

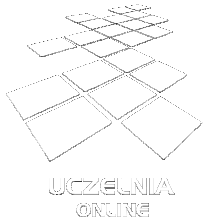
Grupowanie

Wprowadzanie

Definicja problemu

**Klasyfikacja metod
grupowania**

Grupowanie hierarchiczne

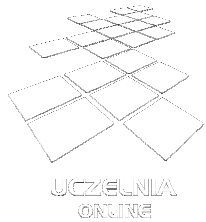


Sformułowanie problemu

Dany jest zbiór obiektów (rekordów).
Znajdź naturalne pogrupowanie obiektów w klasy
(klastry, skupienia) obiektów o podobnych cechach

- **Grupowanie**

proces grupowania obiektów, rzeczywistych bądź abstrakcyjnych, w klasy, nazywane klastrami lub skupieniami, o podobnych cechach



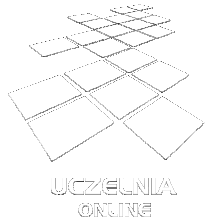
Czym jest klaster?

- Istnieje wiele definicji

Zbiór obiektów, które są “podobne”

Zbiór obiektów, takich, że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza aniżeli odległość pomiędzy dowolnym obiektem należącym do klastra i dowolnym obiektem nie należącym do tego klastra

Spójny obszar przestrzeni wielowymiarowej, charakteryzujący się dużą gęstością występowania obiektów



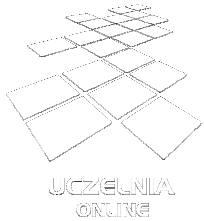
Przykłady (1)

- **Zbiór dokumentów**

Zbiór punktów w przestrzeni wielowymiarowej, w której pojedynczy wymiar odpowiada jednemu słowu z określonego słownika

Współrzędne dokumentu w przestrzeni są zdefiniowane względną częstością występowania słów ze słownika.

Klastry dokumentów odpowiadają grupom dokumentów dotyczących podobnej tematyki

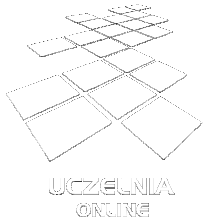


Przykłady (2)

