

Rとクラスター分析(2)

1. 樹形図の切断とコーフェン相関係数

先月号では、階層的クラスター分析の基本概念や樹形図の作成などについて説明した。

クラスター分析結果を分析する際には、どの個体がどのクラスターに属するかを確認することが必要である。階層的クラスター分析では、クラスターの数を指定し、樹形図を切断すると個体が属するクラスが決定される。

(1) 樹形図の切断

R には個体が属するクラスターの情報を返す関数 `cutree` が用意されている。関数 `cutree` はクラスターの数を指定すると、個体がどのクラスターに属するかに関する情報を返す。

次に `iris` データの 51~100 行 (`versicolor` 品種) と 101~150 行 (`virginica` 品種) の 2 品種のデータを用いた関数 `cutree` の使用例を示す。

```
>iris2<-iris[51:150,1:4]
>iris2.hc<-hclust(dist(iris2),"ward")
>(iris2.cl<-cutree(iris2.hc,k=2))
 51  52  53  54  55  56  57  58  59  60
  1   1   1   1   1   1   1   1   1   1
<中略>
141 142 143 144 145 146 147 148 149 150
  2   2   1   2   2   2   1   2   2   1
```

クラスター分類の結果の精度を確認するため、既知のクラス情報と分類結果のクロス表を次に

示す。

```
>iris2.lab<-c(rep(1,50),rep(2,50))
>table(iris2.lab, iris2.cl)
      iris2.cl
iris2.lab  1  2
      1  50  0
      2  14 36
```

この方法では、14 個の `virginica` 品種が `versicolor` 品種に誤分類されている。

階層的クラスター分析では、同一の距離測度を用いても、用いるクラスター分析の方法によって結果が異なる可能性がある。その際、どの結果を信頼すべきであるかが 1 つの問題である。樹形図は、距離データから求めたコーフェン行列の図示である。そこで、与えられた距離とコーフェン行列との相関係数(コーフェン相関係数)を用いて結果を評価することが提案されている(西田・他 1992)。

(2) コーフエン相関係数

R にはコーフェン行列を返す関数 `cophenetic` が用意されている。前項で求めた `iris2.hc` のコーフェン行列は次のように返す。

```
>iris2.cop<-cophenetic(iris2.hc)
```

コーフェン相関係数は、距離の行列と用いた方法で生成したコーフェン行列とのピアソン相関係数である。次にユークリッド距離とウォード法によるコーフェン相関係数を求める例を示

す。

```
>cor(iris2.cop,dist(iris2))  
[1] 0.6184636
```

このコーフエン相関係数の値が大きいほど、距離行列と用いた方法のコーフエン行列との歪みが小さいと、ある側面から言える。しかし、歪みが小さいことと分類の結果がより妥当であることは等価ではない。

ここでは、データ iris2 のユークリッド (Eucliden) 距離とキャンベラ (canberra) 距離を用いたコーフエン相関係数と正しく分類された比率(正解率)を表 1 に示す。

正解率は、関数 cutree を用いて各個体のクラスの属性を求め、既知のクラス属性情報とのクロス表から次のように求めた。

```
>iris2.ta<-table(iris2.lab,iris2.cl)  
>sum(diag(iris2.ta))/100  
[1] 0.86
```

表 1 コーフエン相関係数と正解率

方 法	ユークリッド距離		キャンベラ距離	
	相関係数	正解率	相関係数	正解率
最近隣法 (single)	0.575	0.52	0.462	0.51
最遠隣法 (complet)	0.615	0.24	0.533	0.94
群平均法 (average)	0.658	0.86	0.548	0.94
重心法 (centroid)	0.661	0.52	0.544	0.40
McQuitty 法 (mcquitty)	0.633	0.85	0.544	0.94
メディアン法 (median)	0.552	0.85	0.616	0.58
ワード法 (ward)	0.618	0.86	0.605	0.85

表 1 から分かるように、必ずしもコーフエン相関係数が高い方法の分類結果が良いとは言えない。

データの構造が複雑になると経験上ワード法は、妥当と思われる結果を返す確率が高いが、

コーフエン相関係数はいつも比較的に低い。ワード法によるコーフエン行列の生成は、最近隣法、群平均法などと異なり、距離の測度ではなく、群内の分散と群間の分散の情報を用いて求めているのが 1 つの原因でもある。

距離によるクラスター分析を行う際には、クラスター分析の方法以外に、用いる距離の測度も結果に大きい影響を与える。どのような距離測度、どのような方法を用いるべきであるかは、データに依存するので、経験に頼るのが現状である。

