

クラスター分析

富島 諒

2022 年 2 月 15 日

1 クラスター分析

クラスター分析とは

クラスターとは, ”群れ”や”集団”という意味を持つ. そして, クラスター分析とは, 与えられたデータを”似たものどうしの群れに分ける方法”である. クラスター分析ではデータのことを”個体”と呼び, 個体と個体とが集まって, クラスターを構成することになる.

しかし, このままではクラスターに分類する基準が曖昧であるため, ”似ている”とは何かを数学的に定義する必要がある. そこでまず, ”似ている程度”を測る方法として, 以下のようなものがあげられる.

- ユークリッド距離
- ユークリッド距離の 2 乗
- マハラノビスの距離
- 相関係数

これらの方法は, 距離の概念を一般化したものと考えられるので, これらを広い意味で”距離”と呼ぶこととする.

クラスター間の距離

分析の際, ”2 つのクラスター間の距離 D をどのように決めるか”という問題が発生する. もし, 各クラスターの成分が 1 個だけならば, 個体間の距離をそのまま D とすればよい. では, 各クラスターの成分が 2 個以上から成る場合は, どのように距離 D を測ればよいだろうか? (図 1)

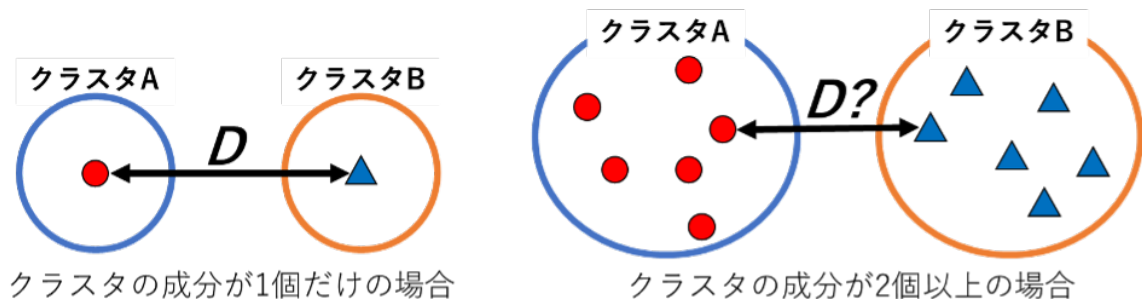


図 1: 成分の個数によるクラスター分析の違い

この”2 つのクラスター間の距離 D の決め方”には, 多くの方法が存在しており, §2 では, そのうち

の 6 つの手法について説明する.

2 クラスタ間の距離の決め方

最短距離法

クラスタ A の個体とクラスタ B の個体とのすべての組み合わせについて距離を求め、その中で最も短い距離をクラスタ間の距離 D と定義する. (図 2)

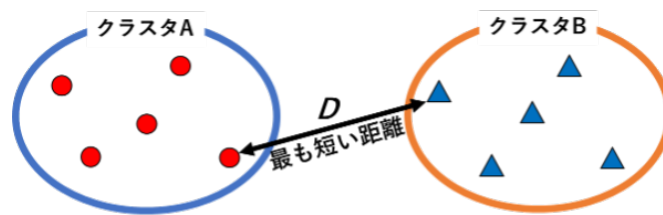


図 2: 最短距離法

最長距離法

クラスタ A の個体とクラスタ B の個体とのすべての組み合わせについて距離を求め、その中で最も長い距離をクラスタ間の距離 D と定義する. (図 3)

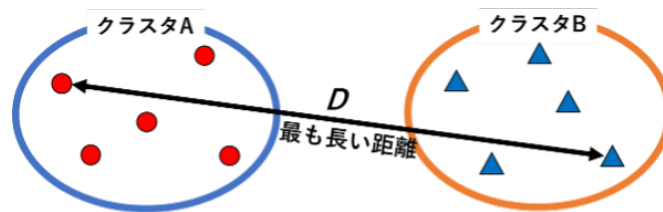


図 3: 最長距離法

群平均法

クラスタ A の個体とクラスタ B の個体との全ての組み合わせについて距離を求め、その距離の平均値をクラスタ間の距離 D と定義する. (図 4)

