R 言語を用いたクラスタリングに関する実験レポート

5419080 関一樹
日本大学文理学部情報科学科

概要

本実験レポートでは, まず最初にクラスタリングに関する知識及び手法について説明を行う. その後, 与えられたデータに対しクラスタリングを行う. 最後に得られた結果に関して考察を行う.

1 目的

本実験レポートは,R 言語を用いたデータの分析を通じ, クラスタリングの考え方を確認・復習すること, および具体的な操作を習得することを目的とする.

2 理論

クラスタリングとは, データを類似性の高いもの同士で分類することである. 特に, 予めグループの意味づけを行った後に分類を行うクラス分類と異なり, クラスタリングは設定した基準に基づいた分類を行う. そのため, 分類結果の意味はあとから解釈する. また, クラスタリングには階層型と非階層型の 2 種類が存在する.

2.1 階層的クラスタリング

階層的クラスタリングとは, 各データについて類似性が高い順に 1 つずつグルーピングを行う手法のことである. 本実験レポートでは最短距離法とウォード法を採択した. また, 得られた結果をデンドログラム (樹形図) で表示し, 任意の基準を設定することで最終的なクラスタが決定する.

2.1.1 最短距離法

最短距離法とは各データ間の距離について着目し, 最もデータ間の距離 (以後クラスタ間距離とする) が小さいペアについてグルーピングする手法である. 判断に用いる距離はユークリッド距離を用いる. また結合を行ったクラスタについては, クラスの重心を元にクラスタ間距離を算出する. メリットとして用いる値がクラスタ間距離のみであるため, 非常に計算量が少ない. 一方, クラスタ間に着目する性質上, 外れ値に弱いデメリットがある.

2.1.2 ウォード法

ウォード法とは,2 つのクラスタ C_1, C_2 の併合前と併合後の差異について着目する手法であり, 下記の式で表現される. また, 判断に用いる距離はユークリッド距離を用いる.

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$E(C) = \sum_{x \in C} dist(x, M(C))^2$$
$$dist(x, M(C)) = \sqrt{\sum_{x_i \in C} (x_i - M(C))^2}$$

上記の式より,$E(C)$(2 乗平方和) はクラスタの重心 $(M(C))$ からのユークリッド距離を用いている為,$L(C)$ において得られる値は各クラスタの散らばり具合を表している. 即ち, 上記の式では, 併合後の散らばり具合から各クラスタの散らばり具合を引くことで, 各データがどの程度併合後の重心に集中しているかを表すことができる. その後, 得られた値について比較を行い, 順にクラスタリングを繰り返していく. メリットとして, データの散らばり具合 (平方和) について着目しているため分類精度が高い. 一方計算量が多くなるデメリットがある.

2.2 非階層的クラスタリング

非階層的クラスタリングとは, 任意のクラスター数を元に各データをクラスタリングする手法である. 階層的クラスタリングと異なり,1 つずつ分類するのではなく, 各手法による基準を元に分類を行う. 本実験レポートでは K-平均法を採択した.

2.2.1 K-平均法

K-平均法とはクラスタの重心と各データ間の距離について着目し繰り返しクラスタリングする手法である. 手順として, まず任意のクラスタ数と同じ数ランダムな初期値を設定する. その後, 与えられた初期値と各データについてユークリッド距離を用いたクラスタ間算出し, 最も距離が短い初期値のクラスタに分類を行う. その後, 分類されたクラスタについて重心を算出し, 得た重心について再びクラスタ間距離を算出し分類を繰り返す. この時, 再分類の前後にてクラスタのメンバが一致するまで分類を繰り返す. メリットとして, 一般に重心と各データの計算のみとなるため, 計算量が少なく収まる点がある. 一方, デメリットとして, 分類結果が初期値に依存しているため, 実行により結果が異なる点がある. この問題に対し, 本実験レポートでは k-平均法を複数回行い, ウォード法で用いた 2 乗平方和が最小となる実行結果を最適解とし,k-平均法の実行結果とする.

2.3 評価手法

得られたクラスタリング結果に関しての評価として, 内的妥当性尺度と外的妥当性尺度の手法が存在する. 内的妥当性尺度とは, クラスタリングそのものに対する性質に関しての評価手法に対し, 外的妥当性尺度はクラスタリング結果と正解となるクラスタリングと比較し整合性について評価を行う手法である. 本実験レポートでは内的妥当性尺度による評価を採択し, 特にコーフェン相関係数及びシルエット値による評価を用いる.

2.3.1 コーフェン相関係数

コーフェン相関係数とは, クラスタ間距離とデンドログラム作成時の高さの相関係数について着目する手法である. 以下にコーフェン相関係数の式を示す.

