クラスタ分析

基本的な考え方と階層的方法

村田 昇

講義の内容

- **第 1 回 : 基本的な考え方と階層的方法**
- 第 2 回 : 非階層的方法と分析の評価

クラスタ分析の考え方

クラスタ分析

- クラスタ分析 (**cluster analysis**) の目的
  個体の間に隠れている**集まり=クラスタ**を個体間の “距離” にもとづいて発見する方法
- 個体間の類似度・距離 (非類似度) を定義
  - 同じクラスタに属する個体どうしは似通った性質
  - 異なるクラスタに属する個体どうしは異なる性質
- さらなるデータ解析やデータの可視化に利用
- 教師なし学習の代表的な手法の一つ

クラスタ分析の例

- 総務省統計局より取得した都道府県別の社会生活統計指標の一部
  - 総務省 https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0

都道府県名
地方区分
森林面積割合 ： 森林面積割合（%）2014 年
農業産出額 ： 就業者 1 人当たり農業産出額（販売農家）（万円）2014 年
人口割合 ： 全国総人口に占める人口割合（%）2015 年
土地生産性 ： 土地生産性（耕地面積 1 ヘクタール当たり）（万円）2014 年
商品販売額 ： 商業年間商品販売額［卸売業＋小売業］（事業所当たり）（百万円）2013 年

クラスタ分析の考え方

- 階層的方法
  - データ点およびクラスタの間に **距離** を定義
  - 距離に基づいてグループ化
    - 近いものから順にクラスタを **凝集**
    - 近いものが同じクラスタに残るように **分割**
- 非階層的方法

