岡本安晴　2002.1、2004.7

クラスター分析

N 個の対象とそれらの間の距離が与えられているとき、階層的クラスター分析を次の方法で行う。

（1）個々の対象、x_i、$i=1,\cdots,N$、1つ1つからなるクラスター、$C_i=\{x_i\}$、を作成する。$N_C=N$（現在のクラスターの数）とおく。

（2）N_C 個のクラスターのなかで、お互いの距離が最小のもの、C_q と C_r とする、を選ぶ。クラスター間の距離 $d(C_q,C_r)$ の算出方法は、この手続きの説明の後で解説する。クラスター C_q と C_r を1つのクラスター $C_s=C_q\cup C_r$ にまとめ（merge）、$N_C\leftarrow N_C-1$ とする。

（3）$N_C>1$ ならば、（2）に戻る。
$N_C=1$ ならば（4）にすすむ。

（4）（2）でのクラスター生成の過程を樹形図に表す。（2）でのクラスター生成の過程は、各クラスターをオブジェクトとして表しておき、C_s と C_q および C_r の間の親子関係をポインタを用いて表しておく。

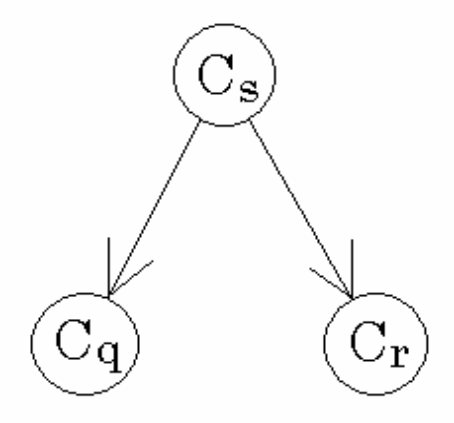


図1　クラスターの親子関係

クラスター間の距離の算出方法として、プログラム PClusterCWDist.exe では次のものが用いられている。

（１）重心法（centroid method）

$$d(C_q, C_r) = \left\| M(C_q) - M(C_r) \right\|^2$$

ここで、$M(C_q)$はクラスターC_qの重心を表し、$\| \quad \|^2$はユークリッド距離の 2 乗を表す。

（２）Ward 法（Ward's method）

次式で与えられる$\Delta E(C_q, C_r)$を距離の代わりとして用いる。

$$\Delta E(C_q, C_r) = E(C_q \cup C_r) - E(C_q) - E(C_r)$$

ただし、

$$E(C_q) = \sum_{x \in C_q} \left\| x - M(C_q) \right\|^2$$

である。

図１で表されているクラスターの親子関係の描画は、再帰的方法を用いると簡単に行うことができる。

まず、クラスターを表すオブジェクトのクラス型 TCluster が次のように宣言されているとする。

```
PntrC    = ^TCluster;          //   クラスターへのポインタ
TCluster = record              //   クラスターを表すクラス型
             L, R : PntrC;     //   クラスターL と R の結合クラスター
             x,                //   樹形図におけるシフト量
             y : extended;     //   クラスタ形成基準の距離（類似度）
                   .
                   .
                   .
           end;
```

すなわち、その子であるクラスターを指し示すために２つのポインタ L と R を用意する。樹形図におけるクラスタの縦方向の位置を x、横方向の位置を y で表す。横方向の位置は、L と R の指し示すクラスター間の距離（クラスターを生成したときの距離）に対応させる。

上のオブジェクトによって表されているクラスター間の関係（ポインタ L および R によ

るリスト構造）を再帰的手続き（DrawTree と名付けておく）により描画する。再帰的手続き DrawTree のヘッダーは

```
procedure DrawTree( C : PntrC );
```

と宣言されているとする。

プリンタ用紙の上からの位置を表す変数を cpos とし、cpos の初期値を用紙の上部の位置を表す値に設定する。最後のクラスター（すべての対象を含むもの）を表すオブジェクトを指し示すポインタが C0 であれば、まず手続き DrawTree を次の形

```
DrawTree( C0 );
```

