Algorytmy grupowania - cd, wykł. 11

Joanna Jędrzejowicz

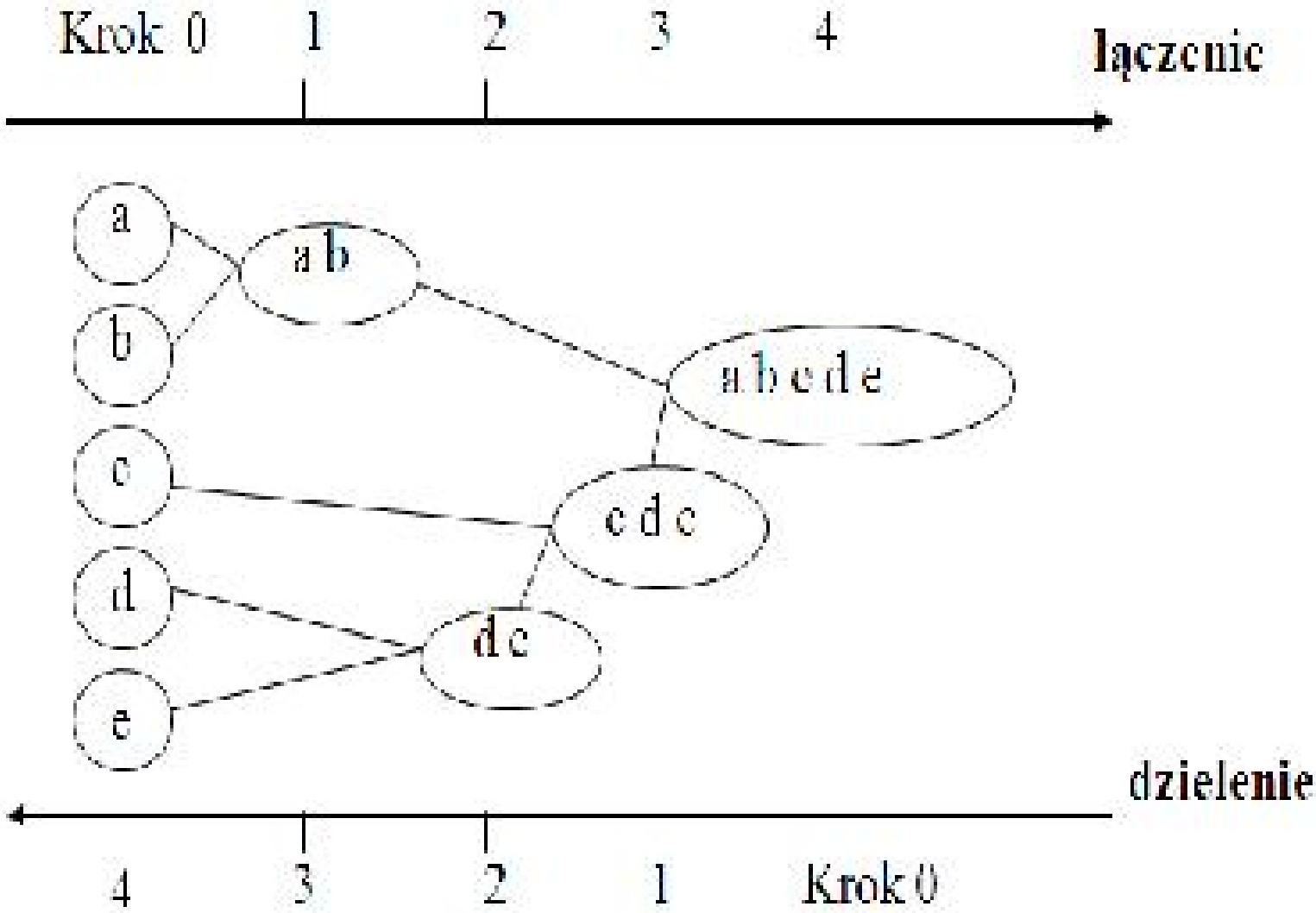
Instytut Informatyki

Grupowanie hierarchiczne

- W grupowaniu hierarchicznym tworzona jest struktura drzewiasta poprzez rekurencyjne dzielenie (metody rozdzielające) lub łączenie (metody aglomeracyjne).
- W metodach aglomeracyjnych początkowo zakłada się, że każdy rekord jest grupą składającą się z pojedyńczego elementu. W kolejnych krokach, dwie grupy które są najbliżej siebie, łaczy się w nową wspólna grupę.
- Metody rozdzielające zaczynają ze wszystkimi rekordami zawartymi w jednej grupie, z najbardziej niepodobnymi rekordami rozdzielanymi rekurencyjnie w oddzielne grupy.

