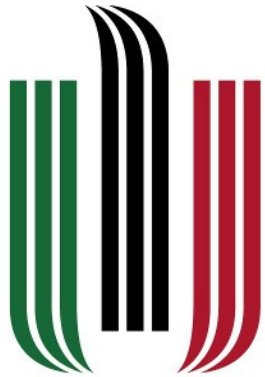


AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

AGH UNIVERSITY OF SCIENCE
AND TECHNOLOGY

AGH

Asocjacyjny system wydajnej automatycznej klasteryzacji i eksploracji danych

Autor: Agata Socha
Promotor: dr hab. Adrian Horzyk

Plan prezentacji

1. Cel i zakres pracy.
2. Wprowadzenie do tematyki pracy:
   - eksploracja danych,
   - klasteryzacja danych,
   - asocjacyjne grafowe struktury danych AGDS.
3. Zastosowane rozwiązanie.
4. Omówienie uzyskanych wyników.
5. Podsumowanie.

Cel i zakres pracy

Cel pracy: projekt i implementacja systemu automatycznej klasteryzacji przy użyciu asocjacyjnych grafowych struktur danych.

Zakres pracy:

- interfejs prezentujący dane i wyniki eksperymentów,
- wizualizacja procesów zachodzących w grafie,
- porównanie szybkości i skuteczności działań na strukturach tabelarycznych i AGDS,
- oszacowanie złożoności obliczeniowej zaimplementowanych algorytmów.

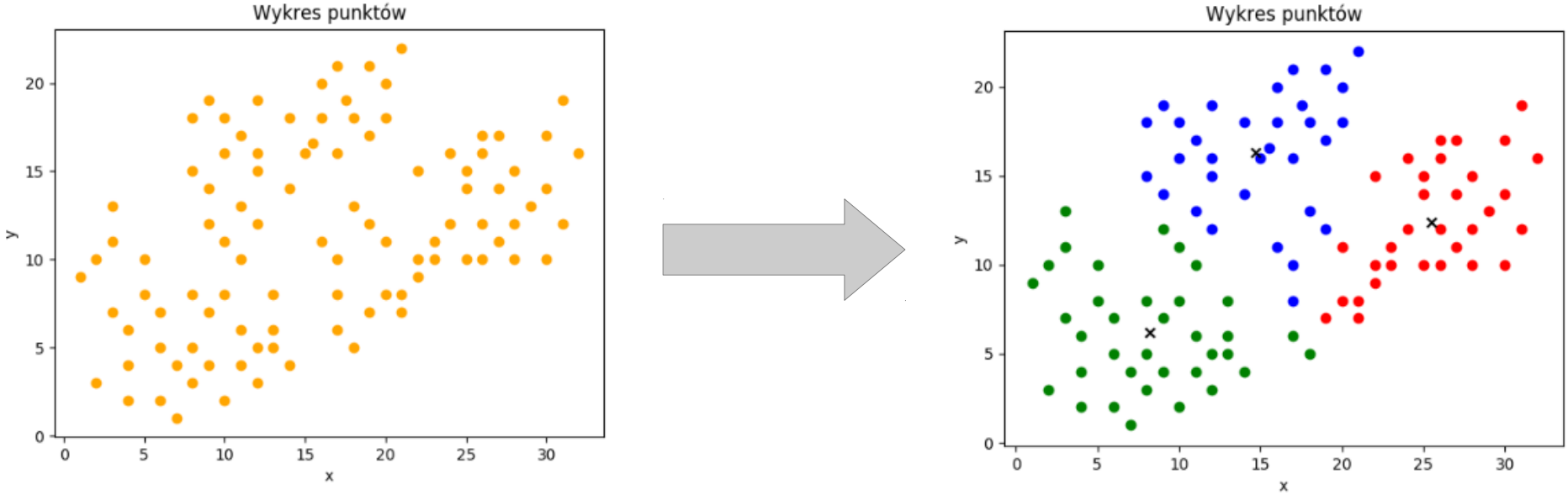
Eksploracja danych

Eksploracja danych to analiza (często ogromnych) zbiorów danych obserwacyjnych w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób tak, aby były zarówno zrozumiałe, jak i przydatne dla ich właściciela.

D. Hand „Principles of data mining”

Klasteryzacja danych

Klasteryzacja to dzielenie zbiory danych na grupy obiektów (klastry), które mają wspólne cechy i są podobne względem wybranej miary podobieństwa.



