非階層クラスター分析：k-平均法

渡邊直樹

2021 年 10 月 20 日

クラスター分析とは類似した特徴を持ついくつかのグループにデータを分類する手法であり，階層クラスター分析と非階層クラスター分析に大別される．ここでは，主成分分析によって生成された第 2 主成分と第 2 主成分をデータとして，それらを k-平均法による非階層クラスター分析を用いて分類する手続きについて説明する．

非階層クラスター分析では，各グループ内のデータが類似した特徴を持つように，分析者によって予め決められた数のグループにデータを分類する．あるグループに属するデータは他のどのグループにも属さない．このような条件の下で，データはいくつかのグループ（クラスター）に分類される．グループの数を予め設定することにより，データのサイズが大きくなっても，明快な結果を得ることができる．ただ，クラスターの数をいくつに設定すれば良いかという判断は分析者に委ねられる．

階層クラスター分析では，データが持ついくつかの特徴の間にどの程度の類似性があるか，または，それぞれが特徴を持つ個人はどの程度類似してるかが入子構造を持つグループ（クラスター）を表現することに適した樹形図（デンドログラム）によって示される．ただ，分類によって生成されるグループの数が予め決まっていないため，データが持つ特徴やサンプルとしての個人の数が多いときには，却って，分類が不明確になることがしばしばある．

k-平均法

k-平均法（k-means method）には，いくつかの計算方法が提案されているが，いずれの方法でも，次のような 4 つのステップを踏む．まず，分類によって生成されるクラスターの数を k とする．

(1) k 個のクラスターの種（seeds）を何らかの方法で与える.（初期値の設定）

(2) 各データについて，k 個のクラスターの種との距離をそれぞれ計算し，最も近いクラスターに分類する.

(3) 形成されたクラスターにおける新しい種を計算によって与える.

(4) (2) と (3) のステップを繰り返す．この繰り返しは，新たに計算されたすべてのクラスターの種が直前の手続きにおける計算結果と同じになるか，事前に指定した繰り返しの回数に達するまで続く.

主成分データを 4 つのクラスターに分ける

ここには，EZR（Easy R）を用いるとして，主成分分析とそれに続く主成分データのクラスタリングの手順をまとめてノートしておく．主成分分析については，別の資料を参照してほしい．また，自分の計算機にデータセット（data2.csv）が保存されているとして，それを EZR で読み込むことにする.

1. データの読み込み : リボンに表示されているボタンを次の順で押す.
    - データ＞データのインポート＞テキストファイルまたはクリップボード，URL から
    - ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている.（データファイルの場所はデフォルトでローカルファイルシステムが選択されているので，そのままにしておく.）
    - 新たなポップアプウィンドウが現れるので，そこで data2.csv を選択し、Open ボタンを押す.
    - ホーム画面上で「データセットを表示」ボタンを押して，データが読み込まれていることを確認する.
2. 主成分分析 : リボンに表示されているボタンを次の順で押す.
    - 統計量＞次元解析＞主成分分析

- ポップアップウィンドウの「データ」の欄にて 2 つ以上の変数を選択する.
- オプションで「データセットの主成分得点を保存」の項目にチェックを入れておく.
- デフォルトでは「相関行列の分析」の項目にチェックが入っているので，それを外すと，分散共分散行列の分析になる.
  （相関行列での分析では，データの下処理が必要となることがある.）
- 「適用」ボタンを押す.
- 新しいポップアップウィンドウが現れるので，保存する主成分得点の数をスライダーで選択する.
- OK ボタンを押す.

3. クラスター分析：リボンに表示されているボタンを次の順で押す.

- 統計量＞次元解析＞クラスター分析＞ k-平均クラスター分析
- ポップアップウィンドウの「データ」の欄にて第 1 主成分（PC1）と第 2 主成分（PC2）を選択する.
- オプションで「クラスタ数」を 4 とする.「シード初期値の数」と「最大繰り返し数」を選択する.（スライダーを動かして操作する.）
- 「クラスタを保存する変数」のセルに変数名を書き込み，クラスター番号を格納する変数の名前をつける．デフォルトでは KMeans となっている.
- デフォルトでは「クラスタのサマリの表示」と「クラスタのバイプロット」の項目にチェックが入っている．クラスタのバイプロットにチェックが入っていると，クラスター分析の結果が散布図でも表示される.「データセットにクラスタを割り当てる」の項目にチェックを入れる．これにより，クラスターの番号をデータとして保存し，エクセルなどで読み出して作図を行うことができるようになる.
- 「適用」ボタンを押す.

EZR の出力は,「シード初期値の数」と「最大繰り返し数」の設定値によって異なるが，次のようになる．#の右側に結果の意味が書かれている．各クラスターのデータ数は 20，36，31，13 であり，その中心となっている種の座標が (new.x.PC1, new.x.PC2) である．各クラスターの中心となっている種からの距離の自乗和とそれらの総和，クラスターの中心となっている種同士の距離の自乗和が表示される．

```
> .cluster$size  # Cluster Sizes
[1] 20 36 31 13

> .cluster$centers  # Cluster Centroids
   new.x.PC1   new.x.PC2
1  2.1252883  0.89206278
2 -2.2564000 -0.23232312
3 -0.4041804  0.07264912
4  3.9426331 -0.90228816

> .cluster$withinss  # Within Cluster Sum of Squares
[1] 31.83451 30.27055 38.97158 10.74044

> .cluster$tot.withinss  # Total Within Sum of Squares
[1] 111.8171

> .cluster$betweenss  # Between Cluster Sum of Squares
[1] 509.3719
```

EZR のリボンにある「データセットを表示」ボタンを押して，データが書き込まれているかどうかを確認する．

4. 分析結果の保存，主成分スコアとクラスター番号を書き加えたデータセットの出力：リボンに表示されているボタンを次の順で押す．

   - ファイル＞出力ファイルに保存．ポップアップウィンドウでファイルに名前をつけ，計算機に csv 形式で分析結果を保存する．（EZR の画面にも分析結果は出力されるが，やや見にくい．）

- データ＞アクティヴデータセット＞アクティブデータセットのエクスポート.（アクティヴデータセットの保存を押すと，RData 形式で保存されてしまうので，Excel などで読み込めない.）ポップアップウィンドウにおいて，フィールドの区切り記号として「カンマ [,]」を選択し，OK ボタンを押す．デフォルトでは「空白」が選択されている．