

Zaawansowane uczenie maszynowe:
wykład 12

Paweł Cichosz

- 1 Zadanie grupowania
- 2 Algorytmy k -środków
- 3 Grupowanie gęstościowe
- 4 Ocena jakości grupowania

Zadanie grupowania

Dziedzina, przykład, atrybut, zbiór trenujący: jak dla klasyfikacji i regresji, ale bez wyróżnionego atrybutu docelowego.

Model: $h : X \rightarrow C_h$, gdzie C_h oznacza zbiór grup dla modelu grupowania h – nie tylko podział zbioru trenującego na grupy, lecz także możliwość przypisania do tych grup dowolnych przykładów z dziedziny (choć nie zawsze ta możliwość jest oferowana).

Grupy a podobieństwo: maksymalizacja podobieństwa wewnątrzgrupowego, minimalizacja podobieństwa międzygrupowego.

Zadanie grupowania

Scenariusze użycia:

- odkrycie wiedzy o wzorcach podobieństwa w dziedzinie,
- grupowanie na podstawie atrybutów obserwowalnych (dostępnych zawsze) w celu wnioskowania o atrybutach ukrytych (dostępnych rzadko),
- nienadzorowana detekcja anomalii jako przykładów niepodobnych do żadnej grupy,
- dekompozycja dziedziny w celu uproszczenia innych zadań modelowania.

Typy algorytmów grupowania

- Na podstawie **niepodobieństwa**: wykorzystujące jawnie zadaną miarę niepodobieństwa do ustalenia przynależności przykładów do grup.
- Na podstawie **gęstości**: wykorzystujące gęstość przykładów w przestrzeni atrybutów do wyznaczenia podziału na grupy.
- Na podstawie **rozkładu**: dopasowujące mieszaninę rozkładów prawdopodobieństwa do danych (nie będą omawiane).

Miary niepodobieństwa

Odległość euklidesowa:

$$\delta_{\text{euc}}(x_1, x_2) = \sqrt{\sum_{i=1}^n (a_i(x_1) - a_i(x_2))^2}$$

Odległość Minkowskiego:

$$\delta_{\text{mink}}(x_1, x_2) = \left(\sum_{i=1}^n |a_i(x_1) - a_i(x_2)|^p \right)^{\frac{1}{p}}$$

Odległość miejska (Manhattan): odległość Minkowskiego dla $p = 1$:

$$\delta_{\text{man}}(x_1, x_2) = \sum_{i=1}^n |a_i(x_1) - a_i(x_2)|$$

Odległość Czebyszewa (maksimum): odległość Minkowskiego dla $p \rightarrow \infty$:

$$\delta_{\text{max}}(x_1, x_2) = \max_i |a_i(x_1) - a_i(x_2)|$$

Miary niepodobieństwa

Odległość Mahalanobisa: rozszerzony wariant odległości euklidesowej, kompensujący zróżnicowaną wariancję atrybutów oraz korelacje między atrybutami:

$$\delta_{\text{mah}}(x_1, x_2) = \sqrt{(\mathbf{a}(x_1) - \mathbf{a}(x_2))^T \text{Cov}^{-1}(\mathbf{a})(\mathbf{a}(x_1) - \mathbf{a}(x_2))}$$

gdzie:

- $\mathbf{a}(x)$ – (kolumnowy) wektor różnic wartości atrybutów a_1, a_2, \dots, a_n dla przykładu x ,
- $\text{Cov}^{-1}(\mathbf{a})$ oznacza odwróconą macierz kowariancji atrybutów a_1, a_2, \dots, a_n .

Miary niepodobieństwa

Odległość Hamminga:

$$\delta_{\text{ham}}(x_1, x_2) = \sum_{i=1}^n \mathbb{I}_{a_i(x_1) \neq a_i(x_2)}$$

gdzie

$$\mathbb{I}_{a_i(x_1) \neq a_i(x_2)} = \begin{cases} 1 & \text{jeśli } a_i(x_1) \neq a_i(x_2) \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Niepodobieństwo kosinusowe:

$$\delta_{\text{cos}}(x_1, x_2) = 1 - \frac{\mathbf{a}(x_1) \circ \mathbf{a}(x_2)}{\|\mathbf{a}(x_1)\| \cdot \|\mathbf{a}(x_2)\|}$$

Miary niepodobieństwa

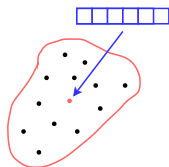
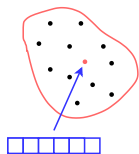
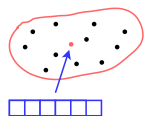
Transformacja atrybutów: w przypadku użycia miar opartych na różnicach wartości atrybutów zazwyczaj wskazana standardyzacja (lub normalizacja).