Wybrane algorytmy klasteryzacji:

- Algorytm klasteryzacji hierarchicznej
- Algorytm k-średnich

Asocjacyjne grafowe struktury danych AGDS

AGDS to struktura grafowa pozwalająca na przechowywanie wartości danych i ich kombinacji oraz relacji je łączących.

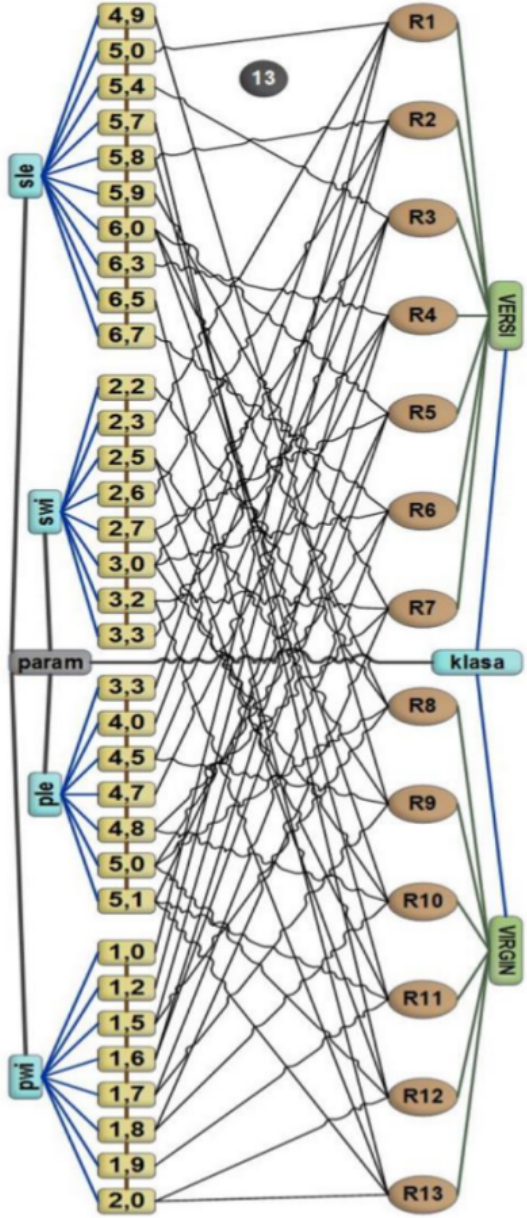
AGDS gwarantuje m.in.:

- przechowywanie posortowanych wartości dla atrybutów,
- kompresję danych poprzez eliminację duplikatów,
- usunięcie nadmiarowych wartości atrybutów i obiektów,
- błyskawiczny dostęp do danych.

Asocjacyjne grafowe struktury danych AGDS

WZORCE IRIS

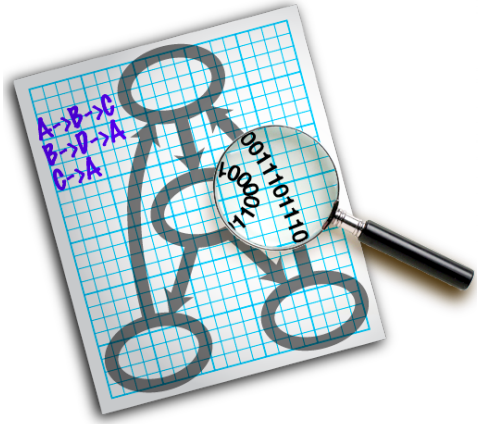
param	sle	swi	ple	pwi	klasa
R1	5,0	2,3	3,3	1,0	VERSI
R2	5,8	2,6	4,0	1,2	VERSI
R3	5,4	3,0	4,5	1,5	VERSI
R4	6,3	3,3	4,7	1,6	VERSI
R5	6,0	2,7	5,1	1,6	VERSI
R6	6,7	3,0	5,0	1,7	VERSI
R7	5,9	3,2	4,8	1,8	VERSI
R8	6,0	2,2	5,0	1,5	VIRGIN
R9	4,9	2,5	4,5	1,7	VIRGIN
R10	6,0	3,0	4,8	1,8	VIRGIN
R11	5,8	2,7	5,1	1,9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



