

2012年7月17日(火)

問題発見技法

6. クラスタ分析

情報学部 堀田敬介

クラスタ分析

- Contents
 - クラスタ分析
 - 1. クラスタ分析概要
 - 2. 類似度の測定
 - 3. クラスタ化の方法の決定(類似度更新法)
 - クラスタ分析[階層的方法]の実施
 - 4. Excelで計算したクラスタ分析, Rによるクラスタ分析
 - 5. クラスタ分析実施上の注意点
 - クラスタ分析[非階層的方法]
 - 6. 非階層的クラスタ分析[K-means法]
 - 7. Rによるクラスタ分析[K-means法]

1. クラスタ分析概要

- クラスタ分析とは？
 - 複数の対象(もの, 変数など)を, その属性によって類似度(similarity)をはかり, 均質な集団(cluster)に分類する方法の総称

どれとどれが似てる？
(同じクラスター？)

1. クラスタ分析概要

- クラスタ分析の種類
 - 階層的方法
 - 樹形図(デンドログラム)を作成
 - 目的により高さを決めてクラスタリング
 - 非階層的方法
 - 予めクラスタ数を決め(or決まっていれば), クラスタリングを行う

例: 3つのクラスターに分類

1. クラスタ分析概要

- 例: 階層的 method (対象の属性が2つの場合)

1. クラスタ分析概要

- どうやって **類似度** を測るのか？

		3	1	2	3	4	6	6
	x_1	1	2	3	5	5	5	3
	3	1						
	1	2						
	2	3						
	3	5						
	4	5						
	6	5						
	6	3						

2. 類似度の測定

類似度は尺度により距離や相関で測る
(距離: 近いほうが類似)
(相関: 高いほうが類似)

距離【間隔尺度】

- ユークリッド距離
- ユークリッド平方距離
- 重み付きユークリッド距離
- マンハッタン距離
- ミンコフスキー距離
- マハラノビス汎距離

距離【名義尺度 {0, 1}】

- 類似比
- 一致係数
- Russel-Rao係数
- Rogers-Tanimoto係数
- Hamann係数
- ファイ係数

相関【間隔尺度】

- Pearsonの積率相関係数
- ベクトル内積

相関【順序尺度】

- Spearmanの順位相関係数
- Kendallの順位相関係数

変量間類似度【名義尺度】

- 平均平方根一致係数
- グッドマン・クラスカルの λ

クラスタ分析ノート.pdf

2. 類似度の測定

- データと尺度

	順序尺度	名義尺度	名義尺度	名義尺度	名義尺度	名義尺度	順序尺度
学術番号	氏名	性別	生年月日	身長	体重	問題発見技法成績	...
1	文教太郎	男	1987.5.6	175cm	69kg	B	...
2	湘南花子	女	1988.1.4	163cm	48kg	AA	...
3	...	...	...	...	...	...	...

量的データ
(数値データ)

比率尺度
比に意味がある(絶対原点が存在する)
例) 身長 180cmのAさんは息子(100cm)の1.8倍背が高い

間隔尺度
差に意味がある
例) 温度 気温20℃より30℃の方が10℃高い

質的データ
(カテゴリデータ)

順序尺度
順序関係がある
例) 成績評価 (A > B > C > D)

名義尺度
単なる分類(区別ができる)
例) 名前, 性別

2. 類似度の測定

- 個体間類似度
 - ユークリッド距離
(cf. l_2 -ノルム)
 - マンハッタン距離
(cf. l_1 -ノルム)
 - ミンコフスキー距離
(cf. l_p -ノルム)
 - マハラノビス汎距離
(cf. l_∞ -ノルム)
 - ユークリッド平方距離

クラスタ分析でよく使われる

(注: 各ノルムとは2変量の差ベクトルに対するノルム)

$l_2(C, D) = 5 = \sqrt{(3-7)^2 + (1-4)^2}$ $l_1(C, D) = 7 = |3-7| + |1-4|$ $l_3(C, D) = 4.498 = \sqrt[3]{|3-7|^3 + |1-4|^3}$ $l_\infty(C, D) = 4 = \max\{|3-7|, |1-4|\}$ $l_2(C, D)^2 = 25 = (3-7)^2 + (1-4)^2$