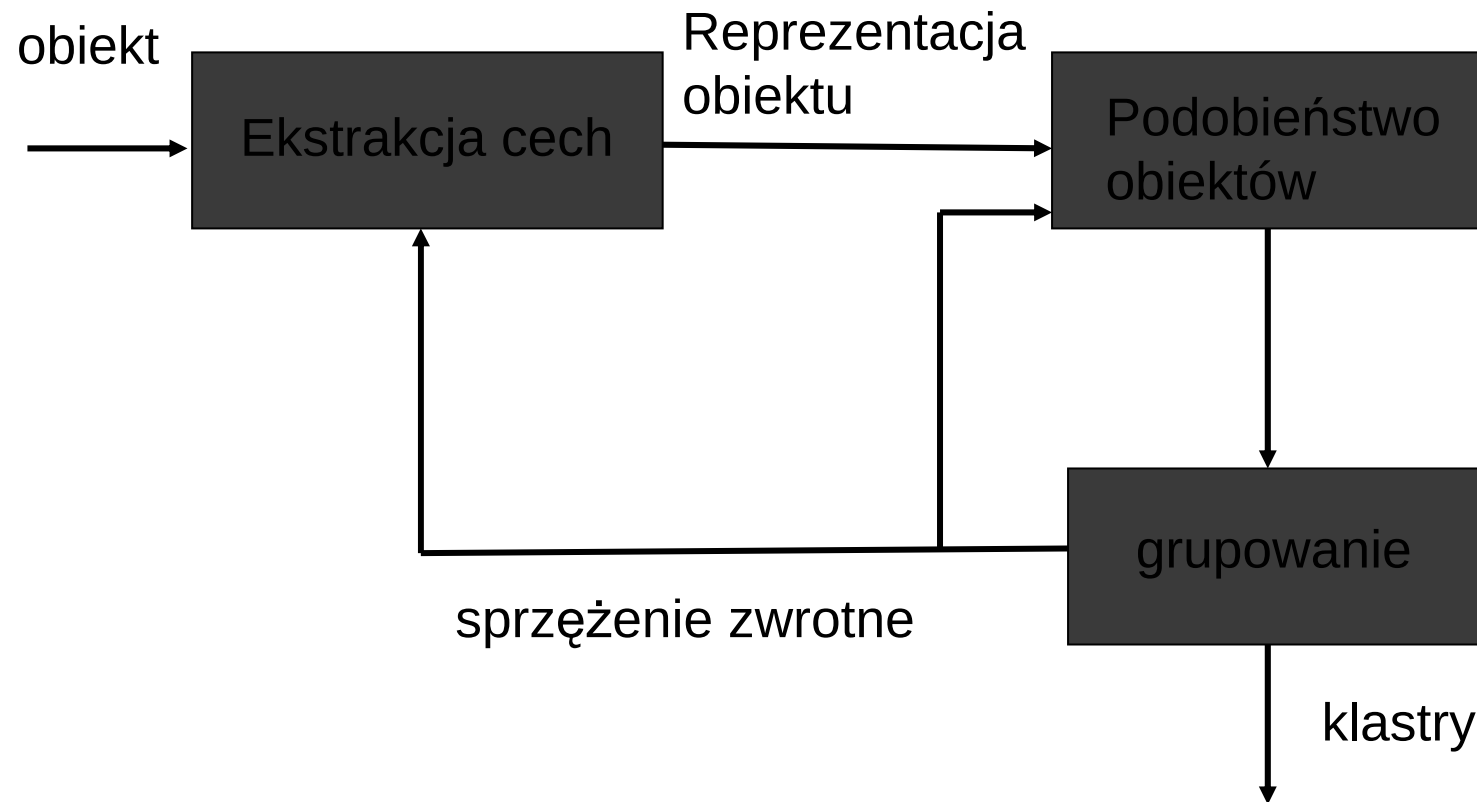
- **Zbiór sekwencji stron WWW**

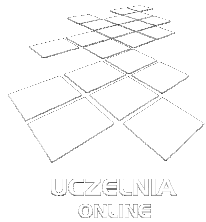
Pojedyncza sekwencja opisuje sekwencję dostępów do stron WWW danego serwera realizowaną w ramach jednej sesji przez użytkownika

Klastry sekwencji odpowiadają grupom użytkowników danego serwera, którzy realizowali dostęp do tego serwera w podobny sposób



Składowe procesu grupowania (1)





Składowe procesu grupowania (2)

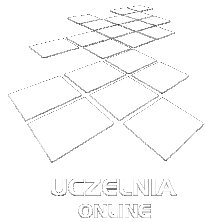
- Proces grupowania

Reprezentacja obiektów
(zawiera ekstrakcję/selekcję cech obiektów)

Definicja miary podobieństwa pomiędzy obiektami
(zależy od dziedziny zastosowań)

Grupowanie obiektów (klastry)

Znajdowanie charakterystyki klastrów



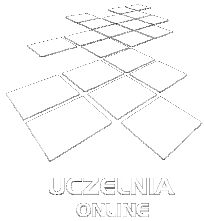
Miary odległości (1)

- Dyskusja dotycząca podobieństwa, lub odległości, dwóch obiektów wymaga przyjęcia miary odległości pomiędzy dwoma obiektami x i y reprezentowanymi przez punkty w przestrzeni wielowymiarowej
- Klasyczne aksjomaty dla miary odległości

$$1. D(x, y) = 0 \Leftrightarrow x = y$$

$$2. D(x, y) = D(y, x)$$

$$3. D(x, y) \leq D(x, z) + D(z, y) \text{ (nierówność trójkąta)}$$



Miary odległości (2)