階層的クラスター分析は、データの構造が複雑になると少数個の個体を入れ替えるだけで、結果が大きく変わるケースも少なくない。また、用いる方法の違いで大きく異なる結果が得られることも珍しくない。どの結果を信じるべきであるかに関しては、階層的クラスター分析の結果だけではなく、他のデータ解析方法をも用いて探索的に様々な角度からデータを分析して総合的な判断することが必要である。

2. 非階層的クラスター分析

階層的クラスター法は、個体数が多いと計算量が膨大になり、大量のデータ解析には向いていない。大規模のデータセットのクラスター分析には、非階層的クラスター法が多用されている。非階層的クラスター法の代表的な方法として k 平均 (k-means) 法がある。

(1) k 平均法のアルゴリズム

k 平均法も幾つかの方法が提案されているが、その大まかな流れは同じである。そのアルゴリズムを次に示す。

① k 個の初期クラスターの中心 (seeds) を何ら

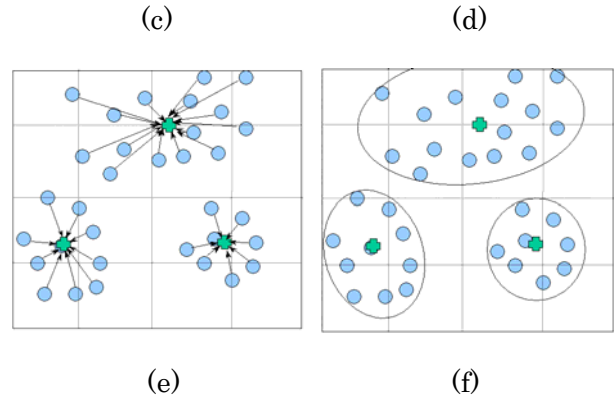
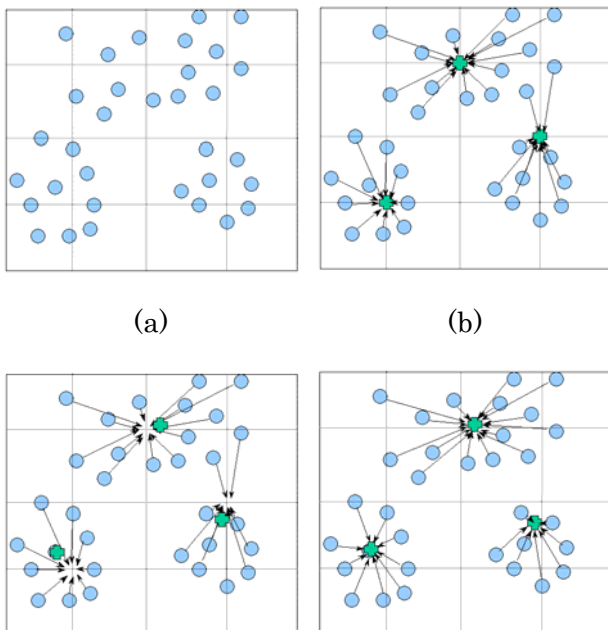
かの方法で与える。

- ② すべてのデータと k 個のクラスター中心との距離を求め、最も近いクラスターに分類する。
- ③ 新たに形成されたクラスターの中心を求める。
- ④ クラスターの中心がすべて前の段階の結果と同じになる、あるいは事前に指定している繰り返しの回数に達するまで②、③を繰り返す。

理解を助けるため、2次元平面上の散布図を用いて、 k 平均法のアルゴリズムのイメージを説明する。例えば、図 1(a)のような散布図があるとする。これを3つのクラスターに分類する場合、まず図 1(b)のように何らかの方法で3つの中心($k=3$)を与え、この中心のクラスターを求める。

次は図 1(c)のように新しく求めたクラスターの中心を求め、図 1(d)のように新しいクラスターを求める。このような作業を、クラスターの中心が変わらなくなるまで繰り返す。

図 1 k 平均法のクラスター形成過程



k 平均法の中で、多く用いられているのは、Lloyd 法、Forgy 法、MacQueen 法、Hartigan-Wong 法である。

(2) 関数 kmeans と例

R には k 平均法の関数 **kmeans** がある。関数 **kmeans** の書き方を次に示す。

```
kmeans(x, centers, iter.max = 10, nstart = 1,  
       algorithm = c("Hartigan-Wong",  
                     "Lloyd", "Forgy", "MacQueen"))
```

引数 x はデータ、 $centers$ はクラスターの数あるいはクラスターの中心、 $iter.max$ は繰り返しの最大値、 $nstart$ は初期中心の与え方を指定する。 $centers$ がクラスターの数である場合、初期中心値はランダムに与える。引数 $algorithm$ は、4 つの方法 ("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen") から 1 つ選んで指定する。デフォルトには、Hartigan-Wong が指定されている。

