

# Analiza Wielowymiarowa

## Hierarchiczna analiza skupień

Maciej Nasiński, Paweł Strawiński

Zajęcia 9  
9 grudnia 2021

## 1 Hierarchiczna analiza skupień

- Wprowadzenie
- Metody
- Podsumowanie

# Definicja

- Hierarchiczna analiza skupień jest wariantem klasyfikacyjnej analizy danych
- Polega na utworzeniu klasyfikacji obiektów takiej, w której grupa wyższego stopnia jest sumą grup niższego stopnia
- Jeżeli zbiór liczy  $n$  obserwacji, to uzyskana hierarchia liczy  $n$  klasyfikacji składających się odpowiednio z  $1, 2, 3, \dots, n$  klas.
- Klasyfikacja zawierająca jedną klasę stanowi zbiór wszystkich obserwacji, natomiast złożona z  $n$  klas zawiera wyłącznie klasy jednoelementowe.

# Kiedy warto przeprowadzić analizę hierarchiczną

- Niech  $n$  będzie liczbą obserwacji w zbiorze
- Niech  $k$  będzie liczbą skupień
- Złożoność obliczeniowa metody  $k$ -średnich jest proporcjonalna do  $n \times k$
- Złożoność obliczeniowa analizy hierarchicznej jest proporcjonalna do  $n^k$  (aglomeracja) lub  $k^n$  (podział)

# Sposób analizy

- W odróżnieniu od metod niehierarchicznych nie jest wymagane określenie liczby grup, na które zbiór będzie dzielony
- Liczba grup jest rezultatem interpretacji wyników zastosowania algorytmu
- Utworzone grupy powinny być homogeniczne i odseparowane
- Typ analizy zależy od dwóch wyborów
  - miary odległości
  - metody tworzenia skupień

# Metody skupień

- Metody analizy hierarchcznej w zależności od punktu początkowego analizy można podzielić na:
  - aglomeracyjne - punktem początkowym jest  $n$  zbiorów jednoelementowych. W każdym kroku łączone są skupienia powstałe w poprzednich krokach
  - podziału - punktem początkowym jest jeden zbiór  $n$  elementowy. Wykorzystując miary (nie)podobieństwa (formalnie metryki lub semi-metryki) zbiór jest w kolejnych krokach dzielony

# Metody aglomeracyjne

- Załóżmy, że zbiór liczy  $n$  obserwacji
- Obliczana jest macierz odległości o wymiarze  $n \times n$
- Początkowo każda obserwacja tworzy osobną klasę
- Algorytm:
  - 1 Szukana jest para klas, między którymi odległość jest najmniejsza (podobieństwo jest największe).
  - 2 Klasy są łączone w jedną
  - 3 Modyfikowana jest macierz odległości
  - 4 Kroki (1)-(3) są powtarzane do uzyskania 1 klasy

# Dendrogram

- Graficzną ilustracją algorytmu jest dendrogram
- Jest to drzewo binarne, którego węzły reprezentują skupienia, a liście obiekty.
- Liście są na poziomie zerowym
- Węzły znajdują się na wysokości odpowiadającej wartości miary braku podobieństwa pomiędzy skupieniami reprezentowanymi przez oddzielne węzły



# Metody podziału

- Załóżmy, że zbiór liczy  $n$  obserwacji
- Obliczana jest macierz odległości o wymiarze  $n \times n$
- Początkowo wszystkie obserwacje tworzą jedną klasę
- Algorytm:
  - 1 Szukany jest taki podział jednej klasy na dwie klasy, aby odległość między nimi była największa.
  - 2 Klasy są dzielone
  - 3 Modyfikowana jest macierz odległości
  - 4 Kroki (1)-(3) są powtarzane do uzyskania  $n$  klas

# Algorytm DIANA

- DIANA jest metodą hierarchiczną która tworzy skupienia w odwrotnym porządku
- Jest to algorytm odwrotny do algorytmu aglomeracyjnego.
- Na początku jest jeden  $n$  elementowy zbiór
- W każdym kroku jest dzielony  $n$  dwa
- Zbudowanie hierarchii zajmuje  $n-1$  kroków.
- Ale w pierwszym kroku są rozpatrywane wszystkie możliwe dzielenia dwóch jednostek, liczba możliwych kombinacji wynosi  $2^{n-1} - 1$ .