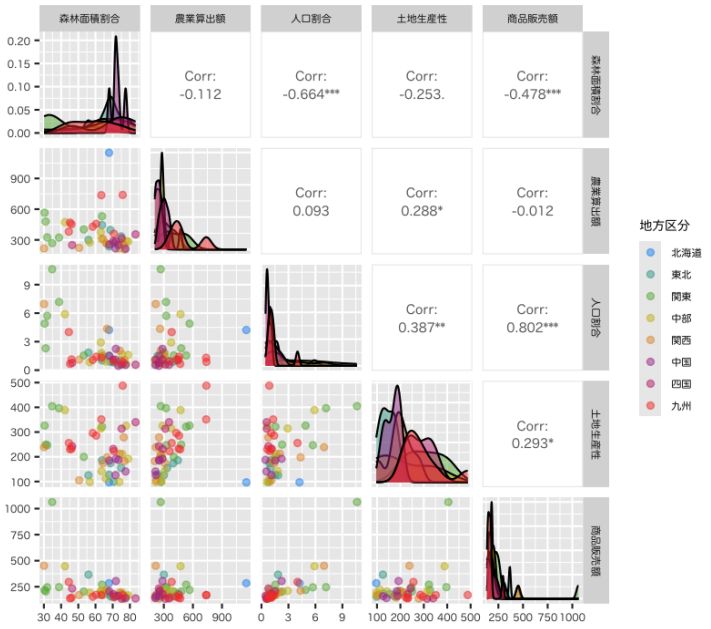

Figure 1: 散布図

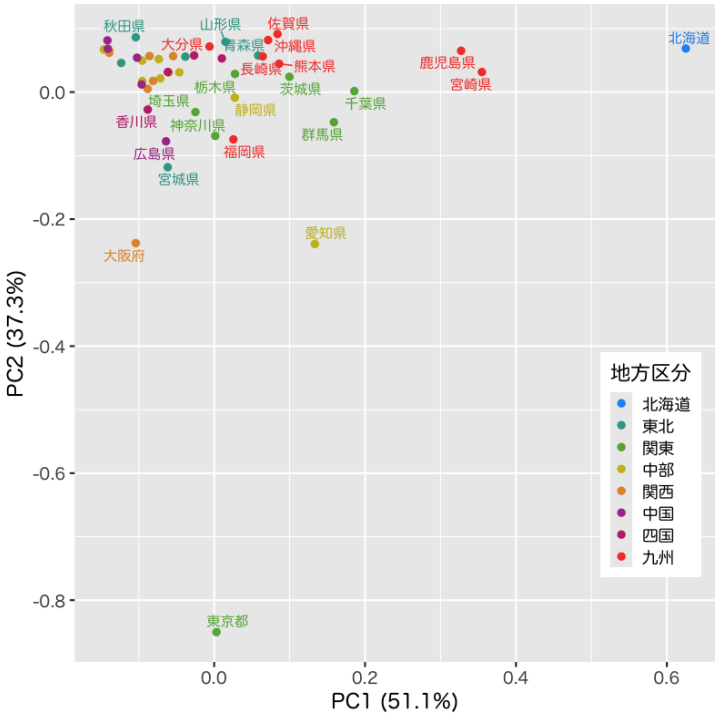

Figure 2: 主成分得点による散布図

都道府県名	地方区分	森林面積割合	農業算出額	人口割合	土地生産性	商品販売額
北海道	北海道	67.9	1150.6	4.23	96.8	283.3
青森県	東北	63.8	444.7	1.03	186.0	183.0
岩手県	東北	74.9	334.3	1.01	155.2	179.4
宮城県	東北	55.9	299.9	1.84	125.3	365.9
秋田県	東北	70.5	268.7	0.81	98.5	153.3
山形県	東北	68.7	396.3	0.88	174.1	157.5
福島県	東北	67.9	236.4	1.51	127.1	184.5
茨城県	関東	31.0	479.0	2.30	249.1	204.9
栃木県	関東	53.2	402.6	1.55	199.6	204.3
群馬県	関東	63.8	530.6	1.55	321.6	270.0
埼玉県	関東	31.9	324.7	5.72	247.0	244.7
千葉県	関東	30.4	565.5	4.90	326.1	219.7
東京都	関東	34.8	268.5	10.63	404.7	1062.6
神奈川県	関東	38.8	322.8	7.18	396.4	246.1
新潟県	中部	63.5	308.6	1.81	141.9	205.5
富山県	中部	56.6	276.1	0.84	98.5	192.4
石川県	中部	66.0	271.3	0.91	112.0	222.9
福井県	中部	73.9	216.1	0.62	98.5	167.3
山梨県	中部	77.8	287.4	0.66	325.3	156.2
長野県	中部	75.5	280.0	1.65	211.3	194.4
岐阜県	中部	79.0	283.7	1.60	192.1	167.9
静岡県	中部	63.1	375.8	2.91	314.5	211.4
愛知県	中部	42.2	472.3	5.89	388.9	446.9
三重県	関西	64.3	310.6	1.43	174.3	170.1
滋賀県	関西	50.5	222.8	1.11	104.9	170.7
京都府	関西	74.2	267.8	2.05	212.5	196.7
大阪府	関西	30.1	216.3	6.96	238.8	451.2
兵庫県	関西	66.7	261.2	4.35	197.7	212.5
奈良県	関西	76.8	207.0	1.07	182.7	147.0
和歌山県	関西	76.4	251.1	0.76	278.4	136.4
鳥取県	中国	73.3	249.9	0.45	187.6	162.2
島根県	中国	77.5	214.1	0.55	140.8	141.1
岡山県	中国	68.0	254.8	1.51	184.9	207.8
広島県	中国	71.8	286.2	2.24	192.2	304.6
山口県	中国	71.6	216.9	1.11	125.8	158.9
徳島県	四国	75.2	315.4	0.59	313.5	134.5
香川県	四国	46.4	249.5	0.77	242.9	232.9
愛媛県	四国	70.3	288.5	1.09	231.6	179.4
高知県	四国	83.3	354.2	0.57	339.9	137.9
福岡県	九州	44.5	381.0	4.01	255.6	295.7
佐賀県	九州	45.2	468.7	0.66	230.3	137.9
長崎県	九州	58.4	428.9	1.08	296.0	154.0
熊本県	九州	60.4	456.6	1.41	285.5	172.5
大分県	九州	70.7	360.1	0.92	222.8	148.3
宮崎県	九州	75.8	739.1	0.87	487.7	170.6
鹿児島県	九州	63.4	736.5	1.30	351.2	169.4
沖縄県	九州	46.1	452.4	1.13	232.8	145.4

  - クラスタの数を事前に指定
  - クラスタの **集まりの良さ** を評価する損失関数を定義
  - 損失関数を最小化するようにクラスタを形成

階層的方法

凝集的クラスタリング

1. データ・クラスタ間の距離を定義する
    - データ点とデータ点の距離
    - クラスタとクラスタの距離
2. データ点およびクラスタ間の距離を求める
3. 最も近い 2 つを統合し新たなクラスタを形成する
    - データ点とデータ点
    - データ点とクラスタ
    - クラスタとクラスタ
4. クラスタ数が 1 つになるまで 2-3 の手続きを繰り返す

