[PADR] Praca domowa nr 4 – raport

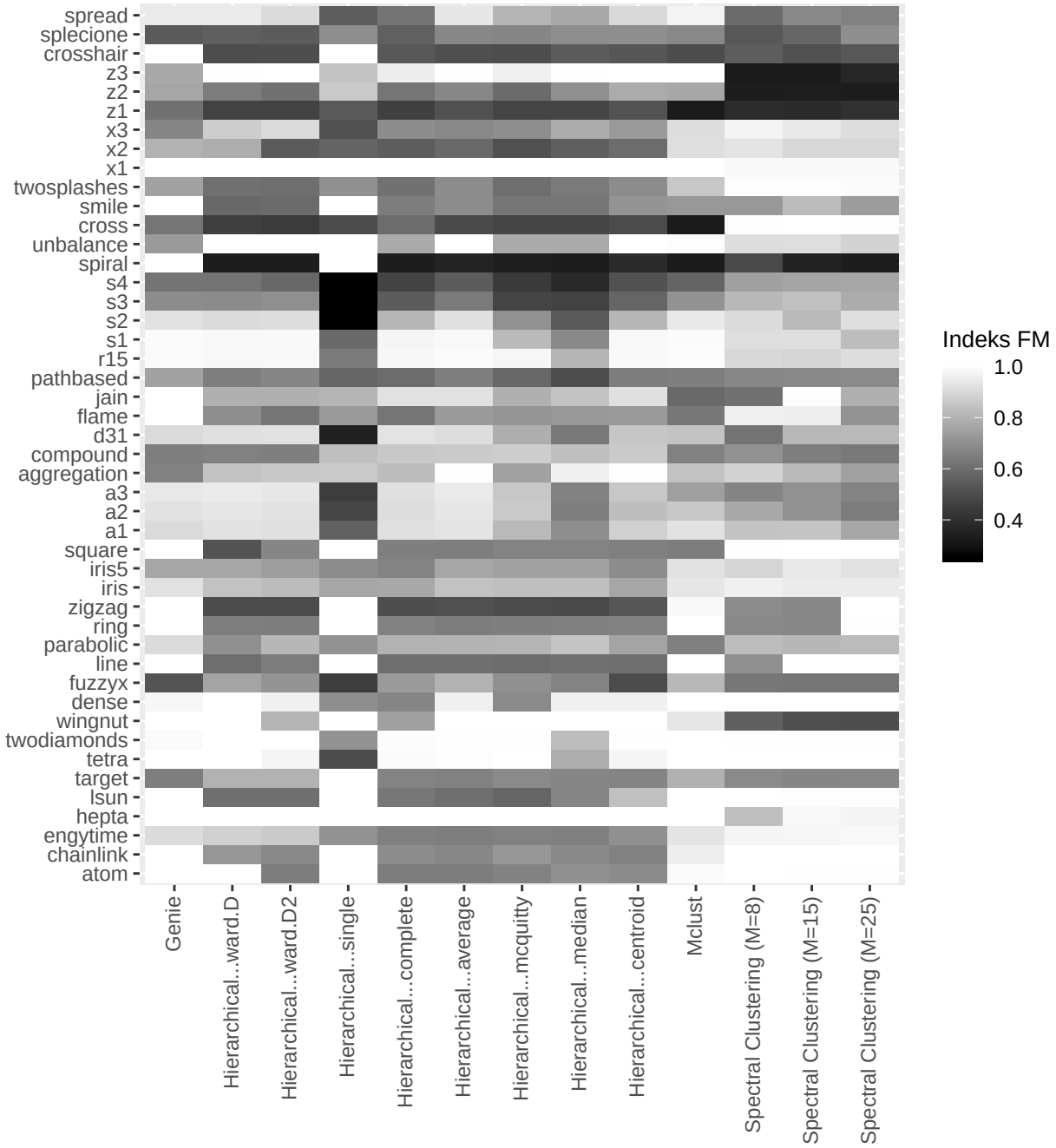
Wstęp

W niniejszym raporcie analizuję działanie własnej implementacji spektralnego algorytmu analizy skupień. Algorytm ten uruchomiłem z trzema różnymi wartościami parametru M, czyli liczby najbliższych sąsiadów, to jest 8, 15 oraz 25 i będę porównywał z szeregiem innych algorytmów analizy skupień.

