Metody Inżynierii Wiedzy

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

Mateusz Burcon

Kraków, czerwiec 2017

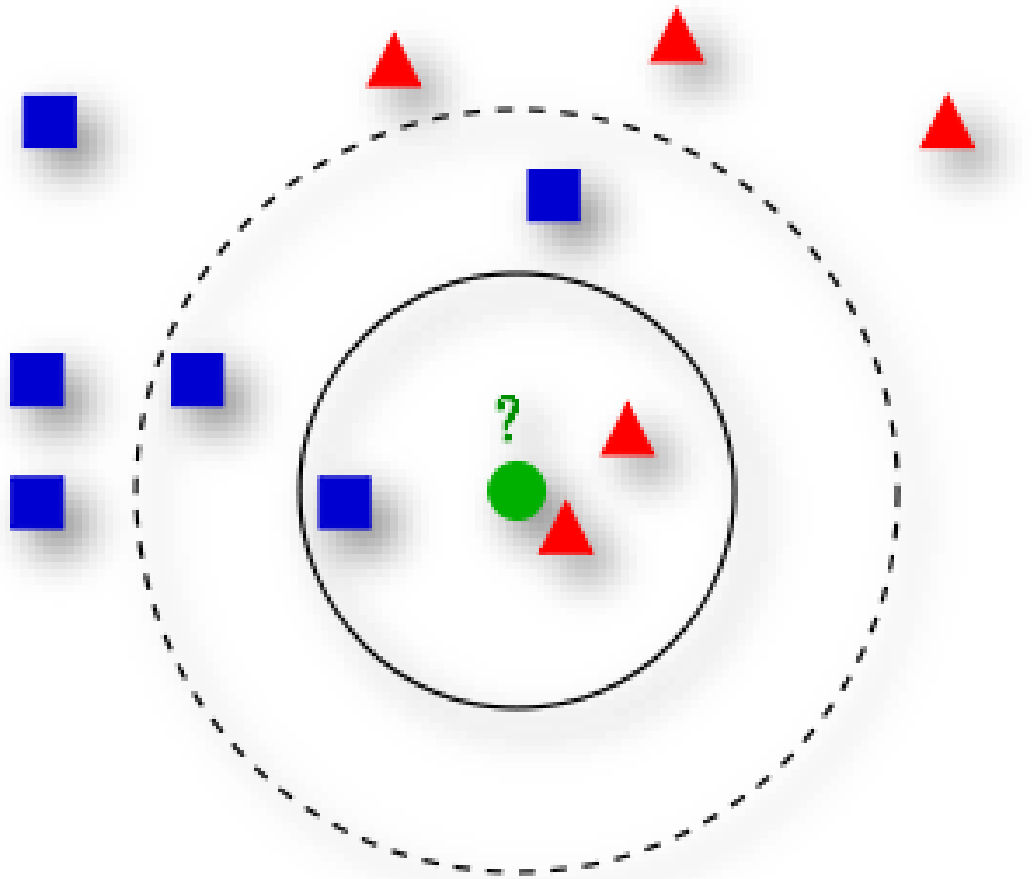
Wykorzystane technologie

Python 3.4

PyCharm 2016.3.2

Metoda K Najbliższych Sąsiadów

Metoda k Najbliższych Sąsiadów wyznacza k sąsiadów do których badany element ma najbliżej dla wybranej metryki, a następnie wyznacza wynik w oparciu o głos większości.



Metoda K Najbliższych Sąsiadów

```
Parametr K: 2   Poprawność proste KNN: 94.67.   Poprawność zmodyfikowane KNN: 96.0
Parametr K: 3   Poprawność proste KNN: 96.0.    Poprawność zmodyfikowane KNN: 96.0
Parametr K: 4   Poprawność proste KNN: 96.0.    Poprawność zmodyfikowane KNN: 96.0
Parametr K: 5   Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 6   Poprawność proste KNN: 96.0.    Poprawność zmodyfikowane KNN: 96.0
Parametr K: 7   Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 8   Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 9   Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 10  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 96.0
Parametr K: 11  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 12  Poprawność proste KNN: 96.0.    Poprawność zmodyfikowane KNN: 96.67
Parametr K: 13  Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 14  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 15  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 16  Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 17  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 18  Poprawność proste KNN: 97.33.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 19  Poprawność proste KNN: 98.0.    Poprawność zmodyfikowane KNN: 98.0
Parametr K: 20  Poprawność proste KNN: 98.0.    Poprawność zmodyfikowane KNN: 97.33
Parametr K: 21  Poprawność proste KNN: 98.0.    Poprawność zmodyfikowane KNN: 98.0
Parametr K: 22  Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 97.33
Parametr K: 23  Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
Parametr K: 24  Poprawność proste KNN: 96.67.   Poprawność zmodyfikowane KNN: 96.67
```

⇦ Wyjście programu

Najlepszy rezultat dla K=19:

- Poprawność niezmodyfikowanej metody Knn - 98%
- Poprawność zmodyfikowanej metody Knn - 98%

Metoda K Najbliższych Sąsiadów

```
Wprowadź parametry nowego obiektu:
Leaf length:
2.3
Leaf width:
4.2
Petal length:
1.7
Petal width:
1.2
Ilość sąsiadów:
12
Najlepsze dopasowania:
Iris-setosa:    12
Iris-versicolor:    0
Iris-virginica: 0
```