2. 類似度の測定

- 個体間類似度
 - ユークリッド距離
(cf. l_2 -ノルム)
 - マンハッタン距離
(cf. l_1 -ノルム)
 - ミンコフスキー距離
(cf. l_p -ノルム)
 - マハラノビス汎距離

左側の対象内での、A-B間距離と
右側の対象内でのA-B間距離が
異なる! (ユークリッド距離などでは同じ)

マハラノビス汎距離(2変量 x_1, x_2 版)

$$D \equiv \sqrt{\frac{u_1^2 + u_2^2 - 2\rho u_1 u_2}{1 - \rho^2}}$$

ただし、 u_1, u_2 は x_1, x_2 の標準化変量で、 $u_1 = \frac{x_1 - \mu_1}{\sigma_1}, u_2 = \frac{x_2 - \mu_2}{\sigma_2}$

また、 μ_1, μ_2 はそれぞれ、変量 x_1, x_2 の平均、
 σ_1, σ_2 は x_1, x_2 の標準偏差、 ρ は x_1, x_2 の相関係数

2. 類似度の測定

- どうやって類似度を測るか?
 - 例: ユークリッド平方距離

	3	1	2	3	4	6	6		
x_1	1	2	3	5	5	5	3		
x_2	1	2	3	5	5	5	3		
	3	1		5	5	16	17	25	13
	1	2			2	13	18	34	26
	2	3				5	8	20	16
	3	5					1	9	13
	4	5						4	8
	6	5							4
	6	3							

2. 類似度の測定

- どうやって類似度を更新するか?

	3	1	2	3	4	6	6		
x_1	1	2	3	5	5	5	3		
x_2	1	2	3	5	5	5	3		
	3	1		5	5	16	17	25	13
	1	2			2	13	18	34	26
	2	3				5	8	20	16
	3	5					1	9	13
	4	5						4	8
	6	5							4
	6	3							

2. 類似度の測定

● どうやって類似度を更新するか？

		3	1	2	3,4	6	6	
	x_1	1	2	3	5,5	5	3	
	3	1		5	5	16,17	25	13
	1	2			2	13,18	34	26
	2	3				5,8	20	16
	3,4	5,5				1	9,4	13,8
	6	5						4
	6	3						

3. クラスタ化の方法

● 新たなクラスタ生成時の類似度の更新方法

● クラスタ p , クラスタ q が一つのクラスタ t になる場合、他のクラスタ r との類似度をどう更新する？

1. 最短距離法
2. 最長距離法
3. 群平均法
4. 重心法
5. 中央値法
6. ウォード法

「最短」か「最長」か
何らかの「平均」

(s_{pr} : クラスタ p, r の類似度)

3. クラスタ化の方法

1. 最短距離法 (nearest neighbor method)

[単連結法 (single linkage method)]

$$s_{tr} = \min\{s_{pr}, s_{qr}\}$$

あるクラスタにおいて、クラスタ内の各対象が、そのクラスタ外の任意の対象よりも、そのクラスタ内の少なくとも1つの対象とより近接している。

※類似度は、対象間の類似度の大小関係だけで決まる。よって、類似度(距離)は順序尺度ならばよい。

3. クラスタ化の方法

1. 最短距離法

$$s_{tr} = \min\{s_{pr}, s_{qr}\}$$

3. クラスタ化の方法

2. 最長距離法 (furthest neighbor method)
〔完全連結法 (complete linkage method)〕

$$s_{tr} = \max\{s_{pr}, s_{qr}\}$$

あるクラスタにおいて、クラスタ内の**全て**の対象が、そのクラスタ外の任意の対象との距離よりも**常に近接**している。

※類似度は、対象間の類似度の大小関係だけで決まる。
よって、類似度(距離)は**順序尺度**ならばよい。

3. クラスタ化の方法

2. 最長距離法

$$s_{tr} = \max\{s_{pr}, s_{qr}\}$$

3. クラスタ化の方法

3. 群平均法 (group average method)

