Eksploracja danych: Wykład 6

Wojciech Kempa

Politechnika Śląska
Wydział Matematyki Stosowanej

Miary podobieństwa dokumentów tekstowych (1)

Oceniając podobieństwo dokumentów tekstowych lub też stron www, wykorzystuje się następujące trzy miary niepodobieństwa:

- **Miara cosinusowa:** dla obiektów X_i i X_j definiuje się ją jako

$$D(X_i, X_j) = 1 - \frac{X_i \circ X_j}{|X_i| \cdot |X_j|}, \tag{1}$$

gdzie symbol "$\circ$" oznacza iloczyn skalarny, a $|\cdot|$ jest długością wektora.

Zauważmy, że odjemnikiem w powyższej różnicy jest wartość cosinusa kąta pomiędzy obiektami X_i i X_j, traktowanymi jako wektory. Im większa wartość miary (przyjmuje ona wartości z przedziału $[0, 1]$), tym większe niepodobieństwo obiektów.

Miary podobieństwa dokumentów tekstowych (2)

- **Odległość Tanimoto**: jest pewną modyfikacją odległości cosinusowej i ma postać

$$D(X_i, X_j) = 1 - \frac{X_i \circ X_j}{|X_i|^2 + |X_j|^2 - X_i \circ X_j}. \tag{2}$$

Miary podobieństwa dokumentów tekstowych (2)

- **Odległość Tanimoto**: jest pewną modyfikacją odległości cosinusowej i ma postać

$$D(X_i, X_j) = 1 - \frac{X_i \circ X_j}{|X_i|^2 + |X_j|^2 - X_i \circ X_j}. \qquad (2)$$

- **Odległość Levensteina** (zwana także **odległością edycyjną**): jest równa minimalnej liczbie operacji prostych (takich jak: wstawienie nowego symbolu, usunięcie symbolu, zamiana jednego symbolu na inny), które przeprowadzają jeden obiekt w drugi.

Metody grupowania obiektów

Ogólna klasyfikacja metod grupowania obiektów, których celem jest ich podział na pewną liczbę rozłącznych skupień (grup, klastrów itp.), zawierających obiekty „podobne” do siebie, jest następująca:

- **metody hierarchiczne**, w wyniku których otrzymujemy tzw. **dendrogram**, który stanowi wizualizację procesu grupowania; metody hierarchiczne dzielimy dodatkowo na

Metody grupowania obiektów

Ogólna klasyfikacja metod grupowania obiektów, których celem jest ich podział na pewną liczbę rozłącznych skupień (grup, klastrów itp.), zawierających obiekty „podobne" do siebie, jest następująca:

- **metody hierarchiczne**, w wyniku których otrzymujemy tzw. **dendrogram**, który stanowi wizualizację procesu grupowania; metody hierarchiczne dzielimy dodatkowo na
  - **metody aglomeracyjne**, w których „kierunek" przeprowadzania grupowania przebiega od maksymalnego rozproszenia obiektów do jednego skupienia, obejmującego wszystkie badane obiekty;

Metody grupowania obiektów

Ogólna klasyfikacja metod grupowania obiektów, których celem jest ich podział na pewną liczbę rozłącznych skupień (grup, klastrów itp.), zawierających obiekty „podobne" do siebie, jest następująca:

- **metody hierarchiczne**, w wyniku których otrzymujemy tzw. **dendrogram**, który stanowi wizualizację procesu grupowania; metody hierarchiczne dzielimy dodatkowo na
  - **metody aglomeracyjne**, w których „kierunek" przeprowadzania grupowania przebiega od maksymalnego rozproszenia obiektów do jednego skupienia, obejmującego wszystkie badane obiekty;
  - **metody podziałowe**, w których proces grupowania przebiega w odwrotnym „kierunku": od jednego skupienia do maksymalnego rozproszenia obiektów; techniki podziałowe są stosowane dużo rzadziej niż techniki aglomeracyjne.