で呼び出す。

このとき、DrawTree を以下のように再帰的に構成しておくと、クラスター間の関係を表す樹形図を描くことができる。

（１）C^.L あるいは C^.R が nil であれば、C^は対象が１つからなるクラスターを表す。このとき、（１a）にすすむ。C^.L あるいは C^.R が nil でないときは（２）にとぶ。

（１a）cpos の値を単位量だけ増やした後、その cpos の値を C^.x に設定する。上からの位置が cpos の表す位置、左からの位置が距離０に対応する位置を C の指し示すクラスターC^の位置として、そこに C^のラベルを印字する。

（１b）この手続きの実行を終了する。

（２）DrawTree(C^.L)を呼び出した後、DrawTree(C^.R)を呼び出す。

（３）C^のプリンタ用紙での上からの位置を

$$C^{\wedge}.x \leftarrow (C^{\wedge}.R^{\wedge}.x+C^{\wedge}.L^{\wedge}.x)/2$$

と設定する。C^の位置を、上から C^.x、左から C^.y として、C^.L^の位置および C^.R^の位置と C^の位置を線分で結ぶ。

（４）この手続きの実行を終了する。

プログラム PClusterCWDist.exe を起動すると図２のフォームが表示される。

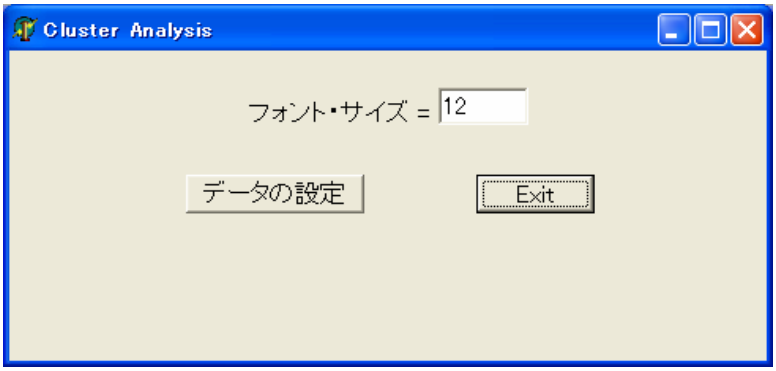


図２　PClusterCWDist.exe の実行開始時のフォーム

フォントサイズに対応して、クラスタ分析の結果を表す樹状図の大きさが決まる。「データの設定」ボタンのクリックで図３のフォームが表示される。

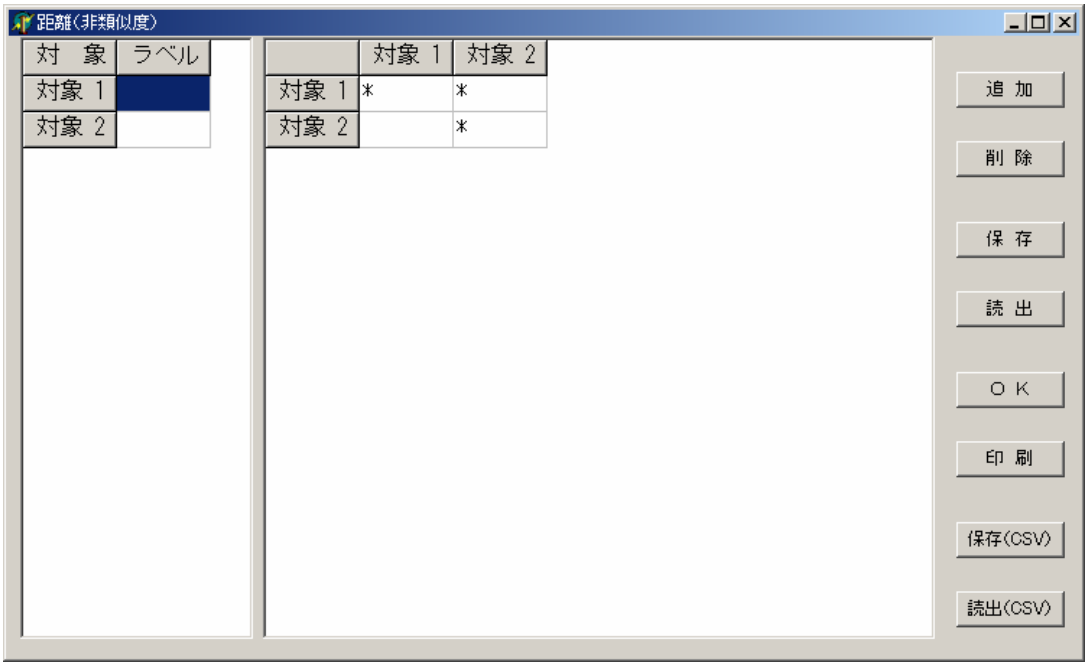


図３　距離データ入力用フォーム

左側の StringGrid コンポーネントにクラスタ分析を行う対象のラベルを設定する。このラベルは樹形図の出力で用いられる。右側の大きな StringGrid に距離データを設定する。クラスタ分析にかける対象の数は、「追加」ボタンのクリックで増やすことが出来る。追加は、右側の StringGrid 内のアクティブなセルの下側に空白行が挿入される。セルはそのセル内のクリックによりアクティブになる。アクティブなセルの下側に空白行が設定されるのに合わせて、列の挿入も行われ、左側のラベル設定用 StringGrid 内の行も追加される。「削除」ボタンをクリックすると、アクティブなセルを含む行が削除され、削除された行の対

象に対応する列とラベルの設定行も削除される。

「追加」ボタン、「削除」ボタンのクリックによって StringGrid 内に必要な行数、列数を設定した後、距離データを設定する。距離データは、下３三角行列の形式で設定する。対角成分とその上側の部分は＊印が設定されていて、ここにはデータを設定する必要はない。

