HW-R-02

Динара Руслановна Файзуллина

3 ноября 2018 г.

Кластерный анализ

Были проанализированны данные protein из пакета ClustOfVar о потреблении белковой пищи в странах Европы, состоящие из 25 записей и 9 переменных и собранные A. Weber and cited in Hand et al., A Handbook of Small Data Sets, (1994, p. 297).

Иерархический кластерный анализ

Для разбиения данных на кластеры был произведен иерархический кластерный анализ с помощью методов полных и средних связей. Соответствующиие кластерные диаграммы изображены на Рис. 1.

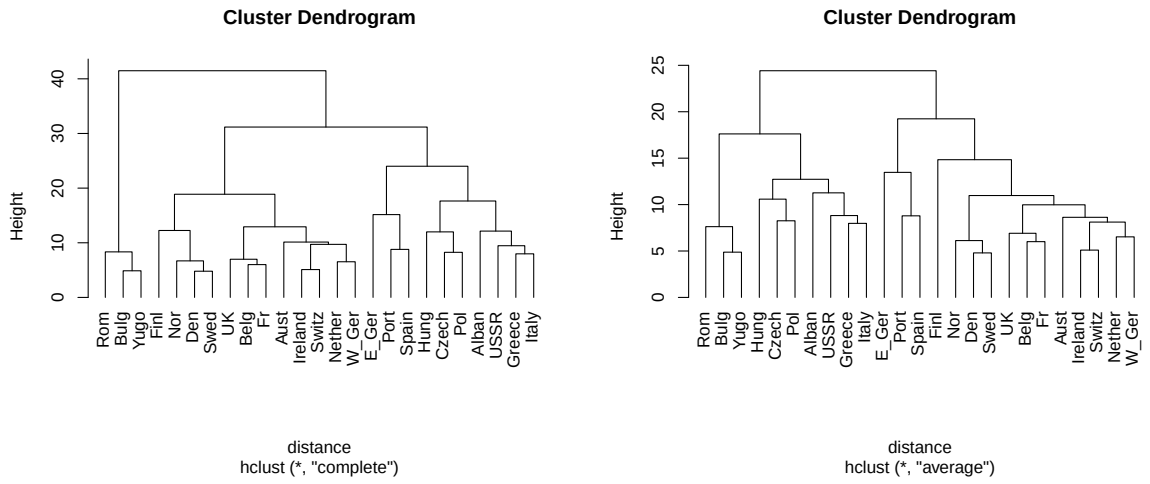

Рис. 1: Дендограммы для полных и средних связей

С помощью графика каменистой осыпи, изображенного на Рис. 2, было определено оптимальное количество кластеров как 3. Однако после анализа результатов

кластеризации и выделенных регионов было принято решение взять количество кластеров как 5, так как они лучше представляют регионы Европы по местоположению и культурной общности.

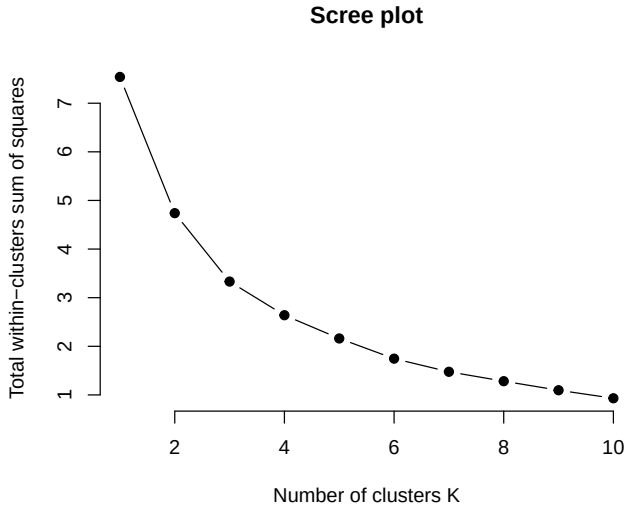

Рис. 2: График каменистой осыпи для данных protein

Также, сравнив результаты разбиения методов полных и медианных связей с помощью Silhouette plot, было определено, что разбиение по методу полных связей лучше выделяет регионы по географическому признаку для данного набора данных, поэтому именно этот метод использовался в дальнейшем. В результате было выделено 5 групп среди стран Европы по потреблению белковой пищи:

1. *Группа 1.1 (Южный и восточный регион)*
   Албания, Чехия, Греция, Венгрия, Италия, Польша, СССР
   Высокое потребление круп, орехов, фруктов и овощей

2. *Группа 1.2 (Центральный и западный регион)*
   Австрия, Бельгия, Франция, Ирландия, Нидерланды, Швейцария, Великобритания, Западная Германия
   Высокое потребление красного и белого мяса, яиц, молока, крахмалистой пищи

3. *Группа 1.3 (Центрально-восточный регион)*
   Болгария, Румыния, Югославия
   Высокое потребление круп и орехов

4. *Группа 1.4 (Северный регион)*
   Дания, Финляндия, Норвегия, Швеция
   Высокое потребление красного и белого мяса, яиц, рыбы

5. *Группа 1.5 (Юго-западный регион)*
   Испания, Португалия, Восточная Германия
   Высокое потребление рыбы, крахмалистой пищи, фруктов и овощей

На Рис. 3 представлен Silhouette plot – метод интерпретации и валидации согласованности данных внутри кластеров. Как мы видим, данные достататочно согласованны, хотя в первом кластере Silhouette width для 1 (Албания) достаточно невелика, а в пятом кластере для 7 (Восточная Германия) данное значение даже отрицательное.

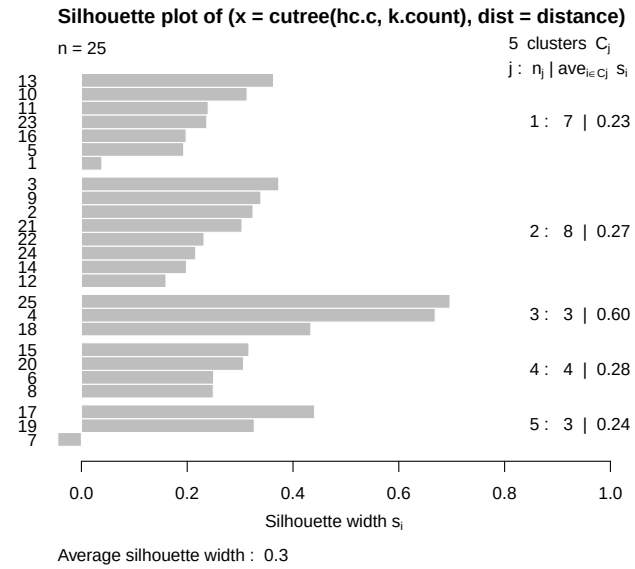

Рис. 3: Silhouette plot

Ниже представлена таблица, содержащая средние значения белка по типу пищи в различных группах:

№	Red.Meat	White.Meat	Eggs	Milk	Fish	Cereals	Starchy	Nuts	Fruit.veg.
1	8.64	6.87	2.39	14.04	2.54	39.27	3.74	4.21	4.66
2	13.21	10.64	3.99	21.16	3.38	24.70	4.65	2.06	4.17
3	6.13	5.77	1.43	9.63	0.93	54.07	2.40	4.90	3.40
4	9.85	7.05	3.15	26.68	8.22	22.68	4.55	1.18	2.12
5	7.23	6.23	2.63	8.20	8.87	26.93	6.03	3.80	6.23

Анализ методом k-средних

С помощью того же графика каменистой осыпи, изображенного на Рис. 2, было определено оптимальное количество кластеров как 3. Однако для сравнения методов разделим набор данных также на 5 кластеров. Результатом выделения кластеров методом к-средних стали 5 групп:

1. *Группа 2.1*
   Восточная Германия, Португалия, Испания
2. *Группа 2.2*
   Албания, Чехия, Греция, Венгрия, Италия, Польша, СССР
3. *Группа 2.3*
   Дания, Финляндия, Норвегия, Швеция
4. *Группа 2.4*
   Болгария, Румыния, Югославия
5. *Группа 2.5*
   Австрия, Бельгия, Франция, Ирландия, Нидерланды, Швейцария, Великобритания, Западная Германия

Разделение на кластеры методом иерархической кластеризации и методом к-средних оказалось идентичным. На Рис. 4 можно наблюдать найденные кластеры на диаграмме рассеивания по типам источников белка: крупы и рыба.

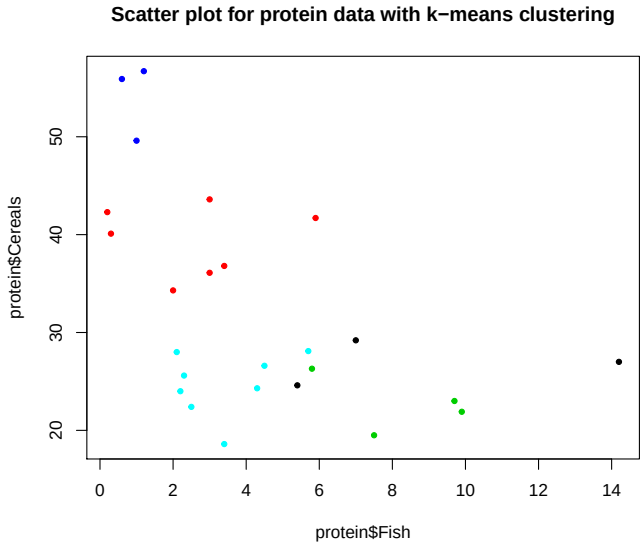

Рис. 4: Scatter plot for Cereals and Fish

Множественная линейная регрессия

Были проанализированы данные Advertising, состоящие из 200 записей и 6 переменных, на зависимость продажи товара от расходов на рекламу по разным каналам.

Для оценки результатов регрессии с помощью метода кросс-валидации исходный набор данных был разбит на тренировочный (80%) и тестовый (20%) наборы. На основе тренировочного набора было проведен регрессионный анализ:

	Estimate	Std.Error	t-value	p-value	Signif.level
(Intercept)	2.950072	0.343926	8.578	9.07e-15	0.001
TV	0.046395	0.001544	30.046	$<$ 2e-16	0.001
radio	0.187679	0.009401	19.963	$<$ 2e-16	0.001
newspaper	-0.003059	0.006430	-0.476	0.635	1

Из таблицы видно, что продажи достаточно сильно зависят от рекламы на ТВ и радио и не зависят от рекламы в газетах.

Зависимость между переменной sales (y) и переменными TV(x_1), radio(x_2), newspaper(x_3) выражается следующим уравнением регрессии:

$$y = 0.046395 * x_1 + 0.187679 * x_2 - 0.003059 * x_3 + 2.950072$$

Заметим, что данная модель объясняет 90% изменчивости данных (Multiple R-squared: 0.9051, Adjusted R-squared: 0.9033)

Доверительные интервалы для каждого коэффициента уравнения с доверительной вероятностью 95%:

	2.5 %	97.5 %
(Intercept)	2.27	3.63
TV	0.04	0.05
radio	0.17	0.21
newspaper	-0.02	0.01

Для оценки качества регрессии проведем корреляционный анализ для ожидамого и полученного значений переменной sales на тренировочном и тестовом наборах. Квадрат корреляции на тренировочном наборе равен 0.9025, что практически совпадает с Adjusted R-squared, равным 0.9033. Квадрат корреляции на тестовом наборе равен 0.8464, что довольно близко к Adjusted R-squared.

Следовательно, можно сделать вывод, что данное уравнение регрессии предсказывает значение зависимой переменной достаточно точно.