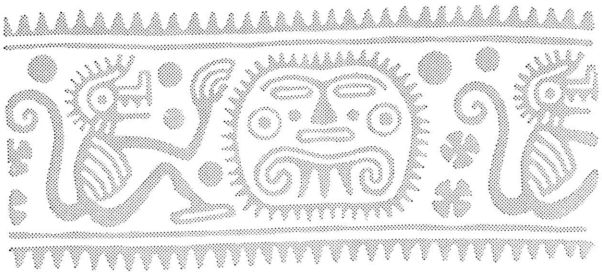
クラスター分析はどう使われるか

大隅　昇



1. まえがき

最近，“探査的”あるいは“手探り”のデータ解析という表現がよく目にとまる．また“ダータロジー(datalogy)”という新造語まで現われまさに多次元データ解析の分野は百花斉放の感がある．

なかでも“ものを分ける”という問題はこの多次元データ解析のあらゆる面で顔を出す．多次元データで扱うデータの形式はあるときは（個体）×（変数）の形のデータ行列であり，また解析対象（個体または変数）間の関連性を示す類似性の行列であることもある．またその種類も質的あるいはカテゴリカル・データ，量的データ，さらに両者の混在したものと様々である．さらに社会調査，マーケティングの分野のように電子計算機の発達に伴い取り扱うデータの量が非常に増えている面があるかと思うと，医学，生物学などのように多量データの収集は難しいが測定特性項目が多いというデータを扱う場合もある．このように扱うデータの性格が多様化していることに対応して分ける対象も複相化する．つまり従来の多変量解析法の重点が変数側にあったのに対し，データ解析の利用者は個体，変数の両者に注目したデータ要約化の方法論を要求するようになっている．そして個から集団への要約と集団から個への情報還元という繰り返しが要請されているのである．こうした場面で確かに分類のはたす役割は大きい．しかもなるべく主観をまじえずにより客観的に“もの”を分ける手続きが必要であろう．こうした分類手法の総称がいわゆる“クラスター分析”である．すでに本誌の Vol. 16-7, 1978 に階層的手法を中心にこの話題を紹介したが，ここで改めて総括することにする．

クラスター分析が多次元データ解析法の新しい研究課題として注目を集めるようになったのは主として1960年以後のことであるから比較的新しい方法論であるといえよう．Blashfield ら[46]の最近の報告によると 1960 年以後の急速な進展の理由として，

a）急激な研究発表の増加．これは計算機の利用可能性の進歩と関連していること，

b）クラスター分析に対して関心を示す研究分野が学際的かつ広範であること，

c）計算に必要なプログラム・ソフトウェアの増大，

などを指摘している．確かに文献数を追ってみても1970

表 1

年度	著者名　（　）の数字は参考文献番号
1963	Sokal and Sneath (33)
1965	Mcnaughton-Smith (28)
1969	A. J. Cole (ed.) (8)
1969	W. D. Fisher (15)
1970	Lerman (24)
1970	Tryon and Bailey (37)
1971	Jardine and Sibson (20)
1973	Anderberg (1)
1973	Benzécri (2)
1973	Bijenen (3)
1973	Sneath and Sokal (34)
1974	Duran and Odell (11)
1974	Everitt (14)
1974	Estabrook (13)
1974	Bock (4)
1975	Hartigan (17)
1975	Clifford and Stephenson (7)
1975	Späth (35)
1975	Vogel (38)
1977	Späth (36)
1977	Ryszin (ed.) (32)

年までに現われた数の数倍近くがそれ以後に集中して現われているという傾向がみられる．ちなみに著作物に限っていま筆者の手元にあるものを一覧にしてみるとこの傾向がよりはっきりする．(表1)がその一覧であるがこれをみると“数値分類法 (Numerical Taxonomy)”の糸口を作った Sokal, Sneath の1963年の著書が重要な役割をはたしていることがわかる．また年を追って次第にクラスター分析の包括する内容が多様化し分類の広い総称として使われるようになってきたことがわかる．ここに挙げた著作物は数値分類，クラスター分析を中心に紹介した本であって，レビューや本の一節をクラスター分析の紹介にあてたものまでを含めるとさらに多くのものがみられる．

クラスター分析では分類対象（ここでは個体と呼ぶことにする）に潜む類似性または差異性（非類似性）をある程度定量的に把えて各個体に標識を付与してクラスター化することであり，分析目的に応じてデータからクラスターを生成してみせる一つの道具と考えられる．したがって扱う問題の性質，データの性格にあわせてそれぞれ固有の使い方が考えられねばならないのであるが，現状の使い方をみると多くは既存の計算プログラムにデータを入力し結果を適当に解釈するという傾向がみられる．確かに計算手続きの上からは“自動分類”であっても扱うデータの種々の付帯条件との絡みでおこる様々の現象が分析の重大な極め手となる．クラスター分析にはしっかりしたモデルの考え方がない，あるいは理論的裏づけがないという理由からこれを批判する向きもあるが，より単純に考えて，データの要約表示の便利な道具とみるべきであろう．つまり分析者の問題に対する専門的見地にもとづく裏づけと，データにそれを求める手探りの部分との接点にあって情報をより見易い形に要約する方法なのである．

クラスター分析の手続きの上で，具体的に大切な点は次のようなことである．

（1） 観測特性の選択と数値化（コード化）

これはデータ解析全般にいえることであるが，分析目的に適した特性（変数）の選択は重要である．これは単に選んだ特性の適否の問題だけではなくそれが目的とする分析に役立つか否か，つまり測定項目の代表性の問題とも関連する．より一般的にいえば“データの質”に関わることでこれが分類結果を大きく左右することはいうまでもない．

