第9回データ分析勉強会

Rによるクラスター分析実践の紹介

Yuki Watada

2019年2月23日

クラスター分析とは

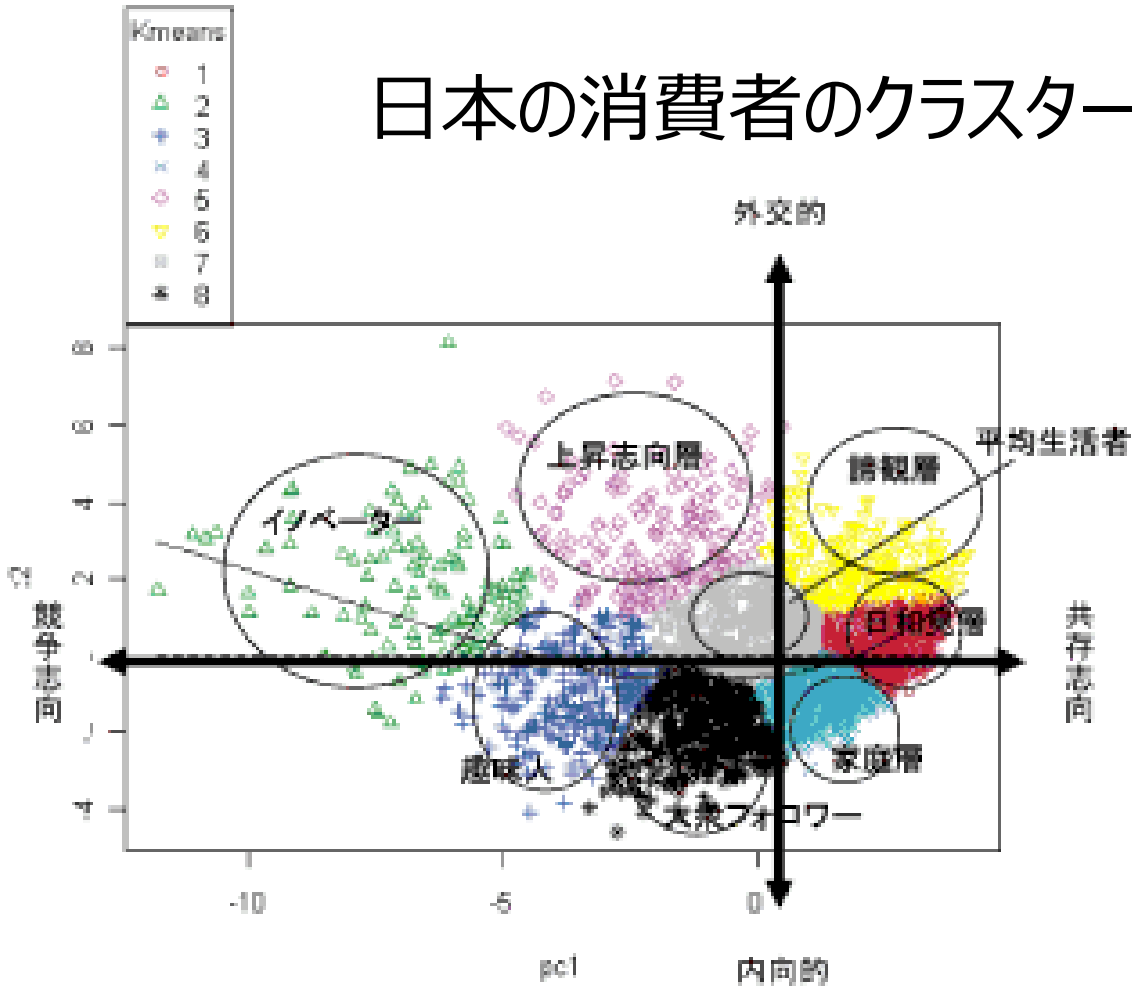
クラスター分析とは、

何かしらの基準や方法で「似た者同士」で括って
分類する方法のことです。

この括った単位を クラスター といいます。
※クラスター(cluster) ＝ 房、集団、群れ

主に**データを分類**したいときに使います。

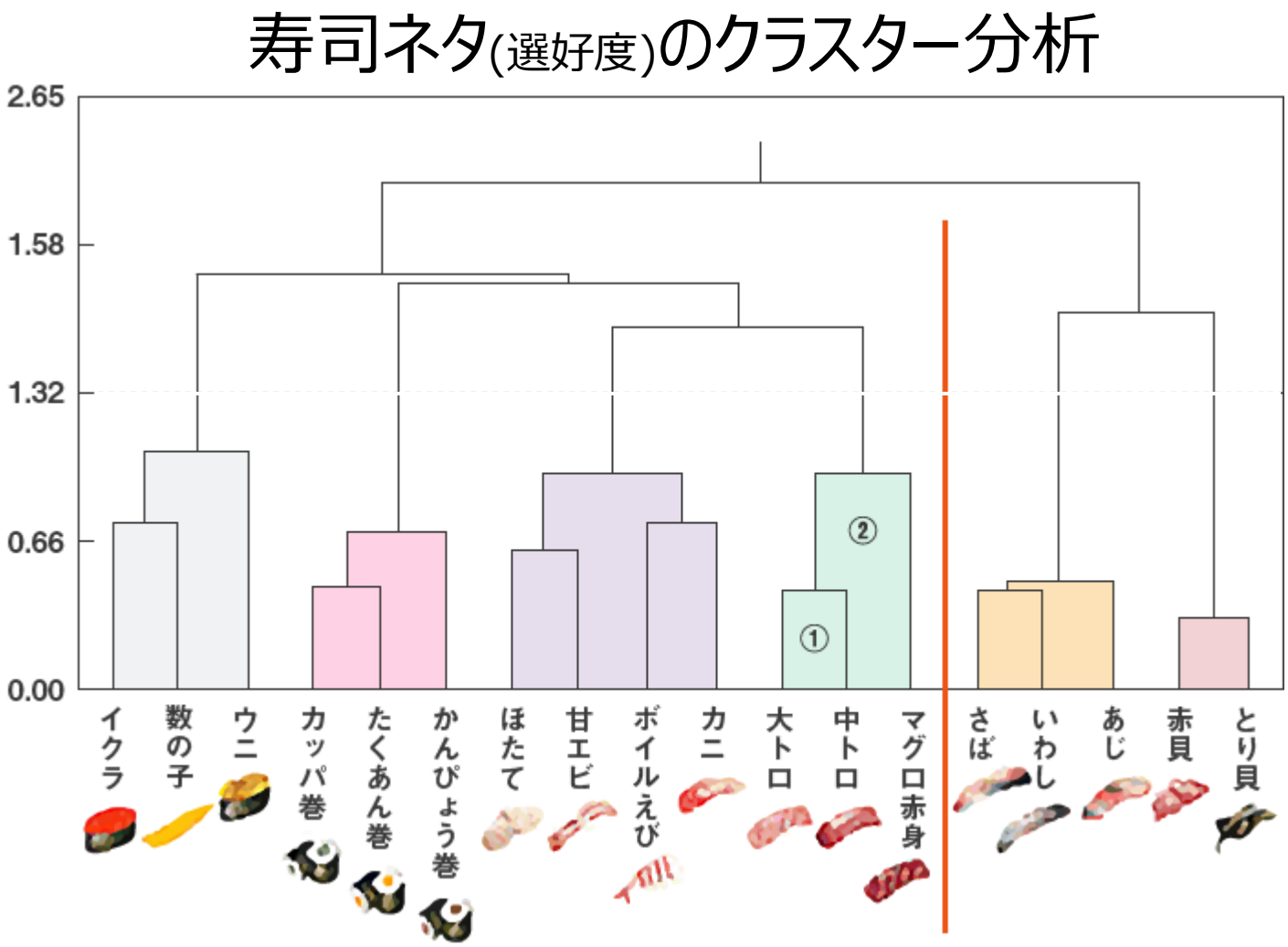
クラスター分析のイメージ (非階層表示型の例)



資料：経済産業省「アジア消費トレンドマップ報告書」。

引用：http://www.meti.go.jp/report/tsuhaku2009/2009honbun/html/i3220000.html

クラスター分析のイメージ (階層表示型の例)



引用：https://www.macromill.com/service/data_analysis/cluster-analysis.html

分類する前に (分類のルール決め)

