Special Edition

多次元データ解析における分類手法の役割
—分けて知ることの効用と難しさ—

■1972年、文部省統計数理研究所、第4研究部研究員。79年理学博士取得。85年同調査実験解析研究系助教授。91年より現職。

大隅 昇 | Ohsumi Noboru

統計数理研究所調査実験解析研究系
パターン解析研究部門教授

1. 分類とは？

多次元データ解析手法の中で分類手法（とくにクラスター化法）がはたす役割は重要と思われるが、現実にはこれが適切に利用されてきたとは限らない。理由は多々あるが、分類に関わる手法が多すぎること、適用方法があまり明らかでないこと（どんなデータにどう用いるべきか）、アルゴリズムの中味が不透明なこと等が考えられる。加えて分類問題自体がかなりの難問であるということもある。ここでは、そもそも分類とは何か、分類手法をどう考え、何が問題とされてきたか、実用上はどう対処すべきかを述べてみたい。

そもそも分類とは、物事や事象を体系的に理解するための基本操作である。ソクラテスを持ち出すまでもなく、古来より科学史、哲学、博物学、植物学、本草学等々、多くの学問・思想の歴史の中で**分類の概念**が占めてきた役割は大きい。ジョルジュ・ペレックは著書「考える／分類する」の中で、「『分類する前に考えるか』あるいは『考える前に分類するか』」とも問うている[12]。このように分類操作は、それを意識する・しないにかかわりなく、人の思考の基本形態である。またそれが自然のことであり必要な行為であるがために、分類を一つの研究分野や「学」として捉える姿勢が希薄であったと言える。これはデータ解析で「データ」を扱うことの重要性が当然であるがゆえに真剣に考えられなかったことに相通じるものがある。

話題を絞って、統計的データ解析に関わる方法論の中での分類操作の位置付けを考えよう。言うまでもなく分類に関わる方法論は多数ある。例えば、ヒストグラムを描くこと（層別化）から、各種層別化手法、判別分析、クラスター化法、その他多くの方法論がある。分析対象の様相を、**分けて理解するため**の操作・方法を工夫することが分類手法の目標で

あるが、あまりに当たり前のことであるがために、データ解析方法論としての構築が軽視されてきたきらいがある。一方では、分類操作があらゆる分野、諸現象に深く関わることから、分類の目的も様々であり、結果として多種多様な分類手法が存在する。

とくに1960年代後半から、コンピュータ利用環境の進展に連動して、単純な分類操作を越えたコンピュータ利用を前提とするアルゴリズム的な分類法、とくに**数値分類法**(numerical taxonomy)が、生物系統分類学の分野で多数登場した。また、1970年代～80年代にかけて、現在利用されているほとんどの分類手法が登場し、これらが商用の統計ソフトウェアに組み入れられることで、多数の利用者が自由に利用できる環境が整ってきた。与えられたデータセットから、しかるべき分類アルゴリズムに従ってコンピュータを用いてほぼ自動的に分類するという意味では**自動分類法**と呼ぶのがふさわしい。しかし、他の多次元データ解析手法と同様に、抱える問題は多々ある。忘れてならないことは、現存する多くの分類手法が、生物分類学、計量心理学、生態学、社会科学等の諸分野の現場における必要性に迫られてそれぞれの研究分野の目的・要請に応じて登場してきたことである。従って各分野固有の「方言」を用いた記述となり、このことが一般のデータアナリストにとっては混乱の元となる。階層的分類法が良い例で、類似や同等の手法が多数あり、しかもそれぞれが異なる名称を持つ。

2. 分類手法の分類

このようなことから、分類手法を語る際に必ず登場するキーワードが「分類手法の分類」である。実際に多くの紹介記事、総合報告、書物がある[3][4]。これを「**階層的分類法—非階層的分類法**」、「**教師あり分類**(判別分析的)—**教師なし分類**(クラスタリング的)」と対比させる場合、また階層的分類法をさらに「**凝集型—分枝型**」と区分する、さらには、分割最適化型手法、ダイナミック・クラスタリング、ファジイ・クラスタリング、分布の混合問題等々、依拠する分類の規範のありようによって様々の視点がある。最近は、ソフトデータ解析、ファジイ理論やラフ集合、シンボリック・データ解析(SDA)、さらにはデータマイニングとの関連でコホーネン自己組織化マッピング(SOM)やニューラルネットワーク等々、様々の技法を分類手法構築に適用することが流行である。

しかし、問題の本質は手法の分類のみにあるのではない。分類操作を必要とする場面では、分類対象や事象の解明に必要なより処として、どのような分類方法が適切であるかを知り、引き続く解析のための指針を与えることにあるが、これは同時に与えられた問題の数だけ分類の道筋があるという見方もできる。このようなことで、分類手法を実際に用いる場合の留意事項や未解決の課題にどう対処するか、それらの主な事項について述べる。