（2） クラスターの定義，あるいは類似性，差異性を示すものさしの設定・具体的には分類対象の間に約束する類似度，非類似度，クラスターの等質性基準などを定めることである．これは分析者の立場からみてどのようなクラスター化が望ましいのかある程度の仮説設定の用意が必要であるということに他ならない．何もわからぬから分類するのではなく分類すべき必然性なり意図があるから分類するのであるからクラスターに対するある程度の約束ごとが必要となる．

（3） クラスターを生成するクラスター化の技法，つまり具体的な算法とその処理を可能にするプログラムの作成．このことだけを取り上げるならば“自動分類”そのものといってもよい．ここでは，計算効率や処理手続きの容易性などが重要な点である．

（4） 得られた分類結果の評価方式．

分類したままではどうにもならないわけで，その後どうするかが問題である．実はクラスター分析の多くは(3)の段階の手法，算法の提案に終始し，この分類後の検討方法については分析者の専門知識にもとづく適当な解釈におわってしまうことが多い．しかし，本来は分類結果を客観的に比較検討する何らかのものさしが必要であって，いたずらに分類してみただけでは解釈に苦しむことになる．社会調査データなどであれば各種のデモグラフィク要因を利用して分類結果を再編集してみる．より一般的には分類に利用した特性以外の副次特性要因を使ってプロファイル分析を行うなど分類後の処理が大切である．つまり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むすびつくのではなく次の分析への手掛りを得て別の局面へ切り込む手段としての働きが大きい．つまり，(3)の段階がハードな側面とすれば，(4)の手続きはソフトな側面でありその時点でのデータ処理はきわめて手作業的である．この意味でも問題ごとに分類手続きが異なるあるいは工夫が必要であるとみるのが至当で，この段階で汎用的ということはあまり期待できないのである．

さてここでクラスター分析の発展の経過をあらためて追ってみることにしよう．そこには大きな流れがいくつかあるが，その1つは前述の Sokal らによる数値分類法の考え方である．本特集の埴原，森島両氏の報告にみられるような系統樹（デンドログラム）による分類表示法がその代表的なものである．これは系統分類の自動化，計量化の思想のもとに発展した手法である．生物学，人類考古学，エコロジー，博物学，その他の多くの分野で盛んに利用されてきた階層的あるいは系統的分類法の大部分の手法の源流はここにある．

一方，それより早く1940年前後にすでに“クラスター分析”という用語は現われている．これは因子分析など

の分析法に対応させて，変数側情報の集約化の意味で用いられている．また1960年代半ばから，それまでの統計理論，とくに多変量解析法，分散分析法，標本抽出論（とくに層別化）などを裏づけとし，算法上の工夫を主とする自動分類法が多く現われる．この代表的な考え方はクラスターの等質性基準としてクラスター内平方和，非等質性基準としてクラスター間平方和（平方和のかわりに分散としても同じ）を利用するもので，こうした基準そのものよりもそれを最適化する算法上の工夫に焦点がある．この代表的な手法が MacQueen の k-means 法であり，Ball, Hall の ISODATA 法 (Iterative Self-Organizing Data Analysis Techniques—A) である．また数量化Ⅰ類型のデータを用いてカテゴリカル・データの説明変数群の情報を利用して量的データの外的基準を自動層別する AID (Automatic Interaction Detector) 法は分散分析のアイディアの利用である．また計量経済学の分野でクラスタリングを積極的に導入した W. Fisher は数理計画法（線形，二次計画法，動的計画法など）を使って時系列データのクラスター化，産業連関表の分類などの試みを行っている．このように流れを追ってみるとこうした現象がほぼ同時期にそれぞれ異なる分野に現われている点が興味深い．つまり分類という目的は同じでありながら，それぞれの分野での要請に応じた工夫がなされた上で各手法が生まれている．ところが最近はこれらがクラスター分析の名のもとに統合されたばかりか利用の形態があらゆる分野に波及したために，個々の手法が本来扱いうるデータの性格や，制約の範囲をこえた使い方がなされている．その上計算プログラムの氾濫がこれに拍車をかけて誤用，乱用が目立つようになっているのである．

2. 手法の概略

以上のようにクラスター分析の守備範囲は非常に広い．しかし表1にあげた文献の中でほぼ共通して使われている分類法の分類手続きに従って整理すると手法をいくつかに大別できる．その典型的な見方は“階層的手法”と“非階層的手法”に二分するものであるが，ここではさらに細分して次のように整理してみよう．

1) 階層的手法 (hierarchical techniques)
2) 分割最適化法 (partitioning-optimization techniques)
3) モード探索型手法，分布の混合 (mode-seeking techniques, mixture problems)
4) 重複を許すクラスタリング (overlapping clustering)
5) ファジイ・クラスタリング (fuzzy clustering)
6) 図的表示法 (graphical representation, exhibit)
7) その他

どのような分類に従っても個々の手法を1つのカテゴリーで代表することはむずかしく，いくつかにまたがることが多い．その意味では上の分け方はあくまで1つの目安である．たとえば，階層的手法でデータの一部を利用して初期分類を行ない，その結果を初期条件として残りのデータの分類を行うなどがそれであり，この種の手法の変形が数多くみられる．

