Analiza Wielowymiarowa

Niehierarchiczna analiza skupień

Dorota Celińska, Paweł Strawiński

Zajęcia 8
24 listopada 2015

1. Analiza skupień
   - Klasyfikacja
   - Analiza skupień

2. Niehierarchiczna analiza skupień
   - Analiza rozproszenia
   - Optymalizacja
   - Liczba 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 danych
- Podstawą grupowania jest podobieństwo pomiędzy obiektami

Definicja

- Analiza skupień jest dziedziną eksploracji danych
- Polega na dzieleniu wielowymiarowego zbioru danych na grupy w taki sposób, by elementy w tej samej grupie były do siebie podobne, a jednocześnie jak najbardziej odmienne od elementów z pozostałych grup
- 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)

Cele analizy

- Uzyskanie 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

Metody analizy skupień

- Analiza niehierarchiczna traktująca grupy danych w sposób równoważny
- Analiza hierarchiczna zakładająca, że grupy danych są zagnieżdżone

Wprowadzenie

- Polega na podziale zbioru danych na określoną *a-priori* liczbę grup (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

Analiza rozproszenia

- Kryteria podziału na grupy oparte są o równość analizy wariancji

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

gdzie

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

Analiza rozproszenia

- Całkowite rozproszenie

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

- Rozproszenie wewnątrzgrupowe

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

- Rozproszenie międzygrupowe

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

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 $\mathbf{BW}^{-1}$

Własności kryteriów podziału na grupy

- Wartość kryterium minimalizacji śladu macierzy $\mathbf{W}$ 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

$$N_{n,g} = \frac{1}{g!} \sum_{m=1}^{g} (-1)^{g-m} \binom{g}{m} m^n$$

- 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

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

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

$$\mathtt{opt}g = \frac{\mathtt{tr}(\mathbf{B})}{g-1} \Big/ \frac{\mathtt{tr}(\mathbf{W})}{n-g}$$

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

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

Sposób mierzenia odległości

- L_1, czyli metryka miejska
- L_2, czyli metryka Euklidesowa
- L_∞, czyli największa odległość
- odległość Mahalanobisa