- Dana jest k-wymiarowa przestrzeń euklidesowa, odległość pomiędzy dwoma punktami $x=[x_1, x_2, \dots, x_k]$ oraz $y=[y_1, y_2, \dots, y_k]$ można zdefiniować następująco:

Odległość euklidesowa:

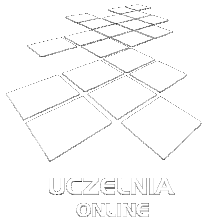
(„norma L_2 ”)

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Odległość Manhattan:

(„miara L_1 ”)

$$\sum_{i=1}^k |x_i - y_i|$$



Miary odległości (3)

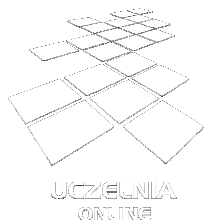
Odległość max z wymiarów:
(„norma L^∞ ”)

$$\max_{i=1}^k |x_i - y_i|$$

Odległość Minkowskiego

$$\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

W przypadku, gdy obiekty nie poddają się transformacji do przestrzeni euklidesowej, proces grupowania wymaga zdefiniowania innych miar odległości (podobieństwa): sekwencja dostępów do stron WWW, sekwencje DNA, sekwencje zbiorów, zbiory atrybutów kategoriycznych, dokumenty tekstowe, XML, grafy, itp..



Inne miary odległości (1)

- Strony WWW: punkty w przestrzeni wielowymiarowej, w której pojedynczy wymiar odpowiada jednemu słowu z określonego słownika
- Idea: Podobieństwo (odległość) $D(x, y)$ stron x i y zdefiniowane jako iloczyn skalarny wektorów reprezentujących strony x i y . Współrzędne dokumentu w przestrzeni są zdefiniowane jako względna częstość występowania słów ze słownika

Np. przyjmijmy słownik składający się z 4 słów:

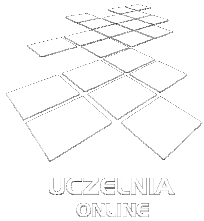
$$x=[2,0,3,1]$$

$$y=[5,3,2,0]$$

$$x \circ y = 16$$

$$\text{długość } x = \sqrt{14}$$

$$\text{długość } y = \sqrt{38}$$



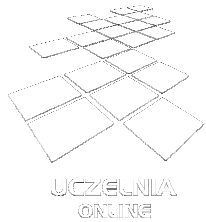
Inne miary odległości (2)

- Miara odległości

$$D(x, y) = 1 - \frac{x \circ y}{|x| * |y|}$$

- $D(x, y) = 1 - \frac{16}{\sqrt{14}\sqrt{38}} \cong 0.3$

Założmy $x = [a_1, a_2, a_3, a_4]$ i $y = [2a_1, 2a_2, 2a_3, 2a_4]$
(strona y jest kopią x), wówczas $D(x, y) = 0$

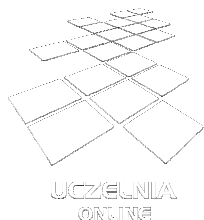


Inne miary odległości (3)

- Sekwencje DNA, sekwencje dostępu do stron WWW: definicja odległości (podobieństwa) sekwencji symboli, powinna uwzględniać fakt, że sekwencje mogą mieć różną długość oraz różne symbole na tych samych pozycjach, np.: $x = abcde$ $y = bcdxye$
- Miara odległości

$$D(x, y) = |x| + |y| - 2 |LCS(x, y)|$$

- gdzie LCS oznacza najdłuższą wspólną podsekwencję (*ang. longest common subsequence*) ($LCS(x, y) = bcde$).
Stąd, $D(x, y) = 3$



Zmienne binarne (1)