Ważenie atrybutów: uwzględnianie różnicy wartości $a_i(x_1) - a_i(x_2)$ z wagą $\eta_i \geq 0$.

Uwzględnianie atrybutów dyskretnych: np. zastąpienie różnicy wartości $a_i(x_1) - a_i(x_2)$ przez binarny indyktor $\mathbb{I}_{a_i(x_1) \neq a_i(x_2)}$ (jak w odległości Hamminga).

- 1 Zadanie grupowania
- 2 Algorytmy k -środków**
- 3 Grupowanie gęstościowe
- 4 Ocena jakości grupowania

Reprezentacja grupowania



Liczba grup: k – ustalona.

Reprezentacja grupy: wektor środkowy ζ_d dla każdej grupy $d = 1, 2, \dots, k$ – wektor wartości atrybutów używanych do grupowania.

Elementy grupy: zbiór przykładów trenujących T_d zaliczonych do grupy d dla $d = 1, 2, \dots, k$ na podstawie niepodobieństwa od wektora środkowego ζ_d .

Schemat algorytmu

- 1 wybierz początkowe wektory środkowe $\zeta_1, \zeta_2, \dots, \zeta_k$;
 - 2 **powtarzaj**:
 - 1 **dla wszystkich** $x \in T$:
przypisz x do grupy $d_x = \arg \min_d \delta(x, \zeta_d)$;
 - 2 **dla** $d = 1, 2, \dots, k$:
modyfikuj wektor środkowy ζ_d na podstawie zbioru przykładów T_d ;
- jak długo** nie są spełnione kryteria stopu.

Schemat algorytmu

Inicjalizacja wektorów środkowych: np. losowy wybór k przykładów trenujących.

Modyfikacja wektorów środkowych: różne warianty.

Kryteria stopu:

- zbieżność (brak zmiany przynależności przykładów do grup),
- przybliżona zbieżność (mała liczba przykładów zmieniających przynależność do grup),
- zadana liczba iteracji.

Dobór k : na podstawie oceny jakości grupowania.

Warianty

k -średnich:

wektory środkowe: wektory *średnich* wartości atrybutów dla przykładów z poszczególnych grup,

właściwości: wrażliwość na wartości odstające, gwarancja zbieżności z odległością euklidesową, lokalna minimalizacja sumy kwadratów odległości euklidesowych przykładów od odpowiednich środków grup.

k -median:

wektory środkowe: wektory *median* wartości atrybutów dla przykładów z poszczególnych grup,

właściwości: większa odporność na wartości odstające, gwarancja zbieżności z odległością miejską, lokalna minimalizacja sumy odległości miejskich przykładów od odpowiednich środków grup.

Warianty

k -medoid:

- wektory środkowe:** faktyczne przykłady trenujące minimalizujące sumę niepodobieństw od przykładów z poszczególnych grup,
- popularna realizacja:** PAM (*partitioning around medoids*),
- właściwości:** największa odporność na wartości odstające i zależności między atrybutami, gwarancja zbieżności z dowolną miarą niepodobieństwa, lokalna minimalizacja sumy niepodobieństw przykładów od odpowiednich środków grup.

- 1 Zadanie grupowania
- 2 Algorytmy k -środków
- 3 Grupowanie gęstościowe**
- 4 Ocena jakości grupowania

DBSCAN

DBSCAN: *Density-Based Spatial Clustering of Applications with Noise.*

Liczba grup: dobierana automatycznie.

Reprezentacja grupowania: przypisanie przykładów trenujących do grup lub oznaczenie ich jako „szumu” (przykładów odstających).

Zasada działania: grupy powstają wokół przykładów o wystarczająco gęstym ϵ -sąsiedztwie:

$$T_{x,\epsilon} = \{x' \in T \mid \delta(x', x) \leq \epsilon\}$$

Schemat algorytmu

1 oznacz wszystkie przykłady jako *nieodwiedzone*;

2 **powtarzaj:**

1 wybierz *nieodwiedzony* przykład x_s ;

2 oznacz x_s jako *odwiedzony*;

3 **jeżeli** $|T_{x_s, \epsilon}| < m$:

 oznacz x_s jako *szum*;

w przeciwnym przypadku:

1 utwórz nową grupę d_{x_s} ;

2 $T_{d_{x_s}} := \{x_s\}$;

3 *rozszerzenie*($d_{x_s}, T_{x_s, \epsilon}$);

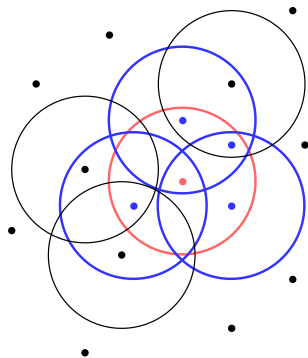
dopóki wszystkie przykłady nie zostaną oznaczone jako *odwiedzone*.

