多次元データの処理分析　No.5

クラスター分析

統計数理研究所
大隅　昇

1. 概　観

数量化理論、多次元尺度解析法と続いた本連載の最後の話題としてクラスター分析を取り上げよう。これに関連した話題は、すでに本誌の'77.12月号で紹介したが、今月は利用上の注意、手法の限界、扱うデータの性格との関連など、主として使用上の問題に焦点をあてて考えよう。

クラスター分析が多次元データ解析の新しい手法として注目を集めるようになったのは1960年代以後である。とくに国内では70年代に入ってからのことでその利用の範囲も限られていた。たとえば、1つに Lance, Williams らの組み合わせ的な手法を中心とする代表的な階層的手法を取り上げた D. Wishart の作成になるプログラム CLUSTAN-1A の紹介と適用例がある。また、マーケティング分野での調査データ分析、とくにライフスタイル分析やマーケティング・セグメンテーションの分析などに積極的に利用されてきた AID(Automatic Interaction Dctector)、Association Analysis なども同じ枠組の中に入れてよい。そして Sokal, Sneath に代表される生物学の分野の研究者の提案になる数値分類法の術語で代表される一群の手法、それは主として系統的あるいは階層的分類法であるが、を適用した生物学、生態学、医学などの分野での応用例も多く見られる。さらに、気象衛星、資源探査衛星などから送信された画像データを扱うリモート・センシングの分野のデータ解析の手法として、いまやクラスタリングは欠かせないものとなっている。

このように国内に限っても次第にその適用分野、扱う問題の性格、分析の目的は多様となりつつある。これはとりも直さず〝クラスター分析〟という用語の解釈が幅広いものであることを示している。いわゆる〝クラスター分析〟とは主観をまじえずなるべく客観的に〝もの〟を分けることを目的とする手法の総称といえる。そこではコンピューターの役割は不可欠で、クラスター分析の歴史的変遷はまさにコンピューターのそれと並行しているという事実に何ら不思議はない。コンピューターとの関連で単なる計算法としてクラスター分析をとらえるならば〝自動分類〟とでもいいかえてもよい。事実現在のクラスター分析の手法と称しているものの大部分は算法中心である。しかし現実はデータとの絡みでおこる手探りの部分が相当にあることが問題なのである。とくに昨今のように取り扱うデータ量が増大するとコンピューターを使って分類の情報を瞬時に提供することが要請される。そしてあまりにも容易に情報が手に入ることから、ともするとクラスタリングそのものによって何か高度の情報、結論が引き出せると錯覚して、この新しい未熟な手法に過大の期待をかける向きも少なくない。またクラスター分析を多変量解析の手法、たとえば因子分析、判別分析、主成分分析などと同列に置いて論ずる人も多いが、データ構造の解釈に対してこうした手法よりもずっと柔軟な姿勢で臨む、あるいははっきりしたモデルがない場合が多い、という点でこれらの方法論とは多少趣きを異にする。

データ解析という大きな枠組の中で考えると、〝ものを分ける〟という問題は分析過程のあらゆる場面で顔を出す。これに臨機応変に立ち向かう

には、算法、技法（ハードな側面）の取り扱いにとどまらず、それらを状況に応じて有効に使いこなす、つまりデータの手探りの道具としての使い方(ソフトな側面)の比重がはるかに高い。にもかかわらず、クラスター分析の現状は、算法（またはプログラム）が先行し、これを利用するソフトの技術が未熟である。もちろんコンピューターの支援は不可欠であるから、これを自由に使いこなせる技術と環境も大切である。しかし一般には湯水のごとく利用することは難しいので自然にある限られた範囲内の処理に終わることが多い。ところがクラスター分析は試行錯誤的、発見的な部分が多い上に、理論、応用面とも発展過程にあるため活用する上でさまざまの支障が生ずるのである。

2. クラスターとは

〝クラスター（cluster＝集落、集塊の意)〟とはそもそも何か、ごく常識的な表現をすれば、似ているもの、等質なものの集まりである。そしてこの〝似ている〟、〝等質な〟という言葉の解釈をめぐってさまざまの算法が生まれたといってよい。等質なものを考えることは相対的に非等質なものが存在することであり、この似ていること(類似性）と離れていること（差異性）を定量的に把握しようという方法の1つがクラスター分析である。クラスター分析では次の点が重要である。

(1) 観測特性の数値化（コード化)。
(2) クラスターの定義、あるいは類似性、差異性を表すものさしの作成（具体的には類似度、距離、等質性基準といったもの)。
(3) それを使ってクラスターを生成するクラスター化の技術（いわゆる算法)。
(4) 具体的な計算プログラム。

