Analiza Wielowymiarowa

Niehierarchiczna analiza skupień

Paweł Strawiński

Zajęcia 9
21 grudnia 2023

1. Analiza skupień
   - Klasyfikacja
   - Analiza skupień
2. Niehierarchiczna analiza skupień
   - Analiza rozproszenia
   - Optymalizacja
   - Liczba, podobieństwo, poprawność skupień

Definicja

Cluster analysis is the art of finding groups in data

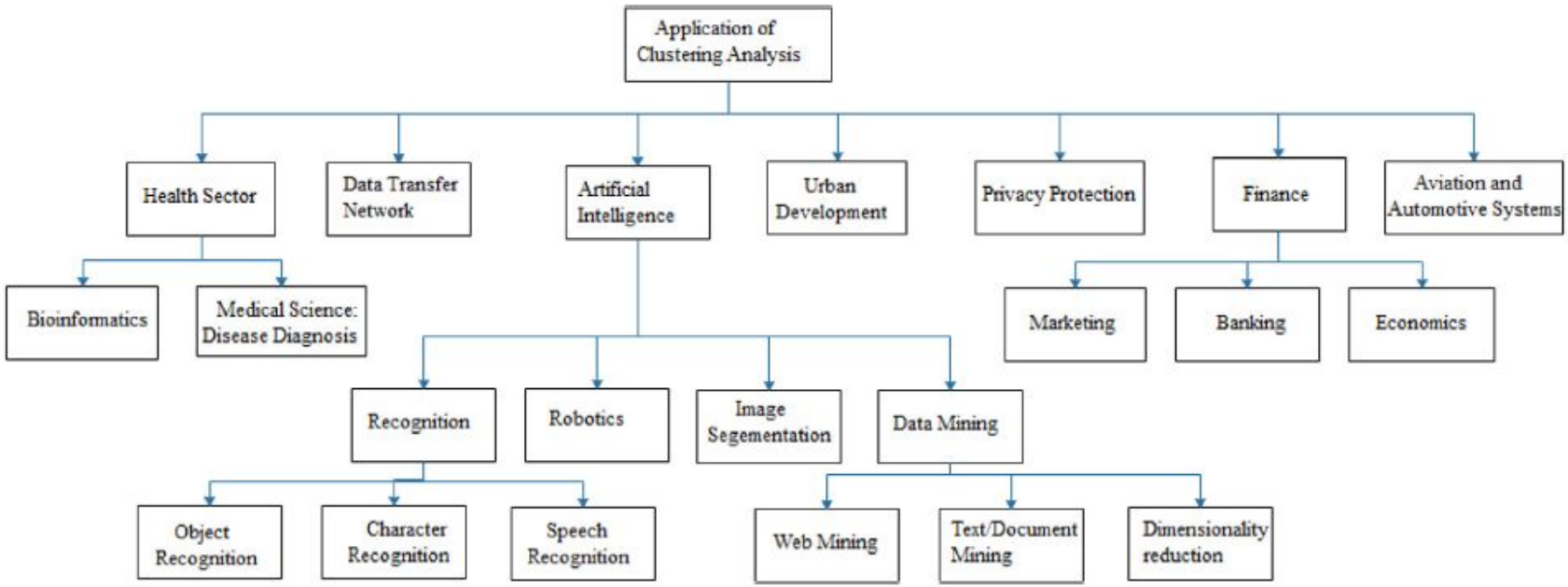
Kaufmann i Rousseeuw (1990)

- Jest to statystyczna metoda pozwalająca na znajdowanie grup podobnych obiektów w zbiorze danych
- Podstawą łączenia obiektów w grupy jest podobieństwo pomiędzy obiektami

Definicja

- Analiza skupień jest dziedziną eksploracji danych
- Polega na dzieleniu wielowymiarowego zbioru danych na grupy (podzbiory) w taki sposób, by elementy w tej samej grupie były do siebie podobne, a jednocześnie jak najbardziej odmienne od elementów z pozostałych grup (podzbiorów)
- Analiza skupień znalazła wiele zastosowań w różnych dziedzinach, jak np. klasyfikacja dokumentów (analiza internetu), odkrywanie grup klientów o podobnych zachowaniach (marketing), czy wykrywanie oszustw kredytowych (banki)

Zastosowania



Źródło: Ezugwu et al. (2022)

Cele analizy

- Uzyskanie grup jednorodnych obiektów, które ułatwiają wyodrębnienie ich cech
- Redukowanie dużej liczby danych pierwotnych do kilku podstawowych kategorii, które mogą być traktowane jako przedmioty dalszej analizy
- Ograniczenie czasu analizy, których przedmiotem będzie uzyskanie cech obiektów typowych
- Poznanie struktury analizowanych danych wielowymiarowych

