POLITECHNIKA WARSZAWSKA

METODY EKSPLORACJI DANYCH

Projekt

GRUPOWANIE ZE WZGLĘDU NA MIARY: KOSINUSOWĄ, 1-MIARA KOSINUSOWA ORAZ (DLA ZNORMALIZOWANYCH POSTACI WEKTORÓW) ODLEGŁOŚĆ EUKLIDESOWĄ Z UŻYCIEM NBC I TI-NBC

Wykonała:

Zuzanna Górecka

Numer albumu: 305280

Warszawa 2022/2023

1. Opis przydzielonego zadania projektowego
   NBC jest metodą grupowania gęstościowego. Do zastosowania tej metody konieczne było, aby najpierw wyznaczyć k+ najbliższych sąsiadów i liczność odwróconego sąsiedztwa |Rk+NN|. Na podstawie tych informacji można wyznaczyć gęstość podprzestrzeni zgodnie ze wzorem:

$$NDF(p) = \frac{|Rk^{+}NN(p)|}{|k^{+}NN(p)|}$$

   Jeżeli wartość NDF ≥ 1 oznacza to, że punkt jest punktem rdzeniowym. Taki punkt wraz ze swoimi sąsiadami tworzy nową grupę lub staje się częścią już istniejącej.
   Podczas grupowania było również wyznaczane jakiego typu jest każdy punkt. Istnieją 3 rodzaje punktów:
   - punkty rdzeniowe (oznaczane 1)
   - punkty brzegowe (oznaczane 0) – punkt, który należy do grupy, ale nie jest rdzeniem
   - punkty szumu (oznaczane -1) – punkty nie należące do żadnej grupy

   Projekt składał się z następujących etapów:
   - przygotowanie danych i wczytanie ich
   - normalizacja danych
   - implementacja algorytmu znajdowania k+NN dla postaci znormalizowanych. Powinno to zostać wykonane w 3 wariantach:
     - bez wykorzystania nierówności trójkąta
     - z wykorzystaniem nierówności trójkąta
     - z wykorzystaniem nierówności trójkąta i ostateczną weryfikacją odległości pomiędzy punktami z wykorzystaniem miary kosinusowej
   - wyznaczenie odwrotnego k+ sąsiedztwa
   - wyliczenie NDF
   - grupowanie punktów
   - zapisanie wyników do odpowiednich plików

2. Przyjęte założenia
   Dane uczące wcześniej poddawane są wstępnemu przetwarzaniu polegającemu na usunięciu brakujących danych, w razie potrzeby zmianie identyfikatorów grup na numeryczne i odpowiedni zapis do plików. Same dane (bez nagłówków) są zapisywane do 2 plików:
   - pierwszy ze współrzędnymi punktów
   - drugi ze wzorcowymi klasami punktów

   Dla poprawnego działania programu wszystkie atrybuty muszą być numeryczne. Dane powinny być zapisane w formacie csv rozdzielanym przecinkami.

   Przyjęto k =3.

   Pojedynczy wiersz danych jest wczytywany do klasy DataSample. Wczytywanie danych obsługuje klasa CSVReader. Zaimplementowano 3 klasy do wyznaczenia k+NN: klasa KNNAlgorithm, TikNNAlgorithm i TiCoskNNAlgorithm. Klasa NBCAlgorithm wyznacza podział na grupy.

Podczas wyznaczania k+NN sortowany jest pomocniczy vector zawierający informacje o id punktu i jego odległości od drugiego punktu, zamiast sortowania vectora z pełnymi danymi.

Liczność Rk+NN wyznaczana jest podczas wyznaczania k+NN, więc przyjęto czas wykonania wyliczenia |Rk+NN| równy czasowi wyliczenia k+NN.