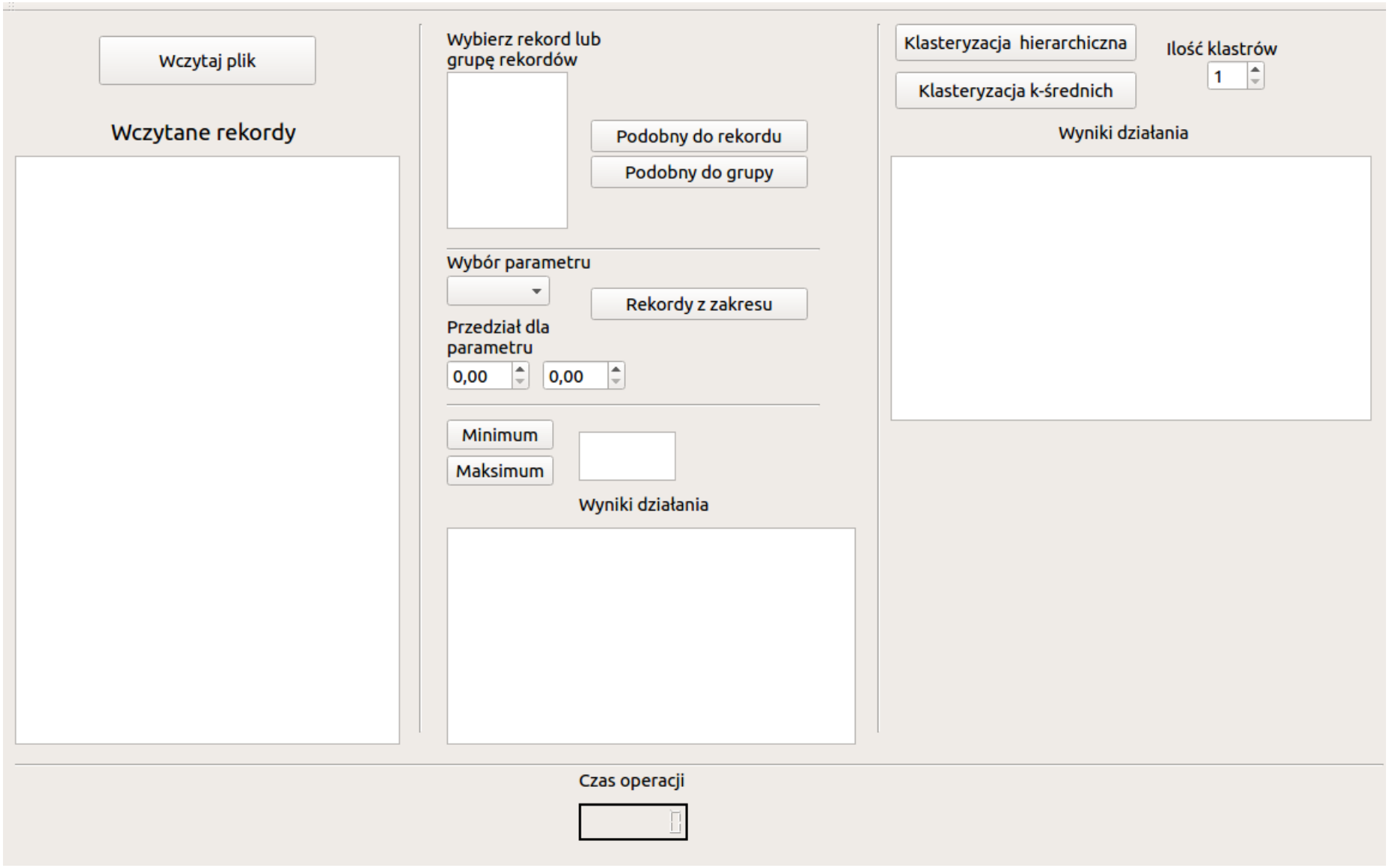
Źródło: A. Horzyk „Struktury asocjacyjne oraz asocjacyjne grafy neuronowe do eksploracji wiedzy z danych”

