

Asocjacyjna reprezentacja danych i wnioskowanie

Metody Inżynierii Wiedzy

Patrycja Strugała

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
AGH University of Science and Technology

Kraków, 2016 r.

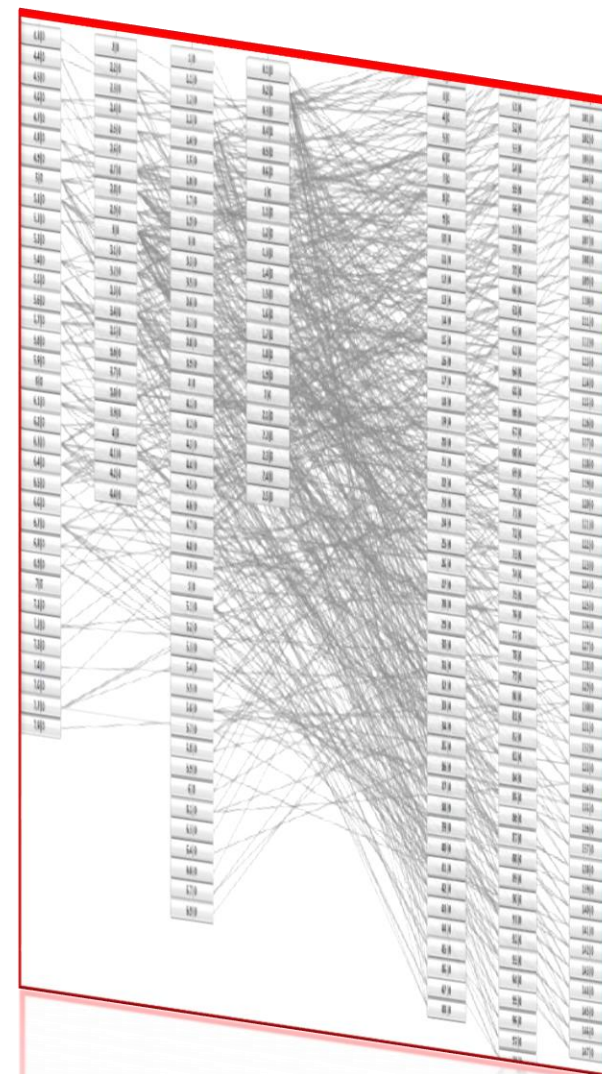
Cel projektu







	X	Y	month	day	FFMC	DMC	DC	ISI	temp
1	7	5	3	5	86.2	26.2	94.3	5.1	8.2
2	7	4	10	2	90.6	35.4	669.1	6.7	18
3	7	4	10	6	90.6	43.7	686.9	6.7	14.6
4	8	6	3	5	91.7	33.3	77.5	9	8.3
5	8	6	3	7	89.3	51.3	102.2	9.6	11.4
6	8	6	8	7	92.3	85.3	488	14.7	22.2
7	8	6	8	1	92.3	88.9	495.6	8.5	24.1
8	8	6	8	1	91.5	145.4	608.2	10.7	8
9	8	6	9	2	91	129.5	692.6	7	13.1
10	7	5	9	6	92.5	88	698.6	7.1	22.8
11	7	5	9	6	92.5	88	698.6	7.1	17.8
12	7	5	9	6	92.8	73.2	713	22.6	19.3
13	6	5	8	5	63.5	70.8	665.3	0.8	17
14	6	5	9	1	90.9	126.5	686.5	7	21.3
15	6	5	9	3	92.9	133.3	699.6	9.2	26.4
16	6	5	9	5	93.3	141.2	713.9	13.9	22.9
17	5	5	3	6	91.7	35.8	80.8	7.8	15.1
18	8	5	10	1	84.9	32.8	664.2	3	16.7
19	6	4	3	3	89.2	27.9	70.8	6.3	15.9