3. 自動分類における課題とは？

B. Everittが1979年に指摘したクラスター化法が抱える問題を改めて眺めてみる[2]。この頃には既に多数の階層的分類法を含む数値分類法(群平均法、メジアン法、最短距離・最長距離法等)が登場している。また、基本的

な教師なし分類法（分割最適化型手法、ダイナミック・クラスタリング、ファジイ・クラスタリング）、分布の混合問題等も現れている。さらに、当時のコンピュータの主流であったメインフレーム機上で動作する多数の応用ソフトが登場している。こうした当時の状況を背景に、Everittが指摘した主な問題は以下のようなことであった。

①用いる変数はいかに尺度化すべきか？
②類似度や非類似度（あるいは距離）の測度はどれを用いるべきか？
③（得られた）クラスターの安定性や妥当性をどう検証するのか？
④クラスターの有意性・意味をどう評価するのか？
⑤クラスター化の手法としてどれを用いることが適切か？
⑥クラスター数（群の数）をどう決めるのか？
⑦（クラスター化を）実現するための具体的なソフトウェアの存在は？

その後二十数年、多種多様な試みがなされてきたものの、現状はほとんど未解決のままである。分類問題はそれほど難解な課題と言うべきかもしれない。とくに、③や④への具体的な解は未だないと言って良い。しかし実用上は何らかの対処策を必要とし、事実、経験的な対応方法が多数提案されてきた。多くの統計ソフトウェアでは、経験則を多様なオプション機能として取り入れることで、利用者の要求に応えようと努めている。

最近の統計ソフトウェア利用環境では、オブジェクト指向や高度のプログラミング技術、GUI技術を駆使して、データ入力はもとより解析手法の適用、結果の出力、そのグラフィカル表現と、多様な機能がユーザの意図するままに操作できる能力を備えている。ユーザフレンドリなインターフェースはどうあるべきかというかつての議論がまるで嘘のようである。しかし、何もかもがあまりに平易にできることから（実はそのように見えるだけである）、分類手法の具体的な処理・計算手続きの内容を不透明化させる結果となり、簡単に出力結果を得たものの、その意味解釈や理解が徹底しない、誤用濫用があっても看過されるという現象が見られる。階層的分類で良くある例だが、同じ手法名を掲げているものの、類似度・非類似度の選択や階層化手順の違い、独自に設けたオプションの差異などから、同一データに対して、用いたソフトウェアの出力結果がまるで異なることさえある。

また、分類手法に限らず多変量解析一般、統計科学の知識だけではカバーしきれない新たな問題として、コンピュータと統計ソフトウェア利用の接点で生じるリテラシーの再構築を求められている。多様なデータ処理ができるようになったことが、かえって統計手法への正しい理解を遠いものとしてしまったという皮肉な現象となったかに見える。

(1)尺度化、標準化等の影響

測定値の単位や桁の変更（乗除）はバラツキと間隔を変えることに他ならない。例えば、平方和あるいは分散（言い換えると平方ユークリッド距離）を考えてみれば良い。また、主成分分析や対応分析・数量化法Ⅲ類等で得

た成分スコアを用いて分類操作を行う場合にも類似の問題がある。つまり元の変数や成分スコアの標準化を行うか否かで、分類結果が異なるからである。こうした自明の操作が分類では無視できないが、統計ソフトウェアを用いる際にはユーザはほとんど無関心のまま、デフォルト値を用いて機械的に処理する。また、別の問題として、定性データや質的データに対して、こうした操作を行うという誤った処理が行われることもある。

(2)データタイプと用いる測度

分類手法も他の多次元データ解析手法と同様に、扱うデータタイプや出発データ行列にいろいろな型がある。階層的分類であれば、多くは分類対象間の関係を示す**類似度・非類似度行列**を必要とする。あるいは**距離**（ユークリッド、平方ユークリッド）と分散または平方和の関係を利用した測度が必要となる。その事前手当として、用いる特性に合った類似度・非類似度、距離として用いる指標を決めねばならない（かなりの難問である）。

一方、分割最適化型手法（例えばk-平均法やその変形手法）のように、**クラスターの等質性基準**（例えばクラスター内平方和の和）を仮定し、これの最適化を行う場合もある。つまりはどれを用いるにしても、与えられたデータのタイプと測度との関係を無視しては成り立たない。また、多くの場合は**量的データ**（区間尺度、比例尺度）への適用が多い。結果として、初めに行うコーディングの問題、類似度（非類似度）の選択、分類手法の選択、最適化の方法、…とその組み合わせは膨大な数となり、一体どれをどう用いれば良いかが不透明となる。断片的な研究はあるものの十分ではなく、利用者はこの壁を越えて、十分に内容を理解し対応する必要があるが通常は提供情報が十分でない。

