講義

クラスター分析の概要

吉原一紘[1*]，徳高平蔵[2]
[1]オミクロンナノテクノロジージャパン(株)
140-0002 東京都品川区東品川 3-32-42　IS ビル
[2]（有）SOM ジャパン
680-0941 鳥取市湖山町北四丁目 637 番地
*k.yoshihara@omicron.oxinst.com

（2014 年 4 月 4 日受理）

多変量解析の一種であるクラスター分析は，与えられたデータ群の中から似たもの同士を集めて群（クラスター）に分類する方法である．多量のデータをクラスター分析で分類することは既に各分野で頻繁に実施されており，表面分析にも今後多用されることが予測される．ここでは，クラスター分析を基礎から解説するとともに，実例として TOF-SIMS データをクラスター分析により解析した結果を示す.

Introduction of Cluster Analysis

Kazuhiro Yoshihara[1*] and Heizo Tokutaka[2]
[1]Omicron Nanotechnology Japan
3-32-42 HigashiShinagawa, Shinagawa, Tokyo 140-0002, Japan
[2]SOM Japan
637, Kita 4 Chome, Koyamacho, Tottori, 680-0941, Japan
*k.yoshihara@omicron.oxinst.jp

(Accepted: April 4, 2014)

The cluster analysis is one of multivariate analyses, and classifies the data into clusters of similar characteristics.　To classify a large amount of data by the cluster analysis is now widely applied in many fields. The cluster analysis is also expected to be applied in the surface analysis area.　This paper introduces the basis of the cluster analysis, and then shows the analytical results using TOF-SIMS data as an example.

1. クラスター分析とは

クラスター分析は，多変量解析の一種であるが，与えられたデータ群の中から似たもの同士を集めて群（クラスター）に分類する方法である．多量のデータをクラスター分析で分類することは既に各分野で頻繁に実施されており，例えば人工衛星画像の解析では，表面から反射される電磁波の波長スペクトルをクラスター分析し，都市域，森林，水面などに分類して画像表示させることが行われている．著者らは TOF-SIMS スペクトルをクラスター分析法の一種である SOM 法による分類分けを行った結果を紹介した（JSA, vol.20, No2, A-68(2013)）．その際に簡単な SOM 法の紹介を行ったが，今回はクラスター分析の基礎を簡単に紹介する．

クラスター分析には，大きく分けて階層的クラスター分析と非階層的クラスター分析の２種類がある．そのなかで，階層的クラスター分析という一群の方法が，概念的にも理解しやいので，この方法を主と

して紹介する．なお，研究会で発表した SOM 法はこれらのクラスター分析とは異なるアルゴリズムを用いた新しいクラスター分析である．

2. 階層的クラスター分析

この分析法は，データ群の中で似ている物から順番に群（クラスター）としてまとめていく方法で，クラスター分析というと，通常はこの方法を指す．

図１は７人の生徒の国語と英語の点数をプロットしたものである．図を見れば，国語も英語も出来るグループ, 国語は出来るが英語は出来ないグループ, 国語は出来ないが英語は出来るグループ，国語も英語も出来ないグループに分かれることが分かるが，これをクラスター分析という手法で分類してみる．階層的クラスター分析では，結果は樹状図（デンドログラム）として表される．手順を次に示す．

