Metody i narzędzia eksploracji danych

Porównanie metod grupowania skupień

Ewelina Kamrowska

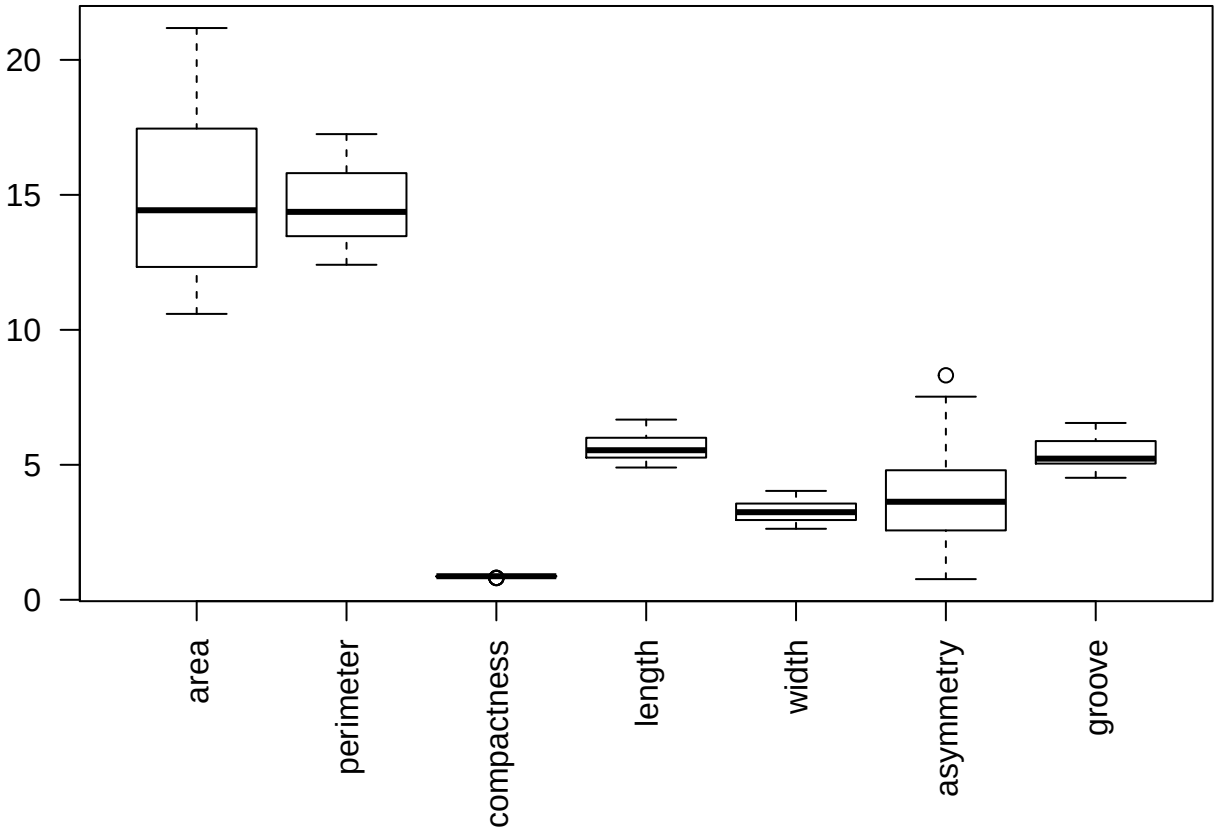
Dane

Przygotowany do analizy zbiór danych składa się z pomiarów właściwości geometrycznych ziarniaków należących do trzech różnych odmian pszenicy, kolejno: *Kamy*, *Rosy* i *Kanady*. Zbiór danych przechowuje zmienne, które opisują właściwości ziaren takie jak: powierzchnia, obwód, spójność, długość i szerokość ziarna, współczynnik asymetrii oraz długość bruzdy ziarna. Ostatnia kolumna zawiera informację o przynależności obserwacji do gatunku ziarna. Dane zostały oczyszczone z braków oraz przeskalowane.