これに加えて、実用上は手法をどのように使いこなすかという使い方が重要であり、本連載の意図もこの点にあると思うので、個々の手法の話は類書にゆずり、ミニチュアなデータによる実験を通してクラスター分析の性質の一端を紹介することにしたい。

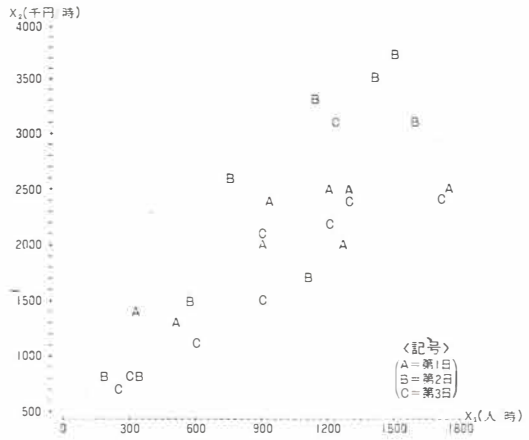
3. 階層的分類法の特徴

（表1）は都心の某商店の調査データである。測定項目として、年末のある3日間に、1時間ごとに入店者数 X_1(人/時)、売上金額 X_2(千円/時)、金銭出納機の使用回数 X_3(回/時）を記録してある。日、時は数値化されカテゴリカル・データとなっている。これに対し入店者数、金額などは計量的データである。まず分析の常套手段としてヒストグラム、散布図などの図的表示、簡単な統計量の算出などを行うであろう。また、日、時などの情報も積極的に利用するであろうから、これを

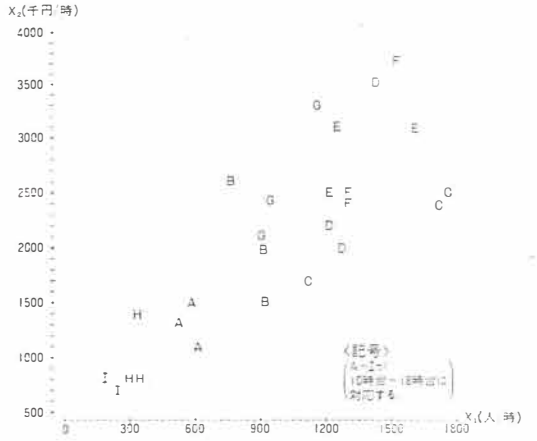
(表1)

（A）							（B）			（C）		
日・曜日		時間		個体番号	入店者数(人/時)	売上金額(千円/時)	金銭出納機回数(回/時)	正規化したデータ			X_1、X_2、X_3をカテゴリー化したデータ		
日	コード	時間	コード		X_1	X_2	X_3	$U_1=(X_1-m_1)/S_1$	$U_2=(X_2-m_2)/S_2$	$U_3=(X_3-m_3)/S_3$	X_1	X_2	X_3
12/13(月)	1	10	1	1	524	1343	332	−0.96	−0.86	−1.11	1	1	1
	1	11	2	2	900	1998	456	−0.15	−0.09	−0.64	2	2	2
	1	12	3	3	1728	2525	964	1.65	0.53	1.29	3	2	4
	1	13	4	4	1254	1971	676	0.62	−0.12	0.19	2	2	3
	1	14	5	5	1211	2491	819	0.53	0.49	0.74	2	2	4
	1	15	6	6	1287	2494	730	0.69	0.50	0.40	2	2	3
	1	16	7	7	938	2375	714	−0.06	0.36	0.34	2	2	3
	1	17	8	8	329	1369	379	−1.39	−0.83	−0.93	1	1	1
12/18(土)	2	10	1	9	566	1489	318	−0.87	−0.69	−1.16	1	1	1
	2	11	2	10	759	2624	514	−0.45	0.65	−0.42	2	2	2
	2	12	3	11	1114	1715	694	0.32	−0.42	0.26	2	2	3
	2	13	4	12	1421	3454	929	0.99	1.63	1.15	3	3	4
	2	14	5	13	1577	3110	962	1.33	1.22	1.28	3	3	4
	2	15	6	14	1500	3683	1049	1.16	1.90	1.61	3	3	4
	2	16	7	15	1146	3303	932	0.39	1.45	1.16	2	3	4
	2	17	8	16	336	837	292	−1.37	−1.45	−1.26	1	1	1
	2	18	9	17	173	804	260	−1.73	−1.49	−1.38	1	1	1
12/22(水)	3	10	1	18	609	1138	317	−0.78	−1.10	−1.17	1	1	1
	3	11	2	19	888	1508	446	−0.17	−0.66	−0.68	2	2	2
	3	12	3	20	1724	2370	919	1.65	0.35	1.11	3	2	3
	3	13	4	21	1207	2152	712	0.52	0.09	0.33	2	2	3
	3	14	5	22	1222	3103	806	0.55	1.21	0.69	2	3	4
	3	15	6	23	1291	2460	892	0.70	0.46	1.01	2	2	4
	3	16	7	24	914	2080	655	−0.12	0.01	0.11	2	2	3
	3	17	8	25	312	826	282	−1.43	−1.47	−1.30	1	1	1
	3	18	9	26	225	651	198	−1.61	−1.67	−1.62	1	1	1
			平均		$m_1=967.5$	$m_2=2072.0$	$m_3=624.9$						
			標準偏差		$s_1=468.96$	$s_2=866.4$	$S_3=269.1$						
			変動係数		$CV_1=0.485$	$CV_2=0.418$	$CV_3=0.431$						
											$(X_1\leqq 650)=1$ $(650<X_1<1300)=2$ $(X_1\geqq 1300)=3$	$(X_2\leqq 1500)=1$ $(1500<X_2<3000)=2$ $(X_2\geqq 3000)=3$	$(X_3\leqq 400)=1$ $(400<X_3\leqq 600)=2$ $(600<X_3<800)=3$ $(X_3\geqq 800)=4$