似た者同士に分類するためには、以下のような決め事を
定義しておく必要があります。

■分類の対象
～何を分類する？～

■分類の形式
～どのように表現する？～

■分類時の対象間の距離(＝類似度)
～どのくらい似てる？似てる感はどのように計る？～

類似度と非類似度

類似度＝似てる感の度合い。ざっくり、観測値間の距離。

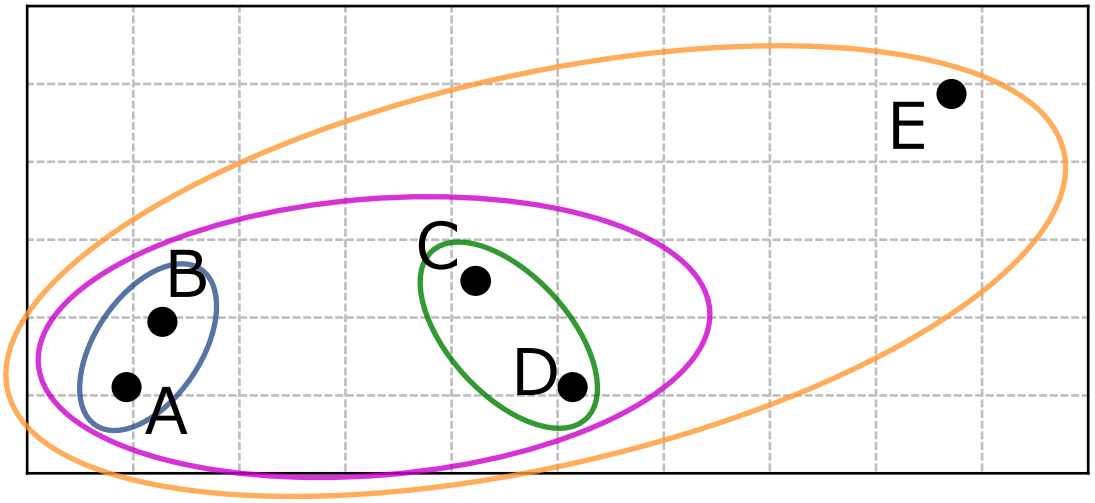
この類似度(観測値間の距離)が大きいほど
観測値同士は似ていないと解釈でき 非類似度 と
呼ばれます。

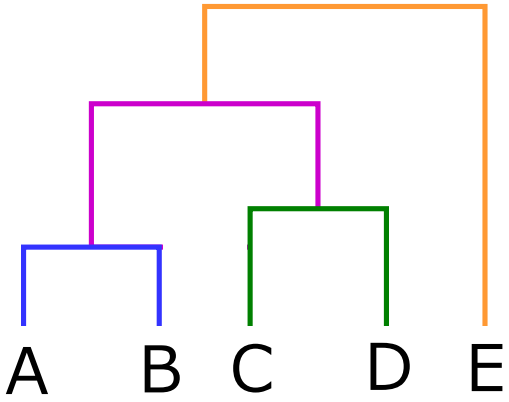
(参考) ユークリッド距離
　類似度を測る指標のひとつ。
　$D_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + \cdots}$

クラスター分析の手順 (階層型クラスター分析の例)

全ての観測値同士の距離を測って分類していきます。

手順1 全ての観測値の組に対して非類似度を計算する。
手順2 非類似度が最小となる観測値の組をひとつのクラスターとしてまとめ、その非類似度を記録する。
手順3 全ての観測値とクラスターの組に対して非類似度を再計算する。
手順4 非類似度が最小となる観測値もしくはクラスターの組をひとつのクラスターとしてまとめ、その非類似度を記録する。
手順5 クラスター数が1個になれば終了、そうでなければ手順3に戻る。





引用：Rによるデータマイニング入門(オーム社)

クラスター分析実践 (概要)

