

Analiza klasyfikacyjna

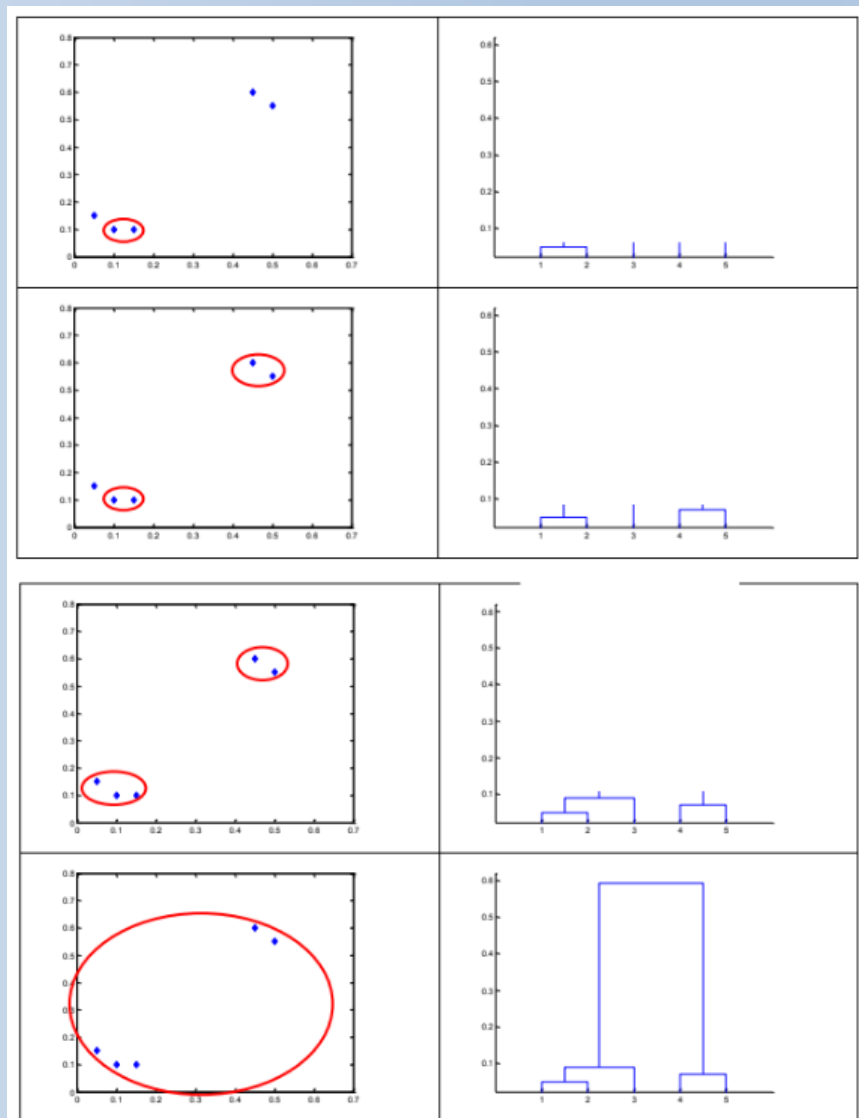
Analiza klasyfikacyjna

- Stosowanie analizy skupień ma na celu określenie czy wartości zmiennych tworzą jakąkolwiek strukturę w zbiorze danych, oraz czy grupy obserwacji są blisko siebie.
- głównym zadaniem klastrowania jest wyznaczenie naturalnego podziału, a następnie sprawdzenie, czy odpowiada ono narzuconym kryteriom.
- Najważniejszym celem analizy skupień jest możliwość przewidywania na podstawie stworzonego podziału

GRUPOWANIE Hierarchiczne

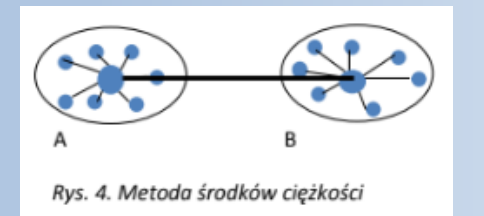
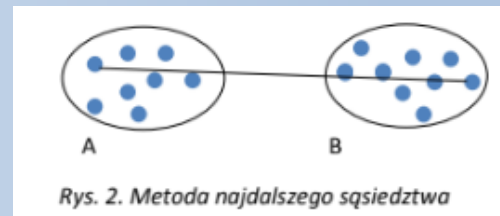
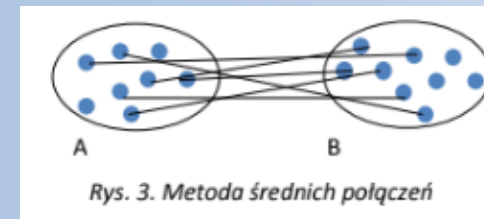
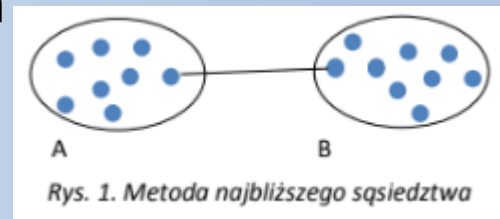
- Hierarchiczne
 - Aglomeracyjne –
wstępujące/łączące
 - Deglomeracyjne –
zstępujące/dzielące