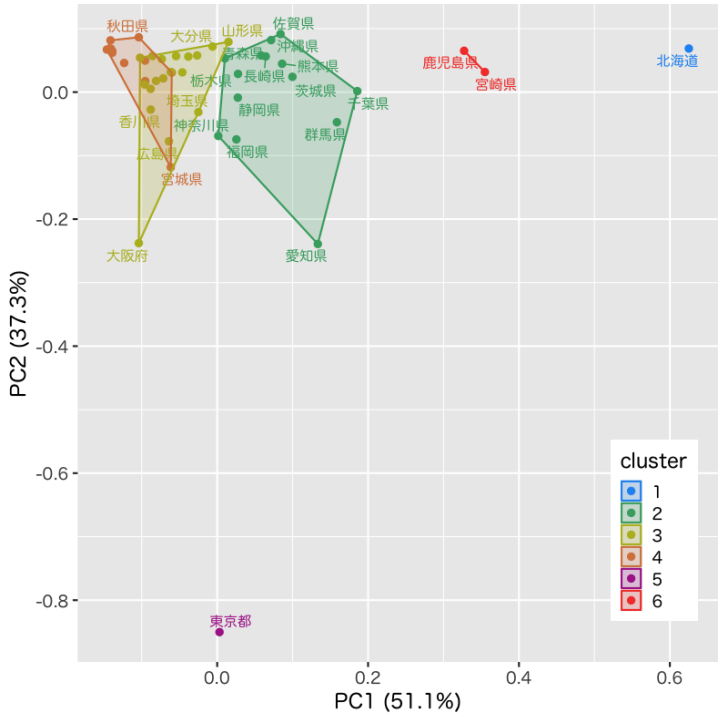

Figure 3: 散布図上のクラスタ構造 (クラスタ分析の概念図)

都道府県名	森林面積割合	農業算出額	人口割合	土地生産性	商品販売額
茨城県	31.0	479.0	2.30	249.1	204.9
栃木県	53.2	402.6	1.55	199.6	204.3
群馬県	63.8	530.6	1.55	321.6	270.0
埼玉県	31.9	324.7	5.72	247.0	244.7
千葉県	30.4	565.5	4.90	326.1	219.7
東京都	34.8	268.5	10.63	404.7	1062.6
神奈川県	38.8	322.8	7.18	396.4	246.1

事例

- 社会生活統計指標の一部 (関東地方)

データ間の距離

データ間の距離

- データ : 変数の値を成分としてもつベクトル

$$\boldsymbol{x} = (x_1, \ldots, x_d)^\mathsf{T}, \boldsymbol{y} = (y_1, \ldots, y_d)^\mathsf{T} \in \mathbb{R}^d$$

- 距離 : $d(\boldsymbol{x}, \boldsymbol{y})$
- 代表的なデータ間の距離
  - Euclid 距離 (ユークリッド ; Euclidean distance)
  - Manhattan 距離 (マンハッタン ; Manhattan distance)
  - Minkowski 距離 (ミンコフスキー ; Minkowski distance)

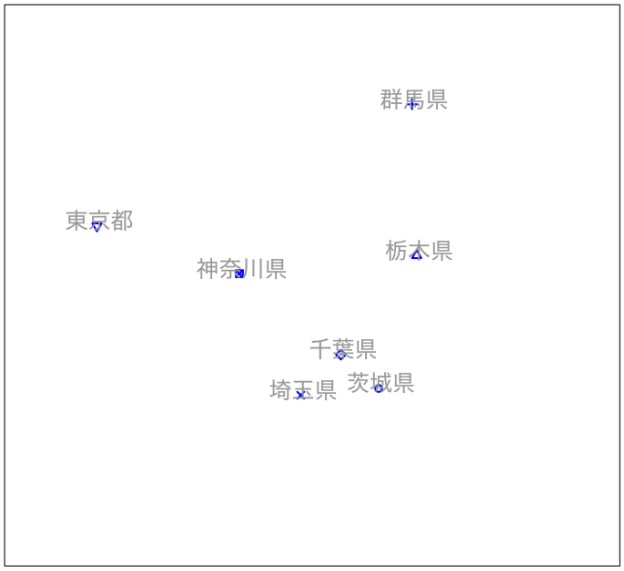

Figure 4: 凝集的クラスタリング

Figure 5: クラスタリングの手続き (その 1)

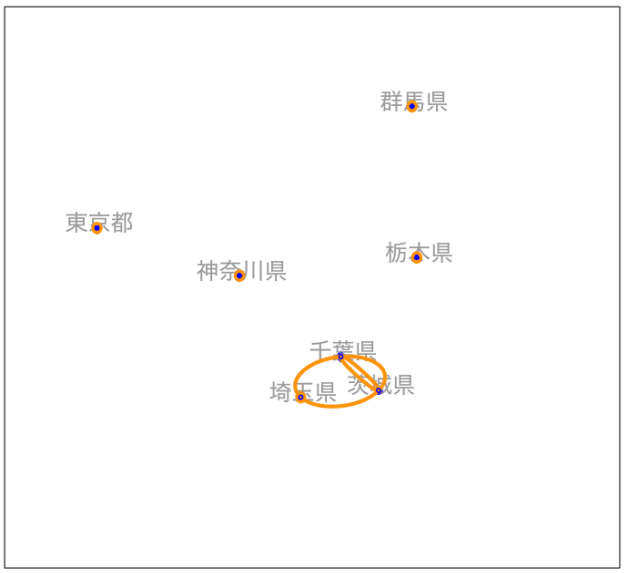

Figure 6: クラスタリングの手続き (その 2)