3. Zbiór danych i dane wyjściowe

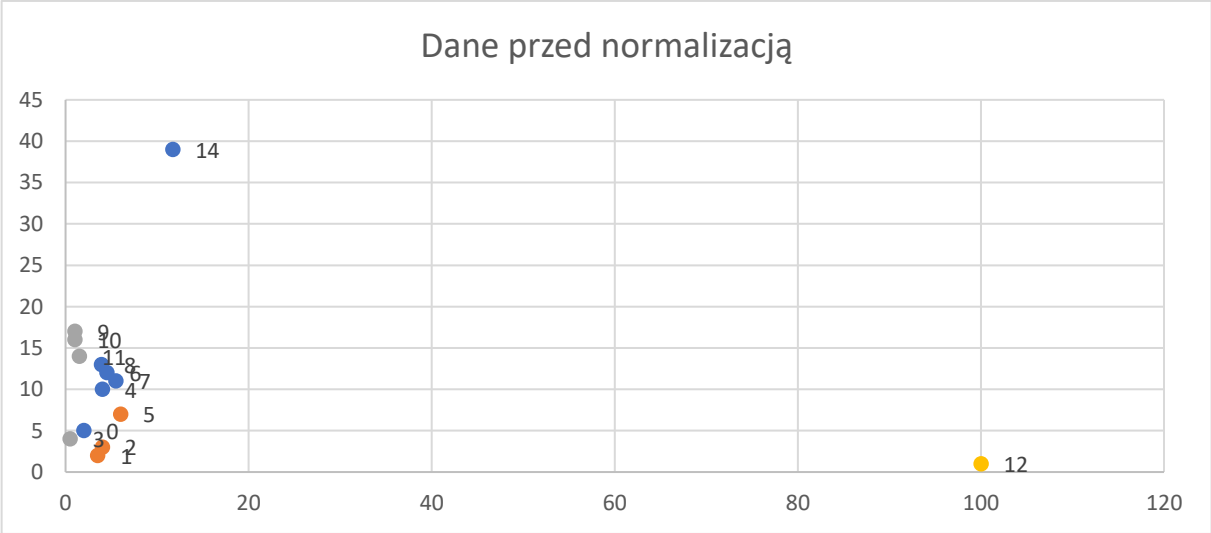
3.1. Własny, mały zbiór danych

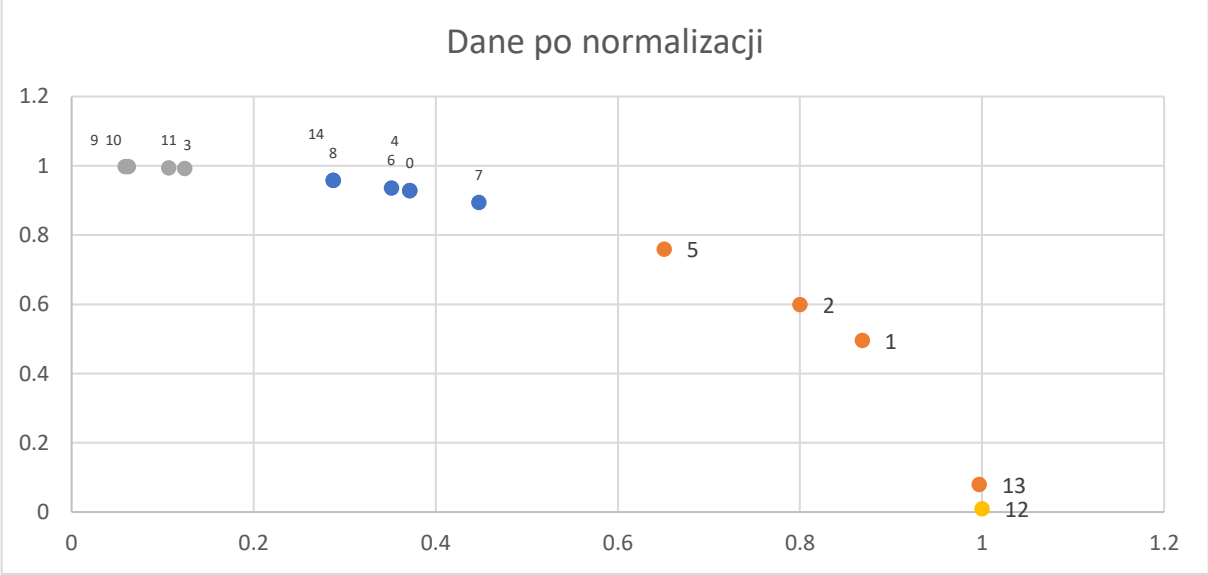
Przygotowano własny zbiór danych, który był używany do walidacji działania algorytmu.