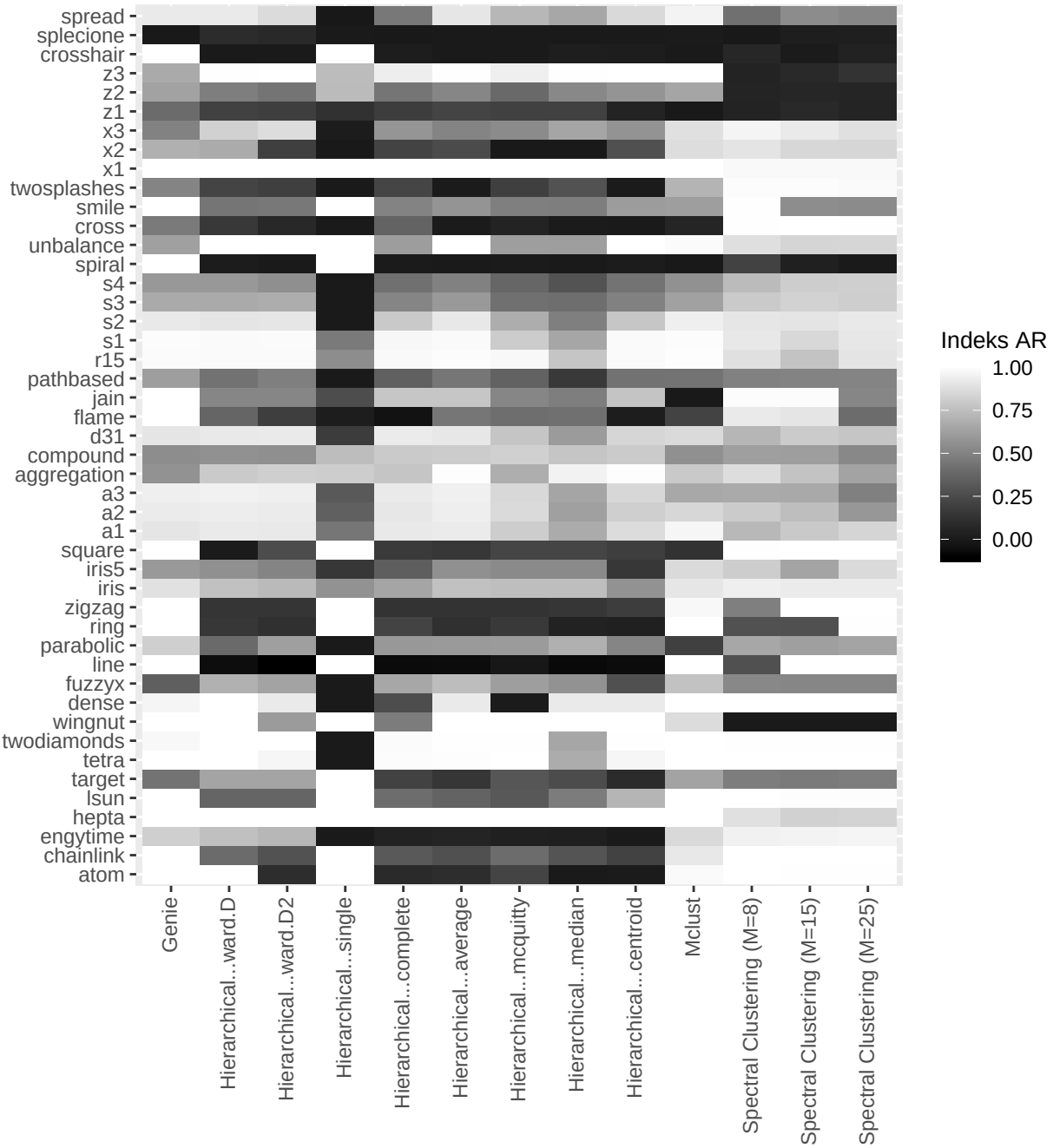
Z początku przeanalizuję skuteczność wyznaczoną przez indeks Fowlkesa-Mallowsa i skorygowany indeks Randa każdego z algorytmów. Następnie zbadam wpływ standaryzacji danych na skuteczności tych metod.