メディアン法

クラスタ A の個体とクラスタ B の個体との全ての組み合わせについて距離を求め、その距離を順番に並べたときの中央値 (メディアン) をクラスタ間距離 D と定義する. (図 5)

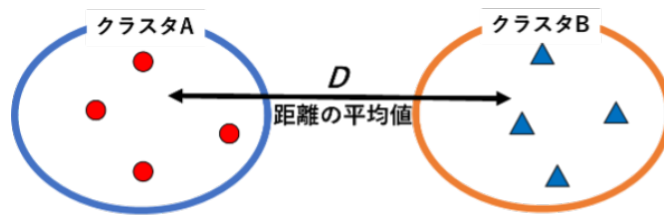


図 4: 群平均法

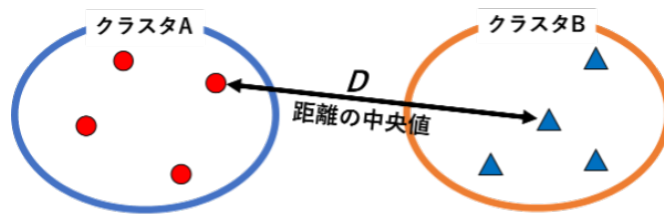


図 5: メディアン法

重心法

クラスタ A の重心とクラスタ B の重心との距離を、クラスタ間距離 D と定義する. (図 6)

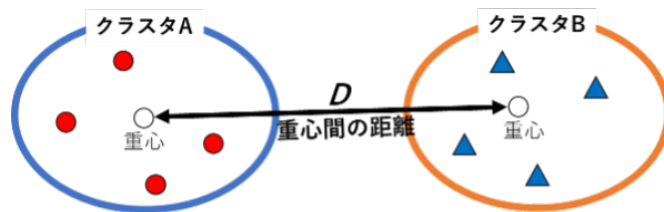


図 6: 重心法

ワード法

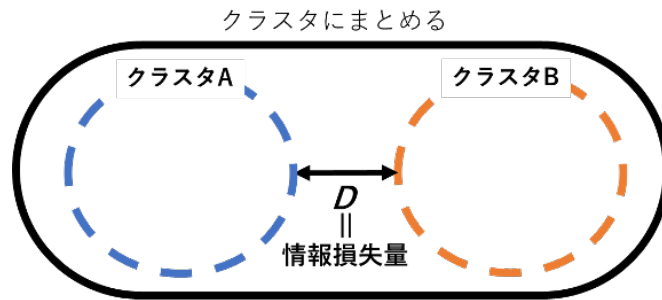
例えば, シャムネコとペルシャネコをまとめてネコたちと呼んでしまうと, もともとどんなネコいたのかわからなくなってしまう. このように, 異なるものを 1 つにまとめると, 元の情報が少し失われてしまう. これをクラスタの情報損失量と呼ぶこととする.

ワード法では, 2 つのクラスタ A, B を 1 つのクラスタにまとめたとき, その情報損失量をクラスタ間の距離 D とする. (図 7)

具体的に, クラスタ間の距離 D は, 以下のような式で定義される.

$$\text{クラスタ間の距離 } D = L(A \cup B) - L(A) - L(B)$$

ここで $L(A)$ は, クラスタ A の各個体から重心までの距離の 2 乗和を計算したもので, クラスタ内



でのデータの散らばり具合を表現している. $L(B)$ 及び $L(A \cup B)$ 同様である.

3 クラスター分析の手順

表 1 のデータを使って, 実際にクラスター分析をしてみる. クラスター分析は, 以降のような手順で進んでいき, 次々にまとまっていくクラスタをデンドログラム (樹形図) というグラフで表現する. なお今回, 距離は平方ユークリッド距離, クラスタ間距離は重心法を用いて求めていく.