Specyfikacja:

- 15 przykładów
- 2 współrzędne
- 4 grupy

Poniżej przedstawiono na wykresie dane. Kolory odpowiadają grupom wzorcowym.

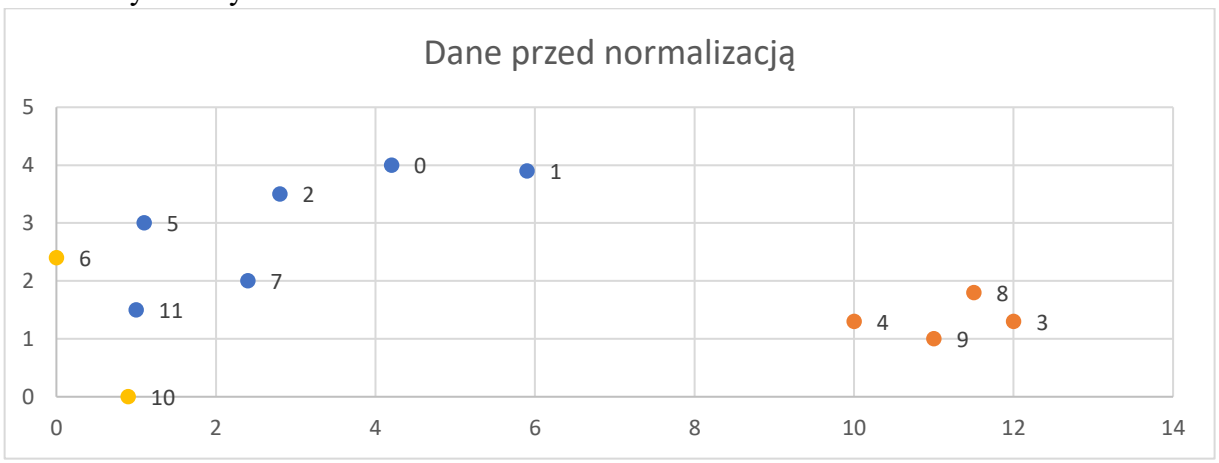


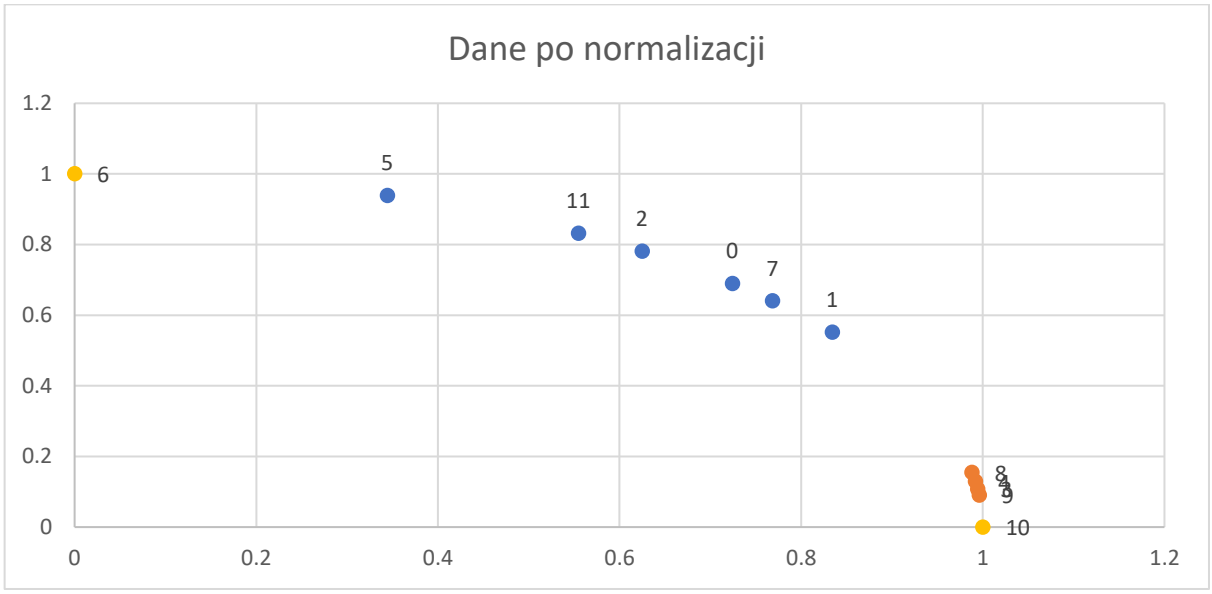


Po normalizacji część punktów na siebie nachodzi, jednakże nie przeszkadza to w obliczeniach, a dzięki temu łatwiej uzyskać liczbę sąsiadów większą niż k.

Dzięki temu, że zbiór był mały można było ręcznie sprawdzić poprawność działania algorytmu na wszystkich jego etapach. W tym celu w excelu wyliczono potrzebne dane dla każdego punktu.

3.2. Zbiór danych z wykładu





3.3. Wine Quality Dataset
Specyfikacja:

- 1143 przykładów
- 11 atrybutów
- 6 grup

Atrybutami są własności fizykochemiczne wina, a klasami jakość tego wina.

3.4. Glass Identification
Specyfikacja:

- 214 przykładów
- 9 atrybutów
- 7 grup

Atrybuty wynikają ze składu chemicznego.

3.5. Iris

Specyfikacja:

- 150 przykładów
- 4 atrybuty
- 3 grupy

Dane zawierają informacje o długości kwiatów i płatków irysów.

3.6. Forest Cover Type

Specyfikacja:

- 581011 przykładów
- 53 atrybuty
- 7 grupy

Celem było sprawdzenie działania algorytmu na różnych zbiorach danych i jak wydłuża się czas działania wraz ze zwiększaniem liczby przykładów i atrybutów. Dla tych danych sam czas wczytania ich zajął 8 min, więc ostatecznie z niego zrezygnowano.

3.7. Dane wyjściowe

