(2) 最終判別関数

$$D = W'1X1 + W'2X2 + \cdots\cdots + W'nXn + W'n+1Xn+1$$

2．判別分析手法の分類"パラメトリックおよびノンパラメトリック"

判別関数の求め方には様々なアプローチがある。 これらのアプローチは前章で述べたようにパラメトリック、ノンパラメトリックの2種類に分類できる。 この分類は、判別関数を求める過程で解析母集団のパターン分布密度等のデータを用いるか否かが基準となる。 パラメトリック手法はこれらのデータを利用して判別関数を求めるものであり、ノンパラメトリック手法はこれらのデータを必要としない手法の総称である。 以下、順を追って説明する。

□ パラメトリック手法

使用データセットに関する様々な統計的パラメータ（パターン分布密度関数等）を利用して判別関数を求める手法の総称である。 この統計的パラメータの差異や取扱手法の差異に従って様々な判別分析手法が存在する。

□ ノンパラメトリック手法

この手法は統計的パラメータを用いずに判別関数を求める手法の総称である。 現在、このアプローチの一つとして、"学習"で判別関数を求めるアプローチがある。 先に述べた線型学習機械法は、このアプローチの代表的手法である。 "学習"とは、分類とウエイトベクトルの修正を交互に繰り返しつつ、最終判別関数を求める手法である。

学習は、完全に誤分類がなくなった（収束した）時点、或いはこれ以上誤分類パターンを少なくする事が不可能（収束不可能）な時点を学習完了点とする。

3．ノンパラメトリック手法による判別分析手法

□ パーセプトロンによる手法

生体における神経細胞（ニューロン）の働きをシミュレーションし、生体の得意とする画像認識や音声認識等の問題を解決しようとする学問から展開された。 この分野の学問はパターン認識の代表的アプローチとして展開されてきた。 現在、ニューラルネットワークと称されて展開されており、第2世代目に突入している。 先に述べたパーセプトロンは第1世代のニューラルネットワークである。 このアプローチの基本は生体における神経細胞のネットワーク構造を基本とする点であり、従って様々なネットワーク構造を持つ手法が展開されている。 実際の計算はＨｅｂｂの学習則に従った学習（フィードバックトレイニング）によりウエイトベクトル（ネットワーク強度）を変化させることで最終判別関数が求められる。

線型学習機械法は2クラス分類を基本とする解析手法であり、対象とするデータセットが2分割可能な時に極めて有効な手法である。 この手法の詳細については前章を参照されたい。

（参考文献）
J. T. TOU AND R. C. GONZALEZ, "PATTERN RECOGNITION PRINCIPLES", ADDISON-WESLEY, READING, MASS., 1974

□ シンプレックスアルゴリズムによる手法

最適化等の問題に利用されるシンプレックスアルゴリズムを用い、予め設定された応答関数を最適化することで判別関数を求める手法である。

この手法は対象とするデータセットが2分割不可能な時、極めて有効な手法である。詳細な説明（特にシンプレックスアルゴリズムについて）は2章4節を参照されたい。

（参考文献）
S. L. MORGAN AND S. N. DEMING, "SIMPLEX OPTIMIZATION OF ANALYTICAL CHEMICAL METHODS", ANAL. CHEM., 46, 1170 (1974)
G. L. RITTER, S. R. LOWRY, C. L. WILKINS, T. L. ISENHOUR, "SIMPLEX PATTERN RECOGNITION" ANAL. CHEM., 47, 1951 (1975)

□ 最小二乗アルゴリズムによる手法

最適化等の問題によく利用される最小二乗アルゴリズムを用いて判別関数を求める手法である。

この手法も前項のシンプレックス法と同様に、対象とするデータセットが2分割不可能である時に極めて有効な手法である。

最適化する時に利用される関数は解析目的により様々な関数が利用されるが、以下にその1例を示す。

$$Q = \sum_{i=1}^{n} [Y_i - F(S_i)]^2 \quad \text{———— ( )}$$

n ： データ（サンプル）数
Y_i ： 教師データである。 一つのクラスが＋1、もう一方のクラスは－1、に設定される
S_i ： i番目のパターンに対するWEIGHT VECTOR の内積である
$F(S_i)$ ： S_iのハイパーボリックTANGENTである

この式ではある判別関数を用いた分類で、誤分類が生じた時にQの値が大きくなり、分類が正しく、且つウエイトベクトルとパターンベクトルとの内積（Si）の値が大きい（パターンが識別平面から遠く離れている）程Qの値が小さくなるように設定されている。この関数の値が小さくなるように最小二乗アルゴリズムを用いて判別関数の修正が行われる。

（参考文献）

P. C. JURS AND T. L. ISENHOUR, "CHEMICAL APPLICATIONS OF PATTERN RECOGNITION" WILEY-INTERSCIENCE, NEW YORK, 1975

□ ＡＬＳ法（ADAPTIVE LEAST SQUARES METHOD)

この手法は森口らにより開発された手法である。 ここまでに述べた解析手法は、基本的には2クラス分類を基本とするものであるが、ＡＬＳ法は多クラス分類を前提として開発された。 このＡＬＳ法で用いられるクラス識別の為の判別関数もフィードバック学習により獲得される。

$$L = \overline{W} \cdot \overline{X}$$

L：判別得点
W：ウエイトベクトル
X：パターンベクトル

最初にWEIGHT VECTOR が以下の式を用いて初期化される

$$W = (X'X)^{-1}X'S$$

但し、Ｓは個々のクラスのメンバー数から計算されるFORCING FACTORである。
繰り返し計算の後にＬ値とＳ値とが比較される。
Ｌ≠Ｓの時は修正項がＳに加えられる。 修正項は以下の式にて表現される。

$$C = 0.1/(\alpha + d)^2 + \beta(\alpha + d)^2$$

但し、ｄはＬと実際のクラス間の境界値との差を示す。
上式のαとβの値とを変化させながら最適の判別関数を算出する。

（参考文献）

IKUO MORIGUCHI, KATSUICHIRO KOMATSU, AND YASUO MATSUSHITA, J. MED. CHEM., 23, 20 (1980

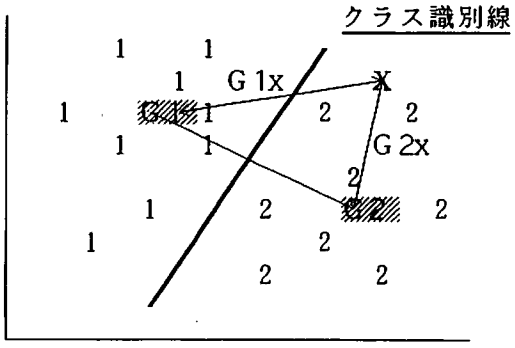
4．パラメトリック手法による判別分析手法

パラメトリック手法は多変量解析の基本となる手法である。 ここでは多変量解析の代表的判別分析手法についてのべる。 これらの手法は学習を行わず、統計的パラメータを利用する事で判別関数をもとめるものである。