Dane pochodzą ze strony: https://archive.ics.uci.edu/ml/datasets/seeds#.

Podgląd danych

area	perimeter	compactness	length	width	asymmetry	groove	type
15.26	14.84	0.8710	5.763	3.312	2.221	5.220	1
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
14.29	14.09	0.9050	5.291	3.337	2.699	4.825	1
13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1

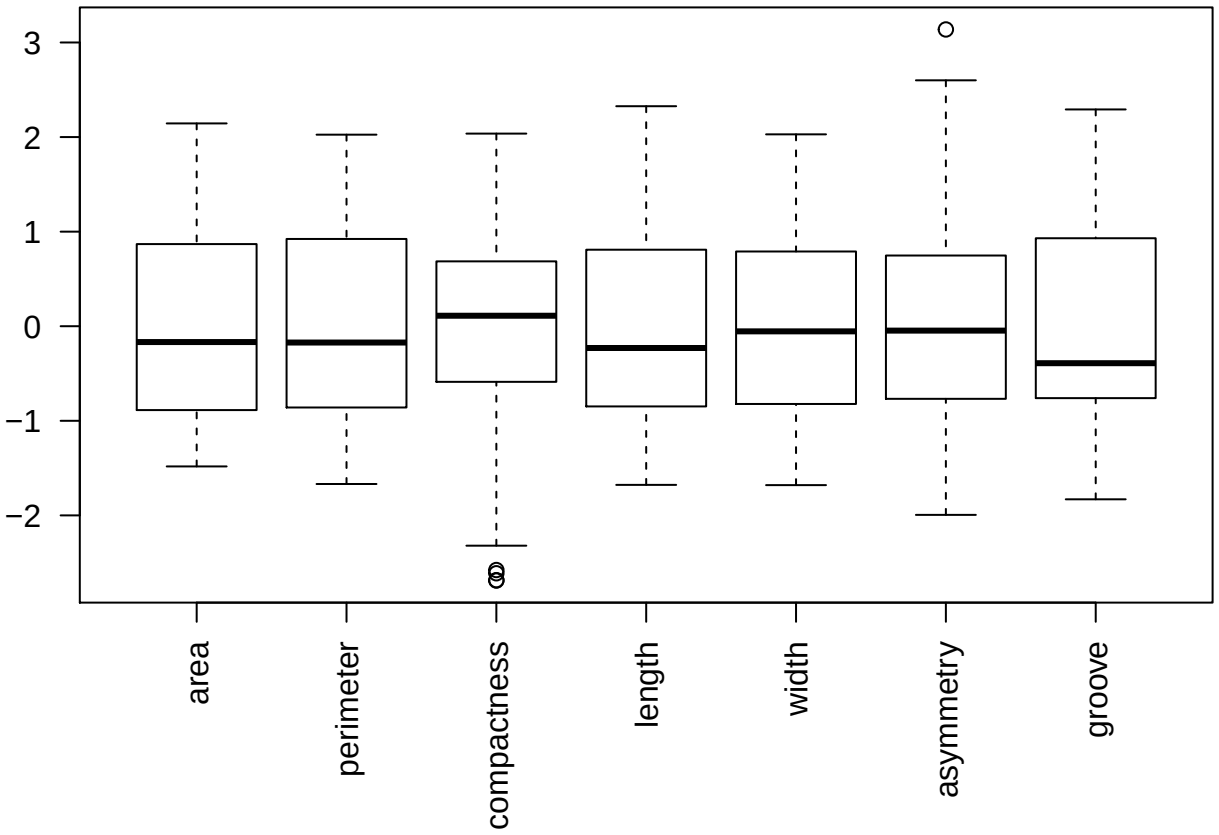


Struktura danych

area	perimeter	compactness	length	width	asymmetry	groove	type
double	double	double	double	double	double	double	double

