GDAŃSK UNIVERSITY
OF TECHNOLOGY

Sztuczna Inteligencja

Wizualizacja działania algorytmu klasteryzacji K-means
w zależności od wybranej metryki mierzenia odległości.
[na podstawie samodzielnie zaimplementowanego algorytmu]

Maciej Adryan 175854

26 maja 2021

Streszczenie

Projekt ma na celu przedstawienie działania krok po kroku algorytmu klasteryzacji K-means w zależności od wybranej metryki mierzenia odległości.

Zaimplementowane zostały:

- metryka Chebysheva
- metryka Euklidesowa
- metryka Manhattan
- metryka Minkowskiego z $p = 3$

Ze względu na błąd popełniony na etapie projektowania rozwiązania, poszczególne klastry nie mają przypisane na stałe określonego koloru, co w pewnym stopniu obiża intuicyjność odczytywania wykresu, dlatego zwracam na to uwagę.

Dane początkowe dla każdego z algorytmów są identyczne.
Punkty oznaczone kolorem **czerwonym** symbolizują centroidy, punkty o pozostałych kolorach symbolizują losowo rozmieszczone punkty.
Celem zwiększenia czytelności wykresy pozbawione zostały legend, opisów osi oraz tytułów.

Symulacje przeprowadzono dla ziarna generatora = **175854** (Nr. Indeksu)
Ilość losowo rozmieszczonych punktów: **100**
Ilość centroid: **6**

Kod źródłowy wraz z wykresami znaleźć można na moim profilu na platformie **GitHub**:
LINK

1 Obliczenia

1.1 Metryka Euklidesowa

Wzór służący do obliczenia metryki:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

1.2 Metryka Chebysheva

Wzór służący do obliczenia metryki:

$$D(x,y) = max(|x_i - y_i|)$$

1.3 Metryka Manhattan

Wzór służący do obliczenia metryki:

$$D(x,y) = \sum_{i=1}^{k}|x_i - y_i|$$

1.4 Metryka Minkowskiego

Wzór służący do obliczenia metryki (w projekcie wykorzystałem $p = 3$):