Klasyfikacja

- Klasyfikacja jest grupowaniem podobnych obiektów
- Klasyfikacja jest ważnym elementem wielu dziedzin nauki, np. biologia, geologia, chemia, astronomia, itd.
- Klasyfikacja pozwala na zrozumienie zależności między obiektami
- Klasyfikacja może być traktowana jako wygodny sposób porządkowania dużych zbiorów danych
- Gdy dane można zaprezentować w formie małej liczby grup obiektów, wówczas etykiety obiektów w zwięzły sposób opisują wzory podobieństw i różnic między obiektami

Numeryczne metody klasyfikacji

- Wykorzystanie metod statystycznych do problemu klasyfikacji wywodzi się z nauk o ziemi
- Celem jest dostarczenie obiektywnej i stabilnej klasyfikacji obiektów
- Zasadniczym celem analizy skupień jest odkrywanie grup w danych

Analiza skupień

- Analiza skupień jest ogólną nazwą eksploracyjnej analizy danych, której celem jest odnalezienie, bądź wyodrębnienie grup lub skupień (ang. *clusters*) w danych
- Głównym celem analizy jest wykrycie w zbiorze danych, tzw. „naturalnych" skupień, czyli skupień, które można w sensowny sposób interpretować
- Analiza skupień jest metodą stworzoną raczej do formułowania hipotez na podstawie danych niż ich statystycznej weryfikacji
- Służy również opisowi wyodrębnionych podzbiorów

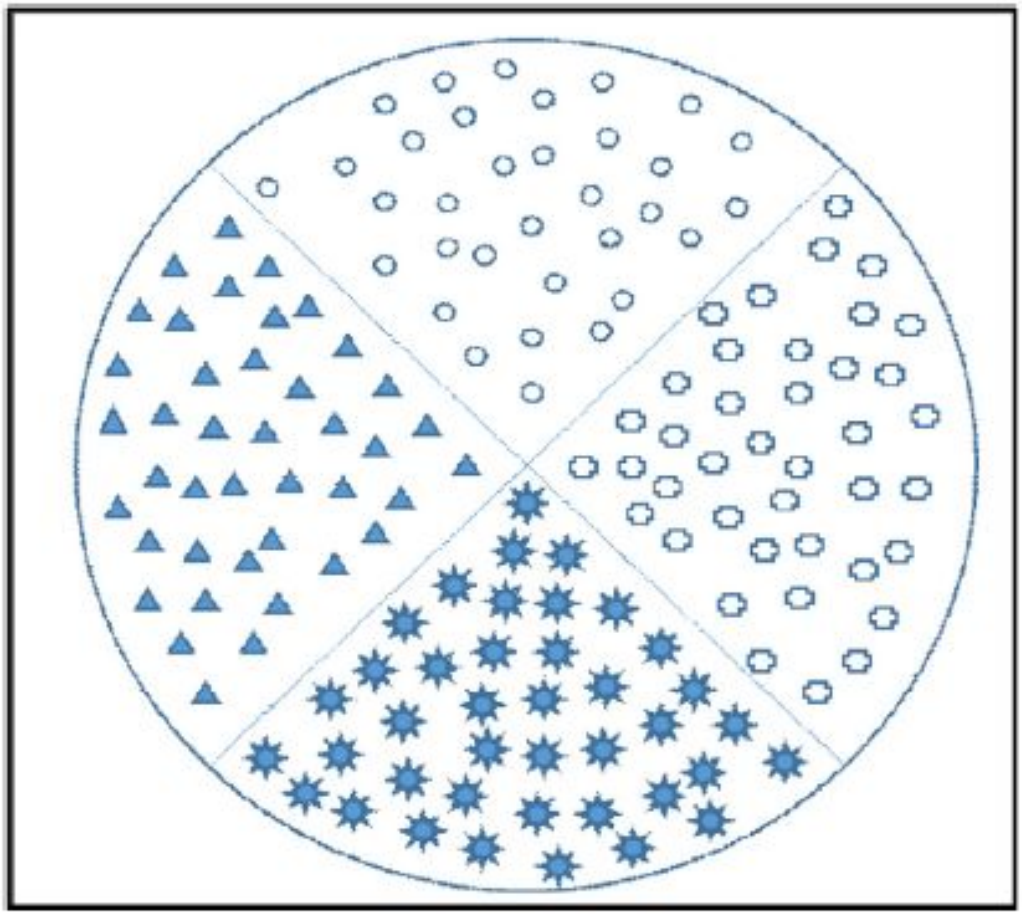
Algorytmy analizy skupień

- Algorytmy analizy skupień dzielą się na:
  - Algorytmy hierarchiczne zakładające, że grupy danych są zagnieżdżone
    - aglomeracyjne
    - podziału
  - Algorytmy podziału (niehierarchiczne) traktujące grupy danych w sposób równoważny
    - rozdzielające
    - mieszane (*ang. mixed*)
    - zamazane (*ang. fuzzy*)