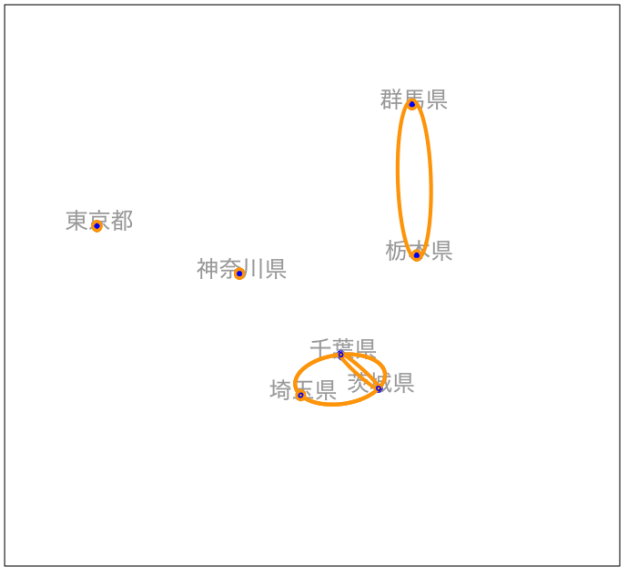

Figure 7: クラスタリングの手続き (その 3)

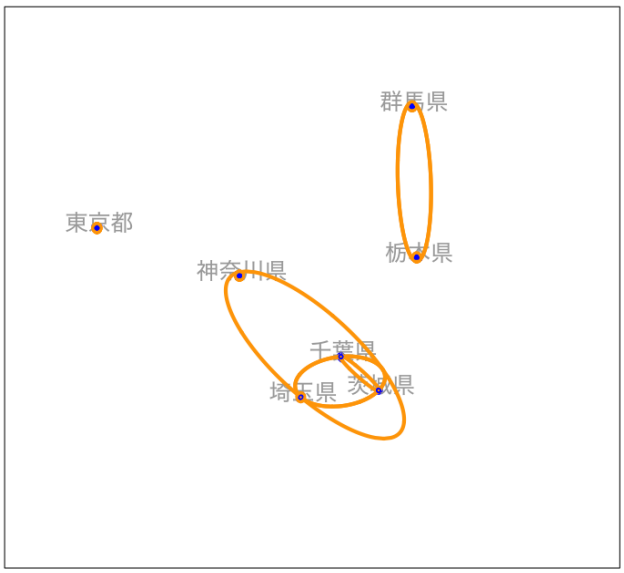

Figure 8: クラスタリングの手続き (その 4)

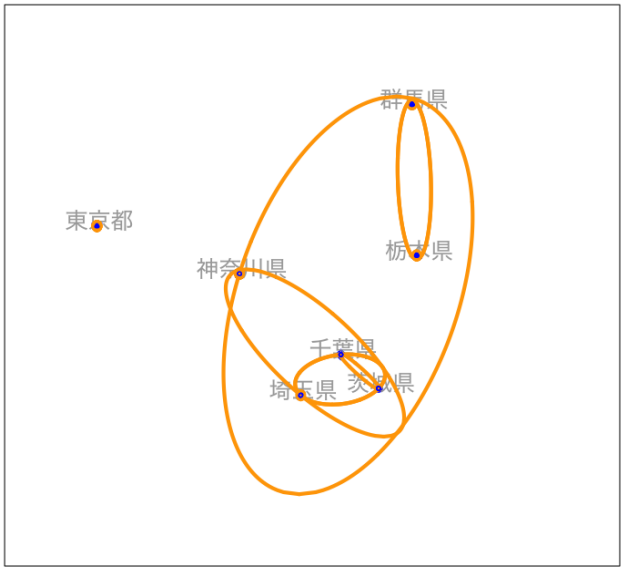

Figure 9: クラスタリングの手続き (その 5)

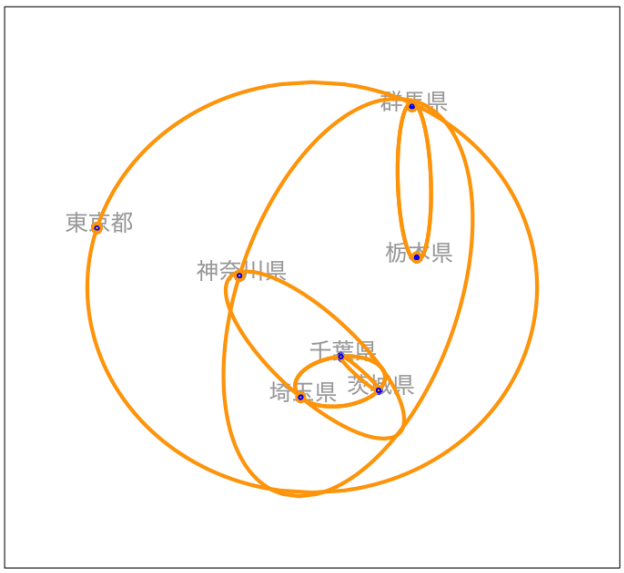

Figure 10: クラスタリングの手続き (その 6)