Metody grupowania obiektów

Ogólna klasyfikacja metod grupowania obiektów, których celem jest ich podział na pewną liczbę rozłącznych skupień (grup, klastrów itp.), zawierających obiekty „podobne" do siebie, jest następująca:

- **metody hierarchiczne**, w wyniku których otrzymujemy tzw. **dendrogram**, który stanowi wizualizację procesu grupowania; metody hierarchiczne dzielimy dodatkowo na
  - **metody aglomeracyjne**, w których „kierunek" przeprowadzania grupowania przebiega od maksymalnego rozproszenia obiektów do jednego skupienia, obejmującego wszystkie badane obiekty;
  - **metody podziałowe**, w których proces grupowania przebiega w odwrotnym „kierunku": od jednego skupienia do maksymalnego rozproszenia obiektów; techniki podziałowe są stosowane dużo rzadziej niż techniki aglomeracyjne.
- **metoda k-średnich**, której istota polega na określeniu *a priori* docelowej liczby k skupień; proces grupowania w metodzie k-średnich przebiega iteracyjnie: w każdym kroku iteracji możliwa jest zmiana „lokalizacji" obiektu.

Metody pomiaru odległości skupień (1)

Metody pomiaru odległości pomiędzy skupieniami są scharakteryzowane poniżej.

- **metoda pojedynczego wiązania** (zwana także **metodą najbliższego sąsiedztwa**), w myśl której odległość między dwoma skupieniami jest równa odległości pomiędzy ich elementami najbliżej siebie położonymi;

Metody pomiaru odległości skupień (1)

Metody pomiaru odległości pomiędzy skupieniami są scharakteryzowane poniżej.

- **metoda pojedynczego wiązania** (zwana także **metodą najbliższego sąsiedztwa**), w myśl której odległość między dwoma skupieniami jest równa odległości pomiędzy ich elementami najbliżej siebie położonymi;
- **metoda pełnego wiązania** (zwana również **metodą najdalszego sąsiedztwa**), w której odległość między dwoma skupieniami jest równa odległości pomiędzy ich elementami najbardziej od siebie odległymi;

Metody pomiaru odległości skupień (1)

Metody pomiaru odległości pomiędzy skupieniami są scharakteryzowane poniżej.

- **metoda pojedynczego wiązania** (zwana także **metodą najbliższego sąsiedztwa**), w myśl której odległość między dwoma skupieniami jest równa odległości pomiędzy ich elementami najbliżej siebie położonymi;
- **metoda pełnego wiązania** (zwana również **metodą najdalszego sąsiedztwa**), w której odległość między dwoma skupieniami jest równa odległości pomiędzy ich elementami najbardziej od siebie odległymi;
- **metoda średnich połączeń**, w myśl której odległość pomiędzy skupieniami jest definiowana jako średnia arytmetyczna odległości pomiędzy elementami należącymi do jednego i drugiego skupienia.

Metody pomiaru odległości skupień (2)

- **Metoda ważonych średnich połączeń** jest pewną modyfikacją metody średnich połączeń. Wprowadza się w niej wagi poszczególnych skupień, którymi są ich liczebności.

Metody pomiaru odległości skupień (2)

- **Metoda ważonych średnich połączeń** jest pewną modyfikacją metody średnich połączeń. Wprowadza się w niej wagi poszczególnych skupień, którymi są ich liczebności.
- Jeżeli na danym etapie doszło do połączenia dwóch skupień S_i i S_j (często o różnej liczebności obiektów), to odległość tak powstałego większego skupienia od dowolnego innego skupienia S_k jest równa $\frac{1}{2}\big(D(S_i, S_k) + D(S_j, S_k)\big)$ (czyli osobno obliczana jest odległość pomiędzy S_k a każdym ze skupień S_i, S_j tworzących nowe skupienie, co nie ma miejsca w przypadku metody nieważonej).