Grupowanie hierarchiczne

Używa się macierzy odległości do tworzenia drzewa. Metoda nie wymaga podania liczby klastrów, tylko warunek końca!



Określanie odległości między grupami A i B

- metoda pojedynczego połączenia (ang. single linkage) - oparta na minimalnej odległości pomiędzy dowolnym rekordem z grupy A i dowolnym rekordem z grupy B,
- metoda całkowitego połączenia (ang. complete linkage) - oparta na maksymalnej odległości pomiędzy dowolnym rekordem z grupy A i dowolnym rekordem z grupy B, łączy się grupy dla ktorych ta odległoć jest najmniejsza
- metoda średniego połączenia (ang. average linkage) - stworzona, aby ograniczyć wpływ ekstremalnych wartości, takich jak najbliższe i najbardziej oddalone rekordy, na kryterium połączenia klastrów. Bierze się pod uwagę średnią odległośc wszystkich rekordów z klastra A do wszystkich rekordów z klastra B
- metoda odleglosci centroidów - w każdej iteracji łączone są klastry A i B, które charakteryzują się najmniejszą odległoscią centroidów.

Przykład

Rozpatrujemy jednowymiarowy zbiór danych:

$$2 \quad 5 \quad 9 \quad 15 \quad 16 \quad 18 \quad 25 \quad 33 \quad 33 \quad 45$$

Zaczynamy od klastrów jednoelementowych.

Metoda pojedynczego połączenia

1. krok: elementy $33, 33$ tworzą jeden klaster, pozostałe bez zmiany, czyli 9 klastrów
2. elementy $15, 16$ tworzą jeden klaster, razem 8 klastrów
3. zostają połączone klastry $\{15, 16\}$ z $\{18\}$ - zostaje 7 klastrów,
4. zostają połączone klastry $\{2\}$ z $\{5\}$ - zostaje 6 klastrów,
5. połączone $\{2, 5\}$ z $\{9\}$ - 5 klastrów,
6. połączone $\{2, 5, 9\}$ z $\{15, 16, 18\}$ - 4 klastry,
7. połączone $\{2, 5, 9, 15, 16, 18\}$ z $\{25\}$ - minimalna odległość jest 7 - zostają 3 klastry,
8. połączone $\{2, 5, 9, 15, 16, 18, 25\}$ z $\{33, 33\}$ - pozostają 2 klastry,
9. krok: dwa klastry połaczone w jeden.

Metoda całkowitego połączenia

1. krok: elementy $33, 33$ tworzą jeden klaster, pozostałe bez zmiany, czyli 9 klastrów
2. elementy $15, 16$ tworzą jeden klaster, razem 8 klastrów
3. inaczej niż w metodzie poj. połaczenia - można połaczyć 2 i 5 lub $\{15, 16\}$ i 18, wybieramy to pierwsze - zostaje 7 klastrów,
4. zostają połączone klastry $\{15, 16\}$ z $\{18\}$ - zostaje 6 klastrów,
5. połączone $\{2, 5\}$ z $\{9\}$ - 5 klastrów,
6. połączone $\{25\}$ z $\{33, 33\}$ - 4 klastry,
7. połączone $\{2, 5, 9\}$ z $\{15, 16, 18\}$ - zostają 3 klastry,
8. połączone $\{25, 33, 33\}$ z $\{45\}$ - pozostają 2 klastry,
9. krok: dwa klastry: $\{2, 5, 9, 15, 16, 18\}$, $\{25, 33, 33, 45\}$ połaczone w jeden.

Metoda średniego połączenia

Kryterium - średnia odległość pomiędzy wszystkimi rekordami z jednej grupy z wszystkimi rekordami z drugiej grupy. W przykładzie otrzymuje się tą samą strukturę hierarchiczną co w całkowitym połączeniu. Ogólnie, metoda średniego połączenia prowadzi do grup bardziej podobnych kształtem do grup uzyskanych metodą całkowitego połączenia niż do grup otrzymanych metodą pojedynczego połączenia.

Zalety i wady grupowania hierarchicznego

- metody hierarchiczne umożliwiają stosowanie dowolnych miar odległosci lub podobienstwa obiektów, może być stosowana do grupowania dowolnych obiektóe: dokumentów tekstowych, dokumentów XML itd,
- umożliwia elastyczny wybór klastrów o dowolnej ziarnistosci,
- metody hierarchiczne są zwykle słabo skalowalne - zwykle złożonoć $O(n^2)$

Grupowanie oparte na gęstości

- Klastrem obiektów jest obszar w przestrzeni obiektów charakteryzujący się duzą gęstoscią obiektów, klastry są odseparowane od siebie obszarami o małej gęstosci obiektów,
- odkrywa grupy o dowolnym kształcie,
- odkrywa szumy,
- jedno przeglądanie danych

DBSCAN - metoda oparta na gęstości

- parametry:
  ϵ - promień definiujący sąsiedztwo punktu
  MinPts - minimalna liczba punktów w ϵ sąsiedztwie,

DBSCAN cd

Kilka definicji..