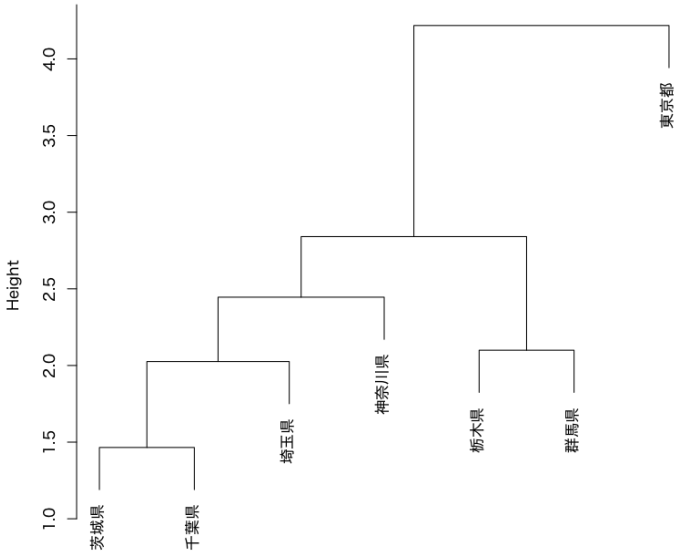

Figure 11: デンドログラムによるクラスタ構造の表示

Euclid 距離

- 最も一般的な距離
- 各成分の差の 2 乗和の平方根 (2 ノルム)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}$$

Manhattan 距離

- 後述する Minkowski 距離の $p = 1$ の場合
- 格子状に引かれた路に沿って移動するときの距離

$$d(\boldsymbol{x}, \boldsymbol{y}) = |x_1 - y_1| + \cdots + |x_d - y_d|$$

Minkowski 距離

- Euclid 距離を p 乗に一般化した距離
- 各成分の差の p 乗和の p 乗根 (p-ノルム)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left\{|x_1 - y_1|^p + \cdots + |x_d - y_d|^p\right\}^{1/p}$$

その他の距離

- 類似度や乖離度などデータ間に自然に定義されるものを用いることは可能
  - 語句の共起 (同一文書に現れる頻度・確率)
  - 会社間の取引量 (売上高などで正規化が必要)
- 擬似的な距離でもアルゴリズムは動く

演習

問題

- 以下の問に答えなさい
  - 距離の定義を述べなさい
  - Minkowski 距離において $p \to \infty$ とするとどのような距離となるか答えなさい

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left\{|x_1 - y_1|^p + \cdots + |x_d - y_d|^p\right\}^{1/p}$$

解答例

- 2 変数の実数値関数で以下の 3 つの条件を満たす
  - 非退化性

$$x = y \Leftrightarrow d(x, y) = 0$$

  - 対称性

$$d(x, y) = d(y, x)$$

  - 劣加法性 (三角不等式の成立)

$$d(x, y) + d(y, z) \geq d(x, z)$$

- 非負性 $d(x, y) \geq 0$ は 3 つの条件から自然に導かれる

$$\begin{aligned} & d(x, y) + d(y, x) \geq d(x, x) && \text{(劣加法性)} \\ & d(x, y) + d(x, y) \geq d(x, x) && \text{(対称性)} \\ & 2d(x, y) \geq 0 && \text{(非退化性)} \\ & d(x, y) \geq 0 \end{aligned}$$

- 最大の要素に着目して計算すればよい

$$\begin{aligned} \lim_{p\to\infty} d(\boldsymbol{x}, \boldsymbol{y}) &= \lim_{p\to\infty} \left\{ |x_1 - y_1|^p + \cdots + |x_d - y_d|^p \right\}^{1/p} \\ &= \lim_{p\to\infty} \max_k |x_k - y_k| \left\{ \left(\frac{|x_1 - y_1|}{\max_k |x_k - y_k|} \right)^p \right. \\ &\qquad \left. + \cdots \left(\frac{|x_d - y_d|}{\max_k |x_k - y_k|} \right)^p \right\}^{1/p} \\ &= \max_k |x_k - y_k| \lim_{p\to\infty} (\text{1 以上の有限値})^{1/p} \\ &= \max_k |x_k - y_k| \end{aligned}$$

