Sztuczna inteligencja 9999
pages 1–7

MEODY GRUPOWANIA DANYCH

PB

1 CWICZENIE I

1. Ze zbioru danych iris.tab wybrać nastepujące obiekty:

ID	SL	SW	PL	PW	C
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
51	7.0	3.2	4.7	1.3	Iris-ver
52	6.3	3.2	4.5	1.5	Iris-ver
53	6.9	3.1	4.9	1.5	Iris-ver
100	5.7	2.8	4.1	1.3	Iris-vir
101	6.3	3.3	6.0	2.5	Iris-vir
102	5.8	2.7	5.1	1.9	Iris-vir

Nazwy atrybutów: SL- sepallength, SW - sepalwidth, PL- petallength, PW - petalwidth. C - klasyfikacja.

2. Wybrać do obliczeń dwa atrybuty - sepallength oraz petallength, obliczyć macierz podobieństwa między parami wszystkich 9 obiektów. Na tym etapie, mamy 9 jednoelementowych grup

Pobobieństwo między dwoma obiektami p_i oraz p_j, obliczamy jako odległość euklidesową:

$$d_{ij} = \sqrt{\sum_{a=1}^{N_a}(p_{ia} - p_{ja})^2}$$

gdzie N_a - liczba atrybutów, p_{ia} - wartość atrybutu o numerze a dla obiektu o indeksie i.

3. Wykonac grupowanie obiektow, poprzez połączenie dwu najbardziej podobnych obiektów w jedną grupę, w wyniku otrzymujemy 8 grup obiektów, jedną grupę dwuelementową, oraz 7 grup jednoelementowych.

4. Powtórzyć grupowanie obiektów do momentu otrzymania trzech grup obiektów (ew. jednej 9-elementowej grupy obiektów).

Miara podobieństwa obiektów /grup obiektów. W przypadku, gdy wykonujemy obliczenia odległości między dwiema grupami obiektów złożonych z k obiektów -

pierwsza grupa, oraz l obiektów - druga grupa. Wykonujemy obliczenie odległości między wszystkimi parami obiektów x_i, y_j, gdzie
x_i - i-ty obiekt z grupy I, $i = 1, ..., k$
y_j - j-ty obiekt z grupy II, $j = 1, ..., l$
oraz

$$d_{x_i,y_j} = \sqrt{\sum_{a=1}^{N_a}(p_{x_i a} - p_{y_j a})^2}$$

gdzie N_a - liczba atrybutów, $p_{x_i a}$ - wartość atrybutu o numerze a dla obiektu x_i z grupy pierwszej oraz $p_{x_i a}$ - wartość atrybutu o numerze a dla obiektu y_i grupy drugiej.

Metody łączenia grup (scalania)

1. Miara (ang. nearest first) - jako odległość dwu grup rozumiemy najmniejszą z odległości d_{x_i,y_j}, wykonujemy łączenie dwu grup, których odległość jest najmniejsza.
2. Miara (ang. farest first)- łączymy grupy, dla których odległość między dwoma najdalszymi obiektami jest najmniejsza.
3. Miara (ang. average first)- łączymy grupy, których średnia odległość jest najmniejsza,

Nadawanie wag poszczególnym atrybutom

5. Zastosować ważenie wartości atrybutów według schematu:
Niech
$min = min_{PL}$ - wartość minimalna dla petallength
$max = max_{PL}$ - wartość maksymalna dla petallength
oraz w podobny sposób:

$$min_{SL}, min_{SW}, min_{PW}, max_{SL}, max_{SW}, max_{PW},$$

$$d_{SL} = max_{SL} - min_{SL}$$

$$d_{SW} = max_{SW} - min_{SW}$$

$$d_{PL} = max_{PL} - min_{PL}$$

$$d_{PW} = max_{PW} - min_{PW}$$

Wagę dla danego atrybutu obliczamy jako iloraz zakresu danego atrybutu do maksymalnego zakresu atrybutów, przykładowo dla atrybutu petallength:

1.

$$W_{PL} = \frac{d_{PL}}{max(d_{PL,PW,SW,SL})}$$

2.

$$\frac{d_{PL}}{sum(d_{PL,PW,SW,SL})}$$

$$W_{PL} = \frac{1.0}{W_{PL}}$$

Wzór na odleglość między dwoma atrybutami przybiera postać:

$$d_{ij} = \sqrt{\sum_{a=1}^{N_a} W_a * (p_{ia} - p_{ja})^2}$$

2 CWICZENIE II

Powtórzyć zadanie dla atrybutów:

1. SW, PW
2. SW, PW, SL
3. SW, PW, SL, PL.

oraz obiektów:

1. 10, 15, 20, 60, 65, 70, 120, 125, 130.
2. 20, 25, 30, 85, 90, 95, 130, 135, 140.
3. 35, 40, 45, 70, 75, 80, 135, 140, 145.

3 CWICZENIE III

1. Uruchomić aplikację WEKA - Explorer
2. Wczytać dane iris.arff
3. Wybrać dwa atrybuty SL, PL
4. Przejść do zakladki Clusterer
5. Wybrać algorytm SimpleKMeans
6. Dwukrotnie wcisnąć przycik myszy na nazwie algorytmu
7. Wprowadzić liczbę grup-klas na jakie ma zostać podzielony badany zbiór - ustawić na trzy klasy.
8. Wykonać grupowanie (przycisk Start)
9. Zapamiętać środki klas

W arkuszu kalkulacyjnym przydzielić wszystkie obiekty do najbliższych środków klas.

Powtórzyć zadanie dla atrybutów:

1. SW, PW
2. SW, PW, SL

3. SW, PW, SL, PL.

oraz obiektów:

1. 10, 15, 20, 60, 65, 70, 120, 125, 130.
2. 20, 25, 30, 85, 90, 95, 130, 135, 140.
3. 35, 40, 45, 70, 75, 80, 135, 140, 145.

4 CWICZENIE IV

1. Uruchomić aplikację WEKA - Explorer
2. Wczytać dane iris.arff
3. Wybrać dwa atrybuty SL, PL
4. Przejść do zakladki Clusterer
5. Wybrać algorytm EM
6. Dwukrotnie wcisnąć przycik myszy na nazwie algorytmu
7. Wprowadzić liczbę grup-klas na jakie ma zostać podzielony badany zbiór - ustawić na trzy klasy.
8. Wykonać grupowanie (przycisk Start)
9. Zapamiętać środki klas

Powtórzyć zadanie dla atrybutów:

1. SW, PW
2. SW, PW, SL
3. SW, PW, SL, PL.

oraz obiektów:

1. 10, 15, 20, 60, 65, 70, 120, 125, 130.
2. 20, 25, 30, 85, 90, 95, 130, 135, 140.
3. 35, 40, 45, 70, 75, 80, 135, 140, 145.