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

n_p : クラスタ p に含まれる対象数
 n_q : クラスタ q に含まれる対象数

※類似度は、**間隔尺度**ならばOK

3. クラスタ化の方法

3. 群平均法

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

3. クラスタ化の方法

4. 重心法 (centroid method)

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} s_{pq}$$

n_p : クラスタ p に含まれる対象数
 n_q : クラスタ q に含まれる対象数

※導出過程より, 類似度 S_{ij} はユークリッド平方距離の時のみ妥当. → cf. ファイル「クラスタ分析ノート.pdf」

$\bar{x}_t = \frac{n_p \bar{x}_p + n_q \bar{x}_q}{n_p + n_q}$ ※ x はベクトル

3. クラスタ化の方法

4. 重心法

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} s_{pq}$$

3. クラスタ化の方法

5. 中央値法 (median method)

$$s_{tr} = \frac{1}{2} s_{pr} + \frac{1}{2} s_{qr} - \frac{1}{4} s_{pq}$$

(重心法の簡易版, 重心の代わりに中央値を取る
重心法で $n_p := 1, n_q := 1$ に相当)

※導出過程より, 類似度 S_{ij} はユークリッド平方距離の時のみ妥当. → cf. ファイル「クラスタ分析ノート.pdf」

3. クラスタ化の方法

5. 中央値法

$$s_{tr} = \frac{1}{2} s_{pr} + \frac{1}{2} s_{qr} - \frac{1}{4} s_{pq}$$

3. クラスタ化の方法

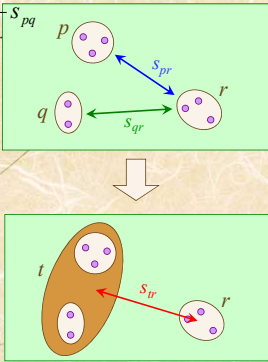
6. ウォード法 (Ward method)

$$S_{tr} = \frac{n_p + n_r}{n_p + n_q + n_r} S_{pr} + \frac{n_q + n_r}{n_p + n_q + n_r} S_{qr} - \frac{n_r}{n_p + n_q + n_r} S_{pq}$$

n_p : クラスタ p に含まれる対象数

n_q : クラスタ q に含まれる対象数

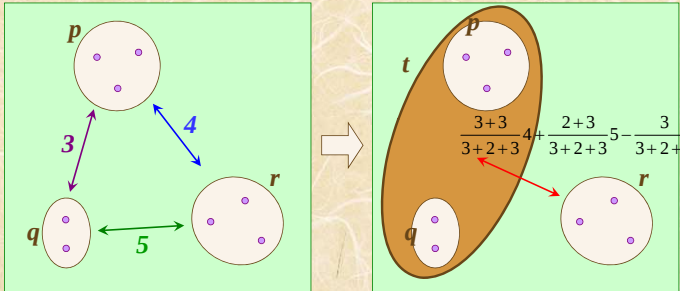
n_r : クラスタ r に含まれる対象数



※導出過程より、類似度 S_{tr} は
ユークリッド平方距離の時のみ妥当。
→ cf. ファイル「クラスタ分析ノート.pdf」

3. クラスタ化の方法

6. ウォード法

$$S_{tr} = \frac{n_p + n_r}{n_p + n_q + n_r} S_{pr} + \frac{n_q + n_r}{n_p + n_q + n_r} S_{qr} - \frac{n_r}{n_p + n_q + n_r} S_{pq}$$


3. クラスタ化の方法

● どうやって類似度を更新するか？

			3	1	2	3:4	6	6	
	x_1	x_2	1	2	3	5:5	5	3	
	3	1		5	5	16:17	25	13	
	1	2			2	13:18	34	26	
	2	3				5:8	20	16	
	3:4	5:5				1	9:4	13:8	
	6	5						4	
	6	3							