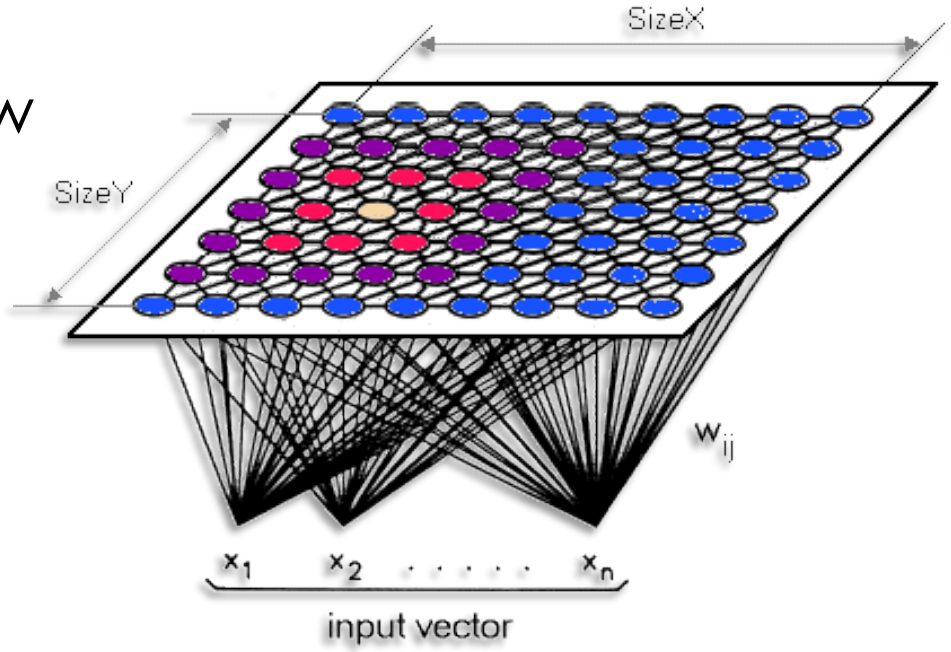
Możliwość zdefiniowania **własnego** wzorca testowego

Zwycięzca:

Iris-setosa

Neuronowe mapy samoorganizujące się

- Umożliwiają reprezentację wielowymiarowych danych w przestrzeni o mniejszym wymiarze
- Przykład sieci neuropodobnych uczony metodą uczenia nienadzorowanego

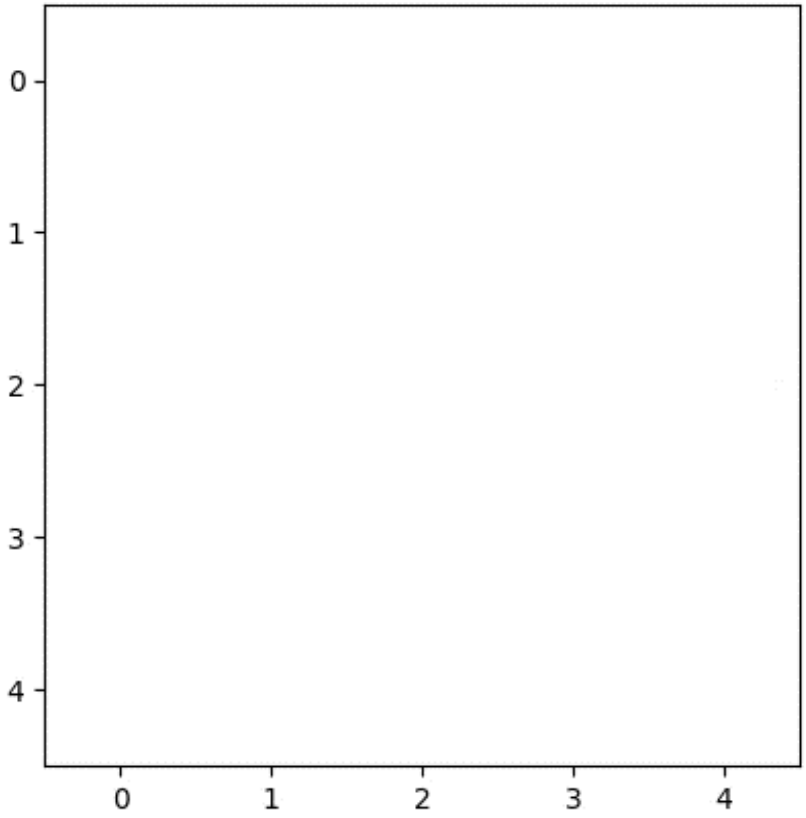


Neuronowe mapy samoorganizujące się

1. Inicjalizacja wag każdego pseudoneuronu niewielkimi liczbami losowymi.
2. Obliczanie wartości wyjściowej pseudoneuronów względem wektora wejściowego.
3. Aktualizacja wag pseudoneuronu najmocniej pobudzonego.
4. Aktualizacja wag sąsiednich pseudoneuronów w podobny sposób, ale z malejącym współczynnikiem.

$$d(X_k, W_{i,j}(t) = \sqrt{\sum_{i=1}^{I} \sum_{j=1}^{J} \left(X_k - W_{i,j}(t)\right)^2}$$

Neuronowe mapy samoorganizujące się



Eksploracja Danych

Proces automatycznego odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie reguł, zależności, wzorców schematów, podobieństw lub trendów w dużych repozytoriach danych.

Celem eksploracji danych jest analiza danych i procesów w celu lepszego ich rozumienia.

Eksploracja Danych

- Wsparcie s to ilość transakcji zawierających zarówno X jak i Y
- Pewność c to prawdopodobieństwo warunkowe, że transakcja zawierająca X zawiera również Y
- Wzorzec częsty, to taki, który ma wsparcie większe niż ustalony próg
- Wzorzec zamknięty, to taki wzorzec, który jest częsty i nie istnieje żaden nadwzorzec, który miałby takie samo wsparcie
- Wzorzec maksymalny, to taki wzorzec, który jest częsty i nie istnieje żaden częsty nadwzorzec

