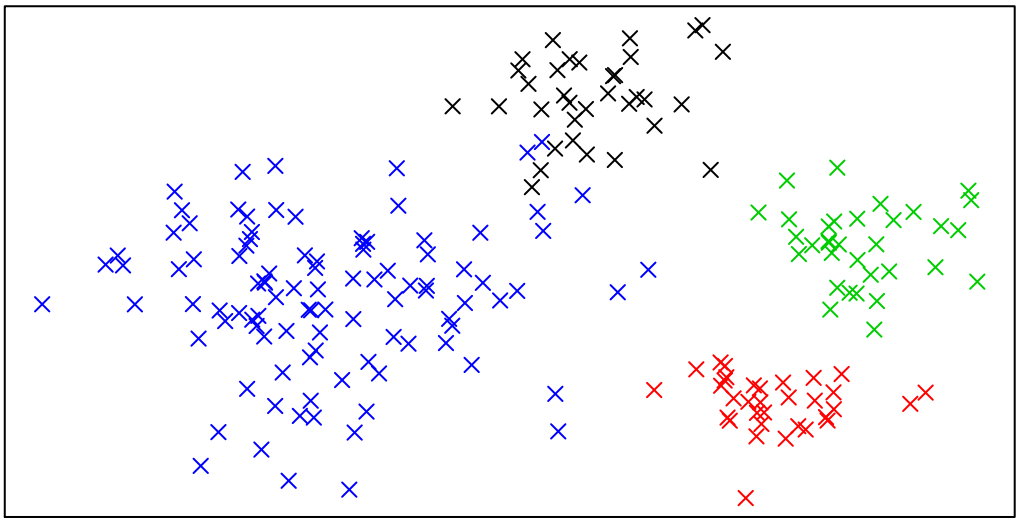
Testy funkcji `spectral_clustering`

Do testowania wykorzystałem zbiory, które wygenerowałem na czwartą pracę domową z PADPY. Porównam pobieżnie wyniki otrzymane przez implementację w Pythonie z tą z R dla takich samych parametrów.

Zbiór nr 1 – *rozległe skupienia*

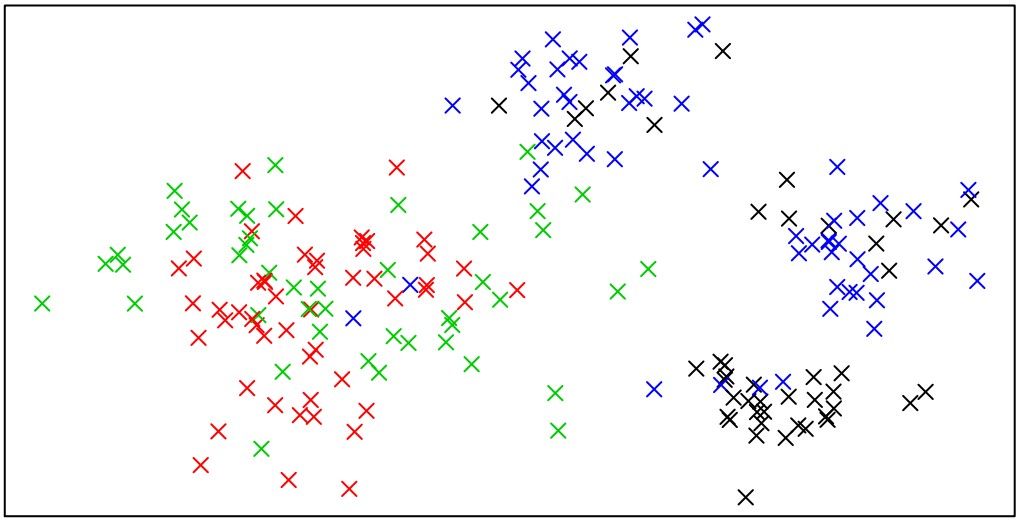
Wygląd zbioru

Zbiór składa się z czterech skupień delikatnie „zahaczających" o siebie. Dodatkowo jeden ze zbiorów jest mocniej rozproszony niż pozostałe.



Testy algorytmu

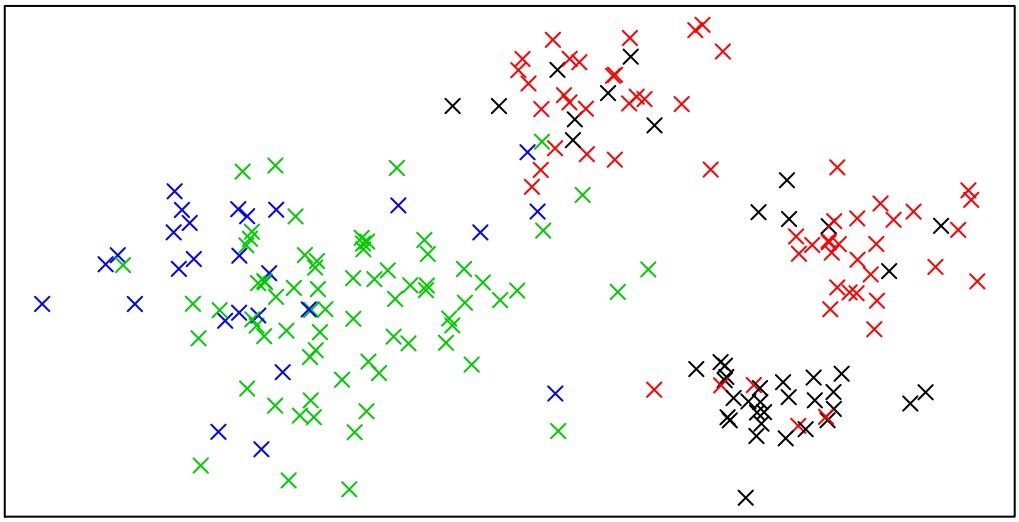
M=5



FM: 0.597735884534735

AR: 0.430131062583777

M=15



FM: 0.680315012510301

AR: 0.539042659894577

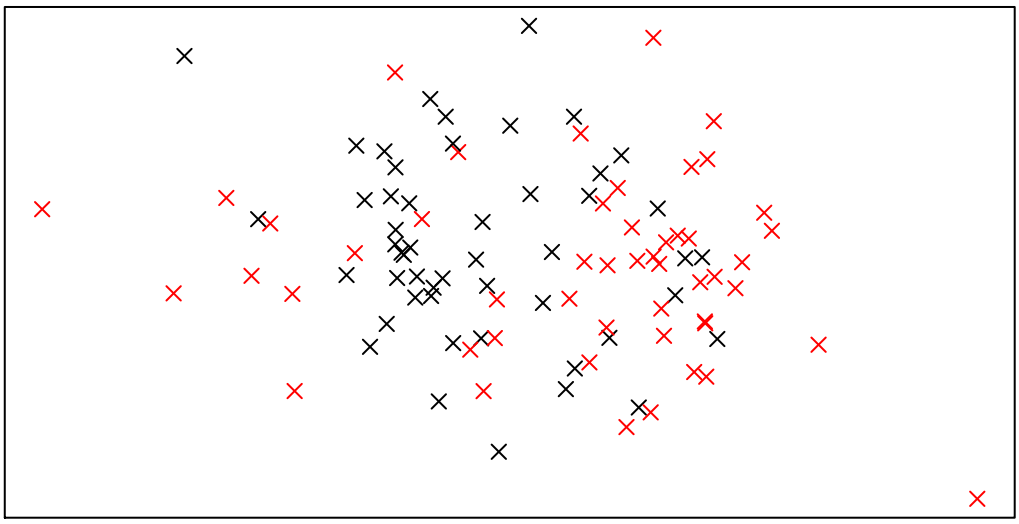
Interpretacja testu

Widać, że algorytm w tym wydaniu nie radzi sobie najlepiej – klasyfikacja jest dość chaotyczna, a zwiększenie liczby sąsiadów niekoniecznie pomaga.

Zbiór nr 2 – splecione zbiory

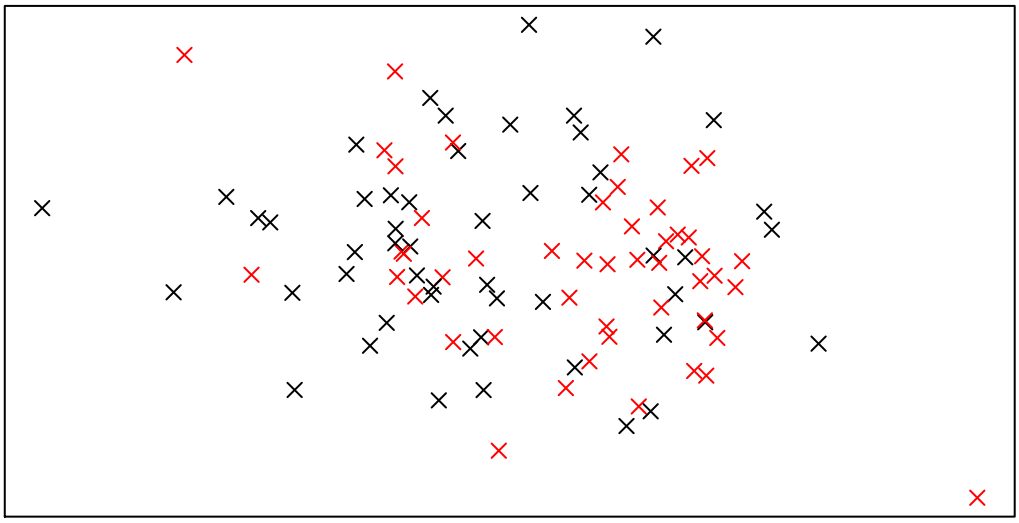
Wygląd zbioru

Zbiór składa się z dwóch prawie w pełni pokrywających się zbiorów znaczników.



Testy algorytmu

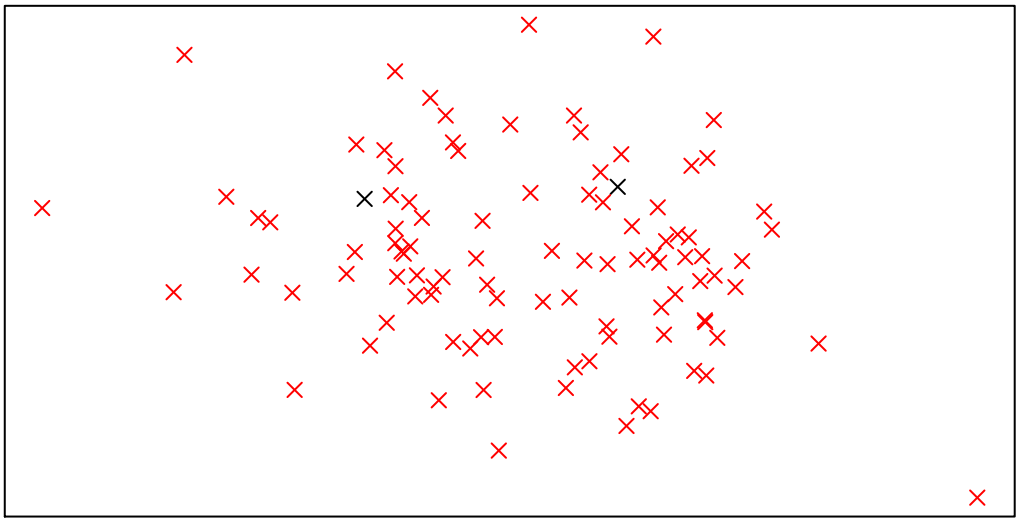
M=3



FM: 0.503262810151614

AR: 0.0156732134985552

M=85



FM: 0.689167337943901

AR: -0.000792546824960259

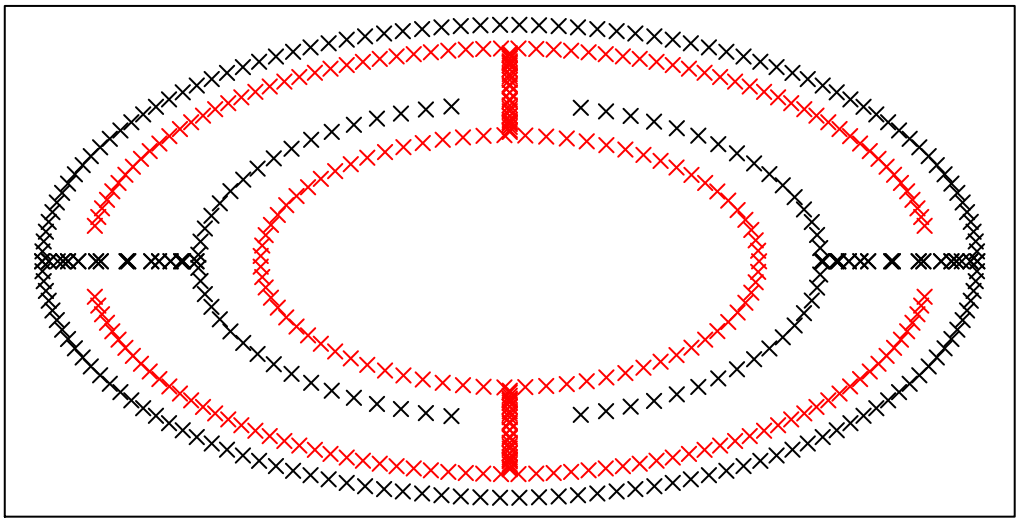
Interpretacja testu

W porównaniu do testów na zbiorze nr 3 algorytm zaskakująco dobrze radzi, gdy skupienia są na siebie bardzo nałożone. Przykład dla $M = 85$ bardzo dobrze pokazuje, że mierzenie dokładności algorytmu tylko jednym indeksem, bez weryfikacji z innymi jest nie najlepszym pomysłem.

Zbiór nr 3 – crosshair

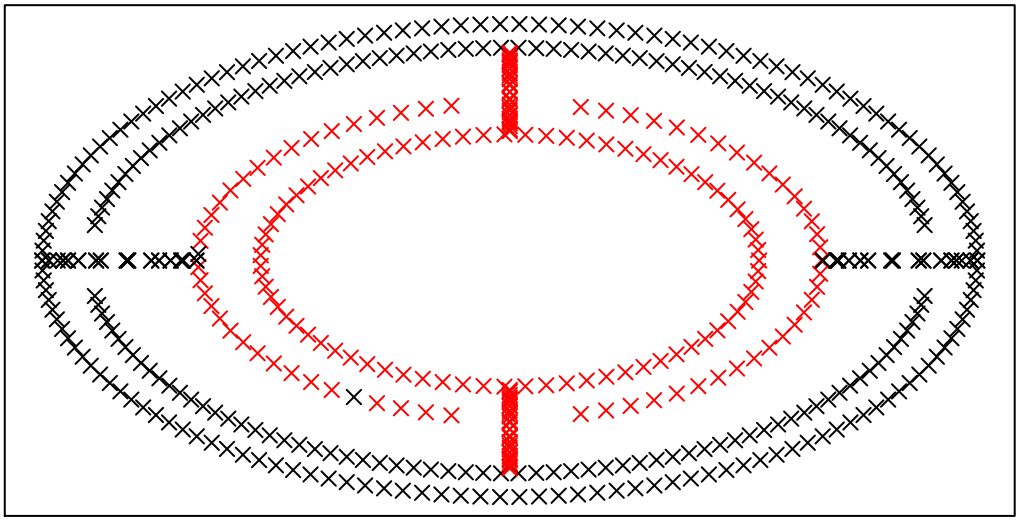
Wygląd zbioru

Wygląd zbioru można porównać do celownika z gier komputerowych. Docelowo zbiór miał składać się z poprzeplatanych różnych skupień, które są blisko siebie. Dodatkowo te same skupienia są połączone ze sobą.



Testy algorytmu

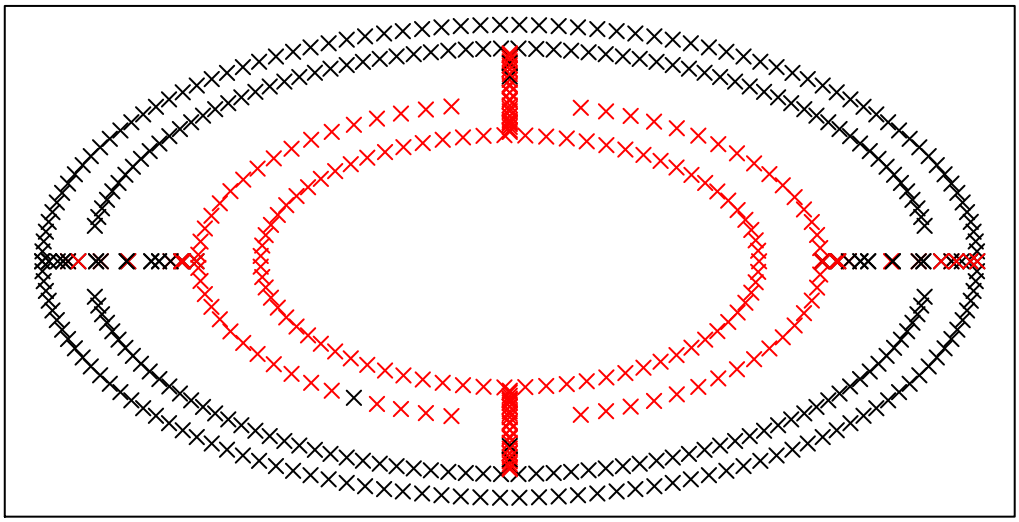
M=3



FM: 0.550619792755094

AR: 0.0693962618172412

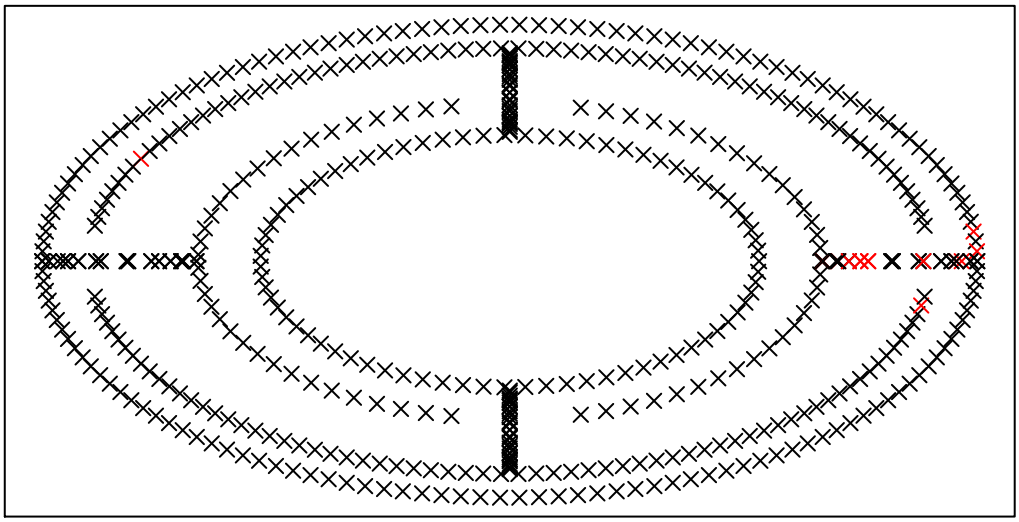
M=85



FM: 0.524054352282709

AR: 0.0289857270200506

M=500



FM: 0.688788784044259

AR: 0.000362811982642558

Interpretacja testu

Widzimy, że zmiana liczby najbliższych sąsiadów M może negatywnie wpływać na wynik działania algorytmu. Jedyny efekt jest taki, że dla $M = 500$ indeks FM wzrósł, aczkolwiek etykiety nadal nie były zbyt dobrze dopasowane. Przypuszczam, że algorytm priorytetyzuje odległości między punktami w rozpoznawaniu skupień, przez co otrzymany wynik jest taki, a nie inny.