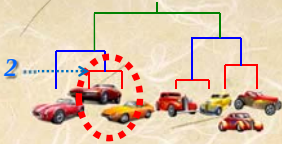
3. クラスタ化の方法

● どうやって類似度を更新するか？

			3	1	2	3:4	$\frac{1+1}{1+1+1}16 + \frac{1+1}{1+1+1}17 - \frac{1}{1+1+1}1$	$\frac{1+1}{1+1+1}16 + \frac{1+1}{1+1+1}17 - \frac{1}{1+1+1}1$	
	x_1	x_2	1	2	3	5:5	$\frac{1+1}{1+1+1}13 + \frac{1+1}{1+1+1}18 - \frac{1}{1+1+1}1$	$\frac{1+1}{1+1+1}13 + \frac{1+1}{1+1+1}18 - \frac{1}{1+1+1}1$	
	3	1		5	5	21.7	25	13	
	1	2			2	20.3	34	26	
	2	3				8.3	20	16	
	3:4	5:5	$\frac{1+1}{1+1+1}5 + \frac{1+1}{1+1+1}8 - \frac{1}{1+1+1}1$			1	8.3	13.7	
	6	5	$\frac{1+1}{1+1+1}9 + \frac{1+1}{1+1+1}4 - \frac{1}{1+1+1}1$					4	
	6	3	$\frac{1+1}{1+1+1}13 + \frac{1+1}{1+1+1}8 - \frac{1}{1+1+1}1$						

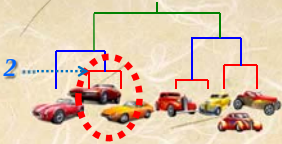
3. クラスタ化の方法

● どうやって類似度を更新するか？



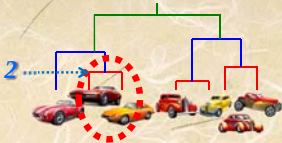
3. クラスタ化の方法

● どうやって類似度を更新するか？



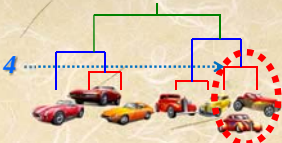
3. クラスタ化の方法

● どうやって類似度を更新するか？



3. クラスタ化の方法

● どうやって類似度を更新するか？



3. クラスタ化の方法

● どうやって類似度を更新するか？

3. クラスタ化の方法

● どうやって類似度を更新するか？

	x_1	x_2
		47.4

4. クラスタ分析の実施

● Excelを用いて計算するクラスタ分析: 例2

- 対象: 5人の学生
- 対象の属性: 7つ

	属性1	属性2	属性3	属性4	属性5	属性6	属性7
太郎	13	12	7	1	13	13	12
次郎	6	5	8	4	9	5	15
三郎	13	14	5	15	2	19	17
四郎	13	5	8	7	9	3	13
五郎	1	18	6	1	3	1	20

● 距離: ユークリッド平方距離

$$l_2(Taro, Jiro)^2 = (13-6)^2 + (12-5)^2 + \dots + (12-15)^2$$

● クラスタ間の類似度更新方法: 群平均法

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

4. クラスタ分析の実施

● Excelで計算によるクラスタ分析: 例2

	属性1	属性2	属性3	属性4	属性5	属性6	属性7
太郎	13	12	7	1	13	13	12
次郎	6	5	8	4	9	5	15
三郎	13	14	5	15	2	19	17
四郎	13	5	8	7	9	3	13
五郎	1	18	6	1	3	1	20

↓ 類似度の測定: ユークリッド平方距離による

$$l_2(Taro, Jiro)^2 = 197 = (13-6)^2 + \dots + (12-15)^2$$

	太郎	次郎	三郎	四郎
太郎				
次郎	197			
三郎	386	509		
四郎	203	66	475	
五郎	489	284	691	442

↓ 類似度の更新: 群平均法による

	太郎	次&四	三郎
次&四	200		
三郎	386	492	
五郎	489	363	691

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

4. クラスタ分析の実施

● Excelで計算によるクラスタ分析: 例2

	太郎	次&四	三郎
次&四	200		
三郎	386	492	
五郎	489	363	691

↓ 類似度の更新: 群平均法による

	太郎	次&四	三郎
次&四	200		
三郎	456.67	405	
五郎	489	405	691

↓ 類似度の更新: 群平均法による

	五&(太&次&四)	三郎
五&(太&次&四)		
三郎	515.25	

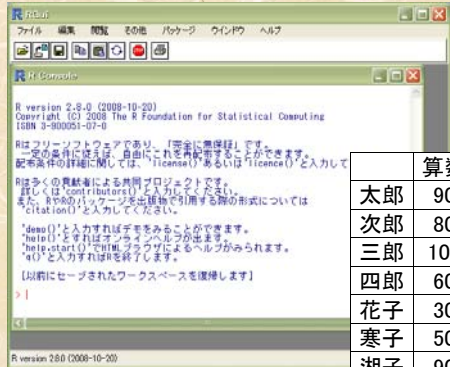
$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

樹形図(デンドログラム)

4. クラスタ分析の実施

- Rによるクラスタ分析:1. 起動画面とデータファイル

R起動時画面



データをcsvファイルで用意
(Excelやeditorで作成)

ファイル「data-seiseki.csv」

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70

4. クラスタ分析の実施

- Rによるクラスタ分析:2. クラスタ分析の実施例

```
> seiseki <- read.csv("C:/data-seiseki.csv", header=T, row.names=1)
> seiseki
```

← csvファイルを読み込み、変数seisekiに格納

← 変数seisekiの中身確認

```
太郎  算数 理科 国語 英語 社会
次郎  80  60  70  70  20
三郎  100  40  30  70  80
四郎  60  30  40  80  80
花子  30  60  80  90  90
寒子  50  60  40  30  60
湘子  90 100  90  80  70
```

```
> seiseki.d <- dist(seiseki, "manhattan")
> seiseki.d
```

← 対象間の類似度をmanhattan距離で測定し、変数seiseki.dに格納

← 変数seiseki.dの中身確認

```
次郎  太郎  次郎  三郎  四郎  花子  寒子
80      180  140
三郎  190  150  70
四郎  170  150  170  120
花子  200  140  140  110  150
寒子  70  130  150  160  140  190
湘子  70  130  150  160  140  190
```

```
> (seiseki.hc <- hclust(seiseki.d, "ward"))
```

← ward法でクラスタ分析を実施し、変数seiseki.hcに格納

Call: hclust(d = seiseki.d, method = "ward")

Cluster method : ward
Distance : manhattan
Number of objects : 7

← クラスタ化: ward法
← 類似度: manhattan距離
← 対象の数: 7

```
> plot(seiseki.hc, hang=-1)
```

← 結果を樹形図で表示

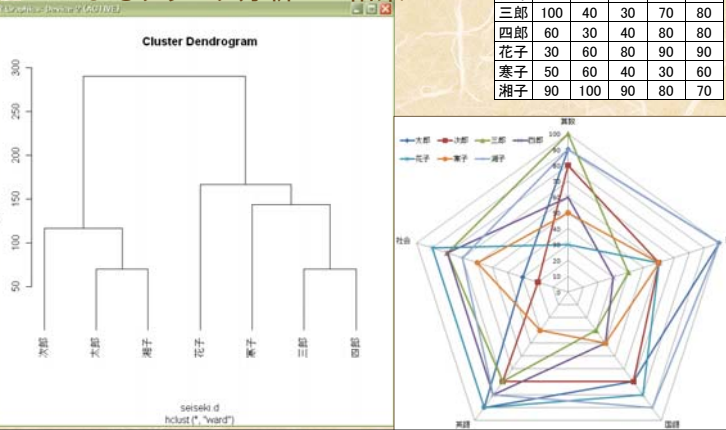
注) ward法を用いる場合、距離はユークリッド平方距離を使うのが妥当

4. クラスタ分析の実施

- Rによるクラスタ分析:3. 結果

cf. 元データ

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70



4. クラスタ分析の実施

- Rによるクラスタ分析:4.手法選択について

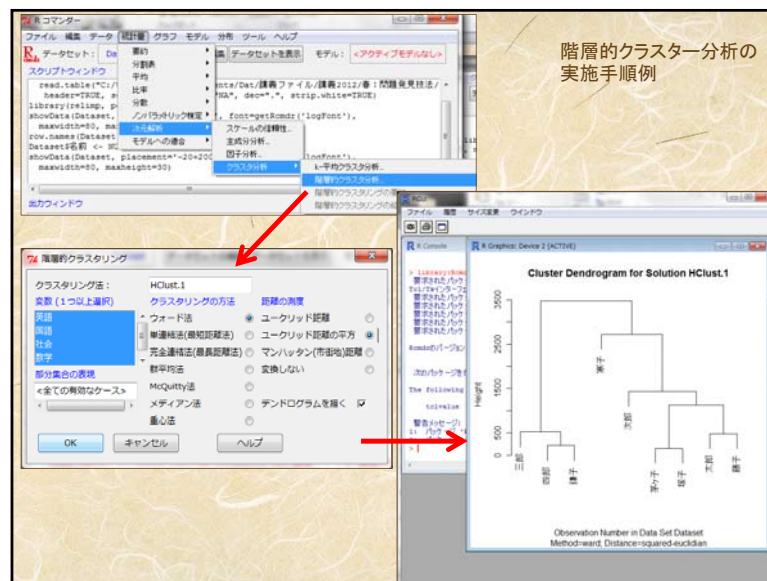
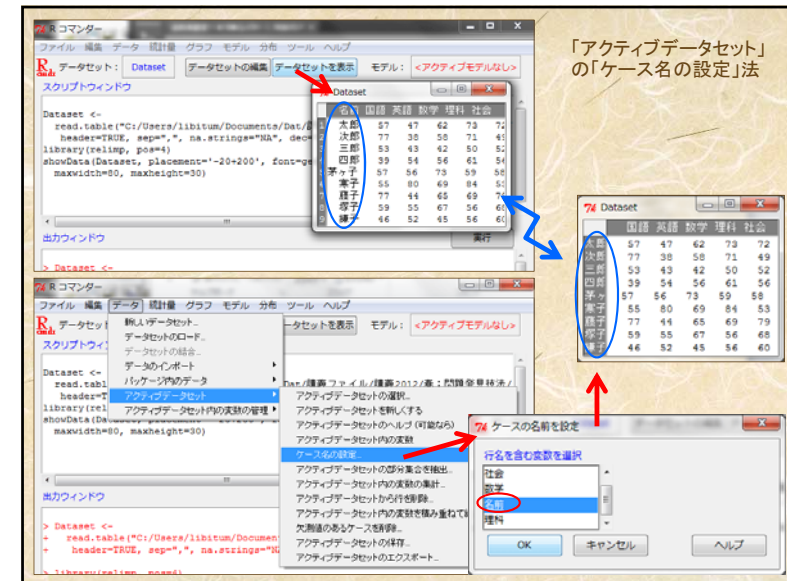
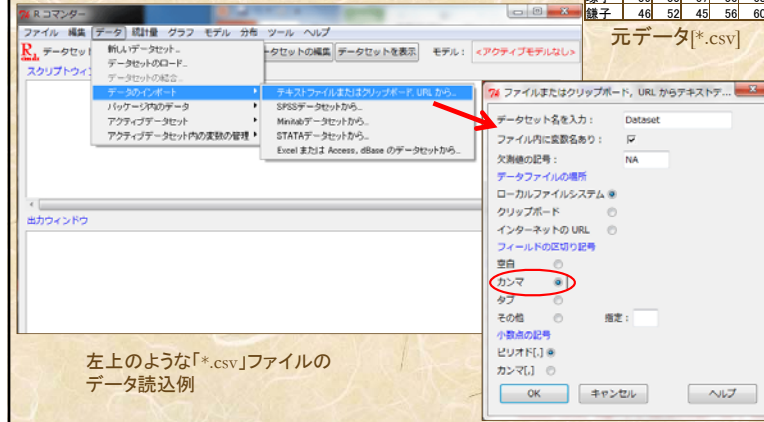
- 距離の測定:関数dist() 【書式:dist(data, "method")】
 - methodの部分に距離の測定方法を指定
 - euclidean ... ユークリッド距離 (l_2 ノルム) ex) dist(data) ← 指定無しだとこれ
 - manhattan ... マンハッタン距離 (l_1 ノルム) ex) dist(data, "manhattan")
 - minkowski ... ミンコフスキー距離 (l_p ノルム) ex) dist(data, "minkowski", p=4)
 - maximum ... l_∞ ノルム ex) dist(data, "maximum")
- 注) ユークリッド平方距離は、ユークリッド距離の計算後、2乗する
- クラスタ化の方法:関数hclust() 【書式:hclust(data.d, "method")】
 - methodの部分にクラスタ化の方法を指定
 - single ... 最短距離法 ex) hclust(data.d, "single")
 - complete ... 最長距離法 ex) hclust(data.d, "complete")
 - average ... 群平均法 ex) hclust(data.d, "average")
 - centroid ... 重心法 ex) hclust(data.d^2, "centroid")
 - median ... 中央値法 ex) hclust(data.d, "median")
 - ward ... ウォード法 ex) hclust(data.d^2, "ward")