(a) あらゆるクラスター，対象間の距離（通常用いるのはユークリッド距離）を求め，最も近いもの同士を新しいクラスターとする．

(b) 新しく形成されたクラスターとその他の距離を求める．全てのクラスター，対象間の距離のうち最も近い２つを結合して新しくクラスターを作る．

(c) 全てのデータ群が一つのクラスターに結合されるまで繰り返す．

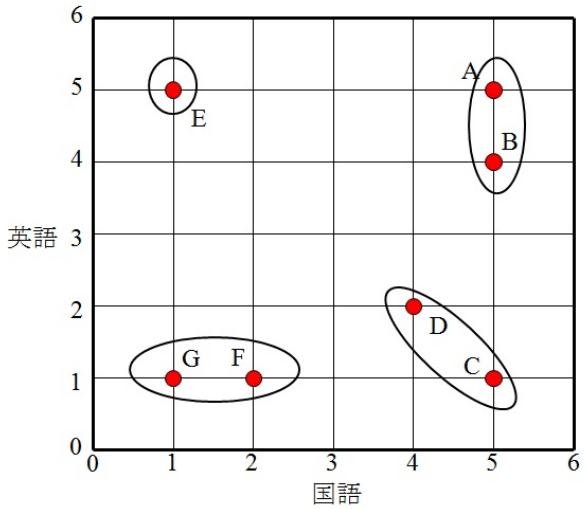

図１　生徒の国語と英語の成績の分布

この手順に従って，図１に示されるデータをクラスター分析する．クラスター間の距離の計算方法はいくつかあるが，最も普通に用いられる方法は群平均法と Ward 法である．

（１）群平均法（group average method）

例えば, 図１の生徒 C と生徒 D のユークリッド距離は

$$\sqrt{(\text{国語の成績の差})^2+(\text{英語の成績の差})^2}$$
$$=\sqrt{(5-4)^2+(2-1)^2}=1.4142$$

となる．これを全ての生徒の組み合わせについて計算する. 各生徒間の距離の計算結果を表１に示す．

表１　各生徒間のユークリッド距離

	A	B	C	D	E	F	G
A	0.00	1.00	4.00	3.16	4.00	5.00	5.66
B	1.00	0.00	3.00	2.23	4.12	4.24	5.00
C	4.00	3.00	0.00	1.41	5.66	3.00	4.00
D	3.16	2.23	1.41	0.00	4.24	2.24	3.16
E	4.00	4.12	5.66	4.24	0.00	4.12	4.00
F	5.00	4.24	3.00	2.24	4.12	0.00	1.00
G	5.66	5.00	4.00	3.16	4.00	1.00	0.00

この表の中から最もユークリッド距離が小さい組み合わせを最初のクラスターとして選定する．図１の場合には生徒 A と生徒 B の間の距離と，生徒 F と生徒 G の間の距離が「1」となり，生徒 A と生徒 B, または生徒 F と生徒 G を組み合わせたクラスターが最もユークリッド距離が小さい組み合わせとなる．したがって, まず生徒 A と生徒 B の組み合わせで最初のクラスターを形成する．次に，形成されたクラスターを一つの単位（A,B）として，残りの生徒との距離を計算して，最もユークリッド距離の短い組み合わせを次のクラスターとする．

例えば，点 A と点 C 間の距離を $d_{A,C}$，点 B と点 C 間の距離を $d_{B,C}$ とすると，点 A と点 B を含むクラスター（A,B）とデータ点 C との距離 $d_{(A,B)C}$ を以下のように定義する．

$$d_{(A,B)C}=\frac{d_{A,C}+d_{B,C}}{2}=\frac{4+3}{2}=3.5$$

このようにして AB をクラスターとしたときの各生徒間の距離を計算した結果を表２に示す.

表２　AB をクラスターとしたときの各生徒間の距離

	A,B	C	D	E	F	G
A,B	0.00	3.50	2.70	4.06	4.62	5.33
C	3.50	0.00	1.41	5.66	3.00	4.00
D	2.70	1.41	0.00	4.24	4.12	3.16
E	4.06	5.66	4.24	0.00	4.12	4.00
F	4.62	3.00	2.24	4.12	0.00	1.00
G	5.33	4.00	3.16	4.00	1.00	0.00

表２から距離の最も小さい組み合わせは生徒 F と生徒 G の間で，距離は「1.00」と最も小さい．したがって，次は（A,B）クラスターに加えて，（F,G）がクラスターになり，同様に各生徒間，及び（A,B）クラスター間の距離を求める．

二つの点を含むクラスター同士（A,B）と（F,G）の距離は以下のように定義する．

$$d_{(A,B)(F,G)} = \frac{d_{A,F} + d_{A,G} + d_{B,F} + d_{B,G}}{4}$$
$$= \frac{5 + 5.66 + 4.24 + 5}{4} = 4.98$$