Wprowadzenie

- Algorytm jednoznacznego podziału polega na podziale zbioru danych na określoną *a-priori* liczbę grup (podzbiorów) określanych mianem skupień
- Podział dokonywany jest w rezultacie optymalizacji funkcji kryterium podziału
- Utworzone grupy powinny być homogeniczne i odseparowane
- Homogeniczność oznacza, że grupy powinny być wewnętrznie spójne
- Separowanie oznacza, że grupy powinny być rozłączne i różnić się pod względem cech obiektów

Wynik algorytmu jednoznacznego podziału



Źródło: Ezugwu et al. (2022)

Omawiane metody

- Opracowano wiele metod/technik jednoznacznego (twardego) podziału danych między grupy
- Podczas zajęć omówione zostaną
  - metody medoidowe (k-średnich, k-median)
  - metody gęstości (DBSCAN)
  - metody rozkładu (Gaussian Mixture Models)

Analiza rozproszenia

Kryteria podziału na grupy oparte są o równość analizy wariancji

$$T = W + B$$

gdzie

- T oznacza całkowite rozproszenie,
- W oznacza rozproszenie wewnątrzgrupowe
- B oznacza rozproszenie międzygrupowe

Analiza rozproszenia

- Całkowite rozproszenie

$$\mathrm{T} = \sum_{m=1}^{g} \sum_{i=1}^{n_g} (x_{mi} - \bar{x})(x_{mi} - \bar{x})'$$

- Rozproszenie wewnątrzgrupowe

$$\mathrm{W} = \sum_{m=1}^{g} \sum_{i=1}^{n_g} (x_{mi} - \bar{x}_m)(x_{mi} - \bar{x}_m)'$$

- Rozproszenie międzygrupowe

$$\mathrm{B} = \sum_{m=1}^{g} n_m (x_m - \bar{x})(x_m - \bar{x})'$$

Kryteria podziału na grupy

- Minimalizacja śladu macierzy W lub maksymalizacja śladu macierzy B. Jest to odpowiednik minimalizacji wariancji wewnątrzgrupowej
- Minimalizacja wartości wyznacznika macierzy W
- Maksymalizacja wartości śladu macierzy BW^{-1}

Własności kryteriów podziału na grupy

- Wartość kryterium minimalizacji śladu macierzy W jest zależna od skali pomiarowej zmiennych oraz dodatkowo wymusza sferyczność struktury skupień
- Wartości dwóch pozostałych kryteriów są niezależne od skali pomiarowej

Wybór algorytmu optymalizacji

- Liczba możliwych podziałów zbioru n obiektów na g grup wynosi

$$N_{n,g} = \frac{1}{g!} \sum_{m=1}^{g} (-1)^{g-m} \left(\begin{array}{c} g \\ m \end{array} \right) m^n$$

- Na przykład
  - $N_{5,2} = 15$
  - $N_{10,3} = 9330$
- Wykonanie obliczeń dla wszystkich możliwych podziałów jest praktycznie niemożliwe przy dzisiejszej mocy obliczeniowej komputerów

Praktyczny algorytm optymalizacji

1. Znalezienie początkowego podziału n obiektów na g grup
2. Obliczenie zmiany wartości kryterium przy przesunięciu obiektu między grupami
3. Wybranie przesunięcia, które skutkuje największym przyrostem wartości funkcji kryterium
4. Powtarzanie kroków (2)-(3) do chwili, gdy przyrost wynosi zero.

Początkowe rozwiązanie

Sposoby ustalenia rozwiązania początkowego

1. Na podstawie wiedzy *a-priori*
2. W sposób losowy
3. Na podstawie statystycznej analizy danych
4. W praktyce najczęściej wykorzystywana jest metoda k-średnich albo k-median
5. Problem. Wybór początkowego rozwiązania może wpływać na rozwiązanie końcowe

Algorytm PAM

- Najczęściej wykorzystwaną w praktyce metodą wyboru rozwiązania początkowego jest algorytm PAM (ang. Partitioning Around Medoids)
- PAM jest metodą suboptymalną
- PAM wykorzystuje zachłanne wyszukiwanie, tzn. najlepiej rokujące w danym momencie wyboru rozwiązanie częściowe, które może nie prowadzić do optymalnego rozwiązania, ale jest szybsze niż pełne wyszukiwanie.

