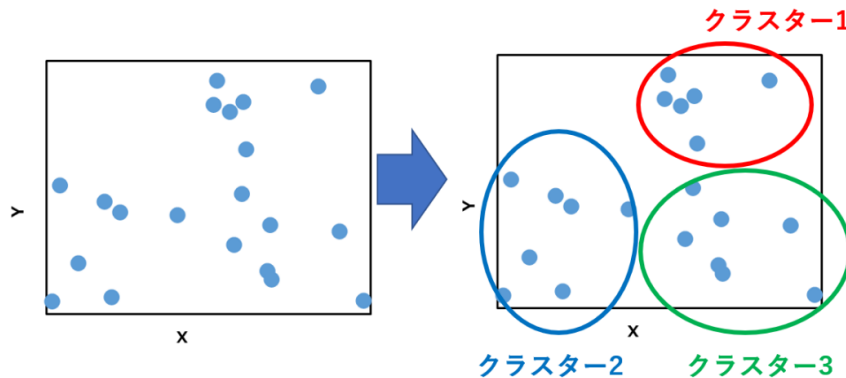


クラスター分析を最短距離法、最長距離法、群平均法の3手法の違いを理解しながらマスターできる！

【1】クラスター分析とは

(1) データをクラスターで分ける

データ群をある規模のクラスターで分類することですね。下図のようなイメージが簡単にできますよね！



ここで問題になるのが、**どうやってクラスターに分類するの？**

(2) クラスター分析の主な2つの手法

よくあるのが、

1. 階層的方法(手計算で考えて解く方法)←これを解説！
2. 非階層的方法(計算機で解く方法)

本来は、非階層的方法で、計算機とプログラムを使って解きたいですが、**何を解いているかがわからないので、手計算で理解できる階層的方法を使ってクラスター分析を理解しましょう。**

【2】最短距離法、最長距離法、群平均法とは

階層的方法はさらに3つの方法に分類できます。**比較しながら3手法をマスターしましょう！**

1. 最短距離法(最も基本的)
2. 最長距離法
3. 群平均法

まずは、最短距離法でクラスター分類して、結果を可視化して納得いけばいいですが、結果がいまいちな場合は、最長距離法、群平均法を使っていきます。

(1) 最短距離法

クラスターに含まれる対象の対の中で、**最短距離なもの**を選びます。式で書くと

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

「min」から最短とわかればOKです。

(2) 最長距離法

クラスターに含まれる対象の対の中で、**最長距離なもの**を選びます。式で書くと

$$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

「max」から最長とわかればOKです。

(3) 群平均法

最短でも最長でもなく、平均的な値で定義したい場合に使います。式で書くと

$$d(C_i \cup C_j, C_k) = \frac{n_i \times d(C_i, C_k) + n_j \times d(C_j, C_k)}{n_i + n_j}$$

「平均」を計算しているとわかればOKです。

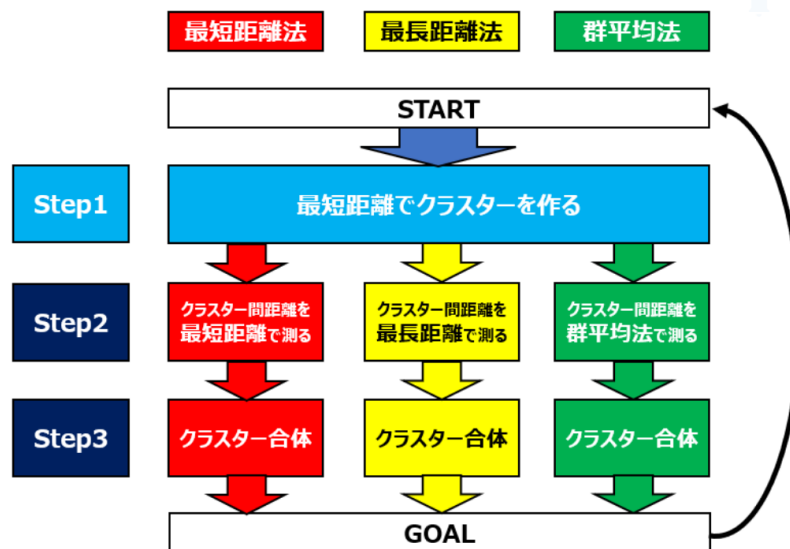
【3】 クラスタ分析の解法

(1) 共通の解き方(最短距離法、最長距離法、群平均法)