int / double / float

Wykorzystane narzędzia programistyczne



Interfejs użytkownika

AGDS

File

D:/Projekt AGDS/build-AGDS-Desktop_Qt_5_5_0_MSVC2013_32bit-Debug/DataSe

	X	Y	month	day	FFMC	DMC	DC	ISI	temp
1	7	5	3	5	86.2	26.2	94.3	5.1	8.2
2	7	4	10	2	90.6	35.4	669.1	6.7	18
3	7	4	10	6	90.6	43.7	686.9	6.7	14.6
4	8	6	3	5	91.7	33.3	77.5	9	8.3
5	8	6	3	7	89.3	51.3	102.2	9.6	11.4
6	8	6	8	7	92.3	85.3	488	14.7	22.2
7	8	6	8	1	92.3	88.9	495.6	8.5	24.1
8	8	6	8	1	91.5	145.4	608.2	10.7	8
9	8	6	9	2	91	129.5	692.6	7	13.1
10	7	5	9	6	92.5	88	698.6	7.1	22.8
11	7	5	9	6	92.5	88	698.6	7.1	17.8
12	7	5	9	6	92.8	73.2	713	22.6	19.3
13	6	5	8	5	63.5	70.8	665.3	0.8	17
14	6	5	9	1	90.9	126.5	686.5	7	21.3
15	6	5	9	3	92.9	133.3	699.6	9.2	26.4
16	6	5	9	5	93.3	141.2	713.9	13.9	22.9
17	5	5	3	6	91.7	35.8	80.8	7.8	15.1
18	8	5	10	1	84.9	32.8	664.2	3	16.7
19	6	4	3	3	89.2	27.9	70.8	6.3	15.9
20	6	4	4	6	86.2	27.4	87.1	5.1	8.2

Functions

- ☒ Find 'k' similar objects to existing one id: 50 neighbourhood: 70 k: 10
- ☐ Find 'k' similar objects to: Enter values of attributes (comma separated) n: 1 k: 1
- ☐ Find 'k' similar object to group: Enter ids of existing objects (comma separated) n: 1 k: 1
- ☐ Find all objects with attribute: X value: value: X value: value:
- ☐ Find objects for attribute value in range: X value: value: -
- ☐ Find 'k' minimum objects for attribute: X k: 1
- ☐ Find 'k' maximum objects for attribute: X k: 1

Run function

Results

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
1	4	4	3	1	87.6	52.2	103.8	5	11	46	5.8	0	0
2	5	4	3	1	90.1	39.7	86.6	6.2	13.2	40	5.4	0	0.95
3	3	4	3	1	90.1	39.7	86.6	6.2	10.6	30	4	0	0
4	4	4	3	1	87.2	23.9	64.7	4.1	11.8	35	1.8	0	0