(図1—1)入店者数(X_1)と売上金額(X_2)の散布図(日別)



(図1—2)入店者数(X_1)と売上金額(X_2)の散布図(時間別)



考慮して X_1 と X_2 などの散布図を表示すると（図 1-1、1-2）がえられる。また、X_1 と X_3 についても同様の散布図を書くことができよう。これを観察すると、

(1)「X_1 が増えると X_2、X_3 がふえる」という常識的な傾向；(2) 日、時の組み分けの情報からa）日別にみると、第1、3日と第2日はやや傾向が異なる、b）時間別にみると各日とも類似している；などが指摘できるので、取りあえず回帰分析などによりこうした傾向を定量的に確かめるというのが定石であろう。上のように事前に与えられた日、時の情報を利用する考え方はいわゆる〝層別化〟に相当する。社会調査データなどを、個有、属性（性、年齢、学歴、……）により仕分けして観察するなどもそれである。しかし、こうした標識情報の利用が難しい、あるいは積極的に取り入れて分析するだけの根拠が稀薄である。さらには、それらを分析過程で一時伏せておいてしかるべき段階で改めて取り出して照合比較したい……など、分析の段階に応じてさまざまの局面がありうる。こうした場合に、ある程度客観的にものを分けて観察を容易にする方法が必要となる。

いま、特性の数を m とし、大きさが n の観測データの第 i 個体の観測データ・ベクトルを $\boldsymbol{x}_i=(x_{i1}, x_{i2}, \cdots\cdots x_{im})$ で表す。このとき分けるもの（対象）は個体、特性、あるいはその両者と目的に応じていろいろであるが、ここでは個体側の分類を中心に話を進める。個体間の距離あるいは類似度を考えるという点ではMDSと類似するが、手法との関連で若干の注意がいる。クラスター分析ではMDSのように単に対象間の関連の度合を表す数値であれば何でもよい場合から、必ずある種の距離でなければ手法の適用が意味を持たない場合まで幅が広い。しかも、データの種類が質的か量的かによっても軽重があることに、十分注意がいるのである。

いわゆる階層的手法は若干の例外はあるが、多くは個体 $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ 間の類似度 s_{ij}、または距離 d_{ij} が与えられたものとして計算に入る。たとえば、凝集型の階層的手法と名づけられた典型的な方法の算法はきわめて簡単である。

