Przestrzeń algorytmów klastrowania

Michał Woźniak

20 listopada 2008

Plan prezentacji

1. Wprowadzenie
   - Podstawowe pojęcia
   - Przykłady algorytmów klastrowania
2. Formalizacja podejścia
   - Odległość algorytmów klastrowania
   - Odległość podziałów
3. Rezultaty badań
   - Dane wejściowe
   - Eksperymenty

Praca źródłowa

Landscape of Clustering Algorithms

Anil K. Jain, Alexander Topchy, Martin H.C. Law,
and Joachim M. Buhmann *
Michigan State University, East Lansing, USA
** Institute of Computional Science, ETH Zentrum, Zurich, Switzerland*

Co to jest klasteryzacja?

Reprezentacja podziału na klastry

Podział na klastry możemy reprezentować poprzez funkcję $P_X : X \to \mathbb{N}$, która przypisuje każdej obserwacji ze zbioru X liczbę naturalną oznaczającą numer klastru. Dla danego X przez $\mathbb{P}_X$ oznaczamy przestrzeń podziałów zbioru X.

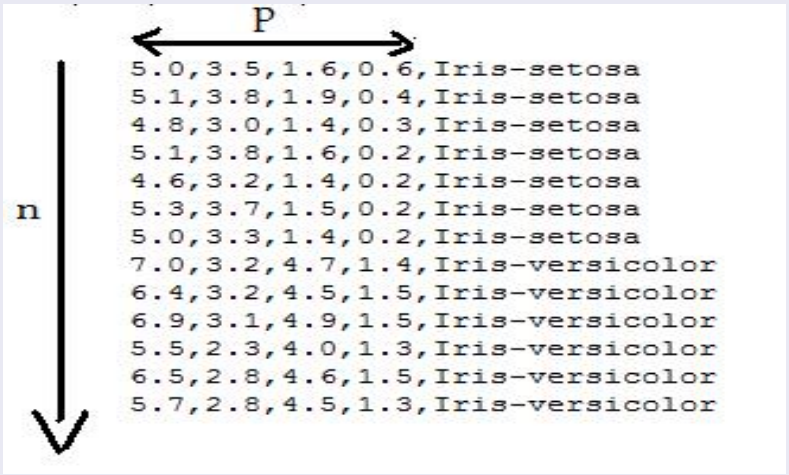
Zadanie

Z algorytmem klasteryzacji możemy utozsamiać funkcję celu $F_X : \mathbb{P}_X \to R$, która dla danego zbioru obserwacji X, ocenia podział. Algorytmy klasteryzacji znajdują maksimum funkcji F_X.

Reprezentacja danych

Przykłady według cech

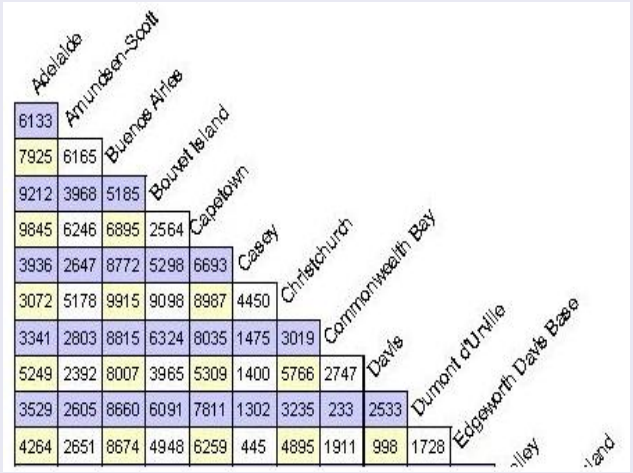
Dane o obserwacjach są w macierzy X o wymiarach n na p, gdzie n to liczba przykładów, a p to liczba cech (atrybutów).



Reprezentacja danych II

Macierz odległości

Macierz o wymiarach n na n, gdzie n to liczność zbioru danych.
Posiadamy jedynie dane o odległościach (mniej informacji).



