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。

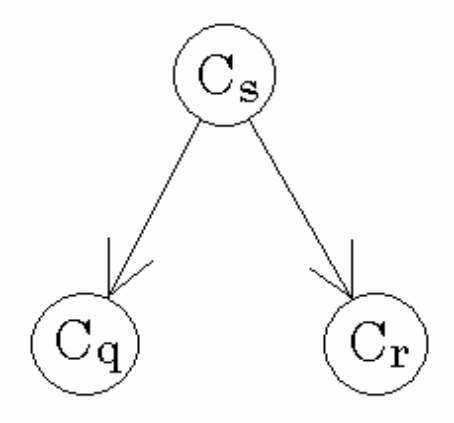

図1　クラスターの親子関係

クラスター間の距離の算出方法として、ここでは次のものを用いている。

（1）最短距離法（single linkage method または nearest neighbor method）

$$d(C_q, C_r) = \min_{x \in C_q, y \in C_r} d(x, y)$$

（２）最長距離法（complete linkage method または furthest neighbor method）

$$d(C_q, C_r) = \max_{x \in C_q, y \in C_r} d(x, y)$$

（３）群間平均法（average linkage between the merged group）

$$d(C_q, C_r) = \frac{1}{|C_q| \cdot |C_r|} \sum_{x \in C_q, y \in C_r} d(x, y)$$

（４）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。

（５）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。

N 個の対象とそれらの間の類似度が与えられているときは、類似度が大きいほど距離が小さいと考えて上の手続きを適用する。

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 PClusterRev.exe と PClusterCWRev.exe は、上の方法でクラスター分析を行うものである。PClusterRev.exe は、最長距離方法、群間平均法、最短距離法のいずれかによってクラスター分析を行うものである。PClusterCWRev.exe では、重心法あるいは Ward 法によってクラスター分析を行う。

プログラム PClusterRev.exe を実行すると図１のフォームが表示される。

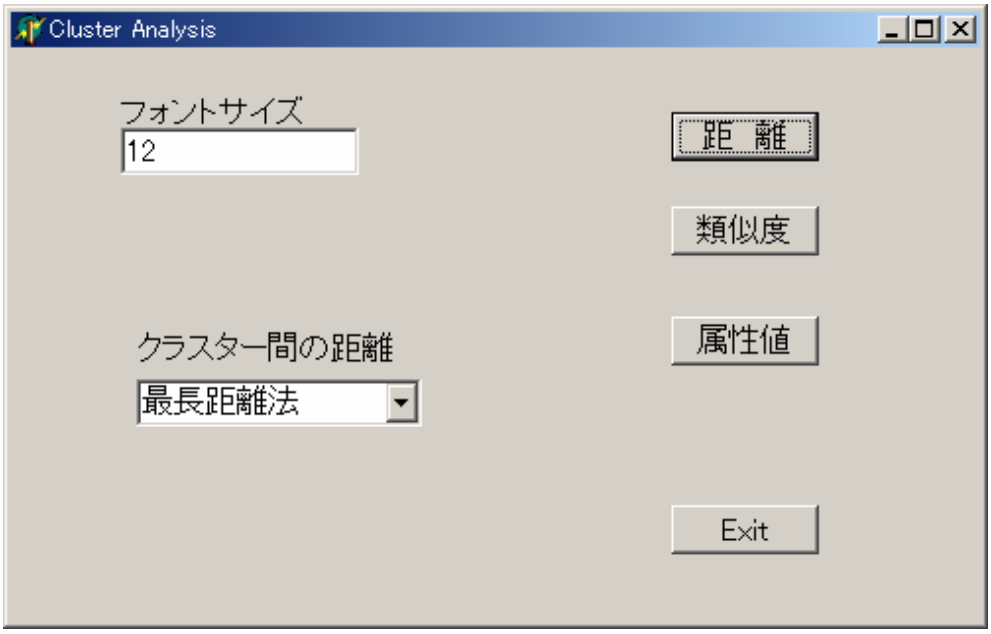

図１　PClusterRev.exe の起動時のフォーム

フォントサイズの値によってクラスター間の関係を表すツリーの描画サイズが決まる。「クラスター間の距離」の下にある ComboBox コンポーネントにおいて、最長距離法、群間平均法、最短距離法のいずれかを選ぶ。最長距離法以外を選ぶときは、ComboBox コンポーネントの右端のボタンをクリックすると項目のドロップダウンリストが表示されるので、そこから選ぶ（図２）。