- ϵ sąsiedztwo punktu p - taki zbiór punktów, do których odległość z p jest nie większa od ustalonego promienia ϵ,
- punkty tworzące jedna grupę mozna podzielić na dwie klasy: punktów rdzenia (ang. core points) i punktów brzegowych (ang. border points),
- rdzeń - punkt, który ma co najmniej MinPts w ϵ sąsiedztwie,
- punkt brzegowy - ma mniej niż MinPts w ϵ sąsiedztwie,
- punkt p jest bezpośrednio osiagalny gęstościowo z punktu q, gdy p znajduje sie w ϵ sąsiedztwie punktu q i q jest punktem rdzeniowym,

DBSCAN cd

- punkt p jest osiągalny gęstościowo z punktu q, gdy istnieje taki ciąg punktów $p_1,\ldots,p_n$, gdzie $p_1 = p$, $p_n = q$ że p_{i+1} jest bezpośrednio osiągalny gęstościowo z p_i,
- punkt p jest połączony gęstościowo z punktem q, wtedy i tylko wtedy gdy istnieje taki punkt r, że zarówno p jak i q są osiągalne gęstościowo z r,
- grupa C jest niepustym zbiorem punktów, takich że dla każdej pary punktów p i q zachodzą dwa warunki: oba punkty sa połączone gęstościowo i jeżeli $p \in C$ i p jest osiągalny gęstościowo z punktu q, to $q \in C$.

Algorytm DBSCAN

Dane wejściowe: zbiór punktów, ϵ, MinPts
Wynik: grupy
1 **while** istnieje jakiś punkt nierozpatrzony **do**
2 wybierz dowolny nierozpatrzony punkt p;
3 Q $\leftarrow$ wszystkie punkty osiagalne gęstosciowo z p;
4 **if** p jest rdzeniem **then**
5 zwróć Q jako nową grupe;
6 **od**

Algorytm DBSCAN

Algorytm DBSCAN jest wydajny dla dużych baz danych. Jego czas działania to $O(n \cdot \log n)$. Potrafi wykryć grupy o dowolnych kształtach w przeciwieństwie do metod klasycznych, które wykrywają tylko grupy wypukłe.

Algorytm PAM (partitioning around medoids)

Algorytm PAM jest wersją algorytmu k-medoidów

- Idea: algorytm w kolejnych iteracjach wykonuje dwie fazy. W fazie BUILD wybiera k medoidów (srodków klastrów), w fazie SWAP próbuje poprawic wybór medoidów, analizujac wszystkie możliwe pary obiektów, takich, że jeden z obiektów jest medoidem, natomiast drugi z obiektów nie jest medoidem
- Jakość grupowania, dla każdej kombinacji par, jest szacowana i wybierany jest najlepszy zbiór medoidów. Otrzymany zbiór medoidów stanowi punkt wyjścia do obliczeń w kolejnej iteracji,

Algorytm PAM - definicje

Niech O oznacza zbiór wszystkich elementów. Będą tworzone dwa podzbiory: S (selected) - zbiór elementów wyróżnionych, U (unselected) - pozostałe.
Dla dowolnego elementu p:
D_p - odległość p od najbliższego elementu z S,
E_p - odległość p od drugiego, najbliższego elementu z S,
obie wartości są modyfikowane po każdej zmianie U i S
mamy $D_p \leq E_p$ oraz $p \in S \longleftrightarrow D_p = 0$

Faza BUILD

1. dodaj do S element, dla którego suma odległości od innych elementów jest minimalna,
2. rozpatrz możliwość umieszczenia elementu i ze zbioru U w zbiorze S,
   1. dla $j \in U - \{i\}$ oblicz D_j,
   2. jeżeli $D_j > d(i,j)$, to może świadczyć, że warto dołączyć i do S;
   3. zysk z umieszczenia i w S: $g_i = \sum_{j \in U} C_{ji}$, gdzie $C_{ji} = max\{D_j - d(j,i), 0\}$
3. wybierz i ktore maksymalizuje g_i, umiesc w S i usun z U

   powtarzaj powyższe kroki, aż do umieszczenia k elementów w S

Faza SWAP

SWAP służy do poprawy zbioru S: rozpatruje się wszystkie pary $(i, h) \in S \times U$, wylicza wartość T_{ih}, która ocenia rezultat zamiany roli obu elementów i oraz h