Algorytmy klastrowania hierarchicznego

Ogólny schemat działania algorytmów hierarchicznych:

```
Inicjalizacja: każdy obiekt umieść w oddzielnym klastrze

for(i in 1..(n-1))
      Połącz 2 najbliższe klastry
      Odnów macierz odległości pomiędzy klastrami
```

Podstawowe metody liczenia odległości pomiędzy klastrami:

- single-link: $D_{k.ij} = min(D_{k.i}, D_{k.j})$
- complete-link: $D_{k.ij} = max(D_{k.i}, D_{k.j})$
- average-link: $D_{k.ij} = \frac{n_i}{n_i+n_j} D_{k.i} + \frac{n_j}{n_i+n_j} D_{k.j}$

Algorytm k-średnich

Inicjalizacja: C-podział startowy

```
repeat
      for(k in 1..K)
```

$$m_k = \underset{m}{\operatorname{argmin}} \sum_{i \in C^k} \|X_{i\cdot} - m\|^2 = \frac{\sum_{i \in C^k} X_{i\cdot}}{n}$$

```
      for(i in 1..N)
```

$$C(i) = \underset{k}{\operatorname{argmin}} \|X_{i\cdot} - m_k\|^2$$

```
until OK
```

Uwaga!

"Środek" nie musi należeć do zbioru obserwacji.

Jak klastrować algorytmy klastrowania?

- Porównywanie optymalizowanej funkcji celu
- Porównywanie wyników algorytmów

Odległość algorytmów klastrowania

Założenie

Odległość $D(F_1, F_2)$ pomiędzy algorytmami klastrowania na zbiorze danych X można przybliżać jako odległość $d(\cdot, \cdot)$ pomiędzy podziałami P_1^X oraz P_2^X zbioru X na klastry.

Zgodnie z założeniem definiujemy:

$$D_X(F_1, F_2) = d(P_1^X, P_2^X)$$

gdzie:

$$P_i^X = \underset{P^X \in \mathbb{P}^X}{\operatorname{argmax}} F_i(P^X)$$

Odległość podziałów

Odległość dwóch podziałów zbioru X będziemy liczyć ze wzoru (jest to tak zwany *Rand index*):

$$d(P_1^X, P_2^X) = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

gdzie:

- a- liczba par elementów X, które należą do tego samego klastra dla obu podziałów
- b- liczba par elementów X, które należą do różnych klastrów w podziale P_1^X oraz P_2^X
- c- liczba par elementów X, które należą do tego samego klastra w P_1^X, ale do różnych w P_2^X
- d- liczba par elementów X, które należą do różnych klastrów w P_1^X, ale do tego samego klastra w P_2^X
- n- liczba elementów X

Zbiory danych

Użyto rzeczywistych zbiorów danych pobranych z bazy *UCI Machine Learning Repository*.
Atrybuty z NULL'owymi wartościami zostały pominięte.

Dermatology	Galaxy	Glass
Heart	Ionosphere	Iris
Letter recognition (A,B,C)	Segmentation	Texture
Letter recognition (X,Y,Z)	Wdbc	Wine

Tabela: Wykorzystane zestawy danych

Algorytmy

Zbadano 35 algorytmów klastrowania, numerowanych liczbami od 1..35. Niektóre z nich:

- Algorytm k-średnich (29)
- Algorytmy klastrowania hierarchicznego z użytymi metodami: SL(30), AL(5), CL(13) oraz Ward(35).
- Dwie wersje klastrowania spektralnego z dwoma różnymi parametrami odpowiedzialnymi za współczynniki skalowania

Algorytmy II

- Algorytmy oparte na 7 funkcjach celu zaimplementowane w programie CLUTO:
  - $\max I_1 = \sum_{i=1}^{k} \frac{S_i}{n_i}$ (27)
  - $\max I_2 = \sum_{i=1}^{k} \sqrt{S_i}$ (28)
  - $\min E_1 = \sum_{i=1}^{k} n_i \frac{R_i}{\sqrt{S_i}}$ (18)
  - $\min G_1 = \sum_{i=1}^{k} \frac{R_i}{\sqrt{S_i}}$ (19)
  - $\min E_2 = \sum_{i=1}^{k} n_i^2 \frac{R_i}{\sqrt{S_i}}$ (20)
  - $\max H_1 = \frac{I_1}{E_1}$ (25)
  - $\max H_2 = \frac{I_2}{E_1}$ (26)

  gdzie n_i to rozmiar klastra C_i oraz

$$S_i = \sum_{x,y \in C_i} sim(x,y), \qquad R_i = \sum_{j} \sum_{x \in C_i, y \in C_j} sum(x,y)$$