Implementacja

Implementacja: C++
Wizualizacja: Python, graphviz
GUI: Qt

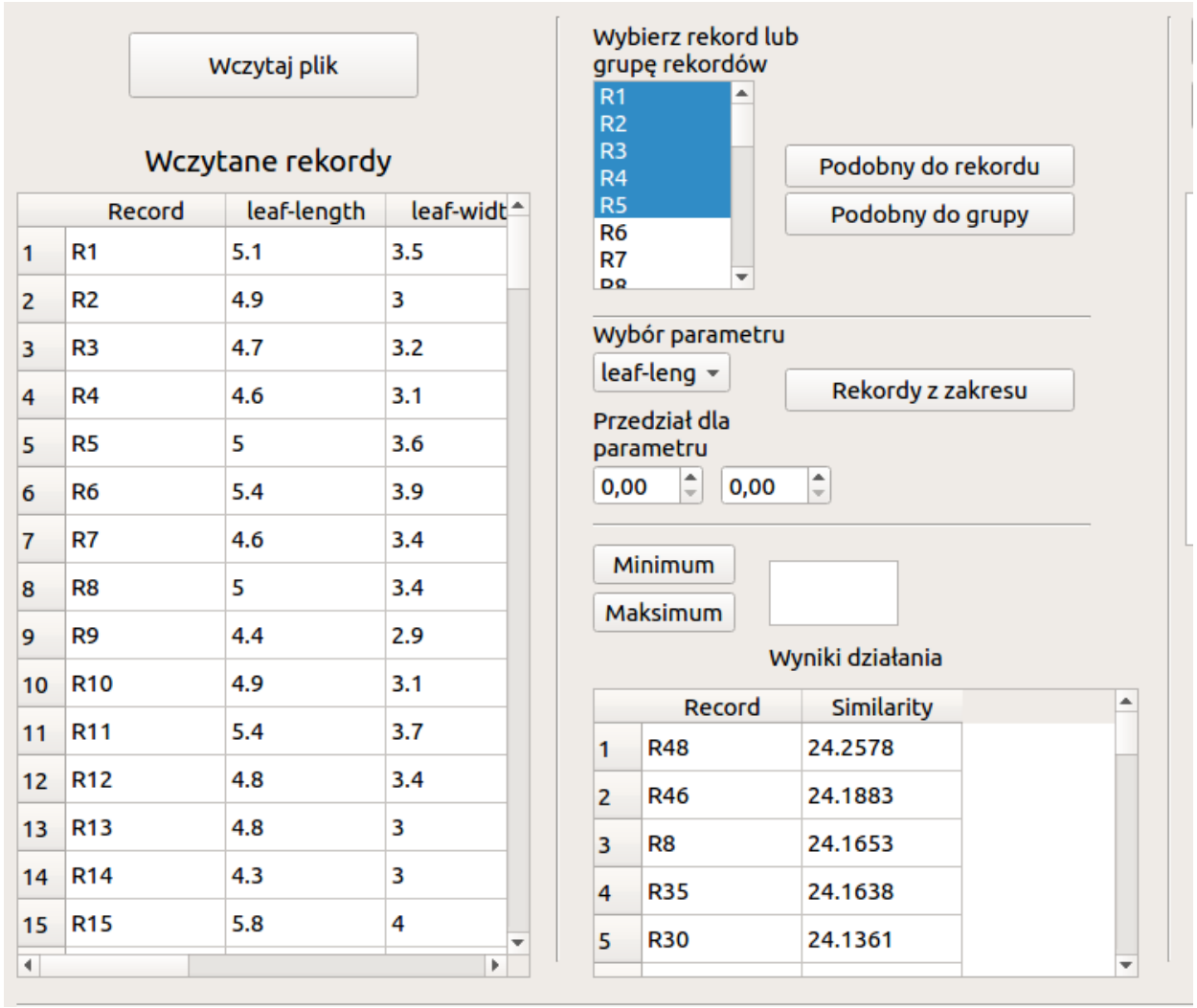


Panel użytkownika



Panel użytkownika

Operacja wyszukiwania rekordów podobnych do grupy



	Record	leaf-length	leaf-widt
1	R1	5.1	3.5
2	R2	4.9	3
3	R3	4.7	3.2
4	R4	4.6	3.1
5	R5	5	3.6
6	R6	5.4	3.9
7	R7	4.6	3.4
8	R8	5	3.4
9	R9	4.4	2.9
10	R10	4.9	3.1
11	R11	5.4	3.7
12	R12	4.8	3.4
13	R13	4.8	3
14	R14	4.3	3
15	R15	5.8	4

	Record	Similarity
1	R48	24.2578
2	R46	24.1883
3	R8	24.1653
4	R35	24.1638
5	R30	24.1361

