

2007年6月19日

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

1. クラスタ分析概要

2. 類似度の測定

3. クラスタ間の近さの決定

4. クラスタ分析の実施[SPSS, 手計算]

5. クラスタ分析実施上の注意点

6. 演習: やってみよう!

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,

類似度(similarity)を定義し,

均質な集団(cluster)に分類する

方法の総称

1. クラスタ分析概要

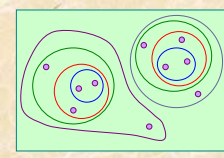
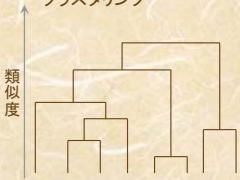
クラスタ分析の種類

階層的な方法

- 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

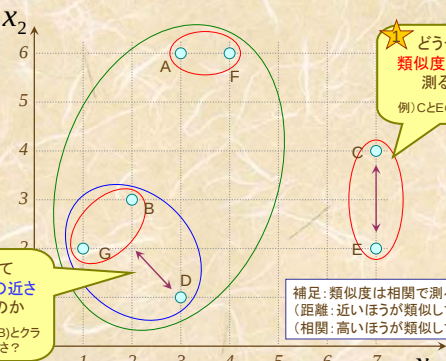
非階層的な方法

- 予めクラスタ数を決め(or決まっています), クラスタリングを行う



1. クラスタ分析概要

例: x_2



★ どうやって クラスタ間の近さを決めるのか
例) クラスタ(G, B)とクラスタ(D)の近さ?

★ どうやって 類似度(距離)を測るのか
例) CとEの類似度?

補足: 類似度は相関で測る場合もある
(距離: 近いほうが類似している)
(相関: 高いほうが類似している)

2. 類似度の測定

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

2. 類似度の測定

データと尺度

学籍番号	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	...	...	...	...	...	...	...

比率尺度

比に意味がある(絶対原点が存在)
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

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
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) (cf. l_∞ -ノルム)
- マハラノビス汎距離
- ユークリッド平方距離

クラスター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)

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) (cf. l_∞ -ノルム)
- マハラノビス汎距離

マハラノビス汎距離 (2変量 x_1, x_2 版)

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

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

3. クラスタ化の方法の決定

新たなクラスタ生成時の近さの決定

- 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 (距離) はどうなる?

(s_{pr} : クラスタ p, r の類似度 [距離])

3. クラスタ化の方法の決定

クラスタ間の近さ決定方法

(事前にクラスタ数を決める必要はない方法群)

- 最短距離法 (nearest neighbor method)
- 最長距離法 (furthest neighbor method)
- 群平均法 (group average method)
- 重心法 (centroid method)
- 中央値法 (median method)
- ウォード法 (Ward method)

3. クラスタ化の方法の決定

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一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. クラスタ化の方法の決定

1. 最短距離法

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

3. クラスタ化の方法の決定

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. クラスタ化の方法の決定

2. 最長距離法

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

3. クラスタ化の方法の決定

3. 群平均法 (group average method)

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

(n_p : クラスタ p に含まれる対象数)

※類似度は、間隔尺度ならばよい。

3. クラスタ化の方法の決定

3. 群平均法

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

3. クラスタ化の方法の決定

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)

※導出過程 (tex-file参照) より、類似度 S_{tr} はユークリッド平方距離の時のみ妥当。

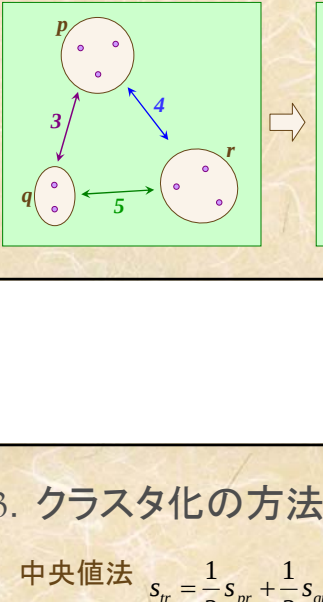
$$\bar{x}_t = \frac{n_p \bar{x}_p + n_q \bar{x}_q}{n_p + n_q}$$