□ グループ（クラス）の重心を利用した分類

パターンベクトルから容易にクラス毎の重心を求める事が出来る。 クラス未知パター

ンXのクラス決定は、①この未知パターンから各クラスの重心点への距離を求める、②未知パターンに最も近接している重心を決定する、③最近接重心が帰属するクラスに、未知パターンXのクラスを帰属させる、という手順を踏む。　従って、クラス識別の為の判別関数はクラス毎の重心を結ぶ直線の垂直2等分線として引かれる。



X：クラス未知パターン
G1：クラス1の重心
G2：クラス2の重心
G1x：重心G1とパターンXとの距離
G2x：重心G2とパターンXとの距離

図には2クラス分類について示してあるが、以下では多クラス分類を対象とした一般的なアプローチを簡単にのべる。

解析に当たり、クラス毎の重心を求める事が必要となる。　クラスc の重心Gc は以下の式により容易に求まる。

$$\overline{Gc} = \frac{1}{N_c} \sum_{k=1}^{N_c} \overline{Pck}$$

$$Gc(Xcj) = \frac{1}{N_c} \sum_{k=1}^{N_c} Pck(Xkj) \qquad (次元数；\ j= 1,2,\cdots\cdots,d)$$

ここで$\overline{Gc}$はクラスcの重心を示すベクトルであり、要素Xcjより構成される。　N_cはクラスcの総パターン数、$\overline{Pck}$はクラスcのパターンでk番目のパターンに関するパターンベクトルを示し、Xkjの要素から構成される。　Xkjはクラスcのパターン中のk番目のパターンのj番目の数値データ（記述子）を示す。

この重心とクラス未知パターンXとの距離Dcxを以下の式により求める。　尚、Xjはクラス未知パターンXの要素データである。

$$Dcx = \sqrt{\sum_{j=1}^{d} (X'j - Xj)^2}$$

このクラス毎の重心とパターンXとの距離を用い、最小距離を持つ重心と同じクラスにパターンXのクラスを帰属させる。

$$Class(X) = Min_{c=1}^{l} (Dcx)$$

ここでlはクラスの数を、$Min_{c=1}^{l}(Dcx)$は取り出されたDcxの値の内、最も小さい値を持つDcxを取り出す事をしめす。　またClass（X）は、取り出された値が属するクラスにパターンXのクラスを帰属させる事を示す。

□　BAYES線型判別分析

BAYESのとなえた分類手法であり、確立分布を基本とした分類手法である。

$$S = d_1(X) - d_2(X) = \ln p_1 - \ln p_2 + X' C^{-1}(m_1 - m_2) - \frac{1}{2} m_1' C^{-1} m_1 + \frac{1}{2} m_2' C^{-2} m_2$$

$S \geqq 0$　クラス1
$S < 0$　クラス2

□　BAYES非線型判別分析

$$d_k(X) = \ln p_k - \frac{1}{2} \ln |C_k| - \frac{1}{2} [(X - m_k)^T C_k^{-1} (X - m_k)]$$

但し、K＝1，2

p_k：クラスKに対する先見確率
C_k：クラスKに対する共分散マトリクス
m_k：クラスKに対する平均値

5．様々な判別分析手法の特徴について

前記ノンパラメトリックおよびパラメトリックな判別分析手法は、それぞれに特徴を持っている。　この特徴を知る事で、判別分析を正しく使う事が出来る。　以下にはこの点について簡単にのべる

ノンパラメトリック手法では最終的に得られる判別関数に手法間の差異が出てくる。この特徴により、特にデータセットが2分割可能時と不可能時とで判別関数の落ち込み点が若干異なってくる。

- 線型学習機械法は手法的な特徴（特に学習（エラーフィードバック）過程）の為、最終識別平面は各クラス間の接線方向に落ち着く傾向が強い。　特に最終時点で修正に用いられたパターンの位置に影響を受けやすい。
- 他のノンパラメトリック手法には線型学習機械法のような特徴はみられず、クラス間に存在するスペースの中央付近に落ち着く傾向がある。

この特徴から、データセットのパターン分布状態の差異により、適用する手法を変える事が必要であることがわかる。

①完全に分割可能で分類が重要な時は、線型学習機械法が適している。
②予測を重視する時は線型学習機械法以外のノンパラメトリック手法が望ましい。
③完全に分割不可能なデータセットに線型学習機械法は不適である。

データセットの2分割の可能性	解　析　手　法
線型2分割可能	線型学習機械法（パーセプトロン）
線型2分割不可能	最小二乗アルゴリズムによる判別分析 シンプレックスアルゴリズムによる判別分析 ＢＡＹＥＳ線型／非線型判別分析

6. クラスタリング

クラスタリングは対象となるパターンを、類似度または非類似度に関する基準に従ってパターンをグループ（クラスター）化する手法の総称である。 この手法は、クラスに関する教師データを必要としない教師無し学習に属する。

クラスタリングの目的は対象パターン群を似たもの同士でグループ化する事にある。従って、クラスタリング手法を用いた解析はクラスタリングの結果得られたクラスターの様子やクラスターを構成するパターンを検討して行われる。

解析はクラスタリングにより得られたクラスターの内容を吟味し、そのグループ分け（クラスター化）が妥当なものであるか、否かを判定する。 クラスターが非妥当な時、クラスタリングに用いた距離基準、融合手法、分割手法等を変える事で、妥当と思われる結果が出るものを探す事が必要となる。 クラスターが妥当な時、個々のクラスターを構成するパターンについて吟味し、これらのパターンが何故同じクラスターに分類されたかについて考察する。 これはクラスタリングに用いた距離基準が代表する情報の意味を読み取ることを意味する。

6．1． クラスタリング手法の種類

クラスタリング手法には様々な手法が存在する。 これらの手法はクラスタリングの実行で得られる結果が階層的（ＨＩＥＲＡＲＣＨＩＣＡＬ）なものか非階層的（ＮＯＮ－ＨＩＥＲＡＲＣＨＩＣＡＬ）なものかで２種類に分類される。

これらの２手法はパターンをクラスター化する時の手続きの差異により、さらに２つに分類される。 つまり、個々のパターンレベルから出発し、少しづつパターンをグループ化して最終的に一つの大きなクラスターとする、小から大へのアプローチ。 もうひとつは、全体を一つのクラスターとして出発し、このクラスターを小さく分割し、全パターンをクラスター化して行く、大から小へのアプローチの２つである。