Eksploracja Danych

» Przeszukiwanie zbioru Iris

```
Wzorce częste dla progu 12%:
    Wartość:(1.4,)
        Wsparcie:13.33
        Numery transakcji:[0, 1, 4, 6, 8, 12, 17, 28, 33, 45, 47, 49, 50, 59, 63, 65, 75, 76, 91, 134]
    Wartość:(0.2,)
        Wsparcie:18.67
        Numery transakcji:[0, 1, 2, 3, 4, 7, 8, 10, 11, 14, 20, 22, 24, 25, 27, 28, 29, 30, 33, 35, 36, 38, 39, 42, 46, 47, 48, 49]
    Wartość:(3.0,)
        Wsparcie:18.0
        Numery transakcji:[1, 12, 13, 25, 38, 45, 61, 66, 75, 77, 84, 88, 91, 95, 98, 102, 104, 105, 112, 116, 127, 129, 135, 138, 145, 147, 149]
    Wartość:(1.3,)
        Wsparcie:13.33
        Numery transakcji:[2, 16, 36, 38, 40, 41, 42, 53, 55, 58, 64, 71, 74, 87, 88, 89, 94, 96, 97, 99]
    Wartość:(1.5,)
        Wsparcie:17.33
        Numery transakcji:[3, 7, 9, 10, 15, 19, 21, 27, 31, 32, 34, 37, 39, 48, 51, 52, 54, 61, 66, 68, 72, 78, 84, 86, 119, 133]
-----------------------------------
Wzorce zamknięte dla progu 12%:
{(1.5,), (0.2,), (3.0,), (1.4,), (1.3,)}
-----------------------------------
Wzorce maksymalne dla progu 12%:
{(1.5,), (0.2,), (3.0,), (1.4,), (1.3,)}
-----------------------------------
```

Eksploracja Danych

» Przeszukiwanie Baśni Grimm

```
Wzorce częste dla progu 40%:
dict_keys(['and', ('and', 'to'), ('the', 'and', 'to'), 'a', 'the', 'he', ('the', 'and'), ('the', 'to'), 'to'])

Ilość wzorców 9
----------------------------------
Wzorce maksymalne dla progu 40%:
{'a', 'he', ('the', 'and', 'to')}

Ilość wzorców 3
----------------------------------
```

Oznacza to, że 40% zdań w ‚Baśniach Grimm' zawiera kombinacje słów takich jak ‚a' lub ‚and' lub ‚the' i ‚and' lub ‚to' itd.

Eksploracja Danych

» Przeszukiwanie zbioru transakcji

```
Wzorce częste dla progu 33%:
dict_keys([('chleb', 'ser'), 'chleb', ('cukier', 'kawa'), 'maslo', 'ser', 'mleko', 'orzeszki', 'cukier', 'kawa', 'jajka'])

Ilośc wzorców: 10
----------------------------------
```

Oznacza to, że co trzecie zakupy dokonywane w sklepie, zawierają produkty takie jak:

- Chleb oraz ser
- Masło
- Jajka
- Kawa oraz cukier
- itd

Eksploracja Danych

Ekwiwalentna Transformacja klas ECLAT to algorytm przeszukiwania w głąb wykorzystujący przecięcie zbiorów. Służy do eksploracji częstych wzorców poprzez badanie ich wertykalnego formatu:

Transakcja 1:	kawa	mleko	cukier	orzeszki
Transakcja 2:	kawa	cukier	jajka	
Transakcja 3:	kawa	cukier	chleb	maslo ser
Transakcja 4:	cukier	orzeszki	jajka	miod platki
Transakcja 5:	mleko	jajka	maslo	
Transakcja 6:	kawa	orzeszki	chleb	
Transakcja 7:	mleko	miod	platki	
Transakcja 8:	jajka	chleb	maslo	ser
Transakcja 9:	mleko	chleb	ser	

```
{'chleb': ['3', '6', '8', '9'],
 'cukier': ['1', '2', '3', '4'],
 'jajka': ['2', '4', '5', '8'],
 'kawa': ['1', '2', '3', '6'],
 'maslo': ['3', '5', '8'],
 'miod': ['4', '7'],
 'mleko': ['1', '5', '7', '9'],
 'orzeszki': ['1', '4', '6'],
 'platki': ['4', '7'],
 'ser': ['3', '8', '9']}
```

Eksploracja Danych

Wyszukiwanie reguł asocjacyjnych w zbiorze transakcji:

```
Reguły asocjacyjne:
[('ser', 'chleb', 0.33, 1.0),
 ('cukier', 'kawa', 0.33, 0.75),
 ('chleb', 'ser', 0.33, 0.75),
 ('kawa', 'cukier', 0.33, 0.75)]
```

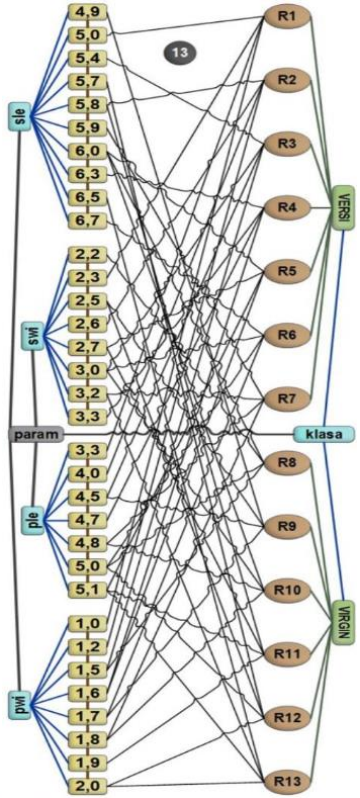
Można wywnioskować, że jeżeli klient zakupi ser, to ze 100 procentową pewnością zakupi także chleb. Jeżeli klient kupi kawę, to z 75 procentową pewnością zakupi także cukier.

AGDS

Systemy asocjacyjne umożliwiają zarówno horyzontalne jak i wertykalne powiązanie danych połączone z agregacją duplikatów, co generuje spore oszczędności.