表３　AB と FG をクラスターとしたときの各生徒間の距離

	A,B	C	D	E	F,G
A,B	0.00	3.50	2.70	4.06	4.98
C	3.50	0.00	1.41	5.66	3.50
D	2.70	1.41	0.00	4.24	2.70
E	4.06	5.66	4.24	0.00	4.06
F,G	4.98	3.50	2.70	4.06	0.00

表３から距離の最も小さい組み合わせは生徒 C と生徒 D の間の距離で,「1.41」と最も小さい. したがって，次は（A,B），（F,G）クラスターに加えて（C,D）がクラスターになり，残った生徒 E，各クラスター間の距離を求める．全てが一つのクラスターになるまでこの作業を繰り返す. 階層クラスター分析では，クラスター形成過程を示すために，デンドログラムを用いる．これは，横軸にクラスター間の距離をとり，縦軸に対象を適宜並べ，結合した時の距離の大きさで対象を結び，デンドログラムを作ったものである．図１の例では，群平均法を用いてクラスター分析すると，図２のようなデンドログラムが出来上がる．

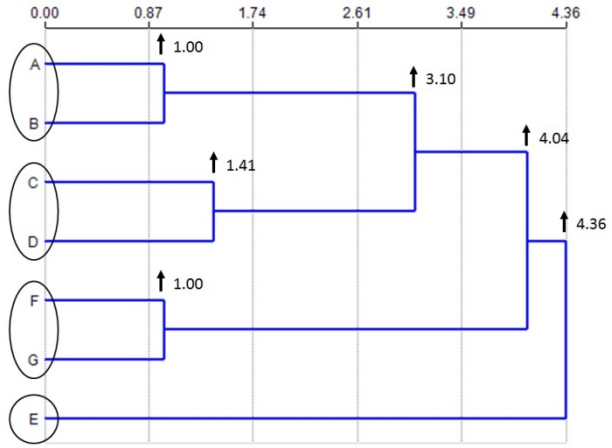

図２　生徒の成績のデンドログラム. 横軸はクラスター間の距離. 例えば, A と B は距離「1.00」, C と D は距離「1.41」だけ離れていることが分かる．縦軸は生徒の名称

図１で直感的にグループ分けした結果と同様の結果が得られている．横軸の距離はクラスター間のユークリッド距離を表しており，クラスター間の関連の度合いは距離という客観的指標で決められる．図に示した例は２変量のデータなので「見れば分かるだろう」ということになるが，スペクトルデータの様な多変量データの解析にはクラスター分析を用いなければ分類は非常に困難である．

（２）Ward 法

Ward はこの方法を提案した人の名前である. クラスター内のデータのバラツキの大小をクラスター間の距離と考える．例えば，点 A（座標：x_A,y_A）と点 B（座標：x_B,y_B）がクラスターを形成したとすると，クラスターの平均座標（x_{ave}, y_{ave}）を以下のように定義する．

$$x_{ave} = (x_A + x_B)/2 = (5+5)/2 = 5$$
$$y_{ave} = (y_A + y_B)/2 = (5+4)/2 = 4.5$$

クラスター内のデータのバラツキ（*div*）は

$$div = (x_A - x_{ave})^2 + (x_B - x_{ave})^2 + (y_A - y_{ave})^2 + (y_B - y_{ave})^2 = (5-5)^2 + (5-5)^2 + (5-4.5)^2 + (4-4.5)^2 = 0.5$$

これを全ての生徒の組み合わせについて計算する．各生徒間の距離（バラツキの大きさ）の計算結果を表４に示す．

表４　各生徒間の組み合わせのバラツキの大きさ

	A	B	C	D	E	F	G
A	0.00	0.50	8.00	8.00	17.00	12.50	16.00
B	0.50	0.00	4.50	2.50	8.50	9.00	12.50
C	8.00	4.50	0.00	1.00	16.00	4.50	8.00
D	5.00	2.50	1.00	0.00	9.00	2.50	5.00
E	8.00	8.50	16.00	9.00	0.00	8.50	8.00
F	12.50	9.00	4.50	2.50	8.50	0.00	0.50
G	16.00	12.50	8.00	5.00	8.00	0.50	0.00