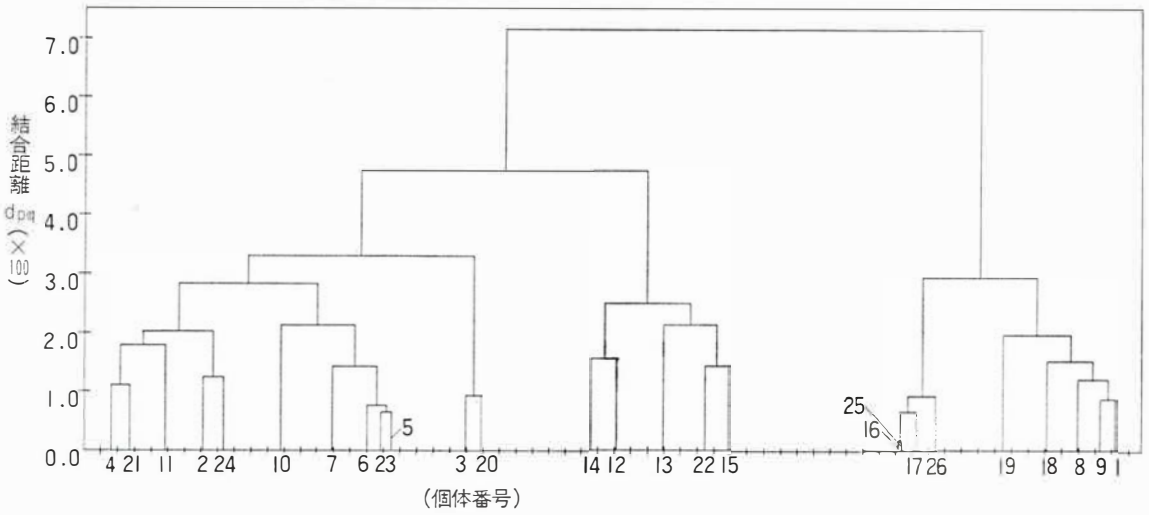
〔算法〕

S 1) n 個の全個体を n 個のクラスターとみなして、この個体に1から n までの番号を付与する。

S 2) 距離行列 $D=(d_{ij})$（または、類似度行列 $S=(s_{ij})$）を計算して、最も近い（あるいは似ている）クラスター（または個体）の組をさがす。いまこうして選ばれたクラスターのラベルが p と q ($p<q$) であったとする。そしてその距離（または類似度）を d_{pq}（または s_{pq}）とする。

S 3) クラスター C_p と C_q とを結合して、クラスターの数を1だけ減らす。このとき、クラスター C_p とそれ以外のすべてのクラスターとの間の距離（類似度）を更新する。そしてクラスター q に属する D

(図2)樹木図(デンドログラム)



(または S) の行と列を消去する。

S 4) S2)、S3) を ($n-1$) 回 (または適当に決めた回数) だけ繰り返す。各ステップで結合したクラスターの組み合わせ、それらの距離といった情報をリストにとる。

S 5) 必要があれば、結果をデンドログラム (dendrogram (樹木図)、phenogramなどとも呼ばれる) として図的表示する。

似ている、あるいは距離の近い個体同士を、逐次結合させながら寄せ集めるので凝集型という。手順S 3) でクラスター間の距離の更新を行うときの規則が要であって、それが個々の手法に相当する。数多くの手法があるが、その多くは数値分類法を源とすることも特徴である。この種の手法の主な性質として、

(1) 原理が簡単である。

(2) 広汎な種類のデータが扱える。

(3) (2)とクラスターの基準との関連が不明確となりやすく、結果の判断が難しいことがある。

(4) 計算機向きの手法が多い、つまりプログラム化が容易である。しかし半面、組み合わせ的要素が強いので、大量のデータを扱うことが難しい。

(5) 専門的知識に裏づけされた経験を結果に反映しやすい、などが挙げられる。

では実際にこの方法で計算した結果はどうなるか。答えの1つが (図2) 樹木図である。ここでは個体間の距離として次のユークリッド距離を使った。

$$d(x_i,\ x_j) \equiv d_{ij} = \{\sum_{k=1}^{3}(x_{ik}-x_{jk})\}^{\frac{1}{2}}$$

$(i,\ j=1,\ 2,\cdots\cdots,\ 26)$

たとえば、個体番号16と25の3次元データの間の離れ具合を距離、

$$d(x_{15},\ x_{25}) = \{(336-312)^2+(877-826)^2 + (292-282)^2\}^{\frac{1}{2}} = \sqrt{797} \fallingdotseq 28.231$$

で表す。この手法では図にみるように、特に指定のない限りは全個体が1個のクラスターになるまで結合を繰り返す。そして、何がクラスターかその判断は利用者に委ねられるのである。つまり3個のクラスターが欲しいならば3個と考えてみるのである。樹木図の結合距離をみて、これが大きく増大するところで切断してクラスターを決めるとよいとされているが、むしろクラスターの個数を何個にしたいという目安を分析の初期段階で決めて取り組むほうがよいように思う。それは手法によって樹木図の見映えが変わるのが常で、単に視覚だけに頼っては危険であることによる。もちろん多少は客観的に、何らかの指標を算出し、それを目安にクラスター数の見当をつける方法もいくつか考えられてはいる(次回に解説予定)。(図2)の樹木図からかりに4個のクラスターを作るとしよう。4つのクラスターを C_1, C_2, C_3, C_4 で表し、それぞれのクラスターの構成を個体番号によ

