \in C_j\} \\ = \min [\min \{d_{im} | O_i \in C_i, O_m \in C_m\}] \quad (1) \\ (1 \leq l, m \leq n-k, l \neq m) \end{aligned}$$

CL 法のと看

$$\begin{aligned} \max \{d_{ij} | O_i \in C_i, O_j \in C_j\} \\ = \min [\max \{d_{im} | O_i \in C_i, O_m \in C_m\}] \quad (2) \\ (1 \leq l, m \leq n-k, l \neq m) \end{aligned}$$

(1) はクラスター間の個体間の非類似度のうちの最小値を考える。一方 (2) は個体間のもっとも離れた、最大の非類似度を採用する。こうしたことから、これら2つの手法の性質は全く相反するものであることが予想される。またいずれも個体間の連結性だけに注目し、クラスター内の個体数 (クラスター・サイズという) やチラバリといった情報は考えに入れていない。

ここで、手法としてはもはや古典的ともいえる2つの方法を挙げたわけは、これらがいろいろな意味で、クラスター分析のあらゆる面を代表しているように思われるからである。これは、クラスター分析の適用分野の多様性という点1つをとってみてもこれが理解できる。たとえば上の2つの手法と同義あるいは同義とみてよい手法として、

SL 法の場合

- (a) nearest neighbor method (最近隣法)
- (b) minimum method (Johnson)
- (c) elementary linkage analysis (McQuitty)
- (d) connected method (Bonner 他)
- (e) minimal spanning tree (MST) のあるもの

CL 法の場合

- (a) furthest neighbor method (最遠隣法)
- (b) maximum method (Johnson)
- (c) rank order typal analysis (McQuitty)
- (d) diameter analysis (Hubert, Ling, 他)

と、容易にいくつかを列挙できる。そしていずれもグラフや関係 (主にブール代数) の概念を利用する点や、算法中心であるところも似ている。また、Lance と Williams は組み合わせ的な表現を借りると、こうした手法も含めて、一連の階層的な手法をまとめて表わすことが可能になることを提案した。それはクラスター C_i と C_j が結合

して C_i が新たに作られるとき ($C_i \cup C_j = C_{i+j}$ と書く)、 C_i と C_i, C_j 以外のクラスターとの距離 d_{ir} を、つぎの式で与えるのである。

SL 法のと看

$$\begin{aligned} d_{ir} &= \min \{d_{ir} | O_i \in C_i, O_r \in C_r\} \\ &= \frac{1}{2} d_{ir} + \frac{1}{2} d_{jr} - \frac{1}{2} |d_{ir} - d_{jr}| \\ &= \min \{d_{ir}, d_{jr}\} \quad (3) \end{aligned}$$

CL 法のと看

$$\begin{aligned} d_{ir} &= \max \{d_{ir} | O_i \in C_i, O_r \in C_r\} \\ &= \frac{1}{2} d_{ir} + \frac{1}{2} d_{jr} + \frac{1}{2} |d_{ir} - d_{jr}| \\ &= \max \{d_{ir}, d_{jr}\} \quad (4) \end{aligned}$$

つまり式 (1), (2) は C_i と C_j の関係を示し、式 (3), (4) は C_i と C_j が結合したとき、 C_i, C_j 以外の他のクラスターと、結合でえられた C_i との関係を示している。式 (3), (4) をまとめて、水準 l_{k+1} における結合時の距離の更新を、関係式

$$\begin{aligned} d_{ir} &= \alpha_i d_{ir} + \alpha_j d_{jr} + \beta d_{ij} + \gamma |d_{ir} - d_{jr}| \\ &\quad (r=1, 2, \dots, n-k; r \neq i, j) \end{aligned} \quad (5)$$

で与えることができる。この式で $\alpha_i = \alpha_j = 1/2$, $\beta = 0$, $\gamma = 1/2$ または $\gamma = -1/2$ と与えると確かに CL 法, SL 法となる。また、 C_i, C_j のクラスター・サイズをそれぞれ n_i, n_j で表わすと、