2.1 階層的手法

おそらく最もよく知られまた広く利用されている手法群であるがそれだけに利用法や解釈に混乱があることは否めない．階層的手法はさらに次のように大別できる．

表 2

- 凝集型 (agglomerative type)
  - i）**グラフ理論的な考え方によるもの**（個体あるいはクラスター間の連結性に注目する．ノンメトリックなデータに適する）．
  - ii）**統計的な基準にもとづくもの**（クラスターのセントロイド間距離，平方和などの統計量を使う．メトリックなデータ向き）．
- 分枝型 (divise type)
  - i）**変数側の関連性を利用して個体側を分類する**（相関係数などによる分類，正準相関分析の階層的利用など）．
  - ii）**数量化Ⅰ類，重回帰型データの分類**
  - iii）**統計的基準にもとづくもの**

凝集型の手法は，個々の個体を近いもの類似したものを逐次寄せ集めて最終的に1個のクラスターとする，あるいは適当な段階で結合を止めてクラスターとする，という点で一致するが，クラスターおよびクラスター間（または個体間）距離をどう約束するかによって様々の変形が考えられる．ごく簡単にいえば，この種の手法の入力データは個体間の類似度あるいは非類似度行列 $D=(d_{ij})$ (d_{ij} は個体 i, j 間の非類似度，または類似度）であり，出力，つまりクラスタリングの解は系統樹（デンドログラム）あるいは，その上に現われた新たな個体間距離 $\Delta=(\delta_{ij})$ である．こうした手法の性質として，

（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られた経験を結果に反映しやすい，などが挙げられる．

これらの手法の源はやはり Sokal, Sneath らにあると思われるが，同時に全く同義かほとんど同義とみられる手法が他の分野にもみられることも特徴の1つである（これについては本誌 '78 年7月号を参照）．こうした中で Lance, Williams らが考えた，手法の統一的表現は算法上は非常に有益なものといえるが同時に弊害も生んでいる．彼等の提案した“組み合わせ的手法”とはクラスター間の距離関係をある回帰的表現で定式化するものでこれにより多くの手法が統一表現できるというものであるがあまりにデータ処理志向の考え方であるために誤用の危険を含んでいる．この考え方はきわめて単純で，ある結合の段階でクラスター C_p, C_q が結合して新たに C_t が生成されるとき C_t と C_p, C_q 以外のクラスター C_r との結合距離 d_{tr} を，

$$d_{tr} \triangleq \alpha_p d_{pr} + \alpha_q d_{qr} + \beta \cdot d_{pq} + \gamma |d_{pr} - d_{qr}| \quad (1)$$

で与えるものである．ここで d_{ij} $(i, j = p, q, r)$ はクラスター C_i と C_j との距離を示す．この方式によると α_p, α_q, β, γ の各パラメータと類似度あるいは非類似度行列があれば逐次クラスター化の過程が追跡できる．Lance らの考え方は計算プログラム作成上はきわめて都合がよい．しかし個々の手法で何をクラスター間距離としているかが問題である．たとえば single linkage 法では（1）で，$\alpha_p = \alpha_q = 1/2, \beta = 0, \gamma - 1/2$ と与えて $d_{tr} \triangleq \min\{d_{pr}, d_{qr}\}$ とおく．つまりクラスター間の個体間距離の最小値を C_t, C_r 間の距離と定義するものである．一方，complete linkage 法では $\alpha_p = \alpha_q = 1/2, \beta = 0, \gamma = 1/2$，すなわち $d_{tr} \triangleq \max\{d_{pr}, d_{qr}\}$ と与える．つまり single linkage 法とは相反する関係にある．またウォード法とは C_t, C_r のクラスター間距離を重みつきセントロイド間距離の平方つまり $d_{tr} \triangleq [n_t n_r/(n_t + n_r)] \| \boldsymbol{m}_t - \boldsymbol{m}_r \|^2$ で定義すると（ここで $\boldsymbol{m}_t, \boldsymbol{m}_r$ は C_t, C_r の平均ベクトル，$n_t = n_p + n_q, n_p, n_q$ は C_p, C_q, n_r は C_r それぞれのクラスター・サイズ，つまりクラスター内の個体数），これがやはり（1）式の形式で表わせる．このときの d_{ij} はもちろん，$d_{ij} = [n_i, n_j/(n_i + n_j)] \| \boldsymbol{m}_i - \boldsymbol{m}_j \|^2$ $(i, j = p, q, r)$ となる（これらの詳細は（1），(14)などを参照）．この3つは非常に極端な例であるがこれらがいずれも（1）式で示されることが問題なのである．同じ距離行列を入力データとして与えてパラメータを変えると何通りもの結果が得られるが実はそこで約束している距離がそれぞれ異なるからクラスター生成の過程も自ずと異るし，しかも入力時の行列の制約条件がそれぞれ違っているのである．つまり得られた解（系統樹）の上にみられる個体間の関連はあくまである約束のもとに生成された近似表現である（これはもともとの個体間の関係に階層的連結性があるとは考えにくいことからも明らかなことである）．こう考えるともとの個体間の関係（つまり類似度，非類似度行列）の情報をなるべく損わないようにその関連を表示する，つまりよく適合する系統樹が描けることが，望ましいあるいはうまい算法であるという見方が出てくる．この見地から手法の性質を調べているのが Jardine, Sibson あるいは Lerman, Roux といった欧州の研究者達である．生物学の分野でも適合のよい系統樹を探すという方法として，Farris, Rohlf, Camin といった人達が精力的に展開してきた考え方も類似の考えに基づくものとみられる（本特集，森島氏の報告にもこれらの一部が紹介されている）．要するに系統樹の上にみる距離は一種の“擬似距離”でありそこに現われる情報は数値が強い意味を持つというよりは，連結のパターン，系統樹間の比較，連結の順位性などの比較的やわらかいものである．つまり系統樹とは個体間の関連性をもとにその集団構造を近似的に階層表現してそれらの間の意味を理解する道具である．また，階層的手法で利用される類似度，非類似度には無数のものがあるがこれがすべてメトリックな性質をもつとは限らない．むしろノンメトリックであることが多い．こうした場合でも多くの階層的手法は適用可能であるがそれだけに得られる結果は大まかな情報であり，手法自体にも限界があることはいうまでもない．