$$D(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}$$

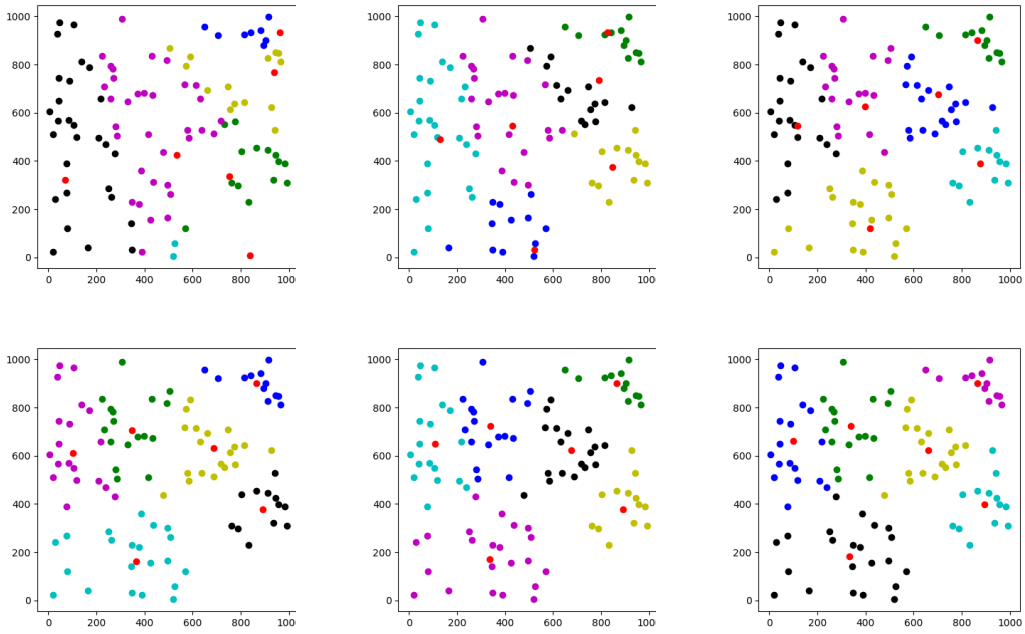
2 Wyniki

2.1 Metryka Euklidesowa

Optymalny wynik osiągniety został w **5** iteracjach

2.1.1 Przebieg algorytmu

Rysunek 1: pozycja początkowa i kolejne iteracje dla metryki Euklidesowej



2.1.2 Obserwacje

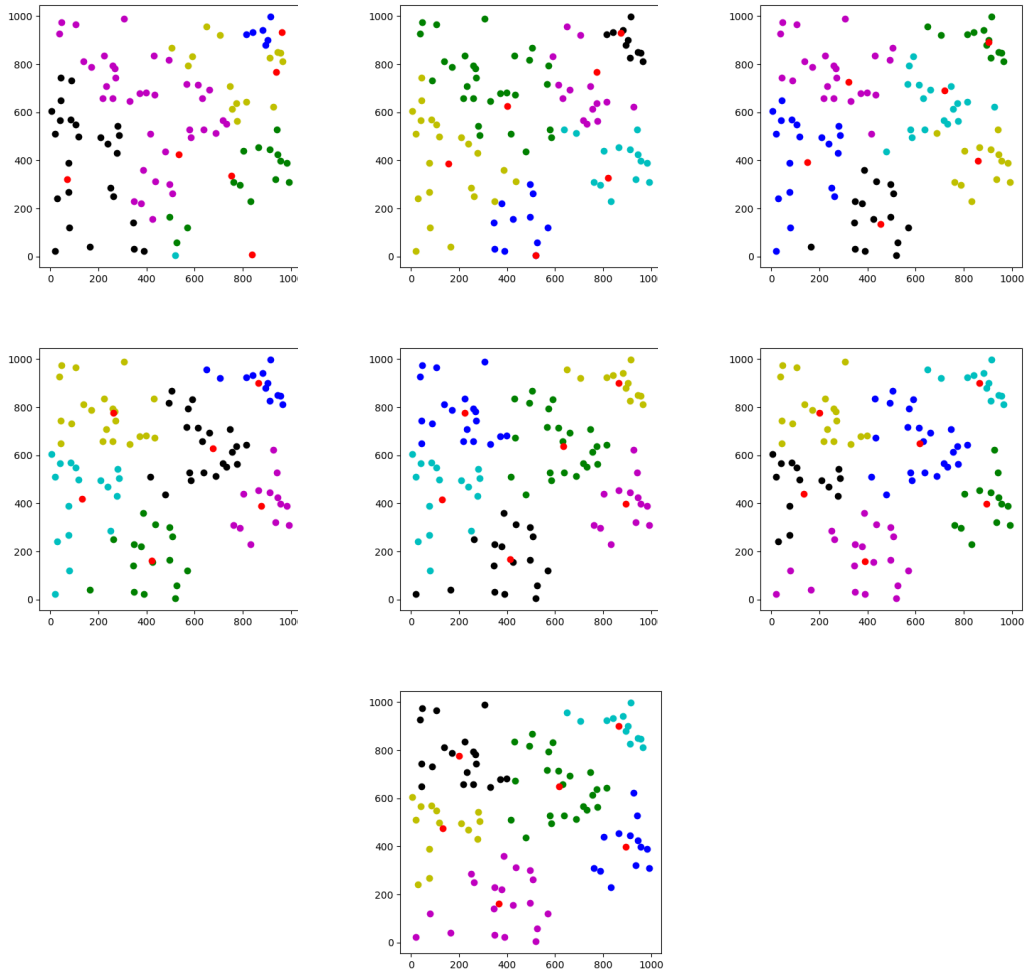
Algorytm w sposób optymalny podzielił zadane dane na odpowiednie klastry. Ze względu na swoją charakterystykę znajduje on zawsze “(”najbliższy) (w rozumieniu powszechnym) centroid. Wynik jego działania traktować będę jako punkt odniesienia dla kolejnych porównań.