生徒 A と生徒 B の組み合わせと，生徒 F と生徒 G の組み合わせのバラツキが 0.50 と最も小さい．そこで，まず生徒 A と生徒 B の組み合わせで最初のクラスターを形成する．次からのプロセスは，群平均法と全く同様に，バラツキの小さいクラスターを見つけながらクラスターを形成していく．

クラスター同士のバラツキの大きさは，Ward 法では，（A,B）と（C,D）が一つのクラスターになって，（A,B,C,D）クラスターになったとすると，以下のように定義する．まず（A,B,C,D）という一つのクラスターになった時の平均座標（x_{ave}, y_{ave}）を求める．

$$x_{ave} = \frac{x_A + x_B + x_C + x_D}{4}$$

$$y_{ave} = \frac{y_A + y_B + y_C + y_D}{4}$$

次に各データ点と平均座標の差の二乗和（div）を以下のように求める．これが，Ward 法におけるクラスター間の距離となる．

$$div = (x_A - x_{ave})^2 + (y_A - y_{ave})^2 + \cdots + (x_D - x_{ave})^2 + (y_D - y_{ave})^2$$

図１のデータを Ward 法で解析した結果を図３に示す．図３の横軸は表示を見やすくするために，バラツキの大きさの平方根にしてある．図２と分類結果はほぼ同じであるが，生徒 E と，（生徒 F, 生徒 G）クラスターとの関係が若干異なっていることが分かる．

群平均法や Ward 法以外には，最短距離法（２つのクラスターに属する対象のうち，最も近い対象間の距離をクラスター間の距離とする方法），最長距離法（２つのクラスターに属する対象のうち，最も遠い対象間の距離をクラスター間の距離とする方法）などがある．ただし，距離の算定方法によって分類結果が異なる場合がある．

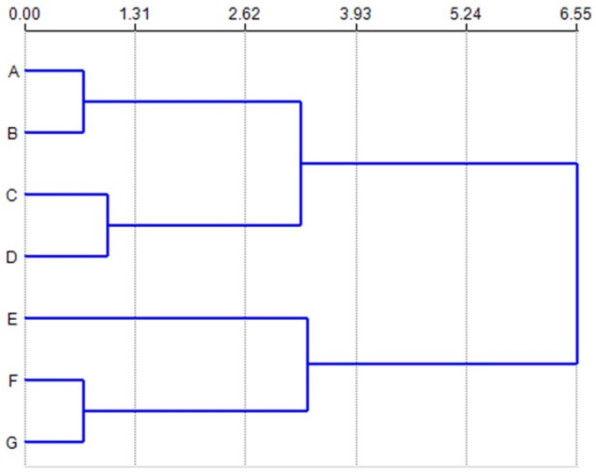

図３　Ward 法により解析した生徒の成績のデンドログラム．横軸は，見やすくするためバラツキの大きさの平方根にしてある．

3. 非階層的クラスター分析

非階層クラスター分析では，階層的クラスター分析とは異なり，結果はデンドログラムとしては得られず単にクラスターが生成される．通常は，あらかじめいくつのクラスターに分類するかを決めておき，それぞれのクラスターに代表点を初期値（任意）として与え，各データ点を最も近い代表点に割り当ててクラスターを形成し，代表点（重心）を計算し直す．これを繰り返し，全てのデータ点の割り当てが一つ前のステップと等しい場合に終了する．非階層クラスター分析は階層クラスター分析に比べて計算

容量が小さく，大量のデータを処理するには向いているが，直感的には階層クラスター分析の方が理解しやすい．表面分析の場合には階層的クラスター分析の方が適しているのではないかと思われる．

4. SOM 法（Self-Organizing Maps: SOM）

データが p 個の変量を持つとき，データ（例えば i 番目のデータと j 番目のデータ）間の相違をデータ間のユークリッド距離：$\sqrt{\sum_{k=1}^{p}\left(x_{k,i}-x_{k,j}\right)^2}$ で代表させ，ユークリッド距離が近いデータを同種のデータとみなして分類する手法が SOM（自己組織化マップ）である．