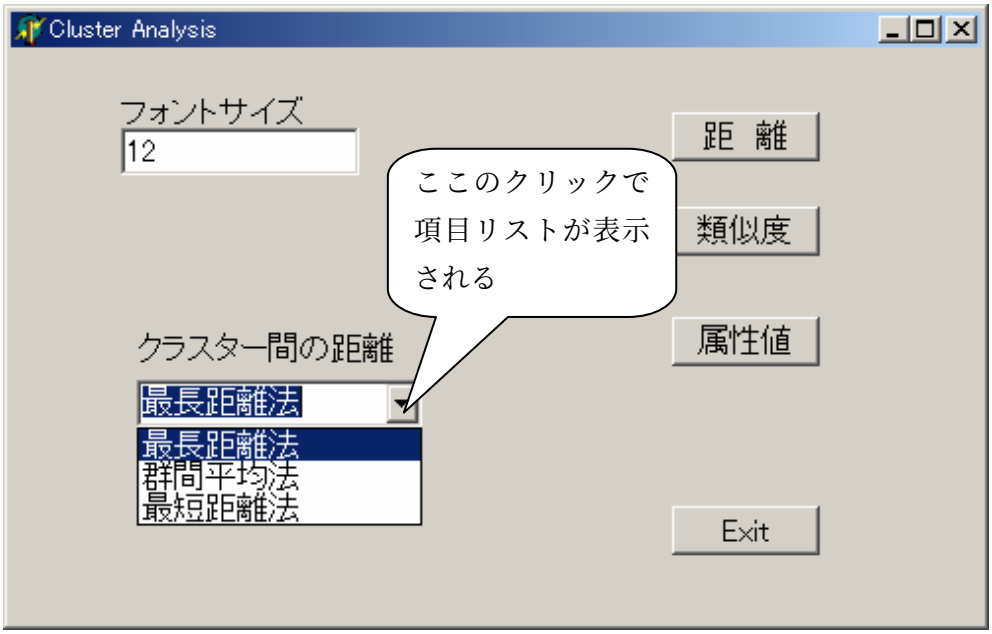

図２　ドロップダウンリストの表示

「クラスタ間の距離」の選択後、「距離」、「類似度」、「属性値」のいずれかのボタンをクリックする。これら３つのボタンは、データの入力形式を選択するものである。「距離」ボタンのクリックで、距離データを入力するフォームが表示される（図３）。