Wyniki w podziale na zbiór i algorytm

Indeks Fowlkesa–Mallowsa



Skorygowany indeks Randa



Interpretacja

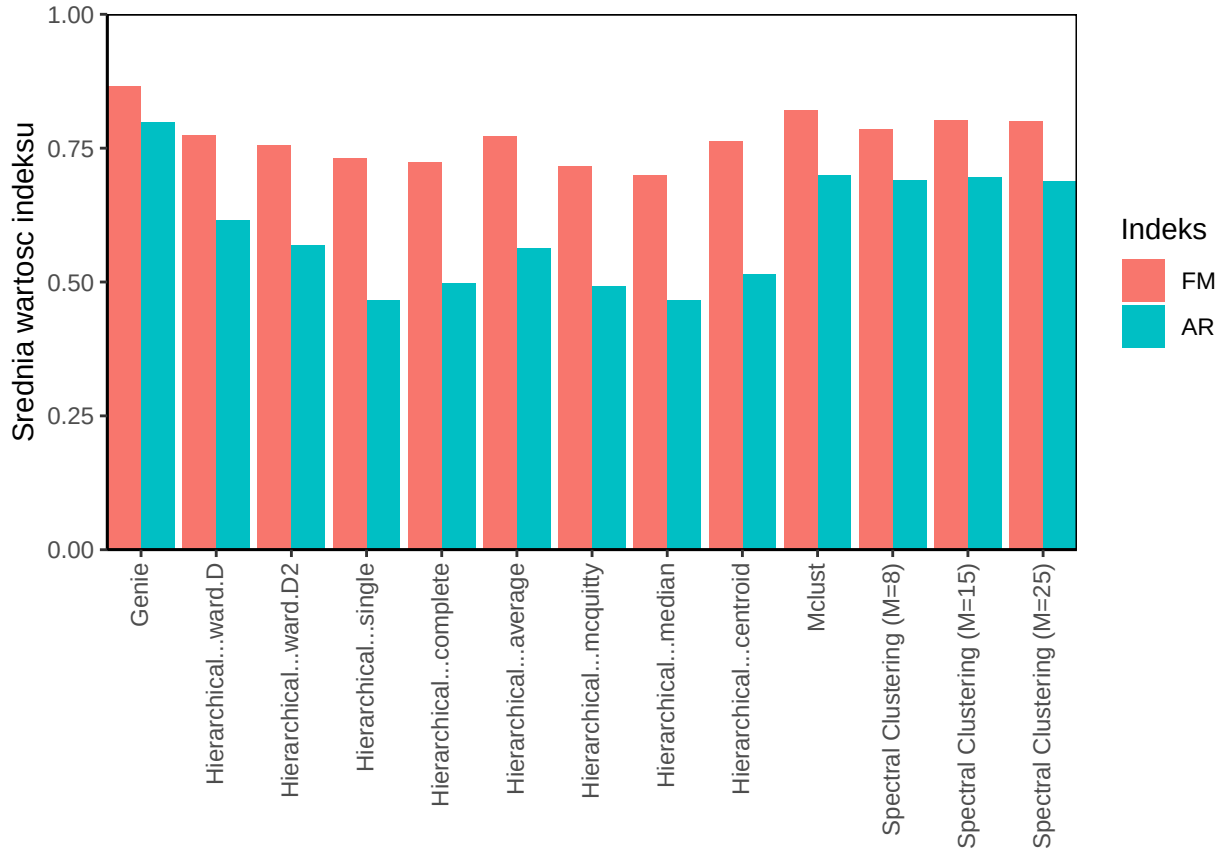
Pierwsze co rzuca się w oczy po spojrzeniu na powyższe mapy ciepła, to fakt, że indeks Fowlkesa-Mallowsa bardziej wybacza błędy, ponieważ daje wyższe noty.

Drugą widoczną rzeczą jest to, że niektóre zbiory są „prostsze" od innych, to znaczy każdy algorytm działa na takim zbiorze dość dobrze. Do zbiorów „prostych" możemy zaliczyć zbiory takie jak: `hepta`, `tetra`, `twodiamonds`, `r15` i `x1`.