表 1: エイズ患者数と新聞の発行部数

| 国名 | エイズ患者 | 新聞の発行部数 |
|----|-------|---------|
| A | 6.6 | 35.8 |
| B | 8.4 | 22.1 |
| C | 24.2 | 19.1 |
| D | 10.0 | 34.4 |
| E | 14.5 | 9.9 |
| F | 12.2 | 31.1 |
| G | 4.8 | 53.0 |
| H | 19.8 | 7.5 |
| I | 6.1 | 53.4 |
| J | 26.8 | 50.0 |
| K | 7.4 | 42.1 |

手順 1

はじめに, すべての組み合わせにおける”距離”を計算すると, 以下の表 2 のようになる. この中で, G と I の間の距離が

$$(4.8 - 6.1)^2 + (53.0 - 53.4)^2 = 1.85 \simeq 1.9$$

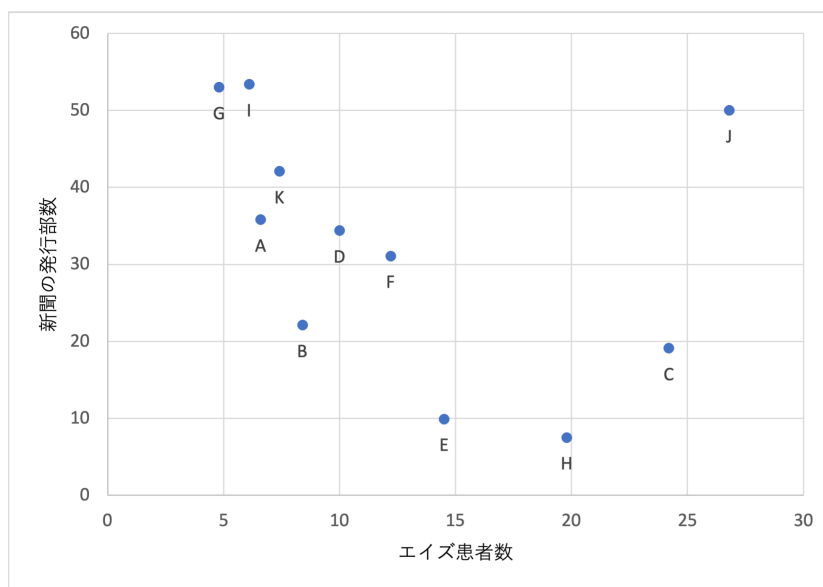


図 8: エイズ患者数と新聞の発行部数

表 2: 手順 1 による距離の計算結果

| | B | C | D | E | F | G | H | I | J | K |
|---|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|
| A | 190.9 | 588.7 | 13.5 | 733.2 | 53.5 | 299.1 | 975.1 | 310.0 | 609.7 | 40.3 |
| B | | 258.6 | 153.9 | 186.1 | 95.4 | 967.8 | 343.1 | 985.0 | 1117.0 | 401.0 |
| C | | | 435.7 | 178.7 | 288.0 | 1525.6 | 153.9 | 1504.1 | 961.6 | 811.2 |
| D | | | | 620.5 | 15.7 | 373.0 | 819.7 | 376.2 | 525.6 | 66.1 |
| E | | | | | 454.7 | 1951.7 | 33.9 | 1962.8 | 1759.3 | 1087.3 |
| F | | | | | | 534.4 | 614.7 | 534.5 | 570.4 | 144.04 |
| G | | | | | | | 2295.3 | 1.9 | 493.0 | 125.6 |
| H | | | | | | | | 2294.5 | 1855.3 | 1350.9 |
| I | | | | | | | | | 440.1 | 129.4 |
| J | | | | | | | | | | 438.8 |

となり、すべての組み合わせの中で最小になる。よって、G と I が最初のクラスタ {G, I} を構成する。これをデンドログラムに描くと、図 9 のようになる。

また、クラスタ {G, I} の重心を求めると (5.45, 53.2) であり、以降の手順ではこの重心を基点として、クラスタ {G, I} との距離を計算していく。

手順 2

次に残りすべての組み合わせにおける”距離”を計算すると、以下の表 3 のようになる。

この中で A と D の組み合わせが最小となる。よって、A と D が 2 つ目のクラスタ {A, D} を構成する。これをデンドログラムに描き加えると、図 10 のようになる。

また、クラスタ {A, D} の重心を求めると (8.3, 35.1) であり、以降の手順ではこの重心を基点として、クラスタ {A, D} との距離を計算していく。

