Metody grupowania

dr inż. Tomasz Piłot

1

Literatura

• Rutkowski Leszek: Metody i techniki sztucznej inteligencji, PWN, 2005

2

Pojęcie[1]

Celem grupowania jest podział na grupy podobnych danych.

Grupowania wykonywane przez człowieka:

- ograniczone do dwu-, trójwymiarowych danych,
- ograniczenie liczby grupowanych przypadków

=> automatyczne metody grupowania.

4

Metody

- ogólny algorytm k-means (k-średnich)
- algorytm hard k-means
- algorytm fuzzy k-means
- algorytm possibilistic k-means
- algorytm Gustafsona-Kessela

5

Cechy prawidłowych grup

1. Homogeniczność w grupach
   duże podobieństwo elementów w grupie

2. Heterogeniczność pomiędzy grupami
   małe podobieństwo elementów należących do różnych grup

6

Sposoby definicji podobieństwa

1. Zależne od typu danych
2. W przypadku cech numerycznych: miara odległości np. miara euklidesowa
3. W przypadku cech jakościowych: miara rangowa
4. Problem wyboru reprezentacji grupy
   – centralny punkt
   – kształty grup
   – zależności logiczne

8

Reprezentacja danych

Niech x_k bedzie wektorem danych do grupowania:

$x_k = [x_{k1}, x_{k2}, \ldots x_{kn}], x_k \in R^n$, *j = 1..n*

k=1..*p*– numer przykładu

zbiór *p* przykładów tworzy macierz **X** :

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

9

Reprezentacja

W wyniku grupowania uzyskuje się *c* grup, czyli wektorów środków grup w przestrzeni R^n:

$$v_i = [v_{i1}, v_{i2}, \ldots, v_{in}], i = 1, \ldots, c$$

Typy podziału danych [1]:

- podział ostry
- podział rozmyty
- podział posybilistyczny

10

Przestrzenie podziału danych

Definicja ostrego podziału:

$$P_o = \left\{ \boldsymbol{U} \in \boldsymbol{R}^{c\times p} \mid \mu_{ik} \in \{0,1\};\ \forall k \sum_{i=1}^{c} \mu_{ik} = 1 \right\}$$

Definicja podziału rozmytego:

$$P_R = \left\{ \boldsymbol{U} \in \boldsymbol{R}^{c\times p} \mid \mu_{ik} \in [0,1];\ \forall k \sum_{i=1}^{c} \mu_{ik} = 1 \right\}$$

Definicja podziału posybilistycznego:

$$P_P = \{ \boldsymbol{U} \in \boldsymbol{R}^{c\times p} \mid \mu_{ik} \in [0,1] \}$$

11

Przykłady podziałów danych na grupy

Podział ostry: całkowita przynależność, grupa główna

Podział rozmyty: przynależność częściowa z sumowaniem do 1, grupa główna i grupy poboczne

Podział posybilistyczny: przynależność rozmyta – suma nieograniczona, wielogrupowa przynależność

12

Metody

- ogólny algorytm k-means (k-średnich)
- algorytm hard k-means
- algorytm fuzzy k-means
- algorytm possibilistic k-means
- algorytm Gustafsona-Kessela

13

Algorytm k-średnich (ang. *k-means*)

Celem alg. jest znalezienie grup w danych wejściowych

Założenia: dana jest liczba grup *k*, na które chcemy dokonać podziału

Algorytm:

1. Liczba grup k
2. Przypisz w sposób losowy k przykładów jako początkowe środki (centroidy) grup
3. Dla każdego przykładu znajdź centroid (środek) grupy, które jest najbliższy (miara odległości euklidesowej lub innej).
4. Uaktualnij położenie centroidów (środków) grup (średnia z położeń poszczególnych przykładów w przestrzeni cech wejściowych przypisanych/przynależących do danej grupy).
5. Kroki 3-5 należy powtórzyć do momentu braku zmian położenia środków grup lub kiedy sumaryczny błąd kwadratowy dany wzorem:

$$SSE = \sum_{i=1}^{c} \sum_{x_k \in P_{oi}} d(x_k, v_i)^2 \quad , SSE < \xi$$

gdzie: $\mathbf{x}_k$ – punkt danych w i-tej grupie,

$\mathbf{v}_i$ – centroid i-tej grupy

14

Algorytm hard k-means

1. Inicjalizacja
   - wprowadzenie liczby grup *k*
   - przypisanie położeń poszczególnych grup **V** inicjalizowanych losowymi przykładami **X**, określenie macierzy **U** ∈ P_O
2. Wyznaczenie grup dla poszczególnych przykładów - obliczenie przynależności do grupy względem wybranej miary odległości:

$$\mu_{ik} = \begin{cases} 1, & \|x_k - v_i\| < \|x_k - v_j\|, \forall j, i \neq j \\ & 0 \end{cases}$$