結果を表示するために，地区（地区の数は任意）を分割した白地図を用意する．最初に地区ごとにデータの次元数と同じ次元数を持つ値を番地として付ける．すなわち，データが p 個の変数を持つときには，各地区に $(w_1, w_2, \cdots, w_p)$ という値をランダムに番地として割り付ける．次に第一番目のデータの値：$(x_1, x_2, \cdots, x_p)$ に最もユークリッド距離が近い番地を探して，もとの番地を $(w_1, w_2, \cdots, w_p)$ から $(x_1, x_2, \cdots, x_p)$ に変え，その近傍の番地も $(x_1, x_2, \cdots, x_p)$ に近い値に変える．変え方の程度はガウス関数で計算する．第２番目のデータはこの番地が変わった地図を対象に，同様に最も近い番地の地区を探して，作業を繰り返す．これにより，同種のデータは同じ領域に集まるようになり，グループごとに地区が分離された地図が形成される．（以上，　JSA, vol.20, No2, A-68(2013)からの転用）

SOM 法は非階層的クラスター分析の一種であり，あらかじめいくつのクラスターに分類するかを決める必要は無く，データ点が次第に“似たもの同士”でまとまっていき，自然にクラスター分類されるという特徴がある．

5. クラスター分析に用いるデータ間の距離

データ間の距離としては，通常，４章に示したように，次式に定義されるユークリッド距離を用いる．

$$\sqrt{\sum_{k=1}^{p}\left(x_{k,i}-x_{k,j}\right)^2}$$

この式は，データがp個の変量を持つとき，データ（例えばi番目のデータとj番目のデータ）間のユークリッド距離を表す．しかし，データの大きさが変量によって大きく異なる場合には，ユークリッド距離をそのまま用いるとデータ群を正しく分離することが出来ないことがある．図１の場合には国語と英語の２教科について，５段階の成績結果を用いたが，これに算数の成績が素点で加わった結果を表５に示す．

表５　生徒の成績表

生徒	国語	英語	算数
A	5	5	85
B	5	4	80
C	5	1	45
D	4	2	40
E	1	5	90
F	2	1	25
G	1	1	95

表５のデータを用いて群平均法によるクラスター分析すると図４のデンドログラムが得られる．図４から生徒 E と生徒 G は同じクラスターに属し，そのクラスターは生徒 A と生徒 B のクラスターと近いと分類されている．これは明らかに算数の点数が素点であるために，国語と英語の成績よりも過大に評価されていると考えられ，成績表を見て，直感的に得られる感覚とは異なっている．

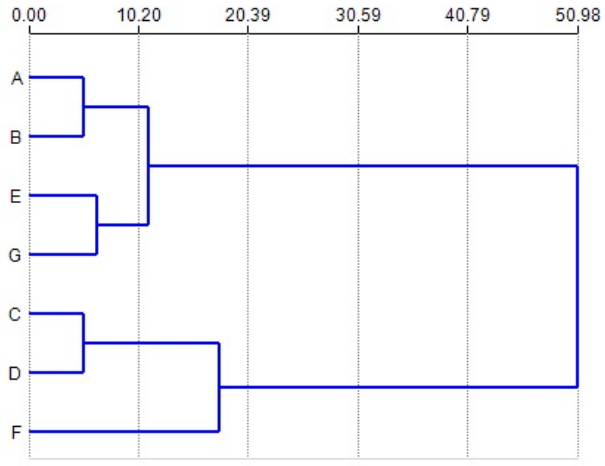

図４　表５のデータのデンドログラム

このような現象を避けるためには「データの標準化」を行えば良い．データの標準化とは平均が[0]，分散が[1]となるようにデータを変換することである．すなわち，変量 k の i 番目のデータ $\bar{x}_{k,i}$ を次式のように変換する．

$$\bar{x}_{k,i} = \frac{\left(x_{k,i} - x_{k,\mathrm{ave}}\right)}{\sqrt{\sum_{i=1}^{p}\left(x_{k,i} - x_{k,\mathrm{ave}}\right)^2 \Big/ \left(p-1\right)}}$$