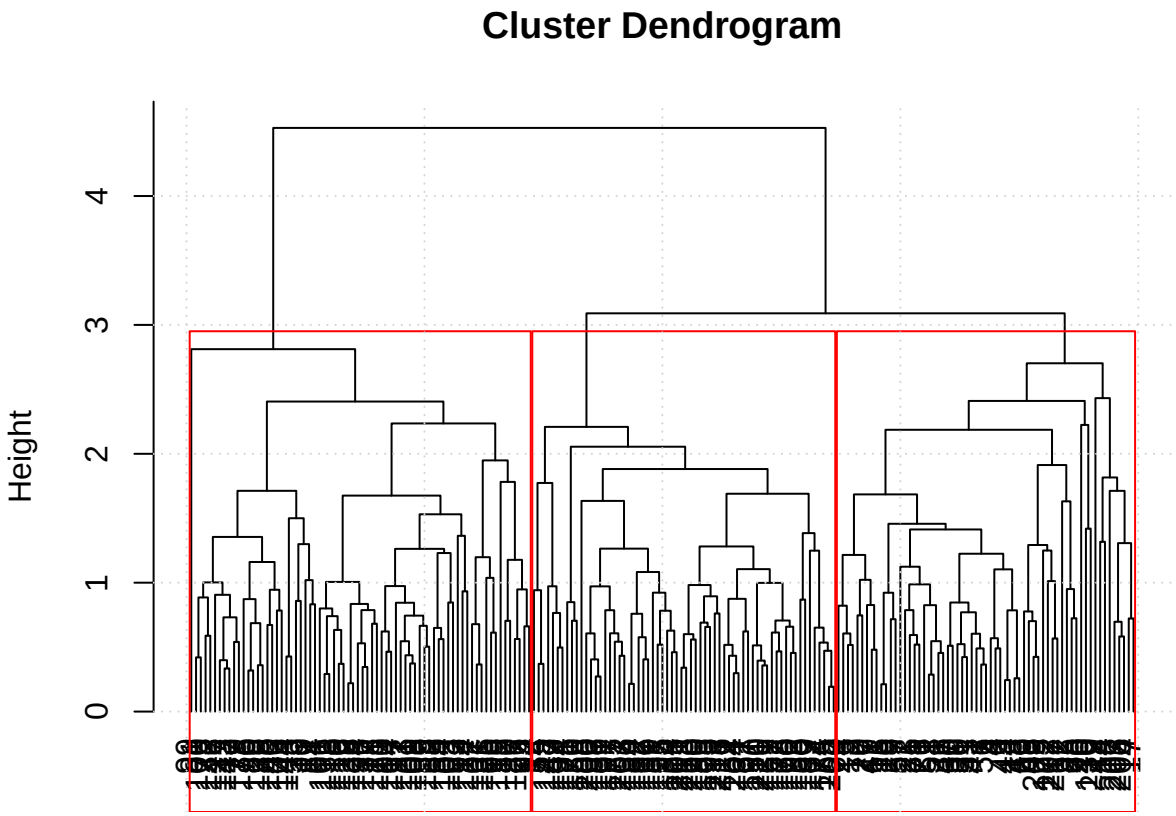
Dane po przeskalowaniu

area	perimeter	compactness	length	width	asymmetry	groove
0.1168696	0.1863267	0.0081238	0.2701781	0.1228250	-1.0048363	-0.4072377
-0.0132685	-0.0197102	0.4412286	-0.2009740	0.1783333	-1.8225905	-0.9430413
-0.2153250	-0.3859981	1.4661003	-0.7938592	0.1889063	-0.6799099	-1.2089135
-0.3694358	-0.5004631	1.0587245	-0.7194668	0.2999230	-0.9790054	-1.2495047
0.4182419	0.3007917	1.3974896	0.0334749	0.7836384	-1.5935106	-0.4985678



Grupowanie hierarchiczne

Ze względu na ilościowy typ wszystkich zmiennych, macierz niepodobieństwa została wyznaczona metodą euklidesową. Jako miara odmienności pomiędzy skupieniami została zastosowana metoda średnich. Następnie dendogram przycięto do trzech skupień.