ところで凝集型とは逆に，与えられた全データを逐次分割して系統樹を作るという方式も考えられる．これが分枝型といわれるもので，系統樹を凝集型とは逆の向きに生成する方法と思えばよい．後述の分割型と異なる点はクラスターが互いに排反的に分割され階層構造を作ることである．代表的な手法として知られているものに Association Analysis, AID などがある．前者は Lance, Williams, McNaughton-Smith らの考案によるもので，0-1 型データの自動分類法の1つである．AID は Sonquist, Morgan らの開発した手法で数量化Ⅰ類で扱うような形式のデータを，項目のカテゴリーの情報を利用して外的基準にあたる量的なデータを分割し樹木図を作り，同時に項目の間に潜む交絡情報を検出しようというもの（それゆえ，Interaction Detector という）である．国内ではこれらの手法をマーケティングの分野でいち早く取り入れてライフスタイル分析などに利用したが，そこでの用い方はこれらの手法の発案者の本来の意図した

狙いとはやや異なるためにこれらの手法が誤解されている面もある．とくに Association Analysis は二値化データについて変数間の相関を利用して個体間の分類をはかるもので，一種のソーターとみることができる．数量化Ⅲ類や，ガットマン尺度解析などの事前処理法として利用して，パターンの分布傾向を捕えるときなど大変便利な方法といえよう．

2.2 非階層的手法とくに分割最適化型手法

少ない個体数に対して変数（特性）の数が多い，あるいは多値データであってそれからえた類似度行列が与えられるだけである，といった場合には階層的手法による大まかな情報要約化が有効であろう．しかし計量的データで量が比較的多いとき，特定の加工を経たデータ（主成分得点，因子得点など）を分類するときなどは分割型の手法が利用できる．クラスター自体のとらえ方に変化があった階層的手法に対し，分割型の手法の多くは従来の多変量解析の応用であり，重点はむしろ算法の工夫にある．階層的手法では特殊な場合を除いては，同一算法を適用すれば同一データを使う限り解は同じである．これに対し，分割型手法ではそうならない上に，階層的手法とは別の制約がいろいろと起る．分割型の算法の基本は非常に簡単で次のように要約できる．

（1） 全データを適当な個数の群に分割する（初期分割）．あるいはクラスターの目安とする代表点（核という）をデータ空間内に与える（初期代表点の指定）．

（2） 群ごとの平均ベクトルを算出し個々のデータをもっとも近い平均ベクトルの群にわりあてる．あるいは，もっとも近い代表点に各個体を配置する（所属の決定，配置）．

（3） 各群の平均ベクトル，その他の統計量の更新．

（4） クラスターのまとまりをはかる適当な基準を用意し（最適化基準の設定），これを満たすまで計算を繰り返す（反復計算）．最後に基準の改良がみられなくなったら，えられた群をクラスターとして採用する．

この算法にどんな工夫をとり入れるかで，さまざまの手法が考えられる．たとえば最も代表的な手法として，k-means 法があるがこれはクラスター内平方和を上の算法で最小化する典型的な方法である．こうした手法の共通の問題として次の点があげられる．

（1） 初期分割，初期代表点の選び方

（2） 各個体をクラスターに配置，再配置する際の方法と平均ベクトル更新の時期

（3） クラスター・サイズが不均衡であるときの手当ての方法

（4） 異常値に対する手当ての有無

（5） クラスター数の決定法（固定か，可変か）

（6） 最適化の基準とそれを達成する算法

いずれも問題であるが(1)，(5)，(6)などがとくに重要な鍵を握っている．たとえば，初期分割の与え方である．かりにクラスター数を k と定めたとしても，n 個の個体を k 個の空でない互いに重なりのない集団に分ける仕方，それを $S(n, k)$ とかくと，その総数は大変な数になる．$k=2$ として2分割を考えても $2^{n-1}-1$ 通りとなる．また，$S(10, 5)=179487$，$S(20, 2)=524287$，$S(100, 3)\approx 8590\times 10^{46}$，……というようにとても実用の計算可能な範囲内にはない．そこでふつうかりに k 個のクラスターがあるとして計算を始め，反復計算によって，しかるべきクラスター基準の最適化をはかるのである．この最適化基準として最も利用されるものは次のよく知られた関係から導びかれるものである．

