11. Analiza skupień

dr inż. Urszula Libal

Politechnika Wrocławska

2015

1. Analiza skupień

Określenia:

— analiza skupień (*cluster analysis*),

— klasteryzacja (*clustering*),

— klasyfikacja nienadzorowana (*unsupervised classification*).

Idea działania - grupowanie obiektów podobnych.

2. Hierarchiczne metody klasteryzacji

Polegają na automatycznym wiązaniu skupień.

Do takich metod należą:

— metoda **pojedynczego wiązania** (ang. *single linkage*, SLINK),

— metoda **pełnego wiązania** (ang. *complete linkage*, CLINK),

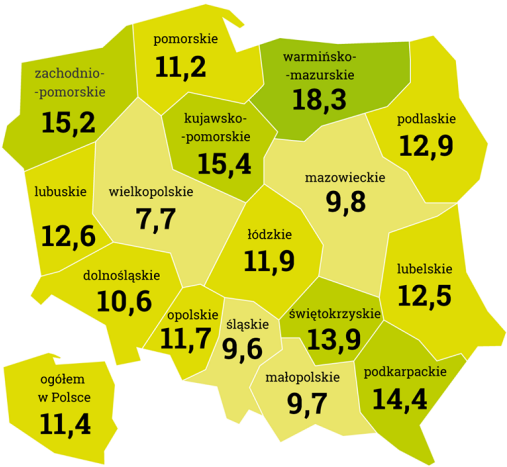
— metoda **wiązania średniego** (ang. *average linkage*, UPGMA).

Najczęściej stosuje się dwa podejścia do tworzenia klastrów:

— **aglomeracyjne** (ang. *agglomerative*) - za pomocą łączenia skupień wyodrębnionych w poprzednich krokach w większe skupienia,

— **rozdzielające** (ang. *divisive*) - za pomocą podziału skupień na mniejsze skupienia.

3. Przykład klasteryzacji hierarchicznej



Rysunek 1. Przykładowe dane: stopa bezrobocia według województw (listopad 2014r.).

Źródło: GUS

4. Metoda wiązania średniego (UPGMA)

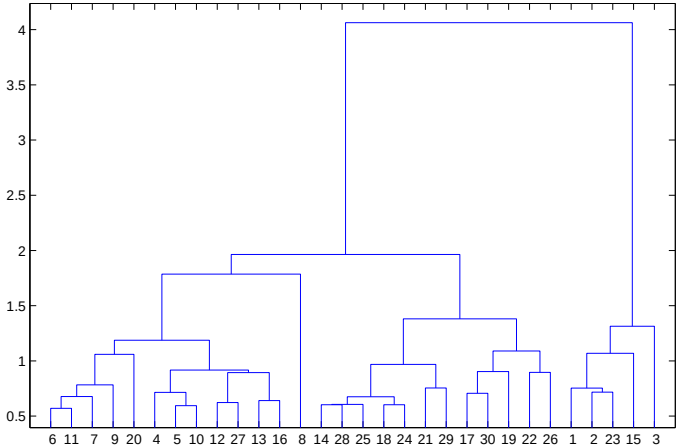
— Metoda średniego wiązania (ang. *average linkage*, lub *unweighted pair-group method using arithmetic averages* - w skrócie UPGMA) [2,3] opiera się na średniej mierze niepodobieństwa między parami obiektów pochodzących z różnych klastrów.

— Miarę niepodobieństwa ρ_c między klastrami C_1i C_2 wyliczamy na podstawie wzoru

$$\rho_c(C_1, C_2) = \frac{1}{N_1 N_2} \sum_{i \in C_1} \sum_{j \in C_2} \rho(x_i, x_j), \tag{1}$$

gdzie N_1 i N_2 to odpowiednie liczności klastrów C_1 i C_2.

— Jest to najpopularniejsza metoda wyznaczania klastrów.



Rysunek 2. Drzewo binarne uzyskane metodą średniego wiązania.
Na osi poziomej zaznaczone są numery obiektów.

Źródło: opracowanie własne na podstawie danych [1]

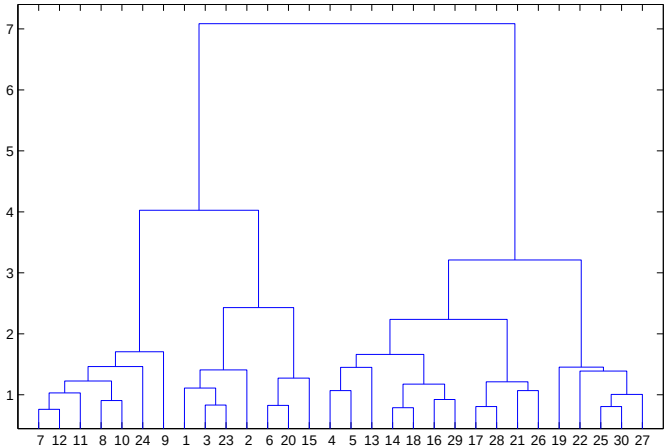
5. Metoda pełnego wiązania (CLINK)

— Drugie podejście to metoda pełnego wiązania (ang. *complete linkage*) [2,3,4,5], która jest znana również jako metoda **najdalszego sąsiedztwa** lub **najdalszej odległości**.