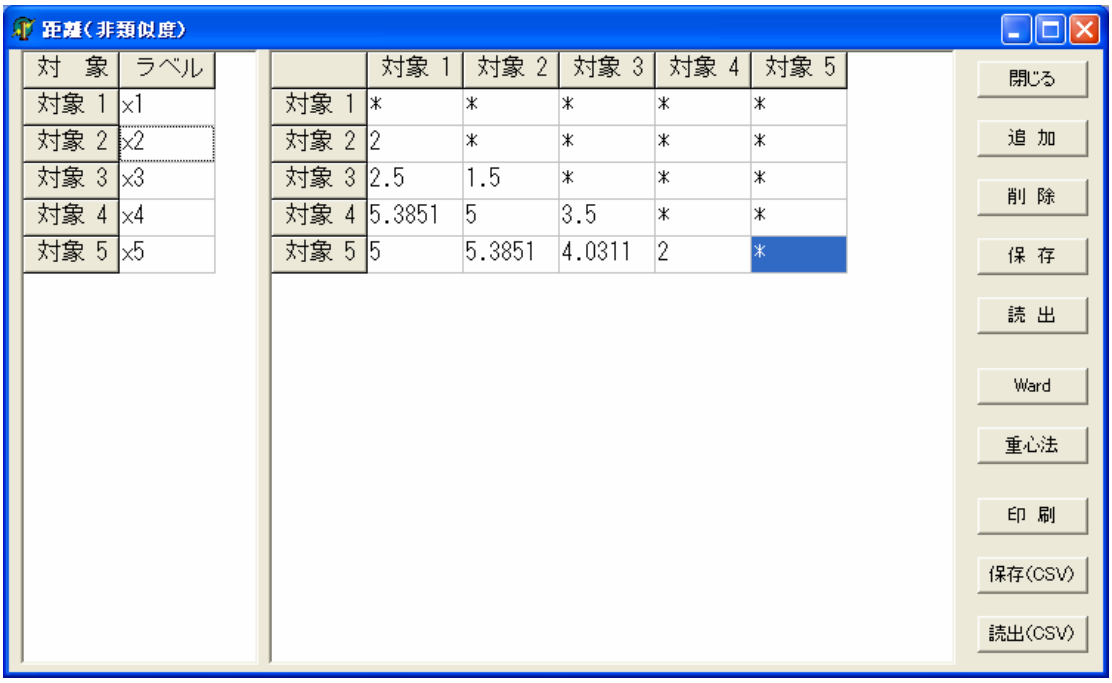


図４　距離データの設定例

図４は、４つの対象、x、y、z、w、の距離データの設定例である。設定されたデータは、「保存」あるいは「保存（CSV)」ボタンのクリックでファイルに保存することができる。「保存」ボタンのクリックで保存したデータは、「読出」ボタンのクリックで読み出すことができる。「保存（CSV)」ボタンのクリックで保存したデータは、「読出（CSV)」ボタンのクリックで読み出すことができる。「保存（CSV)」ボタンのクリックで保存されたデータは CSV 形式で保存されているので、Excel で開くこともできる。例えば、図４の状態で「保存（CSV)」ボタンをクリックすると、まずデータ保存のためのファイル名の設定を求めるダイアログボックスが図５のように表示される。