Schemat algorytmu

rozszerzenie(d, S):

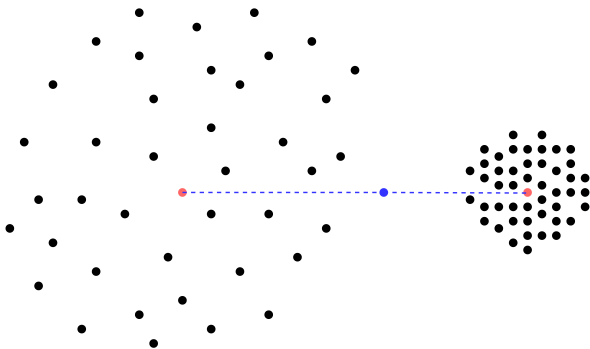
dla wszystkich $x \in S$:

- 1 oznacz x jako odwiedzony;
- 2 $T_d := T_d \cup \{x\}$;
- 3 *rozszerzenie*($d, T_{x,\epsilon}$).



Właściwości

- Możliwość identyfikacji grup o dowolnym „kształcie”, niekoniecznie sferycznych jak w przypadku grupowania k -środków.
- Wykrywanie przykładów odstających.
- Brak optymalizacji jawnie określonej miary jakości grupowania.
- Dobór parametrów ϵ , m trudniejszy niż dobór k .



- 1 Zadanie grupowania
- 2 Algorytmy k -środków
- 3 Grupowanie gęstościowe
- 4 Ocena jakości grupowania

Ocena wewnętrzna

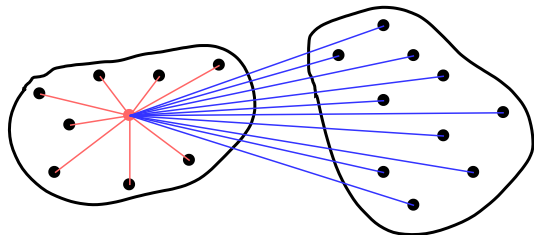
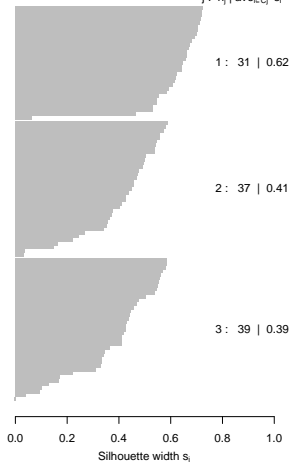
Wiele wskaźników jakości opartych na podobieństwie wewnątrzgrupowym i zróżnicowaniu międzygrupowym.

Szerokość sylwetki: miara dopasowania przykładu do jego grupy, której średnia wartość może być miarą jakości grupowania:

$$sw(x) = \frac{\min_d \Delta(x, d) - \Delta(x)}{\max\{\min_d \Delta(x, d), \Delta(x)\}}$$

gdzie $\Delta(x, d)$ – średnie niepodobieństwo x do przykładów z innej grupy d , $\Delta(x)$ – średnie niepodobieństwo x do przykładów z jego własnej grupy.

Ocena wewnętrzna

Training set, $k=3$ $n = 107$ 3 clusters C_j
 $j : n_j \mid \text{ave}_{i \in C_j} s_i$ 

Average silhouette width : 0.46

Ocena zewnętrzna

Ocena zgodności z zewnętrznym etykietowaniem.

Indeks Randa:

$$\text{rand}_c(h) = \frac{\text{TP}_c(h) + \text{TN}_c(h)}{\text{TP}_c(h) + \text{TN}_c(h) + \text{FP}_c(h) + \text{FN}_c(h)}$$

gdzie

$$\text{TP}_c(h) = |\{\langle x_1, x_2 \rangle \mid c(x_1) = c(x_2) \wedge h(x_1) = h(x_2)\}|$$

$$\text{TN}_c(h) = |\{\langle x_1, x_2 \rangle \mid c(x_1) \neq c(x_2) \wedge h(x_1) \neq h(x_2)\}|$$

$$\text{FP}_c(h) = |\{\langle x_1, x_2 \rangle \mid c(x_1) \neq c(x_2) \wedge h(x_1) = h(x_2)\}|$$

$$\text{FN}_c(h) = |\{\langle x_1, x_2 \rangle \mid c(x_1) = c(x_2) \wedge h(x_1) \neq h(x_2)\}|$$