Wyniki

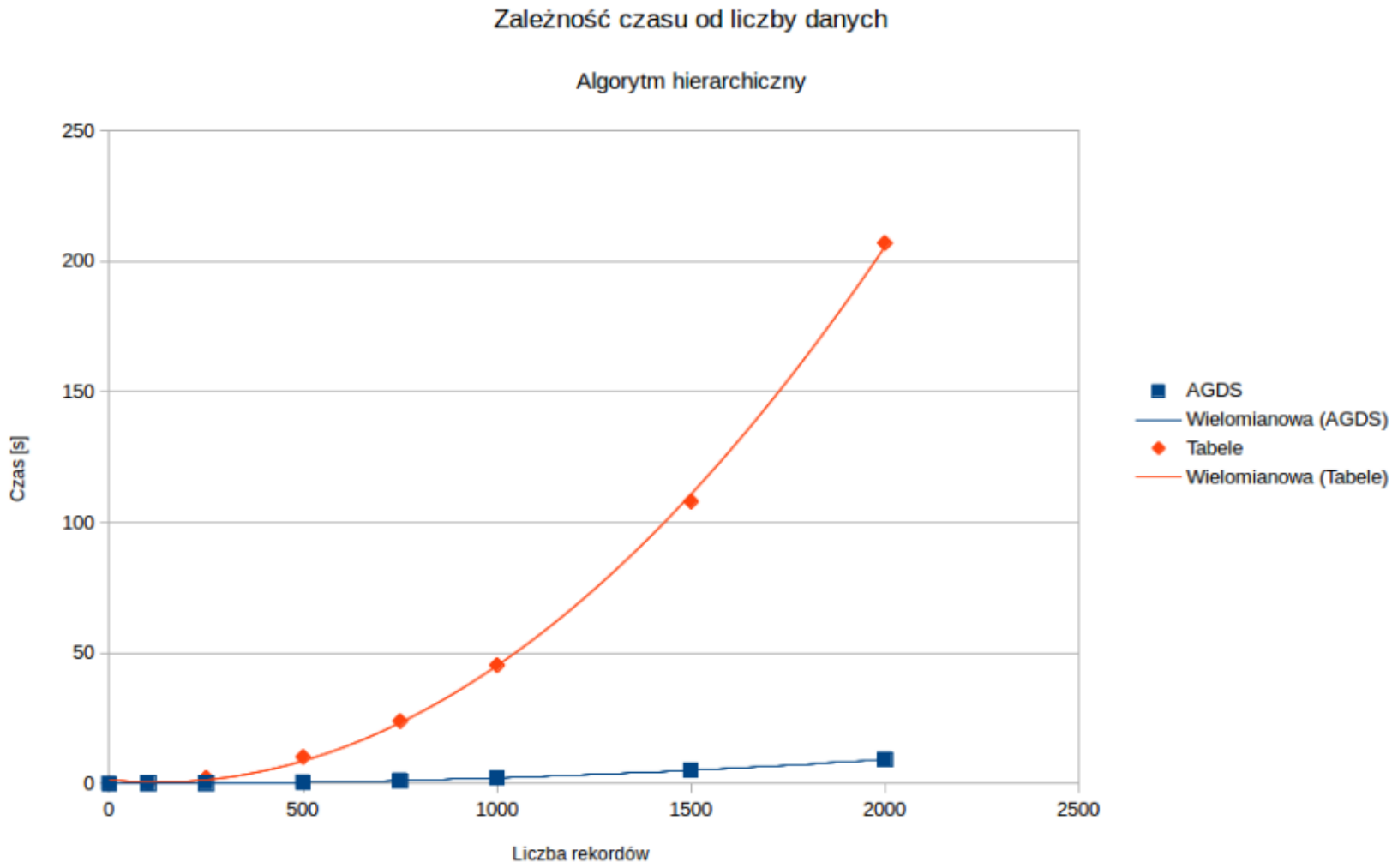
Porównanie czasów działania algorytmów na strukturach tabelarycznych i AGDS

Algorytm hierarchiczny

Rozmiar danych wejścio-wych	Czas działania na struktu-rach tabelarycznych [s]	Czas działania na struktu-rach AGDS [s]
100	0.3	0.017
250	2.13	0.11
500	10.14	0.46
750	23.86	1.14
1000	45.6	2.07
1500	108	5
2000	207	9.22