関数 **kmeans** はクラスターの分類結果 ($\$cluster$)、クラスターの中心ベクトル ($\$centers$)、各クラスター内の個体数 ($\$size$) などを返す。

次に *iris2* データを用いた関数 **kmeans** の使用例を示す。

```
>iris2.km<-kmeans(iris2,2)
```

iris2.km\$cluster にクラスターの分類結果が記録されている。次のコマンドで、クラスターの精度を確認することができる。

```
>table(iris2.lab,iris2.km$cluster)/50
```

```
      1      2
1 0.96 0.04
2 0.28 0.72
```

用いたデータの k-means 法の正解率は、個体 51~100 番は 0.96 (96%) で、個体 101~150 番は 0.72 (72%) である。

3. モデルに基づいたクラスター分析

モデルに基づいたクラスター分析は、通常モデルに基づいたクラスタリング (Model-Based Clustering)、混合分布によるクラスター分析、潜在クラスター分析とも呼ばれている。

モデルに基づいたクラスター分析は、観測データが異なる分布の混合分布であると仮定し、個体が属するクラスのラベルをも隠れ変数として推定する。混合分布のパラメータおよびクラスのラベルの推定は EM(expectation maximization) アルゴリズムが多く用いられている [5]。

パッケージ `mclust` には関数 **EMclust**、**hc**、**hclust** などモデルに基づいたクラスタリングに関連する豊富な関数が用意されている。

関数 `EMclust` は、最大尤度推測法を用いた EM アルゴリズムでパラメータを推定する際に必要となる、ガウス混合分布モデルの情報量規準 BIC (Bayesian Information Criterion) 値を求める。通常情報量規準は、値が小さいモデルが真のモデルに、より近似していると判断するが、パッケージ `mclust` では、BIC

の値が大きいモデルが真のモデルに、より近似すると判断する。つまり、情報量規準の式の正負が通常の式とは逆である。

関数 `hc` は、モデルに基づいた階層的クラスタリングを行う。関数 `hc` では、混合分布の型 (球、楕円球)、体積、形と軸の向きが同じであるかどうかを引数 `modelName = "` で指定するようになっている。多変量の場合は次の 4 種類から 1 つを選択できる。

"EII": 球型、同体積

"VII": 球型、異なる体積

"EEE": 楕円球型、同体積・形・向き

"VVV": 楕円球型、異なる体積・形・向き

関数 `hclass` は、関数 `hc` の結果に、クラスの数指定すると個体が属するクラスの推測結果を返す。

ここでも `iris2` のデータを用いて、その使用例を示すことにする。

パッケージ `mclust` では、関数 `plot` を用いて関数 `EMclust` が返すモデルの BIC 値の折れ線グラフを作成することができる。

```
>library(mclust)
```

```
>plot(EMclust(iris2))
```

```
EII VII EEI VEI EVI VVI EEE EEV VEV VVV  
"A" "B" "C" "D" "E" "F" "G" "H" "I" "J"
```

図 2 クラスタ数とモデルの BIC 値

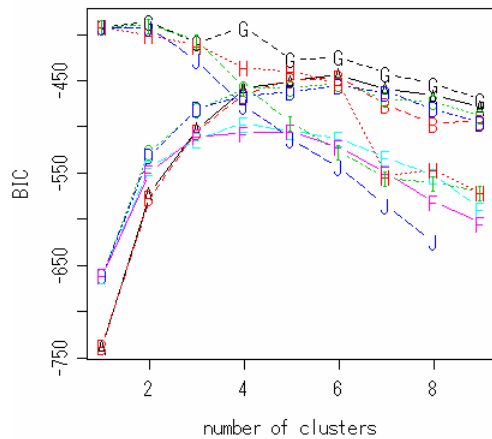


図2では、折れ線Gのクラスター数2,3のBIC値が最も大きい。GはモデルEEE(楕円球型、同体積・形・向き)である。モデルEEEに基づいた階層的クラスタリングは次のように求める。

```
>mhc<-hc(modelName = "EEE", data = iris2)
```

さらに、混合分布EEEモデルに基づいて求めた結果を、次のように関数hclassにクラスターの数を指定することで、個体が属するクラスを決めることができる。

```
>cl<-hclass(mhc,2)
```