ここで，$x_{k,\mathrm{ave}}$ は k 番目の変量に対応するデータの平均値である．データの標準化は，データの分布は Gauss 分布に従うという前提から導かれる．Gauss 分布はμを期待値（平均値），σを標準偏差とすると以下の式で表すことが出来る．

$$G_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

この式から，データから平均値を差し引き，標準偏差で除すと，データは平均値が[0]，分散が[1]となるように規格化できることが分かる．すなわち，データの分布がガウス分布をしていると仮定すれば，標準化することにより，データを同一に取り扱うことが出来ることを示している．

表５のデータ値は表６のように標準化される．

表６　標準化した生徒の成績表

生徒	国語	英語	算数
A	0.9071	1.2095	0.6839
B	0.9071	0.6803	0.5066
C	0.9071	-0.9071	-0.7346
D	0.3780	-0.3780	-0.9119
E	-1.2095	1.2095	0.8612
F	-0.6803	-0.9071	-1.4438
G	-1.2095	-0.9071	1.0385

表６の標準化されたデータを用いて群平均法によるクラスター分析すると図５のデンドログラムが得られる．標準化した値で計算した距離はマハラノビスの距離（Mahalanobis' generalized distance）と呼ばれる．

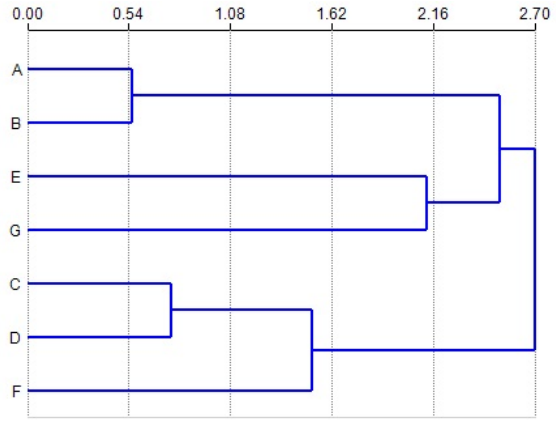

図５　表６に示される標準化されたデータのデンドログラム．

生徒 E と生徒 G は同じクラスターを作るが，生徒 A と生徒 B の作るクラスターとは距離が離れており，異なった性格のクラスターであることが分かる．

ここで述べたマハラノビスの距離（D）は一変量（一次元）の場合であり，以下のように記述できる．

$$D^2 = \frac{\left(x - x_{\mathrm{ave}}\right)^2}{\sigma^2} = \left(x - x_{\mathrm{ave}}\right)\left(\sigma^2\right)^{-1}\left(x - x_{\mathrm{ave}}\right)$$

ここでσ^2は分散である．なお，多変量（p 次元）の場合のマハラノビスの距離はこの式を拡張して

$$D^2 = \left[x_{1,i} - x_{1,\mathrm{ave}}, \cdots, x_{p,i} - x_{p,\mathrm{ave}}\right]\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}^{-1}\begin{bmatrix} x_{1,i} - x_{1,\mathrm{ave}} \\ x_{2,i} - x_{2,\mathrm{ave}} \\ \cdots \\ x_{p,i} - x_{p,\mathrm{ave}} \end{bmatrix}$$

となる．ここで $x_{k,i}$ は i 番目のデータの変量 k の値，$x_{k,\mathrm{ave}}$ は変量 k の全データの平均値，σ_k^2 は変量 k の分散，σ_{km} は変量 k と変量 m の共分散である．多変量のマハラノビスの距離は，多変量解析の一つである判別分析に主に用いられる．

6. 階層的クラスター分析による TOF-SIMS データの解析

SOM 法による TOF-SIMS データの解析結果は既に JSA, vol.20, No2, A-68(2013)に掲載しているので，今回は COMPRO を用いて階層的クラスター分析により解析を行った．COMPRO には[Multivariate analysis]メニューがあり，その中で[Cluster analysis]を実行すると階層的クラスター分析が行える．使用した TOF-SIMS データは 23 種類の PET フィルムのマススペクトルである．表７にデータの一部を示す．