3. Określenie położenia centroidów dla poszczególnych grup – średnia położenia obiektów należących do danej grupy

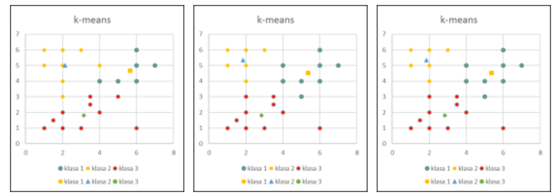
$$v_i = \frac{\sum_{k=1}^{p} \mu_{ik} x_k}{\sum_{k=1}^{p} \mu_{ik}}$$

4. Sprawdzenie warunku zatrzymania algorytmu i ew. przejście do (2)

$$\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)} < \varepsilon$$

15

Przykład działania algorytmu



16

Zalety i wady algorytmu K-means

Zalety:
- można zastosować różne miary odległości
- wynikiem jest podział na grupy istniejących przykładów

Wady:
- nie można dostosować miary odległości do danej grupy, stąd wszystkie grupy mają taki sam kształt
- grupa jest tworzona na zasadzie całkowitej przynależności

17

Algorytm fuzzy k-means

Algorytm jest wynikiem minimalizacji kryterium odległości z uwzględnieniem funkcji przynależności dla zbiorów rozmytych:

$$K = \sum_{i=1}^{c} \sum_{k=1}^{p} \left(\mu_{ik}{}^{m}\right) \left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^2$$

gdzie:

$\mu_{ik} \in P_R$

$\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^2 = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T A (\boldsymbol{x}_k - \boldsymbol{v}_i)$ - jest odległością danego przykładu od środka grupy z uwzględnieniem macierzy wag A w mierze odległości euklidesowej.

m ∈ (1, ∞) – poziom rozmycia przypisania przykładów do grup

Przestrzeń P_R :

$$P_R = \left\{ \boldsymbol{U} \in \boldsymbol{R}^{c \times p} \mid \mu_{ik} \in [0,1];\ \forall k \sum_{i=1}^{c} \mu_{ik} = 1 \right\}$$

18

Algorytm fuzzy k-means

1. Inicjalizacja
   - wprowadzenie liczby grup *k*, rozmycia *m* oraz progu warunku zatrzymania alg. ε
   - przypisanie położeń poszczególnych grup **V** inicjalizowanych losowymi przykładami **X**, określenie macierzy **U** ∈ P_R
   - krok iteracji t = 1
2. Wyznaczenie grup dla poszczególnych przykładów

   obliczenie przyporządkowania względem odległości w wybranej mierze:

$$\mu_{ik}{}^{(t)} = \left\{ \sum_{j=1}^{c} \left(\frac{\left\| x_k - v_i{}^{(t-1)} \right\|}{\left\| x_k - v_j{}^{(t-1)} \right\|} \right)^{\frac{2}{m-1}} \right\}^{-1}, \quad i = 1 \ldots c, \quad k = 1 \ldots p$$

3. Określenie położenia centroidów dla poszczególnych grup

$$v_i = \frac{\sum_{k=1}^{p} (\mu_{ik})^m x_k}{\sum_{k=1}^{p} (\mu_{ik})^m}$$

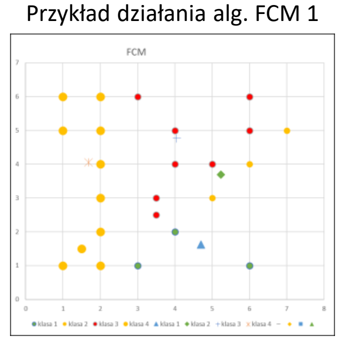
4. Sprawdzenie warunku zatrzymania algorytmu i ew. przejście do (2)