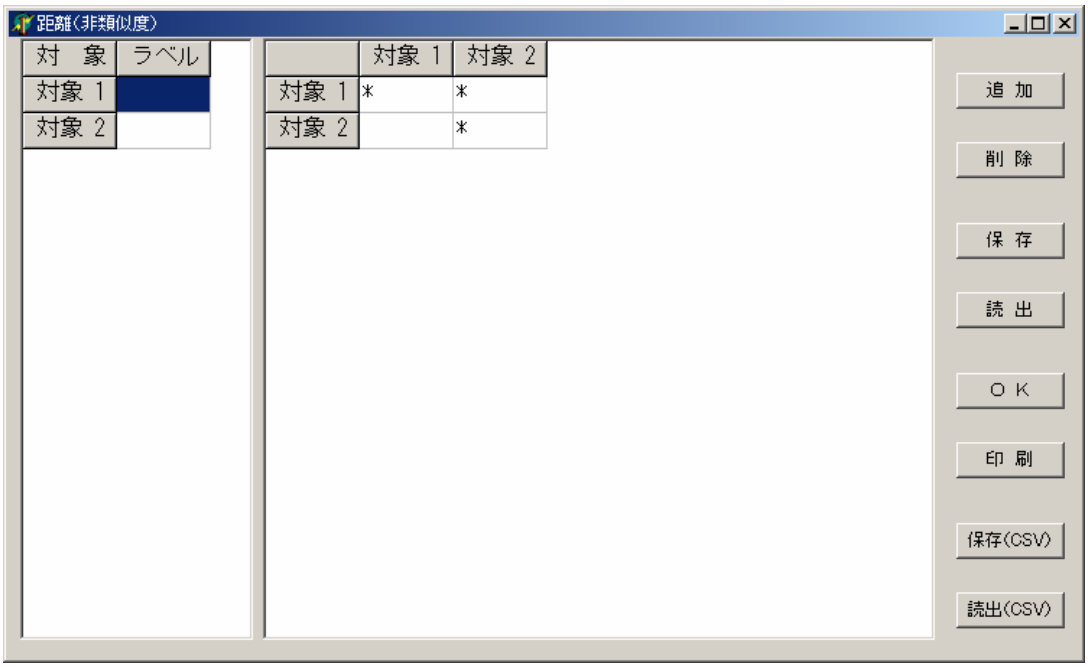

図３　距離データ入力用フォーム

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。

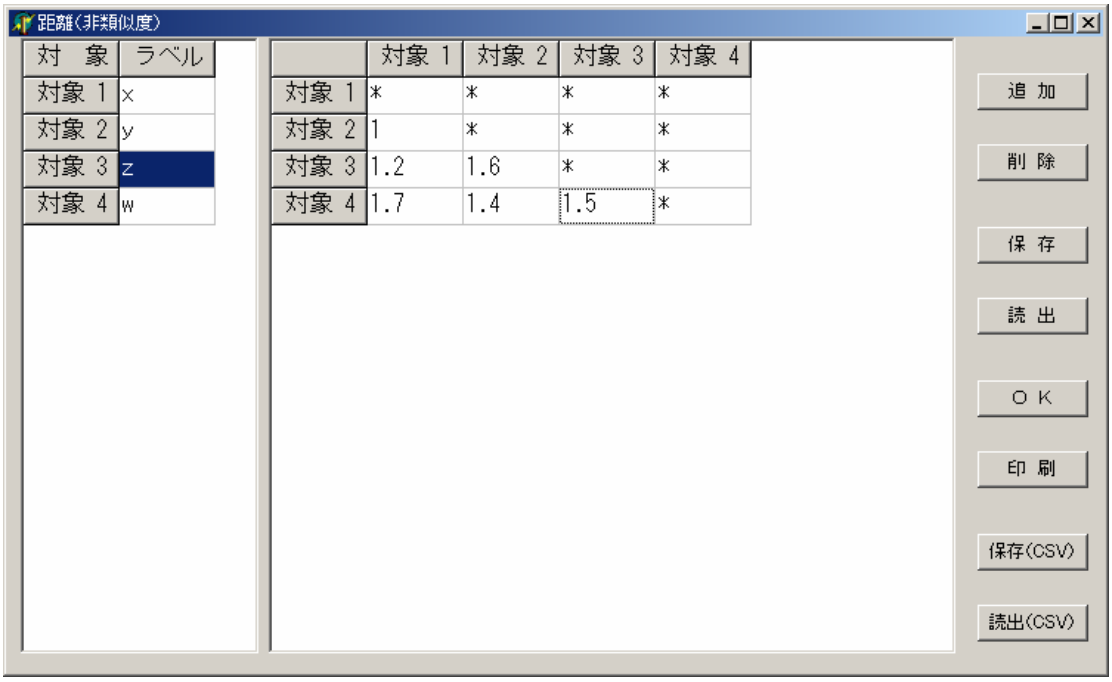

図４　距離データの設定例

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。

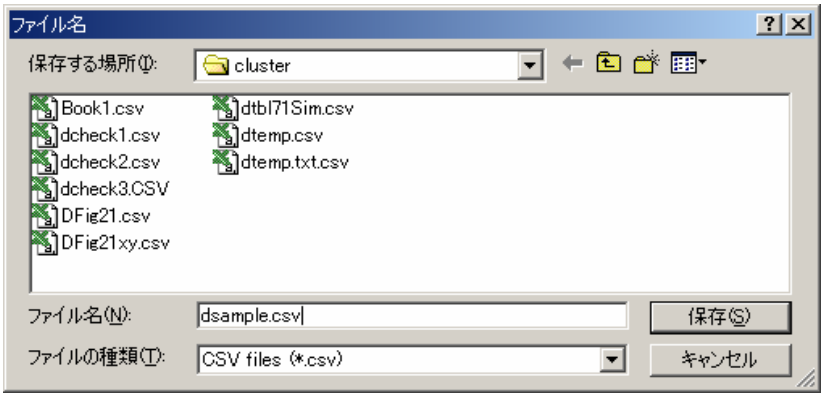

図５　データ保存用ファイル名の設定

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。

dsample.csv

	A	B	C	D	E	F
1		x	y	z	w	
2	x					
3	y	1				
4	z	1.2	1.6			
5	w	1.7	1.4	1.5		
6						
7						

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、「OK」ボタンをクリックするとクラスタ分析が始る。「OK」ボタンをクリックすると、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)。