Metody pomiaru odległości skupień (3)

- **Metoda środków ciężkości**, której idea opiera się na przyjęciu jako miary odległości pomiędzy skupieniami odległości środków ciężkości tych skupień, czyli wielowymiarowych średnich arytmetycznych współrzędnych obiektów.

Metody pomiaru odległości skupień (3)

- **Metoda środków ciężkości**, której idea opiera się na przyjęciu jako miary odległości pomiędzy skupieniami odległości środków ciężkości tych skupień, czyli wielowymiarowych średnich arytmetycznych współrzędnych obiektów.
- **Metoda ważonych środków ciężkości** jest modyfikacją metody środków ciężkości, analogiczną do ważonej metody średnich połączeń.

Metody pomiaru odległości skupień (3)

- **Metoda środków ciężkości**, której idea opiera się na przyjęciu jako miary odległości pomiędzy skupieniami odległości środków ciężkości tych skupień, czyli wielowymiarowych średnich arytmetycznych współrzędnych obiektów.
- **Metoda ważonych środków ciężkości** jest modyfikacją metody środków ciężkości, analogiczną do ważonej metody średnich połączeń.
- **Metoda Warda** wykorzystuje podejście analizy wariancji do oszacowania odległości pomiędzy dwoma skupieniami.

Metoda Warda (1)

- Metoda grupowania Warda uchodzi za bardzo efektywną, choć w praktyce jej zastosowanie może skutkować tworzeniem na kolejnych etapach aglomeracji relatywnie dużej liczby skupień o niewielkich liczebnościach.

Metoda Warda (1)

- Metoda grupowania Warda uchodzi za bardzo efektywną, choć w praktyce jej zastosowanie może skutkować tworzeniem na kolejnych etapach aglomeracji relatywnie dużej liczby skupień o niewielkich liczebnościach.
- Oczywiście, z tego właśnie powodu może być ona w niektórych sytuacjach rekomendowana. Kluczowy w metodzie Warda jest specyficzny sposób wyznaczania odległości pomiędzy skupieniami.

Metoda Warda (1)

- Metoda grupowania Warda uchodzi za bardzo efektywną, choć w praktyce jej zastosowanie może skutkować tworzeniem na kolejnych etapach aglomeracji relatywnie dużej liczby skupień o niewielkich liczebnościach.
- Oczywiście, z tego właśnie powodu może być ona w niektórych sytuacjach rekomendowana. Kluczowy w metodzie Warda jest specyficzny sposób wyznaczania odległości pomiędzy skupieniami.
- Utworzenie pierwszego skupienia następuje w zwykły sposób: łączymy dwa obiekty najbliżej siebie położone.

Metoda Warda (2)

- W kolejnym etapie (i wszystkich następnych) musimy oszacować odległość pomiędzy pojedynczym obiektem a skupieniem obiektów lub też pomiędzy dwoma skupieniami.

Metoda Warda (2)

- W kolejnym etapie (i wszystkich następnych) musimy oszacować odległość pomiędzy pojedynczym obiektem a skupieniem obiektów lub też pomiędzy dwoma skupieniami.
- Załóżmy zatem, że $S = S_i \cup S_j$, gdzie S_i i S_j są pojedynczymi obiektami lub skupieniami.

Metoda Warda (2)

- W kolejnym etapie (i wszystkich następnych) musimy oszacować odległość pomiędzy pojedynczym obiektem a skupieniem obiektów lub też pomiędzy dwoma skupieniami.
- Załóżmy zatem, że $S = S_i \cup S_j$, gdzie S_i i S_j są pojedynczymi obiektami lub skupieniami.
- Odległość $D(S, S_k)$ pomiędzy skupieniem S a skupieniem S_k obliczamy ze wzoru

$$D(S, S_k) \stackrel{def}{=} a_1 \cdot D(S_i, S_k) + a_2 \cdot D(S_j, S_k) + b \cdot D(S_i, S_j), \qquad (3)$$

gdzie a_1, a_2 oraz b są współczynnikami zależnymi od liczebności skupień S_i, S_j oraz S_k i są zdefiniowane następująco:

Metoda Warda (3)

$$a_1 = \frac{|S_i| + |S_k|}{|S_i| + |S_j| + |S_k|},$$
$$a_2 = \frac{|S_j| + |S_k|}{|S_i| + |S_j| + |S_k|},$$
$$b = -\frac{|S_k|}{|S_i| + |S_j| + |S_k|}. \quad (4)$$

Zauważmy zatem, że odległość pomiędzy skupieniem S a skupieniem S_k zależy od odległości pomiędzy skupieniami S_i i S_j, które utworzyły S **w poprzednim etapie aglomeracji.**

Analiza dendrogramu (1)

Poniżej przedstawiamy najważniejsze trzy metody „przecinania” dendrogramu.

- **Metoda 1:** dendrogram „przecinamy” w miejscu, w którym odległość pomiędzy dwoma kolejnymi wiązaniami jest maksymalna.

Analiza dendrogramu (1)

Poniżej przedstawiamy najważniejsze trzy metody „przecinania” dendrogramu.

- **Metoda 1:** dendrogram „przecinamy” w miejscu, w którym odległość pomiędzy dwoma kolejnymi wiązaniami jest maksymalna.
- W analizowanym wcześniej Przykładzie 1 kolejne odległości pomiędzy wiązaniami wynoszą $6 - 2 = 4$ oraz $7 - 6 = 1$. Metoda rekomenduje zatem „przecięcie” dendrogramu pomiędzy wiązaniem pierwszym a drugim. Optymalny wynik grupowania to zatem jedno skupienie $S_1 = \{O_1, O_3\}$ oraz dwa izolowane obiekty (można je traktować jako skupienia jednoelementowe) O_2 i O_4.

Analiza dendrogramu (2)

- **Metoda 2: miernik Grabińskiego**, zaproponowany przez niego w 1992 roku, definiuje wielkość $q_i \stackrel{def}{=} \frac{d_i}{d_{i-1}}$, gdzie d_i oznacza odległość, przy której dochodzi do i-tego z kolei wiązania.

Analiza dendrogramu (2)

- **Metoda 2: miernik Grabińskiego**, zaproponowany przez niego w 1992 roku, definiuje wielkość $q_i \stackrel{def}{=} \frac{d_i}{d_{i-1}}$, gdzie d_i oznacza odległość, przy której dochodzi do i-tego z kolei wiązania.
- Dendrogram „przecinamy" pomiędzy wiązaniem numer $i_0 - 1$ a i_0, gdzie i_0 jest wartością i, dla której q_i jest maksymalne.
- Podstawową wadą tej metody jest fakt, że często w praktyce zdarza się, że q_i przyjmuje wartość maksymalną już dla jednych z pierwszych wiązań dendrogramu.

Analiza dendrogramu (3)

- **Metoda 3: reguła Mojeny**, sformułowana w 1977 roku, stanowi, że dendrogram „przecinamy" pomiędzy wiązaniem numer i a wiązaniem numer $i+1$, gdzie $d_{i+1} > \overline{d} + k \cdot s_d$, przy czym $\overline{d}$ oznacza średnią odległość wiązania, natomiast s_d jest odchyleniem standardowym odległości wiązania. Wielkość k jest stałą.

Analiza dendrogramu (3)

- **Metoda 3: reguła Mojeny**, sformułowana w 1977 roku, stanowi, że dendrogram „przecinamy" pomiędzy wiązaniem numer i a wiązaniem numer $i+1$, gdzie $d_{i+1} > \overline{d} + k \cdot s_d$, przy czym $\overline{d}$ oznacza średnią odległość wiązania, natomiast s_d jest odchyleniem standardowym odległości wiązania. Wielkość k jest stałą.
- Twórca metody zaproponował, by $k \in (2.75, 3.50)$. Później (Milligan, Cooper, 1985) zaproponowano jako optymalną wartość $k = 1.25$.