Poniższe zestawienie ilustruje ile obserwacji zostało zaklasyfikowanych do poszczególnych klastrów.

klaster1	klaster2	klaster3
72	64	63

Klasyfikacja gatunków ziaren do poszczególnych klastrów

Do pierwszego klastra zaklasyfikowano 5 ziaren gatunku *Kamy* oraz 67 ziaren gatunku *Rosy*.

gatunek ziarna	n
1	5
2	67

Do drugiego klastra zaklasyfikowano 7 ziaren gatunku *Kamy* oraz 57 ziaren gatunku *Kanady*.

gatunek ziarna	n
1	7
3	57

Do trzeciego klastra zaklasyfikowano 54 ziaren gatunku *Kamy*, 1 ziarno gatunku *Rosy* oraz 8 ziaren gatunku *Kanady*.

gatunek ziarna	n
1	54
2	1
3	8

Charakterystyka obserwacji zaklasyfikowanych do poszczególnych klastrów

- Podsumowanie obserwacji zaklasyfikowanych do pierwszego klastra:

```
##       area          perimeter      compactness         length     
##  Min.   :15.38   Min.   :14.86   Min.   :0.8452   Min.   :5.718  
##  1st Qu.:17.06   1st Qu.:15.66   1st Qu.:0.8723   1st Qu.:5.979  
##  Median :18.62   Median :16.20   Median :0.8818   Median :6.141  
##  Mean   :18.24   Mean   :16.11   Mean   :0.8823   Mean   :6.150  
##  3rd Qu.:19.13   3rd Qu.:16.53   3rd Qu.:0.8926   3rd Qu.:6.306  
##  Max.   :21.18   Max.   :17.25   Max.   :0.9108   Max.   :6.675  
##      width         asymmetry         groove     
##  Min.   :3.231   Min.   :1.472   Min.   :5.228  
##  1st Qu.:3.500   1st Qu.:2.799   1st Qu.:5.870  
##  Median :3.688   Median :3.572   Median :5.966  
##  Mean   :3.663   Mean   :3.615   Mean   :6.005  
##  3rd Qu.:3.797   3rd Qu.:4.454   3rd Qu.:6.186  
##  Max.   :4.033   Max.   :6.682   Max.   :6.550
```

- Podsumowanie obserwacji zaklasyfikowanych do drugiego klastra:

```
##       area          perimeter      compactness         length     
##  Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899  
##  1st Qu.:11.27   1st Qu.:13.03   1st Qu.:0.8335   1st Qu.:5.175  
##  Median :11.84   Median :13.32   Median :0.8486   Median :5.256  
##  Mean   :11.99   Mean   :13.34   Mean   :0.8457   Mean   :5.277  
##  3rd Qu.:12.47   3rd Qu.:13.57   3rd Qu.:0.8591   3rd Qu.:5.394  
##  Max.   :15.26   Max.   :14.85   Max.   :0.8706   Max.   :5.717  
##      width         asymmetry         groove     
##  Min.   :2.630   Min.   :1.661   Min.   :4.794  
##  1st Qu.:2.739   1st Qu.:4.059   1st Qu.:5.036  
##  Median :2.834   Median :4.764   Median :5.141  
##  Mean   :2.850   Mean   :4.721   Mean   :5.147  
##  3rd Qu.:2.967   3rd Qu.:5.391   3rd Qu.:5.281  
##  Max.   :3.242   Max.   :7.524   Max.   :5.491
```

- Podsumowanie obserwacji zaklasyfikowanych do trzeciego klastra:

```
##       area          perimeter      compactness         length     
##  Min.   :11.23   Min.   :12.63   Min.   :0.8392   Min.   :4.902  
##  1st Qu.:13.18   1st Qu.:13.77   1st Qu.:0.8736   1st Qu.:5.277  
##  Median :14.16   Median :14.21   Median :0.8823   Median :5.439  
##  Mean   :14.10   Mean   :14.15   Mean   :0.8832   Mean   :5.437  
##  3rd Qu.:14.90   3rd Qu.:14.56   3rd Qu.:0.8933   3rd Qu.:5.614  
##  Max.   :17.08   Max.   :15.38   Max.   :0.9183   Max.   :5.833  
##      width         asymmetry          groove     
##  Min.   :2.879   Min.   :0.7651   Min.   :4.519  
##  1st Qu.:3.127   1st Qu.:1.7790   1st Qu.:4.870  
##  Median :3.212   Median :2.5040   Median :5.056  
##  Mean   :3.233   Mean   :2.7565   Mean   :5.031
```

```
##  3rd Qu.:3.357   3rd Qu.:3.2740   3rd Qu.:5.176
##  Max.   :3.683   Max.   :8.3150   Max.   :5.487
```