$$T=W+B \qquad (2)$$

ここで T は全データから得られる総平方和・積和行列であり，W はクラスター内平方和・積和行列の総和，すなわち $W=\sum_{i=1}^{k} W_i$，そして k はクラスター数，W_i はクラスターに対する平方和・積和行列である．また B はクラスター間平方和・積和行列である．ここで T はデータが与えられると確定するので結局，B と W の関係をどう定めるかにかかってくる．普通(2)式のトレースまたは行列式を作ってそれらの組み合わせにより基準を設定することが多い．たとえばトレース基準（平方和基準）としては $\mathrm{tr}W$，$\mathrm{tr}B$，$\mathrm{tr}W^{-1}B$ など，行列式基準として $|W|$，$|W^{-1}B|$ などがある．後者は変数間の相関まで考慮した基準であり前者はそうした考慮のない変数ごとのバラツキだけを考えた基準である．これらの基準をみると従来の多変量解析の考え方と何ら異るところはないようにみえるが，決定的な相違点は，分類対象となる個体に群の所属を示す標識がないことである．つまり最適化の基準は同じであってもクラスター分析では算法の工夫が決め手となる．たとえばクラスター内平方和 $\mathrm{tr}W$ の最小化をはかる方式一つを考えても山登り法により最適化するのか，平方和をクラスター内平方ユークリッド距離の総和とおきかえてこの基準を二次計画法により最適化するのか，その工夫の仕方は様々である．しかもこの種の手法の解は一般に局所最適解となるので1回の計算結果だけから分類結果の解釈が難しいという問題が出てくる．また，上の基準からみてクラスターの形状にある種の制約がある．つまり平方和などのバラツキをはかる量でクラスター内のまとまりの度合をみるのであるからチラバリの形状や向きに大きく左右されること

は明らかである．まして，こうした基準のどれがよいかということはデータの構造によって定まる問題であるから，データに潜む傾向を知ろうとするクラスター分析の狙いからすると，1つの矛盾となり堂々めぐりとなってしまう．こうしたことから，分割型の手法を利用するときは，いきおい試行錯誤的・実験的要素が強くかかわりをもつ．たとえば，1回の計算で結論をつけるのではなく，何回も初期条件をかえては繰り返したり，データ量が多いときには，これをいくつかにランダム分割し，各々のわけたデータで分類を行い，それぞれの結果を比較しあうなどの手続きがいる．あるいは分けた第1の群で判別関数を作り，他の群をそれによって判別してみる，といった方法なども考えられるが，このあたりの工夫の仕方が，分割型手法を使いこなすコツといえよう．

つまり“分類結果の分類”を行う種々の方式が必要となってくる．よく因子得点や主成分得点を利用した分類をこの分割型手法で行うことが多いがこうしたデータはすでに情報が少数次元内に縮約化されているだけでなく模型上の制約（線形モデルである）などもからんで，得点の分布は必らずしもはっきりしたクラスターの存在を示さない（もちろん上の各種の基準の意味でのクラスター）．こうした場合には，1回の分類結果だけからクラスターを確定することはますます危険で何回もの試行が必要なばかりか，その結果も一般にはっきりした傾向を示さぬものである．むしろこうした場面では従来散布図を眺めて主観的に行なっていた仕分けの作業をある程度客観的に出来るよう自動化したというところにクラスター分析の利点がある．

3. その他の方法

階層的，分割最適化型だけに限らずその他の手法も数多くみられる．たとえば“モード探索型手法”としてD. Wishart のモード法や Gitman らのファジイ集合を利用したモード検出方法などがある．

この両者は非常に似ており，その算法の骨子は次のようなものである．まず与えられたデータの各個体の近傍をとりその領域内に含まれる他の個体の数を数えてその個体の密度とする．これをすべての個体について求めその分布状況をみて密度の濃い部分と薄い部分とに仕分けする．そして濃い部分を仮のクラスターとして残し，他の部分つまり密度の薄い部分にある個体を最近隣方式で再配置してクラスターの所属構成を決める，というものである．

また統計理論を積極的に取り入れてパラメトリック・モデルの立場からクラスターを捕えようとする方法もある．Wolfe の提案した確率分布（多変量正規分布）の混合問題がそれであるが，ここでは実験的に導出したある統計量にもとづいて検定を行ない，混合されているクラスター数を見積るというものである．しかし推定すべきパラメータ（平均，分散・共分散行列，クラスター数など）の求め方に難点があって必ずしも完全なものとはいいがたい（これに関連した簡単な実験は（63）を参照.）

また個体が複数個のクラスターに所属することを許す，いわゆる“重複のあるクラスタリング法”についてはグラフ理論的なアプローチが多くみられる．たとえば Jardine, Sibson らの B_k 法などがそれである（20）．こうした手法の多くは組み合わせ的要素が強いので大量のデータを処理することが困難であるという欠点をもつ．

さらに“表示（exhibit）”ということを重視して視覚により分類判定を行うという方法が近頃流行である．たとえば大変はやったものに Chernoff の FACE 法がある．これは多次元データを人の顔の各部分にふりあてて適当に尺度化した上で表示するというものである．これに類似のものとして Frith の方法がある（27）．

この種の手法は顔の各部への特性のわりあて方によって表示の結果（顔の表情）がかなり違ったものとなるのでそれを視覚判定に頼って区分するという使い方はアイディアとしては面白いが結果がきわめて恣意的になるという難点がある．Chernoff の利用法の1つが化石の時系列推移を追跡するという使い方であったことをみても使い方の工夫が難かしい．この他 Andrews のフーリエ変換による多次元データの表示法，多次元尺度解析法の一部の手法，などに図的表示の方法とみられるものがある（これらについては脇本他（84）に詳しい）．

また系統樹も一種の図的表示であることはいうまでもないが，これもその見栄えだけから分類結果を判定するととんでもないことになる．すでに述べたように，これはある基準のもとに生成されたあるクラスター化の過程の一つの表現にすぎないのであるから，その系統樹の評価や比較は別の観点から行なわねばならない．これがいわゆるクラスター分析の評価問題であるが，この種の議論はいまだ少なく今後の研究課題として重要なものの1つといえる．