Zgodnie z zaleceniami wyniki grupowania są zapisywane do 3 plików:
k+NN, OUT i STAT. Podane w nich czasy są w mikrosekundach.

4. Instrukcja uruchomienia

Należy otworzyć plik „main.cpp" w edytorze kodu, np. VS Code. W razie potrzeby w kodzie można zmienić ścieżkę do danych (path), zbiór danych (dataset_name) lub wartość k. Aby uruchomić kod w VS Code można np. wcisnąć F5.

5. Wyniki

5.1. Bez wykorzystania nierówności trójkąta dla różnych wartości k

5.1.1. Własny zbiór danych

Dla każdego punktu wykonano taką samą liczbę operacji wyznaczenia długości = 28.
Plik STAT

nazwa_pliku_wejsciowego	own_dataset.csv	own_dataset.csv	own_dataset.csv	own_dataset.csv
liczba_wym_punktow	2	2	2	2
l_punktow	15	15	15	15
k	2	3	5	10
czas_wczytania_pliku_wej	2154	1560	1699	1758
czas_normalizacji	2.5	1.8	2.8	2.6
calkowity_czas_sortowania	32	17	34	19
czas_obliczania_kNN	563	470	596	406
czas_obliczania_RkNN	563	470	596	406
czas_grupowania	29	31	43	46
czas_zapisywania_plikow_out_kNN	9879	7411	8561	6514
calkowity_czas_dzialania	12645	9486	10921	8738
calkowity_czas_dzialania-czas_zapisywania_plikow	2766	2075	2360	2224
l_odkrytych_grup	4	4	2	2
l_odkrytych_p_szumu	4	1	5	2

l_odkrytych_p_rdzeniowych	8	10	7	9
l_odkrytych_p_brzegowych	3	4	3	4
srednia_l_obl_odl	28	28	28	28
|TP|	13	27	27	24
|TN|	70	78	50	23
l_par_punktow	105	105	105	105
RAND	0.790476	1	0.733333	0.447619