り示すと、

C_1＝{2, 4, 5, 6, 7, 10, 11, 21, 23, 24},

C_2＝{3, 20},

C_3＝{12, 13, 14, 15, 22},

C_4＝{1, 8, 9, 16, 17, 18, 19, 25, 26},

となる。個々のクラスター内の個体数をクラスター・サイズとい。うこの例ではクラスター・サイズはそれぞれ10, 2, 5, 9 個である。

クラスタリングの計算はここで終わるが、分析はさらに進めねばならない。よくクラスター分析をやってみたが結果がさっぱり解釈できない。あるいは理解に苦しむという意見がある。しかし、何事にも打ち出の小槌のような方法はありえないのである。データ解析は、たとえるならば、粘土細工のようなもので初めに何を作ろうかという大まかな狙いがあっても、途中の製作過程でヘラやコテを使ってあれこれ細工をすることによってはじめて構想に近いものが出来るのである。狙いは同じであっても途中の過程はそれぞれ異なるわけで、途中の味つけ、細工の段階のヘラやコテに当たるものが個々の手法であろう。しかもクラスター分析の性格として、計算処理の結果が一挙に結論に結びつく情報を提供するのではなく、むしろ分析の新たな局面に切り込む手がかりを得る手段としての働きが大きい。また、計算処理は自動的にできるがその前後の処理はきわめて泥臭い手作業となるのがつねである。このためには、利用可能な情報は総動員してかかるという姿勢が大切であろう。

例の場合、日、時の情報があるので、これにより（日または曜日）×（時間）の二元表を作り各、セル内にクラスター番号を書き込んでみる。そして適当に行と列を並べかえる。同時に4つのクラスターに対応させて個体を仕分けした散布図（図3）を作り（表2）と比較する。すると、

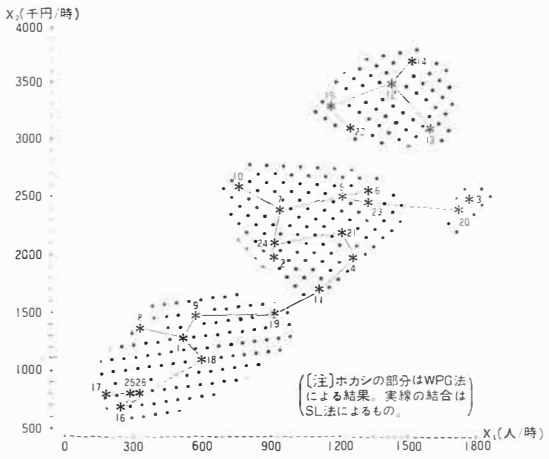
1) 平日と週末の違いがわかる。
2) 昼休み時は特殊な時間帯である。
3) 開店、閉店間際は平日、週末とも差はない
4) 平日の昼休み時は、週末の同時刻より入店者数、金銭出納機使用回数とも多いが売上金額ののびには結びつかない。1人当たりの売上金額は週末のほうが高い。

などの傾向がみえる。入店者数が多ければ売り上げも多いという情報だけではなく、曜日、時間の間にある種の関連がみられ、しかも客種が違うことまで予想させる。ところで、通常はこの例のように単純で当たり前とは限らずデータの量も多く構造も複雑化となる。このように、クラスター分析の結果を他の情報と対比させ、データを吟味する情報集約化の過程の考え方、理想的には自動化が、これからのクラスター分析の課題の1つである。

4. 手法によって解が異なること

1つの手法—WPG法（Weighted Paired Group）というが一を適用して上の例にみたクラスターを得たが、手法がかわると結果(解)はどう変化するのであろうか。WPG法に加えて、SL法（Single-linkage)、CL法（Complete-linkage)、

(図3)分類の結果



(表2)クラスターと日・時の関係

日・曜日＼時間	12	13	14	15	16	11	10	17	18	
12月13日(月)	2	4	4	4	4	4	1	1	*	平日
12月22日(水)	2	4	3	4	4	1	1	1	1	平日
12月18日(土)	4	3	3	3	3	4	1	1	1	週末
	昼休み時		(昼から午後)				(開店時 閉店時)			