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)ver(Y)}}$$

上式により,2 変数間の線形関係の強さを算出することができる. 値は -1 $+1$ の範囲内を取り,1 に近いほど

精度が高くなることが断定できる.

2.3.2 シルエット値

シルエット値とは,「同一クラスタ内の凝集度」,「異なるクラスタへの乖離度」に着目した評価指標である. 一般に下記の通りに定式化される.

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}$$

上式における $a(x_i), b(x_i)$ について以下の通りに表せる.

$$a(x_i) = \frac{1}{|C| - 1} \sum_{x_j \in C \setminus \{x_j\}} d(x_j - x_i)$$

$$b(x_i) = min_{C' \neq C \in CL} \left\{ \frac{1}{|C'|} \sum_{x_j \in C'} d(x_j - x_i) \right\}$$

$a(x_i)$ では x_i が属するクラスタ内のクラスタ間距離平均について算出することで,「同一クラスタ内に関するクラスタ間距離」について着目している. 一方,$b(x_i)$ では異なるクラスタに属するデータへの距離平均の最小値を算出することで,「異なるクラスタへのクラスタ間距離」について着目している. その後, 乖離度と凝集度の差である $b(x_i) - a(x_i)$ に関して $b(x_i)$ または $a(x_i)$ の最大で割ることで取りうる値の範囲を -1 $+1$ の間へ変換を行い, シルエット値が導出された. また, 乖離度と凝集度の差を扱う性質上,$b(x_i) - a(x_i)$ の値が負である場合, 即ち $a(x_i)$ の値が大きい場合,「同一クラスタ内に関するクラスタ間距離」が大きいことより, クラスタリングの精度が低いと断定できる. 一方,$b(x_i) - a(x_i)$ の値が正である場合, 即ち $b(x_i)$ の値が大きい場合,「異なるクラスタへのクラスタ間距離」が高い事よりクラスタリングの精度が高いと断定できる.

3 実験手法

本実験レポートでは e-Stat[1] より入手した「全国家計構造調査（旧全国消費実態調査） 平成 26 年全国消費実態調査 全国 家計収支に関する結果 単身世帯」より食料消費に関する 12 項目を抽出し，各項目の食料消費全体に占める割合を県別に集計したものを用いた. データの各属性については付録に記載した.

本実験レポートでは, 最初にデータに対して各手法を用いクラスタリングを行う. その後, 得られた各結果に対して, 各評価指標を用いた評価を行い, その後, 比較及び考察を行う. また, 分類するクラスタ数に関して本実験レポートでは, 日本の地方数である 7 地方区分より,7 つのクラスタに分類する条件を設定した.

4 実験結果・考察

4.1 最短距離法

食料消費に関するデータ (以後 food データとする) に対し最短距離法を用いてクラスタリングを行う. 以下に実行結果 (デンドログラム) を示す.

得られたデンドログラムを元にクラスタ数を 7 とし, クラスタの分割を行う. 以下に実行結果と各クラスタの属性番号を示す.

```
> food.single <- hclust(dist(food), method="single")
> food.single7 <- cutree(food.single, k=7)
> food.single7
  北海道   青森県   岩手県   宮城県   秋田県   山形県   福島県   茨城県
      1        2        1        3        4        3        5        3
  栃木県   群馬県   埼玉県   千葉県   東京都 神奈川県   新潟県   富山県
      5        3        3        3        3        3        3        3
  石川県   福井県   山梨県   長野県   岐阜県   静岡県   愛知県   三重県
      3        3        3        3        3        3        3        3
  滋賀県   京都府   大阪府   兵庫県   奈良県 和歌山県   鳥取県   島根県
      3        3        3        3        3        3        3        3
  岡山県   広島県   山口県   徳島県   香川県   愛媛県   高知県   福岡県
      3        3        5        3        3        3        3        3
  佐賀県   長崎県   熊本県   大分県   宮崎県 鹿児島県   沖縄県
      3        3        3        3        6        3        7
> table(food.single7)
food.single7
 1  2  3  4  5  6  7
 2  1 37  1  4  1  1
```

また, 最短距離法を用いた際のコーフェン相関係数は以下の通りになる.

```
> cor( cophenetic( food.single ), dist(food) )
[1] 0.4598736
```

以上の結果より, コーフェン相関係数の値が約 0.45 であることからクラスタ間距離とデンドログラム上の高さに弱い相関があることが推測される. 加えて, 上記の各クラスタに属するデータ数に関する結果より, あるグループのみ関してデータが集中している事から,food データを最短距離法を用いて 7 つのクラスタに分類するクラスタリング実行結果は信頼するに足りない事が推測できる. また, デンドログラム表示結果より,「青森県」のデータが最後にクラスタリングされていることから, 青森県に関するデータが外れ値である確率が高い事も推測できる.