Wyniki

Porównanie czasów działania algorytmów na strukturach tabelarycznych i AGDS



Wyniki

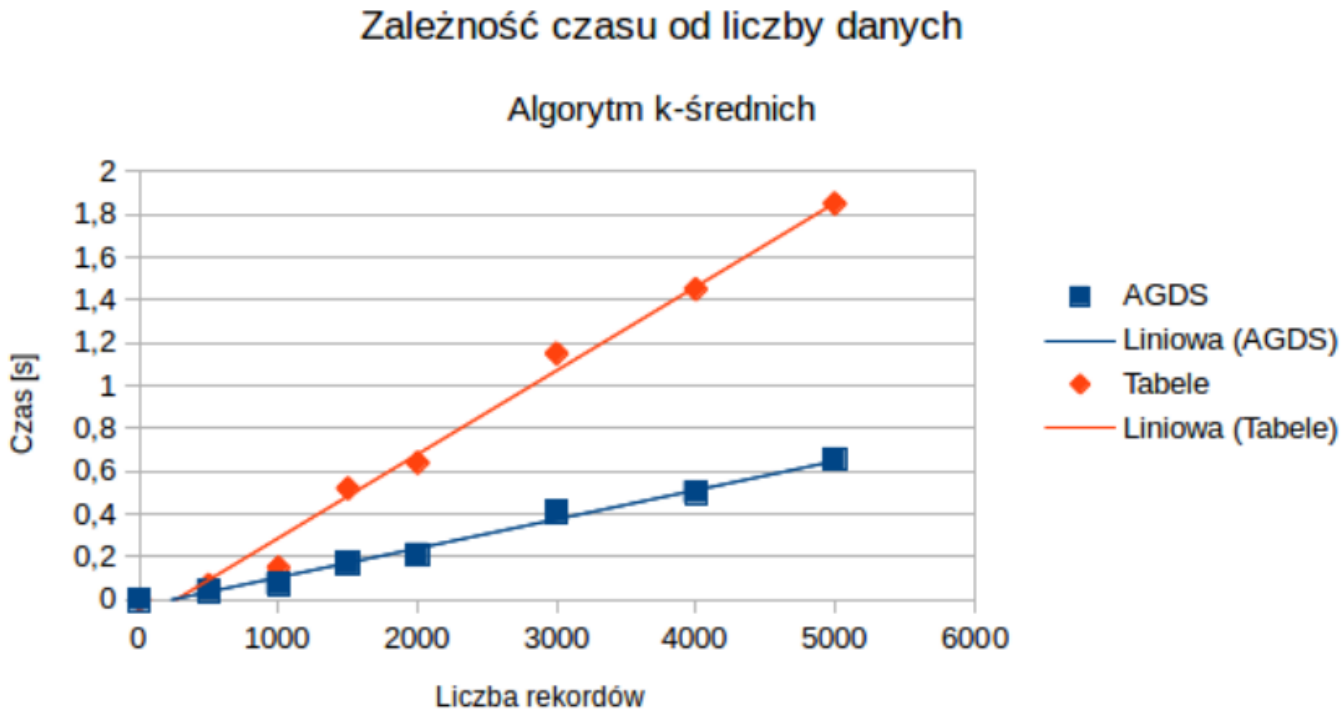
Porównanie czasów działania algorytmów na strukturach tabelarycznych i AGDS

Algorytm k-średnich

Rozmiar danych wejściowych	Czas działania na strukturach tabelarycznych [s]	Czas działania na strukturach AGDS [s]
500	0.07	0.04
1000	0.15	0.07
1500	0.52	0.17
2000	0,64	0.21
3000	1.15	0.41
4000	1.45	0.5
5000	1.85	0.65