また筆者が最近注目している分類法にファジイ・クラスタリングがある．クラスターは本来明確に定義できるものではない，ぼんやりしたあいまいなもの（fuzzy）であるから，そのやわらかな情報を生かしたままに扱うことはできないかという観点に立って問題に接近しようという試みがそれである．Ruspini は“ファジィ理論（fuzzy theory）”を利用して，個々のデータがクラスタ

ーに所属する度合を推定する最適化基準と，それを達成する算法をいへつか提案し，この種の考え方の糸口を作った．また Bezdek, Dunn らは k-means 法にファジィ理論を取り入れた方式 fuzzy ISODATA を考え出した．この他グラフ理論の拡張として位置づけられる“ファジィ・グラフ”あるいは“ファジィ代数”などを応用した研究も目立つようになってきた．これらについては本誌次号（特集 Fuzzy 理論）に掲載の機会を得たのでそこで詳しく述べるつもりである．

4. むすび

主に，実用に結びつく手法を概観し，種々の問題点を指摘してきたがその議論は否定的なものばかりにみえる．実際クラスター分析は実用に耐える方法がいまだ少なくデータ解析の道具として未解決の問題が山積している．

たとえば利用者の最も単純な要求である“クラスター数をいくつとするのか”といった問題1つを考えてもその答えはかなり難かしい．もちろんこうしたことは与えられたデータに大きく依存することであるが，それでも何かの手掛りが欲しいというのが実用の立場からの希望であろう．たとえば最近，リモートセンシングデータの分類などで盛んに利用されている自動分類法にISODATAがある．これが誕生してからすでに10数年も経ているが，それでもなおこれが利用される要素の1つとしてこの手法が一種の学習過程を取り入れた自動分類法であることを指摘できる．原理はきわめて単純で，基本的には前述の k-means 法である．しかし，ISODATA の特徴は，これを基本として，次の諸機能を取り入れているところにある．

1) 異常値とみられる個体あるいはクラスター・サイズの小さいクラスターの一時除去機能．
2) クラスター単位にチラバリの大きさと向きを吟味した上で分割をはかる，いわゆる局所分割方式，(locally splitting).
3) クラスター間距離表を作りこの中の近いクラスターは同時連結する (lumping).
4) クラスターの個数をある程度可変にできる．それも自動的に行なうことができる．

つまりクラスター数をかえたうえで，併合，分割を繰り返しながら trW の最小化をはかり分割のすべての組み合わせを総当りせずになるべくいろいろな分割状態を作り出して最適化しようというのである．したがって単純な k-means 法よりは改良されているし，クラスター数が可変であるというところに面白さがあるが，やはりクラスターの形状に対する制約のあることにはかわりがないから計量データ向きであり，パターン認識とくに画像解析などで多く利用されることもうなずける（たとえば (80)）．

また階層的手法でえられた系統樹の上にみられる分割集合の評価という問題もクラスター数の見積りと関わることである．同一データに複数の手法を適用すると一般に解の系統樹は同じパターンを示すとは限らない．むしろクラスター化の過程が手法によって異るから，つまり分割のすべての可能な組み合わせの総当りを行なって求めた系統樹ではないからこれは当然のことである．しかし，個体内の関連性の強さがしっかりしたものであれば系統樹の上の連結順位も保存されるであろうし，個体関連性がもともと漠然とした類似関係にあるならばクラスターの定義をかえることで結果はふらつき，系統樹はお互いに似ても似つかぬ形となるであろう．つまり同一データから得た複数個の系統樹の比較評価を行なう方式が必要であるが，これはクラスターのパターンの同定化の問題（どの個体がどのクラスターに所属するのか）と関連して重要な課題となりつつある．

このようにクラスター分析の現状も算法中心の考え方から次第に評価方式，手法の適用可能性の判定の問題，安定性・頑健性の検討といった新たな段階に移りつつある．

ところでこれと並行して，各手法をデータ解析の道具としていかに使い，勝手をよくするかという問題が起こる．つまりプログラム・パッケージの必要性がここに生じてくる．近頃の汎用統計プログラム・パッケージには必らずクラスター分析の手法が取り入れられるようになってきている（これについては (81)，(83) 参照）．さらにクラスター分析の計算処理を主とするパッケージもいくつか現われている．最もよく知られたものが Wishart の CLUSTAN-1C である．1969 年に CLUSTAN-I として公表され，以後 1A，1B と改良を続け 1978 年の CLUSTAN-1C の改良版に至るまでに約 10 年近くの時をかけている．最近国内でも日本科学技術研修所がこれを使った計算サービスを始めている．また Sokal の流れをくむ Rohlf らの開発した NTSYS はいくつかの多変量解析手法までを含んだクラスター分析用のパッケージである．

こうしたパッケージは独自の“コマンド(命令語)”を持つシステム化されたものであって単なるる個別プログラムの集合ではない．残念ながら国内では，個別プログラムやサブルーチン集合はあっても，こうしたパッケージ・スタイルのものはまだないようである．試作の段階であるが 10 種程度の手法を連結利用できるパッケージ

を筆者のところで現在作成中であるが，すでにこのために2年近くの時間を費している．このように時と労力が大変に必要であることと，たえず改編を続け利用者の要請を受け入れていかねばならないことがシステム化されたパッケージが誕生しにくい理由の1つであろう．