Time Comparison

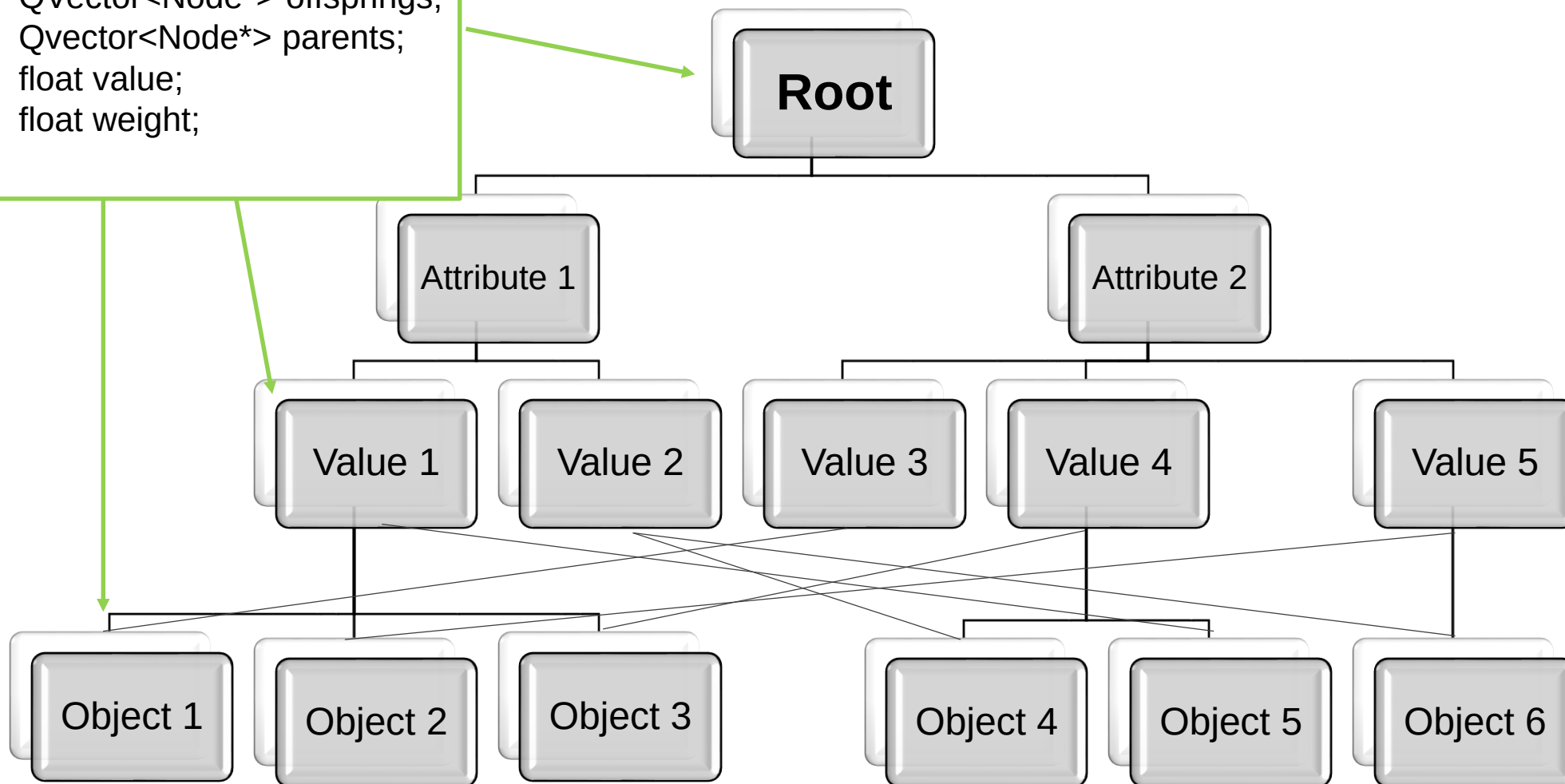
Graph searching 1 ms Table searching 115 ms

AGDS Graph created!



Struktura grafu asocjacyjnego

```
class Node {
    QVector<Node*> offsprings;
    QVector<Node*> parents;
    float value;
    float weight;
}
```



Zaimplementowane funkcje

1. Znalezienie obiektów podobnych
2. Zależenie obiektów o danych atrybutach (skorelowanych / sąsiednich)
3. Znalezienie minimów i maksimów

Functions

☒ Find 'k' similar objects to existing one id: neighbourhood: k:

☐ Find 'k' similar objects to: n%: k:

☐ Find 'k' similar object to group: n%: k:

☐ Find all objects with attribute: value: value:

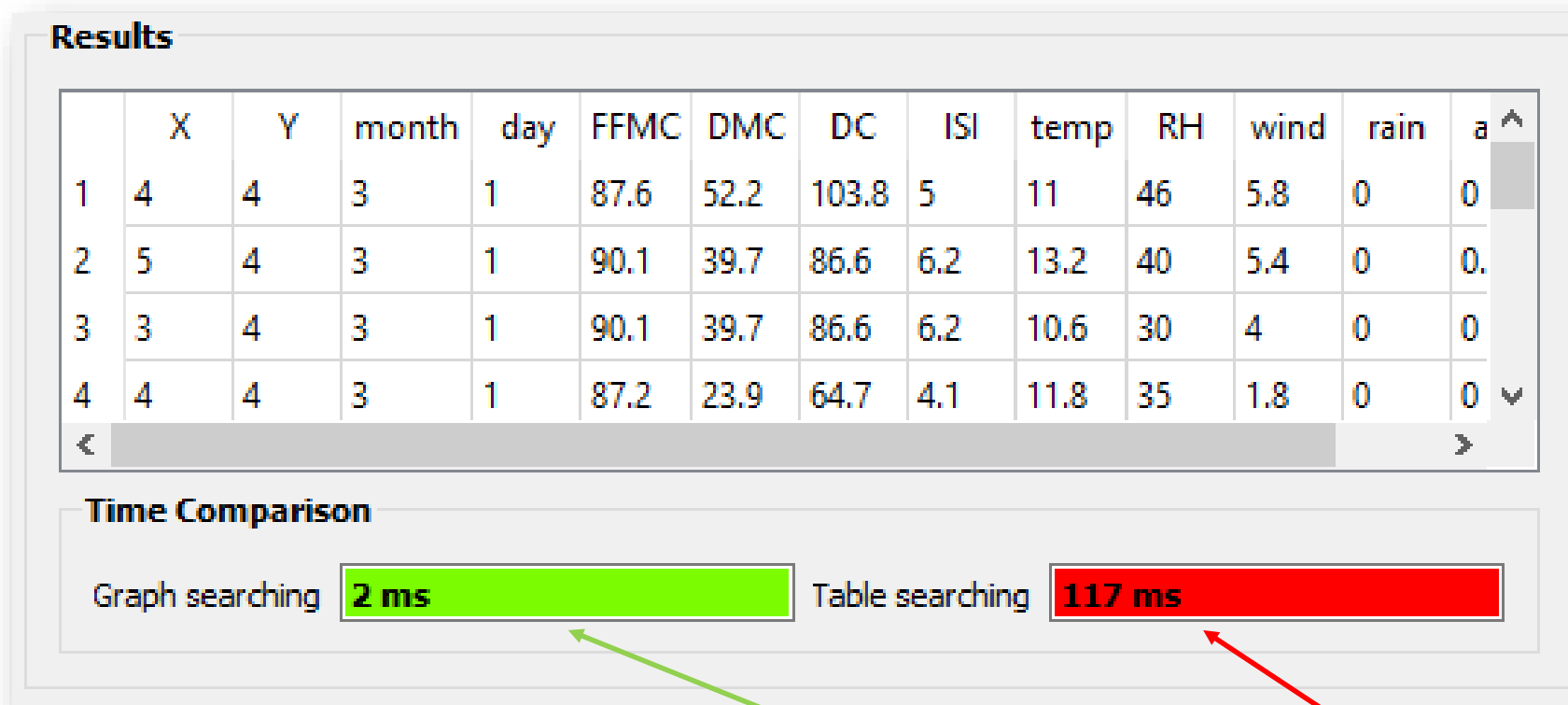
☐ Find objects for attribute value in range: value: -

☐ Find 'k' minimum objects for attribute: k:

☐ Find 'k' maximum objects for attribute: k:

Run function

Rezultaty



#include <QDateTime>

AGDS

File

D:/Projekt AGDS/DataSets/Forest Fires/forestfires.csv

	X	Y	month	day	FFMC	DMC	DC	ISI	temp
1	7	5	3	5	86.2	26.2	94.3	5.1	8.2
2	7	4	10	2	90.6	35.4	669.1	6.7	18
3	7	4	10	6	90.6	43.7	686.9	6.7	14.6
4	8	6	3	5	91.7	33.3	77.5	9	8.3
5	8	6	3	7	89.3	51.3	102.2	9.6	11.4
6	8	6	8	7	92.3	85.3	488	14.7	22.2
7	8	6	8	1	92.3	88.9	495.6	8.5	24.1
8	8	6	8	1	91.5	145.4	608.2	10.7	8
9	8	6	9	2	91	129.5	692.6	7	13.1
10	7	5	9	6	92.5	88	698.6	7.1	22.8
11	7	5	9	6	92.5	88	698.6	7.1	17.8
12	7	5	9	6	92.8	73.2	713	22.6	19.3
13	6	5	8	5	63.5	70.8	665.3	0.8	17
14	6	5	9	1	90.9	126.5	686.5	7	21.3
15	6	5	9	3	92.9	133.3	699.6	9.2	26.4
16	6	5	9	5	93.3	141.2	713.9	13.9	22.9
17	5	5	3	6	91.7	35.8	80.8	7.8	15.1
18	8	5	10	1	84.9	32.8	664.2	3	16.7
19	6	4	3	3	89.2	27.9	70.8	6.3	15.9
20	6	4	4	6	86.2	27.4	97.1	5.1	8.2