クラスタリングの精度を確認するため、既知のクラスのラベルと推測結果とのクロス表を次に示す。

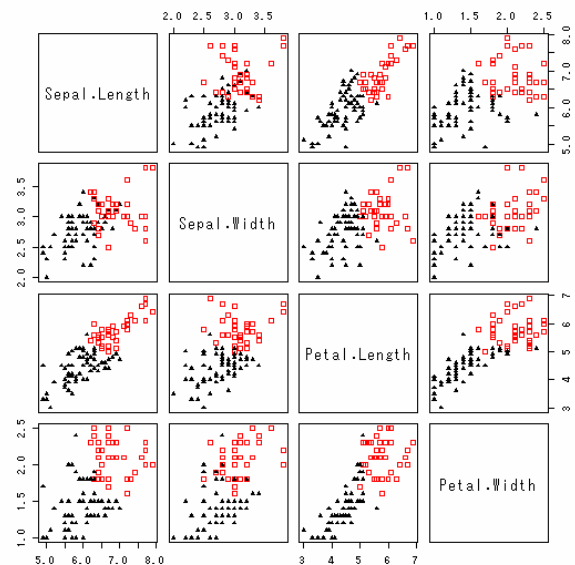
```
> table(iris2. lab, cl)
      cl
iris2. lab  1  2
      1 49  1
      2 15 35
```

関数 **clPairs** を次のように用いることにより、

推定されたクラスターの対散布図を作成することができる。

```
>clPairs(iris2,cl=cl)
```

図3 混合分布(EEE, k=2)に基づいた iris2 のクラスタリンクの対散布図



4 その他

クラスター分析に関しては、パッケージ **cluster** に関数 **clara**、**pam**、**fanny**、**agnes**、**diana**、**mona** など豊富な関数がある。また、パッケージ **amap** に関数 **hcluster** がある。

関数 **pam**、**clara**、**fanny** はk平均法と同じく、分割化の方法による非階層的クラスター分析の結果を返す。関数 **pam** は、k平均法よりロバストで、距離データを扱うこともできる。これらの関数の結果は、関数 **plot** を用いると図4のように散布図にクラスターごとの楕円が自動的に描かれる。

次のコマンドラインを実行すると3つの選択項目が返され、「選択:」右に数値1、2、3いずれか1つを入力し、[Enter]キーを押すとグラフが作成される。数値2を入力し、実行すると図4が得られる。作図の終了は、メニューのアイコン[STOP]をクリック

クする。

```
> plot(pam(iris[,1:4],3),ask = TRUE)
```

Make a plot selection (or 0 to exit):

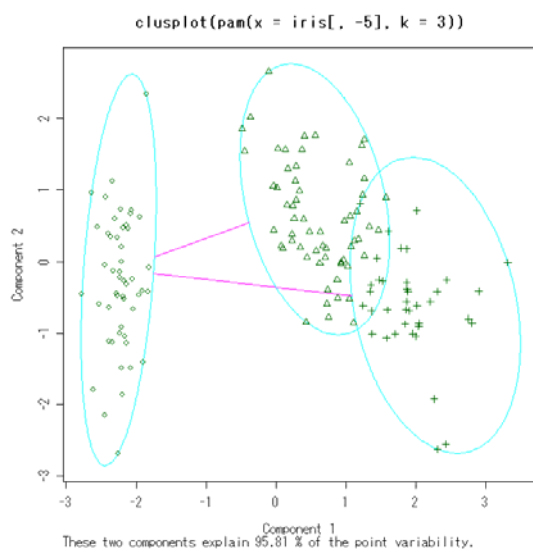
1: plot All

2: plot Clusplot

3: plot Silhouette Plot

選択 : 2

図4 関数 pam の結果の散布図



関数 agnes、diana、mona は階層的クラスター分析関数である。関数 mona は、2 値データの階層的クラスター分析の専用関数である。

オーソドックスな多変量データ解析の書物のほとんどは階層的クラスター分析を扱っているが、非階層的クラスター分析、モデルに基づいたクラスタリング法に関する内容を扱っている書物は比較的少ない。クラスター分析の初心者向きには参考文献[7]がある。より深く追求したい方には、クラスター分析全般に関しては[2]、[3]、[4]が、k 平均法については[1]が、モデルに基づいたクラスタリングについては[5]、[6]が参考になるであろう。

参考文献

- [1] 江原淳・佐藤栄作(1999)：データマイニング手法、海文堂
- [2] 西田英郎・他 共訳(1988)：クラスター分析とその応用、内田老鶴圃
- [3] 西田英郎・他 共訳(1992)：実例クラスター分析、内田老鶴圃
- [4] 西田春彦・他 共訳(1983)：クラスター分析、マイクロソフト
- [5] 渡辺澄夫(2001)：データ学習アルゴリズム、共立出版
- [6] 麻生英樹・津田宏治・村田昇(2003)：パターン認識と学習の統計学、岩波書店
- [7] 山口 和範，高橋 淳一，竹内 光悦(2004)；図解入門 よくわかる多変量解析の基本と仕組み—巨大データベースの分析手法入門、秀和システム