Uniwersytet Przyrodniczy we Wrocławiu

Data Mining
Wykład 8

Analiza skupień (grupowanie)
Grupowanie hierarchiczne O-Cluster

Plan wykładu

- 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

Uniwersytet Przyrodniczy we Wrocławiu

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 tematyk

Uniwersytet Przyrodniczy we Wrocławiu

Przykłady (2)

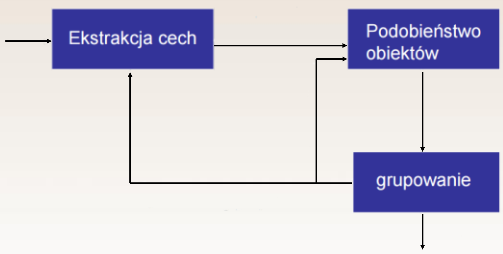
- **Zbiór sekwencji stron WWW:**

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

Uniwersytet Przyrodniczy we Wrocławiu

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. $D(x, y) \leq D(x, z) + D(z, y)$ (nierówność trójkąta)

Miary odległości (2)

- Dana jest k-wymiarowa przestrzeń euklidesowa, odległość pomiędzy dwoma punktami x=[x1, x2, ..., xk] oraz y=[y1, y2, ..., yk] można zdefiniować:

Odległość euklidesowa: $\sqrt{\sum_{i=1}^{k}(xi-yi)^2}$

Odległość Manhattan: $\sum_{i=1}^{k}|xi-yi|$

Miary odległości (3)

Odległość max z wymiarów: $max_{i=1}^{k}|xi - yi|$

Odległość Minkowskiego: $(\sum_{i=1}^{k}(|xi-yi|)^q)^{1/q}$

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 kategorycznych, dokumenty tekstowe, XML, grafy, itp..

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 – 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
...	...	...	...	...	...	...	...

$$d_{asym}(jack, mary) = \frac{2}{4} = 0.5 \qquad d_{sym}(jack, mary) = \frac{2}{7} = 0.29$$

$$d_{asym}(jack, jim) = \frac{2}{4} = 0.5 \qquad d_{sym}(jack, jim) = \frac{2}{7} = 0.29$$

$$d_{asym}(jim, mary) = \frac{4}{5} = 0.8 \qquad d_{sym}(jim, mary) = \frac{4}{7} = 0.57$$

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:

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

- 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

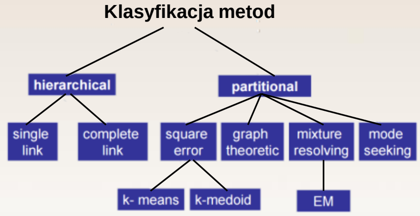
Metody grupowania - Typy metod

- 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

Uniwersytet Przyrodniczy we Wrocławiu

Metody grupowania (1)

Klasyfikacja metod



Uniwersytet Przyrodniczy we Wrocławiu

Metody grupowania (4)

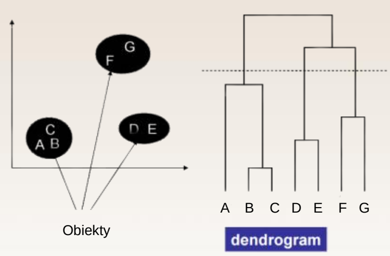
- Dwa podstawowe podejścia do procesu grupowania obiektów:
- **Metody hierarchiczne**

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)



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

Metody grupowania hierarchicznego (2)

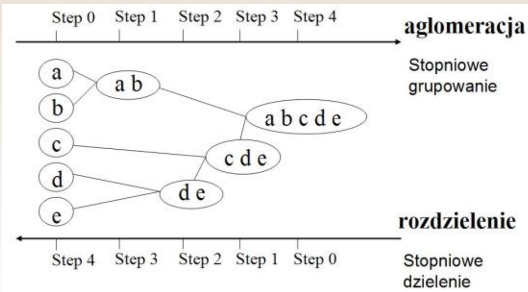
- **Podejście podziałowe**

(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

Metody grupowania hierarchicznego (2)