<注>＊印はデータなし

GA法(Group-average)、Ward 法の4つの方法を取り上げる。たとえば、SL法の結果を示すと（表3）となる。個体番号16と25が最も近いので最初に結合する。その時の距離は28,231であり、25は16に吸収されたとみて、番号25を削除する（以後の反復の中に25の数字は現れない)。 以下同じように結合が進む。これを順につないで樹木図がえられる（読者は試すとよい)。 他の手法についても同じ計算を行い、クラスター数をかりに4個と定めて各手法で作られるクラスターの構成を表にまとめる。（表4）をみると、少しずつ各クラスターの構成が異なる。事前に与えられた所属の標識（名札）をもっていない、むしろ名札をつけることがクラスター分析であるが、この名札の番号のつけ方は1通りとは限らない。したがって手法ごとに解が異なるということは付与したクラスターの番号が必ずしも対応していないことにもなるので、比較するときに約束を決めておく必要がある。かりに、手法ごとのクラスターを順次重ねあわせて、共通の個体番号が多いものに、同じクラスター番号を付与するという方式で整合した結果が(表4)である。(したがって、必ず番号の整合が可能とは限らない。クラスターを同定化する方法は一般に面倒である)。 Ward 法とCL法、WPG法とGA法の結果が大体一致しているが、一般には必ずしも同じ結果にはならない。もちろん、4つのクラスターの個体構成が同じであっても結合の距離は異なる。

では、どれが正しいのか。それぞれの手法で約束したクラスターを作り出すように計算したのであるから、答えはいずれも正しいのである。つまり、利用者は計算の初めにどんなクラスターを望むか構造に対するある程度の仮説を用意しておくことが必要である。

たとえば、なるべく個体間の隣接状況があいまいな部分は積極的に結合し、とび離れた個体を検出したい、大勢は大きくまとめ仲間はずれを切り離してよく観察したい、という方針でクラスターを作りたいときにはSL法が向いている。また、個体をある程度同じクラスター・サイズになるように仕分けし塊状の団子型のクラスターを作るのであればCL法、Ward法などがよい。 さらに、何はともあれどう見てもよく似ているという群を

(表3)SL法の計算結果

反復	P	q	dpq
1	16	25	28.231
2	5	23	112.650
3	5	6	117.073
4	16	17	142.439
5	1	9	152.565
6	3	20	161.450
7	16	26	173.081
8	4	21	190.436
9	1	8	202.262
10	2	24	215.687
11	1	18	222.430
12	15	22	248.298
13	12	14	270.337
14	4	11	292.335
15	2	7	301.798
16	2	4	307.054
17	12	15	313.743
18	2	5	314.658
19	1	19	347.029
20	2	10	366.117
21	12	13	379.158
22	1	2	394.245
23	1	16	407.130
24	1	3	443.078
25	1	12	604.177

(表4)各クラスターの所属構成(メンバーシップ・リスト)

クラスター＼手法	SL法	CL法、Ward法	WPG法、GA法
C_1	1,9,8,18,19 2,24,7,4,21,11,5,23,6,10	2,24,4,21,11,19	5,23,6,7,10 2,24,4,21,11
C_2	3,20	3,20 5,23,6,7,10	3,20
C_3	12,14,15,22,13	12,14,15,22,13	12,14,15,22,13
C_4	16,25,17,26	1,9,8,18 16,25,17,26	1,9,8,18,19 16,25,17,26

さがしたいのであれば、複数個の手法、それも相当に性質の異なるもの、によって何通りかの解を作り、どの方法でもある群となる。つまり非常によく似ていることを確かめてクラスターとする。(表4)でいえば、クラスター C_3 などがそれにあたる。そして手法によってあちらのクラスターこちらのクラスターと移動するあいまいな個体はクラスター間の境目（間隙）が明確でないと考える。例でいえば{19}、{1, 9, 8, 18}、{5, 23, 6, 7, 10}などがそれである。それぞれある程度のまとまりをもった群を構成してはいるが、手法次第でどんなクラスターに成長するかふらつくのである。そして、たいていの手法はこうした境界部分の仕分けはにが手であることが多い。結局、利用者は試行錯誤の上に立ってデータのクセを読みとる手間をおろそかにはできず、まさに手探りのデータ解析となる。

5. データの種類との関連

計量データを扱う手法や類似度は比較的豊富である。しかし、カテゴリカル・データの扱いは一般にクラスター分析は不得手である。その理由の1つとして、分析の目的にあった適切な類似度、距離が作りにくい、あるいは種類はたくさんあってもその性質と手法との関連があまり明確でないことを指摘できる。形式的にはたくさんの類似度を使って、数多くの算法で計算が可能であっても結果の解釈で苦しむことになる。