5.1.2. Wine Quality Dataset

Wyniki pliku STAT dla różnych wartości k

nazwa_pliku_wejsciowego	WineQT.csv	WineQT.csv	WineQT.csv	WineQT.csv	WineQT.csv
liczba_wym_punktow	11	11	11	11	11
l_punktow	1143	1143	1143	1143	1143
k	2	3	5	10	30
czas_wczytania_pliku_wej	67564	89022	75230	71704	117172
czas_normalizacji	13.3	21.7	16.8	14.2	14.6
calkowity_czas_sortowania	456215	248159	539403	608754	250408
czas_obliczania_kNN	42407319	17427467	43407142	133250535	1456764
czas_obliczania_RkNN	42407319	17427467	43407142	133250535	1456764
czas_grupowania	12428	3482	5232	24544	4132
czas_zapisywania_plikow_out_kNN	431206	131974	1779342	1468242	142545
calkowity_czas_dzialania	42988631	17670239	45327061	134883937	1738312
calkowity_czas_dzialania-czas_zapisywania_plikow	42557425	17538265	43547719	133415695	1595767
l_odkrytych_grup	209	124	39	2	2
l_odkrytych_p_szumu	310	247	201	167	74
l_odkrytych_p_rdzeniowych	710	677	648	638	647
l_odkrytych_p_brzegowych	123	219	294	338	422
srednia_l_obl_odl	2284	2284	2284	2284	2284
|TP|	17066	12848	24092	178789	208833
|TN|	386307	396312	377788	108071	54229
l_par_punktow	652653	652653	652653	652653	652653
RAND	0.618051	0.626918	0.615764	0.439529	0.403066

5.1.3. Glass Identification

nazwa_pliku_wejsciowego	glass.csv	glass.csv	glass.csv	glass.csv
liczba_wym_punktow	9	9	9	9
l_punktow	214	214	214	214
k	2	3	5	10
czas_wczytania_pliku_wej	5390	9641	11019	6742
czas_normalizacji	18.9	9.5	15.5	6
calkowity_czas_sortowania	8098	7949	7564	6840
czas_obliczania_kNN	333722	300454	319731	279487
czas_obliczania_RkNN	333722	300454	319731	279487
czas_grupowania	505	320	283	517

czas_zapisywania_plikow_out_kNN	22630	18359	15038	24628
calkowity_czas_dzialania	363157	329347	346678	311967
calkowity_czas_dzialania-czas_zapisywania_plikow	340527	310988	331640	287339
l_odkrytych_grup	41	21	11	4
l_odkrytych_p_szumu	51	50	40	38
l_odkrytych_p_rdzeniowych	133	121	124	112
l_odkrytych_p_brzegowych	30	43	50	64
srednia_l_obl_odl	426	426	426	426
|TP|	500	965	1517	3014
|TN|	15702	15239	14928	12386
l_par_punktow	22791	22791	22791	22791
RAND	0.710895	0.710982	0.721557	0.675705