Miary odległości (1)

- W obu podejściach, aglomeracyjnym i podziałowym, liczba klastrów 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', mi oznacza średnią wartość klastra Ci,
- ni oznacza liczbę obiektów należących do klastra Ci

Miary odległości (2)

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

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

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)

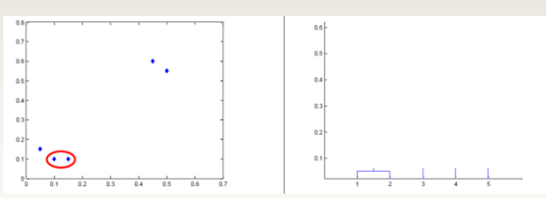
Dane wyjściowe: dendrogram reprezentujący grupowanie obiektów

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 klastrów, w którym każda para klastrów, której wzajemna odległość jest mniejsza niż dk, jest połączona lukiem aż wszystkie klastry utworzą graf spójny

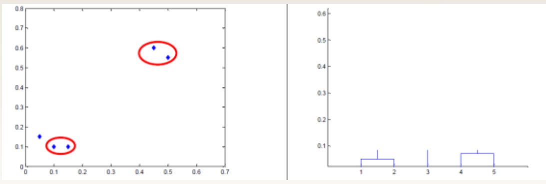
Hierarchiczny aglomeracyjny algorytm grupowania

- 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

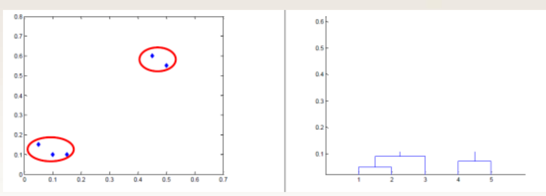
Przykład (1)



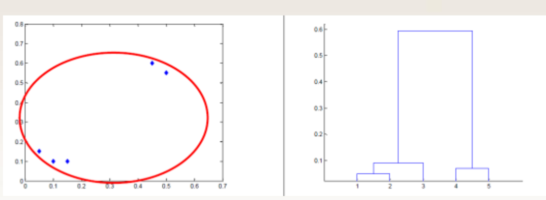
Przykład (2)



Przykład (3)



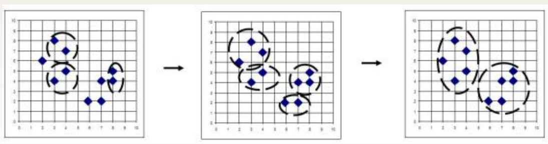
Przykład (4)



Aglomeracyjny algorytm grupowania danych

Algorithm 2: Aglomeracyjny algorytm grupowania danych

wejście: P - zbiór przykładów P,
$D(C_i, C_j)$ – funkcja do mierzenia odległości między dwoma skupieniami C_i i C_j.
wyjście: Dendogram skupień.
begin
  while *pozostało więcej niż jedno skupienie* **do**
    Niech C_i i C_j będą skupieniami minimalizującymi odległość $D(C_i, C_j)$ między dwoma skupieniami;
    $C_i = C_i \bigcup C_j$;
    usuń skupienie C_j;
end



Rozdzielający algorytm grupowania danych

Algorithm 3: Rozdzielający algorytm grupowania danych

wejście: P - zbiór przykładów P,
$D(C_i, C_j)$ – funkcja do mierzenia odległości między dwoma skupieniami C_i i C_j.
wyjście: Dendogram skupień.
begin
  while *istnieje skupienie składające z więcej niż dwóch elementów* **do**
    Niech C_i będzie skupieniem, którego po dzieleniu tworzą się dwa nowe skupienia C_j i C_k o maksymalnej odległości $D(C_j, C_k)$
    Podziel C_i na dwa skupienia C_j i C_k;
    Usuń skupienie C_i;
end