（表1）のデータで、X_1, X_2, X_3 をそれぞれ（表1）—（C）のようにカテゴリー化して、これを例に考えよう。こうした処理は社会調査データなどで、年齢、所得、などの特性を扱う場合によく行われる。むしろ初めからカテゴリカルデータである場合のほうがはるかに多い。

距離の例として、平均偏差

$$d_{ij}=\frac{1}{m}\sum_{k=1}^{m}|x_{ik}-x_{jk}|$$ を考えよう。すると、

個体 $\boldsymbol{x}_1$ と $\boldsymbol{x}_3$ の距離は、

$$d_{13}=\frac{1}{3}\{|1-3|+|1-2|+|1-4|\}=2.0$$

個体 $\boldsymbol{x}_2$ と $\boldsymbol{x}_5$ の距離は、

$$d_{25}=\frac{1}{3}\{|2-2|+|2-2|+|2-4|\}\fallingdotseq 0.67$$

.....

.....

のように距離が計算できる。また次のように(0, 1)型データに変換するとたくさんの類似度、非類似度が利用できる。

個体番号	X_1	X_2	X_3	X_1 1	2	3	X_2 1	2	3	X_3 1	2	3	4
$\boldsymbol{x}_1$	1	1	1	1	0	0	1	0	0	1	0	0	0
$\boldsymbol{x}_2$	2	2	2	0	1	0	0	1	0	0	1	0	0
$\boldsymbol{x}_3$	3	2	4	0	0	1	0	1	0	0	0	0	1
$\boldsymbol{x}_4$	2	2	3	0	1	0	0	1	0	0	0	1	0
$\boldsymbol{x}_5$	2	2	4	0	1	0	0	1	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$\boldsymbol{x}_i$ で1、$\boldsymbol{x}_j$ 1となる組（これを1, 1とかく）の数を a、同様に（1, 0）の数を b、（0, 1）を c、（0, 0）を d とする。例のデータで $\boldsymbol{x}_1$、$\boldsymbol{x}_3$ をとると、$a=0$、$b=3$、$c=3$、$d=4$、$p=a+b+c+d=10$（カテゴリーの総数）となる。このとき、

（i） a、d に注目して、

$$\frac{a+d}{p}=\frac{0+4}{10}=0.4$$（一致率）

（ii） b、c の差異に注目して、

$$\frac{b+c}{p}=\frac{3+3}{10}=0.6$$(不一致率、1−[一致率])

（iii） a だけを強調して、

$$\frac{a}{p}=\frac{0}{10}=0.0$$(Russel と Raoの係数)

（iv） d は無視して、a を強調する、

$$\frac{a}{a+b+c}=\frac{0}{6}=0.0$$

（v） 条件つき確率の平均値、

$$\frac{1}{2}\left\{\frac{a}{a+b}+\frac{a}{a+c}\right\}=\left(\frac{-4}{7}+\frac{-4}{7}\right)\Big/2\fallingdotseq 0.5714$$

など、この手の指標は無数にあるが、どれを使うかは、利用者が特性のどの点を強調して採用するかによる。数値分類法では階層的手法、それもある範囲に限定された手法に、利用が集中している。それは生物分野で扱うデータにはカテゴリカル・データが多く、しかも個体数にくらべて特性の数が多いことや、距離そのものの大きさよりも

デンドログラムの連結情報、比較、距離の大小関係（順位）などに主な関心があることなどにその理由がある。したがってこの種の方法は、もともと情報量が少ない、量も少ないデータの中からぼんやりとではあるが何かをつかむといった場合に有効な、非常に柔軟な使い方に向いている。

次にデータがすでに加工ずみであるかどうかも重要なところである。特性値の測定単位が異なるので適当な方法で尺度化、無次元化する、特性数（次元）が多いので主成分分析、数量化分析などを事前に行い、えられた主成分得点などを使ってクラスタリングする……、などがそれである。測定単位の尺度変換を例にとると、変換の方法によっては、変換前の距離関係が変換後に保存されるとは限らない。最も単純に、平均0、分散が1となるような正規化を例にとる。（表1）のデータで X_2 を例にとると、平均は $m_2=2072$（千円）、標準偏差 $s_2=849.6$（千円）であるから、第1のデータは、$u_1=(1343-2072)/849.6\fallingdotseq-0.858$、第2のデータは、$u_2=(1998-2072)/849.6\fallingdotseq-0.087$、…… となる（距離の丸めなどによる計算誤差の介入も当然おこる）。この方法では、特性間の相関の情報まで考慮に入っていないから、変換前の距離の関係は変換後に保存されない。つまり、同じ手法で分類しても結果が違ってくる。もし、そういう距離が必要であれば、それに見合った変換をすればよい（たとえばマハラノビス距離など）。たとえば正規化したデータにＳＬ法を適用し4個のクラスターを作ると、