Wynik graficzny : dendrogram



GRUPOWANIE hierarchiczne

- Definicje:
 - Centroid (klastroid) : środek grupy
 - Metryki pomiaru odległości : miary odległości wykorzystywane przez algorytm do obliczenia odległości między centroidami albo obserwacjami
 - Metody łączenia grup :
 - Pojedynczego połączenia / najbliższego sąsiada
 - Maksymalnego połączenia / najdalszego sąsiada
 - Centroidalnego połączenia
 - Średniego połączenia / przeciętnego sąsiada
 - Kryterium Warda



GRUPOWANIE Hierarchiczne

- Model :
 - Wybór podejścia: aglomeracyjne / deglomeracyjne
 - Decyzja dotycząca miary odległości
 - Decyzja dotycząca metody łączenia obserwacji
- Wady modelu
 - Metoda jest wrażliwa na obserwacje odstające
 - Jest skuteczna na nie dużych zbiorach
- Zalety modelu
 - Intuicyjność oraz łatwość interpretacji

Analizy>Analizy klasyfikacyjne>hierarchiczna analiza skupień

GRUPOWANIE Niehierarchiczne

- Algorytm k – średnich

Schemat działania :

- Losowanie k pierwszych centroidów
- Przypisanie wszystkich obserwacji do centroidów (miara euklidesowa)
- Dla każdej grupy wyznaczony jest nowy centroid
- Iteracja aż do :
 - Sytuacji w której przydział do grup się nie zmienia
 - Wykonanie określonej ilości iteracji
 - Osiągnięcie zbieżności

GRUPOWANIE Niehierarchiczne

Wady :

- Należy z góry zadać liczbę klastrów
- Wpływ losowości doboru punktów startowych
- Wrażliwość na obserwacje odstające
- Równoliczność klastrów

Zalety :

- Można go stosować na dużych zbiorach

Analizy>Analizy klasyfikacyjne> grupowanie met. K-średnich

Dwustopniowa analiza skupień

Narzędzie eksploracyjne służące do zdefiniowania występujących w zbiorze danych naturalnych zgrupowań (klastrow)

Jest to miara odległości oparta na wiarygodności zakładająca, że zmienne w modelu skupień są niezależne (testowanie: korelacje parami dla zmiennych ilościowych, tabele krzyżowe dla zmiennych jakościowych).

Analizy>Analizy klasyfikacyjne> dwustopniowa analiza skupień

Dwustopniowa analiza skupień

Automatyczne wyodrębnianie skupień				
Liczba skupień	Kryterium bayesowskie Schwarza (BIC)	Zmiana BIC ^a	Iloraz zmiany BIC ^b	Iloraz miary odległości ^c
1	58,044			
2	57,014	-1,030	1,000	1,852
3	64,724	7,710	-7,482	2,146
4	77,915	13,191	-12,801	2,575
5	94,032	16,117	-15,640	1,718
6	110,925	16,894	-16,394	1,263
7	128,044	17,119	-16,612	1,766
8	145,534	17,490	-16,972	1,040
9	163,042	17,509	-16,990	1,041
10	180,569	17,527	-17,008	1,859
11	198,302	17,734	-17,209	1,174
12	216,071	17,769	-17,243	1,781
13	233,931	17,859	-17,331	1,622
14	251,834	17,903	-17,374	1,089
15	269,743	17,909	-17,379	1,434

a. Zmiany odnoszą się do poprzedniej liczby skupień podanej w tabeli.

b. Ilorazy zmian odnoszą się do zmiany przy dwóch skupieniach.

c. Ilorazy miar odległości są obliczane na podstawie ilorazu aktualnej i poprzedniej liczby skupień.

Wyznaczanie liczby klastrów:

- niskie BIC
- wysoki Iloraz zmiany BIC
- wysoki Iloraz zmiany odległości