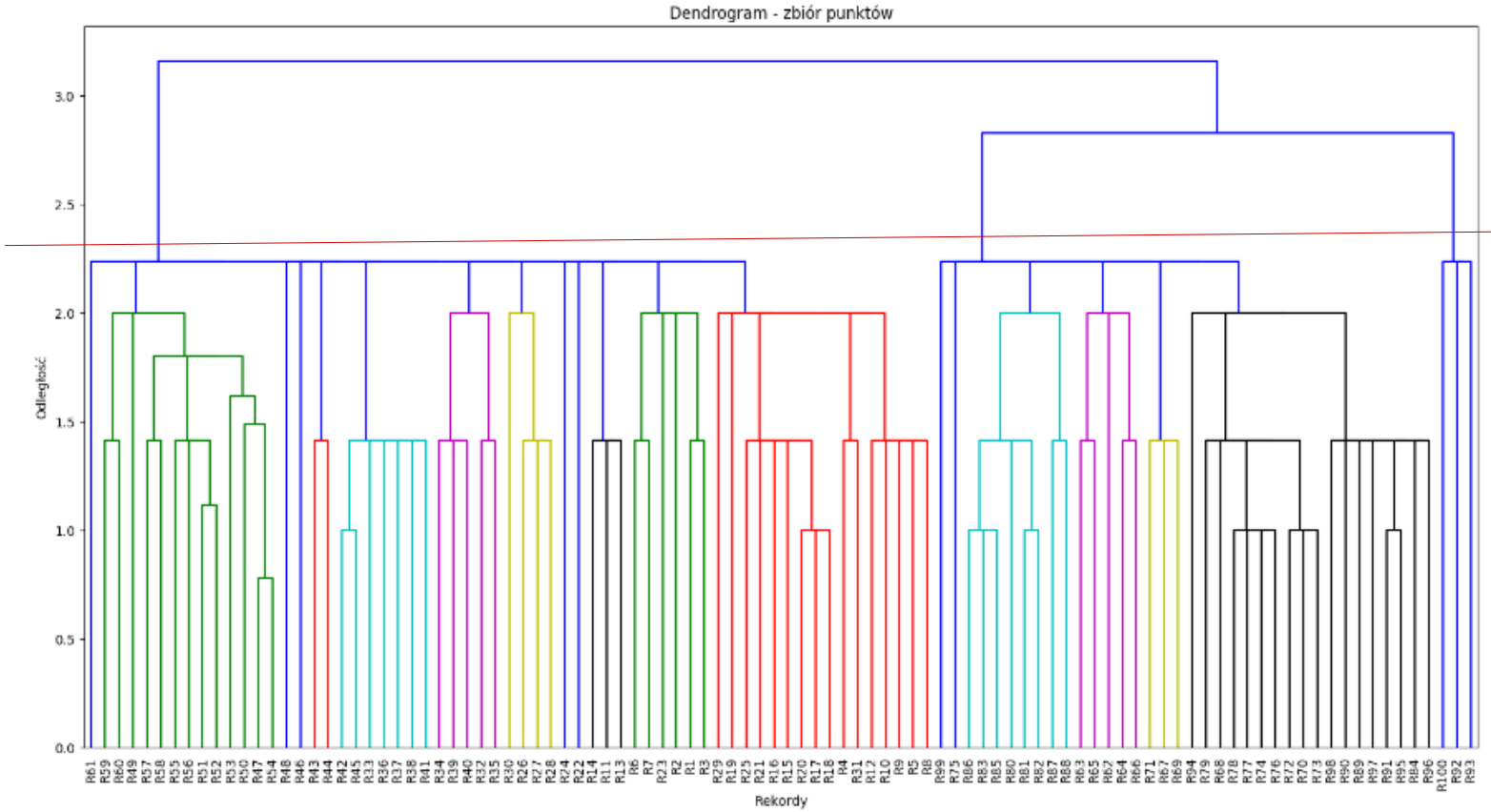
Wyniki

Porównanie czasów działania algorytmów na strukturach tabelarycznych i AGDS



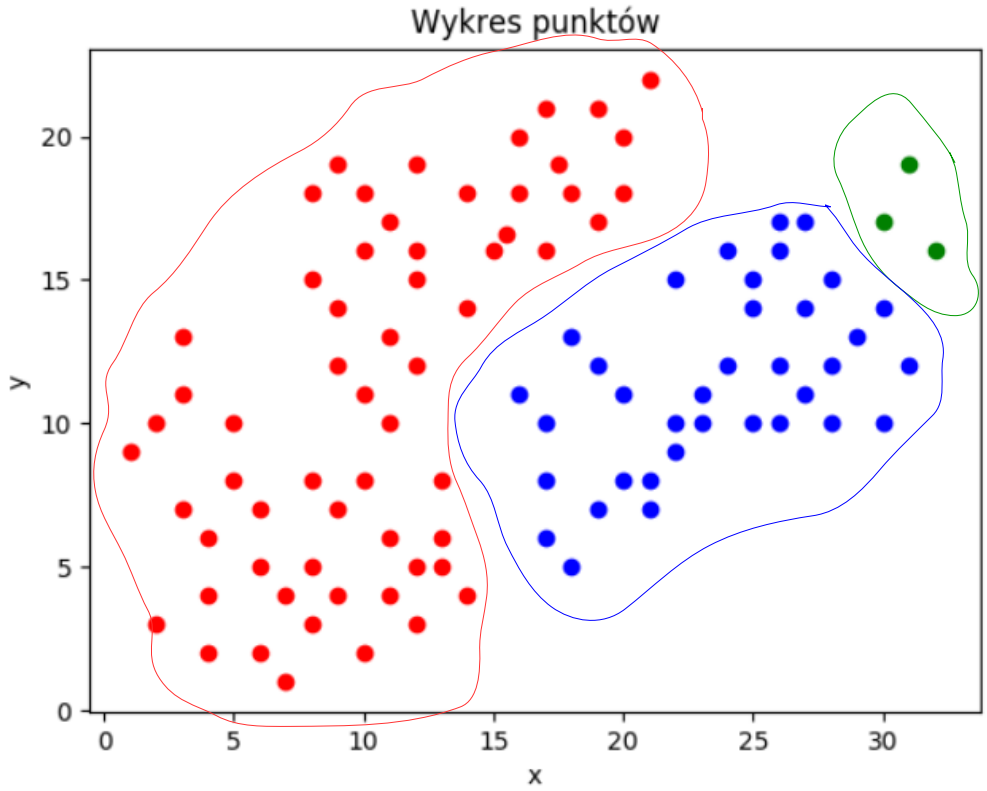
Wyniki

Algorytm hierarchiczny - dendrogram



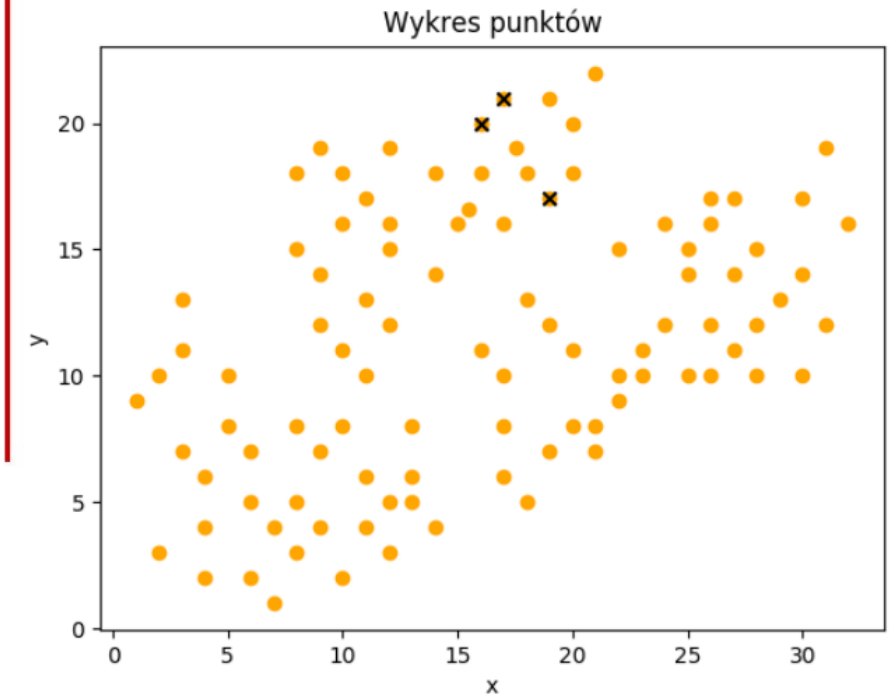
Wyniki