Algorytmy III

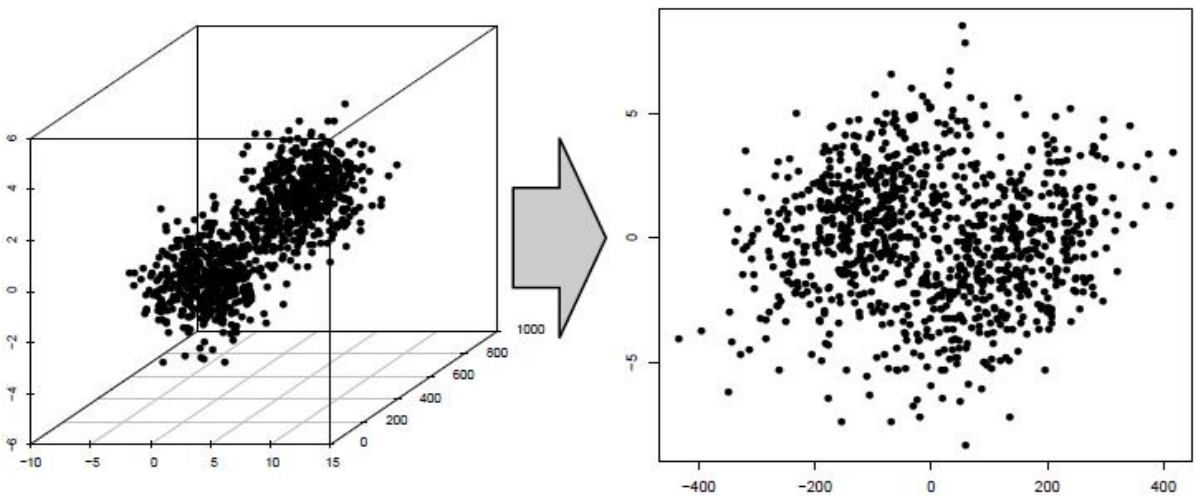
- Rodzina algorytmów realizujących ideę "algorytmu kameleonu", wykorzystujące 7 algorytmów opisanych na poprzednim slajdzie.
  Algorytm kameleon posiada 2 fazy: *podziału* i *łączenia*. Do fazy podziału wykorzystano algorytm *k-średnich*, do wygenerowania dużej liczby małych klastrów. W zależności od algorytmu użytego w fazie łączenia otrzymano 7 algorytmów (6-12).

Algorytmy IV

- 4 algorytmy grafowe, dokonujące podziału poprzez procedurę odcięcia w grafie najbliższych sąsiadów. Algorytmy te używają 4 definicji odległości, które kształtują strukturę sąsiedztwa w grafie:
  - współczynnik korelacji
  - funkcja cosinus
  - odległość euklidesową
  - odległość Jaccarda ($d_{AB} = \frac{|A \cap B|}{|A \cup B|}$)