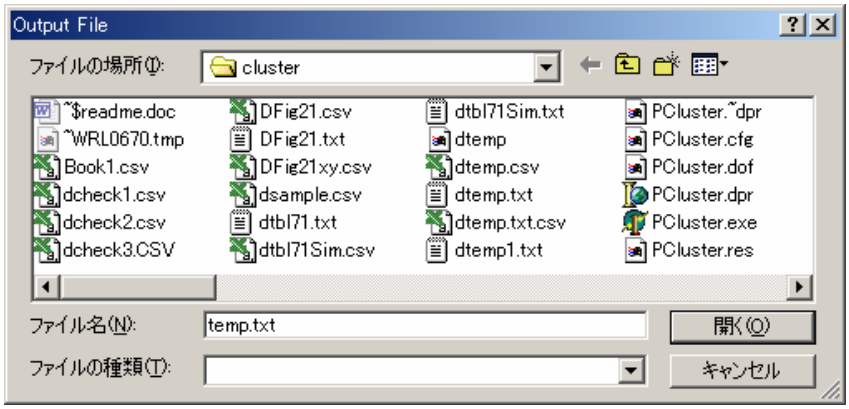

図７　出力用ファイル名の設定

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）。

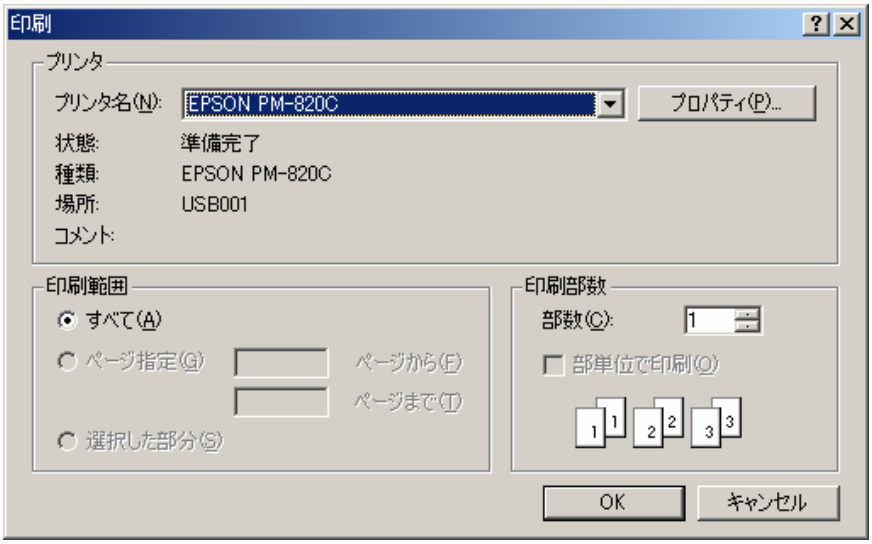

図８　プリントダイアログボックス

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。

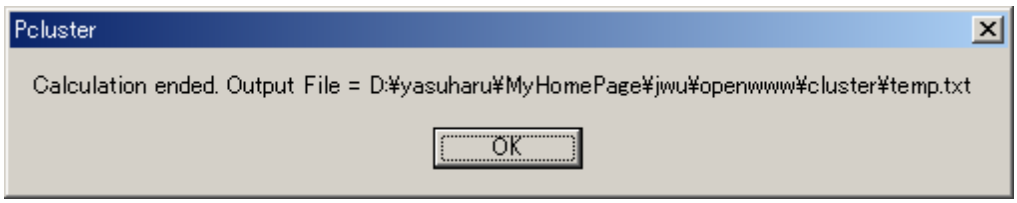

図９　プログラムの終了時に表示されるダイアログボックス

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のデータの場合の樹形図は、図１０のようにプリンタ出力される。