(3)クラスター数の問題

クラスタリングで必ず生じる疑問が「クラスター数はいくつとすべきか」ということであるが、これの答えはない。とくに深刻な問題は階層的分類で生じる。分類対象間の類似関係を階層化表現するわけであるから（例えばデンドログラム－樹木図）、すでにこの操作が一種の近似表現である。加えて、多数の手法があるから適用した手法で解（つまり階層構造）が異なる結果を与える。従ってクラスター数も自ずと異なった結果を与える。用いた類似度（分類対象間の関連性）と分類手法（アルゴリズム）の関連を様々の見地から数理的に説明する試みがあるものの、実用上は確定的にうまい方法などはあり得ない。ウォード法（Ward's method）が良いとの報告が多いが、それはこの手法が平方ユークリッド距離のみに適用され、そのことがクラスター化等質基準の平方和の最適化と同等であるという数理的に自明の仕組みの中で構築された手法であるがゆえに、“いかなるクラスターが生成されたか”が明らかであるというからに他ならない。

つまり、クラスター数の問題は与えられたデータ（分類対象）の性格や分析者の意図する仮説に依存して決まることであり、一般的な解があるとは思われない。対処法としては、同一データに種々の手法を適用し、クラスター化の結果を相互に比較する方法や、クラス

ター数を変化させたときのクラスター生成過程（履歴）を追跡する合理的な方式を検討することが有効である。つまり、きわめて**探索的な操作**が求められる。

しかも多くの場合（少なくとも筆者の経験では)、本来は分かれているとは思われない“ほとんど差異が識別できない”対象を区分すること（いわゆるdissection）も含めて、そもそも存在の有無が明らかでないクラスターなるものの探査を行う手順が重要であり、それゆえ**クラスター化**（クラスターは生成されるもの）と考えるべきである。むしろ分析者が、分類対象に対してどのようなクラスター構造を想定するかの仮説設定が求められる。この点で判別手法（discrimination）とは考え方が異なる（群があることを想定して判別することと群の存在も分からない中で群を生成することの違いは大きい)。さらに、はずれ値の影響も無視できないし、場合によってははずれ値の検出に用いることが有効なこともある。

分布混合モデル（mixture models）と考えて、混合するコンポーネント数（つまりクラスター数に相当）を推定する方法も考えられている。これはコンピュータの計算力を借りたシミュレーションや高度な計算処理が可能になったからである。しかし、理論的な興味は別として、データ構造が正規分布や特定な分布と仮定することに既に無理があり、そのまま実用レベルで利用するには未だ問題がある。

こう考えると、クラスター数の問題は、そのままクラスター化生成手順の問題でもある。さらに遡って、分析対象となった事象の解明にとって必要な分類法とは何かという個別的な問題となってくる。さらにはデータ科学の見地から、いかなるデータ取得を工夫せねばならないかということに行き着く。ともあれ分類手法の利用価値は高いがゆえに、様々なアプローチが登場したとも言える。現在できることは、内容の良く分かった既存の方法を要素技術として、これを分析目的に合わせて結合化（システム化）することであろう。多くの場合、分類操作はデータ解析の出発点であり、分析対象の中に潜在する共通性や類似性を探査し、次の分析に進むガイドとなる情報を提供する。この意味で、上に提起された様々な問題への対処策や留意事項を知っておくことは、分類手法の利用者にとって無駄なことではない。問題のすべてに解答を与えることは当面は無理な話であるが、実用的見地から、少なくとも誤ってはならない初歩的な手当だけは心得るべきであるが、これを簡単な数値例で眺めよう。

4．分類手法の適用例

(1)データセットの特徴

用いる例は、筆者等が独自に掘り起こしたデータセットである。オーストラリアに生息する蟻（キバハリアリ）の生態や形質分類の

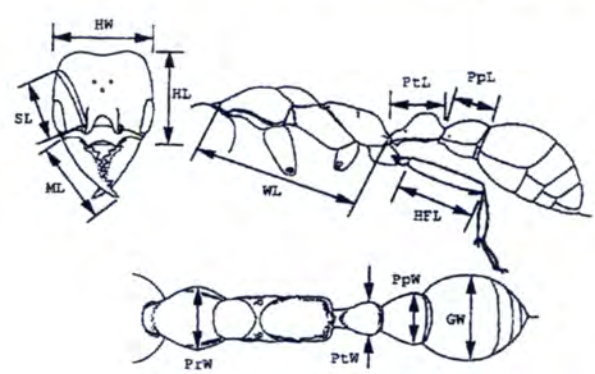

図1　キバハリアリの計測部位

研究過程で、野外観測測定で得た記録データがある。このアーカイブの観測シート上にあった測定値を新たに数値データとして書き起こしたものである[5][7]。計測部位を前ページ図1に示したが、この中の頭幅（HW）、頭長（HL）、柄節長（SL）、大あご長（ML）、腹柄節幅（PtW）、後柄節幅（PpW）、腹部幅（GW）の測定値から得た五つの加工変数（X_1=HW/HL、X_2=SL/HL、X_3=ML/HL；X_4=PtW/GW、X_5=PpW/GW）を用いる。比率とした理由は、個体の大きさの因子と測定単位の影響除去のためである。また元のデータは9の種群と約250個体からなるが、ここでは、この中から4種、約190個体を用いた。また、分析には、SAS/JMPを用いた。