WZORCE IRIS

param	sle	swi	ple	pwi	klasa
R1	5,0	2,3	3,3	1,0	VERSI
R2	5,8	2,6	4,0	1,2	VERSI
R3	5,4	3,0	4,5	1,5	VERSI
R4	6,3	3,3	4,7	1,6	VERSI
R5	6,0	2,7	5,1	1,6	VERSI
R6	6,7	3,0	5,0	1,7	VERSI
R7	5,9	3,2	4,8	1,8	VERSI
R8	6,0	2,2	5,0	1,5	VIRGIN
R9	4,9	2,5	4,5	1,7	VIRGIN
R10	6,0	3,0	4,8	1,8	VIRGIN
R11	5,8	2,7	5,1	1,9	VIRGIN
R12	5,7	2,5	5,0	2,0	VIRGIN
R13	6,5	3,2	5,1	2,0	VIRGIN

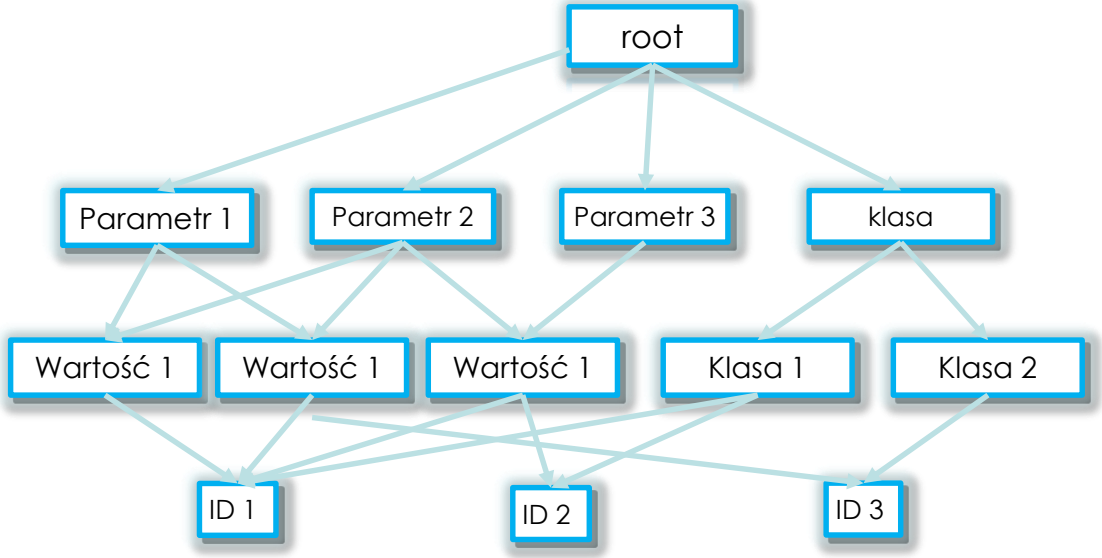


AGDS

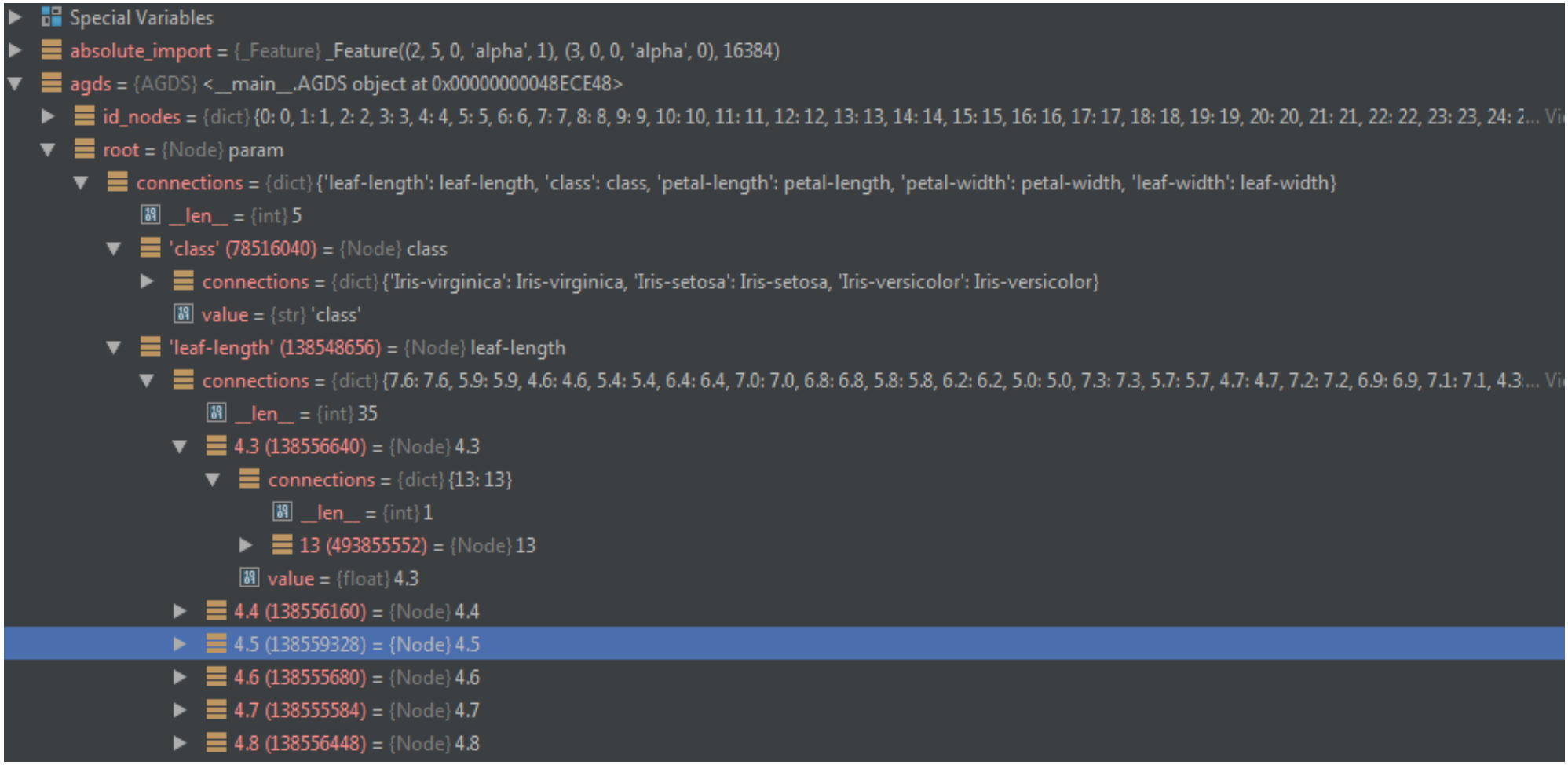
AGDS zaimplementowano w oparciu o hashmapy (słowniki).