5.1.4. Iris

nazwa_pliku_wejsciowego	iris.csv	iris.csv	iris.csv	iris.csv
liczba_wym_punktow	4	4	4	4
l_punktow	150	150	150	150
k	2	3	5	10
czas_wczytania_pliku_wej	2594	3318	2308	2667
czas_normalizacji	1.7	1.7	1.7	1.2
calkowity_czas_sortowania	2273	2116	2109	2282
czas_obliczania_kNN	48967	45508	52452	16780
czas_obliczania_RkNN	48967	45508	52452	16780
czas_grupowania	101	94	200	241
czas_zapisywania_plikow_out_kNN	13470	12490	17923	16804
calkowity_czas_dzialania	65374	61689	73101	36713
calkowity_czas_dzialania-czas_zapisywania_plikow	51904	49199	55178	19909
l_odkrytych_grup	29	15	5	3
l_odkrytych_p_szumu	37	33	37	20
l_odkrytych_p_rdzeniowych	91	91	86	81
l_odkrytych_p_brzegowych	22	26	27	49
srednia_l_obl_odl	298	298	298	298
|TP|	438	834	1907	2821
|TN|	7040	7123	5726	5395
l_par_punktow	11175	11175	11175	11175
RAND	0.669172	0.712036	0.683043	0.735213

5.2. Porównanie wyników dla 3 wariantów wyliczania k+NN

Przedstawione wyniki dla każdego zbioru i każdego wariantu są wynikami uśrednionymi z 10 uruchomień kodu. Podczas działania kodu w tle nie było uruchomionych innych programów. Dzięki temu wyniki będą bardziej miarodajne i łatwiej będzie porównywać czasy wykonywania poszczególnych faz programu.

5.2.1. Własny zbiór danych

	bez Ti	Ti	Ti+cos
nazwa_pliku_wejsciowego	own_dataset.csv	own_dataset.csv	own_dataset.csv
liczba_wym_punktow	2	2	2
l_punktow	15	15	15
k	3	3	3
czas_wczytania_pliku_wej [us]	528.2	561.8	629.4
czas_normalizacji [us]	1.5	2.17	1.79
calkowity_czas_sortowania [us]	15.5	2.9	5.4
czas_obliczania_kNN [us]	439.7	274.4	291.9
czas_obliczania_RkNN [us]	439.7	274.4	291.9
czas_grupowania [us]	16.2	21.4	19.6
czas_zapisywania_plikow_out_kNN [us]	3997.1	4600.8	4108.5
calkowity_czas_dzialania [us]	4545.9	5189.7	4763.2
calkowity_czas_dzialania-czas_zapisywania_plikow [us]	548.8	588.9	654.7
l_odkrytych_grup	4	4	4
l_odkrytych_p_szumu	1	1	1
l_odkrytych_p_rdzeniowych	10	10	10
l_odkrytych_p_brzegowych	4	4	4
srednia_l_obl_odl	28	8.33333	8.33333
|TP|	27	27	27
|TN|	78	78	78
l_par_punktow	105	105	105
RAND	1	1	1

Uzyskany podział na grupy był identyczny dla każdej wersji algorytmu wyznaczającego k+NN.

5.2.2. Zbiór z wykładu

	bez Ti	Ti	Ti+cos
nazwa_pliku_wejsciowego	wyklad.csv	wyklad.csv	wyklad.csv
liczba_wym_punktow	2	2	2
l_punktow	12	12	12
k	3	3	3
czas_wczytania_pliku_wej [us]	273.7	290.6	264.7
czas_normalizacji [us]	0.89	1.05	0.96
calkowity_czas_sortowania [us]	1.7	1.1	1.3
czas_obliczania_kNN [us]	151.6	93.1	91.9
czas_obliczania_RkNN [us]	151.6	93.1	91.9
czas_grupowania [us]	7.7	9	7.7
czas_zapisywania_plikow_out_kNN [us]	1642.4	1877.1	2170.9
calkowity_czas_dzialania [us]	1925.9	2179	2445.6
calkowity_czas_dzialania-czas_zapisywania_plikow [us]	283.5	301.9	274.7
l_odkrytych_grup	3	3	3
l_odkrytych_p_szumu	2	2	2
l_odkrytych_p_rdzeniowych	8	8	8

l_odkrytych_p_brzegowych	2	2	2
srednia_l_obl_odl	22	7	7
|TP|	22	22	22
|TN|	44	44	44
l_par_punktow	66	66	66
RAND	1	1	1

Uzyskany podział na grupy był identyczny dla każdej wersji algorytmu wyznaczającego k+NN.