   - zmiana w przyporządkowaniu

$$\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)} < \varepsilon$$

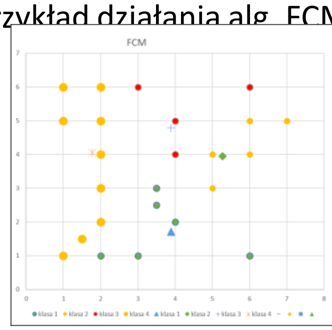
19

Przykład działania alg. FCM 1



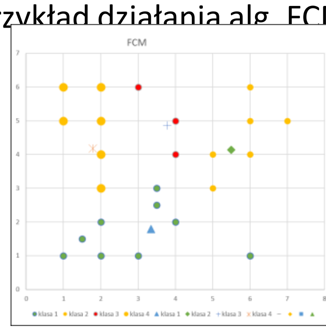
20

Przykład działania alg. FCM 2



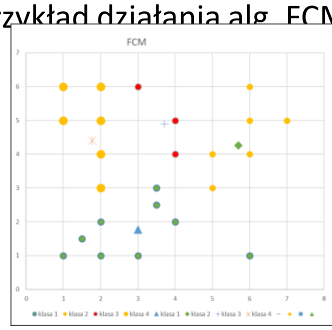
21

Przykład działania alg. FCM 3



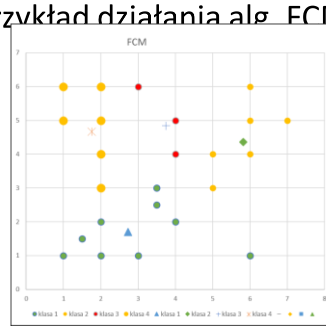
22

Przykład działania alg. FCM 4



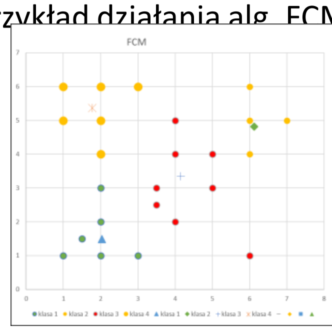
23

Przykład działania alg. FCM 5



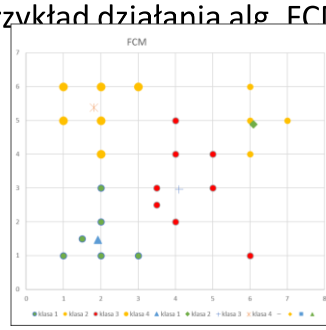
24

Przykład działania alg. FCM 6



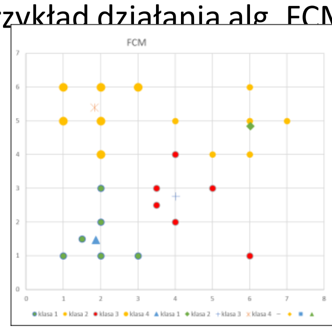
25

Przykład działania alg. FCM 7



26

Przykład działania alg. FCM 8



27

Algorytm possybilistic k-means

Algorytm jest wynikiem minimalizacji kryterium odległości z uwzględnieniem funkcji przynależności dla zbiorów rozmytych:

$$K = \sum_{i=1}^{c}\sum_{k=1}^{p}\left(\mu_{ik}{}^{m}\right)\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^{2} + \sum_{i=1}^{c}\eta_i\sum_{k=1}^{p}(1-\mu_{ik})^m$$

gdzie:

$\mu_{ik} \in P_P$

$\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^{2} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T A(\boldsymbol{x}_k - \boldsymbol{v}_i)$ - jest odległością danego przykładu od środka grupy z uwzględnieniem macierzy wag A w mierze odległości euklidesowej.

$m \in (1, \infty)$ – poziom rozmycia przypisania przykładów do grup

η_i - szerokość wynikowego rozkładu posybilistycznego – kryterium z tym związane wymusza, aby stopnie przynależności były możliwie duże, bez czego rozwiązanie mogłoby zostać osiągnięte przy $\mu_{ik} = 0$

Kryterium K zabezpiecza przed przesuwaniem się środków grup w przypadku występowania w danych szumów lub błędów.

Funkcja celu K może zostać rozbita na szereg podfunkcji osobnych dla każdej grupy, co pozwala minimalizować każdą z nich z osobna.

28

Algorytm possybilistic k-means

Algorytm jest wynikiem minimalizacji kryterium odległości z uwzględnieniem funkcji przynależności dla zbiorów rozmytych:

$$K = \sum_{i=1}^{c}\sum_{k=1}^{p}\left(\mu_{ik}{}^{m}\right)\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^{2} + \sum_{i=1}^{c}\eta_i\sum_{k=1}^{p}(1-\mu_{ik})^m$$

Funkcja celu K może zostać rozbita na szereg podfunkcji osobnych dla każdej grupy, co pozwala minimalizować każdą z nich z osobna:

$$\mu_{ik} = \left(1 + \left(\frac{\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A}{\eta_i}\right)^{\frac{2}{m-1}}\right)^{-1}$$