【実践内容】
都道府県を(お酒の消費支出の傾向で)階層的に分類してみました。

■都道府県別年間収入とその消費支出データ(平成21年度)　統計局統計データより

	A	B	C	D	E	F	G	I	M	O	Q	R	X	AB	AC	AD	AE	AF	AG	AH
1	都道府県	集計世帯数	年間収入	消費支出	米	パン	めん類	生鮮魚介	生鮮肉	牛乳	卵	生鮮野菜	菓子類	日本酒	焼酎	ビール	ウィスキー	ワイン	発泡酒	他の酒
2	北海道	2123	5479	267577	3163	2030	1337	4236	3757	1078	579	5538	5693	506	568	1555	136	302	512	173
3	青森県	705	5662	260126	3225	1830	1506	4450	3960	1194	574	5335	5672	632	709	1554	148	194	722	217
4	岩手県	688	5640	273764	3485	1690	1379	4831	3618	1367	644	5893	5518	820	642	1821	96	223	639	184
5	宮城県	736	6609	311136	4584	1990	1400	4729	4174	1528	718	6164	6467	751	563	1357	135	264	413	152
6	秋田県	705	6045	287995	5093	1738	1652	5388	4589	1269	635	6428	6046	1211	775	1606	175	256	1035	162
7	山形県	702	6664	297262	3868	1661	1640	4700	5135	1573	677	7117	6270	864	601	1425	106	164	487	152
8	福島県	890	6482	302849	4702	1838	1389	4207	4129	1373	777	6013	6591	981	723	1184	125	175	660	152
9	茨城県	1332	6457	306588	4439	2034	1357	4194	4046	1382	636	5612	6365	704	454	1285	126	190	332	146
10	栃木県	838	7045	314425	4279	2263	1594	4125	4082	1504	670	5821	6905	600	443	1466	74	220	384	193
11	群馬県	892	6349	294391	3864	2091	1564	3678	3492	1377	604	5500	5979	690	481	1097	131	147	296	130
12	埼玉県	2632	6747	311595	3137	2444	1564	4013	4720	1402	651	6228	6379	603	487	1281	118	324	418	219
13	千葉県	2500	6739	313685	3395	2504	1389	4397	4701	1451	602	6276	6502	593	496	1246	129	228	401	165
14	東京都	2244	7481	323407	2468	2645	1434	4505	5355	1459	658	6823	6650	599	447	1539	142	410	410	222
15	神奈川県	2609	7094	329004	2801	2649	1484	4438	5261	1469	698	6640	6699	523	428	1372	148	303	454	188
16	新潟県	912	6607	318178	7390	2070	1388	4410	4102	1457	710	7092	6115	1385	631	1848	118	233	624	188
17	富山県	712	7252	344212	5137	2564	1411	5037	4340	1497	609	6136	7004	886	393	1195	56	171	619	143

お酒で都道府県を分類

クラスター分析実践 (前準備)

#統計局のデータの読み込み

```
cons <- read.csv("shohidata.csv", row.name = 1)
```

#お酒データ(日本酒～他の酒の列)**の抽出、消費支出当たりの割合を算出**※

```
cons.p <- (cons %>%
             select(日本酒:他の酒) %>%
             sweep(.,1,cons$消費支出,"/"))*100
```

※都道府県で年間収入が違うため、同じ支出額で比較してもお酒にかける傾向(比率)が違うことになるため、消費支出で割って、同じ条件で大小を比較できるようにしている。

	日本酒 <dbl>	焼ちゅう <dbl>	ビール <dbl>	ウィスキー <dbl>	ワイン <dbl>	発泡酒 <dbl>	他の酒 <dbl>
北海道	0.1891044	0.2122753	0.5811411	0.05082649	0.11286471	0.1913468	0.06465429
青森県	0.2429592	0.2725602	0.5974028	0.05689550	0.07457924	0.2775578	0.08342111
岩手県	0.2995281	0.2345086	0.6651715	0.03506670	0.08145702	0.2334127	0.06721117
宮城県	0.2413735	0.1809498	0.4361437	0.04338939	0.08485035	0.1327394	0.04885323
秋田県	0.4204934	0.2691019	0.5576486	0.06076494	0.08889043	0.3593812	0.05625098
山形県	0.2906527	0.2021785	0.4793751	0.03565878	0.05517019	0.1638285	0.05113334

クラスター分析実践 (前準備～非類似度の計算)

#データの標準化(お酒ごとに平均が0、標準偏差が1になるように計算)

```
cons.std <- scale(cons.p)
```

	日本酒	焼ちゅう	ビール	ウィスキー	ワイン	発泡酒	他の酒
北海道	-0.3066467	0.14495070	1.08901890	1.6955014	2.5390146	0.23392624	0.92517592
青森県	0.4297431	0.86347975	1.25425131	2.1711911	0.7305829	1.25130114	2.22689082
岩手県	1.2032450	0.40994597	1.94283873	0.4602476	1.0554580	0.73034569	1.10252821
宮城県	0.4080617	-0.22841503	-0.38427884	1.1125805	1.2157436	-0.45769898	-0.17082539
秋田県	2.8572804	0.82226032	0.85031446	2.4744781	1.4065786	2.21689842	0.34230088
山形県	1.0818864	0.02460807	0.05498922	0.5066549	-0.1862128	-0.09081622	-0.01267121

#非類似度の計算実行

```
cons.dist <- dist(cons.std)
```

```
>cons.dist
 [1] 2.704085 2.661519 2.567338 4.085009 3.582870 3.958126
```

```
>as.matrix(cons.dist)[1:5,1:5]
```