5.2.3. Wine Quality Dataset

	bez Ti	Ti	Ti+cos
nazwa_pliku_wejsciowego	WineQT.csv	WineQT.csv	WineQT.csv
liczba_wym_punktow	11	11	11
l_punktow	1143	1143	1143
k	3	3	3
czas_wczytania_pliku_wej [us]	193862.7	189618.3	165867.8
czas_normalizacji [us]	10.51	12.04	11.66
calkowity_czas_sortowania [us]	242749.5	241.2	228.6
czas_obliczania_kNN [us]	16534198.3	4237345.8	5989460.9
czas_obliczania_RkNN [us]	16534198.3	4237345.8	5989460.9
czas_grupowania [us]	697.5	679.8	1061.3
czas_zapisywania_plikow_out_kNN [us]	78236.5	73820.1	71188.9
calkowity_czas_dzialania [us]	290656.7	282359	257240
calkowity_czas_dzialania-czas_zapisywania_plikow [us]	212420.2	208538.9	186051.1
l_odkrytych_grup	124	124	76
l_odkrytych_p_szumu	247	247	137
l_odkrytych_p_rdzeniowych	677	677	746
l_odkrytych_p_brzegowych	219	219	260
srednia_l_obl_odl	2284	529.404	592.101
|TP|	12848	12848	33837
|TN|	396312	396312	363099
l_par_punktow	652653	652653	652653
RAND	0.626918	0.626918	0.608188

Uzyskany podział na grupy był inny dla algorytmu wykorzystującego przy ostatecznej weryfikacji odległości miarę kosinusową. Algorytmy niewykorzystujący nierówność trójkąta i wykorzystujący, ale z odległością euklidesową dały taki sam podział na grupy.

5.2.4. Glass Identification

	bez Ti	Ti	Ti+cos
nazwa_pliku_wejsciowego	glass.csv	glass.csv	glass.csv
liczba_wym_punktow	9	9	9
l_punktow	214	214	214
k	3	3	3
czas_wczytania_pliku_wej [us]	7898.2	8035.9	8051.6

czas_normalizacji [us]	13.39	11.31	13.85
calkowity_czas_sortowania [us]	7834.3	70.2	60.8
czas_obliczania_kNN [us]	239558.3	227924.7	448997.6
czas_obliczania_RkNN [us]	239558.3	227924.7	448997.6
czas_grupowania [us]	142.4	147.1	201.4
czas_zapisywania_plikow_out_kNN [us]	16912.3	14766.6	15788.1
calkowity_czas_dzialania [us]	25538.2	23527.5	24613.2
calkowity_czas_dzialania-czas_zapisywania_plikow [us]	8625.9	8760.9	8825.1
l_odkrytych_grup	21	21	20
l_odkrytych_p_szumu	50	50	10
l_odkrytych_p_rdzeniowych	121	121	173
l_odkrytych_p_brzegowych	43	43	31
srednia_l_obl_odl	426	427	427
|TP|	965	965	2025
|TN|	15239	15239	13784
l_par_punktow	22791	22791	22791
RAND	0.710982	0.710982	0.693651

Uzyskany podział na grupy był inny dla algorytmu wykorzystującego przy ostatecznej weryfikacji odległości miarę kosinusową.