※ x はベクトル

3. クラスタ化の方法の決定

4. 重心法

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

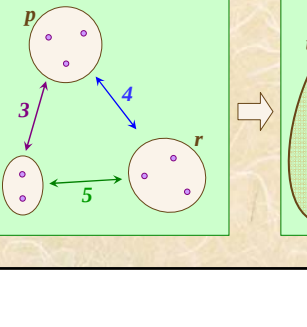


3. クラスタ化の方法の決定

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, 重心ではなく中央値を取る。
よって, 重心法で $n_p := 1, n_q := 1$ に相当する)
※導出過程(重心法参照)より, 類似度 s_{tr} は
ユークリッド平方距離の時のみ妥当。



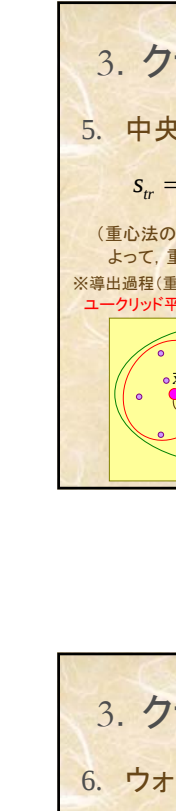
$$\bar{x}_t = \frac{\bar{x}_p + \bar{x}_q}{2}$$

※ x はベクトル

3. クラスタ化の方法の決定

5. 中央値法

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



3. クラスタ化の方法の決定

6. ウォード法 (Ward method)

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

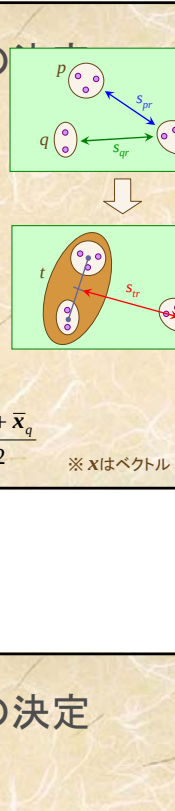
※導出過程 (tex-file参照) より, 類似度 s_{tr} は
ユークリッド平方距離の時のみ妥当。