Obserwacje zaklasyfikowane do pierwszego, najliczniejszego klastra (w odróżnieniu od pozostałych klastrów) charakteryzuje największa powierzchnia oraz obwód ziaren. Obserwacje zaklasyfikowane do drugiego klastra, wyróżniają się najwyższym (pod względem średniej) współczynnikiem asymetrii oraz najmniejszą powierzchnią i obwodem ziaren. Z kolei obserwacje zaklasyfikowane do trzeciego klastra, może wyróżniać najmniejszy, również pod względem średniej, współczynnik asymetrii ziarna.

Grupowanie skupień metodą k-means

W grupowaniu skupień metodą k-means również wybrano podział na trzy klastry. Poniższe zestawienie ilustruje ile ziaren zostało ogólem zaklasyfikowanych do poszczególnych klastrów.

klaster1	klaster2	klaster3
60	70	69

Suma kwadratów odległości wartości od średnich w klastrach wynosi 70.7%.

Klasyfikacja gatunków ziaren do poszczególnych klastrów

Do pierwszego klastra zaklasyfikowano 61 ziaren gatunku *Kamy*, 2 ziarna gatunku *Rosy* oraz 7 ziaren gatunku *Kanady*.

gatunek ziarna	n
1	2
3	58

Do drugiego klastra zaklasyfikowano 2 ziarna gatunku *Kamy* oraz 58 ziaren gatunku *Kanady*.

gatunek ziarna	n
1	61
2	2
3	7

Do trzeciego klastra zaklasyfikowano 3 ziarna gatunku *Kamy* oraz 66 ziaren gatunku *Rosy*.

gatunek ziarna	n
1	3
2	66

Charakterystyka obserwacji zaklasyfikowanych do poszczególnych klastrów

- Podsumowanie obserwacji zaklasyfikowanych do pierwszego klastra:

```
##       area          perimeter      compactness          length
##  Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
##  1st Qu.:11.25   1st Qu.:13.00   1st Qu.:0.8332   1st Qu.:5.173
##  Median :11.81   Median :13.29   Median :0.8480   Median :5.236
```

```
##  Mean   :11.84   Mean   :13.26   Mean   :0.8457   Mean   :5.246
##  3rd Qu.:12.39   3rd Qu.:13.52   3rd Qu.:0.8591   3rd Qu.:5.353
##  Max.   :13.34   Max.   :13.95   Max.   :0.8883   Max.   :5.541
##      width         asymmetry         groove
##  Min.   :2.630   Min.   :2.221   Min.   :4.794
##  1st Qu.:2.719   1st Qu.:4.125   1st Qu.:5.003
##  Median :2.821   Median :4.878   Median :5.132
##  Mean   :2.835   Mean   :4.901   Mean   :5.134
##  3rd Qu.:2.923   3rd Qu.:5.470   3rd Qu.:5.270
##  Max.   :3.232   Max.   :8.315   Max.   :5.491
```