- W jaki sposób obliczyć podobieństwo (lub niepodobieństwo) pomiędzy dwoma obiektami opisanymi zmiennymi binarnymi:
- Podejście: konstruujemy macierz niepodobieństwa

		obiekt j		
		1	0	Sum
obiekt i	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

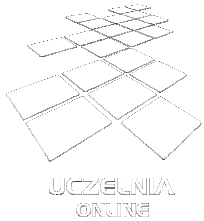
q – liczba zmiennych przyjmujących wartość 1 dla obu obiektów

r – ... 1 dla obiektu i, i wartość 0 dla j

s – ... 0 dla obiektu i, i wartość 1 dla j

t – ... 0 dla obu obiektów

$$p = q + r + s + t - \text{łączna liczba zmiennych}$$



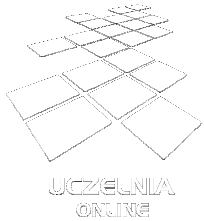
Zmienne binarne (2)

- **Zmienne binarne symetryczne**

Zmienną binarną nazywamy symetryczną jeżeli obie wartości tej zmiennej posiadają tą samą wagę (np. płeć).

Niepodobieństwo pomiędzy obiektami i oraz j jest zdefiniowane następująco

$$d(i, j) = \frac{r + s}{q + r + s + t}$$



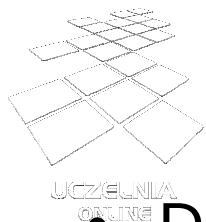
Zmienne binarne (3)

- **Zmienne binarne asymetryczne**

zmienną binarną nazywamy asymetryczną jeżeli obie wartości tej zmiennej posiadają różne wagi (np. wynik badania EKG)

Niepodobieństwo pomiędzy obiektami i oraz j jest zdefiniowane następująco:

$$d(i, j) = \frac{r + s}{q + r + s}$$



Zmienne binarne (4)

- Dana jest tablica zawierająca informacje o pacjentach

imię	ból	gorączka	katar	test1	test2	test3	test4
Jack	Y	Y	N	P	N	N	N
Mary	N	Y	N	P	N	P	N
Jim	Y	Y	Y	N	N	N	N
...	...	...	...	...	...	...	...

$$d_{\text{asym}}(\text{jack}, \text{mary}) = \frac{2}{4} = 0.5$$

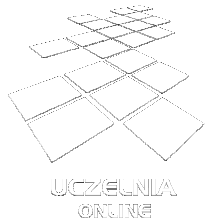
$$d_{\text{asym}}(\text{jack}, \text{jim}) = \frac{2}{4} = 0.5$$

$$d_{\text{asym}}(\text{jim}, \text{mary}) = \frac{4}{5} = 0.8$$

$$d_{\text{sym}}(\text{jack}, \text{mary}) = \frac{2}{7} = 0.29$$

$$d_{\text{sym}}(\text{jack}, \text{jim}) = \frac{2}{7} = 0.29$$

$$d_{\text{sym}}(\text{jim}, \text{mary}) = \frac{4}{7} = 0.57$$

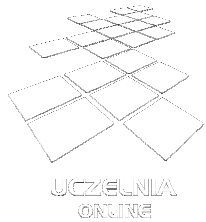


Zmienne kategoryczne

- **Zmienna kategoryczna** jest generalizacją zmiennej binarnej: może przyjmować więcej niż dwie wartości (np. dochód: wysoki, średni, niski)
- Niepodobieństwo (podobieństwo) pomiędzy obiektami i , j , opisanymi zmiennymi kategorycznymi, można zdefiniować następująco:

$$d(i, j) = \frac{p - m}{p} \quad d(i, j) = \frac{p - n}{p} = d(i, j) = \frac{m}{p}$$

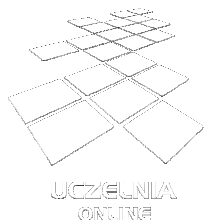
- gdzie p oznacza łączną liczbę zmiennych, m oznacza liczbę zmiennych, których wartość jest identyczna dla obu obiektów, n oznacza liczbę zmiennych, których wartość jest różna dla obu obiektów