3. クラスタ化の方法の決定

6. ウォード法

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

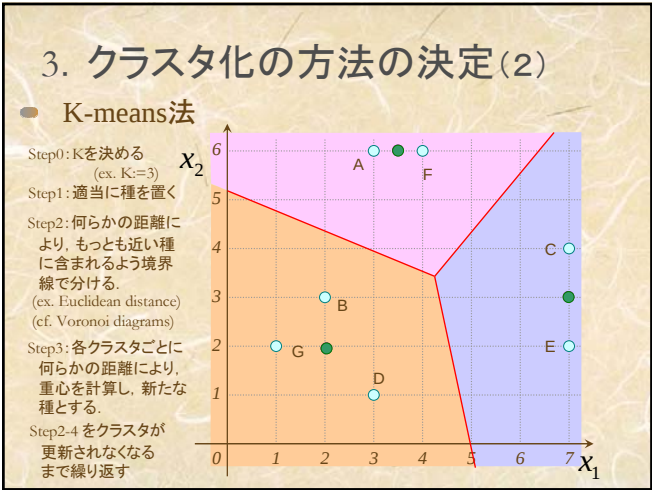
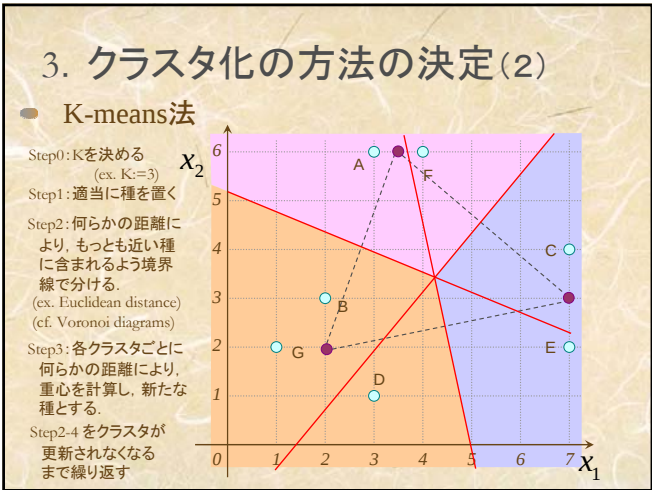
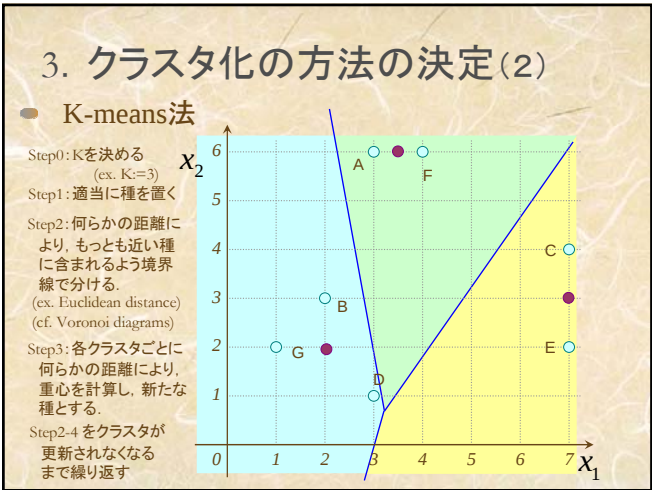
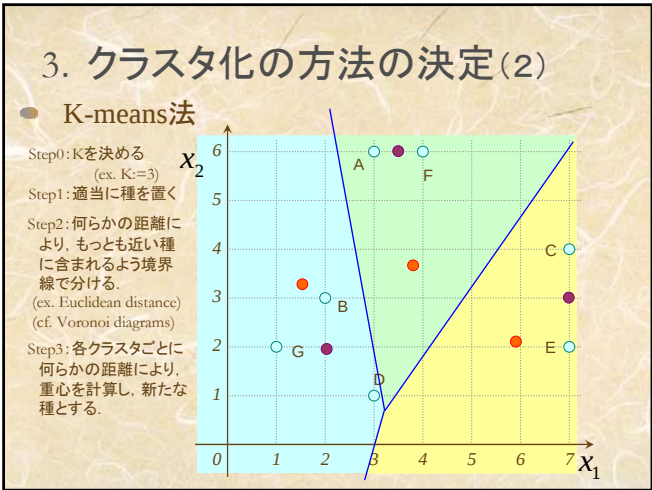
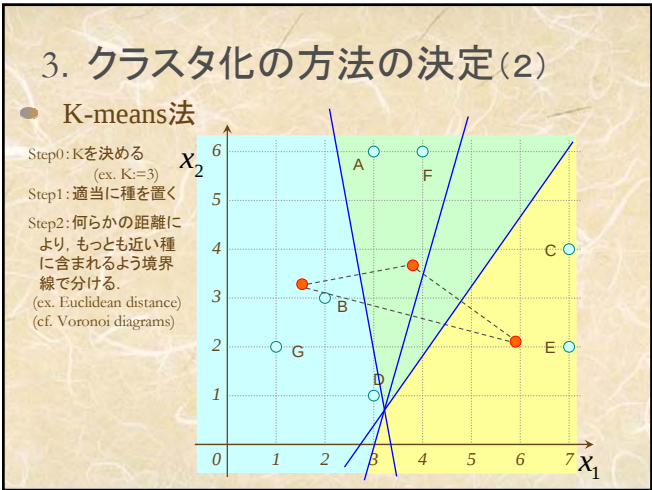


3. クラスタ化の方法の決定(2)

● クラスタ間の近さ決定方法(非階層的 method)