4.2 ウォード法

food データに対してウォード法を用いたクラスタリングを行う. 以下に実行結果 (デンドログラム) を示す.

得られたデンドログラムを元にクラスタ数を 7 とし, クラスタの分割を行う. 以下に実行結果と各クラスタの属性番号を示す.

```
> food.ward <- hclust(dist(food), method="ward.D")
> food.ward7 <- cutree(food.ward, k=7)
> food.ward7
```

北海道	青森県	岩手県	宮城県	秋田県	山形県	福島県	茨城県
1	2	1	3	4	3	5	3
栃木県	**群馬県**	**埼玉県**	**千葉県**	**東京都**	**神奈川県**	**新潟県**	**富山県**
5	3	6	4	4	4	4	2
石川県	**福井県**	**山梨県**	**長野県**	**岐阜県**	**静岡県**	**愛知県**	**三重県**
6	3	3	4	7	6	4	4
滋賀県	**京都府**	**大阪府**	**兵庫県**	**奈良県**	**和歌山県**	**鳥取県**	**島根県**
6	6	3	4	4	5	3	7
岡山県	**広島県**	**山口県**	**徳島県**	**香川県**	**愛媛県**	**高知県**	**福岡県**
4	7	5	3	6	7	2	7

```
  佐賀県   長崎県   熊本県   大分県   宮崎県   鹿児島県   沖縄県
      7        2        2        3        1        2        7
```

```
> table(food.ward7)
food.ward7
 1  2  3  4  5  6  7
 3  6 10 11  4  6  7
```

また, ウォード法を用いた際のコーフェン相関係数は以下の通りになる.

```
> cor( cophenetic( food.ward ), dist(food) )
[1] 0.6511985
```

以上の結果より, コーフェン相関係数の値が約 0.65 であることから, クラスタ間距離とデンドログラムの高さには相関があることが推測できる. 加えて, 各クラスタに属するデータ数に着目すると, 多少の誤差はあるものの大体均等にクラスタリングされている事からウォード法を用いたクラスタリング結果は信用して問題はないことが推測できる.

4.3 k 平均法

K 平均法についてクラスタリングを行う. 以下に $k = 7$ とし K 平均法を 10 回行った際の 2 乗平方和を示す. また, 小数点 3 位で四捨五入する.

1 回目	2 回目	3 回目	4 回目	5 回目	6 回目	7 回目	8 回目	9 回目	10 回目
453.07	436.49	500.25	434.52	434.52	436.49	450.71	449.42	480.49	449.59

上記より第 4 回目, 第 5 回目の結果が最小値であるため, 以後, 第 4 回 k 平均法を用いる. 以下に第 4 回目の k 平均法実行結果を適宜抜粋し, 示す.

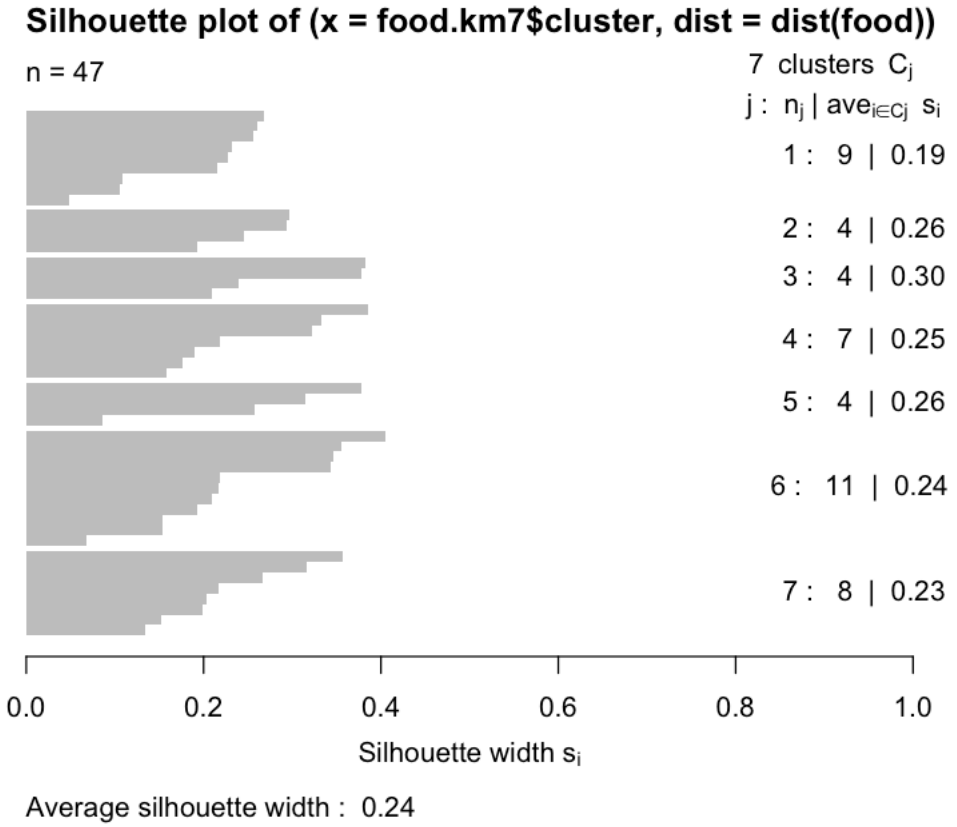
```
  > food.km7 #第 4 回目 k=7 平均法
K-means clustering with 7 clusters of sizes 9, 4, 4, 7, 4, 11, 8

Clustering vector:
  北海道   青森県   岩手県   宮城県   秋田県   山形県   福島県   茨城県
      5        2        5        4        6        1        3        4
  栃木県   群馬県   埼玉県   千葉県   東京都 神奈川県   新潟県   富山県
      3        4        1        6        6        6        6        2
  石川県   福井県   山梨県   長野県 岐阜県   静岡県   愛知県 三重県
      1        4        4        6        7        1        6        6
```