2.2 Metryka Chebysheva

Optymalny wynik osiągniety został w **6** iteracjach

2.2.1 Przebieg algorytmu

Rysunek 2: pozycja początkowa i kolejne iteracje dla metryki Chebysheva



2.2.2 Obserwacje

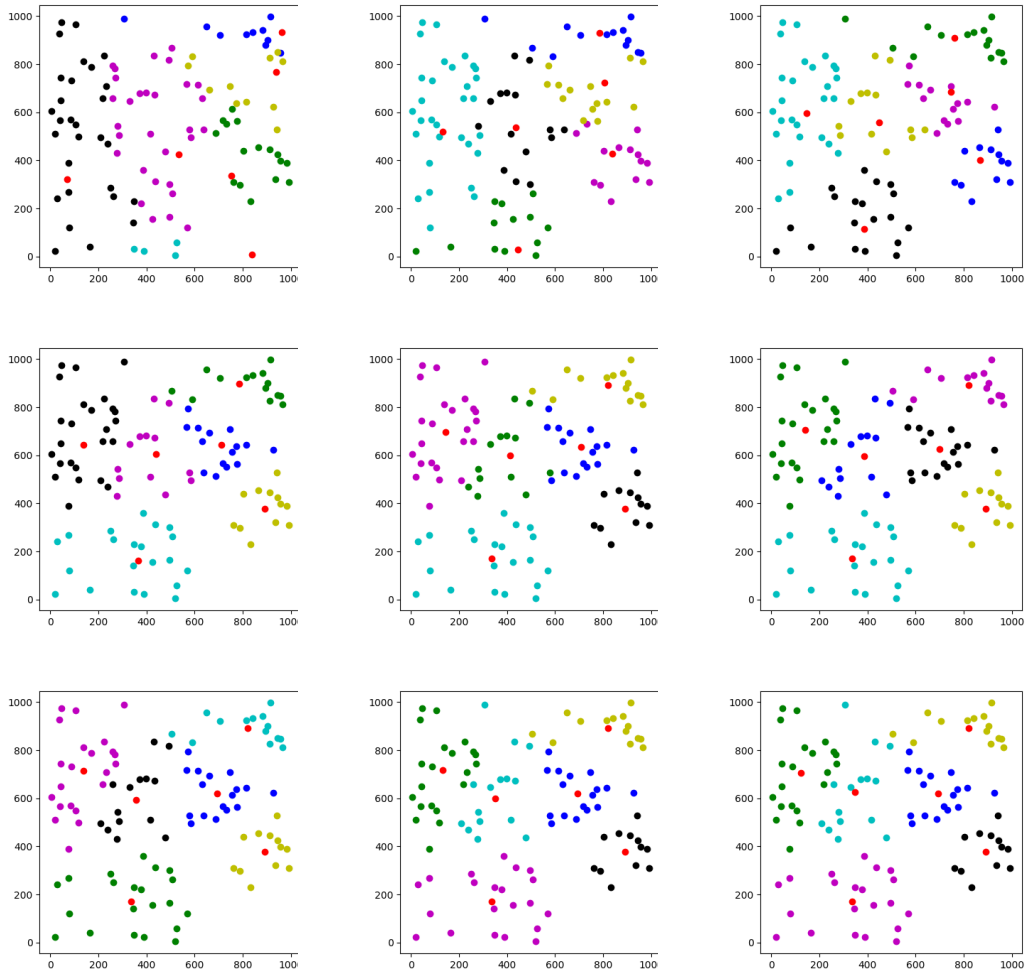
Ze względu na swoją charakterystykę wynik różni się od wyniku otrzymanego w wyniku użycia normy Euklidesowej. Nie jest to natomiast znaczna różnica, różnice w klasyfikacji występują jedynie na krańcach klastrów. Metryka ta bywa często stosowana w logistyce.

2.3 Metryka Manhattan

Optymalny wynik osiągniety został w **8** iteracjach

2.3.1 Przebieg algorytmu

Rysunek 3: pozycja początkowa i kolejne iteracje dla metryki Manhattan



2.3.2 Obserwacje

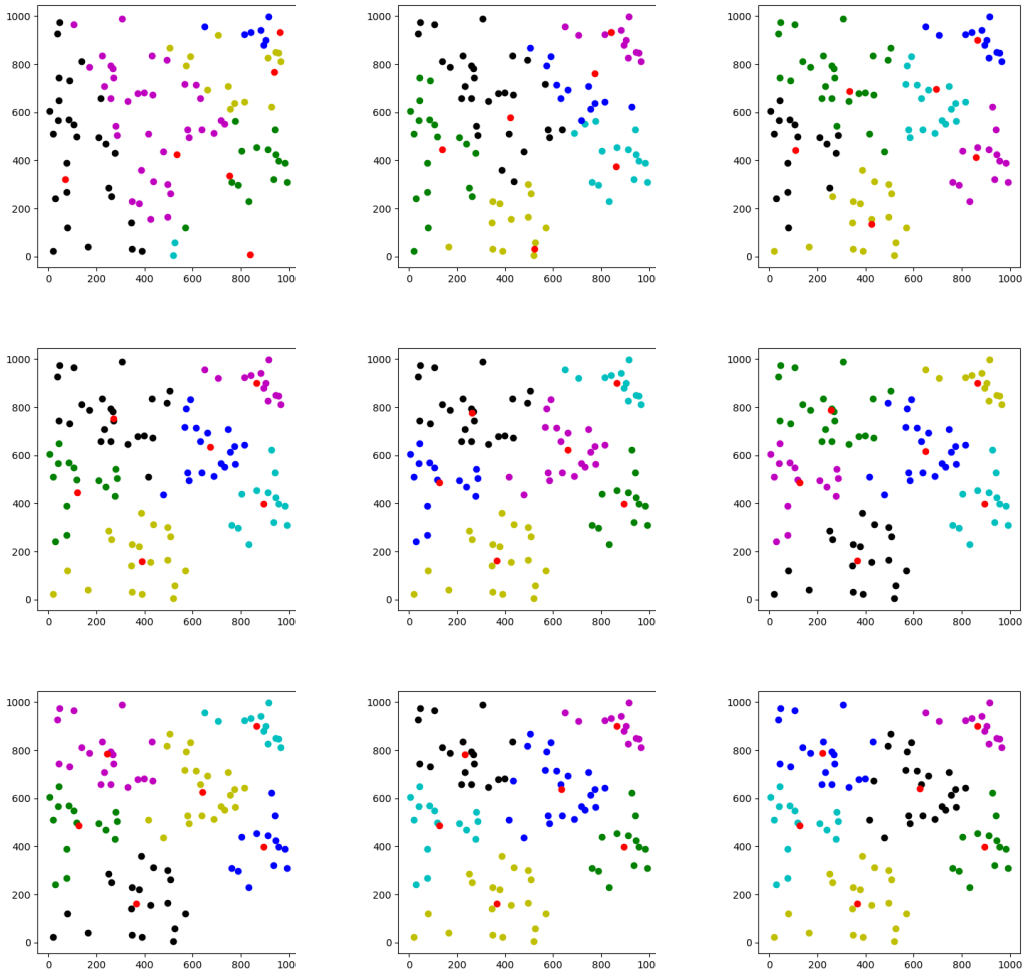
Ze względu na swoją charakterystykę norma Manhattan nie wydaje się być dobrą miarą podziału na klastry, jednak tylko pozornie, znajduje ona zastosowanie przy pracy na danych wielowymiarowych, jest jednak mniej intuicujna i nie odnajduje "najkrótszej drogi" (tu znów w myśleniu powszechnym). Metoda ta jest najbardziej kosztowną obliczeniowo operacją spośród dotychczasowej trójki.