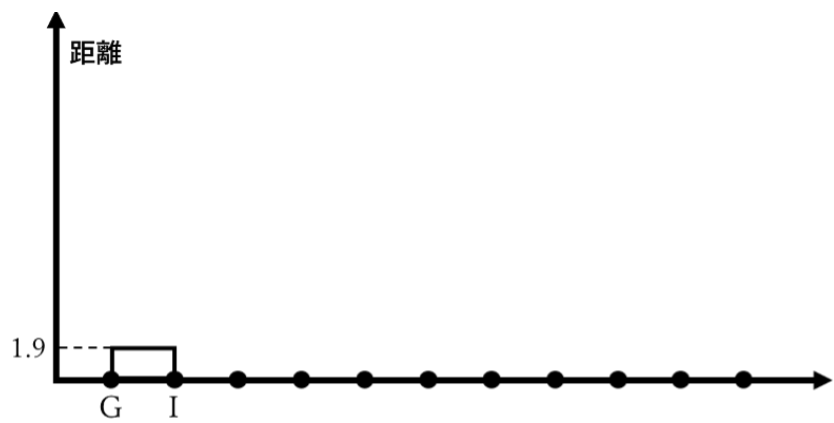


図 9: 手順 1 によるデンドログラム

表 3: 手順 2 による距離の計算結果

| | B | C | D | E | F | G・I | H | J | K |
|-----|-------|-------|-------|-------|-------|--------|--------|--------|--------|
| A | 190.9 | 588.7 | 13.5 | 733.2 | 53.5 | 304.1 | 975.1 | 609.7 | 40.3 |
| B | | 258.6 | 153.9 | 186.1 | 95.4 | 975.9 | 343.1 | 1117.0 | 401.0 |
| C | | | 435.7 | 178.7 | 288 | 1514.4 | 153.9 | 961.6 | 811.2 |
| D | | | | 620.5 | 15.7 | 374.1 | 819.7 | 525.6 | 66.1 |
| E | | | | | 454.7 | 1956.8 | 33.9 | 1759.3 | 1087.3 |
| F | | | | | | 534.0 | 614.7 | 570.4 | 144.0 |
| G・I | | | | | | | 2294.4 | 466.1 | 127.0 |
| H | | | | | | | | 1855.3 | 1350.9 |
| J | | | | | | | | | 438.8 |

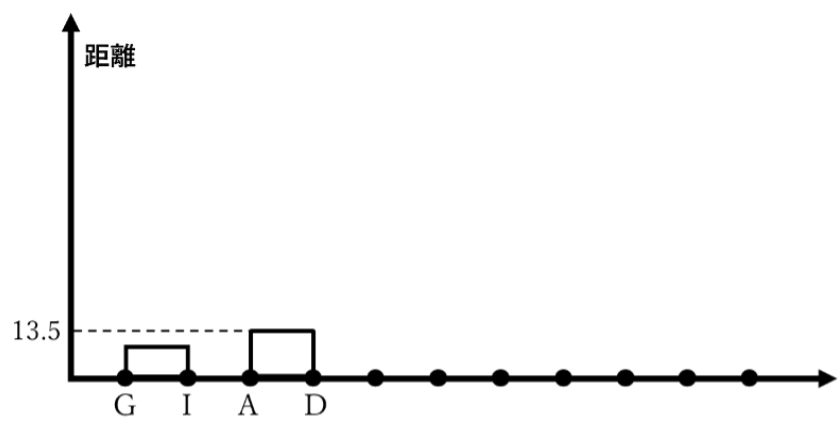


図 10: 手順 2 によるデンドログラム

手順 3

以上の作業を繰り返していき, 10 回目で最後のクラスタが構成されて終了となる. 最終的に完成したデンドログラムは, 次の図 11 のようになる.