□ クラスタリング手法の分類（クラスター構造からの分類）

クラスタリング手法は大きく分けて、(1) 階層的クラスタリング、及び(2) 非階層的クラスタリングとに分類される。

(1) 階層的クラスタリング

一般的にクラスタリングと呼ばれているものは、この階層的クラスタリングを指している。 このクラスタリングの結果は、デンドログラム（ＤＥＮＤＲＯＧＲＡＭ）として表示されるのが特徴である。 このデンドログラムが階層的構造になっており、階層的クラスタリングと呼ばれる。

(2) 非階層的クラスタリング

非階層的クラスタリングでは解析結果がデンドログラムの形で出力されず、単なるクラスターの構成が示されるだけである。 従って、必ずしもクラスター間の上下関係が明確になるとは限らない。

□ クラスタリング手法の分類（融合（Ｆｕｓｉｏｎ）手続きからの分類）

クラスターを形成する時の手続きを、一般的に融合（Ｆｕｓｉｏｎ）と呼ぶ。 クラスタリングで最も大切な手続きがクラスターを形成する為の融合である。 この融合の仕方にも様々なアプローチが提案されているが、出発時のクラスターの状態により大きく２種類に分類される。

全パターンを一つのクラスターとしたものを出発クラスターとし、このクラスターを順次分割する事でクラスター群を形成するアプローチ（大→小への展開）。 もうひとつのアプローチは、個々のパターンを一つのクラスター（即ち、出発クラスターはパターンの数だけ存在する）とし、これらのクラスターを順次統合し、最終的に全体を一個のクラスターとするアプローチ（小→大への展開）とがある。

出発クラスターが全パターンを代表する１個のクラスターである（大→小）のか、個々のパターンを一個のクラスターとみなし、パターン数だけクラスターが存在する状態で出発する（小→大）かで２種類に分類される。

(a) 全体を一つのグループとして出発し、細分化してゆく方法 （大→→→小）
（ＤＩＶＩＳＩＶＥ ＭＥＴＨＯＤ）

(b) 個々のパターンから出発して大きなグループにまとめる方法（小→→→大）
（ＡＧＧＬＯＭＥＲＡＴＩＶＥ ＭＥＴＨＯＤ）

従って、クラスタリングはクラスターの形態と、クラスターを作成する過程の差異により４種類のアプローチに分類される。

階層的クラスタリング ― DIVISIVE METHOD
　　　　　　　　　　 ― AGGLOMERATIVE METHOD
非階層的クラスタリング ― DIVISIVE METHOD
　　　　　　　　　　 ― AGGLOMERATIVE METHOD

６．２．　階層的クラスタリング

□　階層的クラスタリングで用いられる融合手法

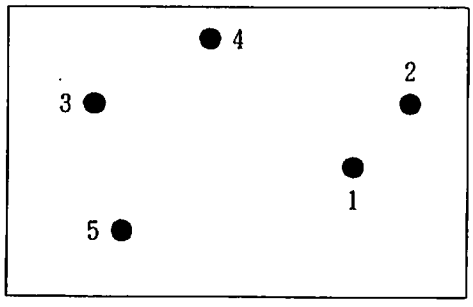
デンドログラムを出力図とするクラスタリングでは、小→大のＡＧＧＲＯＭＥＲＡＴＩＶＥなアプローチが一般的である。　この時、クラスターを形成する為の融合手法として様々なものが提案されている。　以下に融合手法の主なものを示す。

① 最近隣法 (NEAREST NEIGHBOR METHOD)
② 最遠隣法 (FURTHEST NEIGHBOR METHOD)
③ 群平均法 (GROUP-AVERAGE METHOD)
④ 重心法　(CENTROID METHOD)
⑤ メジアン法 (MEDIAN METHOD)
⑥ ワード法 (WARD METHOD)
⑦ 可変法　(FLEXIBLE METHOD)

クラスタリングではここに示す様々な融合手法が利用されているが、用いた融合手法により最終デンドログラムの形が大きく変化する事がある。　特に①と②とでは変化が大きいので注意が必要である。　また、後にのべるがこの融合手法の差異によりパターン空間の変形が起こり、実体を反映しなくなる可能性が強くなる。　このような特徴をよく理解した上で、個々の融合手法を利用する事が必要である。

□　最近隣法及び重心法によるクラスタリング

ここでは、先に示した幾つかの融合手法の内、(1)の最近隣法と(4)の重心法とで実際にクラスタリングを行ない、その結果をデンドログラムとして表示してみる。　解析に用いるデータは２次元データで、図１に示される１～５までの５個のパターンを用いる。



パターンID	1	2	3	4	5
1	□				
2	1	□			
3	8	9	□		
4	4	5	2	□	
5	6	10	3	7	□

図１．５個のパターンの位置関係、及び距離マトリクス（距離の小さい順に番号付け）

最近隣法によるクラスタリング

最近隣法では、最初にパターン間の相互距離を全パターンについて求め、そのうち各パターン間で最も近接したパターン同士を結ぶ。　図２ではこの最近隣パターンを結んだ線が描かれている。　この距離は$D_{12} < D_{34} < D_{53} < D_{41}$という関係にあるので、次に各パターンの融合手続きは求められた距離のうちで、距離の短いものから順に２つずつ融合されていくことになる。

図ではパターン１と２が最初に融合される。　次に短い距離関係にあるパターン３と４が融合される。　次に短い距離はD_{53}であるが、パターン３は既にパターン４と融合されている為、パターン５はパターン３と４のクラスターにつながる事になる。　D_{41}についても同様に手続きを行い、パターン３、４、５のクラスターとパターン１、２のクラスターを結び付けて全体のクラスターが完成する。

図２の右の図がデンドログラムである。　この図の詳細については後に説明する。

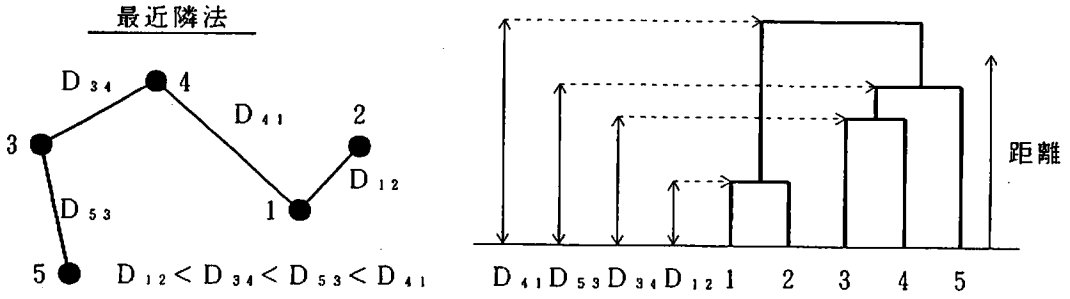


図２．　最近隣法によるクラスタリング手続きとデンドログラム

最近隣法において、クラスターの代表は隣接するパターンか隣接クラスターに最も近接しているパターンを代表としている。　従って、この例ではクラスター（１２）の代表点は、そのクラスターに最近接するパターン４との関係よりパターン１が代表することになる。　又、クラスター（３４）では、パターン５との関係ではパターン３が代表し、クラスター（１２）ととの関係では、それぞれパターン１とパターン４とが代表する。