滋賀県	京都府	大阪府	兵庫県	奈良県	和歌山県	鳥取県	島根県
1	1	4	6	6	3	4	7
岡山県	広島県	山口県	徳島県	香川県	愛媛県	高知県	福岡県
6	7	3	1	1	7	2	7
佐賀県	長崎県	熊本県	大分県	宮崎県	鹿児島県	沖縄県	
7	5	7	1	5	2	7	

```
Within cluster sum of squares by cluster:
[1]  83.74831  40.54963  28.77045  46.59537  36.35838 127.13038  71.36934
 (between_SS / total_SS =  78.8 %)
```

得られたクラスタリングに対しシルエット値を適応する. 以下にシルエット値の計算結果を示す.



上記より各クラスタにおけるシルエット値はいずれも正の範囲に収まっていることから, 異なるクラスタ内データへの距離が大きいことが推測できる. 加えて, 各クラスタに属するデータ数に着目すると, いずれのクラスタも極端に所属データ数が偏っておらず, 大体均等に分配されていることから, クラスタリング結果として問題ないことが推測できる. 一方シルエット値に着目すると, 平均値及び各シルエット値はいずれも 0.3 以内に収まっている事からクラスタリング精度としては信用に値するとは一概に断定出来ないことも推測できる.

以上より, 各クラスタリング結果に対しての評価, 考察を行った. 次に各結果に対して比較を行う.

まず, 各データ間に関して最も近いデータをクラスに分類する手法である最短距離法と k 平均法の比較, 考察を行う. 全体的な評価として,„k 平均法ではクラスタ内データ数のばらつきがある程度均一である事から, 最

短距離法ではクラスタ内データ数のばらつきが目立つ. 各評価指標について, 値のみを参考にした場合, 最短距離法の方が大きいが, 今回用いた food データは 1 つのデータに対し 12 個の属性数があるため, 最短距離法より k 平均法を用いた方が総合的な精度は若干上回ると考察できる.

次に階層的クラスタリング手法である最短距離法とウォード法について比較, 考察を行う. まず, 各クラスタに属するデータ数について着目する. 前述した通り, 最短距離法では突出しているクラスタが目立つ一方, ウォード法ではほぼ均一にクラスタリングされている事から, ウォード法を用いた場合の方が正しくクラスタリングを行えた事が推測できる. 加えて, コーフェン相関係数について比較をした場合でも, ウォード法を用いた場合の方が (データ間距離とデンドログラム上での高さの) 相関が強いという結果が得られた. またデンドログラムについて着目する. 例えば, 各手法の結果に対してクラスタ数を 2 としてクラスタリングを行う. 最短距離法では, 青森県とそれ以外の極端なクラスタ生成された. 一方ウォード法では, ほぼ均等にクラスの生成が行われる. 同様にクラスタ数を 3,4 とした場合にも, ほぼ同様の結果が得られる. 以上より, 階層的クラスタリングを用いた food データのクラスタリングはウォード法を用いる方が精度が高いクラスタを生成できる事が考察された.

5 まとめ

本実験レポートでは, クラスタリングに関する知識の確認と定着を目的とし, 単身世帯の食糧消費に関するデータを対象とした計算機実験を行った. その結果, 今回用いた各手法内でウォード法を用いたクラスタリングが最も精度が高い事が確認された. 加えて,R 言語を用いたクラスタリングに関する一定の経験を得ることができた.

参考文献

[1] e-Stat
https://www.e-stat.go.jp/dbview?sid=0003109850