$\alpha_i = n_i / (n_i + n_j)$, $\alpha_j = n_j / (n_i + n_j)$, $\beta = -\alpha_i \alpha_j$, $\gamma = 0$ のとき、セントロイド法、また、

$$\alpha_i = n_i / (n_i + n_j), \alpha_j = n_j / (n_i + n_j), \beta = \gamma = 0$$

のとき群平均法、というように、他の手法も同一の式で表わすことができる。Lance らはあわせて7つの手法が上の関係で記述できることを示し、Wishart は J. Ward の提案した方法も同じ表現を借りて記述できることを示し、実際に彼の作成になるプログラム CLUSTAN の中で利用できるようにした。彼らの一連の考え方は、計算機向きであって計算法としては非常に卓越したものであるが、個々の手法の性格という観点から眺めると利用に際し十分な注意がいる。たとえば、ノンメトリックな非類似度などに対しては、ウォード法や群平均法の利用は多少むりがある。それは、これらの方法は本来ユークリッド距離のようなメトリックな距離に対して考えられたものだからである。もちろん、形式的には式 (5) により計算が可能であるが、それはあくまで形式的にということわりの上で利用できることである (詳しくは

Anderberg¹⁾, Sneath²⁶⁾ などを見るとよい).

3. SL 法による数値例

ここで、ごく簡単な例を見よう. つぎの行列 D は Peay²⁴⁾ が引用したデータである. 内容はハワイにおける相異なる人種間の結婚の割合を示す数で, 値が大きいほどその人種間の組み合わせに対する結婚の割合は低くなるように, つまり差異を表わすように尺度化してあるという.

$$D = \begin{matrix} & \begin{matrix} O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 \end{matrix} \\ \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_5 \\ O_4 \\ O_7 \\ O_8 \end{matrix} & \begin{bmatrix} 0 & 6 & 21 & 30 & 18 & 27 & 33 & 32 \\ & 0 & 12 & 21 & 14 & 16 & 23 & 23 \\ & & 0 & 20 & 22 & 24 & 24 & 20 \\ & & & 0 & 19 & 37 & 41 & 37 \\ & & & & 0 & 30 & 30 & 28 \\ & & & & & 0 & 24 & 21 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{bmatrix} \end{matrix}$$

ここで8つの人種はそれぞれつぎの通りである.

- O_1 : Hawaiian (ハワイ人)
- O_2 : Part-Hawaiian (ハワイ系混血)
- O_3 : Caucasian (白人)
- O_4 : Puerto Rican (プエルト・リコ人)
- O_5 : Filipino (フィリピン人)
- O_6 : Chinese (中国人)
- O_7 : Japanese (日本人)
- O_8 : Korean (韓国人)

行列 D の観察から, 日本人は他の人種との結婚率が低いこと, 白人は有色系, とくにハワイアンと結婚する割合がやや高い傾向にあることなどがある程度は読みとれる. いま SL 法, CL 法の2つの方法によってクラスタリングを行い結果の階層構造を, いわゆるデンドログラム(樹木図)として書いたものが図2の(a), (b)である. SL 法を例にとると, この図は, つぎのようにして得られる.

まず各個体を1個ずつのクラスター $\{O_1, O_2, \dots, O_8\} \equiv \{C_1, C_2, \dots, C_8\}$ と考える(水準 l_0 にあるとき). つぎに, D の要素のうちの最小値をさがし, それに対応する個体を結合する. いま $d_{12}=6$ であるから O_1 と O_2 を結合しこれを新たに C_1 とする. つまり $O_1 \cup O_2 \equiv C_1 \cup C_2 (\equiv C_1)$ とする. これは式(3)で $i=1, j=2, t=1$ としたことに相当する. そして新しいクラスター $C_t \equiv C_1$ と残りのクラスター $\{C_3, C_4, \dots, C_8\}$ との関係を式(3)により更新する. SL 法(および CL 法)の重要