従って、個々のクラスターは互いに端と端とで結合されることになり、パターン空間におけるクラスター間の相互距離関係を的確に表現する事は困難になる。

このクラスタリングの結果はデンドログラムとして表示されるが、最近隣法によるクラスター化では（（１２）（（３４）５））にクラスター化される事がわかる。

重心法によるクラスタリング

重心法での融合手続きの特徴は、クラスターを代表する点を融合された２個のパターンの重心点をもって代表（図３中○）することである。　図１の問題について実際の手続きを簡単に述べる。

最初に総てのパターン間の距離を求め、距離が最も短いもの（D_1：パターン１及び２）を選び出す。　この２パターンを最初のクラスターとした後、この２パターンの重心点をこのクラスターを代表する新たなパターンP_{12}とする。　取り出された２パターンの残りのパターン（３　４　５）と、新たなパターンP_{12}とで再びパターン間の距離を求めて最短距離（D_2：パターン３及び４）を形成する２パターンを取り出し、２番目のクラスター（３４）とする。　この２パターンの重心P_{34}を新たなパターンとし再び同じ操作を繰り返す。　この手続きをすべてのパターンに関して行い、すべてのパターンがクラスター化された時を終点とする。

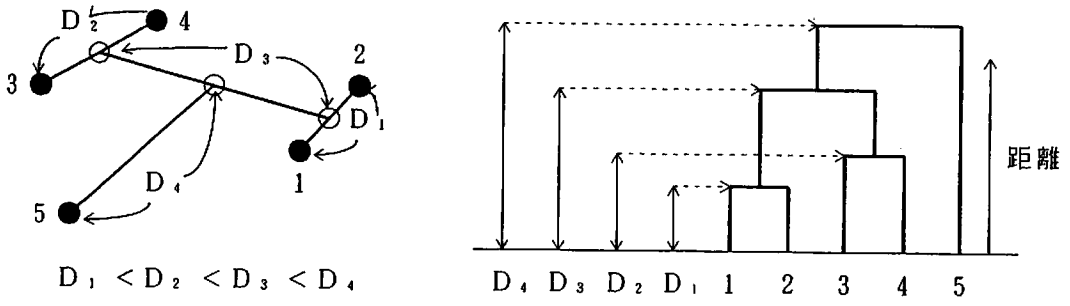


図３．重心法によるクラスタリング手続きとデンドログラム

図３のデンドログラムから、重心法では５つのパターンが（（（１２）（３４））５）のようにクラスター化されている事がわかる。

図２及び図３の結果を比較すると、全く同じパターンデータと距離基準を用いたとしても、最終的に得られるデンドログラムは融合手法の差に従って変化する事がわかる。従って、クラスタリングを行う時は複数の融合手法を試みる事が情報の取りこぼしを防ぐ意味からも重要である。

その他の融合手法については他の専門書を参考にされたい。

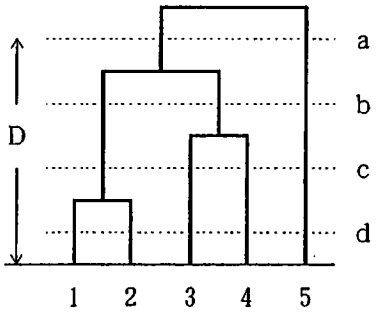
□　デンドログラムの利用法

デンドログラムはクラスタリングの結果を図として出力するもので、この図を用いた手法が階層的クラスタリングとして分類される。　この図はクラスター間の階層関係やクラスターを構成するパターンをまとめて一見する事が可能なので、クラスタリングの結果を解析する時は大変便利なものとなる。

デンドログラムの解読について

デンドログラムはクラスター間の階層関係が一見できるようになっている。　X軸はパターンをリストアップしているだけだが、Y軸の高さはパターン間及びクラスター間の距離の大きさに比例している。　従って、単にクラスターの情報がわかるだけでなく、パターン同士及びクラスター同士の距離関係に関する情報も同時に入手可能である。

このデンドログラムでは互いに結合されているパターンが一つのクラスターを形成すると考える。　また、Y軸上のある高さ（距離）の所でX軸に平行に線を引いた時、この線から下の部分で結合されているクラスターが、Y軸で定められた値以下の距離でつながるクラスターを示している。



①　Y軸のDはパターン／クラスター間の距離を示す。
②　デンドログラムをY軸上のある高さで切断した時、その切断線以下がクラスター化の対象となる。
例）切断線がa、b、c、dの時、クラスターは以下のように決定される。
a：（（1 2 3 4）、（5））
b：（（1 2）、（3 4）、（5））
c：（（1 2）、（3）、（4）、（5））
d：（（1）、（2）、（3）、（4）、（5））

図4．デンドログラムの特徴及び横軸分割によるクラスターの差異について

□　パターン空間の歪み（融合手法の差異がパターン空間に及ぼす影響）

多次元空間上のパターン融合過程で、パターン間の相対的距離関係が変化し、クラスタリングの開始時期と終了時期とでパターン空間が歪んでくる（拡大／縮小）事がある。クラスタリングを行う時は、この特徴を理解しておく事が必要である。

このパターン空間の歪みは、パターンを融合する時、新たに融合されたクラスターをどのような点で代表させ、次の融合に備えるかという手続きに起因する問題である。　この代表点の捕らえ方の差異がそのまま様々な融合手法の差異となる。

パターン空間の拡大（拡散）／縮小（濃縮）という問題は、クラスターのように広がりを持つものをただ一点で代表するという点に起因する。　この問題を根本的に解決することは不可能であるが、この種の問題を生じにくくする融合手法としては群平均法やワード法がある。

融合過程におけるパターン空間の拡大縮小事例

一般的に最近隣法による融合ではパターン空間の縮小が起こり、最遠隣法による融合ではパターン空間の拡大が起こる。　図5を用いてこの関係について論じる。

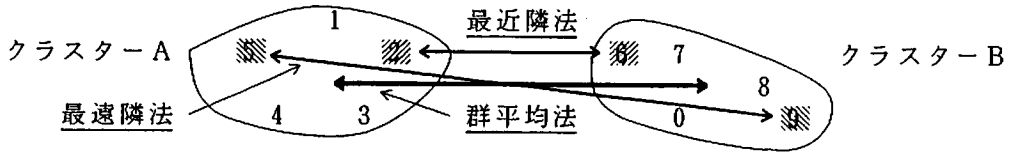


図5．融合手法の差異によるパターン空間の拡大／縮小

図5のようにクラスターA、Bが融合される時（クラスター（1 2 3 4 5）及び（6 7 8 9 0））、最近隣法の融合過程ではA、B、2つのクラスターを代表する点として最も近いパターン（この時はパターン▨とパターン▨）が取られる。　この結果、本来広がりを持つクラスターAとBは一点に凝縮され、結果としてパターン空間の縮小が起こったのと同じ結果となる。