Algorytm PAM

1. Podziel zbiór danych na k skupień z wybranymi k medoidami
2. Oblicz macierz odległości pomiędzy medoidami oraz pozostałymi obserwacjami
3. Przypisz każdą z obserwacji (nie będącą medoidem) do najbardziej zbliżonego skupienia
4. Przy użyciu iteracji zastąp jeden z medoidów jednym z niemedoidów i sprawdź, czy odległości wszystkich elementów niebędących medoidami od najbliższych im medoidów są mniejsze
5. Jeśli nastąpiła przynajmniej jedna zamiana medoidów, wróć do kroku 3, w przeciwnym przypadku zakończ algorytm.

- Po ustaleniu medoidów przypisz każdą pozostałą obserwację do najbliższego jej medoidu

Wybór liczby skupień

- Większość sposobów ustalania liczby skupień ma charakter nieformalny
- Badania właściwości statystycznych kryteriów ustalających optymalną liczbę skupień wykorzystują techniki symulacyjne, przez co ich wyniki nie posiadają ogólnego charakteru
- Literatura wskazuje na dwa kryteria, które posiadają najlepsze właściwości statystyczne

Kryterium Calińskiego i Harabasza

- Caliński i Harabasz (1974) zaproponowali następującą statystykę

$$\mathtt{opt}g = \frac{\mathtt{tr(B)}}{g-1} / \frac{\mathtt{tr(W)}}{n-g} \sim F(g-1, n-g)$$

- Kryterium nie ma dobrych własności w przypadku analiz hierarchicznych

Kryterium Duda i Hart

- Kryterium weryfikuje czy skupienie powinno być podzielone na dwie części
- Niech J_1^2 będzie sumą kwadratów odległości wewnątrz skupienia
- Niech J_2^2 będzie sumą kwadratów odległości wewnątrz optymalnie podzielonego skupienia na dwie części
- Duda i Hart (1973) zaproponowali następującą statystykę

$$L(m) = \left(1 - \frac{J_2^2}{J_1^2} - \frac{2}{\pi p}\right)\left[\frac{n_m p}{2\left(1 - \frac{8}{\pi^2 p}\right)}\right] \sim N(0,1)$$

- Kryterium nie ma dobrych własności w przypadku analiz niehierarchicznych

Podobieństwo skupień

- Odległość i (nie)podobieństwo są podstawą działania algorytmów analizy skupień
- W przypadku danych ilościowych badacze wykorzystują miary odległości w celu ustalenia zależności między obserwacjami
- Miary (nie)podobieństwa są wykorzystywane w analizie danych o charakterze jakościowym

Podobieństwo skupień

- Najpopularniejsze sposoby mierzenia odległości między skupieniami
- L_1, czyli metryka miejska
- L_2, czyli metryka Euklidesowa
- L_∞, czyli największa odległość
- odległość Mahalanobisa
- odległość Minkowskiego
- odległość Jaccarda (1- podobieństwo Jaccarda)

Podobieństwo skupień

- Sposoby określania (nie)podobieństwa. Najpopularniejsze miary (nie)podobieństwa
- podobieństwo Jaccarda (1 - odległość Jaccarda)
- podobieństwo Hamminga
- dla danych mieszanych (ilościowych i jakościowych) - dowolna funkcja o wartościach w przedziale [0,1]
- Szczegółowe informacje dotyczące odległości i miar do znalezienia w Xu, Tian (2015) A Comprehensive Survey of Clustering Algorithms, Annals of Data Science 2(2):165–193

Poprawność skupień

- Statystyka Silhouette zaproponowana przez Rousseeuwa (1987) pokazuje jak dobrze obiekt pasuje do skupienia

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

gdzie a_i to średnia odległość do obiektów w tym samym skupieniu, b_i to średnia odległość do obiektów w kolejnym najbliższym skupieniu

- gdy skupienia są prawidłowo rozdzielone wartość statystyki będzie bliska 0
- ujemne wartości statystyki wskazują, że obiekt został nieprawidłowo przypisany

Poprawność skupień

- Kaufman i Rousseeuw (1990) sugerują wykorzystanie statystyki *szerokość silhouette*. Jest to średnia wartość statystyki Silhouete w zbiorze danych (skupieniu)
- Dane charakteryzujące się średnią wartością statystyki powyżej 0,5 uważają za dobrze rozdzielone
- Zbiory(skupienia), w których średnia wartość statystyki jest poniżej 0,2, określają jako pozbawione solidnej struktury skupień
- Więcej na ten temat: Halpin (2017) Cluster Analysis Utilities for Stata, prezentacja na spotkanie użytkowników Stata.