Metody grupowania

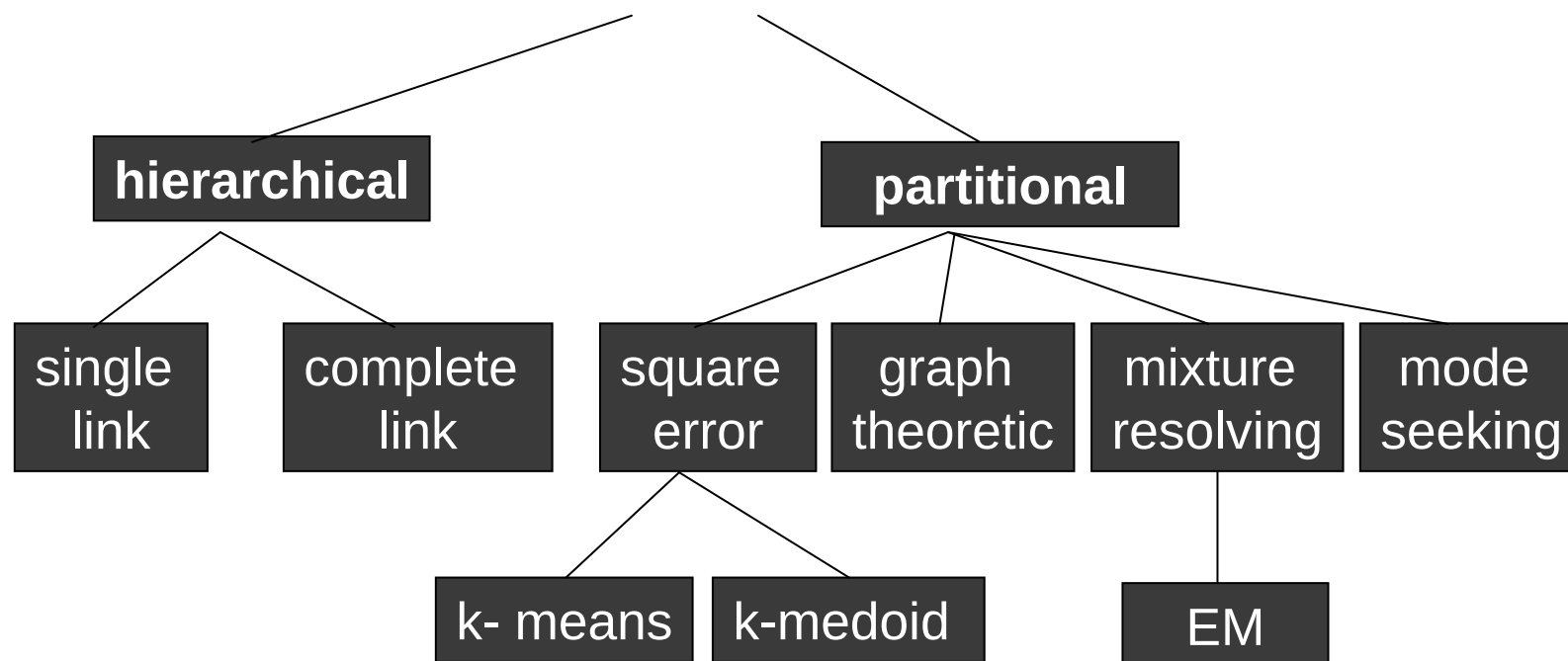
Typy metod

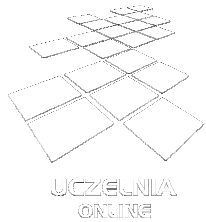
- Istnieje wiele różnych metod i algorytmów grupowania:
 - Dla danych liczbowych i/lub danych symbolicznych
 - Deterministyczne i probabilistyczne
 - Rozłączne i przecinające się
 - Hierarchiczne i płaskie
 - Monoteiczny i politeiczny
 - Przyrostowe i nieprzyrostowe