5.2.5. Iris

	bez Ti	Ti	Ti+cos
nazwa_pliku_wejsciowego	iris.csv	iris.csv	iris.csv
liczba_wym_punktow	4	4	4
l_punktow	150	150	150
k	3	3	3
czas_wczytania_pliku_wej [us]	2961.8	3066.8	3084.3
czas_normalizacji [us]	4.7	4.25	2.94
calkowity_czas_sortowania [us]	3431.6	28.3	31.6
czas_obliczania_kNN [us]	76373.9	23606.5	31628.8
czas_obliczania_RkNN [us]	76373.9	23606.5	31628.8
czas_grupowania [us]	104.8	147.2	184.1
czas_zapisywania_plikow_out_kNN [us]	9891.6	11442	12122.8
calkowity_czas_dzialania [us]	13239.5	15043.4	15754.9
calkowity_czas_dzialania-czas_zapisywania_plikow [us]	3347.9	3601.4	3632.1
l_odkrytych_grup	15	15	15
l_odkrytych_p_szumu	33	33	13
l_odkrytych_p_rdzeniowych	91	91	116
l_odkrytych_p_brzegowych	26	26	21
srednia_l_obl_odl	298	68.64	80.4
|TP|	834	834	1973
|TN|	7123	7123	7243
l_par_punktow	11175	11175	11175
RAND	0.712036	0.712036	0.824698

Uzyskany podział na grupy był inny dla algorytmu wykorzystującego przy ostatecznej weryfikacji odległości miarę kosinusową.

6. Wnioski

Podczas pisania algorytmu przyjęto k=3. Dlatego też dla własnego zbioru dla k=3 otrzymano wskaźnik RAND=1.

Przy wyznaczaniu k+NN bez wykorzystania nierówności trójkąta liczba wyliczeń odległości jest dla każdego punktu równa.

Im większa liczba k tym mniej jest mniej różnych grup.

Spośród sprawdzanych wartości k najlepsze wyniki uzyskuje się zazwyczaj dla małych wartości. Należy jednak pamiętać, że przy doborze najlepszej wartości k zależy wziąć pod uwagę zbiór danych.

Dla mojego, małego zbioru danych najbardziej czasochłonne jest samo wczytanie i zapisanie danych. Przy większych zbiorach najbardziej czasochłonne okazuje się wyznaczanie k+NN i |Rk+NN|. W związku z tym w celu optymalizacji działania algorytmu należałoby się skupić na tym aspekcie.

Podczas badania wpływu wartości k zauważono, że czas wczytania tego samego pliku jest różny dla każdego uruchomienia. Prawdopodobnie było to spowodowane różnym obciążeniem komputera, ponieważ w tle były uruchomione również inne programy.

W związku z tym w dalszej części przedstawiono uśrednione wyniki z 10 uruchomień kodu.

Na podstawie różnych zbiorów danych widać, że wykorzystanie nierówności trójkąta przyśpiesza czas wyznaczenia k+ najbliższych sąsiadów. Średnia liczba obliczeń odległości jest kilkukrotnie mniejsza. W każdym przypadku (poza zbiorem wykładowym, gdzie czasy są do siebie zbliżone) algorytm z miarą kosinusową do ostatecznej weryfikacji odległości działa trochę dłużej niż z odległością euklidesową. Należy również zauważyć, że daje on zazwyczaj inny wynik klasyfikacji. k+ sąsiedztwo wyznaczone tym algorytmem pokrywa się z k+ sąsiedztwem pozostałych algorytmów, z tym, że dla niektórych punktów jest ono powiększone o inne punkty. Najprawdopodobniej przez to algorytm wykonuje więcej obliczeń rzeczywistej odległości. Dodatkowo w tym algorytmie istnieje potrzeba przeliczania wartości Eps. Jednakże wyliczenie miary kosinusowej jest szybsze niż wyliczenie odległości euklidesowej.

Inny podział na klastry dla algorytmu z wykorzystaniem miary kosinusowej może wynikać z faktu, że miara kosinusowa w ogólności nie zachowuje nierówności trójkąta.

7. Bibliografia

7.1. Kryszkiewicz, M., Lasek, P. (2010). A Neighborhood-Based Clustering by Means of the Triangle Inequality. In Intelligent Data Engineering and Automated Learning – IDEAL 2010 (pp. 284–291). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-15381-5_35

7.2. Materiały wykładowe do przedmiotu Metody Ekploracji Danych, Kryszkiewicz, M.