注) この2つの手法では「ユークリッド平方距離」を用いる (data.dがユークリッド距離の計算結果でその2乗を使用)

4. クラスタ分析の実施

- R commanderによるクラスタ分析

名前	国語	英語	数学	理科	社会
太郎	57	47	62	73	72
次郎	77	38	58	71	49
三郎	53	43	42	50	52
四郎	39	54	56	61	56
茅ヶ子	57	56	73	59	58
寒子	55	80	69	84	53
藤子	77	44	65	69	79
塚子	59	55	67	56	68
鯉子	46	52	45	56	60



5. クラスター分析実施上の注意点

- クラスター分析の長所
 - 探索的手法なので、データ構造を事前に知らなくてよい
 - あらゆる種類のデータに適用可能: 数値・カテゴリー
 - 適用が簡単
-
- クラスター分析の短所
 - 迷ったらとりあえず「ユークリッド平方距離」で
 - どんな属性値を選んだらいいのか?
 - どの類似度(距離)測定法を選んだらいいのか?
 - どのクラスタ化更新法を選んだらいいのか?
 - 結果の解釈が困難な可能性がある
 - 迷ったらとりあえず

6. 非階層的クラスタ分析

● K-means法

● 事前にクラスタ数をKとしてクラスタリングを行う



例: 3つのクラスタ(K=3)に分類したい! としよう

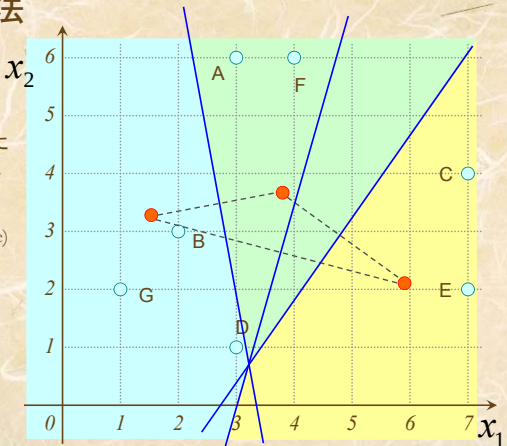
6. 非階層的クラスタ分析

● K-means法

Step0: Kを決める
(ex. $K:=3$)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.
(ex. Euclidean distance)
(cf. Voronoi diagrams)



6. 非階層的クラスタ分析

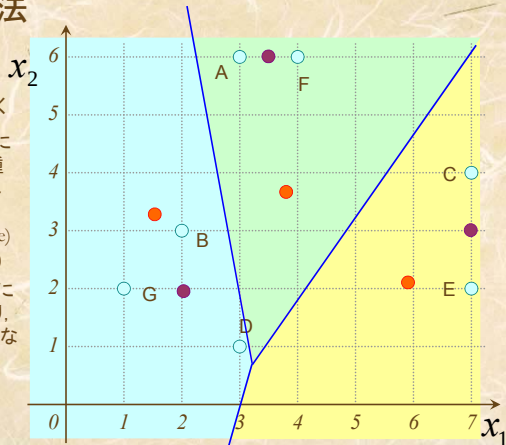
● K-means法

Step0: Kを決める
(ex. $K:=3$)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.
(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスタごとに何らかの距離により、重心を計算し、新たな種とする.