Funkcjonalności:

- Filtrowanie parametrów po zakresie
- Wyszukiwanie minimum, maksimów
- Wyszukiwanie n najmniejszych/największych wartości
- Filtrowanie po wielu parametrach
- Wizualizacja

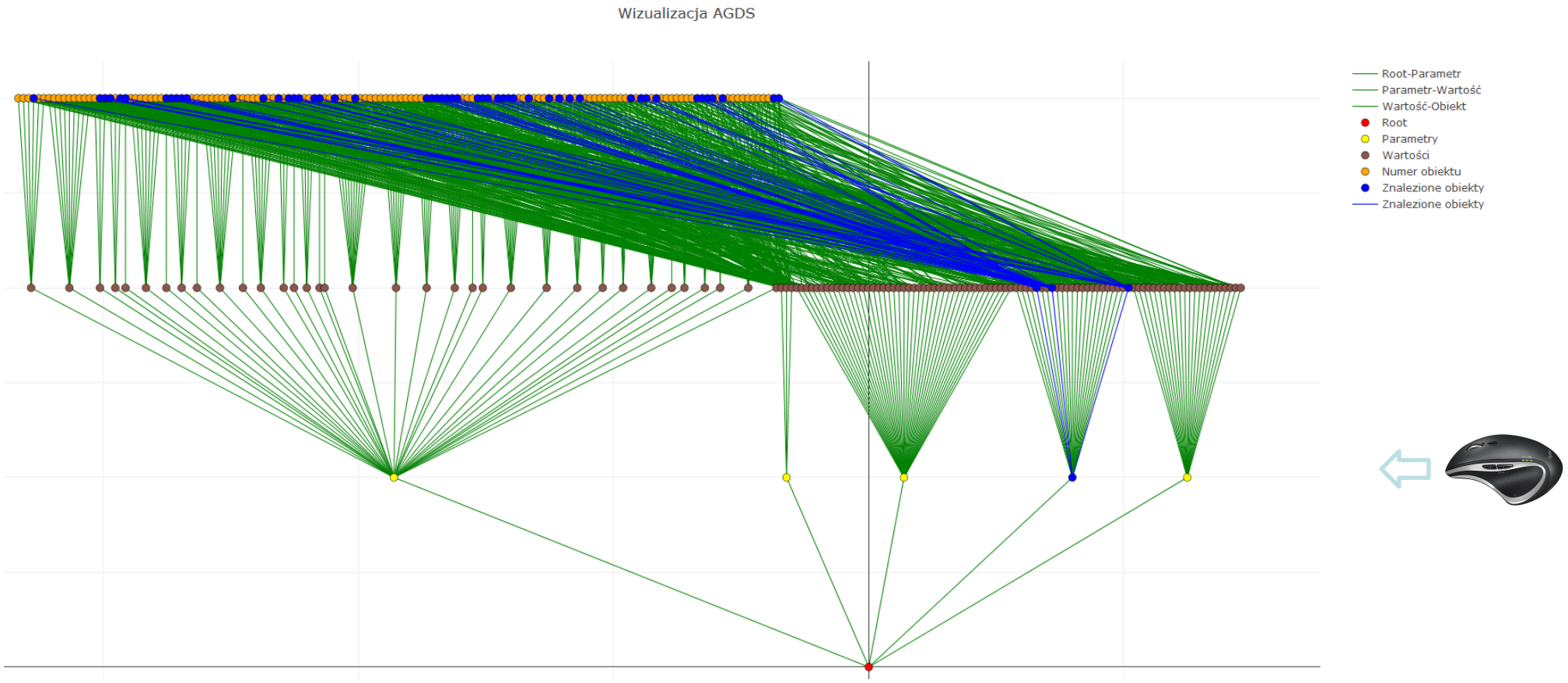


AGDS


Special Variables
absolute_import = {_Feature} _Feature((2, 5, 0, 'alpha', 1), (3, 0, 0, 'alpha', 0), 16384)
agds = {AGDS} <__main__.AGDS object at 0x00000000048ECE48>
id_nodes = {dict} {0: 0, 1: 1, 2: 2, 3: 3, 4: 4, 5: 5, 6: 6, 7: 7, 8: 8, 9: 9, 10: 10, 11: 11, 12: 12, 13: 13, 14: 14, 15: 15, 16: 16, 17: 17, 18: 18, 19: 19, 20: 20, 21: 21, 22: 22, 23: 23, 24:
root = {Node} param
connections = {dict} {'leaf-length': leaf-length, 'class': class, 'petal-length': petal-length, 'petal-width': petal-width, 'leaf-width': leaf-width}
__len__ = {int} 5
'class' (78516040) = {Node} class
connections = {dict} {'Iris-virginica': Iris-virginica, 'Iris-setosa': Iris-setosa, 'Iris-versicolor': Iris-versicolor}
value = {str} 'class'
'leaf-length' (138548656) = {Node} leaf-length
connections = {dict} {7.6: 7.6, 5.9: 5.9, 4.6: 4.6, 5.4: 5.4, 6.4: 6.4, 7.0: 7.0, 6.8: 6.8, 5.8: 5.8, 6.2: 6.2, 5.0: 5.0, 7.3: 7.3, 5.7: 5.7, 4.7: 4.7, 7.2: 7.2, 6.9: 6.9, 7.1: 7.1, 4.3
__len__ = {int} 35
4.3 (138556640) = {Node} 4.3
connections = {dict} {13: 13}
__len__ = {int} 1
13 (493855552) = {Node} 13
value = {float} 4.3
4.4 (138556160) = {Node} 4.4
4.5 (138559328) = {Node} 4.5
4.6 (138555680) = {Node} 4.6
4.7 (138555584) = {Node} 4.7
4.8 (138556448) = {Node} 4.8

