Rによるクラスター分析　　　2014年１月

（金明哲　2009　『Rによるデータサイエンス』第５章）

クラスター分析とは

どの個体がどのグループに属するかに関する事前情報、すなわち、外的基準を持たないデータをグループ分けする方法。データのパターンが似ている個体を同じクループにまとめる分析法。

階層的クラスター分析：

個体間の類似度あるいは非類似度（距離）に基づいて、最も似ている個体から順次に集めてクラスターを作っていく方法。

樹系図テンドログラム：

葉（ラベル部分）と葉の距離（歯から上に伸びている線が連結するまでの高さ）が短いほど個体が似ていると解釈する。いくつかの個体が階層的に集まってクラスター（房、枝）を形成し、複数のクラスターが最終的には一つのクラスター(木)に収束する。

樹系図をある高さ（距離）のところで切断して、クラスターの数と個体の分類が定まる。

階層的クラスター分析のプロセス

１．データから距離（あるいは類似度）を求める（多次元尺度法と同じ）

２．クラスター分析の方法（最近隣法、最遠隣法など）を選択する

３．選択された方法おコーフェン(cophenetic)行列を求める

４．コーフェン行列に基づいて樹系図を作成する

５．結果について検討する

データ行列⇒距離の行列⇒コーフェン行列⇒樹系図

コーフェン距離：もっとも距離が近い２つの個体間の距離

> seiseki<-read.table("J:\\seiseki.dat", header=T)

> seiseki

 math science japanese english social

tanaka 89 90 67 46 50

sato 57 70 80 85 90

suzuki 80 90 35 40 50

honda 40 60 50 45 55

kawabata 78 85 45 55 60

yoshino 55 65 80 75 85

saito 90 85 88 92 95

> seiseki.d<-dist(seiseki)

> round(seiseki.d)

 tanaka sato suzuki honda kawabata yoshino

sato 69

suzuki 34 81

honda 60 64 53

kawabata 28 61 21 47

yoshino 63 12 76 54 56

saito 68 38 88 92 68 46

⇒最短距離12から、ｃ１｛吉野、佐藤｝

最近隣法(nearest)：クラスター間でもっとも近い個体同士の距離をクラスター間距離とする

最遠隣法(furthest)：クラスター間でもっとも遠い個体同士の距離をクラスター間距離とする

群平均法：２つのクラスターのそれぞれの中から１個ずつ個体を選んで個体間の距離を求める。

重心法：クラスターの重心を求め、その重心間の距離をクラスター間距離とする。

メディアン法：２つのクラスターの重心を求めるとき、重みを等しくして計算する。

ウォード法：２つのクラスターを融合する際に、グループ内の分散に対するグループ間の分散を最大化する基準でクラスターを形成していく方法。

Rによる階層的クラスター分析

hclust(d, method=”complete”, …) methodオプションで距離測定法を指定する

　complete（最遠隣法）、single（最近隣法）、average（群平均法）、centroid（重心法）

　median（メディアン法）、ward（ウォード法）、mcquitty(McQuitty法)

hclustのオブジェクトを取る関連関数

　summary（結果のオブジェクトのリストを返す）

　plot（樹系図を作成する）

　plclust（樹系図を作成する）

　cutree（クラスター房の数を指定し、グループ分けする）

　cophenetic（コーフェン行列を返す）

　$merge（クラスター形成の履歴が（個体数－1）行２列）

> summary(sei.hc)

 Length Class Mode

merge 12 -none- numeric

height 6 -none- numeric

order 7 -none- numeric

labels 7 -none- character

method 1 -none- character

call 2 -none- call

dist.method 1 -none- character

> sei.hc$merge

 [,1] [,2]

[1,] -2 -6　　-2（佐藤）と-6（吉野）、マイナスつきは個体番号、符号なしはクラスター番号

[2,] -3 -5　　-3（鈴木）と-5（川端）

[3,] -1 2　　-1（田中）とクラスター２｛鈴木、川端｝がクラスター３を形成

[4,] -7 1

[5,] -4 3

[6,] 4 5

> par(mfrow=c(1,2))

> plot(sei.hc)

> plot(sei.hc, hang=-1) # hang=-1は、歯の高さをそろえる



> plot(hclust(seiseki.d, method="centroid"), hang=-1)

> plot(hclust(dist(seiseki, method="canberra"), method="ward"), hang=-1)

> par(mfrow=c(1,1))



　　ユークリッド距離、重心法　　　キャンベラ距離、ウォード法

※用いる距離、クラスター法によって結果が大きく異なる場合もある

> cutree(sei.hc, 2)　# クラスター数を指定して、各個体がどのクラスターに属するか表示

 tanaka sato suzuki honda kawabata yoshino saito

 1 2 1 1 1 2 2

クラスター分析の結果の妥当性の評価

距離の行列とコーフェン行列との相関係数（コーフェン相関係数）を用いる（西田1992）

> cophenetic(sei.hc) # コーフェン行列

 tanaka sato suzuki honda kawabata yoshino

sato 91.53142

suzuki 33.77869 91.53142

honda 60.13319 91.53142 60.13319

kawabata 33.77869 91.53142 21.30728 60.13319

yoshino 91.53142 12.40967 91.53142 91.53142 91.53142

saito 91.53142 45.58509 91.53142 91.53142 91.53142 45.58509

> cor(seiseki.d, cophenetic(sei.hc)) # 距離行列とコーフェン行列の相関係数

[1] 0.8944869　⇒係数の値が大きいほど、歪みが少ないと判断≠分類結果が妥当

経験的に、ウォード法は妥当と思われる結果を返す確率が高いが、コーフェン相関係数は比較的低い値を返す

文体特性調査による作品クラスター分析

コンラッド　8:19

ドイル　　　30：41

ヘンリ・ジェイムズ　54：63

メルヴィル　68：77

ウッドハウス　90：102

> ConWode<-Zscore.df[c(8:19, 90:102),3:12] # コンラッドとウッドハウス

> ConWode.d <- dist(ConWode)

> ConWode.hc<-hclust(ConWode.d)

> ConWode.hc$merge

 [,1] [,2]

 [1,] -17 -21

 [2,] -14 -16

 [3,] -19 -25

 [4,] -22 -24

 [5,] -10 -12

 [6,] -9 -11

 [7,] -13 -18

 [8,] -20 -23

 [9,] 2 3

[10,] -2 5

[11,] -4 -6

[12,] 7 8

[13,] 1 9

[14,] -1 -7

[15,] -15 4

[16,] 10 14

[17,] -5 -8

[18,] 12 13

[19,] -3 16

[20,] 6 15

[21,] 17 19

[22,] 18 20

[23,] 11 21

[24,] 22 23

> plot(ConWode.hc, hang=-1)



コンラッド作品16と18（だけ）がウッドハウスのクラスターに分類された。