一方、最遠隣法では最も離れたパターンをクラスターの代表点とする為、クラスターAとBはそれぞれパターン▨と▨とで代表される事になる。　従って、実体よりもパターン間の距離が拡大され、パターン空間の拡大が起こる。　このようにクラスタリングではパターンやクラスターの融合手法により、パターン空間の拡大／縮小が起こる。

このようなパターン空間の変形が問題となる時には群平均法とかワード法を用いてクラスタリングを行うと、パターン空間の歪みを最小にすることができる。

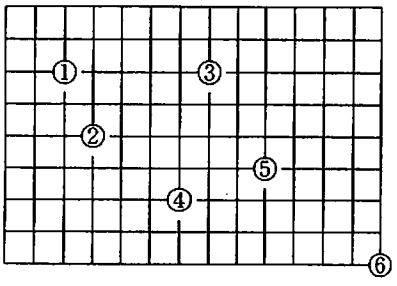
□　類似度及び非類似度を表す距離基準について

クラスタリングを行う時の最も重要な要素としては、先に述べたパターンの融合手法が考えられる。　この他、実際にクラスタリングを適用する時に問題になる事として、類似度及び非類似度を表す為の距離基準が重要な問題となる。　対象とするデータの質により距離基準を変化させないと、必ずしも実体に即した解が得られるとは限らない。　特にバイナリデータを用いる時と連続変数を用いる時とでは距離基準の差異が結果に大きく影響を与えることも考えられる。

実際の解析に用いられる距離基準の詳細については第　章を参照されたい。

ＱＵＩＺ

以下に示されたパターンを①最近隣法　及び　②最遠隣法でクラスター化せよ。
結果はデンドログラムとして表示せよ。



6．3．　非階層的クラスタリング

非階層的クラスタリングは、階層的クラスタリングの特徴であるデンドログラムをもちいないでパターンのグループ化を行う手法の総称である。　この手法として幾つか提唱されているが、ここでは比較的頻繁に利用されるMinimal Spanning Tree 、ＩＳＯＤＡＴＡ手法及びＣ－ＭＥＡＮＳ法について解説する。　さらに、この非階層的クラスタリングを基本とし、最近展開されているファジイを導入したクラスタリングや著者が提唱している「超ボリューム概念」に従ったクラスタリングについても簡単に言及する。

□　Minimal Spanning Tree

この手法は全体を一つのクラスターとし、分割を繰り返す事でより小さなクラスターを形成してゆくものでＤＩＶＩＳＩＶＥ手法に分類される。

このクラスタリングでは最初に全パターンを線で結ぶ事から作業が開始される。　パターン空間上の全てのパターンを線で結ぶ作業時、その結合距離の総和が最も小さな値をとるように結ばれる。　この時、結合される線は分岐点を持ち、末端をもっても構わないが、線が互いに結合して円を形成する事は禁止される。　このようにして得られた樹状図をMinimal Spanning Tree という。

パターンのクラスター化は、予め作成された樹状図を用いて行われる。　樹状図上のある一つの結合線を切断した時、切断された先のデータが繋がっている一群のパターンを一つのクラスターとする。　従って、一回の切断で２つのクラスターが新たに形成される。この過程を繰り返し行うことで、クラスタリングが実行される。

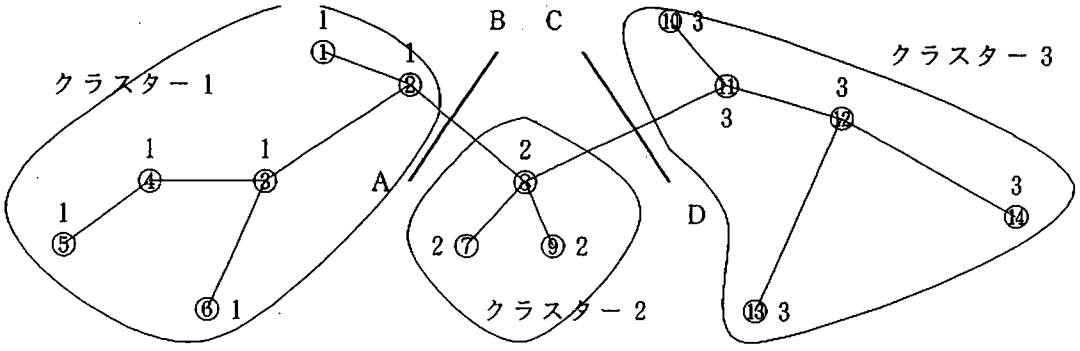


図１．　Minimal Spanning Tree によるパターンの分類

図１に示されたように樹状図が作成された時、ＡＢ及びＣＤとで結合を切断すると、それぞれのパターンはクラス１、２、３とのクラスターに分類される。

クラスター１：パターン①　～　⑥　　ＡＢにて分割
クラスター２：パターン⑦　～　⑨　　ＡＢ及びＣＤにて分割
クラスター３：パターン⑩　～　⑭　　ＣＤにて分割

Minimal Spanning Tree 図の作成

非階層型クラスタリングはデンドログラムのような表示手段を持たない。　Minimal Spanning Tree 図はクラスタリングを行う為の出発図であり、クラスタリングの結果を表すデンドログラムとは内容が異なる。

このMinimal Spanning Tree 図の作成は、他の手法により得られた結果を用いて散布図を作成し、この散布図に樹状図の情報を被せる形式をとる。　従って、作図に利用される散布図はすべてのパターンが広い部分に分散している事が望ましい。　従ってMinimal Spanning Tree 図に利用される散布図は主成分分析で得られた第１、２主成分をＸ、Ｙ軸とした図を利用する事が多い。

このように表示用の基本図として主成分図を用いる時、図上におけるパターン間の距離関係はMinimal Spanning Tree で求められる距離関係とは何の関係も無いので注意が必要である。

□　ＩＳＯＤＡＴＡ手法

ＩＳＯＤＡＴＡ手法概念

この手法も全体を一つのグループとして出発し、細分化してゆく方法　（大→→→小）（ＤＩＶＩＳＩＶＥ　ＭＥＴＨＯＤ）に属する手法である。

この手法の基本は対象となるパターンを２分割し、この操作を繰り返し、総てのパターンについて分割が達成された時点で分割を停止するものである。　この手法に対する具体的な手続きは以下のようになる。