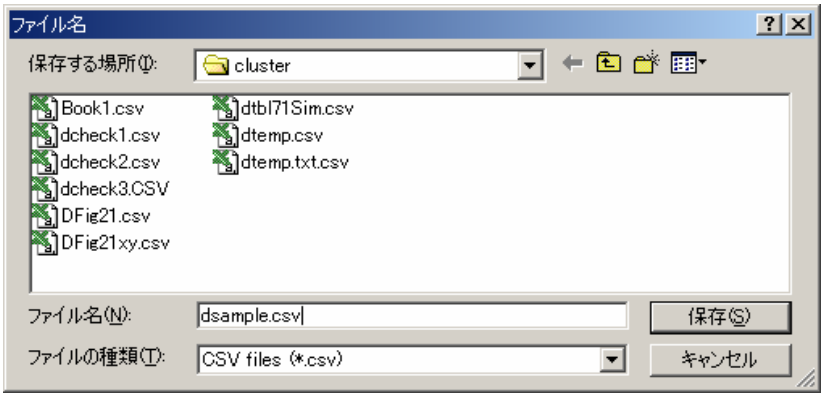
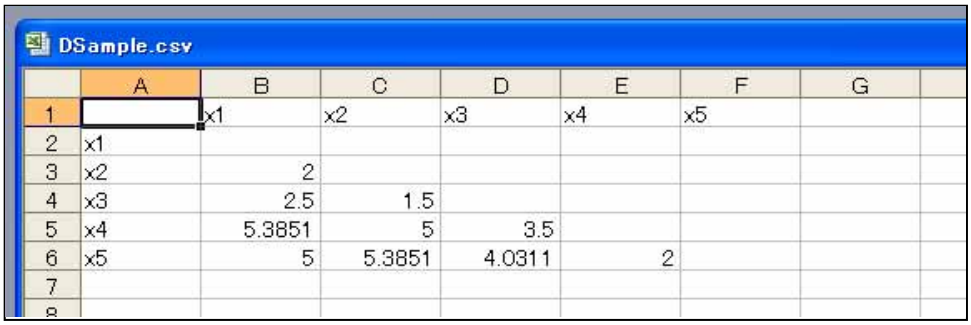


図５　データ保存用ファイル名の設定

図５で設定した名前 dsample.csv で保存されたファイルを Excel で開くと図６のようになる。



	A	B	C	D	E	F	G
1		x1	x2	x3	x4	x5	
2	x1						
3	x2	2					
4	x3	2.5	1.5				
5	x4	5.3851	5	3.5			
6	x5	5	5.3851	4.0311	2		
7							
8							