- Podsumowanie obserwacji zaklasyfikowanych do drugiego klastra:

```
##       area         perimeter      compactness         length
##  Min.   :11.23   Min.   :12.63   Min.   :0.8392   Min.   :4.902
##  1st Qu.:13.39   1st Qu.:13.83   1st Qu.:0.8700   1st Qu.:5.349
##  Median :14.29   Median :14.27   Median :0.8805   Median :5.493
##  Mean   :14.15   Mean   :14.20   Mean   :0.8800   Mean   :5.470
##  3rd Qu.:15.01   3rd Qu.:14.65   3rd Qu.:0.8882   3rd Qu.:5.658
##  Max.   :16.20   Max.   :15.27   Max.   :0.9183   Max.   :5.877
##      width         asymmetry          groove
##  Min.   :2.879   Min.   :0.7651   Min.   :4.519
##  1st Qu.:3.121   1st Qu.:1.8265   1st Qu.:4.883
##  Median :3.207   Median :2.6935   Median :5.091
##  Mean   :3.226   Mean   :2.7501   Mean   :5.069
##  3rd Qu.:3.362   3rd Qu.:3.4022   3rd Qu.:5.209
##  Max.   :3.582   Max.   :6.6850   Max.   :5.752
```

- Podsumowanie obserwacji zaklasyfikowanych do trzeciego klastra:

```
##       area         perimeter      compactness         length
##  Min.   :15.38   Min.   :14.89   Min.   :0.8452   Min.   :5.718
##  1st Qu.:17.32   1st Qu.:15.73   1st Qu.:0.8735   1st Qu.:5.980
##  Median :18.72   Median :16.22   Median :0.8823   Median :6.145
##  Mean   :18.37   Mean   :16.16   Mean   :0.8833   Mean   :6.164
##  3rd Qu.:19.14   3rd Qu.:16.57   3rd Qu.:0.8969   3rd Qu.:6.315
##  Max.   :21.18   Max.   :17.25   Max.   :0.9108   Max.   :6.675
##      width         asymmetry         groove
##  Min.   :3.231   Min.   :1.472   Min.   :5.484
##  1st Qu.:3.552   1st Qu.:2.837   1st Qu.:5.877
##  Median :3.693   Median :3.619   Median :5.971
##  Mean   :3.680   Mean   :3.617   Mean   :6.026
##  3rd Qu.:3.801   3rd Qu.:4.451   3rd Qu.:6.187
##  Max.   :4.033   Max.   :6.682   Max.   :6.550
```

Obserwacje zaklasyfikowane do klastra pierwszego, charakteryzuje najmniejsza powierzchnia i obwód ziarna (pod względem średniej). Obserwacje zaklasyfikowane do klastra drugiego mogą wyróżniać się najmniejszym współczynnikiem asymetrii ziarna. Obserwacje zaklasyfikowane do klastra trzeciego, charakteryzuje natomiast największa powierzchnia i obwód ziarna.

Środki wyznaczone przez metodę grupowania k-means

```
##         area  perimeter compactness     length      width   asymmetry
## 1 -1.0542589 -1.0206428  -1.0771395 -0.8956792 -1.1374344  0.81671154
## 2 -0.2623899 -0.2986777   0.3952839 -0.3893705 -0.1054389 -0.64518185
## 3  1.1829396  1.1905219   0.5356304  1.1738650  1.0960404 -0.05565164
```

```
##       groove
## 1 -0.5810356
## 2 -0.7132082
## 3  1.2287929
```

Badanie jakości klastrów

W celu zbadania jakości powstałych klastrów dla obu metod grupowania posłużono się *indeksem zarysu* oraz *indeksem Calińskiego*.

nazwa metody	indeks zarysu	indeks Calinskiego
hierarchical clustering	0.3925916	220.8677
k-means	0.4043846	236.1354

W obu przeprowadzonych metodach, wartość indeksu zarysu waha się od $[0.39 - 0.41]$. Mimo faktu, iż indeks zarysu jest nieco wyższy dla grupowania skupień metodą k-means od metody skupień grupowania hierarchicznego, można stwierdzić, że wartość indeksu zarysu dla obu metod świadczy o słabej jakości klasyfikacji obserwacji do poszczególnych grup (o słabej strukturze klas). Indeks Calińskiego również jest wyższy dla grupowania metodą k-means. Podsumowując, można stwierdzić, że przy zaproponowanym podziale obserwacji na trzy klasy, nieco lepsze wyniki dała metoda k-means od metody grupowania hierarchicznego.