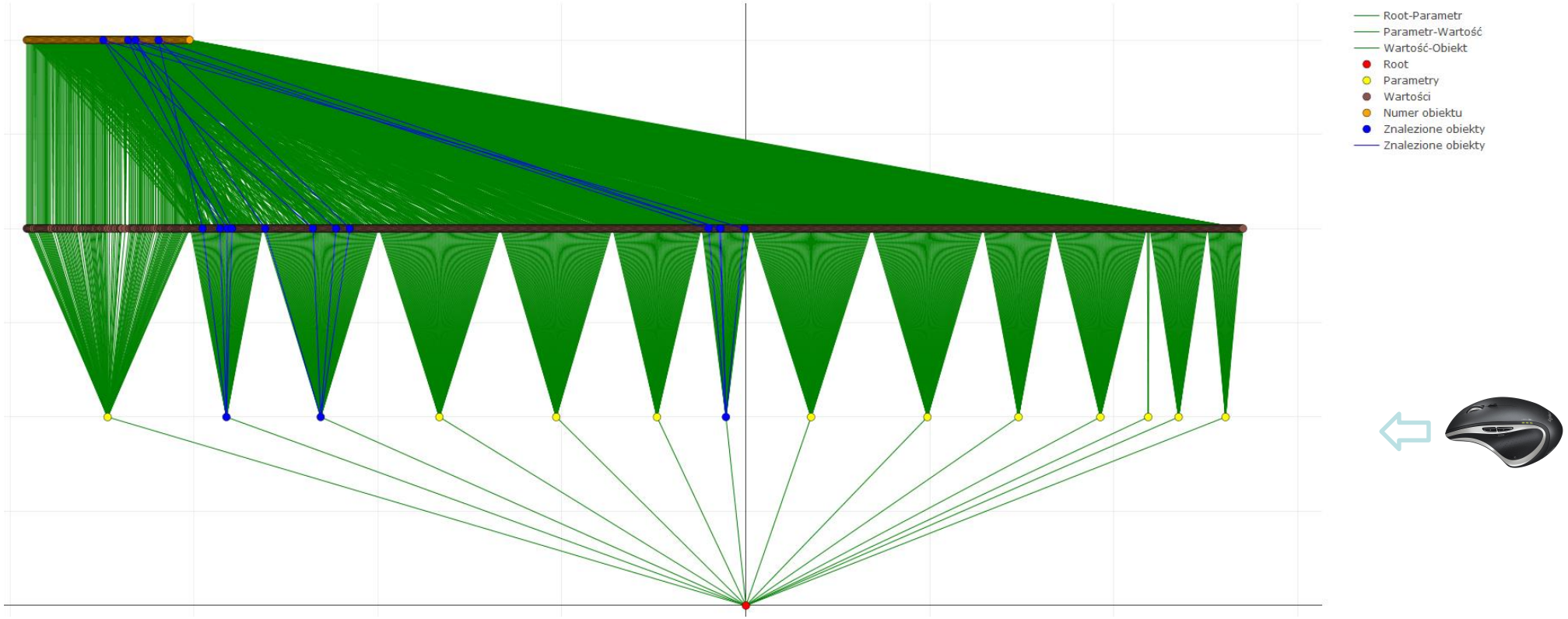
Filtrowanie po zakresie parametru – zbiór Iris

Obiekty z „leaf-length” w zakresie 3.3.2



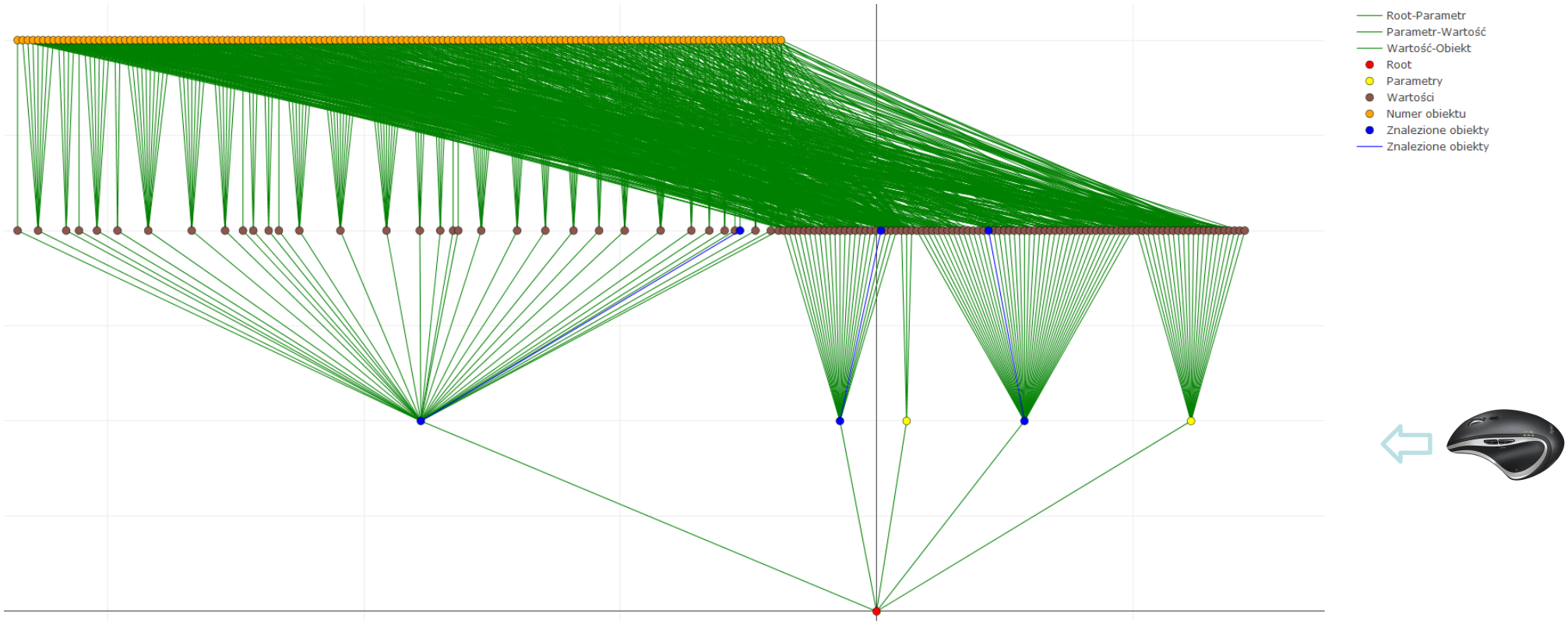
Filtrowanie po zakresie kilku parametrów – zbiór Wine