以上でクラスター分析の最も基本的な部分をごく大まかに総ざらいしたのであるが，こうした方法以外にさらに興味あるものも数多くみられる．しかし実用の見地から眺めるとやはり上に紹介した各手法が挙げられよう．これらはいずれも誕生してから10数年を経たものが多い．つまりそれ位の時間経過をへて淘汰されて残った手法であるともいえる．（おおすみ・のぼる，文部省統計数理研究所）

参考文献

（I）**Books and Monographs**

(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) Bock, H. H.; *Automatishe Klassfikation*, Vandenhoeck und Ruppecht. 1974.
(5) Blackith, R. E. and Reyment, R. A.; *Multivariate Morphometrics*. Academic Press. 1971.
(6) Cailliez, F., Pages, J. P.; *Introduction à l'Analyee des Données*, Smash 1976.
(7) Clifford, H. T. and Stephenson, W.; *An Introduction to Numerical Classification*, Academic Press. 1975.
(8) Cole, A. J. (ed.); *Numerical Taxonomy*, Academic Press. 1969.
(9) Doran, J. E. and Hodson, F. R.; *Mathematics and Computers in Archaeology*, Edinburgh Univ. Press. 1975.
(10) Duda, R. O. and Hart, P. E.; *Pattern Classification and Scene Analysis*, John Wiley. 1973.
(11) Duran, B. S. and Odell, P. L.; *Cluster Analysis: A Survey*, Springer Verlag. 1974.
(12) Enslein, K. Ralston, A. and Wilf, H. S. (eds.); *Statistical Methods for Digital Computers*, Vol. III of *Mathematical Methods for Digital Computer*, John Wiley. 1977.
(13) Estabrook, G. F. (ed.); *Proceedings of the Eighth International Coference on Numerical Taxonomy*, Freeman. 1974.
(14) Everitt, B.; *Cluster Analysis*, John Wiley. 1974.
(15) Fisher, W. D.; *Clustering and Aggregation in Economics*, The Johns Hopkins Press. 1969.
(16) Gnanadesikan, R.; *Methods for Statistical Data Analysis of Multivariate Observations*, John wiley, 1977.
(17) Hartigan, J. A.; *Clustering Algorithm*, John Wiley. 1975.
(18) Heise, D. R. (ed.) *Sociological Methodology*, 1975, Jossey-Bass. 1977.
(19) Hodson, F. R., Kendall, D. G., and Tăutu, P. (eds.); *Mathematics in the Archaeological and Historical Sciences*, Edinburgh Univ. Press. 1970.
(20) Jardine, J. and Sibson, R.; *Mathematical Taxonomy*, John Wiley. 1971.
(21) Kaufmann, A.; *Introduction a la Theorie des sous-ensemble flous*, Masson. Tome I (1973), Tome II, III (1975), Tome IV (1977).
(22) Kendall, M.; *Multivariate Analysis*, charles Griffin, 1975.
(23) K. S. Fu(ed.); *Digital Pattern Recognition*, Springer-Verlag. 1976.
(24) Lerman, I. C.; *Les Bases de la Classification Automatique*, Gauthier-Villas. 1970.
(25) Lockhart, W. R. and Liston, J. (eds.); *Methods for Numerical Taxonomy*, American Society for Microbiology. 1970.
(26) Mather, P. M.; *Computational Methods of Multivariate Analysis in Physical Geography*. John Wiley. 1976.
(27) Maxwell, A. E.; *Multivariate Analysis in Behavioural Research*, Chapman and Hall, 1977.
(28) McNaughton-Smith, P.; *Some Statistical and Other Numerical Techniques for Classifying Individuals*, H. M. S. O. 1965.
(29) Noris, J. R. and Ribbons, D. W.; *Methods in Microbiology*, Academic Press. 1972.
(30) O'Muircheartaigh, C. A. ard Payne. C. (eds.); *The Analysis of Survey Data* vol. I, II, John. Wiley. 1977.
(31) Orlóci, L.; *Multivariate Analysis in Vegetation Research*, Dr. W. Junk B. V.. 1975.
(32) Ryszin, J. (ed.); *Classification and Clustering*, Academic Press. 1977.
(33) Sokal R. R. and Sneath, P. H. A.; *Principles of Numerical Taxonomy*. Freeman. 1963.
(34) Sneath, P. H. A. and Sokal, R. R. *Numerical Taxonomy*, Freeman. 1973.
(35) Späth, H.; *Cluster-Analyse-Algorithmen*, R. Oldenbourg Verlag. 1975.
(36) Späth, H.; *Fallstudien Cluster-Analyse*, R. Oldenbourg Verlag. 1977.
(37) Tryon, R. C. and Bailey, D. E.; *Cluster Analysis*, McGraw Hill. 1970.
(38) Vogel, F.; *Probleme und Verfahren des Numerishen Klassifikation*, Vaudenhoeck und Ruprecht. 1975.
(39) Wang, P. C. C. (ed.); *Graphical Representation of Multivariate Data*, Academic Press. 1978.
(40) Young, T. Y. and Calvert, T. W.; *Classification, Estimation and Pattern Recognition*, Elsevier. 1974.
(41) *Mathematics and Computer Science in Biology and Medicine* 1965, Proceedings of Conference held by MRC, 1964, H. M. S. O.