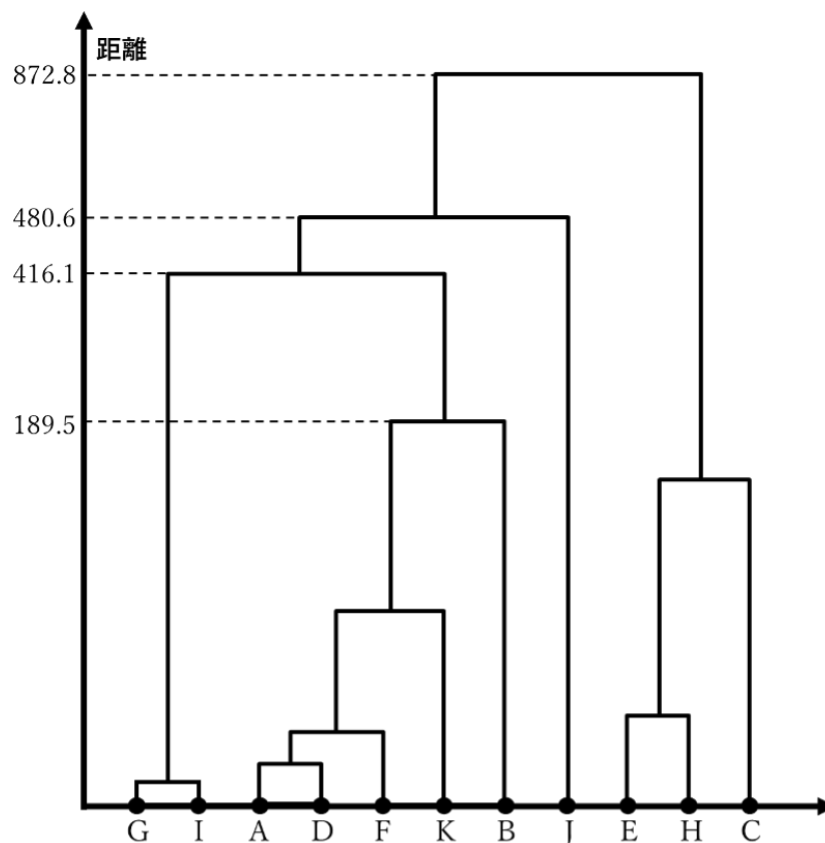


図 11: 完成したデンドログラム

4 デンドログラム

デンドログラムの見方

デンドログラムは個体とクラスタ間の”距離”の関係をまとめたものであり, クラスター分析において非常に重要なグラフ表現である.

縦軸が類似度を表す”距離”となっており, 横軸に平行な線を引いたとき, デンドログラムの縦線とぶつかった個数がクラスタの個数になる. またこのとき, クラスタを構成している個体の内訳をみることができる.

例えば, クラスタの個数を 4 個にしたい場合は, 図 12 のようにオレンジ色の平行線を引けばよい. そして, 4 つのクラスタはそれぞれ $\{G, I\}$, $\{A, D, F, K, B\}$, $\{J\}$, $\{E, H, C\}$ という個体で構