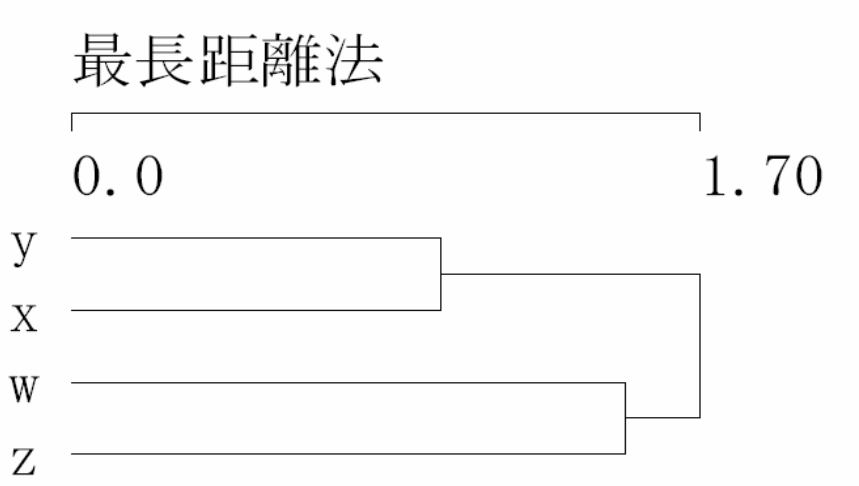

図１０　樹形図

文字列「最長距離法」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、y と x が１つのクラスタ{y,x}とまとめられ、次に w と z がクラスタ{w,z}にまとめられたことが示されている。最後にクラスタ{y,x}とクラスタ{w,z}が結合されているが、このときのクラスタ{y,x}と{w,z}の距離は 1.70 であったことがわかる。

図１において「類似度」ボタンをクリックした場合も、データ値として類似度を設定することを除いて「距離」ボタンのクリックの場合と同じである。

「属性値」ボタンをクリックすると図１１のフォームが提示される。

図１１　属性値設定用フォーム

図１１のフォームは、各対象の属性値（座標値）を入力データとしてクラスタ分析を行うためのものである。設定した座標値から対象間の距離が算出され、クラスタ分析が行われる。必要な行数および列数は、距離データの設定の場合と同様に、「追加」あるいは「削除」ボタンのクリックにより調整する。「追加（行）」ボタンのクリックでアクティブなセルを含む行の下に空白行が挿入・追加され、「削除（行）」ボタンのクリックでアクティブなセルを含む行が削除される。「追加（列）」および「削除（列）」では列単位の挿入・削除が行われる。

図１２は、データの設定例である。

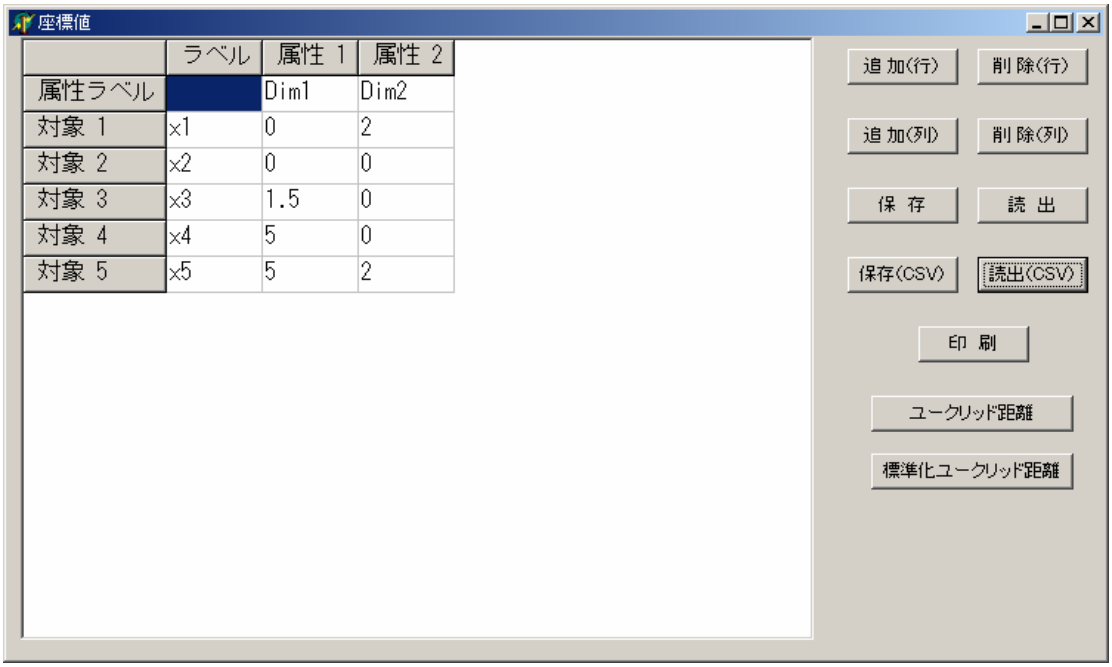

図１２　データの設定例

ラベルの列には各対象のラベルを設定する。このラベルは樹形図の描画において対象を示すために用いられる。属性ラベルの行には各属性のラベルを設定するが、この属性ラベルはクラスタ分析の出力には用いられない。入力の便宜のために設けられている。各対象の属性値（座標値）は、それぞれの行において各属性の列に設定する。図１２に設定されている対象１のラベルは「x1」であり、座標値は(0, 2)である。