Functions

☐ Find 'k' similar objects to existing one id: neighbourhood: k:

☒ Find 'k' similar objects to: n%: k:

☐ Find 'k' similar object to group: n%: k:

☐ Find all objects with attribute: value: value:

☐ Find objects for attribute value in range: value: -

☐ Find 'k' minimum objects for attribute: k:

☐ Find 'k' maximum objects for attribute: k:

Results

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
1	6	5	8	5	63.5	70.8	665.3	0.8	17	72	6.7	0	0
2	6	5	8	6	90.2	96.9	624.2	8.9	14.7	59	5.8	0	0
3	7	5	9	6	92.5	88	698.6	7.1	17.8	51	7.2	0	0
4	7	4	10	5	90	41.5	682.6	8.7	11.3	60	5.4	0	0

Time Comparison

Graph searching Table searching

AGDS Graph created!

AGDS

File

D:/Projekt AGDS/DataSets/Forest Fires/forestfires.csv

	X	Y	month	day	FFMC	DMC	DC	ISI	temp
1	7	5	3	5	86.2	26.2	94.3	5.1	8.2
2	7	4	10	2	90.6	35.4	669.1	6.7	18
3	7	4	10	6	90.6	43.7	686.9	6.7	14.6
4	8	6	3	5	91.7	33.3	77.5	9	8.3
5	8	6	3	7	89.3	51.3	102.2	9.6	11.4
6	8	6	8	7	92.3	85.3	488	14.7	22.2
7	8	6	8	1	92.3	88.9	495.6	8.5	24.1
8	8	6	8	1	91.5	145.4	608.2	10.7	8
9	8	6	9	2	91	129.5	692.6	7	13.1
10	7	5	9	6	92.5	88	698.6	7.1	22.8
11	7	5	9	6	92.5	88	698.6	7.1	17.8
12	7	5	9	6	92.8	73.2	713	22.6	19.3
13	6	5	8	5	63.5	70.8	665.3	0.8	17
14	6	5	9	1	90.9	126.5	686.5	7	21.3
15	6	5	9	3	92.9	133.3	699.6	9.2	26.4
16	6	5	9	5	93.3	141.2	713.9	13.9	22.9
17	5	5	3	6	91.7	35.8	80.8	7.8	15.1
18	8	5	10	1	84.9	32.8	664.2	3	16.7
19	6	4	3	3	89.2	27.9	70.8	6.3	15.9
20	6	4	4	6	86.2	27.4	87.1	5.1	8.2

Functions

☐ Find 'k' similar objects to existing one id: neighbourhood: k:

☐ Find 'k' similar objects to: n%: k:

☐ Find 'k' similar object to group: n%: k:

☐ Find all objects with attribute: value: value:

☒ Find objects for attribute value in range: value: -

☐ Find 'k' minimum objects for attribute: k:

☐ Find 'k' maximum objects for attribute: k:

Run function

Results

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	are
1	7	5	9	2	91	129.5	692.6	7	20.7	37	2.2	0	17.2
2	1	3	9	7	91	276.3	825.1	7.1	21.9	43	4	0	70.7
3	5	4	4	7	91	14.6	25.6	12.3	17.6	27	5.8	0	0
4	6	5	8	5	91	166.9	752.6	7.1	18.2	62	5.4	0	0.43

Time Comparison

Graph searching
Table searching

AGDS Graph created!

Testy – małe zbiory

<i>Wielkość zbioru</i>	<i>Funkcja</i>	<i>Sąsiedztwo</i>	<i>Czas przeszukiwania w grafie</i>	<i>Czas przeszukiwania w tabeli</i>
517	Znajdź podobne do obiektu o id	200	0 ms	68 ms
517	Znajdź podobne do grupy obiektów	80 %	0 ms	45 ms
517	Znajdź podobne do grupy obiektów	100 %	0 ms	99 ms
517	Znajdź minimum	Cały zbiór	0 ms	0 ms
4 898	Znajdź podobne do obiektu o id	1000	1 ms	267 ms
4 898	Znajdź podobne do grupy obiektów	40 %	0 ms	17 ms
4 898	Znajdź w zakresie	Cały zbiór	0 ms	2 ms
4 898	Znajdź maksimum	Cały zbiór	0 ms	2 ms