①　先ず全パターンを適当に２分割する。
②　２分割されたそれぞれのクラスターについて重心を求める。
③　この２つの重心点を結ぶ線分の垂直２等分面を形成する。
④　この垂直２等分面により全データは新たに再分割される。
⑤　再び②、③、④を実行する。
⑥　新たに形成されるクラスターの成分に変化がなくなる迄繰り返し、最終的に得られたクラスターを２分割（２クラスター）の結果とする。
⑦　２分割された個々のクラスターについて、再び①から⑥迄の手順を繰り返す。
⑧　総てのパターンについてクラスター化できた時点でストップする。

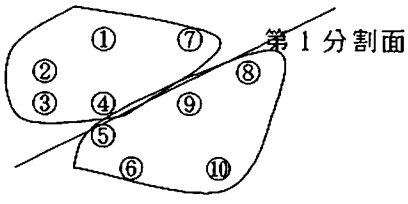
ＩＳＯＤＡＴＡの結果は個々のクラスターを形成するメンバーのリストが出るだけである。　従って、クラスター間の相対関係（例えばクラスター１とクラスター５が近い関係にあり、クラスター１とクラスター２はより遠い関係にある等）に関する情報は入手出来ないので注意が必要である。

前記アルゴリズムからわかるように、この手法ではパターン空間の拡大／縮小は生じない。　また、あらかじめある一定のクラスター数を決めておけば、指定された数のクラスターを形成させた段階で実行を停止する事も可能である。

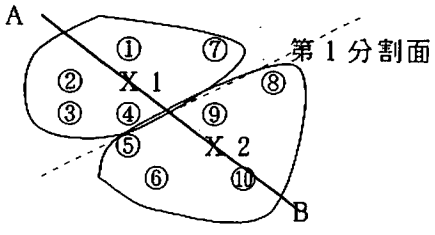
ＩＳＯＤＡＴＡクラスタリング過程シミュレーション

(1)　第１ステップ
適当に２分割面を引き、２つのクラスターを形成する。

クラスター１：　①　②　③　④　⑦
クラスター２：　⑤　⑥　⑧　⑨　⑩

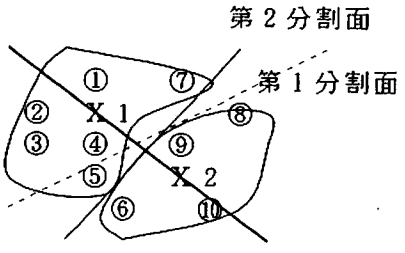


(2)　第２ステップ
新たに形成された２個のクラスターに関し、それぞれの重心点Ｘ１及びＸ２を求める。
重心点Ｘ１及びＸ２を結ぶ線分をＡＢとする。



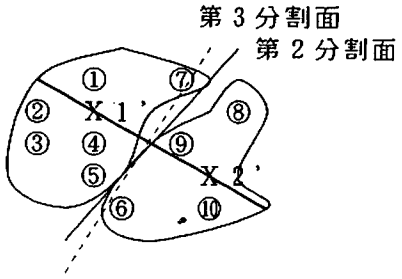
(3)　第３ステップ
重心点を結ぶ線分ＡＢの垂直２等分線を形成し、第２分割面とする。　この分割面で新たに形成されるクラスターを、先に形成されたクラスターと比較する。

クラスター１：　①　②　③　④　⑤　⑦
クラスター２：　⑥　⑧　⑨　⑩



(1)と(3)のクラスターを比較すると、この場合は明らかに変化している。　従って、クラスター決定の過程を継続する。

(4) 第４ステップ
新たに形成された２クラスターに関し、(2)と(3)を繰り返す。 この結果得られた新クラスターの要素をチェックし、変化の有無を判定する。



クラスター１： ① ② ③ ④ ⑤ ⑦
クラスター２： ⑥ ⑧ ⑨ ⑩

(5) 第５ステップ
第４ステップで新たに形成されたクラスター１と２が、第３ステップで形成されたクラスター１、２と全く同じであるので、第１回目の分割を完了する。

(6) 第６ステップ
以下、既にクラスターに分割された２つのクラスター１及び２について、再び(2)～(4)の操作を繰り返す。
予め指定されたクラスター数になるか、又は総てのパターンについてクラスター化が出来た時点を終了点とする。

□ Ｃ－ＭＥＡＮＳ法

その他の非階層的クラスタリングとしてＣ－ＭＥＡＮＳ法がある。 この手法はＣ－分割（Ｃ－ ＰＡＲＴＩＴＩＯＮ）の概念に従った手法である。
この手法では予め設定されたクラスターの数（Ｃ）を目指してクラスター化が試みられる手法であり、先に述べたＡＧＧＲＯＭＥＲＡＴＩＶＥやＤＩＶＩＳＩＶＥ手法のどちらにも属さない第３のアプローチを取る。

「Ｃ分割」について

Ｃ－ＭＥＡＮＳ法はＣ－分割を基本としている。 いま任意の整数値Ｃ（解析目標のクラスター数となる）とパターン数ＮとのＣ×Ｎのマトリクスを考える時、このマトリクスの要素値M_{ik}が以下の３条件を満たす時、これをＣ－分割と呼ぶ。

1. $M_{ik} \in \{0, 1\}$ $\quad 1 \leqq i \leqq C,\ 1 \leqq k \leqq N$
2. $\sum_{i=1}^{C} M_{ik} = 1$ $\quad 1 \leqq k \leqq N$
3. $0 < \sum_{k=1}^{N} M_{ik} < N$ $\quad 1 \leqq i \leqq C$

ここでｉはクラスターを、ｋは個々のパターンを表している。 条件１は、要素値M_{ik}が０と１の値から構成されている事を示す。 条件２は個々のパターンは一つのクラスターにのみアサインされ、複数のクラスターにアサインされる事は無い事を示す。 条件３は全パターンが一つのクラスターとなるものや、パターンが存在しない空のクラスターは認められない事を意味する。

Ｃ－ＭＥＡＮＳ法のシミュレーション

Ｃ－ＭＥＡＮＳ法はＣ－分割の条件下、類似度として距離尺度を用いてパターンのクラスターへの帰属を決定する手法である。 以下にＣ－ＭＥＡＮＳ法を行う為のアルゴリズムを簡単に示す。

① 最初にクラスター数Ｃを決める。 $2 \leqq C < N$
② 初期マトリクスU^{j}を適当に決める。 $j = 1$
③ 個々のクラスターの重心点G_iを求める。

④

$$M_{ik}^{j+1} = \begin{cases} 1: & D_{ik}^{j} = \min_{i} D_{ik}^{j} \quad 1 \leqq i,\ 1 < C \\ 0: & \text{その他} \end{cases}$$