„Alcohol" w przedziale (14.21,15.89), „Magnesium" w przedziale (88,101), „Ash" w przedziale (2.0, 2.8)



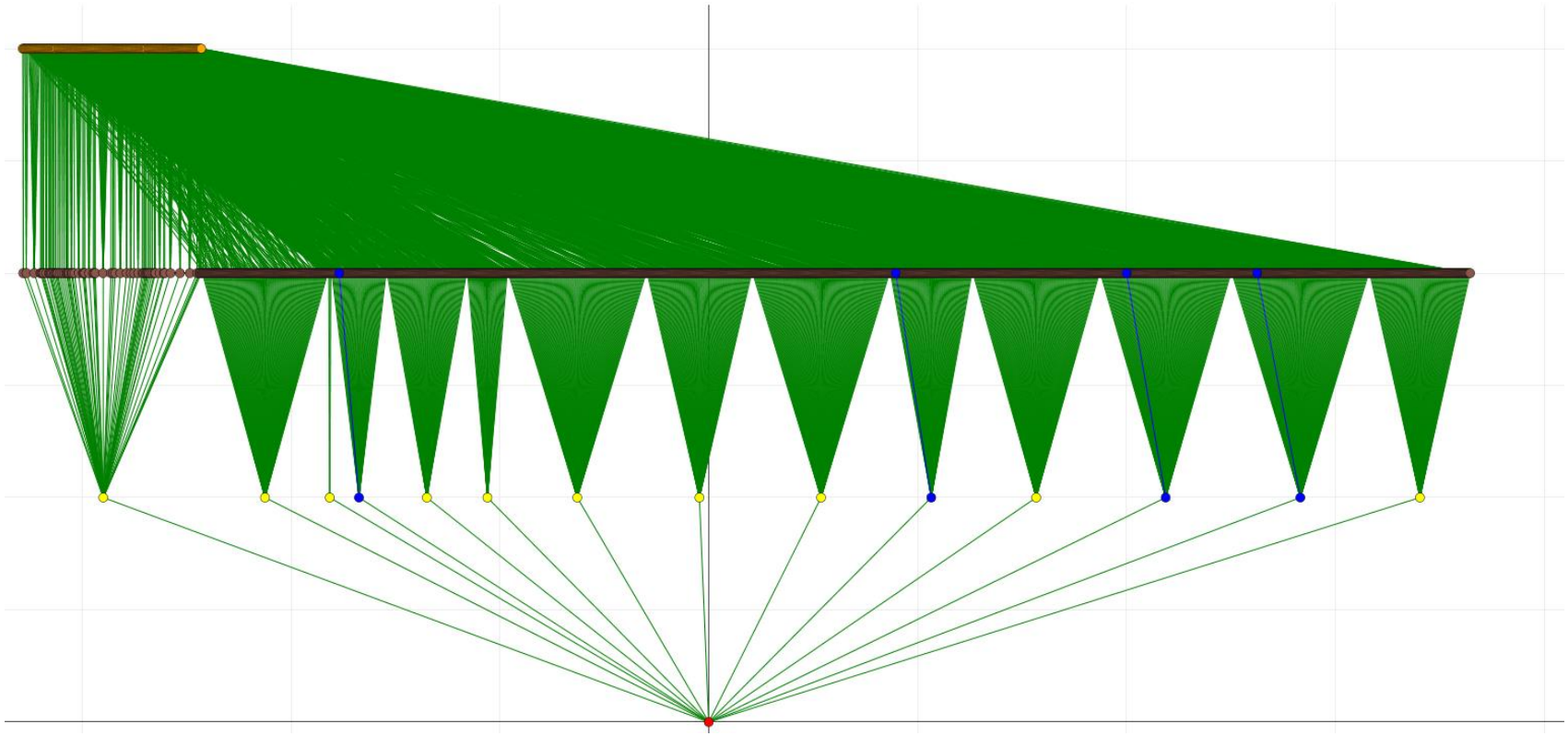
Wyszukiwanie minimów – zbiór Iris

Najmniejsze wartości „petal-witdth", „leaf-width","leaf-length"