表７　クラスター分析に用いたデータの一部，行は質量数（分子種），列は試料ごとのカウント数

	1	2	3	4	5	6	7
1			N1	N2	N3	N4	N5
2	11.99898...	C	22323	14957	33431	26856	28653
3	15.02343...	CH3	104604	61347	124681	27365	10357
4	27.02280...	C2H3	217804	190040	198699	31712	17857
5	27.97532...	Si	59895	25025	54543	580141	70614
6	29.03927...	C2H5	145936	156847	128788	24196	12214
7	30.99780...	CF	3696	3766	8173	15424	8106
8	41.03921...	C3H5	303193	300003	283310	46246	26171
9	46.99202...	COF	755	690	1183	4258	1598
10	49.99767...	CF2	1603	968	2079	4018	1734
11	55.05583...	C4H7	174899	227247	167685	54250	16976
12	68.99449...	CF3	3701	2215	3155	7060	2957
13	73.06246...	SiC3H9	164682	77503	91343	593360	13278
14	77.03882...	C6H5	28753	25549	31431	5695	25298
15	97.00973...	C2OF3	967	627	770	1500	805
16	104.0264...	C7H4O	9615	9422	14174	2304	11839
17	119.0102...	C2F5	1818	1486	2211	15497	2716

クラスター間の距離を計算するには，クラスター内のデータのバラツキを重視して評価する Ward 法が，分析データの分類には適していると考えられるので Ward 法を選択した．ただし，スペクトルデータのカウント数の絶対値は試料間でバラツキがあるので，試料ごとに規格化した値（最大値：1，最小値：０）を用いた．なお，COMPRO では群平均法も選択できるようになっている．また，データの Poisson Scaling 処理や標準化も可能である．

結果は図６に示すようにデンドログラムで表示される．縦軸は試料名，横軸はバラツキの平方根（小さい方がクラスター間の距離がより近い）である．ただし，見やすくするために途中を切断してある．

図６の縦の点線で示す距離で判定すると，試料は６種類に分類される．ただし，どの距離で判定するかは原則として解析者に依存する．

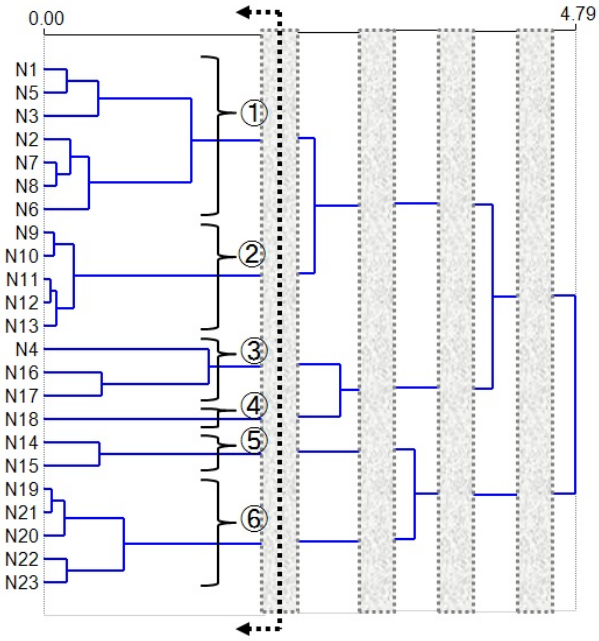

図６　TOF-SIMS データのデンドログラム表示．横軸は Ward 法で計算したクラスター間のバラツキの平方根．縦軸は試料名

４章で述べた SOM 法でも，階層的クラスター分析と同様にデンドログラムが表示できる．図７に SOM 法で求めた TOF-SIMS データのデンドログラムを示す．