1. dla dowolnego $j \in U$, $j \neq h$ będziemy wyliczać K_{jih} ('pożytek' z zastąpienia i przez h)

Mamy

- $d(j, i) > D_j$, lub
- $d(j, i) = D_j$,

Faza SWAP, cd

- jeżeli $d(j,i) > D_j$, to
  - jesli $d(j,h) \geq D_j$, to $K_{jih} = 0$,
  - jesli $d(j,h) < D_j$, to $K_{jih} = d(j,h) - D_j$,

  w obu przypadkach $K_{jih} = min\{d(j,h) - D_j, 0\}$,
- jeżeli $d(j,i) = D_j$, to
  - jesli $d(j,h) < E_j$, to $K_{jih} = d(j,h) - D_j$,
  - jesli $d(j,h) \geq E_j$, to $K_{jih} = E_j - D_j$,

  $K_{jih} = min\{d(j,h), E_j\} - D_j$

ogolny rezulat zamiany roli i oraz h:

$$T_{ih} = \sum_{j \in U} K_{jih}$$

Faza SWAP, cd

2. wybierz parę $(i, h) \in S \times U$, która minimalizuje T_{ih};
3. jeśli $T_{ih} < 0$, to dokonaj zamiany, zmodyfikuj wszystkie D_p oraz E_p oraz wróć do fazy BUILD
4. w przeciwnym przypadku, koniec algorytmu i S zawiera medoidy

Przykład użycia PAM

Dany jest zbiór ciągów symboli:

$$s_1 = A\ B\ C\ D\ E\ F$$

$$s_2 = A\ B\ E\ F\ G$$

$$s_3 = C\ D\ E\ F\ P\ Q\ J$$

$$s_4 = X\ Y\ P\ R\ S\ T\ Z$$

$$s_5 = K\ L\ R\ S\ M$$

$$s_6 = K\ L\ M\ N\ T\ Z$$

z miarą odległosci

$$d(s_i, s_j) = |s_i| + |s_j| - 2|LCS(s_i, s_j)|$$

LCS najdłuższa wspólna sekwencja

Przykład użycia PAM, cd

Przykładowe odległosci:

$$D(s_1, s_3) = 6 + 7 - 2 \cdot 4 = 5$$

$$D(s_2, s_4) = 5 + 7 - 2 \cdot 2 = 8$$

itd
jesli przyjąć liczbę klastrów 2 oraz początkowe medoidy s_3 i s_4, to otrzymujemy dwa klastry:

$$C_1 = \{s_1, s_2, s_3\}\ C_2 = \{s_4, s_5, s_6\}$$

wyliczyć wszystkie odleglosci i sprawdzić podział
po zastosowaniu algorytmu PAM otrzymuje się medoidy s_1, s_5 oraz podział jak wyżej.

Algorytm PAM

- PAM lepiej niż k-średnich radzi sobie z danymi zaszumionymi i wyjątkami (outlier)
- PAM działa dobrze dla małych zbiorów danych, ale jest słabo skalowalny dla dużych zbiorów, wymaga $O(k \cdot (n-k)^2)$ kroków w każdej iteracji (n - liczba rekordów, k - liczba grup),
- dla dużych zbiorów stosuje się próbkowanie - metoda CLARA

CLARA (Clustering LARge Applications)

- zamiast rozważać cały dostępny zbiór obiektów do grupowania, wybieramy reprezentatywny podzbiór obiektów - metoda próbkowania,
- z wybranej próbki, stosujac algorytm PAM, wybieramy zbiór medoidów będących srodkami klastrów,
- jeżeli mechanizm próbkowania obiektów zapewnia losowość próbki, to próbka bedzie dobrze odzwierciedlała rozkład obiektów w oryginalnym zbiorze obiektów,
- wybrany zbiór medoidów bedzie, w miare wiernie, odpowiadał wyborowi medoidów przez algorytm PAM z całego zbioru obiektów,

Algorytm CLARA cechuje się lepszą efektywnoscią niż klasyczny algorytm k-medoidów, jednak nie zmienia zasadniczej wysokiej złożonosci metody k-medoidów. Poprawa efektywnosci jest skutkiem lepszego wyboru poczatkowego medoidów, a co za tym idzie, najczęsciej, mniejszą liczbą iteracji.

Zadanie

Dla ustalonego zbioru danych wykonać grupowanie trzema wybranymi metodami. Dla otrzymanych grup wyliczyć wartosć WCV - suma odległości elementów od odpowiednich centroidów i sporzadzić wykres ilustrujący zależnosć WCV od liczby klastrów.