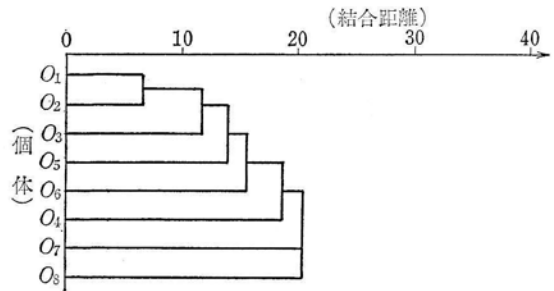


図 2-(a) SL 法の結果

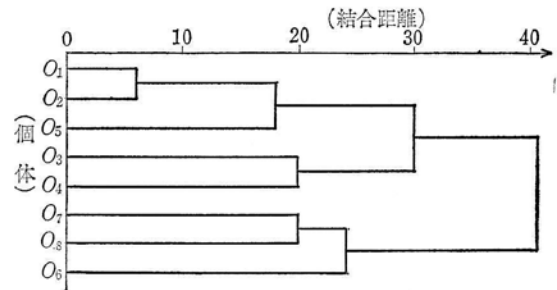


図 2-(b) CL 法の結果

な点は, この更新が単なる数値の大小関係の比較による置きかえにあるということである. つまり, O_1, O_2 の結合時の様子は下のようになる.

r	C_3	C_4	C_5	C_6	C_7	C_8
$i: O_1 \equiv C_1$	21	30	18	27	33	32
$j: O_2 \equiv C_2$	12	21	14	16	23	23
$C_1 \equiv C_1 \cup C_2$ (結合後の新しい数値)	12	21	14	16	23	23

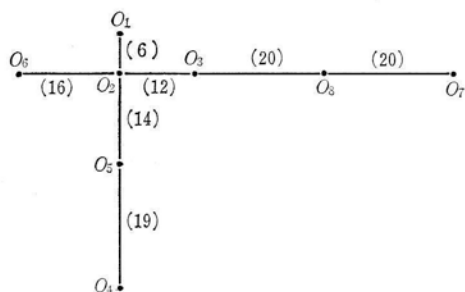
こうして, 次元が1つ退化した7次元の行列ができるから, この行列に対し上と同じ手続きを繰り返すと表1の(a)欄の結合情報と(c)欄の距離をうるから図2-(a)のデンドログラムができる. CL 法の場合は式(4)の関係に従って更新を行えばよい. 図2-(b)が CL 法の結果であるが SL 法にくらべてやや傾向が異なる (D

表 1 SL 法および MST の結果

ステップ	(a) SL 法するとき		(b) MST のとき		(c) 結合距離
	i	j	i	j	
1	1	2	1	2	6
2	1	3	2	3	12
3	1	5	5	2	14
4	1	6	6	2	16
5	1	4	4	5	19
6	7	8	3	8	20
7	1	7	7	8	20

のように、 d_{ij} に同じ値—tie という—があるとき、手法によっては結合の個体番号の選び方によってある種の矛盾がおこることが知られている).

ところで SL 法するとき、結合時の標識 (クラスター番号) のつけかえに少し工夫をすると別の表現も可能である. 上の方式ではクラスター C_i と C_j の結合のとき i と j との大小をくらべて $i < j$ のときに新たに出来るクラスター $C_i = C_i \cup C_j$ の標識 i を i にすると約束した (あるいはその逆でもよいが、とにかく約束しておくことがあとでデンドログラムに表わすときに必要となる). しかし、別の考え方として、式(1)の定義に忠実に従って、クラスター間の最も近い個体のもつ個体番号を利用することにすると別の結合情報がえられる. たとえば、Peay の例の場合に、最初の O_1 と O_2 の結合で $C_1' \equiv \{C_1, C_2\} \equiv \{O_1, O_2\}$ がえられ、つぎにこれと $C_3 (\equiv O_3)$ が結合するとき、 d_{13} と d_{23} をくらべて小さい方の個体番号を採用することにする. D についてこの規則を適用すると、表1の(b)欄の結合リストが得られる. これをもとに各個体を結ぶと図3が得られる. これは Minimal Spanning Tree (MST) として知られるグラフの一種で、spanning tree (張り木) のうち弧の上の距離 (非類似度) の総和が最小になるものを求めたことに相当する. ここで張り木とは閉じた巡回路のないグラフのことである. こうして SL 法とは MST の求め方の1つであることがわかる. したがってデンドログラムと MST は同じ情報を現わしているが、個体数が少ない場合や個体間の関連性を良く観察したいときなどは MST を利用すると具合がよいこともある. 上の例でも、日本人とプエルト・リコ人の間の差異は大きく、ハワイアンと白人の間は親密であるなどの関係を MST の上できれいに読みとることができる.