# Koszt algorytmu DIANA

- Załóżmy, że dysponujemy zbiorem 100 obserwacji
- Złożoność obliczeniowa algorytmu aglomeracji wynosi  $\frac{n(n-1)}{2} \propto n^2$
- Złożoność obliczeniowa algorytmu podziału wynosi  $2^{n-1} - 1 \propto 2^n$
- Ze względu na nieakceptowalną złożoność obliczeniową algorytmów podziału w pakiecie Stata są dostępne wyłącznie algorytmy aglomeracji
- W innych pakietach np. R jest dostępny, lecz w wersji suboptymalnej. Optymalizowany jest jeden krok

# Mierzenie odległości między skupieniami (1)

- Metoda pojedynczego wiązania, nazywana również metodą najbliższego sąsiedztwa (ang. *single linkage clustering*).  
Odległość między dwoma skupieniami jest określona przez odległość między dwoma najbliższymi obiektami (najbliższymi sąsiadami) należącymi do różnych skupień
- Zastosowanie tej odległości prowadzi do tworzenia wydłużonych skupień, tzw. „łańcuchów”. Pozwalają one na wykrycie obserwacji odstających, nie należących do żadnej z grup
- Jest to powód, dla którego warto przeprowadzić początkową klasyfikację za jej pomocą, aby wyeliminować z analizy takie obserwacje

## Mierzenie odległości między skupieniami (2)

- Metoda pełnego wiązania, nazywana również metodą najdalszego sąsiedztwa (ang. *complete linkage clustering*). Odległość między skupieniami jest determinowana przez największą z odległości między dwoma dowolnymi obiektami należącymi do różnych skupień (tzn. „najdalszymi sąsiadami”)
- Metoda ta zwykle daje dobre rezultaty w tych przypadkach, w których obiekty faktycznie formują naturalnie oddzielone „kępki”
- Metoda prowadzi do tworzenia zwartych skupień
- Metoda ta nie jest odpowiednia, jeśli skupienia są „wydłużone” względem jednej lub kilku cech

## Mierzenie odległości między skupieniami (3)

- Metoda średniego wiązania (ang. *average linkage clustering*). Odległość między dwoma skupieniami jest to średnią odległość między wszystkimi parami obiektów należących do dwóch różnych skupień
- Metoda ta jest efektywna, gdy obiekty formują naturalnie oddzielone „kępki”, jednocześnie pozwala uzyskać dobry wynik w przypadku skupień wydłużonych, mających charakter „łańcucha”
- Metoda ta jest swoistym kompromisem pomiędzy metodami pojedynczego i pełnego wiązania. Ma ona jednak zasadniczą wadę. W odróżnieniu od dwóch poprzednich wykorzystywana miara niepodobieństwa nie jest niezmiennicza ze względu na monotoniczne przekształcenia miar niepodobieństwa pomiędzy obiektami

## Mierzenie odległości między skupieniami (4)

- Metoda Warda. Do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Minimalizowana jest suma kwadratów odchyleń dowolnych dwóch skupień, które mogą zostać uformowane na każdym etapie
- Metoda jest traktowana jako efektywna, tworzy skupienia o niewielkim zróżnicowaniu, ale również o małej liczbie obiektów
- Mimo wszystko, często nie jest w stanie zidentyfikować grup o dużym zakresie zmienności poszczególnych cech oraz grup o niewielkiej liczebności

# Własności metod hierarchicznych

- Nie istnieje jedna, najlepsza metoda
- Efektywność poszczególnych metod zależy od statystycznych własności zbioru danych
- Wyniki symulacji pokazują, że najlepsze wyniki daje metoda Warda, następnie metoda średnich połączeń oraz metoda najdalszego sąsiedztwa
- Metody hierarchiczne mają dobre własności w próbach o niewielkiej liczbie obserwacji. Ich efektywność maleje wraz ze wzrostem liczby obserwacji
- Nie zawierają mechanizmu korekty utworzonych skupień. Błędne przypisanie obiektu do skupienia nie może zostać skorygowane w kolejnym kroku analizy