3つのステップがあります。

1. 全手法とも、最初は最短距離なペアでクラスターを作る
2. 手法別にクラスター間距離を計算
3. クラスタを合体

の3ステップを全データが分類し終わるまで繰り返します。



特に注意が必要なのは、

最長距離法、群平均法でも、最初は最短距離なペアを見つける点に注意しましょう。

(2) データ事例

【事例】5つのデータがあり、それぞれの距離がわかっている。

- (1) 最短距離法
- (2) 最長距離法
- (3) 群平均法

を使って、それぞれクラスター分析せよ。

-	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

【4】最短距離法、最長距離法、群平均法を比較しながら解く

分類は3回実施しますので、丁寧に解説します。

(1) 分類1回目

①1回目 step1

最短距離なペアを見つけましょう。CとEの14.1が最短ですね。見ればわかる！

最短距離法					
step1 クラスタ選び 最短					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

最長距離法					
step1 クラスタ選び 最短					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

群平均法					
step1 クラスタ選び 最短					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

②1回目 step2

CEが1つのクラスターになったので、

- AとCEクラスター
- BとCEクラスター
- DとCEクラスター

との距離を最短距離法、最長距離法、群平均法で解きます。

最短距離法					
step2 CEとの距離(最短)					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

最長距離法					
step2 CEとの距離(最長)					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	20	51	-	-	-
D	31.6	28.3	42.4	-	-
E	31.6	63.2	14.1	56.6	-

群平均法					
step2 CEとの距離(群平均)					
	A	B	C	D	E
A	-	-	-	-	-
B	31.6	-	-	-	-
C	25.8	57.1	-	-	-
D	31.6	28.3	49.5	-	-
E	25.8	57.1	49.5	-	-

●最短距離法では、

- ・AとCEクラスター⇒ 黄色の20と31.6から20を選択
- ・BとCEクラスター⇒ 緑色の51と63.2から51を選択
- ・DとCEクラスター⇒ 青色の42.4と56.6から42.4を選択

●最長距離法では、

- ・AとCEクラスター⇒ 黄色の20と31.6から31.6を選択
- ・BとCEクラスター⇒ 緑色の51と63.2から63.2を選択
- ・DとCEクラスター⇒ 青色の42.4と56.6から56.6を選択

●群平均では、

- ・AとCEクラスター⇒ 黄色の20と31.6から平均 $1/2 \times (20+31.6) = 25.8$ を選択
- ・BとCEクラスター⇒ 緑色の51と63.2から平均 $1/2 \times (51+63.2) = 57.1$ を選択
- ・DとCEクラスター⇒ 青色の42.4と56.6から平均 $1/2 \times (42.4+56.6) = 49.5$ を選択

③1回目 step3

step2の計算結果を反映します。

最短距離法 step3 クラスター合体(最短)				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	20	51	-	-
D	31.6	28.3	42.4	-

最長距離法 step3 クラスター合体(最長)				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	31.6	63.2	-	-
D	31.6	28.3	56.6	-

群平均法 step3 クラスター合体(群平均)				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	25.8	57.1	-	-
D	31.6	28.3	49.5	-

ここで1回目が終了です。3手法の違いが見えましたね。2回目も同様に解けます！

(2) 分類 2回目

①2回目 step1

最短距離なペアを見つけましょう。

- 最短距離法では、A-CE 間の 20
- 最長距離法では、B-D 間の 28.3
- 群平均法では、A-CE 間の 25.8

が最短ですね。見ればわかるけど、候補と距離の数字が手法によって変わっていますね。

最短距離法 step1 クラスタ選び 最短				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	20	51	-	-
D	31.6	28.3	42.4	-

最長距離法 step1 クラスタ選び 最長				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	31.6	63.2	-	-
D	31.6	28.3	56.6	-

群平均法 step1 クラスタ選び 最短				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	25.8	57.1	-	-
D	31.6	28.3	49.5	-

② 2回目 step2

- 最短距離法では、ACE クラスターと B,D との距離
- 最長距離法では、A と CE クラスターと BD クラスターとの距離
- 群平均法では、ACE クラスターと B,D との距離

との距離を最短距離法、最長距離法、群平均法で解きます。ここが一番難しい所です！

最短距離法 step2 A,CEとの距離(最短)				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	20	51	-	-
D	31.6	28.3	42.4	-