  - Chebyshev 距離 (最大距離，チェス盤距離) という

- $p \to -\infty$ の場合は以下となることを確認せよ

$$\lim_{p\to-\infty} d(\boldsymbol{x}, \boldsymbol{y}) = \min_k |x_k - y_k|$$

クラスタ間の距離

クラスタ間の距離

- クラスタ : いくつかのデータ点からなる集合

$$C_a = \{\boldsymbol{x}_i | i \in \Lambda_a\}, \; C_b = \left\{\boldsymbol{x}_j | j \in \Lambda_b\right\}, \quad C_a \cap C_b = \emptyset$$

- 2 つのクラスタ間の距離 : $D(C_a, C_b)$
  - データ点の距離から陽に定義する方法
  - クラスタの統合にもとづき再帰的に定義する方法
- 代表的なクラスタ間の距離
  - 最短距離法 (単連結法 ; single linkage method)
  - 最長距離法 (完全連結法 ; complete linkage method)
  - 群平均法 (average linkage method)

最短距離法

- 最も近い対象間の距離を用いる方法

$$D(C_a, C_b) = \min_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \min\bigl\{D(C_a, C_c), D(C_b, C_c)\bigr\}$$

  - $A + B$ は集合 A, B の直和を表す.

最長距離法

- 最も遠い対象間の距離を用いる方法

$$D(C_a, C_b) = \max_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \max\bigl\{D(C_a, C_c), D(C_b, C_c)\bigr\}$$

群平均法

- 全ての対象間の平均距離を用いる方法

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

  - ただし $|C_a|, |C_b|$ はクラスタ内の要素の数を表す

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|}$$

距離計算に関する注意

- データの性質に応じて距離は適宜使い分ける
  - データ間の距離の選択
  - クラスタ間の距離の選択
- 変数の標準化は必要に応じて行う
  - 物理的な意味合いを積極的に利用する場合はそのまま
  - 単位の取り方などによる分析の不確定性を避ける場合は平均 0，分散 1 に標準化
- データの性質を鑑みて適切に前処理

演習

問題

- 以下の問に答えなさい
  - 群平均法におけるクラスタの距離の定義

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

から統合前後のクラスタの距離の関係

$$D(C_a + C_b, C_c) = \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|}$$

を導け

解答例

- 定義に従って計算する

$$\begin{aligned} D(C_a + C_b, C_c) &= \frac{1}{|C_a + C_b||C_c|} \sum_{\boldsymbol{x} \in C_a + C_b,\, \boldsymbol{y} \in C_c} d(\boldsymbol{x}, \boldsymbol{y}) \\ &= \frac{1}{|C_a + C_b||C_c|} \sum_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_c} d(\boldsymbol{x}, \boldsymbol{y}) \\ &\quad + \frac{1}{|C_a + C_b||C_c|} \sum_{\boldsymbol{x} \in C_b,\, \boldsymbol{y} \in C_c} d(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}$$

- (続き)

$$\begin{aligned} &= \frac{|C_a||C_c|}{|C_a + C_b||C_c|} \frac{1}{|C_a||C_c|} \sum_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_c} d(\boldsymbol{x}, \boldsymbol{y}) \\ &\quad + \frac{|C_b||C_c|}{|C_a + C_b||C_c|} \frac{1}{|C_b||C_c|} \sum_{\boldsymbol{x} \in C_b,\, \boldsymbol{y} \in C_c} d(\boldsymbol{x}, \boldsymbol{y}) \\ &= \frac{|C_a||C_c|}{|C_a + C_b||C_c|} D(C_a, C_c) + \frac{|C_b||C_c|}{|C_a + C_b||C_c|} D(C_b, C_c) \\ &= \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|} \end{aligned}$$

解析事例

都道府県別の社会生活統計指標

- 各データを正規化

森林面積割合 ： 森林面積割合（%）2014 年
農業産出額 ： 就業者 1 人当たり農業産出額（販売農家）（万円）2014 年
人口割合 ： 全国総人口に占める人口割合（%）2015 年
土地生産性 ： 土地生産性（耕地面積 1 ヘクタール当たり）（万円）2014 年
商品販売額 ： 商業年間商品販売額［卸売業＋小売業］（事業所当たり）（百万円）2013 年

