

最新醫學・第33卷・第1號（昭和53年1月）別刷

計量診断・F.多次元分類の理論・クラスター分析 1

解 説

大 隅 昇

最 新 醫 學 社

計量診断・F.多次元分類の理論・クラスター分析 1

解 説

大 隅 昇*
お お すみ のぼる

1. 現状と問題点

“ものを分ける”ことは多くの分野の現象解析のもっとも基本的な手続きの1つである。多くの場合は、事前に与えた情報(標識)を利用してものを分けるのであるが、標識そのものを作り、これにより群を作り出すという観点に立った分類のしくみも考えられよう。これがクラスター分析であり、ものを分け、群(集落)を作り出す手続きの総称であるといつてよい。“クラスター分析”の名のもとに沢山の研究報告が発表されているが、“クラスター”という用語の解釈がまちまちである現状では、これに明確な説明を加えることは難しい。しかし、“クラスター”を等質な“もの”の集まりと考えたとき、そこには相対的に非等質なものが存在する。ここに、“似ていること(類似性)”あるいは“離れていること(差異性)”の計量化の必要性が生まれる。これが類似度であり距離(非類似度, 差異度)である。クラスター分析はこれを基礎において展開される分類法の広い呼び名である。

ところで、分けるもの、つまり分類の対象は、多変量特性値を扱うことから自ずと多様になる。それは、あるときは標本(個体)であり、あるときは特性(変数, 項目)、さらにその両者の場合もある。また特性は質的、量的あるいはその混合とこれもさまざまである。したがって類似度、距離の考え方も、相関係数、ユークリッド距離などに代表されるあたりまえのものから、単なる順位や関係の概念のように、ゆるやかな表現をとるものまでさまざまである。

このように、かなり広汎な分類の方法を包括する用語として“クラスター分析”が使われるようになってきたのは最近のことである。たとえばかつては Kendall と Stuart によるとクラスター分析とは変数を分けることであり、分類(Classification)は個体を分けることである、としている。確かにクラスター分析とは

Tryon, Bailey, Cattell らを初めとする心理学, 社会学などの分野の研究者が、因子分析, 主成分分析などの分析法に対し変数を分ける手続きの説明として用意した概念であった。

一方、類語に数値分類, パターン分析, 類型化, 自動分類, 教師なしのパターン認識, などがある。とくに数値分類(Numerical Taxonomy, Numerical Classification)は Sokal らを中心とする生物学者によって独自の発展をみた分野であり、本来のクラスター分析が意味する概念とかなり内容を異にする。こうした大きな流れに、他の多くの分野の分類法が合流しそれがクラスター分析という言葉のもとに広がった。

クラスター分析の動向を知るために、研究の流れを眺めてみると、研究報告の数が1960年を境にして急速に増えているが、このきっかけは Sokal と Sneath (1963) の著書“Principles of Numerical Taxonomy”に負うところが多い。この著作は一般人にはなじみの薄分野の、しかも変わった表題でありながら大衆的で広い分野の研究者に利用可能な内容であったこと、とくにデータ解析を計算機と関連させ実用に結びつく手法を紹介したことがこうした分野の先鞭をつける起因となった。数値分類法は生物学の古典的な分類理論に対する改革案として提唱されたものであり、その根底には系統分類の計量化という思想がある。よく知られた樹木図(デンドログラム)に代表される階層的あるいは系統的分類法の大部分の手法の源流がここにある。1960年代も半ばを過ぎるとそれまでの統計理論(主として多変量解析, 分散分析, 標本抽出論など)に基づいた自動分類の方法が現われる。クラスターの等質性の基準としてクラスター内分散, 非等質性の基準としてクラスター間分散(あるいは分散の代わりに平方和)を考えこれを適当な算法で最適化する方法が、その代表的なものである。MacQueen の k-means 法,

* 統計数理研究所

Friedman と Rubin の方法, パターン分類の Ball と Hall の ISODATA 法, あるいは社会調査データ分析を目的とする Sonquist らの AID, 経済データ分析における W. D. Fisher の考え方などがこうした手法の先駆けとなった。

この2つの大きな流れの他に, 多くの分野で60年代後半から70年代の初めにかけて相当数の論文, とくに手法の提案に関連した文献, が現われている。そしていずれも最適化の算法に工夫があり計算機の助けを借りて初めて利用が可能となる点に特徴がある。

最近はおそらく手法の提案, 応用が中心課題であった時期を過ぎて手法の評価, 比較, 性質の吟味, シミュレーション実験などの報告が増える傾向にある。いずれにせよこのような研究報告の急速な増大はあらゆる点で混乱を引き起こす。こうした状況を要約すると次のようになる。1) クラスタ分析に関心を持ち, それを応用している分野, 研究者が多方面にわたりきわめて学際的である。たとえば, 医学はもちろん, 生物学, 植物学, 心理学, 社会学, 人類・考古学, 生態学, マーケティング, パターン認識, 計算機科学全般, …と際限がない。2) したがって, 沢山の手法がある。3) 専門分野ごとに特殊な術語, 用語を作るので混乱が起こる。4) しかも分析の目的別に方法を考えるのでそれぞれの分野別にみた理論的アプローチがあっても流動的でどれも確立されていない。5) しかし一方では, Ralston⁸⁾の著書に Kruskal の MDSCAL とともに2編の論文が取り上げられ, それが Lance による自動分類一般の解説と Hall による ISODATA の解説, という前述の2つの流れの代表的方法である点是非常に興味深い。Ralston の同シリーズの第1, 2巻が統計解析の歴史, とくに計算面ではたした役割り(重回帰分析の変数選択, 固有値の計算法など)と結びつけて考えると, データ解析ではたすクラスタ分析の位置づけを示唆しているようである。6) 手法間の比較検討の試みはあるが手法ごとの誕生の背景が異なるから一概にどれが良いといった評価が難しい。7) 統計学を志す研究者がクラスタ分析にあまり関心を示さなかった。最近やっと理論, 応用両面から参加する姿勢がみえるが立ち遅れは否定できない。8) 計算機の力を借りなくてはデータ処理が難しい。プログラムがあって初めて利用できる。いずれの分野で

も先鞭をつけた人達はこの作業をしている。しかし, 結果として数多くのプログラムが氾濫し理解不足による乱用が目立っている(国内はとくにこれが激しい)。

こうした現象に対する当然の反動として, 分類法の分類あるいは研究文献の総合報告などが1970年代の半ばから現在に至る間に現われている。レビューの主なものに, Ball, Cormack, Bailey, Gower, Sokal, Bolshvov などの報告がある。著作物としては Anderberg, Hartigan, Sokal らの著書を初め数多くのものが出版されている。また, 数値分類, パターン認識などクラスタ分析に関連のある研究集会も数多く開催されその内容は議事録として出版されている。

一方, 国内に目を転ずると, 1971年の矢島ら²⁶⁾による D. Wishart の CLUSTAN-1A の紹介とその若干の応用例を, また生物測定分野での利用をきっかけに, 一連の組み合わせの階層手法が広く知られるようになった。これとは別に堀内, 井関らにより AID, Association Analysis といった手法によるライフ・スタイル分析がマーケティング分野で利用されてきた。しかしいずれも個別的に特定の範囲内に限定された議論であったために互いの関連を確かめるところにまで至らず, クラスタ分析のある側面だけが強調されるに留まった。このため利用者の一部にクラスタ分析が非常に限られた手法であるとの誤解を生んでいることも事実である。

2. 手法の概観

前述のことからクラスタ分析の手法のすべてをここで述べることは不可能に近い。また過渡的, 流動的な時期にあって独断的に手法を分類, 選択して紹介することも適当ではない。ここでは過去の常識的な分類方法にならって, 手法を2つの枠組, (I) 階層的あるいは系統的分類法(凝集型, 分枝型がある), (II) 分割型(分割型)の分類法, に限って紹介するに留めよう。

2.1. 階層的分類法

なぜ凝集型の階層的分類法(Agglomerative Hierarchical Clustering Method)が広く利用されるのか, そして沢山の手法が提案されているにもかかわらず利用される方法, 支配的な方法はある範囲に限られてしまうのか, この疑問に答えることが階層的手法の特徴を説明することになる。

1) まず Sokal, Sneath; Lance, Williams; McQuitty らの精力的な研究に負うところが多い。彼らはそれまで散在していた手法を整理するだけでなく、データ解析の道具として利用できるように計算機処理に結びつけて考えた。また、その背景にこれを支援する計算機ソフトウェアの急速な発達があった、2) 結果として、多くの文献に彼らの方法が引用され広く普及した、3) 手法の原理と性質が簡明で専門知識にもとづく経験則を分析結果に反映させることが容易である、4) しかし、算法の設定基準がまちまちであることから、結果の評価に対する意見の収束が難しい、などが指摘できよう。

一方手法の乱立と分裂化に刺激され、実用面から系統的に手法を統合整理しようという試みや、理論面から数学的な公理にもとづいて手法を説明しようという機運が出てきた。前者の1つに Lance, Williams らを中心とするオーストラリアの一派による仕事がある。彼らは手法を分類のためのデータ解析の道具として合理化するという立場から算法中心あるいは計算機向きの方法論を展開した。その1つが“組み合わせ的な手法”として知られている一連の手法の一般化である。彼らは、クラスター間の距離の関係をある回帰的表現で定式化すると、それまで提案されていた方法も含めて何通りかの手法が統合化できることを示し、これを組み合わせ的 (Combinatorial) と名づけたのである。

また Jardine, Sibson らに代表される英国の研究者は数学的な理論、とくに代数的公理、グラフ理論などを駆使して手法を評価する基準を用意し、これによって手法の評定を行い、加えていくつかの手法を提案した。彼らの主張は、“類似度の1つの変換法が手法であり解は階層的樹木図 (デンドログラム) で表現できる。したがってもとの類似度に含まれる情報をなるべく忠実に再現する、つまりもっともみずみの少ない変換法 (算法) が良い手法である”，と要約できる。そしてこれを評価する基準とその性質をいくつか提案した。

その他、Gower の類似度と手法の関連についての主張、Ling, Hubert らの一連のグラフ理論にもとづく評価、McQuitty の独自の発見的な方法論の展開 (実は大部分のものが他の手法と同義であるが考え方が面白い)、離散データの著名な研究者でもある I. J.

Good の分類に対する独特な考え方、など枚挙にいとまがないが、いずれも研究分野と考え方の相違から同じ組上での議論は難しい。

また分枝型の系統分類法として、Association Analysis (関連分析という妙な訳がつけられている)、AID (Automatic Interaction Detector) 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、多少趣きのかかった分類法の1つである。両者ともそれなりの事情を反映して登場した方法であるが考え方の基本は次に述べる分割型の手法に類似するところが多い。

2.2. 分割型の分類手法

いくつかの例外を除いて (たとえば Ward 法、k-means 法を利用したものなど) 階層的な手法は類似度 (距離) 行列が計算の出発点となる。この方式による計算の効率性は扱うデータ数に大きく左右される。これに対し、与えられたデータを適当な個数のクラスターに分ける分割型の手法は大量のデータの計算処理が可能であるが、階層的な手法とは別の制約が生ずる。その1つは最適な分類とは何か、それを約束する“最適化基準の設定”の点にある。第2は、その基準の最適化を達成する算法の工夫である。しかし分割型の算法の骨子はきわめて簡単で、i) 全データを適当な個数の群に分割する (初期分割)、ii) 次に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割りあてる (所属の決定、配置)、iii) 各群の平均ベクトルを更新する、iv) ii), iii) を適当な基準 (たとえば群内平方和の総和が最小、など) を満たすまで繰り返し (反復計算)、最後に出来上がった群をクラスターとする、と要約することができる。

このようにあらかじめ指定したクラスター数 (k) と平均ベクトルの更新により基準の最適化 (といっても一般には局所最適化となるが) をはかる方法を k-means 法と総称している。この範疇に入る方法の特徴として 1) 初期分割あるいは代表点 (初期平均ベ

クトルの代用とする目安)の選び方, 2) データをクラスターに配置, 再配置するときの方法と平均ベクトル更新の時期, 3) クラスター・サイズ(クラスター内のデータ数)が異常に小さい場合, あるいは逆にチラバリが大きくなった場合の手当ての方法, 4) 異常値に対する手当ての方法, 5) クラスター数の決め方, 6) 最適化の基準とそれを達成する算法の工夫, などが列挙できる。

たとえば最適化の基準だけを考えても, クラスター内分散・共分散または平方和・積和行列(W), クラスター間分散・共分散あるいは平方和・積和行列(B)などをもとにした基準(Wや $W^{-1}B$ のトレース, 行列式の最小化など)とその変形など沢山のものがある。したがって, 上の条件設定の組み合わせを考えると無数の方法ができる。しかし, 分割型の特徴は, その基準からみて背景に複数個の多変量正規分布の混合またはそれに近い状況を想定した場合に適した方法である。実際のデータ解析では, そうした傾向があるかどうかを知る術がないし, むしろ潜在的な性質を把握したいから利用するのであるから, 利用者は, クラスターがどのような基準のもとに作り出されたものであるかよく承知してかかる必要がある。分割型では大量のデータを扱うことが多いのであらかじめデータを複数個の組にランダム分割し, 各組ごとのクラスタリングを行い結果をくらべる, 初期条件をかって結果を比較する, など手間をかけた試行錯誤が妥当な分類結果を手にする早道である。

3. 新しい動きと今後の課題

こうした大きな流れに対し, 統計理論や新しい理論の積極的導入の兆のあることも見逃がせない。たとえば, クラスター数の決定あるいはクラスター・モデル(仮説)に対する統計的検定論の応用がその1つである。Wolfeは確率分布の混合問題(mixture problem)としてクラスター分析を考え, 検定論を使ってクラスター数を決定するという方式NORMIXを提案している。しかし, 最後の決め手となる検定の手続きで若干の問題が残し, 立ち入った言及を避けている。筆者の若干の実験では何か適当な規準(たとえば赤池の情報量規準など)を取り入れることでかなりよい結果が得られることがわかってきた。また分散分析における検定と類似の提案も多くみられる(Beal, Hartigan,