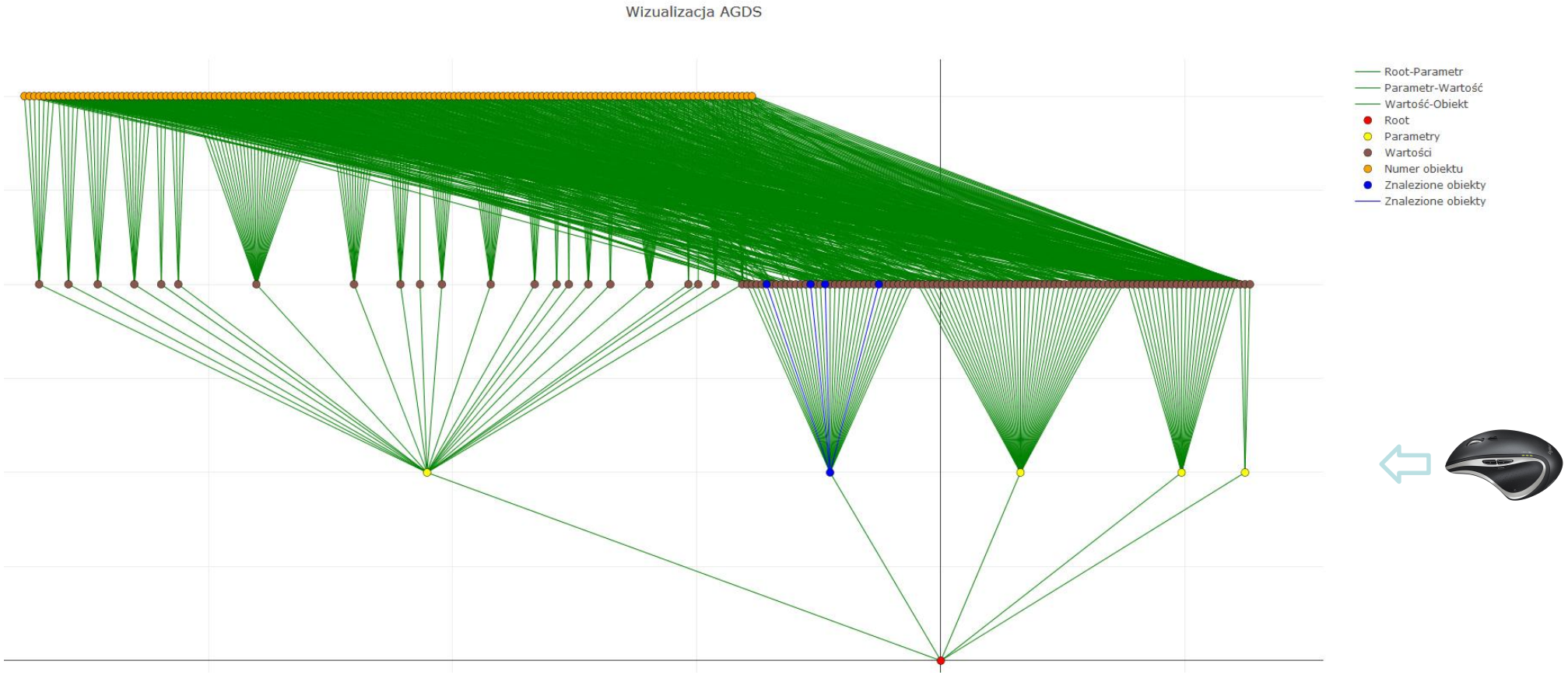
Wyszukiwanie maksimów – zbiór Wine

Największe wartości „Magnesium", „Ash", „Flavanoids", „Alcohol"



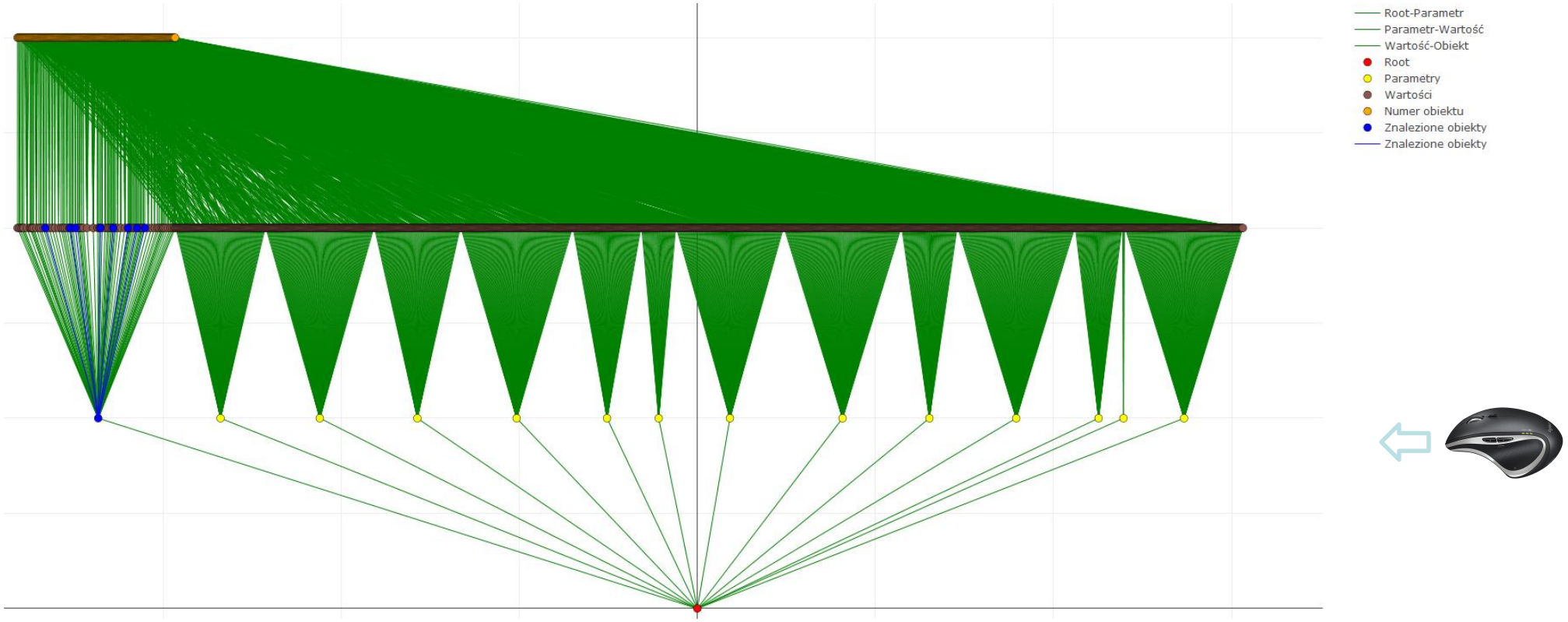
N największych wartości – zbiór Iris

4 największe wartości „leaf-length”



N najmniejszych wartości – zbiór Wine

8 najmniejszych wartości „Ash”



Wiedza to ciągły, społeczny proces dochodzenia do pewnych wniosków, które są uznawane za prawdziwe przez większość lub chociaż część ludzi.

Dziękuję.