- 分析方法 : Euclid 距離 + 群平均法

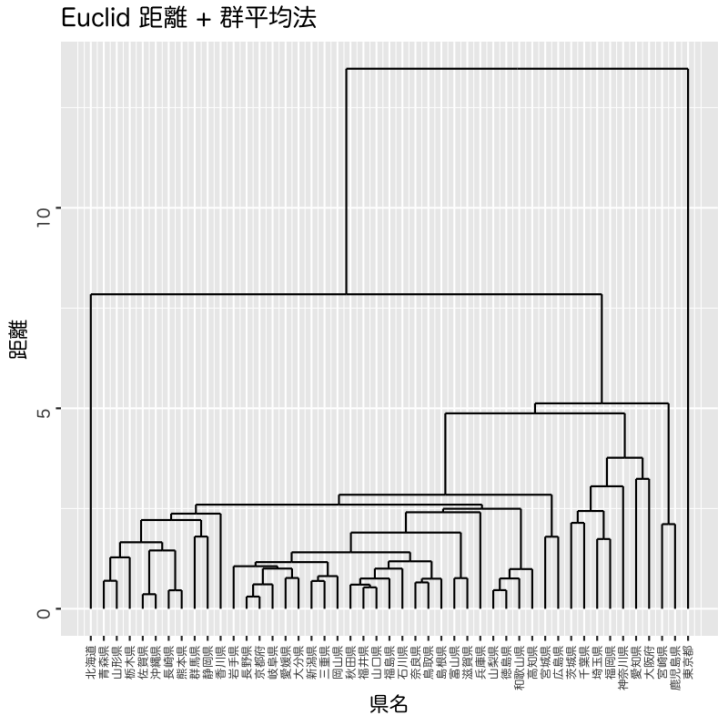

Figure 12: 社会生活統計指標のクラスタ分析 (デンドログラム)

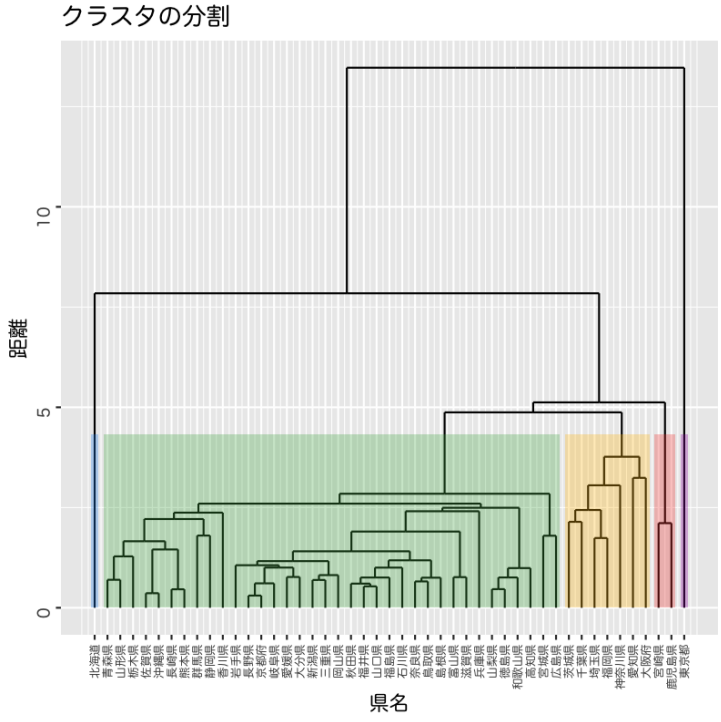

Figure 13: 5 分割の例

都道府県別好きなおむすびの具

- Web アンケート
  - 「ごはんを食べよう国民運動推進協議会」(平成 30 年解散)
    http://www.gohan.gr.jp/result/09/anketo09.html (閉鎖)
  - データ https://noboru-murata.github.io/multivariate-analysis/data/omusubi.csv
- アンケート概要 (Q2 の結果を利用)

  【応募期間】2009 年 1 月 4 日〜2009 年 2 月 28 日
  【応募方法】インターネット、携帯ウェブ
  【内　　容】
  Q1. おむすびを最近 1 週間に、何個食べましたか？
  　　そのうち市販のおむすびは何個でしたか？
  Q2. おむすびの具では何が一番好きですか？
  　　A. 梅 B. 鮭 C. 昆布 D. かつお E. 明太子 F. たらこ G. ツナ H. その他
  Q3. おむすびのことをあなたはなんと呼んでいますか？
  　　A. おにぎり B. おむすび C. その他
  Q4. おむすびといえば、どういう形ですか？
  　　A. 三角形 B. 丸形 C. 俵形 D. その他
  【回答者数】