(注) 形にとくに意味はない. なるべくもとの個体間の関係に近いように書いた. () 内は非類似度の大きさである.

図3 MSTの1つの表現

4. SL 法の別の見方

いま、デンドログラムから各水準における個体間の結合時の非類似度を読みとり、1つの行列を作ろう. たとえば O_1 と O_6 は水準 l_4 で結合し、このときの非類似度は16であるから $d_{16}^* = 16$ とおく. こうして得られる行列はもとの行列とは必ずしも一致しないから d_{ij}^* で O_i と O_j の差異を示すのである. こうして得られた行列がつぎの D^* である.

$$D^* = \begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 \\ \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \end{matrix} & \begin{pmatrix} 0 & 6 & 12 & 19 & 14 & 16 & 20 & 20 \\ & 0 & 12 & 19 & 14 & 16 & 20 & 20 \\ & & 0 & 19 & 14 & 16 & 20 & 20 \\ & & & 0 & 19 & 19 & 20 & 20 \\ & & & & 0 & 16 & 20 & 20 \\ & & & & & 0 & 20 & 20 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{pmatrix} \end{matrix}$$

ここで、任意の O_i, O_j, O_k に対して、つぎの関係がある.

$$d_{ij}^* \leq \max \{d_{ik}^*, d_{jk}^*\} \quad (6)$$

たとえば、 $i=7, j=4, k=1$ としてみると、

$$d_{74}^* = 20, d_{71}^* = 20, d_{41}^* = 19$$

でたしかに(6)が成り立つ. その他の場合も同様である.

(6)は単調な階層構造をもつデンドログラムを解として与えるような手法が共通にもつ性質で、これが成立するとき、ultrametric であるという. つまり、式(i), (ii)の関係にこの(6)を推移律として加えた場合である. このとき明らかに、 $d_{ik}^* + d_{jk}^* \geq \max \{d_{ik}^*, d_{jk}^*\}$ の関係があるから、通常の三角不等式を推移律とするよりも強い意味でメトリックであるとして、 D^* に限らず、(6)が成立する場合を ultrametric と呼んでいる. つまり、SL 法や CL 法は ultrametric な距離関係を作り出す方式のある場合に相当する. これは、もともと階層構造にない個体間の関連を、無理に階層構造を借りて表現したことでもある. したがって、手法としてはなるべくもとの距離関係に矛盾しないようなひずみの少ない階層構造となることが望ましいわけである. Jardine, Sibson といった英国の研究者はこの観点から手法の比較をし、いくつかの基準の用意のもとに、SL 法がそれらの基準をみたす良い方法の1つであるとしている. 一方、実用上からは SL 法は隣接する個体をつぎつぎと結合して増殖する、いわゆる鎖状効果 (chaining effect) が起こるので良い方法とはいえないと指摘する人もいる. しかし、こうした議論で忘れられている点は、手法はデ

ータに依存するものであること、またクラスターをどう考えるかによること、である。似ているもの同士は積極的に寄せ集めて本当にかけ離れているもの、異常値に近いもの、を除きたいという利用目的からは SL 法はかなり良い結果を生むし、上の例に見るように MST の表現をかりてもかなり違った結果となりうるのである。またなるべく塊状に寄せ集めて（実際はそうでなくても）切り離したようなクラスターが作りたいのであれば CL 法が向いている。また、何としても、似ているという群を見つけないのであれば、1つの方法ではなくいくつかの方法でクラスタリングした結果をくらべて、どの手法でも群となるものをクラスターと考えればよい。そのときも、性質がかなり異なる手法を使うようにする。このように、使い方の上での工夫が、モデルという概念がはっきりしない階層的手法などを利用するときには特に必要となるのである。この、手法をどう運用して使いこなすかという部分が、これからのクラスター分析の課題の1つであろう。