Testy – duże zbiory

<i>Wielkość zbioru</i>	<i>Funkcja</i>	<i>Sąsiedztwo</i>	<i>Czas przeszukiwania w grafie</i>	<i>Czas przeszukiwania w tabeli</i>
25 010	Znajdź podobne do obiektu o id	Cały zbiór	4 ms	213001 ms
25 010	Znajdź podobne do obiektu o id	1000	4 ms	19418 ms
25 010	Znajdź podobne do obiektu o id	100	4 ms	2022 ms
25 010	Znajdź minimum	Cały zbiór	0 ms	3 ms
49 414	Znajdź podobne do obiektu o id	5000	9 ms	124912 ms
49 414	Znajdź w zakresie	Cały zbiór	0 ms	16 ms
49 414	Znajdź maksimum	Cały zbiór	0 ms	11 ms
49 414	Znajdź podobne do grupy obiektów	20 %	3 ms	10815 ms

Wnioski

Asocjacyjna grafowa struktura danych AGDS umożliwia:

- Horyzontalne i wertykalne powiązanie danych
- Przechowywanie danych w posortowanej strukturze (równocześnie dla różnych atrybutów)
- Usunięcie duplikatów, redundancji i nadmiarowości
- Łatwiejszy i szybszy dostęp do danych szczególnie dla BIG DATA
- Zdecydowanie szybsze przetwarzanie danych:
 - wyszukanie obiektów wg podobieństwa atrybutów bądź grup atrybutów
 - znalezienie obiektów skorelowanych
 - znalezienie obiektów sąsiednich
 - wyszukanie minimów i maksimów

Metoda K – najbliższych sąsiadów

- » Wczytanie danych zbioru uczącego (Iris)
- » Wprowadzenie testowego obiektu
- » Wyznaczenie odległości Euklidesowych
- » W obszarze k – najbliższych sąsiadów sklasyfikowanie obiektu testowego do odpowiedniej klasy w zależności od k

```
Enter:
Leaf length: 2.4
Leaf width: 1.1
Petal length: 1.8
Petal Width: 0.5
Amount of neighbors: 10
Distance: 2.4779Class: Iris-setosa
Distance: 2.73679Class: Iris-setosa
Distance: 2.80535Class: Iris-setosa
Distance: 2.81957Class: Iris-setosa
Distance: 2.95804Class: Iris-setosa
Distance: 3.00333Class: Iris-setosa
Distance: 3.08221Class: Iris-setosa
Distance: 3.09354Class: Iris-setosa
Distance: 3.11288Class: Iris-setosa
Distance: 3.13528Class: Iris-setosa
Distance: 3.14484Class: Iris-setosa
Distance: 3.1686Class: Iris-setosa
Distance: 3.17962Class: Iris-setosa
Distance: 3.21403Class: Iris-setosa
Distance: 3.2311Class: Iris-versicolor
Distance: 3.24037Class: Iris-setosa
Distance: 3.24037Class: Iris-setosa
Distance: 3.24037Class: Iris-setosa
Distance: 3.24037Class: Iris-setosa
Distance: 3.27109Class: Iris-versicolor
Distance: 3.27261Class: Iris-versicolor
Distance: 7.71946Class: Iris-virginica
Distance: 7.80795Class: Iris-virginica
Distance: 7.89177Class: Iris-virginica
Iris-setosa: 10
Iris-versicolor: 0
Iris-virginica: 0
```