6. 非階層的クラスタ分析

● K-means法

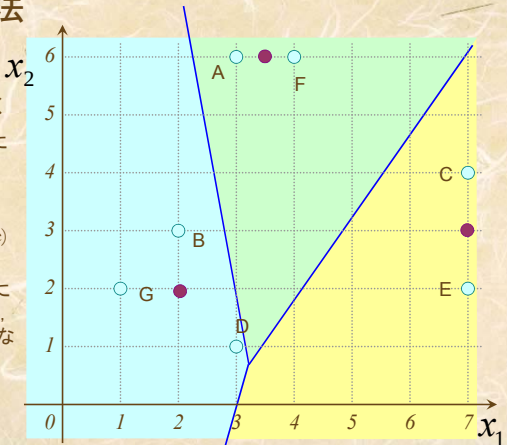
Step0: Kを決める
(ex. $K:=3$)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.
(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスタごとに何らかの距離により、重心を計算し、新たな種とする.

Step2-4をクラスタが更新されなくなるまで繰り返す



6. 非階層的クラスタ分析

- K-means法

Step0: Kを決める
(ex. K:=3)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける。
(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする。

Step2-4をクラスが更新されなくなるまで繰り返す

6. 非階層的クラスタ分析

- K-means法

Step0: Kを決める
(ex. K:=3)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける。
(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする。

Step2-4をクラスが更新されなくなるまで繰り返す

7. クラスタ分析の実施

- Rによるクラスタ分析: 4. K-means法による結果

```
> (seiseki.km <- kmeans(seiseki, 3))  
K-means clustering with 3 clusters of sizes 1, 3, 3  
  
Cluster means:  
  算数  理科  国語  英語  社会  
1 30.00000 60.00000 80.00000 80 90.00000  
2 70.00000 43.33333 36.66667 60 73.33333  
3 86.66667 86.66667 76.66667 80 40.00000  
  
Clustering vector:  
太郎 次郎 三郎 四郎 花子 寒子 湘子  
  3   3   2   2   1   2   3  
  
Within cluster sum of squares by cluster:  
[1] 0 3600 3000  
  
Available components:  
[1] "cluster" "center" "withinss" "size"
```

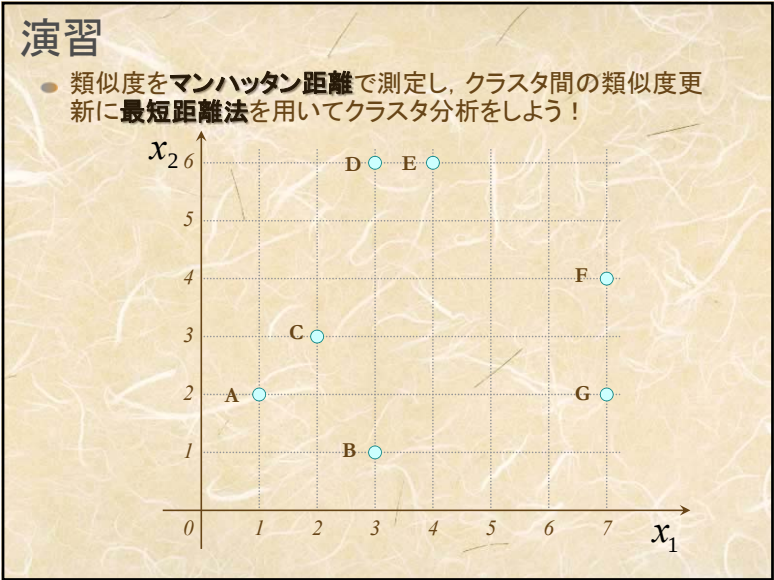
結果:

- cluster1: 花子
- cluster2: 三郎, 四郎, 寒子
- cluster3: 太郎, 次郎, 湘子

cf. 元データ

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70

R commander による
非階層的クラスタ分析
(K-means法)の実施手順例



参考文献

- 田中豊・脇本和昌『多変量統計解析法』現代数学社(1983)
- 河口至商『多変量解析入門Ⅱ』森北出版(1978,2005)
- 金明哲『Rによるデータサイエンス』森北出版(2007)
- 荒木孝治『RとRコマンドーではじめる多変量解析』日科技連(2007)
- 新納浩幸『Rで学ぶクラスタ解析』オーム社(2007)