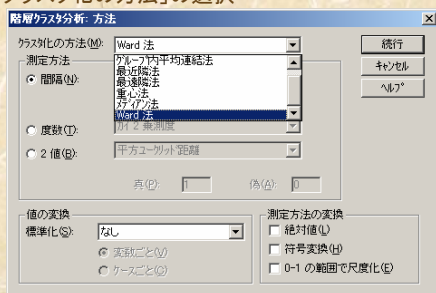
● K-means 法

● 事前にクラスタ数を K としてクラスタリングを行う。



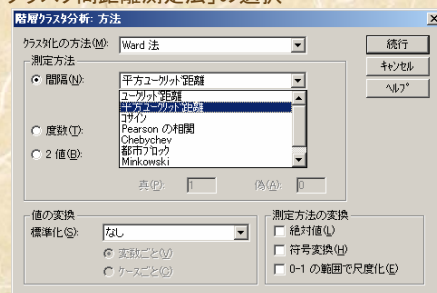
4. クラスタ分析の実施〔SPSS〕

- 統計パッケージ SPSS によるクラスタ分析
- 「クラスタ化の方法」の選択



4. クラスタ分析の実施〔SPSS〕

- 統計パッケージ SPSS によるクラスタ分析
- 「クラスター間距離測定法」の選択



4. クラスタ分析の実施〔例題〕

ある部屋に次の8人(①～⑧)が図7-1の様に座った。座り方は任意なので、似たもの同士、親しい同士ほど接近して座るものと思われる。座った位置のデータは表7-1である。このデータをもとにして8人を分類しデンドログラムをつくる。



表7-1

No	よこ座標	よこ座標
1	4	1
2	2	4
3	5	7
4	5	2
5	3	4
6	6	5
7	2	6
8	7	2

8人相互間のユークリッド平方距離を表7-2から求める。

$$D_{ij} = \sum_{k=1}^l (x_{ki} - x_{kj})^2$$

表7-2

	②	③	④	⑤	⑥	⑦	⑧
①	13	37	2	10	20	29	10
②		18	13	1	17	4	29
③			25	13	5	10	29
④				8	10	25	4
⑤					10	5	20
⑥						17	10
⑦							39



値が小さいほど
類似性を表す

最短距離法で更新！！

表7-3

	②	③	④	⑤	⑥	⑦	⑧
①	10	37	2	20	29	10	
②		13	8	10	4	20	
③			25	5	10	29	
④				10	25	4	
⑤					10	5	20
⑥						17	10
⑦							39

表7-4

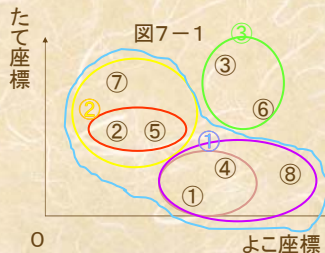
	②	③	④	⑤	⑥	⑦	⑧
①	8	25	10	25	4		
②		13	10		20		
③			5	10	29		
④					17	10	
⑤							39

更新

更新

更新

更新



	②	③	⑥
①	8	25	10
②		13	10
③			5

表7-5

	②	③	③
①	8	10	
②		10	
③			10

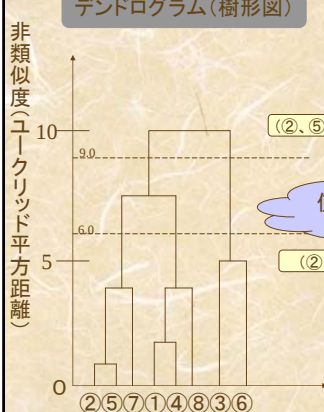
表7-6

	③
①	10

表7-7

デンドログラム(樹形図)

各クラスターの非類似度は、
表7-2～7-7でマークした値である。



使用目的により高さを
変えて判断

樹の高さを変えることにより
クラスターの数も異なる

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: データの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
 - 類似度(距離)測定法の選択が困難の可能性
 - クラスタ化法の選択が困難の可能性
 - 非階層的手法の場合, 事前に決定するクラスタ数の決定が困難の可能性
 - 結果の解釈が困難の可能性

6. やってみよう!

- 演習: 類似度をユークリッド平方距離, クラスタ化を最短距離法でクラスタ分析しよう!



参考文献

- 田中豊ほか『多変量統計解析法』現代数学社
- 河口至商『多変量解析入門Ⅱ』森北出版
- 浅利英吉ほか『パソコンによるデータマイニング』日刊工業新聞社
- 東大教養学部統計学教室編『統計学入門』東京大学出版会
- M.J.A.ベリーほか『データマイニング手法』海文堂