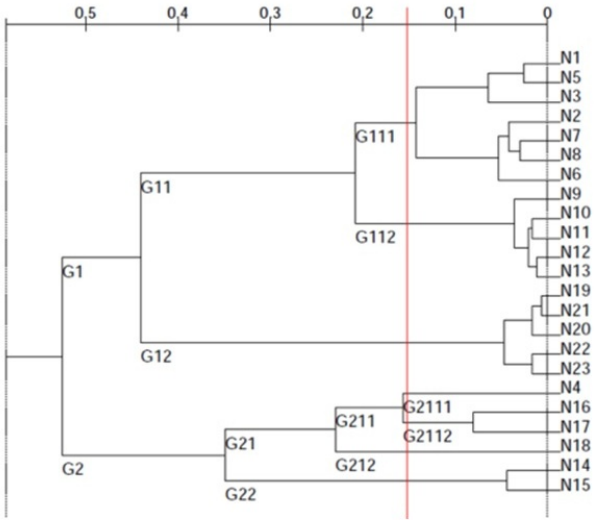

図７　SOM 法で求めたデンドログラム

階層的クラスター分析で求めた結果と SOM 法で得られた結果は良く一致している．

SOM 法の大きな特徴は，同種のクラスターを地図上の近接した領域に配置させることにより，分類分けを地図として見やすく表示できることである．中でも分類結果を球面上に表示する球面 SOM は図８に示すように，分類分けを直感的に見やすく表示できるため，クラスター分析として優れた方法である．また，球面 SOM は表示面を自由に回転できるため，全てのクラスター間の関連が容易に視覚的に理解できるようになっている．

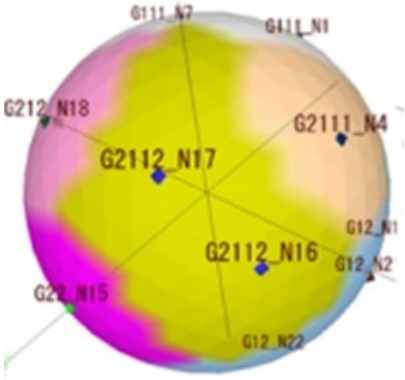

図８　球面 SOM によるクラスター分布の表示例

図６に示されているグループ②に属する N11 と，グループ④に属する N18 のスペクトルを図９に示す．図９から N18 は Si 系と F 系の質量ピークが観測され，N11 とはスペクトルの形状がかなり異なっており，異なるグループに分類される理由が理解できる．

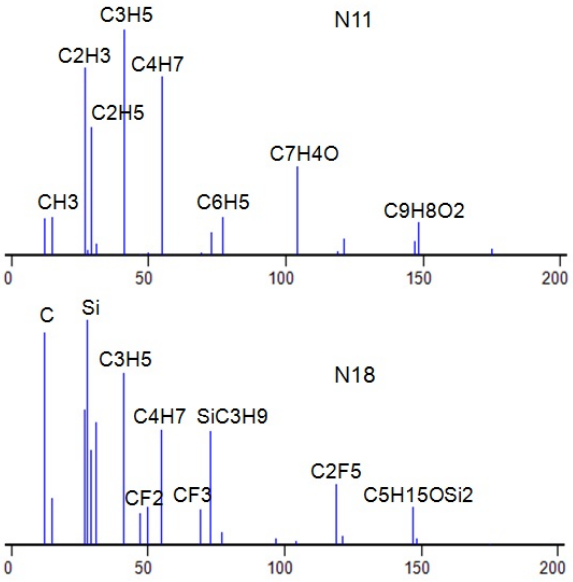

図９　TOF-SIMS スペクトルの代表例

7. 終わりに

今回はクラスター分析の中で，最も多用されている階層的クラスター分析について力点を置いて解説した．TOF-SIMS に代表される多量のデータを解析するには，クラスター分析は有用な手段である．

8. 参考図書

１）上田尚一，*クラスター分析*，朝倉書店，クラスター分析の全体像を説明している．

２）竹内光悦，酒折文武，*Excel で学ぶ理論と技術 多変量解析入門*，ソフトバンククリエイティブ，Excel を使って多変量解析を実習する．クラスター分析は多変量解析の一手法として紹介している．

３）T.コホネン，徳高平蔵，岸田悟，藤村喜久郎訳，*自己組織化マップ*，シュプリンガーフェアラーク東京，SOM 法の考案者であるコホネン氏が SOM の基礎を解説している．