— Podział na klastry odbywa się wieloetapowo, w każdym kroku maksymalizowana jest miara niepodobieństwa ρ_c

$$\rho_c(C_1, C_2) = \max_{i \in C_1,\, j \in C_2} \rho(x_i, x_j). \tag{2}$$

— Skutkuje to silnym skupieniem obiektów wewnątrz klastrów.



Rysunek 3. Drzewo binarne uzyskane metodą pełnego wiązania.

Źródło: opracowanie własne na podstawie danych [1]

6. Metoda pojedynczego wiązania (SLINK)

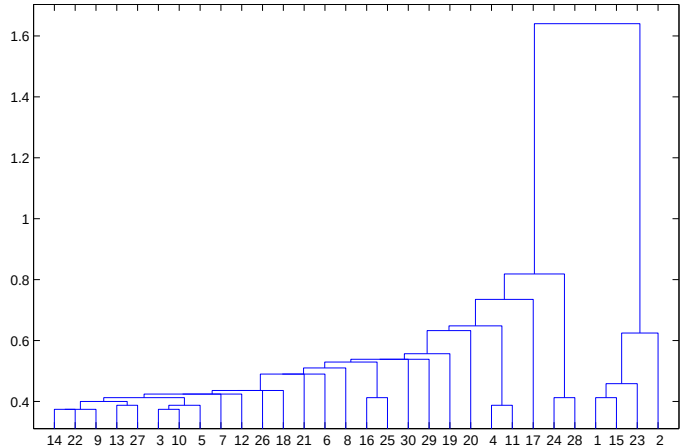
— Metoda pojedynczego wiązania (ang. *single linkage*) [2,3,4,5] jest zwana także metodą **najbliższego sąsiedztwa** lub **najbliższej odległości**. W literaturze funkcjonuje również pod nazwą **„taksonomii wrocławskiej"**.

— Podział obiektów na klastry następuje wieloetapowo. W pierwszym kroku dokonujemy podziału wszystkich obiektów na dwa klastry C_1 i C_2 dzięki maksymalizacji miary niepodobieństwa między klastrami ρ_c

$$\rho_c(C_1, C_2) = \min_{i \in C_1,\, j \in C_2} \rho(x_i, x_j). \tag{3}$$

— W kolejnych krokach podziałowi na dwa nowe klastry ulegają obiekty zakwalifikowane w poprzednim kroku do wspólnego skupiska. Podział na podstawie miary ρ_c można wykonywać aż do osiągnięcia klastrów składających się z pojedynczych obiektów.

— W praktyce nie jest konieczne dokonanie wszystkich możliwych kroków tej metody, a jedynie kilku początkowych.



Rysunek 4. Drzewo binarne uzyskane metodą pojedynczego wiązania.

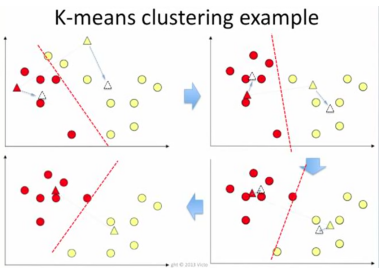
Źródło: opracowanie własne na podstawie danych [1]

7. Metoda k-średnich

— Podział obiektów na k klastrów, skupionych wokół centrów.

— W pierwszym kroku położenie centrów ustalone lub wybrane losowo.

— W kolejnych krokach położenia centrów klastrów są “poprawiane”.



Rysunek 5. Klasteryzacja k-średnich.

Źródło: [6]

Literatura

[1] R.A. Fisher, *The use of multiple measurements in axonomic problems*,

Annals of Eugenics, Vol. 7 (1936) pp. 179-188

[2] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*, WNT, Warszawa (2005)

[3] A.R. Webb, K.D. Copsey, *Statistical Pattern Recognition*, 3rd ed., Wiley, (2011)

[4] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd ed., Wiley, (2000)

[5] M. Krzyśko, W. Wołyński, T. Górecki, M. Skorzybut, *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. WNT, Warszawa (2008)

[6] V. Lavrenko, *Clustering 4: K-means algorithm*,

`https://www.youtube.com/watch?v=_aWzGGNrcic`