Ponadto patrząc na słupki odpowiadające danym metodom widzimy, że najgorzej radził sobie hierarchiczny algorytm skupień z metodą „single" – w kolumnie tego algorytmu widnieje najwięcej czarnych kratek.

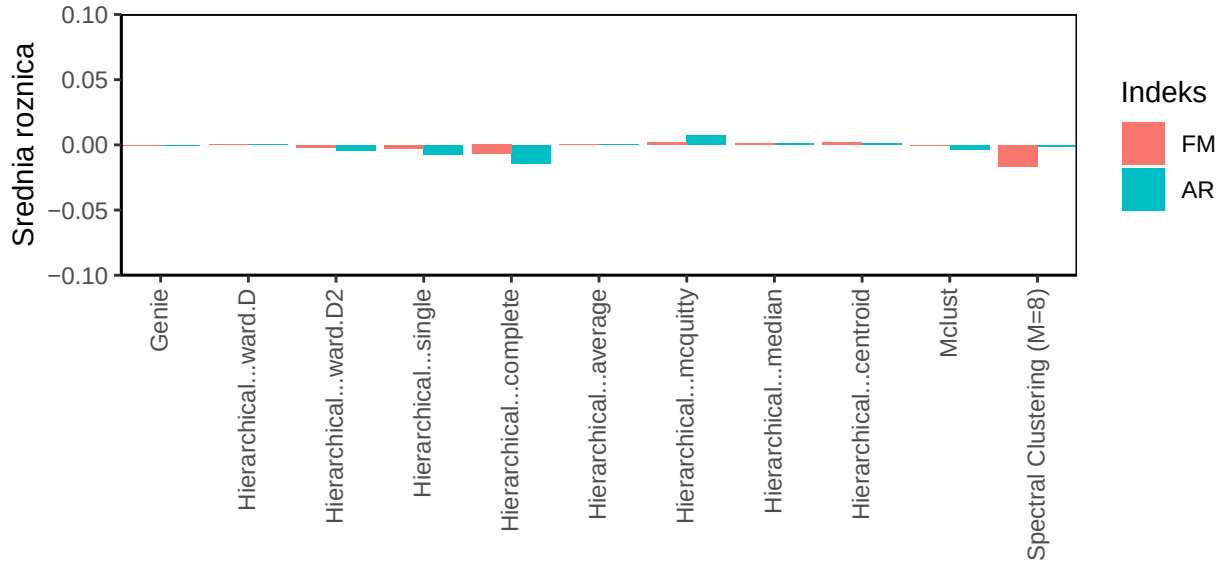
Warto również zwrócić uwagę na trzy kolumny odpowiadające mojej implementacji algorytmu `Spectral Clustering` – widoczne są różnice w zależności od doboru parametru najbliższych sąsiadów M.

Analiza średnich wyników algorytmów



Powyższy wykres obrazuje średnią skuteczność każdej z metod. Widzimy, że na prowadzeniu jest algorytm z pakietu `Genie` a na drugim miejscu z dość zbliżonymi wynikami algorytm `Mclust` i `Spectral Clustering`. Jednocześnie widać, że zwiększenie parametru liczby sąsiadów M z 8 na większe wartości nieco podniosło skuteczność mojej implementacji.

Wpływ standaryzacji na wyniki



Powyższy wykres pokazuje nam, że w zasadzie żaden z testowanych algorytmów nie zyskuje na skuteczności przy standaryzacji zmiennych. Największy zysk jest rzędu 10^{-3}.

Podsumowanie

Algorytm spektralny analizy skupień *(Spectral Clustering)* okazał się być dość dobrym algorytmem. Znajduje się na podium, jeśli chodzi o średnie wartości indeksów. Jednakże jego przewaga nad konkurencją nie jest tak duża jak była w przypadku implementacji w Pythonie.

W przeciwieństwie do implementacji z Pythona, w tej wersji modyfikacja parametru M miała znaczenie. W przypadku moich testów zwiększenie tego parametru podniosło skuteczność algorytmu. Jednakże już na etapie testów zauważyłem, że przesadne zwiększenie tego parametru nie sprzyja wynikom predykcji. Z moich obserwacji wynika, że algorytm najlepiej działa dla „rozsądnych" liczb najbliższych sąsiadów