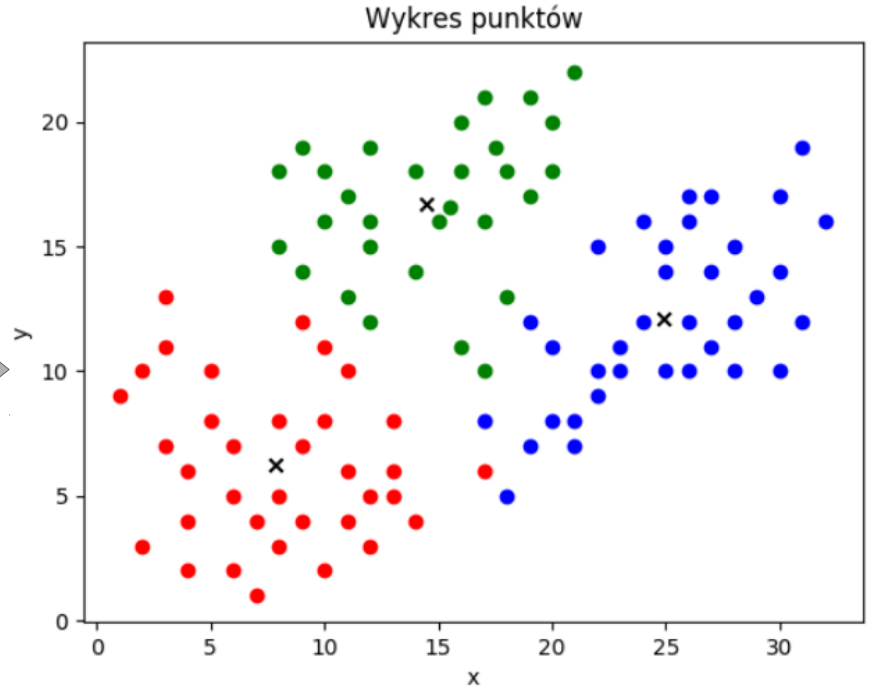
Algorytm hierarchiczny



Wyniki

Algorytm k-średnich





Podsumowanie

- Porównanie czasów działań na strukturach tabelarycznych i AGDS.
- Porównanie działania zaimplementowanych algorytmów.
- Napotkane problemy.
- Dalszy rozwój aplikacji.

Dziękuję za uwagę.