Metoda k-średnich (1)

- Istotą **metody k-średnich** jest ustalenie na samym początku procesu grupowania docelowej liczby skupień k.

Metoda k-średnich (1)

- Istotą **metody k-średnich** jest ustalenie na samym początku procesu grupowania docelowej liczby skupień k.
- Proces wygląda w następujący sposób. Na początku dzielimy wszystkie obiekty na k początkowych skupień.

Metoda k-średnich (1)

- Istotą **metody k-średnich** jest ustalenie na samym początku procesu grupowania docelowej liczby skupień k.
- Proces wygląda w następujący sposób. Na początku dzielimy wszystkie obiekty na k początkowych skupień.
- Na tym etapie wybiera się tzw. **wstępne centra skupień (centroidy)**, czyli punkty, które będą decydować o pierwszej przynależności obiektów (w praktyce stosuje się różne metody wyznaczania takich centrów).

Metoda k-średnich (1)

- Istotą **metody k-średnich** jest ustalenie na samym początku procesu grupowania docelowej liczby skupień k.
- Proces wygląda w następujący sposób. Na początku dzielimy wszystkie obiekty na k początkowych skupień.
- Na tym etapie wybiera się tzw. **wstępne centra skupień (centroidy)**, czyli punkty, które będą decydować o pierwszej przynależności obiektów (w praktyce stosuje się różne metody wyznaczania takich centrów).
- Mając wybrane wstępne centra skupień, obliczamy odległości każdego z obiektów od każdego z centrów i przyporządkowujemy obiekt do tego skupienia, do którego centrum jest mu „najbliżej”.

Metoda k-średnich (2)

- W kolejnym etapie wyznaczamy już rzeczywiste centra utworzonych skupień i odległości każdego z obiektów od nich.

Metoda k-średnich (2)

- W kolejnym etapie wyznaczamy już rzeczywiste centra utworzonych skupień i odległości każdego z obiektów od nich.
- I tu zdarzyć się może, że obiekt znajduje się w skupieniu, którego centrum jest od niego bardziej odległe niż centrum innego skupienia.

Metoda k-średnich (2)

- W kolejnym etapie wyznaczamy już rzeczywiste centra utworzonych skupień i odległości każdego z obiektów od nich.
- I tu zdarzyć się może, że obiekt znajduje się w skupieniu, którego centrum jest od niego bardziej odległe niż centrum innego skupienia.
- W takiej sytuacji dokonujemy odpowiedniego przeniesienia obiektu (lub obiektów, gdy jest ich więcej).

Metoda k-średnich (2)

- W kolejnym etapie wyznaczamy już rzeczywiste centra utworzonych skupień i odległości każdego z obiektów od nich.
- I tu zdarzyć się może, że obiekt znajduje się w skupieniu, którego centrum jest od niego bardziej odległe niż centrum innego skupienia.
- W takiej sytuacji dokonujemy odpowiedniego przeniesienia obiektu (lub obiektów, gdy jest ich więcej).
- Następnie ponownie wyznaczamy nowe centra skupień i powtarzamy całą procedurę.

Metoda k-średnich (2)

- W kolejnym etapie wyznaczamy już rzeczywiste centra utworzonych skupień i odległości każdego z obiektów od nich.
- I tu zdarzyć się może, że obiekt znajduje się w skupieniu, którego centrum jest od niego bardziej odległe niż centrum innego skupienia.
- W takiej sytuacji dokonujemy odpowiedniego przeniesienia obiektu (lub obiektów, gdy jest ich więcej).
- Następnie ponownie wyznaczamy nowe centra skupień i powtarzamy całą procedurę.
- W pewnym momencie obiekty przestaną już być przenoszone, ponieważ znajdą się w skupieniach, do których jest im „najbliżej”. W tym momencie proces grupowania się kończy.