5 CWICZENIE V

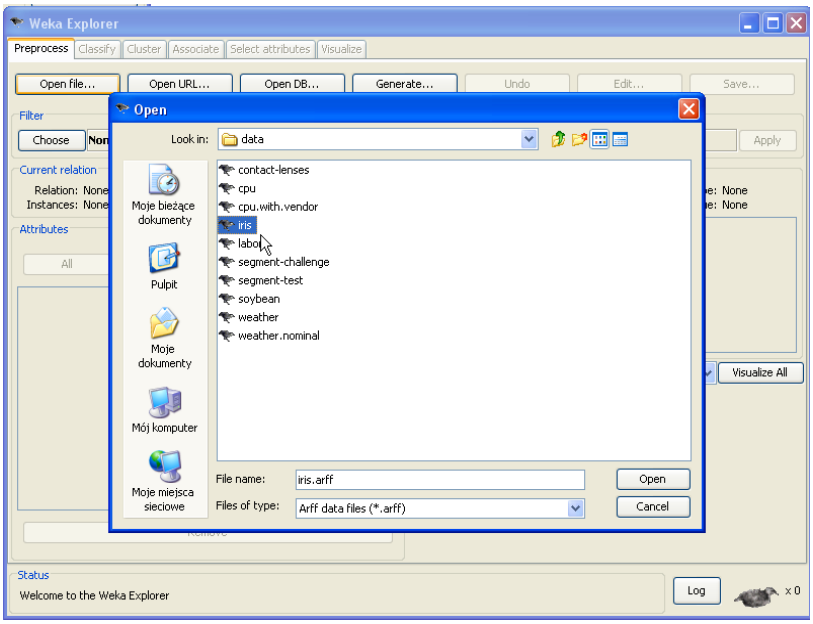
Wykonać polecenia z zadania pierwszego dla poniższych danych:
wybierając atrybuty: Powtórzyć zadanie I, III, IV dla atrybutów:

1. T, H
2. T, H, W (przypisując 1 - dla TRUE, 0 - dla FALSE)
3. T, H, O (przypisując 0 - rainy, 0.5 - overcast, 1 dla sunny)
4. T, H, W, O (przypisując 1 - dla TRUE, 0 - dla FALSE), (przypisując 0 - rainy, 0.5 - overcast, 1 dla sunny)

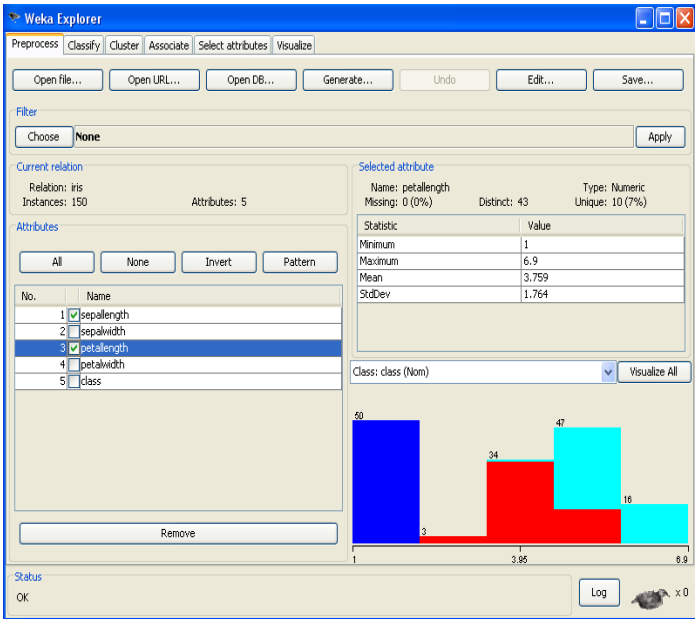
Literatura

1. Dokumentacja systemu WEKA.

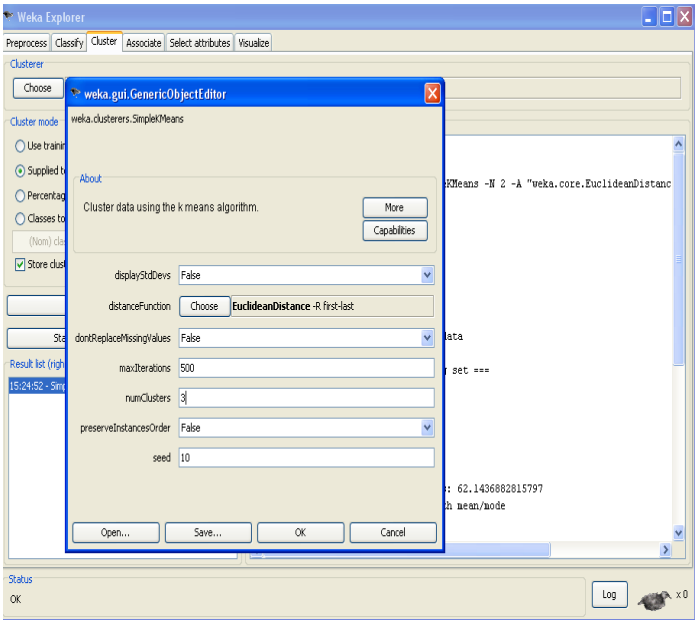
ID	O	T	H	W	P
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	rainy	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	sunny	81.0	75.0	FALSE	yes
14	sunny	71.0	91.0	TRUE	no



