

# **Analiza skupień**

Marzena Nowakowska

Wydział Zarządzania i Modelowania Komputerowego  
Politechnika Świętokrzyska

# Zadania analizy skupień

Analiza skupień (*cluster analysis*) to narzędzie analizy danych, za pomocą którego w m-wymiarowym zbiorze danych wyodrębnia się niepuste, zazwyczaj rozłączne, podzbiory (grupy, klastry skupienia) podstawie niepodobieństwa/podobieństwa ich elementów.

Wyodrębnione skupienia są jak najbardziej różne od siebie i jednocześnie zawierają elementy jak najbardziej podobne sobie – wg. wybranej miary niepodobieństwa/podobieństwa.

Celem analizy skupień jest pomoc w ujawnieniu wzorców i struktur w zbiorze danych, które mogą dostarczyć informacji na temat leżących u ich podstaw relacji i powiązań. Analizę skupień zalicza się do metod uczenia nienadzorowanego.

## Przykłady zastosowań

- Marketing: segmentacja rynku w celu określania grup docelowych i tworzenia spersonalizowanych kampanii
- Medycyna: grupowanie pacjentów o podobnych objawach, potrzebach lub reakcjach na leczenie
- Socjologia: analiza sieci społecznościowych
- Planowanie miejskie: identyfikacja potrzeb różnych dzielnic miasta

# Cechy analizy skupień

Metoda heurystyczna



- Brak jednoznacznego, precyzyjnego modelu statystycznego.
- Wynik zależy od przyjętych ustaleń nt. miary niepodobieństwa obiektów, liczby skupień, algorytmu
- Brak testów statystycznych
- Charakter eksploracyjny
- Wrażliwość na zakres wartości zmiennych
- Wrażliwość na korelację zmiennych

# Miara niepodobieństwa

Miary niepodobieństwa obiektów określają, jak bardzo różnią się między sobą dwa obiekty. Im mniejsza miara niepodobieństwa tym bardziej obiekty są podobne

## Klasyfikacja miar niepodobieństwa

- Dla zmiennych ilościowych – metryka (czyli miara odległość)
- Dla zmiennych jakościowych – pseudometryka (nie spełnia wszystkich warunków trójkąta)

Miara  $d(x_i, x_j)$  jest metryką, jeżeli spełnia warunki:

- Tożsamość nierozróżnialnych  $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$
- Symetria  $d(x_i, x_j) = d(x_j, x_i)$
- Nierówność trójkąta  $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$

W ww. wynika nieujemność miary:  $d(x_i, x_j) \geq 0$

# Miary odległości

$d_{ij} = d(x_i, x_j)$ ,  $m$  - wymiar przestrzeni (liczba zmiennych)

|                                                                                                                                                  |                                                                  |
|--------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------|
| Metryka <b>euklidesowa</b>                                                                                                                       | $d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$               |
| Metryka <b>kwadratowa euklidesowa</b>                                                                                                            | $d_{ij}^2 = \sum_{k=1}^m (x_{ik} - x_{jk})^2$                    |
| Metryka <b>Manhattan</b> (miejska, taksówkowa)                                                                                                   | $d_{ij} = \sum_{k=1}^m  x_{ik} - x_{jk} $                        |
| Metryka <b>Czebyszewa</b> (maksimum)                                                                                                             | $d_{ij} = \max_k  x_{ik} - x_{jk} $                              |
| Metryka <b>Mahalanobisa</b>                                                                                                                      | $d_{ij} = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$               |
| Metryka <b>Minkowskiego</b><br>$p = 1 \rightarrow$ Manhattan<br>$p = 2 \rightarrow$ Euklidesowa<br>$p \rightarrow \infty \rightarrow$ Czebyszewa | $d_{ij} = \left( \sum_{k=1}^m  x_{ik} - x_{jk} ^p \right)^{1/p}$ |

# Miary niepodobieństwa dla cech jakościowych

Miara **Govera**

$$d_{ij} = 1 - \frac{\sum_{k=1}^m s_{ijk}}{m}$$

gdzie:

$$s_{ijk} = \begin{cases} 1, & \text{jeśli } x_{ik} = x_{jk} \\ 0, & \text{jeśli } x_{ik} \neq x_{jk} \end{cases}$$

Uogólnienie miar: **Hamming'a**, **Jaccarda**, **DICE'a (Sørensen)**, **Rogersa i Tanimoto**, **Russella i Rao**; wiele z nich jest dostosowanych do zmiennych binarnych.

Dla cech porządkowych często stosowane rozwiązanie: zamiana wartości na rangi liczbowe (np. 1, 2, 3), a następnie wyznaczenie niepodobieństwa stosując metrykę.

# Klasyfikacja algorytmów analizy skupień