ところで、上の手法を別の概念により説明することもできる。ファジィ集合論を提唱した L. A. Zadeh は、(i), (ii)および(6)が成立する場合を、ファジィ論の立場からの同値関係 (fuzzy equivalence relation) の特別な場合で、非類似度関係 (dissimilarity relation) にあたるとして証明を与えている³⁰⁾。この観点に立つとつぎの演算によるクラスタリングを考えることができる。

いま非類似度（あるいは距離）行列 D に対してつぎの演算を用意する。

$$\begin{cases} D^{(0)} \equiv D \quad (n \times n \text{ 次行列}) \\ D^{(l+1)} = D^{(l)} \otimes D^{(0)} \quad (l=0, 1, 2, \dots) \end{cases} \quad (7)$$

ここで $\otimes$ は、

$$c_{ij} = \min_k \{ \max \{ a_{ik}, b_{jk} \} \} \quad (8)$$

という演算を行うことを表す。また a_{ij} , b_{ij} , c_{ij} はそれぞれ行列 $D^{(0)}$, $D^{(1)}$, $D^{(l+1)}$ の要素である。つまり、

$$\begin{aligned} D^{(1)} &= D^{(0)} \otimes D^{(0)} \\ D^{(2)} &= D^{(1)} \otimes D^{(0)} \equiv (D^{(0)} \otimes D^{(0)}) \otimes D^{(0)} \\ D^{(3)} &= D^{(2)} \otimes D^{(0)} \equiv (D^{(0)} \otimes D^{(0)} \otimes D^{(0)}) \otimes D^{(0)} \\ D^{(4)} &= D^{(3)} \otimes D^{(0)} \\ &\dots\dots\dots \\ D^{(l+1)} &= D^{(l)} \otimes D^{(0)} \\ &\dots\dots\dots \\ &\dots\dots\dots \end{aligned}$$

と繰り返す。一般に l は高々 $l \leq n$ であり、 $l' (\leq n)$ で $D^{(l')} \equiv D^{(l'-1)}$ (一種のベキ等関係) に達したときの $D^{(l')}$ を到達 (可能) 行列 (reachability matrix) と呼んでいる。これは $(0, 1)$ 要素からなるブール行列 (あるいは通常のグラフ) の場合に成り立つ類似の性質を一般化したものに相当する。実は、この方式で作られた行列 $D^{(l')}$ は式(6)の性質を持ち ultrametric となる。つまり、 D からえられた $D^{(l')}$ はファジィ関係という非類似度関係になっているのである (これらの性質の詳細は Kaufmann¹⁸⁾ を見るとよい)。例とした、Peay も式(8)の方式でクラスタリングを行うことを提案した1人であるが、性質の詳しい点には立ち入っていない。Hubert, Matula, Ling といった人達はグラフ理論と組み合わせ理論の観点からこうした問題を扱っている。

こうした議論のうちの興味ある性質の1つに、式(8)からえられる $D^{(l')}$ と、SL 法によってえられる D^* との関係を示すことができる (Peay, Hubert らは両者の一致の可能性を示唆しているが、もう少し詳しい証明が

$$D^{(1)} = \begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 \\ \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \end{matrix} & \begin{bmatrix} 0 & 6 & 12 & 19 & 14 & 16 & 23 & 21 \\ & 0 & 12 & 19 & 14 & 16 & 23 & 20 \\ & & 0 & 20 & 14 & 16 & 20 & 20 \\ & & & 0 & 19 & 21 & 23 & 20 \\ & & & & 0 & 16 & 23 & 22 \\ & & & & & 0 & 21 & 21 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{bmatrix} \end{matrix}$$

$$D^{(2)} = \begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 \\ \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \end{matrix} & \begin{bmatrix} 0 & 6 & 12 & 19 & 14 & 16 & 21 & 20 \\ & 0 & 12 & 19 & 14 & 16 & 20 & 20 \\ & & 0 & 19 & 14 & 16 & 20 & 20 \\ & & & 0 & 19 & 19 & 20 & 20 \\ & & & & 0 & 16 & 22 & 20 \\ & & & & & 0 & 21 & 20 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{bmatrix} \end{matrix}$$

$$D^{(3)} = \begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 \\ \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \end{matrix} & \begin{bmatrix} 0 & 6 & 12 & 19 & 14 & 16 & 20 & 20 \\ & 0 & 12 & 19 & 14 & 16 & 20 & 20 \\ & & 0 & 19 & 14 & 16 & 20 & 20 \\ & & & 0 & 19 & 19 & 20 & 20 \\ & & & & 0 & 16 & 20 & 20 \\ & & & & & 0 & 20 & 20 \\ & & & & & & 0 & 20 \\ & & & & & & & 0 \end{bmatrix} \end{matrix}$$