	北海道	青森県	岩手県	宮城県	秋田県
北海道	0.000000	2.704085	2.661519	2.567338	4.085009
青森県	2.704085	0.000000	2.417976	3.728706	3.330287
岩手県	2.661519	2.417976	0.000000	3.152682	3.326806
宮城県	2.567338	3.728706	3.152682	0.000000	4.234946
秋田県	4.085009	3.330287	3.326806	4.234946	0.000000

これが非類似度
数字が0に近いほど
似てることを示します。

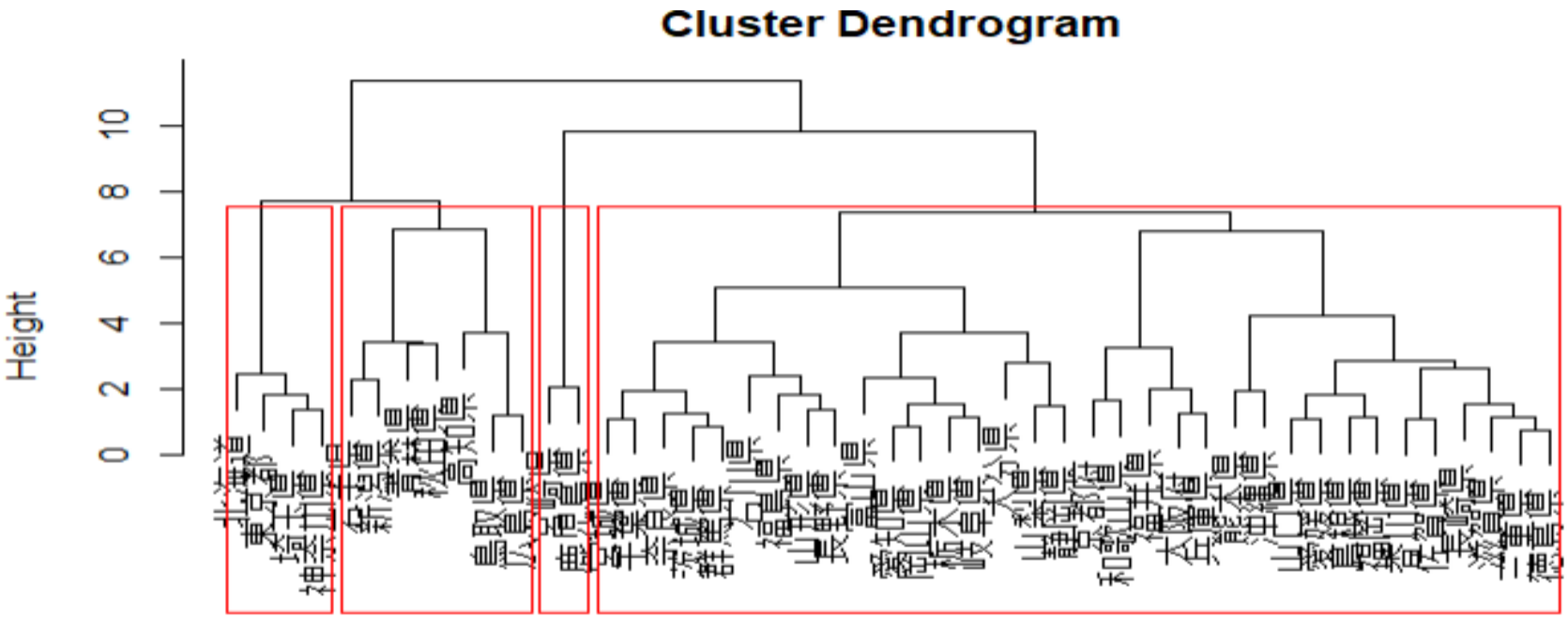
クラスター分析実践 (階層型クラスター分析実行)

#階層型クラスター分析の実行

```
cons.hclust <- hclust(cons.dist, method = "ward.D2")
```

#階層型クラスター分析の実行結果(デンドログラム)の出力

```
plot(cons.hclust)
rect.hclust(cons.hclust, k=4)
```



Enjoy!

クラスター分析＝デンドログラムというわけではありません。
今回は、クラスター分析の結果の表現のひとつを紹介したところまでです。

クラスター分析もまだまだ続きがあります。
他に機械学習に使われる分析手法も多くあります。

次年度もよろしくお願いいたします。