Eksploracja danych

- » Wprowadzenie przykładowego zbioru danych
- » Przekonwertowanie danych z horyzontalnego do wertykalnego formatu danych
- » Możliwość obliczenia wsparcia
- » Możliwość określenia progu dla zbiorów częstych i usunięcie zbiorów rzadkich
- » Określenie reguł asocjacyjnych dla zbiorów częstych bądź wszystkich zbiorów

```
MENU:
1 - Oblicz wsparcie; 2 - Okresl prog dla czestych zbiorow i usun zbiory rzadkie;
3 - Reguly asocjacyjne dla zbiorow czestych; 4 - Reguly asocjacyjne dla wszystkich zbiorow;
5 - Wswietl wszystkie zbiory; q - exit
1
kawa    ->    44.4444 %
mleko    ->    44.4444 %
cukier    ->    44.4444 %
orzechy  ->    33.3333 %
jajka    ->    44.4444 %
chleb    ->    44.4444 %
maslo    ->    33.3333 %
miod     ->    22.2222 %
platki   ->    22.2222 %
ser       ->    33.3333 %

mleko    ->    kawa (78%, 25%)
mleko    ->    cukier (78%, 25%)
mleko    ->    orzechy (67%, 33%)
mleko    ->    jajka (78%, 25%)
mleko    ->    chleb (78%, 25%)
mleko    ->    maslo (67%, 33%)
mleko    ->    miod (56%, 50%)
mleko    ->    platki (56%, 50%)
mleko    ->    ser (67%, 33%)
```

Walidacja krzyżowa

- » Wczytanie zbioru danych (Iris)
- » Wyznaczenie zbioru uczącego i testowego (15 losowych obiektów)
- » Procent poprawnie rozpoznanych klas w zależności od k

```
Procentowe rozpoznanie poprawnej klasy obiektu w zaleznosci licznosci zbioru 'k':  
k = 135 ==> 53.3333 %  
k = 134 ==> 53.3333 %  
k = 133 ==> 53.3333 %  
k = 132 ==> 53.3333 %  
k = 131 ==> 53.3333 %  
k = 130 ==> 53.3333 %  
k = 129 ==> 53.3333 %  
k = 128 ==> 53.3333 %  
k = 127 ==> 53.3333 %  
k = 126 ==> 53.3333 %  
k = 125 ==> 53.3333 %  
k = 124 ==> 53.3333 %  
k = 123 ==> 53.3333 %  
k = 122 ==> 53.3333 %  
k = 121 ==> 53.3333 %  
k = 120 ==> 53.3333 %  
k = 119 ==> 53.3333 %  
k = 118 ==> 53.3333 %  
k = 117 ==> 53.3333 %  
k = 116 ==> 53.3333 %  
k = 115 ==> 53.3333 %  
k = 114 ==> 53.3333 %  
k = 113 ==> 53.3333 %  
k = 112 ==> 53.3333 %  
k = 111 ==> 53.3333 %  
k = 110 ==> 53.3333 %  
k = 109 ==> 53.3333 %  
k = 108 ==> 53.3333 %  
k = 107 ==> 53.3333 %  
k = 106 ==> 53.3333 %  
k = 105 ==> 53.3333 %  
k = 104 ==> 53.3333 %  
k = 103 ==> 53.3333 %  
k = 102 ==> 53.3333 %  
k = 101 ==> 53.3333 %  
k = 100 ==> 53.3333 %  
k = 99 ==> 53.3333 %  
k = 98 ==> 53.3333 %  
k = 97 ==> 53.3333 %  
k = 96 ==> 53.3333 %  
k = 95 ==> 53.3333 %  
k = 94 ==> 53.3333 %  
k = 93 ==> 53.3333 %  
k = 92 ==> 53.3333 %  
k = 91 ==> 53.3333 %  
k = 90 ==> 53.3333 %  
k = 89 ==> 53.3333 %  
k = 88 ==> 53.3333 %  
k = 87 ==> 53.3333 %  
k = 86 ==> 53.3333 %  
k = 85 ==> 53.3333 %  
k = 84 ==> 53.3333 %  
k = 83 ==> 53.3333 %  
k = 82 ==> 53.3333 %  
k = 81 ==> 53.3333 %  