この他、分析に必要な定性情報他、各種のデータが得られているが、ここではこれら五つの特性についての量的データ（比例尺度）を用いる。この5変量データの多変量連関図と統計値の一覧を示した（図2、表1）。図2で四つの種がどう混在するかが見えにくいが、図中では表示記号を変えてある。

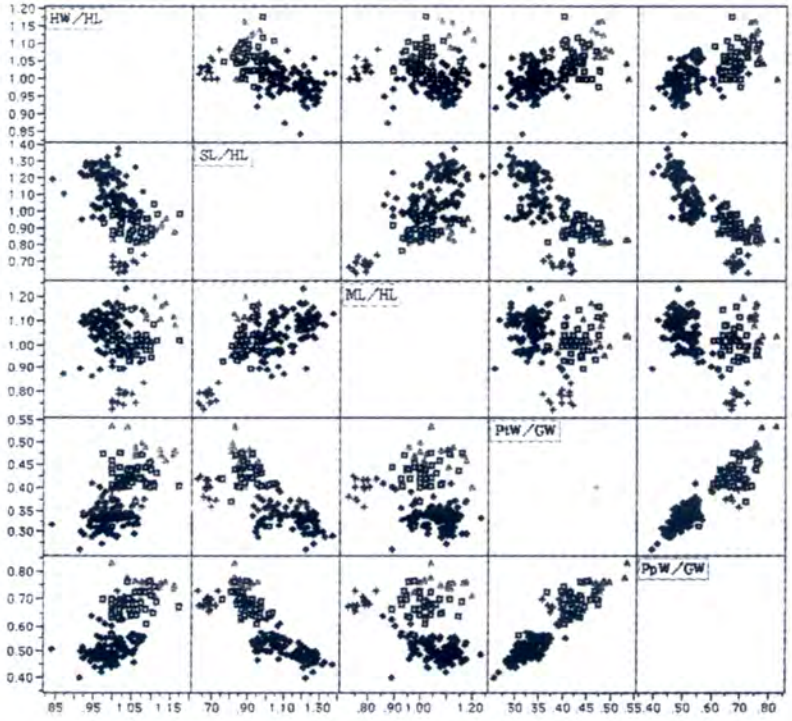

図2　5変量の関係（多変量連関図）

表1　統計値一覧（5変量の統計値）

統計値	変数				
	X_1=HW/HL	X_2=SL/HL	X_3=ML/HL	X_4=PtW/GW	X_5=PpW/GW
有効個体数	191	178	182	187	192
平均値	1.017	1.021	1.029	0.373	0.579
標準偏差	0.05330	0.17399	0.10100	0.05701	0.10243
メジアン	1.0145	1.0218	1.0417	0.3571	0.5447
歪度	0.22908	−0.17416	−1.00350	0.57442	0.43629
尖度	0.45337	−0.65600	0.96306	−0.42677	−1.17003
変動係数(%)	5.2448	17.0477	9.8183	15.2868	17.6856

表2　最長距離法による比較

	完全連結法（標準化なし）					行和
	クラスター番号	1	2	3	4	
完全連結法（標準化あり）	1	13	0	0	0	13
	2	0	39	48	0	87
	3	0	0	14	40	54
	4	0	11	6	0	17
	列和	13	50	68	40	171

【数値例1】変数の尺度化の影響

ここでは、与えられた5変数をそのまま用いた場合とそれぞれを標準化した場合（平均を0、分散を1とする）の2通りのデータセットに、同じ階層的分類法を適用した結果がどうなるかを見よう。用いる階層的手法は、代

表的な完全連結法（CL：Complete Linkage；最遠隣距離法、最長距離法ともいう）を用いる。なお、クラスター数は元の種の種群数に合わせて4群とした。得られた分類結果をクロス表とすると前ページ表2となり、ここに見るように両者の結果はかなり異なり、標準化処理の有無が大きく影響したことが分かる（表1で変動係数が変数によりかなり差があることに注意しよう）。従って一般に測定単位や桁数の変更等は分類結果に影響する。