設定したデータは、「保存」ボタンのクリックで保存することができる。保存されたデータは、「読出」ボタンのクリックで読み込むことができる。「保存（CSV)」ボタンの場合は CSV 形式で保存される。この場合は「読出（CSV)」ボタンのクリックで読み込む。CSV 形式で保存されたデータは、Excel で開くこともできる。図１３は、図１２のデータを「保存（CSV)」ボタンのクリックで保存したものを Excel で開いた場合を示す。

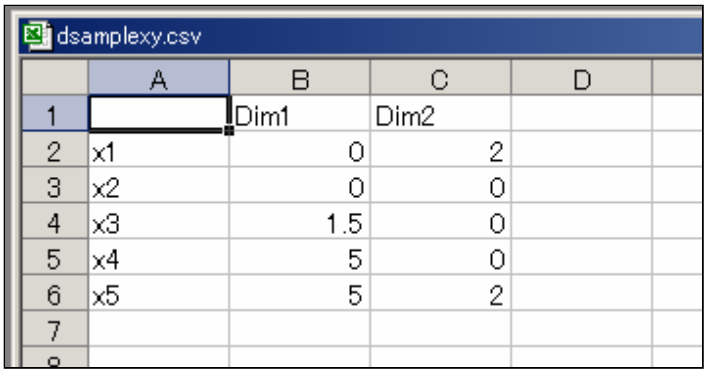

図１３　Excel の場合

逆に、Excel で図１３の形式によって用意されたデータは、ファイルの拡張子を.csv として CSV 形式のファイルとして保存すれば、図１１における「読出（CSV)」ボタンのクリックで読み込むことができる。

図１２のようにデータを設定した後、「ユークリッド距離」ボタンあるいは「標準化ユークリッド距離」ボタンをクリックすると、設定された座標値から距離が算出され、クラスタ分析にかけられる。「ユークリッド距離」ボタンをクリックすると、設定された座標値がそのまま用いられる。「標準化ユークリッド距離」ボタンをクリックすると、座標値の値が各次元（属性）ごとに平均値が０、分散が１になるように標準化された後、その標準化された座標値から距離が算出される。距離が算出された後の処理は、「距離」ボタンのクリックの場合と同じである。

プログラム PClusterCWRev.exe は、重心法あるいは Ward 法による基準によってクラスタの結合を行うものである。このプログラムを起動すると図１４のフォームが表示される。

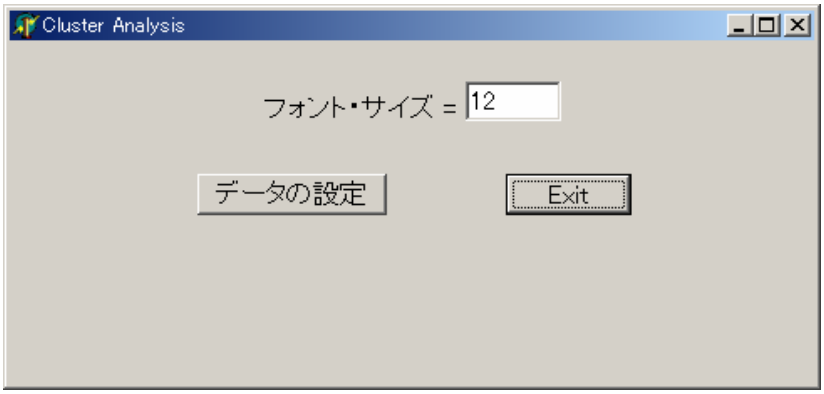

図１４　PClusterCWRev.exe の実行開始時のフォーム

フォントサイズの値に応じて描画されるツリーの大きさが決まる。
「データの設定」ボタンのクリックで図１５のフォームが表示される。

図１５　「データの設定」ボタンのクリックで表示されるフォーム

このフォームにおけるデータの設定は、図１１のフォームと同じである。データ設定後にクリックするボタンが、図 15 の場合「重心法」と「Ward 法」であることが、図１１のフォームとの違いである。「重心法」ボタンをクリックすると、クラスタ間の距離を重心法によって算出した場合のクラスタ分析が行われる。「Ward 法」ボタンのクリックでは、クラスタの結合が Ward 法による基準に従って行われる。「重心法」ボタンあるいは「Ward 法」ボタンのクリック後のプログラムの操作は、PClusterRev.exe の場合と同じである。

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.