図６　「保存（CSV)」ボタンのクリックで保存したデータを Excel で開いたもの

逆に、Excel で図６の形式により距離データを設定したものを CSV 形式で保存したものは、図３のフォームで「読出（CSV)」ボタンをクリックすると読み込むことができる。Excel で CSV 形式により保存するときは、拡張子として.csv を選ぶ。

図４のように設定されたデータは、「印刷」ボタンのクリックでプリンタに出力することができる。

データの設定後、「Ward」あるいは「重心法」ボタンをクリックするとそれぞれの方法でクラスタ分析が始る。まず入力データ値などを出力するテキストファイルの名前の設定を求めるダイアログボックスが表示される（図７)。

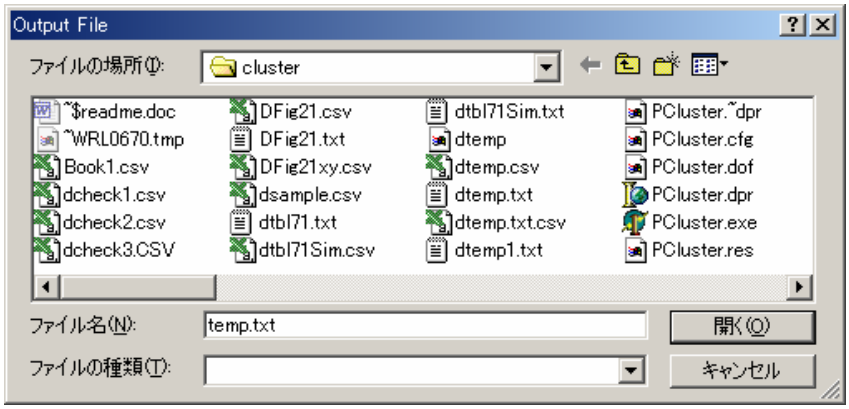


図７　出力用ファイル名の設定

ファイル名の設定後、「開く」ボタンをクリックすると図７のダイアログボックスは閉じられる。続いて、プリントダイアログボックスが表示される（図８）。

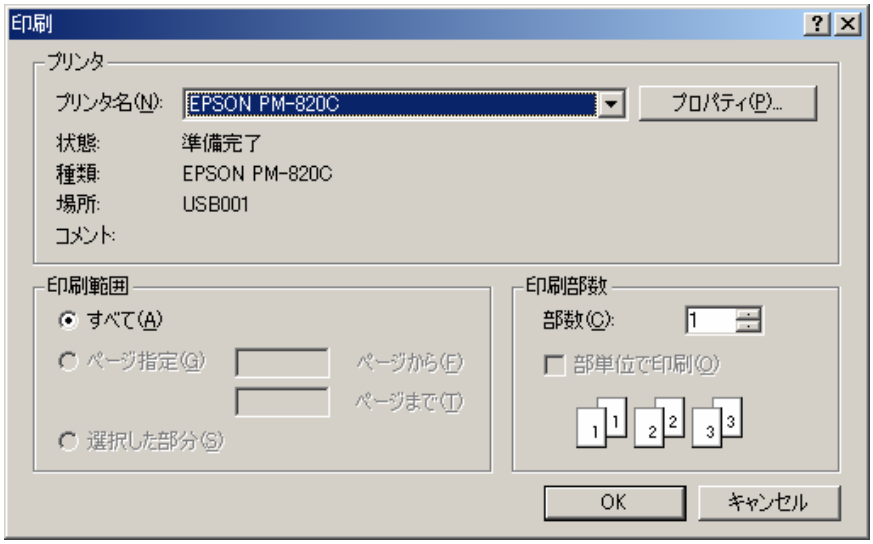


図８　プリントダイアログボックス

図８のダイアログボックスにおいて、必要ならばプリンタの設定・選択などを行うことができる。「OK」ボタンのクリックで、プリンタに樹形図が出力される。

プリンタへの出力が終わると図９のダイアログボックスが表示される。