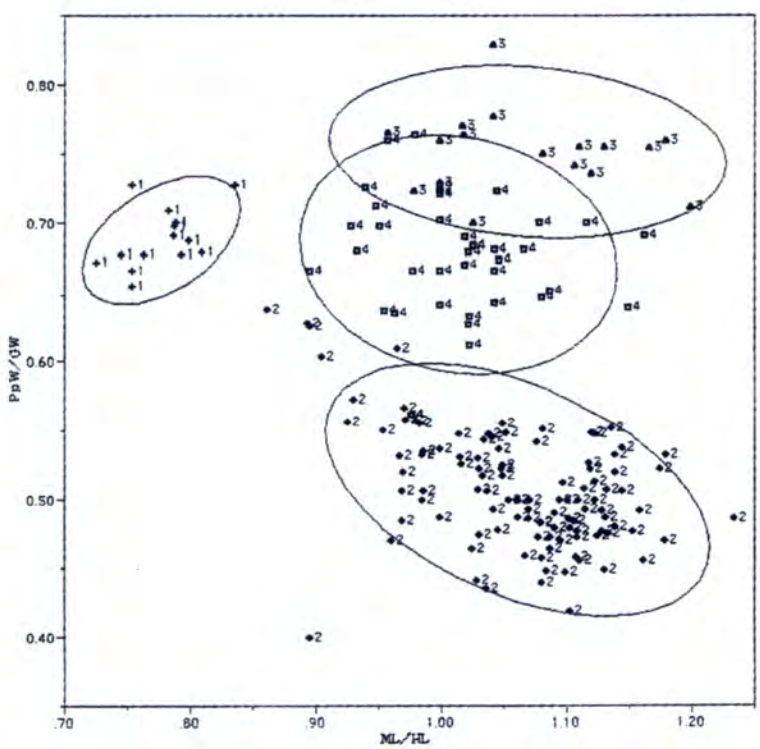

図3-1　元の種群の分類

表3　二つの手法の比較

手法	k-平均法					行和
ウォード法	クラスター番号	1	2	3	4	
	1	0	14	0	0	14
	2	49	0	0	0	49
	3	0	0	1	56	57
	4	0	0	49	2	51
列和		49	14	50	58	171

【数値例2】分類手法間の差異

では、用いる分類手法によって結果がどう異なるだろうか。元のデータに対して階層的手法のウォード法（正確にはWishartの算法によるウォード法）と分割最適化型手法のk-平均法を適用する。この二つの手法は平方和をクラスター等質性基準として用いる代表的な手法である。ここでも群の数は4とした。両者の結果をクロス表とすると表3が得られた。この場合は例1ほどの差異は見られないが、やはり僅かに不一致が生じる。ここでさらにk-平均法の初期設定条件を変更すると、やはり結果が変わる。

【数値例3】分類結果と元の情報の比較

データには蟻の種群を示すコード（分類）が与えられている。この情報は専門の研究者が判定して付与したものであり、また種群の下位の分

表4　4種群と四つの手法のクロス表

種群	ウォード法（4群）				完全連結法（4群）				単連結法（4群）				k-平均法（4群）				種群別和
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
aberrans	13	0	0	0	13	0	0	0	13	0	0	0	0	13	0	0	13
gulosa	0	49	55	0	0	50	54	0	0	1	103	0	49	0	0	55	104
mandibularis	0	0	0	17	0	0	0	17	0	0	17	0	0	0	17	0	17
pilosula	1	0	2	34	0	0	14	23	0	0	36	1	0	1	33	3	37
列和	14	49	57	51	13	50	68	40	13	1	156	1	49	14	50	58	171

類としての種の情報もある。この種群（4群）の情報と、自動分類で得た結果を比べよう。用いた分類手法は階層的分類としてウォード法、完全連結法、単連結法（最短距離法、最近隣法）を、これにk-平均法を加えた四つの手法でそれぞれ4群として得られた結果と、元の4種群の関係をクロス表として要約した（前ページ表4）。なおここでは変数の標準化処理は行わなかった。5変量データであるから視認は難しいが、多変量連関図の中で、四つの種群の区別が比較的分かり易い二つの特性（X_3=ML/HL、X_5=PpW/GW）の散布図を描いて比較しよう。前ページ図3-1～次ページ図3-5は各手法の分類結果に対応する。なお、図中の数字は元の四つの種群を示す番号であり、群のおよその位置・分布を知るために正規確率楕円（90%信頼限界）を書き入れた。

元のデータが5変量であることを考慮して結果を解釈する必要があるが、分類結果はいずれもが正しいのである。ウォード法や完全連結法ではクラスター・サイズが同じようなクラスターが生じるし、単連結法では布置データの接近する曖昧な部分で**連鎖現象**を生じ、結果としてサイズの極端に大きいクラスターとシングルトン（サイズ1個）が生まれる。つまりそれぞれの手法はそれぞれの