Metody grupowania (1)

Klasyfikacja metod





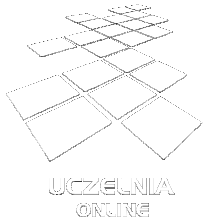
Metody grupowania (4)

- Dwa podstawowe podejścia do procesu grupowania obiektów:
- **Metody hierarchiczne**

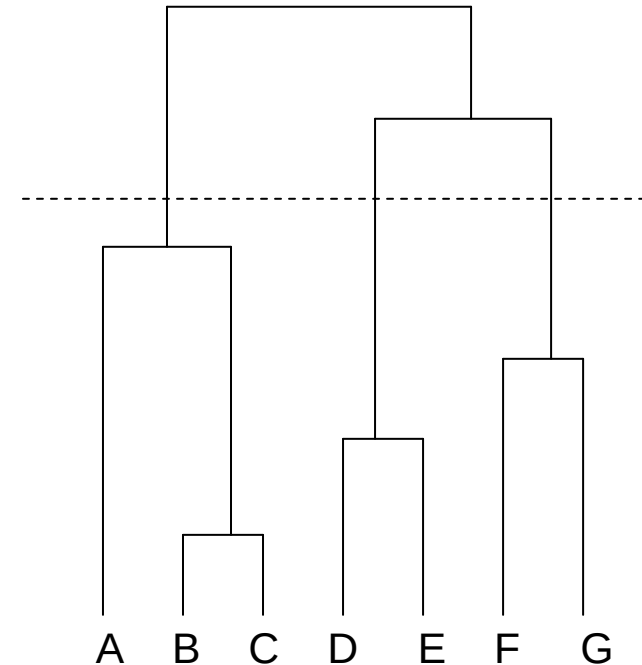
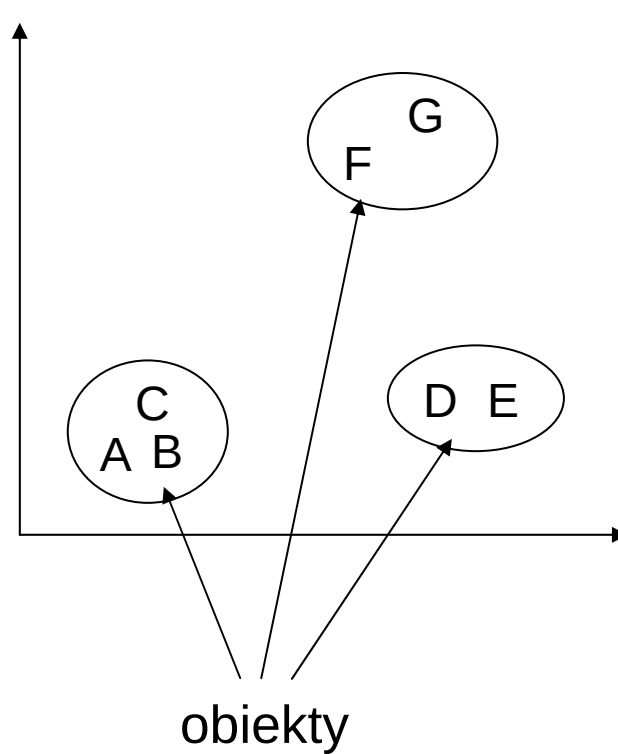
generują zagnieżdżoną sekwencję podziałów zbiorów obiektów w procesie grupowania

- **Metody z iteracyjno-optymalizacyjne**

generują tylko jeden podział (partycję) zbioru obiektów w dowolnym momencie procesu grupowania