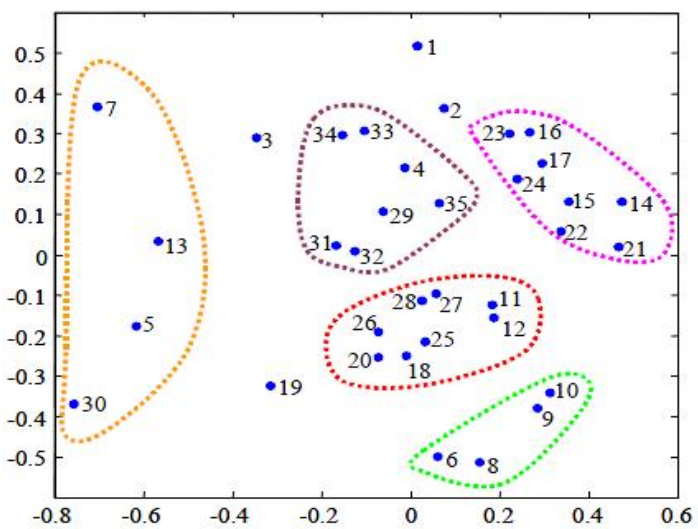
Opis eksperymentu 1

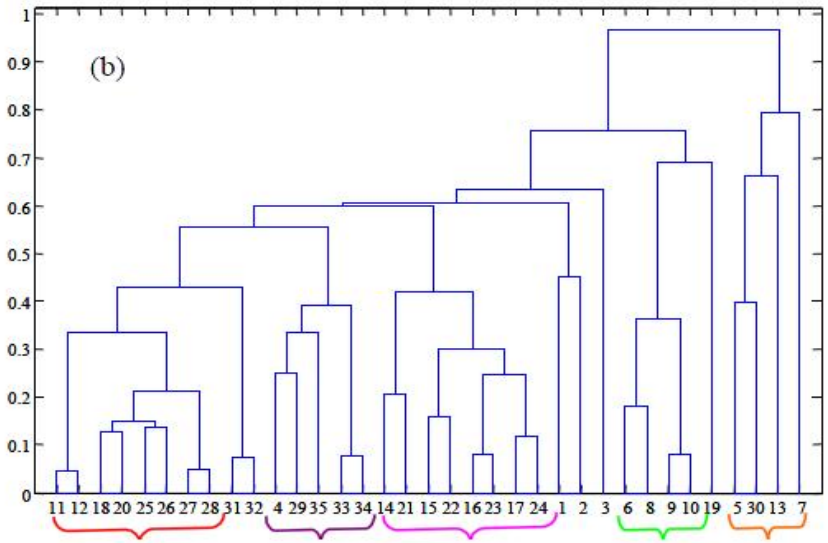
1. Wyznaczenie macierzy odległości 35x35 pomiędzy algorytmami uśrednionej z 12 macierzy dla różnych danych
2. Skalowanie Sammona (stress value=0.0587)
3. Dendogram algorytmów metodą complete-link



Rysunek: Przykład skalowania wielowymiarowego

Wyniki eksperymentu 1





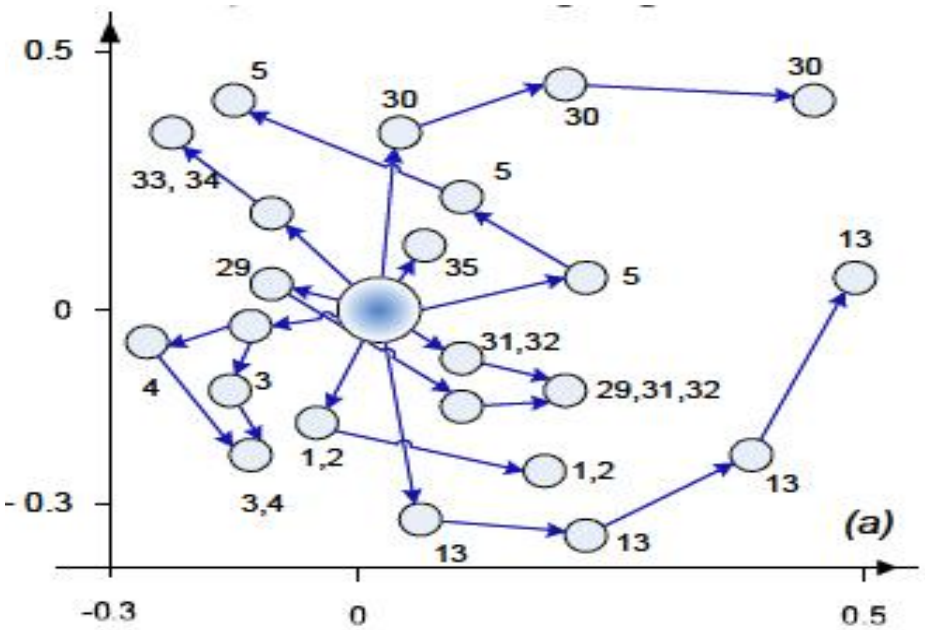
Obserwacje:

- algorytm K-średnich(29) w centrum przestrzeni
- algorytmy typu kameleon blisko siebie (6-12)
- algorytmy spektralne blisko siebie (31-34)

Opis eksperymentu 2

1. Przygotowano 12 sztucznych zestawów danych zawierających po 3 klastry wygenerowane z 2-wymiarowych rozkładów normalnych
2. Kolejne zbiory danych różnią się poziomem separowalności klastrów
3. W wyniku zmiejszania się separowalności klastrów obserwowano przemieszczanie się algorytmów w przestrzeni

Wyniki eksperymentu 2



Rysunek: Przestrzeń algorytmów klastrowania, ścieżki odpowiadają zmianom położenia algorytmów w wyniku zmiejszania odległości pomiedzy trzema klastrami