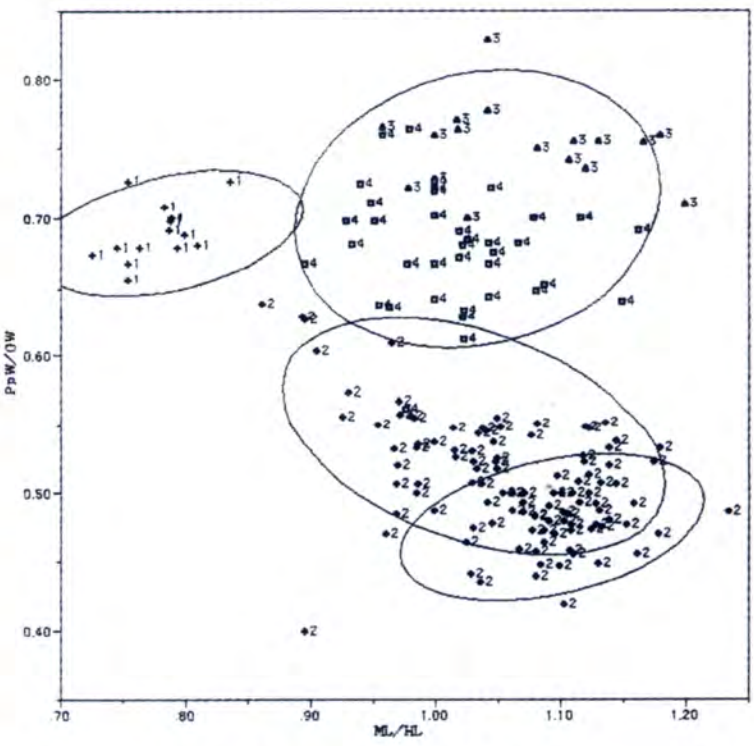

図3-2　k-平均法の場合

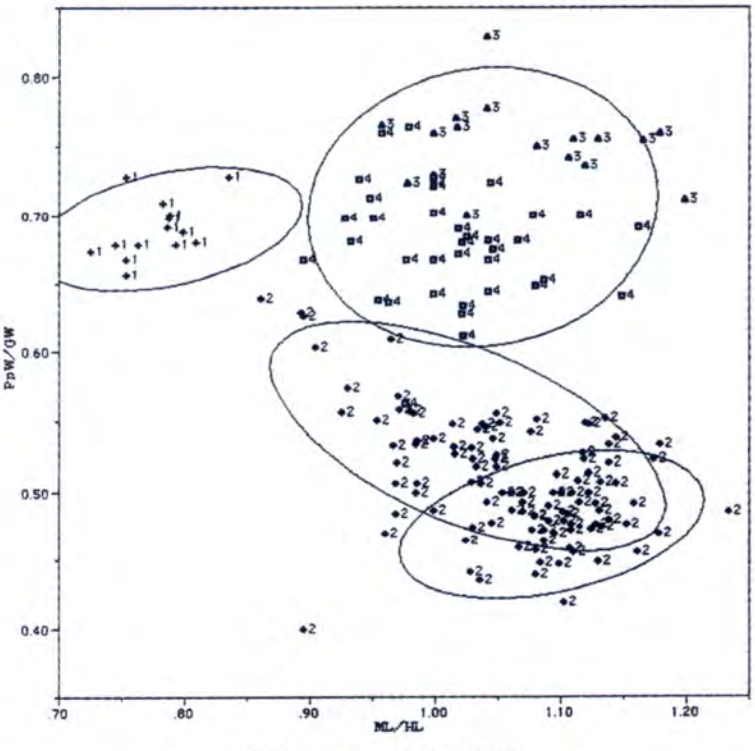

図3-3　ウォード法の場合

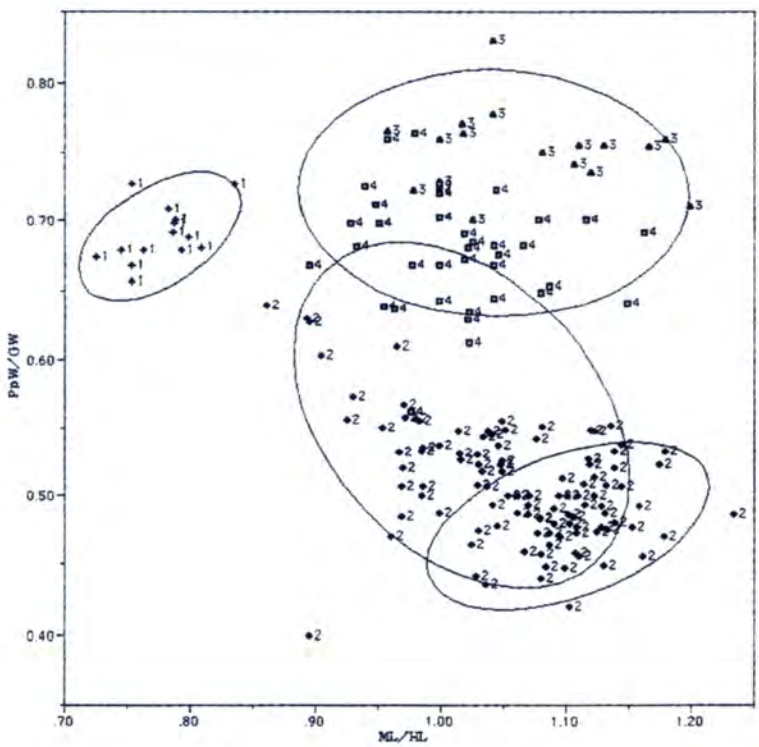

図3-4　完全連結法の場合

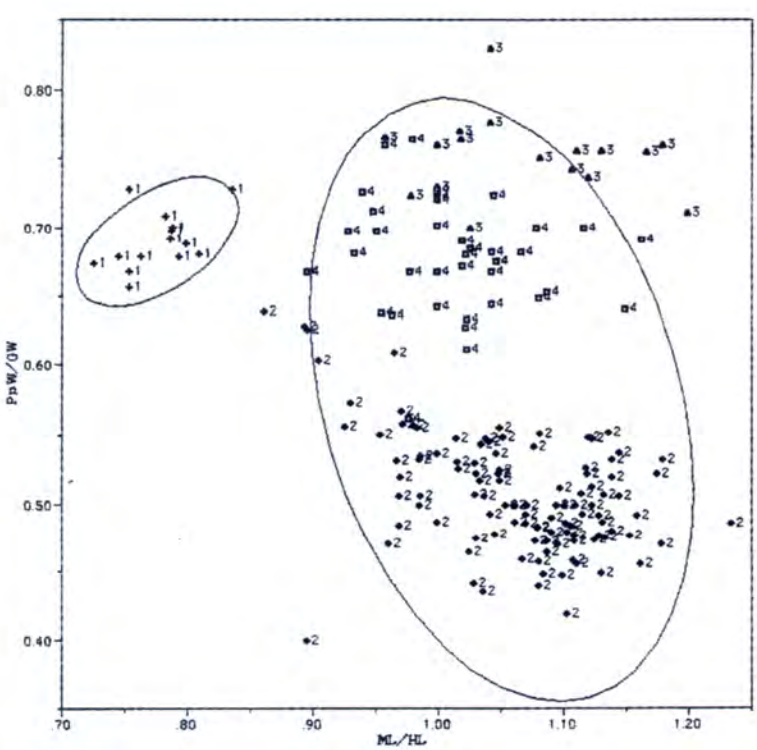

図3-5　単連結法の場合

分類基準に従って忠実に自動分類したわけである。論文などで提案方法の数値検証を行うとき、大抵は良く分かれたクラスター（well-separated clusters）を数値例に用いるが、こうした検証がほとんどナンセンスであることが分かる。現実には、ここで示した簡単な例でも既にかなり複雑な構造となっており分類に難儀するのである。

多くの統計ソフトウェアに組み入れられ、利用頻度が高い代表的な階層的分類やk-平均法による分類結果でも、上のような結果となる。実は、前述のように、このことが分類法の特徴であり、また他の多次元データ解析手法と大いに異なることである。主成分分析を用いるとき、変数の標準化の有無つまり出発行列の指定を行ったり、得られた成分スコアの標準化の有無を指示することはあっても（数値計算上の誤差程度の差異はあっても）、ここで見た分類例の差異とは異なる事象である。この例では、複数の手法を用いることで、