いるように思う)。いま式(8)により $D^{(1)}$ を作ると、そして $D^{(3)} \equiv D^{(4)} \equiv D^{(5)} \equiv \dots$ となつてここで演算は終了する。この場合確かに $D^{(3)} \equiv D^*$ となっており、 $D^{(3)}$ の要素を順に小さい方からひろって個体をつなぐと図2—(a)と同じデンドログラムとなる。また演算の反復回数、いまの例では3、がもとの行列 D が階層構造にどれ位近いかという1つの目安となっている。たとえば、 $D^{(1)} \equiv D^{(0)}$ となったならば、全く矛盾なくもとの個体間の関連は階層構造として表わせることを示している。こうした演算処理はもとの個体間の距離関係をあいまい化する、あるいは平滑化する働きをもっていると考えられることもできる。形式的には、様々の演算を考えることもできるわけで、Hubert らは、

$$\begin{cases} M^{(0)} \equiv D \\ M^{(l+1)} = M^{(l)} \oplus M^{(0)} \quad (l=0, 1, 2, \dots) \end{cases}$$

ここで、 a_{ik} , b_{jk} , c_{ij} はそれぞれ $M^{(0)}$, $M^{(l)}$, $M^{(l+1)}$ の要素であり、

演算 $\oplus$ は、

$$c_{ij} = \min_k \{a_{ik} + b_{jk}\} \quad (9)$$

とする例を示し、これで l に適当な数を与えて出来上がった行列に対して CL 法や SL 法をほどこす1つの折衷案を出している。また式(8)の方法でも類似の手続きによって処理することも考えられよう。

5. むすび

あえて、SL 法という1つの方式に関連した話題を紹介するという立場をとった。こうした狭い枠内に限ってもいろいろな立場から議論が進められている。また次第に算法中心の考え方から、手法相互の関連を吟味する評価方式の検討^{17), 21)}、クラスターをどう考えて手法を使いこなすか、といった手法の運用面の問題などが中心になりつつあるという点を強調しておこう。

参 考 文 献

- 1) Anderberg, M.R.; *Cluster Analysis for Applications*, Academic Press. 1973.
- 2) Benzécri, J.; *L'Analyse des Données: Tom I (La Taxinomie)*, Dunod. 1973.
- 3) Bijnen, E.J. and Stouthard, Ph. C.; *Cluster Analysis: Survey and Evaluation of Techniques*, Tiburg Univ. Press. 1973.
- 4) Blashfield, R.K.; A Consumer Report on the Versatility and User Manuals of Cluster Analysis Soft-