男性	9,702 人	32.0%
女性	20,616 人	68.0%
総数	30,318 人	100.0%

- 分析方法 : Hellinger 距離 (確率分布の距離) + 群平均法

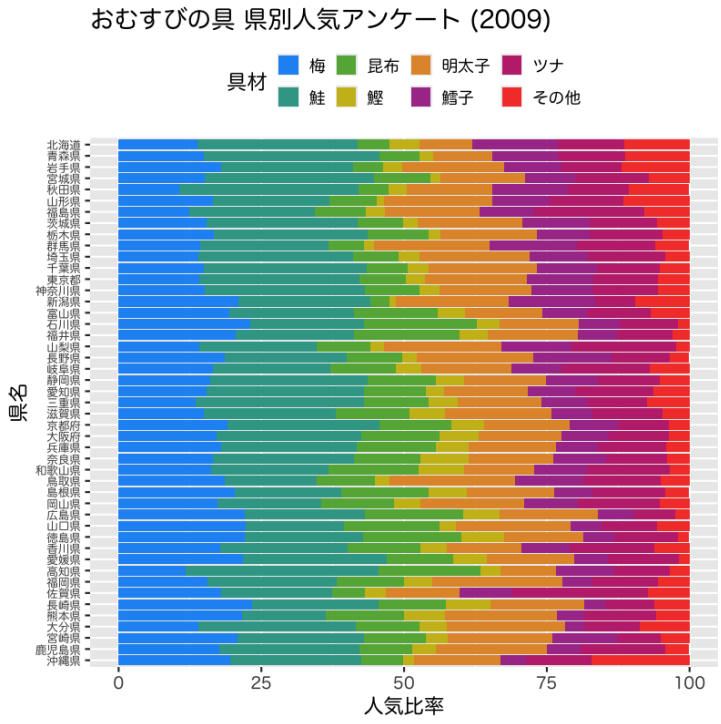

Figure 14: データの概要

次回の予定

- 第 1 回 : 基本的な考え方と階層的方法

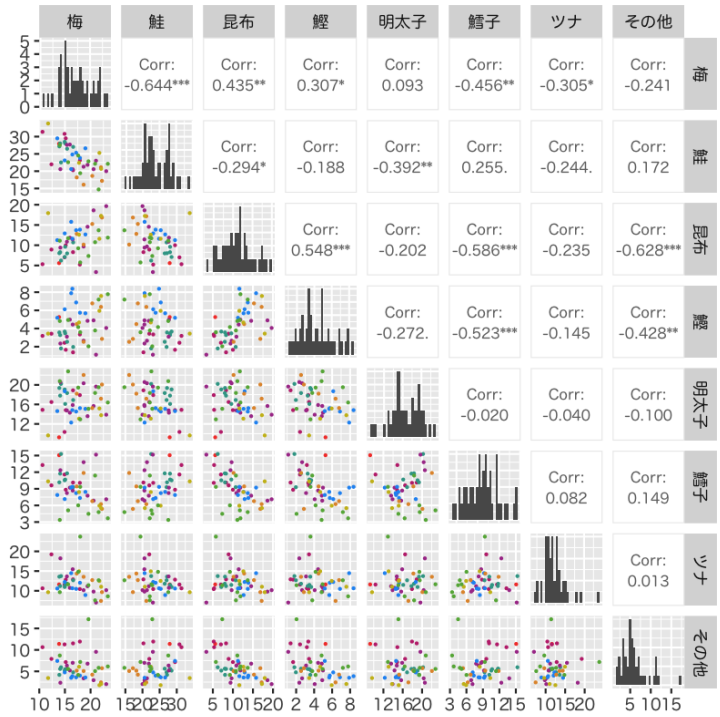

Figure 15: データの散布図

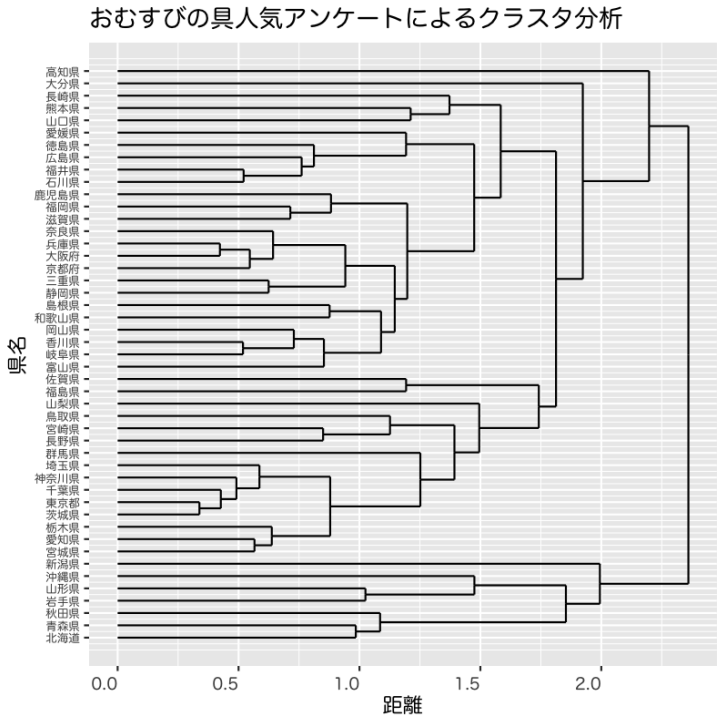

Figure 16: デンドログラム

- **第 2 回 : 非階層的方法と分析の評価**