W celu określenia η_i można przyjąć stałą lub przyjąć dla każdej z grup na podstawie średniej odległości obiektów od środka danej grupy:

$$\eta_i = \frac{\sum_{k=1}^{p}(\mu_{ik})^m\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A^{2}}{\sum_{k=1}^{p}\ \mu_{ik}{}^{m}}$$

29

Algorytm possybilistic k-means

1. Inicjalizacja
   - wprowadzenie liczby grup *k,* rozmycia *m* oraz progu warunku zatrzymania alg. ε
   - przypisanie położeń poszczególnych grup **V** inicjalizowanych losowymi przykładami **X**, określenie macierzy **U** $\in P_R$
   - krok iteracji t = 1
2. Wyznaczenie grup dla poszczególnych przykładów

   obliczenie przyporządkowania względem odległości w wybranej mierze:

   $$\mu_{ik} = \left(1 + \left(\frac{\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_A}{\eta_i}\right)^{\frac{2}{m-1}}\right)^{-1}, \qquad i = 1 \dots c, \qquad k = 1 \dots p$$
3. Określenie położenia centroidów dla poszczególnych grup

   $$\boldsymbol{v}_i = \frac{\sum_{k=1}^{p}\ \mu_{ik}{}^{m}\boldsymbol{x}_k}{\sum_{k=1}^{p}\mu_{ik}{}^{m}}$$
4. Sprawdzenie warunku zatrzymania algorytmu i ew. przejście do (2)

   $$\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)} < \varepsilon$$

30

Wady algorytmu PKM

Wady:

- nie można dostosować miary odległości do danej grupy, stąd wszystkie grupy mają taki sam kształt

31

Algorytm Gustafsona-Kessela

Algorytm ten jest modyfikacją algorytmu FCM, kryterium przyjmuje postać:

$$K = \sum_{i=1}^{c}\sum_{k=1}^{p}\left(\mu_{ik}{}^{m}\right)\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_{A_i}^{2}$$

gdzie:

$\mu_{ik} \in P_P,$

$\left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|_{A_i}^{2} = \left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|^T A_i \left\|\boldsymbol{x}_k - \boldsymbol{v}_i\right\|$

- jest odległością danego przykładu od środka grupy z uwzględnieniem macierzy wag A_i w mierze odległości euklidesowej.

$m \in (1, \infty)$ – poziom rozmycia przypisania przykładów do grup

η_i - szerokość wynikowego rozkładu posybilistycznego.

Macierz wag **A** jest liczona osobno dla każdej grupy. Pozwala to uzyskać różne kształty dla różnych grup i zabezpiecza algorytm przed szukaniem kształtu grupy, którego nie ma w danych.

Wprowadzając ograniczenie na macierz A_i:

$$\det(A_i) = \alpha_i, \qquad \alpha_i > 0$$

gdzie α_i jest stałą zwykle równa 1 lub określa się na podstawie wiedzy o danych podlegających grupowaniu.

32

Algorytm Gustafsona-Kessela

W wyniku minimalizacji kryterium K względem macierzy A_i :

$$K = \sum_{i=1}^{c}\sum_{k=1}^{p}\left(\mu_{ik}{}^{m}\right)\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_{A_i}{}^{2}$$

otrzymuje się:

$$\boldsymbol{A}_i = [\alpha_i \det(\boldsymbol{F}_i)]^{\frac{1}{n}}\boldsymbol{F}_i{}^{-1}.$$

gdzie $\mathbf{F}_i$ – jest macierzą kowariancji i-tej grupy:

$$\boldsymbol{F}_i = \frac{\sum_{k=1}^{p}(\mu_{ik})^m(\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{k=1}^{p}(\mu_{ik})^m}$$

Algorytm znajduje grupy o dowolnych kształtach.

33

Algorytm Gustafsona-Kessela

1. Inicjalizacja
   - wprowadzenie liczby grup *k*, rozmycia *m* oraz progu warunku zatrzymania alg. ε
   - przypisanie położeń poszczególnych grup **V** inicjalizowanych losowymi przykładami **X**, określenie macierzy **U** ∈ P_R
   - krok iteracji t = 1
2. Wyznaczenie grup dla poszczególnych przykładów
- wyznaczenie macierzy F w celu wyznaczenia odległości
- obliczenie przyporządkowania względem odległości:

$$\mu_{ik} = \left(\sum_{j=1}^{c}\left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_{A_i}}{\|\boldsymbol{x}_k - \boldsymbol{v}_j\|_{A_j}}\right)^{\frac{2}{m-1}}\right)^{-1}, \quad i = 1 \dots c, \quad k = 1 \dots p$$