但し、D_{ik}^{j} はマトリクスU^{j}の要素値を求める為の計算で、K番目のパターンとi番目のクラスターの重心G_iとの距離を示す。

⑤ 新たに求めたM_{ik}^{j+1}により、U^{j+1}を求める。 このU^{j+1}とU^{j}との変化値vを求め、この値が予め設定された収束判定値εよりも小さい時は終了し、大きい時はj＝j＋1として手順3にもどる。

6．4． その他のクラスタリング手法

先に述べたクラスタリング手法の他に、従来とは異なる基礎理論や概念に基づいたアプローチが試みられている。 これらの手法は従来手法とくらべた時、幾つかの特徴を持つ事が多い。 ここではこのようなクラスタリング手法として、ファジイ理論に基づいた「ファジイクラスタリング」及び超ボリューム概念に基づいた「超球クラスタリング」について簡単に概要を述べる。 それぞれの詳細については 章の超ボリューム概念及び 章のファジイ理論の章を参照されたい。

6．4．1 ファジイ理論に基づいたクラスタリング

クラスタリング手法にファジイ理論を導入する事で、従来のクラスタリングとは異なったアプローチをする。

ファジイ理論の章で詳しく述べるが、ファジイ理論導入の基本は、1対1対応の絶対的な世界から、1対多対応の自由度を持たせた世界を受け入れる事にある。 従来のクラスタリングでは個々のパターンは常に一つのクラスターに帰属されており、複数のクラスターに帰属される事はあり得なかった。 ファジイクラスタリングでは個々のパターンが複数のクラスターに帰属され、このような状態の方を自然とする手法である。

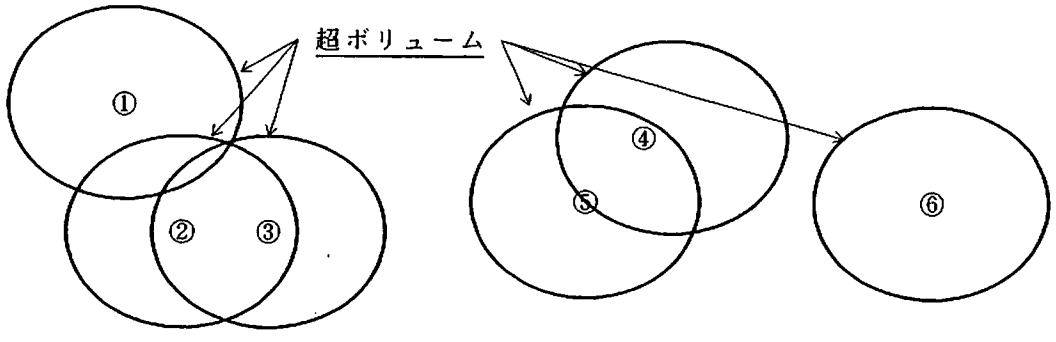
手法の詳細についてはファジイ理論の章にてのべる。

□ 超ボリューム概念

超ボリューム概念はパターンにボリュームを持たせるという考え方に基本をおく新たなパターン認識の概念である。 従来のパターン認識手法の総ては、パターンを単なる点として捕らえており、この超ボリューム概念と比べて大きな差がある。 このように、パターンに対する基本概念を変化させることで、従来手法では実現が困難であった問題を自然な形で解析手法に取り入れる事が可能となる。

□ 超ボリューム概念に基づいた超球クラスタリング

超ボリューム概念をクラスタリングに適用することを考える。 超ボリューム概念に基づいたクラスタリングでは、各パターンを代表する超球の重なりを基本としてクラスターを形成する。



クラスターA（①②③） クラスターB（④⑤） クラスターC（⑥）

図2．超球クラスタリングによる結果

超球クラスタリングでは「互いに超球が重なっているパターンは一つのクラスターとす

る」という極めて単純なルールに従ってクラスタリングが行われる。　従って、ｉ番目のパターンが帰属すべきクラスターは以下の式を満たす事が必要である。

$$Pil = Pkl \quad if \quad 2R \geqq \underset{k=1}{\overset{N}{Dist}}(Pi, Pk) \quad i \neq k$$

ここでＰｉｌはｉ番目のパターンがクラスターｌに帰属している事を示し、従ってその値はｋ番目のパターンが帰属しているクラスターｌの値と同じである。　Ｒは超球の半径を示し、Ｄｉｓｔ(Ｐｉ,Ｐｋ)はパターンｉとパターンｋの距離である。　従って、ｋは自分以外のパターンとの比較が必要で、１～Ｎの値を取り、ｉ≠ｋである。　この条件が満たされた時、パターンｉとｋの超球は互いにかさなりあっていることになる。

実際のクラスタリングでは、超球の半径Ｒの値を任意に設定し、その値を徐々に変化させる事で行われる。　Ｒが大きく、全パターンが一つのクラスターとなっている時は超球Ｒの半径は徐々に小さくされる。　一方、Ｒが小さく、パターン数と同じだけクラスターが存在する時は、半径Ｒを大きくする事でクラスタリングがおこなわれる。

またアルゴリズムでもわかるように、超球クラスタリングの結果は最近隣法で融合を行ったクラスタリングと同じ結果となる。

超球クラスタリングの特徴

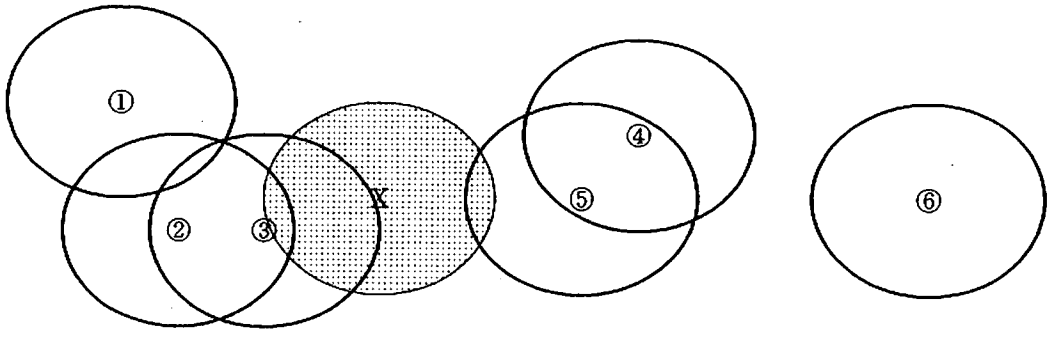
超球クラスタリングでは出発時の超球の大きさにより、ＡＧＧＲＯＭＥＲＡＴＩＶＥ及びＤＩＶＩＳＩＶＥ手法のいずれをとる事も可能である。　全パターンが一つのクラスターになるような大きなサイズを持つ超球から出発し、徐々に超球のサイズを小さくする事でクラスターを新たに形成するアプローチをとるならばＤＩＶＩＳＩＶＥ手法となる。