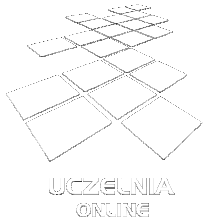


Metody grupowania hierarchicznego (1)



dendrogram

Metoda grupowania hierarchicznego polega na sekwencyjnym grupowaniu obiektów – drzewo klastrow (tzw. dendrogram)



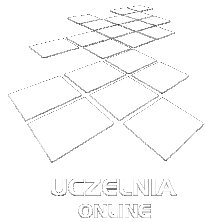
Metody grupowania hierarchicznego (2)

- **Podejście podziałowe**

(top-down): początkowo, wszystkie obiekty przypisujemy do jednego klastra; następnie, w kolejnych iteracjach, klaster jest dzielony na mniejsze klastry, które, z kolei, dzielone są na kolejne mniejsze klastry

- **Podejście aglomeracyjne**

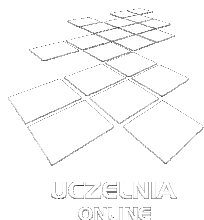
(bottom-up): początkowo, każdy obiekt stanowi osobny klaster, następnie, w kolejnych iteracjach, klastry są łączone w większe klastry



Miary odległości (1)

- W obu podejściach, aglomeracyjnym i podziałowym, liczba klastrow jest ustalona z góry przez użytkownika i stanowi warunek stopu procesu grupowania

4 podstawowe (najczęściej stosowane) miary odległości pomiędzy klastrami są zdefiniowane następująco, gdzie $|p - p'|$ oznacza odległość pomiędzy dwoma obiektami (lub punktami) p i p' , m_i oznacza średnią wartość klastra C_i , i n_i oznacza liczbę obiektów należących do klastra C_i



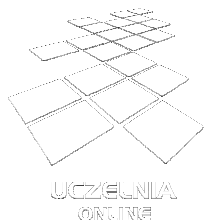
Miary odległości (2)

Minimalna odległość:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|$$

Maksymalna odległość:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \|p - p'\|$$



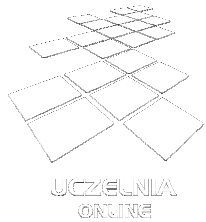
Miary odległości (3)

Odległość średnich:

$$d_{mean}(C_i, C_j) = \|m_i - m_j\|$$

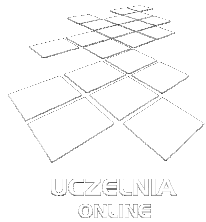
Średnia odległość:

$$d_{ave}(C_i, C_j) = 1 / (n_i n_j) \sum_{p \in C_i} \sum_{p' \in C_j} \|p - p'\|$$



Ogólny hierarchiczny aglomeracyjny algorytm grupowania

- Dane wejściowe: baza danych D obiektów (n - obiektów)
 - Dane wyjściowe: dendrogram reprezentujący grupowanie obiektów
- 1) umieść każdy obiekt w osobnym klastrze;
 - 2) skonstruuj macierz odległości pomiędzy klastrami;
 - 3) Dla każdej wartości niepodobieństwa dk
(dk może się zmieniać w kolejnych iteracjach)
powtarzaj
 - Utwórz graf klastrow, w którym każda para klastrow, której wzajemna odległość jest mniejsza niż dk , jest połączona lukiem aż wszystkie klastry utworzą graf spójny



Hierarchiczny aglomeracyjny algorytm grupowania

- Umieść każdy obiekt w osobnym klastrze.
Skonstruuj macierz przyległości zawierającą odległości pomiędzy każdą parą klastrów
- Korzystając z macierzy przyległości znajdź najbliższą parę klastrów. Połącz znalezione klastry tworząc nowy klaster. Uaktualnij macierz przyległości po operacji połączenia
- Jeżeli wszystkie obiekty należą do jednego klastra, zakończ procedurę grupowania, w przeciwnym razie przejdź do kroku 2