最長距離法 step2 BDとの距離(最長)				
	A	B	CE	D
A	-	-	-	-
B	31.6	-	-	-
CE	31.6	63.2	-	-
D	31.6	28.3	56.6	-

群平均法 step2 A,CEとの距離(群平均)				
	A	B	CE	D
A	-	-	-	-
B	48.6	-	-	-
CE	25.8	48.6	-	-
D	43.5	28.3	43.5	-

- 最短距離法では、
 - ・ACE クラスターと B⇒ 橙色の 31.6 と 51 から 31.6 を選択
 - ・ACE クラスターと D⇒ 緑色の 31.6 と 42.4 から 31.6 を選択
- 最長距離法では、
 - ・A と BD クラスター⇒ 灰色の 31.6 と 31.6 から 31.6 を選択
 - ・BD クラスターと CE クラスター⇒ 紫色の 63.2 と 56.6 から 63.2 を選択
- 最短距離法では、
 - ・ACE クラスターと B⇒ 橙色から $\frac{1}{3} \times 31.6 + \frac{2}{3} \times 57.1 = 48.6$ を選択
 - ・ACE クラスターと D⇒ 緑色の $\frac{1}{3} \times 31.6 + \frac{2}{3} \times 49.5 = 43.5$ を選択

③ 2 回目 step3

step2 の計算結果を反映します。

最短距離法 step3 クラスター合体(最短)				
	ACE	B	D	
ACE	-	-	-	
B	31.6	-	-	
D	31.6	28.3	-	

最長距離法 step3 クラスター合体(最長)				
	A	BD	CE	
A	-	-	-	
BD	31.6	-	-	
CE	31.6	62.2	-	

群平均法 step3 クラスター合体(群平均)				
	ACE	B	D	
ACE	-	-	-	
B	48.6	-	-	
D	43.5	28.3	-	

ここで 2 回目が終了です。3 手法の違いが見えましたね。3 回目も同様に解けます！

(3) 分類 3 回目

①3 回目 step1

最短距離なペアを見つけましょう。

- 最短距離法では、B-D 間の 28.3
- 最長距離法では、A-CE 間の 31.6
- 群平均法では、B-D 間の 28.3

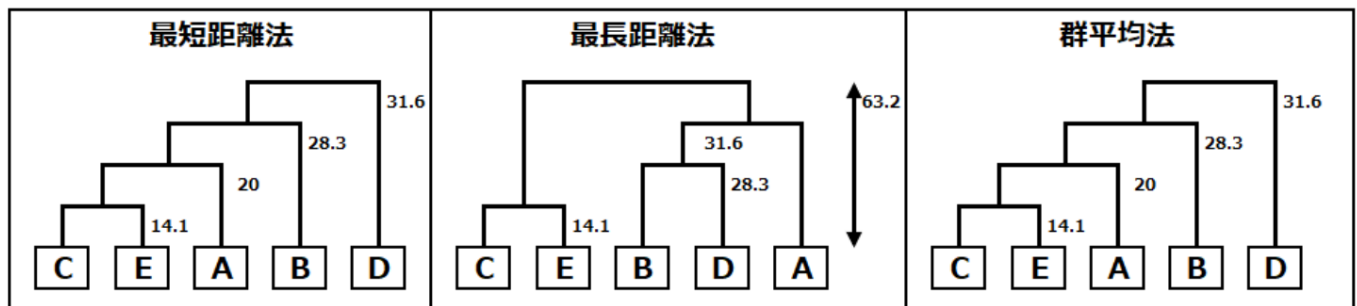
が最短ですね。見ればわかるけど、候補と距離の数字が手法によって変わっていますね。

最短距離法 step1 クラスタ選び 最短				
	ACE	B	D	
ACE	-	-	-	
B	31.6	-	-	
D	31.6	28.3	-	

最長距離法 step1 クラスタ選び 最短				
	A	BD	CE	
A	-	-	-	
BD	31.6	-	-	
CE	31.6	62.2	-	

群平均法 step1 クラスタ選び 最短				
	ACE	B	D	
ACE	-	-	-	
B	31.6	-	-	
D	31.6	28.3	-	

で、ここで、分類が完了したので、結果を比較すると



となりました。手法間で結果が異なりますが、実データと比較してどれを使うかを吟味すれば OK です。

以上、「クラスター分析ができる」を解説しました。