- ware, *Proc. of the Statistical Computing Section, ASA*, 31-37, 1976.
- 5) Bock, H.H.; *Automatische Klassifikation*, Vandenhoeck und Ruprecht. 1974.
- 6) Clifford, H.T. and Stephenson, W.; *An Introduction to Numerical Classification*, Academic Press. 1975.
- 7) Cole, A.J. (ed.); *Numerical Taxonomy*, Academic Press. 1969.
- 8) Cormack, R.M.; A Review of Classification, *J.R. Statist. Soc., ser. A*, **134**, 321-367, 1971.
- 9) Duran, B.S. and Odell, P.L.; *Cluster Analysis: A Survey*, Springer Verlag. 1974.
- 10) Enslein, K., Ralston, A. and Wilf, H.S.; *Statistical Methods for Digital Computers*, Vol. III of *Mathematical Methods for Digital Computers*, John Wiley. 1977.
- 11) Everitt, B.; *Cluster Analysis*, John Wiley. 1974.
- 12) Fisher, W.D.; *Clustering and Aggregation in Economics*, The John Hopkins Press. 1969.
- 13) Gower, J.C.; Classification Problems, *Bull. of I.S.I.* 1973.
- 14) Hartigan, J.A.; *Clustering Algorithm*, John Wiley. 1975.
- 15) Hubert, L.J.; Some Applications of Graph Theory to Clustering, *Psychometrika*, **39**, 3, 283-309, 1974.
- 16) Hubert, L.T.; Hierarchical Clustering and the Concept of Space Distortion, *The Br. J. Math. Statist. Psychol.* **29**, 2, 121-133, 1975.
- 17) Jardine, J. and Sibson, R.; *Mathematical Taxonomy*, John Wiley. 1971.
- 18) Kaufmann, A.; *Introduction a la théorie des sous-ensembles flous*, Masson et Cie. 1973.
- 19) Lerman, I.C.; *Les Bases de la Classification Automatique*, Gauthier-Villars. 1970.
- 20) Matusita, K. and Ohsumi, N.; Evaluation procedure of clustering techniques, *France-Japan Seminar*, Paris, March. 13-20, 1978.
- 21) 大隅 昇; クラスター分析, 現代数学, **10**, 9, 1977.
- 22) 大隅 昇; 多次元データの処理分析, No. 6-7, クラスター分析, 広告月報, No. 214, 215.
- 23) 大隅 昇; クラスター分析, 最新醫學, **33**, 1, 1978
- 24) Peay, E.R.; Nonmetric Grouping: Clusters and Cliques, *Psychometrika*, **40**, 3, 297-313.
- 25) Ryszin, J. (ed.); *Classification and Clustering*. Academic Press. 1977.
- 26) Sneath, P.H.A. and Sokal, R.R.; *Numerical Taxonomy*, Freeman. 1973.
- 27) Tryon, R.C. and Bailey, D.E.; *Cluster Analysis*, McGraw Hill. 1970.
- 28) 矢島敬二, 他; 「クラスター・アナリシス(1)―(5)」, オペレーションズ・リサーチ, Vol. 16, No. 8~No. 11, 1971.
- 29) 矢島敬二, 大隅 昇; 「統計」, bit 増刊, アプリケーション・プログラム, 7月, 1977.
- 30) Zadeh, L.A.; Similarity relations and fuzzy ordering, *Inf. Sciences*, **3**, 177-200, 1971.

(おおすみ・のぼる, 統計数理研究所)



古代ギリシャ 球技に興じる少年 大理石製, BC500年頃

本号に掲載された広告主

日本ユニバック／表4

ビジネスコンサルタント／表2

デンサン／表3

大修館／1

情報処理開発協会／4

岩波書店／15

産業図書／33

同文書院／47

現代数学社／52

朝日出版社／64

海洋科学／61

オペレーションズリサーチ／73

特集——スポーツの力学

《特集》

スポーツと抵抗	5	種子田定俊
射る＝弓道における手の内の力学	10	小林 一敏
打つ＝ゴルフの力学	16	河村 龍馬
泳ぐ＝水にのる	21	高木 隆司
漕ぐ＝自転車の力学	25	吉福 康郎
飛ぶ＝ハンググライダー	35	東 昭
走 る	43	小野 勝次

《連載》

カタストロフィー夜話Ⅵ トムの生物学への挑戦 2	50	野口 広
--------------------------	----	------

基礎暗号学 4 使用度数 その 4	55	加藤 正隆
-------------------	----	-------

リレー連載③

パズル パズル玩具DISCONについて	65	西山 輝夫
---------------------	----	-------

流体数学のすすめⅡ

超関数論 XXI 周期超関数のたたみこみ	70	今井 功
----------------------	----	------

《書評》

M. ガードナー著 一松 信訳 数学カーニバルⅠ,Ⅱ	48	有澤 誠
----------------------------	----	------

青木国夫他著 思い違いの科学史	49	市村 浩
-----------------	----	------

雑談書評 数理論理のいろいろな話題	68	細井 勉
-------------------	----	------

クラスター分析 SL法をめぐる	77	大隅 昇
-----------------	----	------

数理科学前号内容／75

数理科学次号予告／75

別冊「アインシュタイン」内容

数理科学常備店一覧／42

別冊「パズル」Ⅱ内容／67

洋書新刊近刊案内／54

表紙 安野 光雅

図版 武田 幸経