k = 80 ==> 53.3333 %  
k = 79 ==> 53.3333 %  
k = 78 ==> 53.3333 %  
k = 77 ==> 53.3333 %  
k = 76 ==> 53.3333 %  
k = 75 ==> 53.3333 %  
k = 74 ==> 53.3333 %  
k = 73 ==> 53.3333 %  
k = 72 ==> 53.3333 %  
k = 71 ==> 53.3333 %  
k = 70 ==> 53.3333 %  
k = 69 ==> 53.3333 %  
k = 68 ==> 53.3333 %  
k = 67 ==> 53.3333 %  
k = 66 ==> 53.3333 %  
k = 65 ==> 53.3333 %  
k = 64 ==> 53.3333 %  
k = 63 ==> 53.3333 %  
k = 62 ==> 53.3333 %  
k = 61 ==> 53.3333 %  
k = 60 ==> 53.3333 %  
k = 59 ==> 53.3333 %  
k = 58 ==> 53.3333 %  
k = 57 ==> 53.3333 %  
k = 56 ==> 53.3333 %  
k = 55 ==> 53.3333 %  
k = 54 ==> 53.3333 %  
k = 53 ==> 53.3333 %  
k = 52 ==> 53.3333 %  
k = 51 ==> 53.3333 %  
k = 50 ==> 53.3333 %  
k = 49 ==> 53.3333 %  
k = 48 ==> 53.3333 %  
k = 47 ==> 53.3333 %  
k = 46 ==> 53.3333 %  
k = 45 ==> 53.3333 %  
k = 44 ==> 53.3333 %  
k = 43 ==> 53.3333 %  
k = 42 ==> 53.3333 %  
k = 41 ==> 53.3333 %  
k = 40 ==> 53.3333 %  
k = 39 ==> 53.3333 %  
k = 38 ==> 53.3333 %  
k = 37 ==> 53.3333 %  
k = 36 ==> 53.3333 %  
k = 35 ==> 53.3333 %  
k = 34 ==> 53.3333 %  
k = 33 ==> 53.3333 %  
k = 32 ==> 53.3333 %  
k = 31 ==> 53.3333 %  
k = 30 ==> 53.3333 %  
k = 29 ==> 53.3333 %  
k = 28 ==> 53.3333 %  
k = 27 ==> 53.3333 %  
k = 26 ==> 53.3333 %  
k = 25 ==> 53.3333 %  
k = 24 ==> 53.3333 %  
k = 23 ==> 53.3333 %  
k = 22 ==> 53.3333 %  
k = 21 ==> 53.3333 %  
k = 20 ==> 53.3333 %  
k = 19 ==> 53.3333 %  
k = 18 ==> 53.3333 %  
k = 17 ==> 53.3333 %  
k = 16 ==> 53.3333 %  
k = 15 ==> 53.3333 %  
k = 14 ==> 53.3333 %  
k = 13 ==> 53.3333 %  
k = 12 ==> 53.3333 %  
k = 11 ==> 53.3333 %  
k = 10 ==> 53.3333 %  
k = 9 ==> 53.3333 %  
k = 8 ==> 53.3333 %  
k = 7 ==> 53.3333 %  
k = 6 ==> 53.3333 %  
k = 5 ==> 53.3333 %  
k = 4 ==> 53.3333 %  
k = 3 ==> 53.3333 %  
k = 2 ==> 53.3333 %  
k = 1 ==> 53.3333 %
```


AGDS z bazą danych

- » Wprowadzenie przykładowej bazy danych studentów
- » Stworzenie grafu asocjacyjnego dla bazy danych
- » Przeszukania grafu wg wprowadzonych atrybutów

```
valueOfAttribute.push_back("Krakow"); // miasto  
valueOfAttribute.push_back("Reymonta"); // ulica  
valueOfAttribute.push_back(""); // numer mieszkania  
valueOfAttribute.push_back("Automatyka i Robotyka"); // kierunek studiow  
valueOfAttribute.push_back(""); // imie  
valueOfAttribute.push_back(""); // nazwisko  
valueOfAttribute.push_back(""); // stypendium  
valueOfAttribute.push_back(""); // rok studiow  
attributes = 3; // ilosc zgodnych elementow
```

```
Katarzyna Sodej  
Stypendium: 800 Rok: 2  
Automatyka i Robotyka  
Krakow Reymonta 18  
  
Anna Wybicka  
Stypendium: 1200 Rok: 4  
Automatyka i Robotyka  
Krakow Reymonta 18
```

Dziękuję za uwagę 😊