Testy Metody spectral_clustering na własnych zbiorach

Tomasz Suchodolski

13 maja 2019

Wstęp teoretyczny

Metoda **spectral_clustering(X, k, M)** jest algorytmem spektralnym analizy skupień - zaimplementowanym przeze mnie zgodnie opisem algorytmeu przedstawionym w poleceniu zadania. Jej implementacja składa się z:

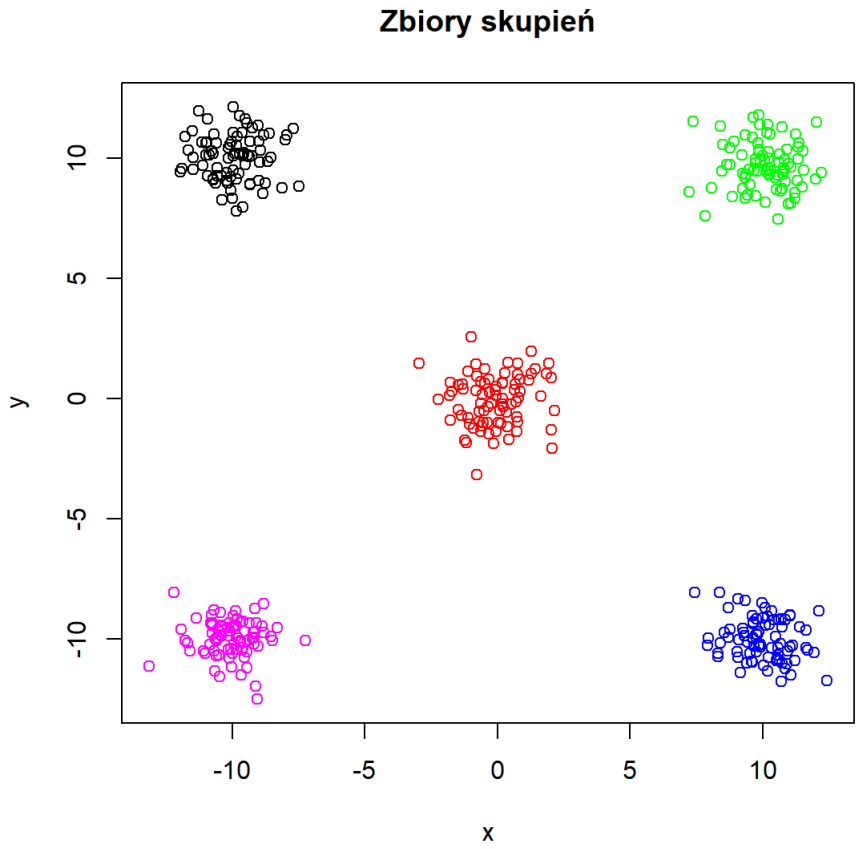
1. **funkcji Mnn(X, M)**, która dla macierzy **X** będącej reprezentacją wierszową n wektorów wyznacza macierz **S** taką, że **S[i,j]** jest indeksem j-tego najbliższego sąsiada wektora odpowiadającego i-temu wierszowi w metryce Euklidesowej.
2. funkcja **Mnn_graph(S)**, której argumentem jest macierz wygenerowana w pkt. 1. Wylicza ona macierz **G** o zbiorze wartości elementów = {0,1}, taką, że **G[i,j]** = 1 <=> Istnieje u takie, że **S[i,u]**=j lub **S[u,j]**=i. Ponadto jeśli Graf reprezentowany przez macierz G jako jego macierz sąsiedztwa nie jest spójny to jest on “uspójaniany” w taki sposób, że pierwszy wierzchołek z pierwszej składowej jest łączony z pierwszymi wierzchołkami z pozostałych składowych.
3. funkcji **Laplacian_eigen(G, k)**, która:
    - wyznacza laplasjan grafu **L=D-G**, gdzie **D** jest macierzą diagonalną taką, że **D[i,i]** = p, gdzie p jest stopniem i-tego wierzchołka w grafie reprezentowanym przez macierz **G**.
    - zwraca jako wynik macierz **E** składającą się z kolumnowo zapisanych wektorów własnych odpowiadających 2, 3,… (k+1) wartości własnej **L**.
4. wyznaczenia skupień dla macierzy **E** z poprzedniego podpunktu za pomocą metody k-średnich.

Zbiory testowe

Przygotowałem 3 testowe zboiory w przestrzeni 2 wymiarowej, za pomocą których mam zamiar przetestować poprawność napisanego agorytmu.

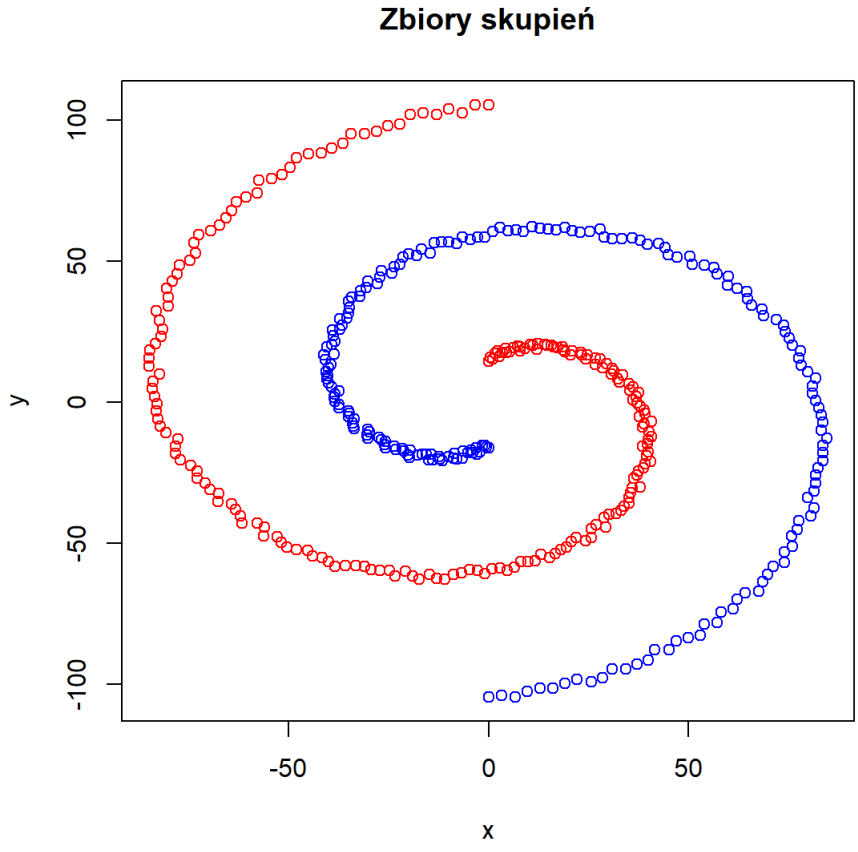
zbiór 1

Pierwszy zbiór składa się ze skupień będących w okolicy pięciu punktów o współrzędnych kolejno: **(-10,-10), (-10,10), (0,0), (10,-10), (10,10)**. Okolica każdego z punktów składa się z 80 punktów, których współrzędne zostały wylosowane za pomocą funkcji **rnorm(p,1)**, gdzie p jest daną “średnią” współrzędną punktu. Rozkłady współrzędnych pochodzą zatem z rozkładu normalnego z odchyleniem standardowym równym 1.



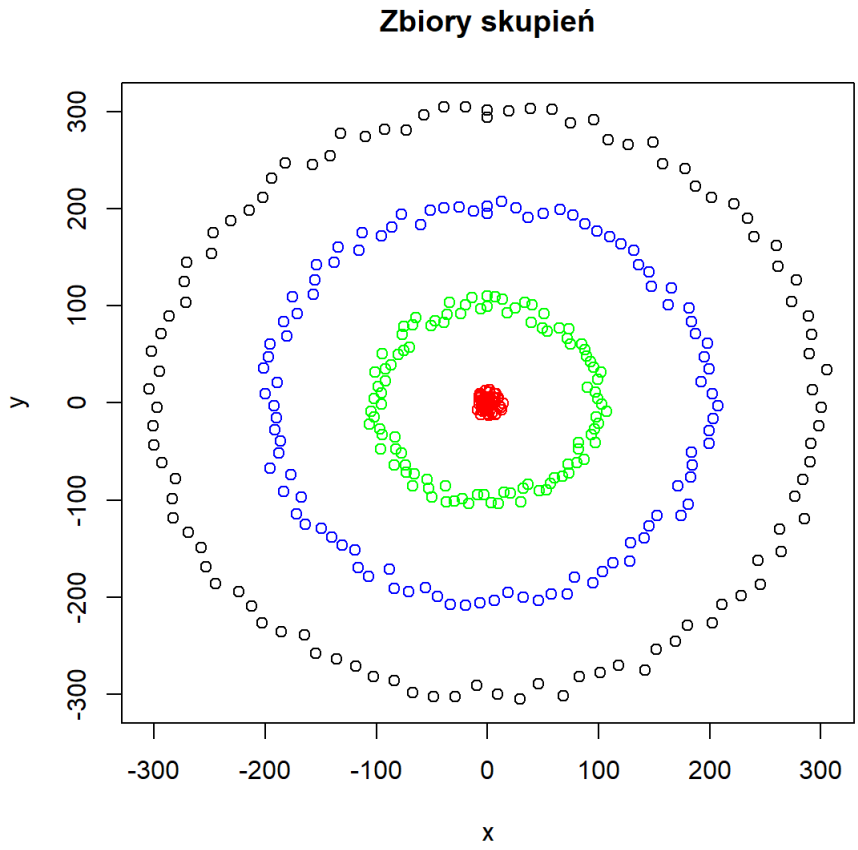
zbiór 2

Zbiór 2 składa się z dwóch spiral w pobliżu, krórych rozmieszczone są punkty należące do skupień. Poprawny algorytm powinien znaleźć w przypadku wyszukiwania dwóch skupień obie spirale jako oddzielne klasy.



zbiór 3

Zbiór 3 składa się z punktów rozlosowanych w okolicy czterech okręgów umieszczonych jeden w drugim. Celem algorytmu jest podział punktów na te cztery okręgi.

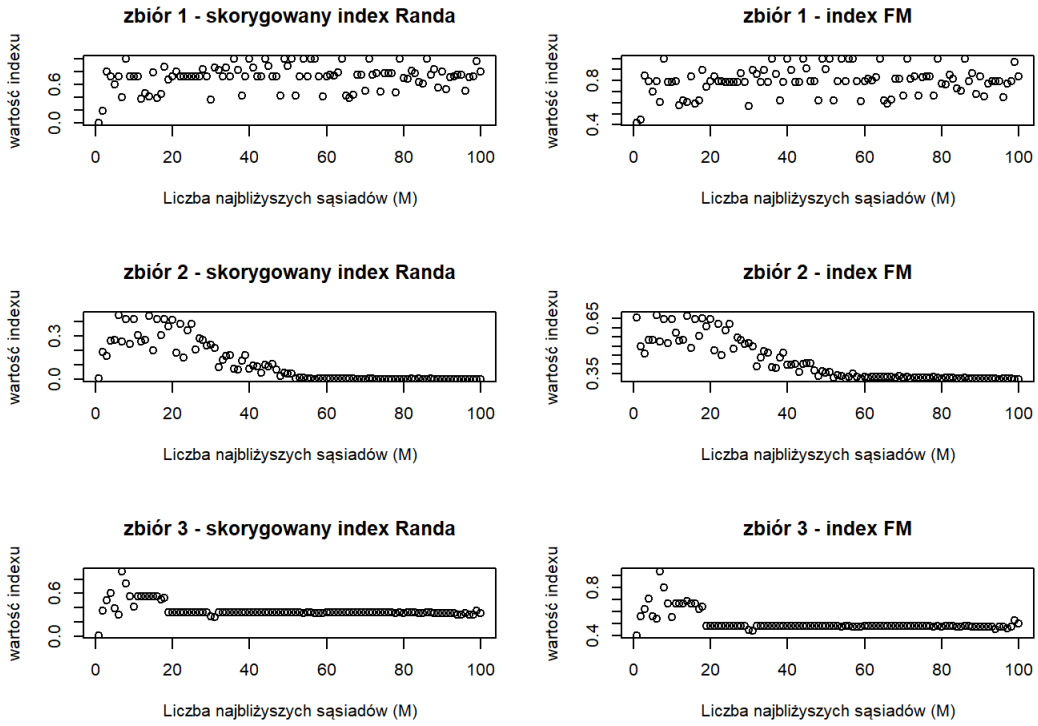


Obliczenie wyników

Dla każdego z trzech zbiorów wywołałem alogrytm 100 razy dla parametru **M** zmieniającego się w zakresie od 0 do 100. Wyniki porównuję do wzorcowych przy pomocy *indeksu Fowlkesa–Mallowsa* oraz *skorygowanego indexu Randa*. Poniżej prezentuję analizę otrzymanych wyników.

Wykres ogólny

Pierwszyni wykresami jakimi przedstawie będą wartości dwóch powyżej wymienionych indexów w zależności od liczby sąsiadów dla testowych zbiorów danych:



Wnioski:

1. Porównując rozkłady zbiorów z danymi z wykresów można dojść do wniosku (raczej niezbyt zaskakującego), że najlepsze dane (indexy bliskie wartości 1) otrzymujemy dla największej liczby sąsiadów takiej, że dla dowolnego punktu z danej klasy wszyscy jego sąsiedzi należą do danej klasy.
2. Zwłaszcza po drugim i trzecim rozkładzie widać, że wrost liczby sąsiadów nie zawsze oznacza lepszy wynik, co więcej może doprowadzać do pogorszenia wyniku a nawet (zbiór 2) do uzyskania podziału zbliżonego do podziału losowego (wartości indexów w okilicach zera).
3. Najlepsze wyniki dla tej metody otrzymamy gdy będziemy znali jaką największą liczbę sąsiadów podawać, tak aby wśród sąsiadów *"nie łapały się"* punkty z innych klas podziału.

z wykresów można empirycznie stwierdzić, że najlepsze dane dla danych zbiorów otrzymujemy dla parametru M w zakresie od 5 do 15. Poniżej zostały zaprezentowane wartości podstawowych parametrów rozkładu dla takich danych.

```
##                       średnia odchylenie standardowe   mediana
## zbior 1 index Randa 0.6290185             0.19735509 0.7174032
## zbior 1 FM index    0.7372111             0.12902136 0.7876743
## zbior 2 index Randa 0.3231835             0.09075722 0.2720185
## zbior 2 FM index    0.5747185             0.06832464 0.5330148
## zbior 3 index Randa 0.5554538             0.16558840 0.5616776
## zbior 3 FM index    0.6723702             0.11376520 0.6646494
```

Z wykresów można też odczytać, że rozkładem, z którym algorytm poradził sobie najgorzej był rozkład z dwiema spiralami zaś najlepiej sobie poradził z rozkładem pierwszym, czyli z czymś co najbardziej przypomina intuicyjnie rozumienie pojęcia *"skupienie"*.