（II）**Review Articles**

(42) Bailey, K. D.; Cluster Analysis (→(18)). 1974.
(43) Ball, G. H.; Data Analysis in the Social Sciences; What about the details?, *Proc. of Fall Joint Computer Conferences*, 533-559. 1965.
(44) Blashfield, R. K.; A Consumer Report on the Versatility and User Manuals of Cluster Analysis Software, *Proc. of the Statistical Computing Section. ASA*, 31-37. 1976.
(45) Blashfield, R. K.; Questionnaire on Cluster Analysis Software, *Classification Society Bulletin*, 3. 4 25-42. 1976.
(46) Blashfield, R. K. and Aldenderfer, M. S.; A Consumer Report on Cluster Analysis Software (1)-(4). 1976.
(47) Bolshev, L. N.; Cluster Analysis, *Bull. I. S. I.*, 43

Book 1, 411-425. 1969.
(48) Cormack, R. M.; A Review of Classification, *J. R. Statist. Soc.*, ser. A, **134**, 321-367. 1971.
(49) Good, I. J.; Categorization of Classification (→(41)) 1965.
(50) Gower, J. C.; Classification Problems, *Bull. of I. S. I.*. 1973.
(51) Ling, R.; Cluster Analysis, Yale Univ. Tech. Rep. No. 18. 1971.
(52) Sokal, R. R.; Clustering and Classification: Background and Current Directions (→(32)). 1977.

(Ⅲ) **References**(本文に関連したごく一部のもののみ)

(53) Ball, G. H. and Hall, D. J.; ISODATA, A Novel Method of Data Analysis and Pattern Classification, Stanford Research Institute, Technical Report. 1965.
(54) Bezdek, J. C.; Mathematical Models for Systematics and Taxonomy (→(13)). 1975.
(55) Dunn, J.; A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well-separated Clusters, *J. of Cybernetics*, 3, 38-57. 1974.
(56) Fisher, W. D.; On Grouping for Maximum Homogeneity, *J. Am. Statist. Ass.*, 53, 789-798. 1958.
(57) Friedman, H. P. and Rubin, J.; On Some Invariant Criteria for Grouping Data, *J. Am. Statist. Ass.*, 62, 1159-1178. 1967.
(58) Gitman, I. and Levine, M. D.; An Algorithm for Detecting Unimodal Fazzy Sets and Its Application as a Clusteriug Technique, *IEEE Trans. on Comp.* C-19, 7, 1970.
(59) Hubert, L. J. and Baker, F. B.; An Empirical Comparison of Baseline Models for Goodness of Fit in r-diameter Hierarchical Clustering (→(32)). 1977.
(60) Kass, G. V.; Significance Testing in Automatic Interaction Detection *Appl. Statist.*, 24, 178-189. 1975.
(61) Ling, R.; A Probability Theory of Cluster Analysis, *J. Am. Statist. Ass.*. 68, 159-169. 1973.
(62) MacQueen, J.; Some Methods for Classification and Analysis of Multivariate Observations, *Proc. 5th Berkely Symp.*, 1, 281-297. 1967.
(63) Matusita, K. and Ohsumi, N.; Evaluation Procedure of Clustering Techniques, *France-Japan Seminar*, Paris, March. 13-20, 1978.
(64) Peay, E. R.; Nonmetric Grouping; Clusters and Cliques, *Psychometrika*, 40, 3, 297-313.
(65) Rohlf, F. J., Kishpaugh, J. and Kirk, D.; NTSYS-User Manual. 1977.
(66) Ruspini, E. H.; A New Approach to Clustering, *Inf. and. Cont.*, 15, 22-32. 1967.
(67) Ruspini, E. H.: Numerical Methods for Fuzzy Clustering, *Information Sciences*. 2, 319-350. 1970.
(68) Sclove, S. L.; Population Mixture Models and Clustering Algorithms, *Comm. Statist.-Theor. Matr.*, A 6 (5). 1977.
(69) Scott, A. J. and Knott, M.; An Approximate Test for Use with AID, *Appl. Statist.*, 25, 103-106. 1970.
(70) Sonquist, J. A. and Morgan, J. N.; The Detection of Interaction Effects, ISR Univ. of Michigan, Monograph No. 35. 1964.
(71) Williams, W. T. and Lance, G. N.; Hierarchical Classificatory Methods (→(12)). 1977.
(72) Wishart, D.; FORTRAN Ⅱ Programs for 8 Methods of Cluster Analysis (CLUSTAN I), Kansas Univ. Computer Contribution 38. 1969.
(73) Wishart, D.; CLUSTAN-User Manual (Third ed.). 1978.
(74) Wolfe, J. H.; NORMIX 360 Computer Program, Research Memo., SRM 72-4. Wolfe, J. H.; NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. Research Memo., SRM 68-2. 1967.
(75) Zadeh, L. A.: Similarity relations and fuzzy ordering, *Inf. Sciences*, **3**, 177-200, 1971.
(76) 日本科学技術研修所; CLUSTAN-1 C プログラム利用の手引き, 1978.
(77) 大隅昇; クラスター分析, 現代数学, **10**, 9, 1977.
(78) 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.
(79) 大隅昇; クラスター分析, 最新醫学, **33**, 1, 1978.
(80) 大隅昇, 渋谷政昭; 数値的地域区分法: NTAP, 統計数理研究所彙報 Vol. 25, No. 1.
(81) 水野欽司, 大隅昇, 桂康一; 統計プログラムパッケージ, bit, 10, No 8～14.
(82) 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.
(83) 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.
(84) 脇本和昌, 後藤昌司, 田栗正章, 松原義弘; 多次元データのグラフ解析法, 応用統計学, **6**, 2-3, 1977.