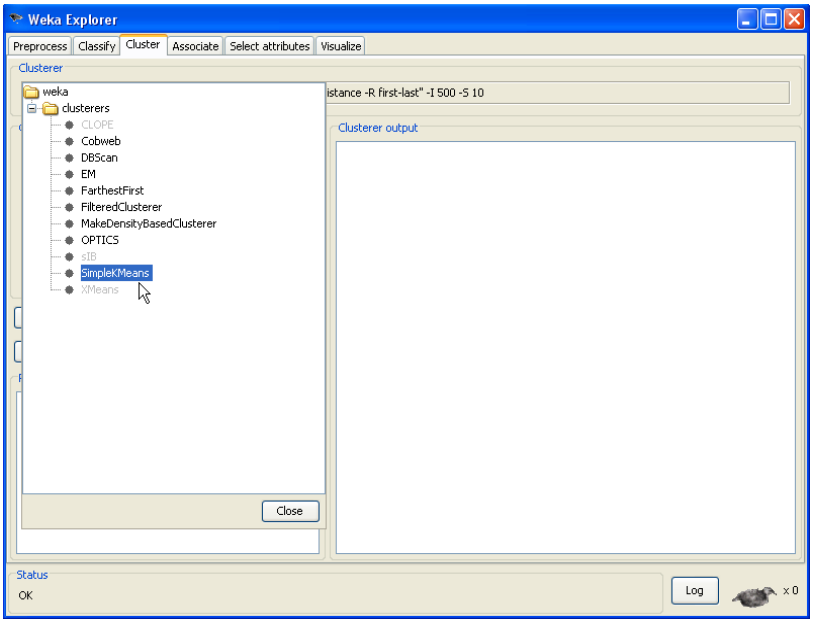
Rysunek 1:



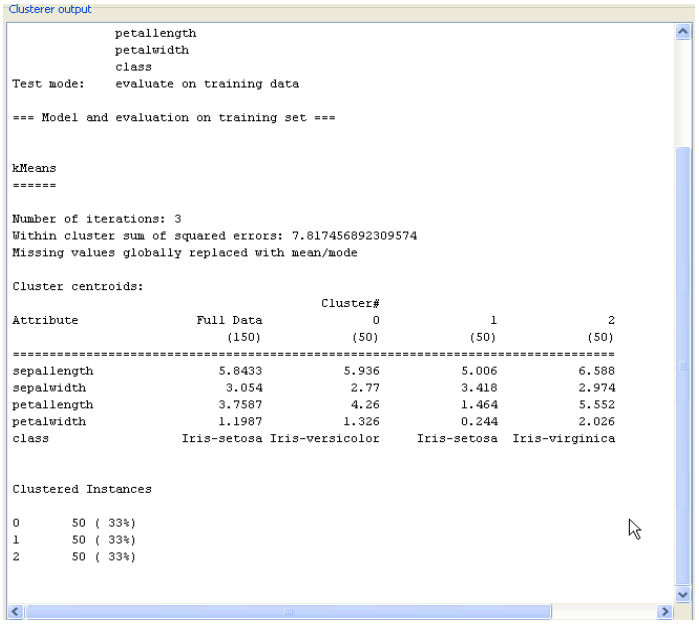
Rysunek 2:



Rysunek 3:



Rysunek 4:



Rysunek 5: