第６章クラスター分析

6.1階層的クラスター分析
6.2非階層的クラスター分析

07t4072f岡崎駿

序論

- クラスター分析とは
  - 観測対象を互いに似たものどうしでグループ分けする手法
  - 例）
    - 車の販売会社の顧客の場合「若くて活発的なグループ」と「幼い子をもの家族連れのグループ」

類似度の定義

- ユークリッド距離

2変数の場合、対称 i と j の座標を(x_{i1}, x_{i2}), (x_{j1}, x_{j2}) としたとき

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$ と定義される

距離が近いほど類似しており、遠いほど似ていない。

多変数の場合は上式を拡張させて

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

動物の分類

以下の手順に従い、階層的クラスター分析を行う。

1.変数を用いて個々の対象間の距離をすべて計算し、その中で距離が最も短い対称同士を併合して最初のクラスターを作成

2.新しく併合させたクラスターと多の対称間の距離を再度計算し、手順1で計算された対象間の距離を含めて最も近いものを併合する。その際、新しく併合させたクラスターと対象間及びクラスター間の距離は、クラスターの重心間の距離を用いる重心法で行う

3. 手順2を繰り返し、全てのクラスターが統合させるまで行う

クラスター間の距離

- 最短距離法：任意の2つのクラスターに含まれる対象すべての組み合わせに関して距離を求め、もっとも近い距離をクラスター間の距離として定義

- 最長距離法：最短距離法とは逆に全ての組み合わせに関して求めた距離のうちで最大距離となる対象間の距離をクラスター間の距離として定義する

- 群平均法：各クラスターに含まれる全ての対象間の距離に関して平均をとったものをクラスター間の距離とする

- 重心法：クラスター間の重心を求め、重心間の距離でクラスター間の距離とする

ウォード法

- 2つのクラスターを併合する際に、クラスター内の平方和を最小にするようにクラスターを併合していく手法。

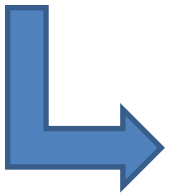
- クラスター間の距離は2つのクラスターを併合したときの平方和（散布度）の増加量
  - 平方和の増加量が大きいほど、2つのクラスターは類似していない
  - 平方和の増加量が小さいものから併合→似たもの同士がまとまる

- 散らばりの変化量＝（AとBを併合した後の平方和）
  ー(A内の平方和)ー(B内の平方和)

ウォード法と重心法

- ウォード法の利点は鎖連鎖が起こりにくい
  - 鎖連鎖・・・ある1つのクラスターに対象が1つずつ順番に吸収されてクラスターの形成がなされていく現象

- 重心法を使った場合クラスター併合後に距離が短くなってしまう可能性がある
  - 距離の単調性が保証されていない



分析方法に迷ったらウォード法がおすすめ

非階層的クラスター分析

- あらかじめ指定したクラスター数で観測対象を分析する
  - クラスター数を指定しなければ分析できない非階層的手法は、それ無しで分析できる階層的手法よりも不利である。
- 膨大なデータを分析する場合にはクラスター数の目星をつけて非階層的手法を数回実施したほうが効率がよい場合がある

K-means法

- 非階層的手法の中で頻繁に利用される手法
- 以下の手順に従ってK個のクラスターに分析
  - 1.N個の観測対象をK個の初期クラスターに任意に分類
  - 2.各クラスターの中心点を計算
  - 3.N個の観測対象のK個のクラスターの中心点への距離（合計N×K個）を計算する
  - 4.全ての観測対象に関して、その時点で各自が所属しているクラスターへの距離がK個のクラスターの中で1番近ければ計算を終了。さもなければ、一番近いクラスターに再割当てして、手順2に戻る

K-means法の確認1

- 4つの観測対象にそれぞれ2変数の状態を保持(N=4)
- この観測対象を2つのクラスターに分ける（K=2）

観測対象	x1	x2
A	6	4
B	−2	2
C	0	−2
D	−2	0

- 初期クラスターとして（AB）と（CD）に分類する（手順1）

K-means法の確認2

- クラスター（AB）の中心点は（2,3）であり、同様にCDの中心点は（-1,-1）である。（手順2）
- 観測対象からクラスターの中心点への距離を計算する。（手順3）
  - Aから（AB）の中心点への距離は $(6-2)^2+(4-3)^2=17$

	（AB）	（CD）
A	17*	74
B	17	10*
C	29	2*
D	25	2*

＊は最小値

Bが誤分類されている

K-means法の確認3

- クラスターを（A）と（BCD）にする（手順4）
- クラスター（A）の中心点は（6,4）であり、（BCD）の中心点は（-4/3,0）である（手順2）

	（A）	（BCD）
A	0*	69.78
B	68	4.44*
C	72	5.78*
D	80	0.44*

- 4つの観測対象が所属しているクラスターへの距離が1番近いので計算終了（手順4）