一方、超球間に重なりの無い小さな超球から出発し、徐々に超球サイズを大きくする事でクラスターを形成するならばＡＧＧＲＯＭＥＲＡＴＩＶＥ手法となる。

超球クラスタリングのアルゴリズムは簡単であり、共通の基本ルーチンを作成するだけで、ＤＩＶＩＳＩＶＥ及びＡＧＧＲＯＭＥＲＡＴＩＶＥ両手法に適用可能である。

ファジイ理論導入による「ファジイ超球クラスタリング」

超球自体に密度勾配の概念を導入する事でファジイ理論を簡単に導入する事が出来る。



クラスターＡ（①②③）　　クラスターＢ（④⑤）　　クラスターＣ（⑥）

クラスターＤ（①②③Ｘ④⑤）

図３．ファジイ超球クラスタリングによる結果

先の超球クラスタリングの結果に新たなパターンＸが導入されたパターン空間を考える。

この時、従来手法によるならばパターンＸの為にクラスターＡとクラスターＢはつながり、新たなクラスターＤとなる。　従ってこの時、パターンＸはクラスターＤに帰属する事になる。

この時ファジイ理論を導入し、ファジイ超球クラスタリングを行うならば、パターンＸは唯一個のクラスターＤに帰属されるだけでなく、複数のクラスターに帰属される事となる。　超球の重なり程度から推算するならばこの場合、パターンＸの帰属度はクラスターＤ＞クラスターＡ＞クラスターＢの順となる。　このように、超球クラスターではファジイ理論の導入が簡単に出来る。

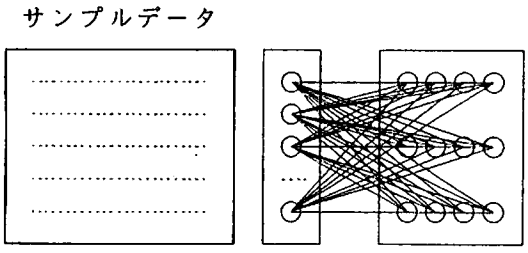
＊このファジイ超球クラスタリングに用いられる、超球内部に密度勾配が存在する超球（傾斜超球と呼ぶ）については、第　章で詳しくのべる。

６．４．２　ニューラルネットワークによるクラスタリング

ニューラルネットワークを用いてもクラスタリングを行うことが可能である。　このクラスタリングはニューラルネットワークのなかの自己学習型のアプローチを取る手法で実行される。

□コホーネンタイプニューラルネットワーク

最終的なネットワークにおいて、似たサンプル（同一クラスター）は同じユニットが発火することになる。　すなわち、このニューラルネットワークをもちいることでサンプルはユニットの数を最大とするクラスターへと分類が行われる。



学習過程：

① 一つのサンプルで、出力ユニットで最も値が小さいか、最大（問題により異なる）のものを探す。
② 探し出されたユニットに結合するウエイトベクトルを修正する。　この修正は出力値が最少または最大になりやすいように行われる。
③ 学習完了は、予め定められた学習回数を目安とする。
④ 同じユニットが発火するものは同じクラスターに帰属させる。

5. 表示手法 (DISPLAY METHOD)

5．1． 表示手法概論

表示手法としては様々なものが存在する。 先に述べた主成分分析も表示手法の一種である。

表示手法とは、人間が認識出来ない多次元情報をさまざまな様式で人間が認識しうる表示形式にして表示する手法を意味する。 この表示手法としては個々のパターンに関する多次元情報をパターン毎に表示する手法、および多次元（４≦N）空間におけるパターン同士の相互位置関係を保ちつつ、人間が認識可能な２、３次元空間に表示しなおす手法との２つがある。

この２つの手法のうち、後者の手法はさらに線型及び非線型の２つに分類する事ができる。 この時、線型手法によるものは投影（ＰＲＯＪＥＣＴＩＯＮ）と呼び、非線型手法によるものはマッピング（ＭＡＰＰＩＮＧ）と呼ばれる。

① 個々のパターンに関する多次元情報の表示
② 多次元空間中におけるパターン間の位置関係の２、３次元空間への投影
　１．線型投影 (LINEAR PROJECTION)
　２．非線型写像 (NON-LINEAR MAPPING)

ここでは①と②について順に説明する。

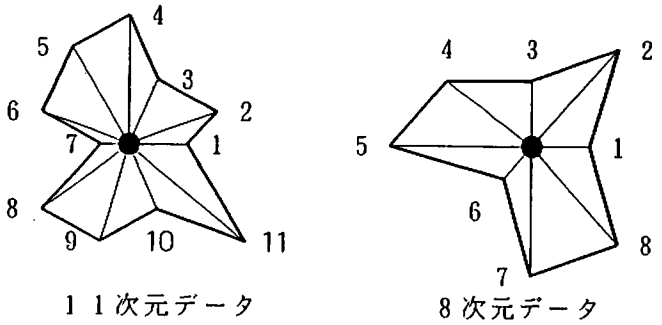
5．2． 個々のパターンに関する多次元情報の表示

個々のパターンに関する多次元情報を表示する手法として様々なアプローチがある。これらの手法は主として統計等の分野で２／３次元グラフ表示の応用として展開されてきた。 これらの手法にはレーダーチャート、ツリーグラフ、星座図等様々なものがある。また、最近展開されてきた手法としてチャーノフ（ＣＨＥＲＮＯＦＦ）の顔型グラフがあり、様々な分野で利用されはじめている。 以下ではこれらの手法のうち、代表的な手法について簡単にのべる。

5．2．1 パターン表示手法

□ レーダーチャート（ＲＡＤＥＲ ＣＨＡＲＴ）／くもの巣グラフ

レーダーチャートは別名“蜘蛛の巣チャート”とも呼ばれる。 気軽に多次元データを表現する事ができる為、頻繁に使われている手法である。 このチャートを用いる時は、レーダーのスポーク部分は総て同じ基準にもとづいてスケーリングされているか、スケールの基準が明確になっている事が必要である。 さもなければ、しばしば情報の読みちがいを起こすという点に注意しなければならない。



１１次元データ　　８次元データ

図５－１ １１次元及び８次元データを持つレーダーチャート

利用形態は幾つもあるが、ただ一個だけのレーダーチャートを用いる時と複数のレーダーチャートを用いる時とがある。

一個のレーダーチャートを用いる時の解析目的は、表示に用いた数値データの比較を行う事にある。 複数のレーダーチャートを用いる時は、複数のパターンの比較が目的となる。 この時当然の事ではあるが、作図時に同一次元は同じ場所に表示されるようにしなければ正しい比較は行われない。