$C_1'=\{1,\ 8,\ 9,\ 18,\ 16,\ 17,\ 25,\ 26\}$、
$C_2'=\{10\}$、
$C_3'=\{3,\ 20\}$、
$C_4'=\{C_1'\sim C_3'$ 以外の個体$\}$

となる。（図3）との比較の結果は説明するまでもなかろう。

次に主成分分析、数量化分析などの解析手法の処理を経たデータを利用するとどうなるか。例のデータの3つの特性（X_1、X_2、X_3）を使って形式的に主成分分析を行い、えられた主成分得点にＷＰＧ法を適用した。計算するまでもなく、散布図の傾向から1つの主成分で大部分の説明がつくことは明らかである。事実、相関行列の第1固有値が $\lambda_1=2.725$ となるから寄与率は 90.83(%)（$\lambda_1/3\fallingdotseq0.9083$）に達する。つまり1成分で原データの90%以上の情報を抽出していることになる。前にならって、4つのクラスターを作ると、

$C_1'=\{4,\ 7,\ 21,\ 5,\ 6,\ 22,\ 23\}$、
$C_2'=\{11,\ 24,\ 10,\ 2,\ 19\}$、
$C_3'=\{12,\ 13,\ 14,\ 15,\ 3,\ 20\}$、
$C_4'=\{1,\ 18,\ 9,\ 8,\ 16,\ 25,\ 17,\ 26\}$

を得る。前の結果とくらべるとクラスターの構成が若干違う。たとえ寄与率で90%以上の情報をひろい出していても、模型の制約（線型モデルである）、失われた情報などの影響は無視できない。ふつうこうした処理を行うと、期待するほど高い寄与率や相関比を少数次元内で得ることは難しいから、加工データの情報量はかなり失われているとみるのが妥当である。したがって、この加工データを使ってクラスター分析を行うと、その制約された中での分類を行うのであって、結果は自ずとぼんやりしたものになる。原データの別の情報（属性など）を求めたクラスターとつきあわせてもはっきりした傾向といえるものが見つからなくても当然であるし、そういうものなのである。むしろこうした場面におけるクラスター分析の役割は、従来散布図を眺めて主観的に仕分けを行なっていた手間をある程度客観的に出来るよう自動化するという点にある。

むしろ場合によっては、分析によって変数間の関連をよく吟味して、少数個の特性を選択し、この特性のもとのデータを使ってクラスタリングするという方法が好ましいと思われる。

おもちゃのようなデータによる経験を通じて、主に凝集型の階層的分類手法の特徴を眺めてきた。いずれも注意深くみれば当たり前の現象である。しかし、データ量が増大しコンピューターで簡単に計算結果が手に入るようになると、おいそれと簡単に目を通すことが困難になり、しかもそれが情報のすべてであると錯覚し、当たり前のことも見落としがちである。また、計算の結果だけに目を奪われて分析の全体の流れを見失うことにもなる。要は、コンピューターに分担させるべき役割と、分析者が創意工夫すべき部分とをはっきり頭に描いて、問題に取り組むことである。

（つづく）

（第6研究部・第1研究室研究員）

昭和51年12月27日第三種郵便物認可　毎月15日発行　広告月報　2月号　昭和53年2月15日発行　通巻214号

広告月報

ASAHI AD. MONTHLY 1978 2 NO.214 朝日新聞社

特集

商品ヒットの切り口と攻め口

february

今月の主な内容

カット：麻生哲郎

昭和51年12月27日第三種郵便物認可　毎月15日発行　広告月報　2月号　昭和53年2月15日発行　通巻214号

朝日新聞社広告部

東京本社広告部
〒100　東京都千代田区有楽町2-6-1
03(212)0131

大阪本社広告部
〒530　大阪市北区中之島3-2-4
06(231)0131

西部本社広告部
〒802　北九州市小倉北区砂津1-12-1
093(531)1131

名古屋本社広告部
〒460　名古屋市中区栄1-3-3
052(231)8131

北海道支社広告課
〒060　札幌市中央区北二条西1-1
011(281)2131

定価　200円