Calinski と Harabasz 等¹⁰⁾¹²⁾). さらに, もっぱら利用中心であったAIDなどを統計理論で補強しようというG. Kass, A. Scottらによる試みも現われている。一方, Hubert, Baker, LingあるいはJardine, Sibsonらによるグラフ理論と統計理論の接点に立った考え方も無視できない。こうした既成の理論をとり入れた方法論の展開に対し, 理論自体が未熟な段階にある新しい概念を積極的に導入しようという動きもある。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, あいまいな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くつか提案し, 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・グラフ”あるいは“ファジ代数”などを応用した研究も目立つようになってきた。

このように多くの研究者が広い視野からクラスター分析を扱い, 理論, 応用の両面から次第に発展しつつある。しかし利用の側からみると, 実用に耐える方法はいまだ少なく, 今後クラスター分析がデータ解析の重要な一翼をになうにはプログラム・ソフトウェアの開発が必須である。これを敏感に予知した一部の研究者はクラスター分析のパッケージを作成, 公開している。たとえばD. WishartのCLUSTANやJ. RohlfらのNTSYS(Numerical Taxonomy System)などは比較的体裁よくまとめられている。また, 規模の大きい統計システムにも次第にクラスター分析が編入されつつある(BMDP, OSIRIS-III, SASなど²⁷⁾).

以上のように“クラスター分析”の言葉で代表される守備範囲は多様である。データ解析に際し, とかく, 何のために分類するのか, 何のために使うのかその必然性についての配慮を忘れがちである。利用者は改めてこれを熟慮すると同時に, 手法の限界をよく心得えて利用することが大切であろう。

- 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 Klassifikation, Vandenhoeck und Rupprecht. 1974.
- 5) Clifford, H. T. and Stephenson, W. ; An Introduction to Numerical Classification, Academic Press. 1975.
- 6) Cole, A. J. (ed.) ; Numerical Taxonomy, Academic Press. 1969.
- 7) Duran, B. S. and Odell, P. L. ; Cluster Analysis : A Survey, Springer Verlag. 1974.
- 8) Enslein, K., Ralston, A. and Wilf, H. S. ; Statistical Methods for Digital Computers, Vol. III of Mathematical Methods for Digital Computers, John Wiley. 1977.
- 9) Estabrook, G. F. (ed.) : Proceedings of the Eighth International Conference on Numerical Taxonomy, Freeman. 1974.
- 10) Everitt, B. ; Cluster Analysis, John Wiley. 1974.
- 11) Fisher, W. D. ; Clustering and Aggregation in Economics, The Johns Hopkins Press. 1969.
- 12) Hartigan, J. A. ; Clustering Algorithm, John Wiley. 1975.
- 13) Jardine, J. and Sibson, R. ; Mathematical Taxonomy, John Wiley. 1971.
- 14) Lerman, I. C. ; Les Bases de la Classification Automatique, Gauthier-Villars. 1970.
- 15) Ryszin, J. (ed.) ; Classification and Clustering, Academic Press. 1977.
- 16) Sneath, P. H. A. and Sokal, R. R. ; Numerical Taxonomy, Freeman. 1973.
- 17) Tryon, R. C. and Bailey, D. E. ; Cluster Analysis, McGraw Hill. 1970.
- 18) Bailey, K. D. ; Cluster Analysis in Sociological Methodology, (ed.) Heise, D. R., Jossey-Bass. 1974.
- 19) Ball, G. H. ; Data Analysis in the Social Sciences ; What about the details?, Proc. of Fall Joint Computer Conferences, 533-559, 1965.
- 20) 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.
- 21) Bolshev, L. N. ; Cluster analysis, Bull. of I. S. I., 43, Book 1, 411-425, 1969.
- 22) Cormack, R. M. ; A Review of Classification, J. R. Statist. Soc., ser. A, 134, 321-367, 1971.
- 23) Good, I. J. ; Categorization of Classification in Mathematics and Computer Science in Biology and Medicine, Proceedings of Conference held by MRC, 1964, H. M. S. O. 1965.
- 24) Gower, J. C. ; Classification Problems, Bull. of I. S. I. 1973.
- 25) Sokal, R. R. ; Clustering and Classification : Background and Current Directions in Classification and Clustering, (ed.) Ryszin, J., Academic Press. 1977.
- 26) 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.
- 27) 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.