i）種群1はかなりはっきりとした群を構成しまた他の種群と異なる、

ii）種群2の個体数が、他の種群より大きいことの影響があり、またこの種群はさらに複数の群に分けられそうである、

iii）種群3と4との差異は分かりにくい、

などの特徴が見えるが、これ以上は他の付加情報と比較分析するか、専門的見地から、そもそも4種群とする意味はどこにあるのか、自動分類の結果を出発点としてさらなる知見を求める他の措置が必要となる。例えば、クラスター数を何通りか変えた自動分類の結果を相互比較する、種群の下位にある種のレベルで比較する、さらに他の利用可能な情報を探査するなどである。

ここで見た数値例の他、クラスター化処理を頻繁に利用する主成分分析や数量化Ⅲ類や対応分析で得た成分スコアの分類の場合、大抵はクラスター生成基準がうまく機能するようなクラスター構造となっているとは限らない。また、はずれ値の影響も大きい。つまりは、事前に分類対象データ（原データか成分スコアか）のはずれ値検出やデータ構造の特徴を慎重に探査せねばならず、さらには「一体、そのデータセットはいかなる方法で測定され」、「その事象の解明に確かに有効か」、「特性の選択は適切であったのか」という疑問にまで遡及する。つまりはデータリテラシーを意識したデータ科学的なアプローチが求められるのである。