2.4 Metryka Minkowskiego

Optymalny wynik osiągniety został w **8** iteracjach

2.4.1 Przebieg algorytmu

Rysunek 4: pozycja początkowa i kolejne iteracje dla metryki Minkowskiego



2.4.2 Obserwacje

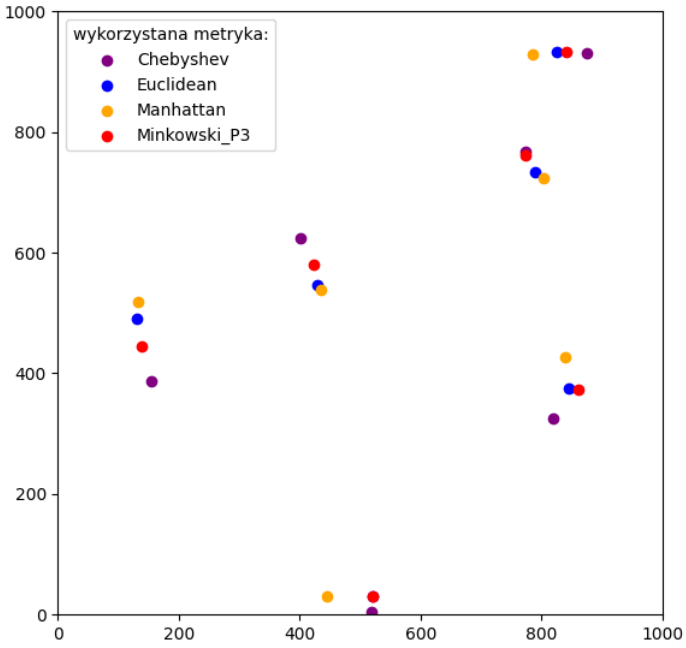
Metoda ta jest podobnie kosztowna obliczeniowo co metoda podziału na klastry przy kalkulacji odległości za pomocą normy Manhattan. Zaletą tej metody jest jej uniwersalność – w zalezności od doboru parametru p, otrzymujemy:

- $p = 1$ – metrykę Manhattan
- $p = 2$ – metrykę Euklidesową
- $p = \infty$ – metrykę Chebysheva

Parametr p dobierać można w zależności od potrzeb lub w drodze eksperymentalnej.

3 Wnioski

3.1 Porównanie rozmieszczenia punktów finalnych



Rysunek 5: Finalne rozmieszczenie centroid w zależności od użytej metryki

Na podstawie wykresu stwierdzić można, iż rodzaj wykorzystanej metryki nie ma szczególnie dużego wpływu na finalne rozmieszczenie centroid, nie jest on jednak zaniedbywalny.
Rodzaj metryki ma jednak wpływ na to do którego klastra zostaną zaliczone poszczególne punkty na płaszczyźnie. Dobór odpowiedniej metryki może zależeć od wielu czynników, m.in.:

- zastosowanie
- rodzaj danych
- wymiarowość danych (metoda Euklidesa nie nadaje do zastosowania przy danych wielowymiarowych)

4 Źródła

9 Distance Measures in Data Science by Maarten Grootendorst
K-means explanation from javatpoint.com