3. Określenie położenia centroidów dla poszczególnych grup

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{p}\mu_{ik}{}^{m}\boldsymbol{x}_k}{\sum_{k=1}^{p}\mu_{ik}{}^{m}}$$

4. Sprawdzenie warunku zatrzymania algorytmu i ew. przejście do (2)

$$\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)} < \varepsilon$$

34

Wady algorytmu GK

Wady:
- wiecej obliczeń wobec alg. FCM

35

Kryteria oceny grupowania

Określenie poprawnej liczby grup jest istotne z punktu widzenia grupowania. Wskaźniki pomagają dokonać oceny czy znalezione grupy są poprawne:

- rozmycie w macierzy podziału
- wskaźnik Fukuyamy-Sugeno
- wskaźnik Xie-Bieni

36

Rozmycie w macierzy podziału Kryteria oceny grupowania

Stopień rozmycia macierzy podziału **U** przyporządkowania poszczególnych przykładów do grup dany jest wzorem:

$$K_1(\mathbf{U}) = \frac{1}{\mathrm{p}}\sum_{i=1}^{c}\sum_{k=1}^{p}(\mu_{ik})^2$$

Maksymalizacja kryterium – najlepszy podział na grupy to taki, w którym istnieje wyraźne przyporządkowanie do jednej grupy, czyli niepewność przydziału jest mała.

37

Rozmycie w macierzy podziału Kryteria oceny grupowania

Z kryterium K_1 związane jest również kryterium entropii podziału danych:

$$K_2(\mathbf{U}) = \frac{1}{\mathrm{p}}\sum_{i=1}^{c}\sum_{k=1}^{p}\mu_{ik}\ln\mu_{ik}$$

K_2 -> minimum

W przypadku gdy μ_{ik} -> 1/c to oznacza wysoki stopień rozmycia grup, wtedy K_2 przyjmuje duże wartości, co oznacza niewłaściwy podział na grupy.

W przypadku gdy stopnie przynależności są bliskie 0 lub 1 to K_2 przyjmuje małe wartości, czyli otrzymujemy poprawne grupy.

38

Wskaźnik Fukuyamy-Sugeno

Kryteria oceny grupowania

Związanie z kształtem grup zawartych w danych:

$$K_3(\mathbf{U}) = \frac{1}{\mathrm{p}}\sum_{i=1}^{c}\sum_{k=1}^{p}(\mu_{ik})^m\left(\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_A^2 - \|\boldsymbol{x}_k - \overline{\boldsymbol{v}}\|_A^2\right)$$

gdzie

$\bar{v}$ - jest średnią wartością po wszystkich danych:

$$\bar{v} = \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{x}_k$$

Kryterium daje najlepszy wynik przy minimalnej wartości.

39

Wskaźnik Xie-Bieni

Kryteria oceny grupowania

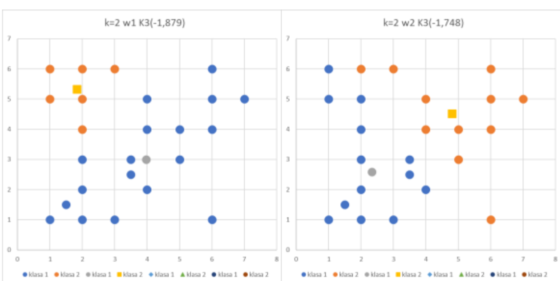
Związanie z kształtem grup zawartych w danych:

$$K_4(\mathbf{U}) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{p}(\mu_{ik})^m\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2}{p\, min_{ij}\{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|\}^2}$$

Najlepszy wynik - minimalizacja kryterium, gdzie c=2,...,p-1, czyli dążymy do podziału, w którym jest minimalna odległość między elementami a środkami grup oraz wszystkie środki będą od siebie jak najdalej.

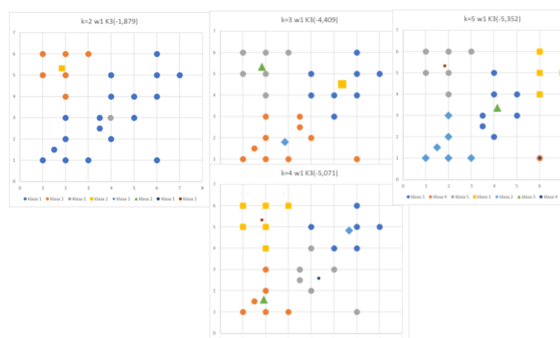
40

Przykłady działania alg. HCM i FCM



41

Porównanie dla k=2,3,4,5 w1(HCM)



42