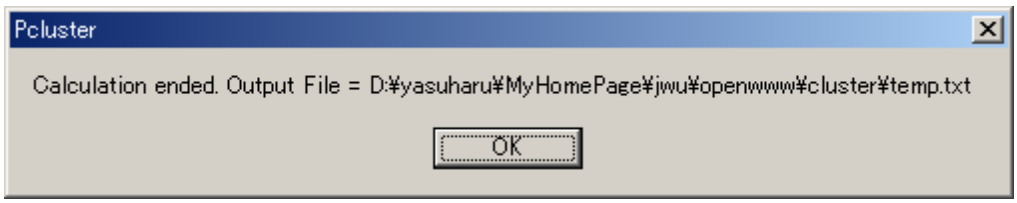


図９　プログラムの終了時に表示されるダイアログボックス

ダイアログボックスには、出力用ファイルの名前として設定したファイル名が表示されている。このファイルはテキストファイルなので、プログラムの実行終了後、エディタなどで開いて見ることができる。図９の「OK」ボタンのクリックでプログラムの実行終了となる。

図４のデータの場合で「Ward」ボタンをクリックしたときの樹形図は、図１０のようにプリンタ出力される。

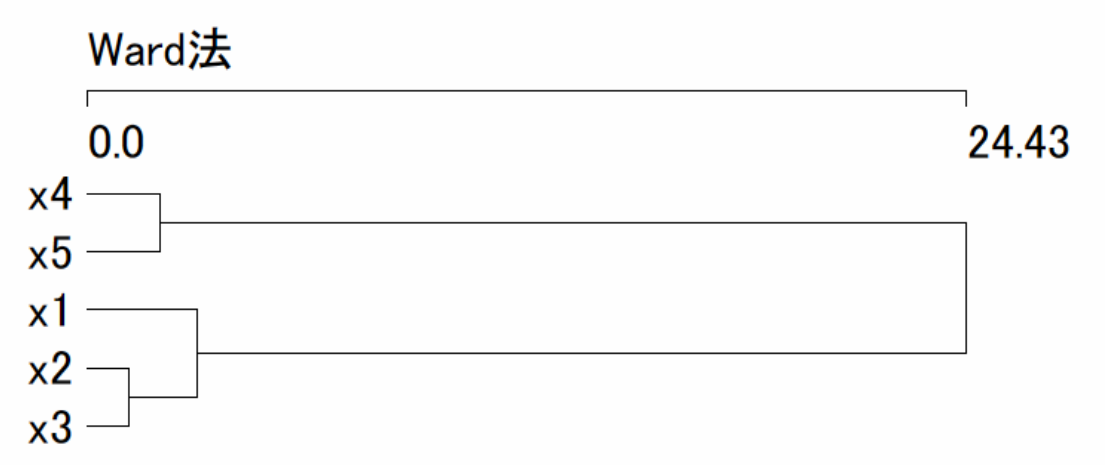


図９　樹形図

文字列「Ward 法」の下に引かれている線分とその下の数値は、２つのクラスタがまとめられた（結合、merge）ときの距離の値を示すためのものである。各対象からの線分は距離０の位置から始っている。図１０では、まず、x2 と x3 が１つのクラスタ{x2,x3}とまとめられ、次に x4 と x5 がクラスタ{x4,x5}にまとめられたことが示されている。その後、x1 とクラスタ{x2,x3}がクラスタ{x1,x2,x3}にまとめられている。最後にクラスタ{x4,x5}とクラスタ{x1,x2,x3}が結合されているが、このときのクラスタ{x4,x5}と{x1,x2,x3}の距離は 24.43 であったことがわかる。

参考文献

宮本定明「クラスター分析入門」森北出版株式会社、1999.