5. 今後のこと

(1)データタイプの多様化と数値化処理上の問題

従来は、データを尺度の概念に従って**量的データ**（区間尺度、比例尺度）と**質的データ**（名義尺度、順序尺度）に分けて考えてきた。こうした区分と利用できる統計手法との間の相性についてはほとんど周知のことである。最近の統計ソフトウェアでは、ユーザが変数（変量、特性）を定義する際に、データタイプを指定しておけば、誤ったデータ処理や統計手法の選択とならぬ手当がなされている。例えば名義尺度と指定した変数について平均値や分散などを算出するようなことはない。二つの名義変数を指定すればクロス表を生成しこれに関連性指標や検定結果の算出が自動的に行われる。これが区間尺度であればただちに相関係数を求め散布図が得られる。

しかし電子技術の進歩でデータ測定環境の多様化現象が見られ、データタイプも上の区分だけでは十分に対応できない状況が生まれてきた。例えば、文字情報（テキスト型データ）、イメージ情報（静止画、動画）、音声情報などの電子的取得が急速に進み、デジタル情報として簡単かつ大量に取得可能である。このことは、分類手法に適した数値情報や数値化処理の方法論、さらには事象解明に必要な「データ取得のあり方」「情報加工の新たな方法論」の議論がデータ科学の観点からなされるべき状況にあることを意味する[10]。

(2)PC利用環境の変化と分類手法

PC利用環境の普及により煩雑な計算処理も、ある程度の環境を整えることで大方が解決されるようになってきた[11]。しかし、これだけ統計ソフトウェアの普及を見たのも1980年代以降である。それ以前はメインフレームやミニコンピュータを必要とする高度の統計解析や大量のデータ処理は、恵まれたコンピュータ環境にある一部の人々に限られていた。

しかし最近は計算機処理の有用性を武器とする**計算機集約型手法**（computer-intensive methods）やデータマイニングが登場し、分

類手法にこうした考え方の導入も盛んである。例えば既存の分類手法が姿を変えてマイニングツールとして登場している（AID、CART等）。その多くは従来の多変量解析的手法とデータベースやデータウェアハウスの要素技術の複合体と言って良い。これにノウハウやモデル特許といった鎧を着せることで、機能の多様化を競っているものの、システムの煩雑化が進み、利用者にとっては益々もって藪の中である。同一データセットに複数のデータマイニングツールを用いたところ、それぞれがまったく異なる結果を示し、利用者は意思決定どころではないという笑えぬ話もある。

6．データ科学から見た分類法のあり方

つまり、いま必要とされる概念はデータ科学（data science）である。我々が提唱してきたデータ科学とは何かは、本号特集で林知己夫先生がその趣旨を述べられている。分類法との関連で考えるならば、分析対象とする事象解明になぜ分類操作が必要とされ、なぜそのデータ特性を選び、そのデータが必要であるかの認識の下に、いかなる手段で必要データの取得を行うかから出立せねばならない。これは単に方法論の数理に精通すれば済むということではない。あまりに通俗的な言い方ではあるが、分類に必要なことは、データリテラシーを重視し、平易なことを侮らず、労を惜しまず、体験と確かな理論に裏付けされた思考が何より重要と考える。

＊参考文献

[1] B.S. Everitt (1996)：*Making Sense of Statistics in Psychology*, Oxford University Press.

[2] B.S. Everitt (1979)：Unsolved Problems in Cluster Analysis：*Biometrics*, 35, 169-181.

[3] B.S. Everitt (1993)：*Cluster Analysis* (3rd edition)：Edward Arnold, London.

[4] A.D. Gordon (1981)：*Classification: Methods for the Exploratory Analysis of Multivariate Data*：Chapman-Hall.

[5] N. Nakamura, N. Ohsumi and others (1996)：Myrmecia measurement data, *Student; Data and Statistics*, 2, 1, 55-66.

[6] 吉田政幸 (1993)：分類学からの出発：中公新書.

[7] 緒方一夫他 (1996)：アリ類における分類学の現状：キバハリアリと亜科の分類を例に：統計数理研究所共同研究リポート67,（6-共研A-57）.

[8] 大隅昇 (1996)：統計ソフトウェアとその利用環境：ESTRELA, 27号, 2-13.

[9] 大隅昇 (1992)：統計的データ解析における分類手法：統計, 43巻, 3号, 31 - 36.

[10] 大隅昇 (2000)：定性情報のマイニングー自由回答データの解析ー：ESTRELA, 74号, 14-26.

[11] 大隅昇 (1998)：統計ソフトウェアーデータの特徴を探査するー：bit別冊, インターネット時代の数学, 11月号別冊, 59-77.

[12] ジョルジュ・ペレック (2000)：「考える／分類する：Penser/Classer」：阪上脩訳, りぶらりあ選書（法政大学出版局）.

ESTRELA

2000年10月号
2000／No.79

CONTENTS　目　次