成されることが読み取れる。

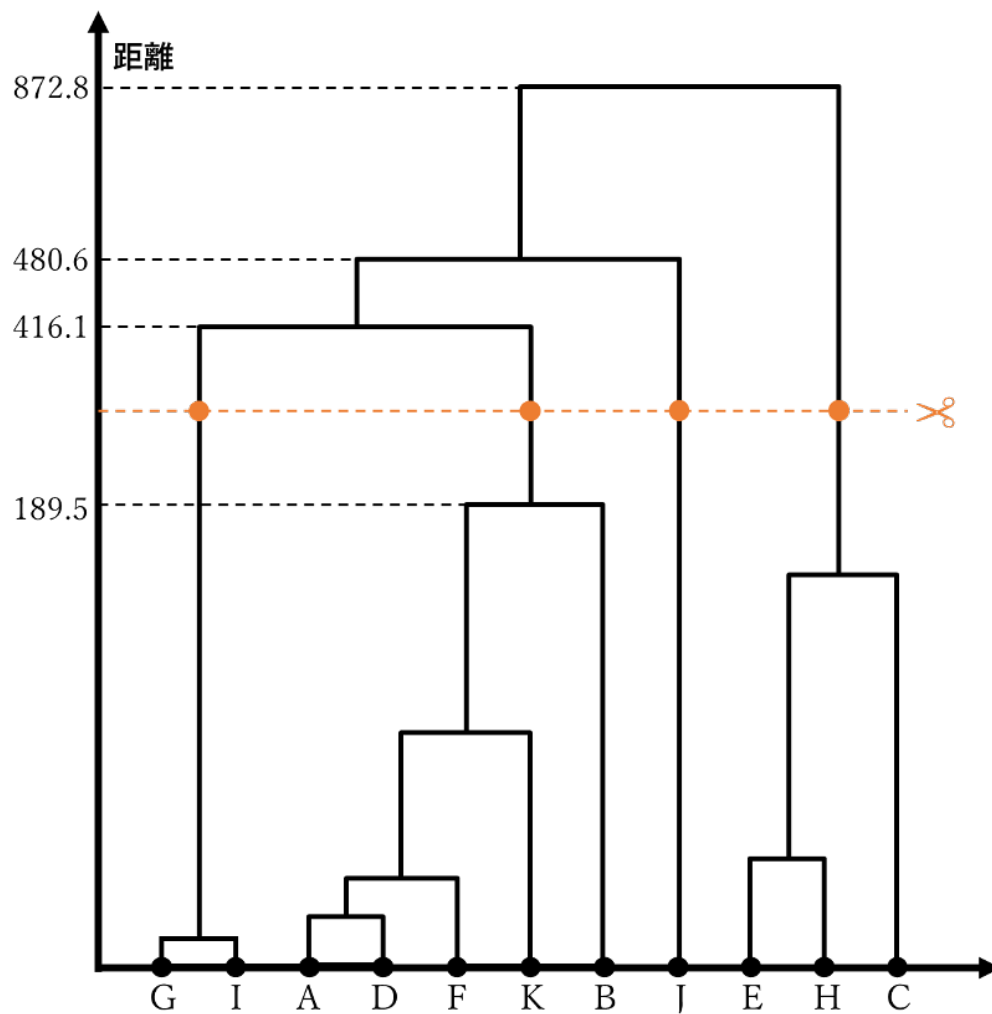


図 12: デンドログラムを用いたクラスター分析

また, 4つのクラスターを散布図に描くと, 次の図 13 のようになる.

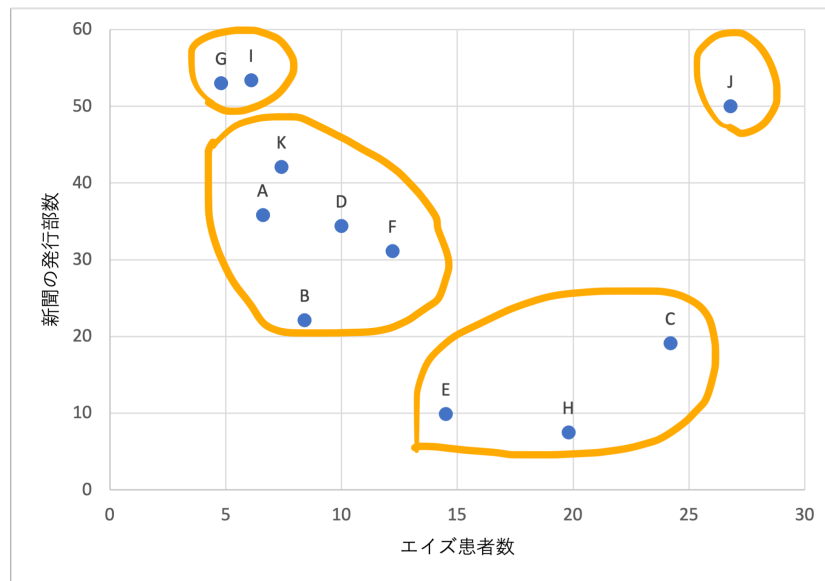


図 13: クラスター分析によって 4 つ分類した結果

最適なクラスターの個数

デンドログラムに平行線を引くことで, 任意の個数のクラスターを求めることができた. しかし, クラスター分析を行う際, ”最適なクラスターの個数は何個なのか?” という問題がある. 実は, はっきりとした基準はなく, 何個のクラスターに分類するかは, そのデータを研究している人次第である.