Raport - praca domowa nr 2

Michał Wdowski

Wprowadzenie

Tematyką drugiej pracy domowej z przedmiotu *"Przetwarzanie Danych Ustruktyruzowanych"* była analiza skupień *(spectral clustering)*. Jest to problem, który polega na automatycznym umieszczeniu punktów w grupach. Poniższy raport stanowi porównanie 11 metod podejścia do tego problemu w języku `R`:

- algorytm *genie* z pakietu `Genie`;
- wszystkie algorytmy hierarchiczne z funkcji `hclust()`;
- własna implementacja algorytmu spektralnego z wykorzystaniem `kmeans()`, czyli `spectral_clustering()` (omówiona szczegółowo w pliku `testy.pdf`);
- funkcja `cmeans()` z pakietu `e1071`.

Poprawność działania będzie mierzona za pomocą dwóch indeksów, które porównują dwa zestawy podziałów, które zwracają wynik: od 0 w najgorszym wypadku, do 1 przy idealnym dopasowaniu do wzorca:

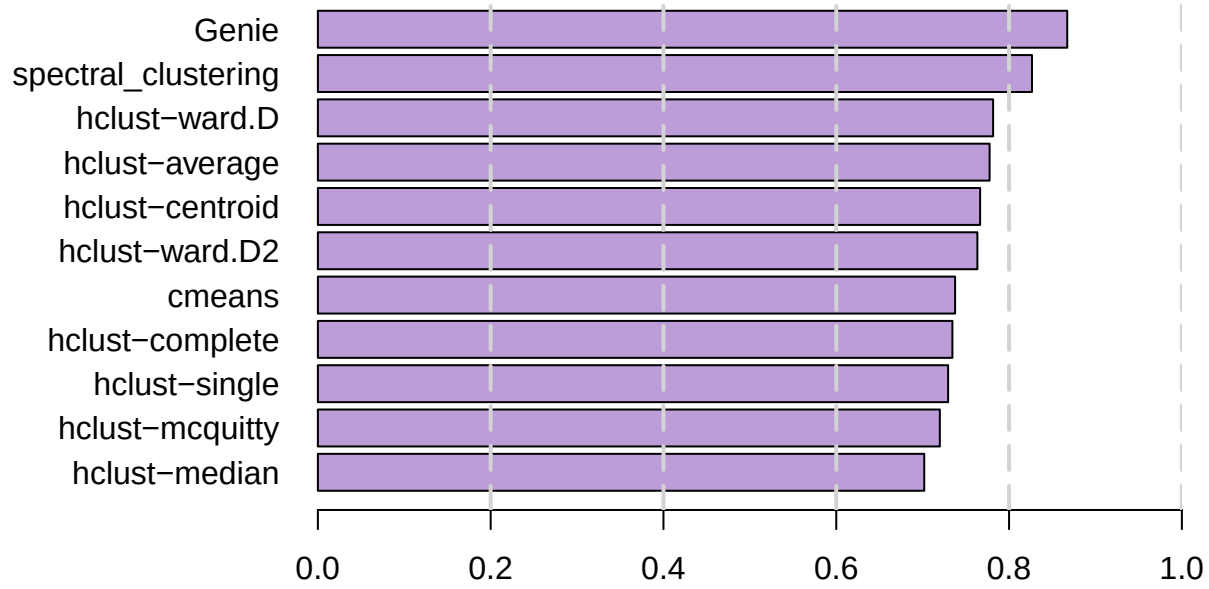
- indeks Fowlkesa–Mallowsa (**FM**)
- skorygowany indeks Randa (**AR**)

Testowe zestawy zadań składały się z 43 zbiorów, w tym 3 autorskich (również opisanych w pliku `testy.pdf`). Każdy z nich zawierał plik `.data`, który był zbiorem punktów z przestrzeni $\mathbb{R}^2$, $\mathbb{R}^3$ lub $\mathbb{R}^4$, zapisanych jako współrzędne, oraz plik `.labels0`, zawierający liczby naturalne, które dla i-tego wiersza oznaczały przynależność punktu w i-tym wierszu w pliku `.data` do danej grupy.

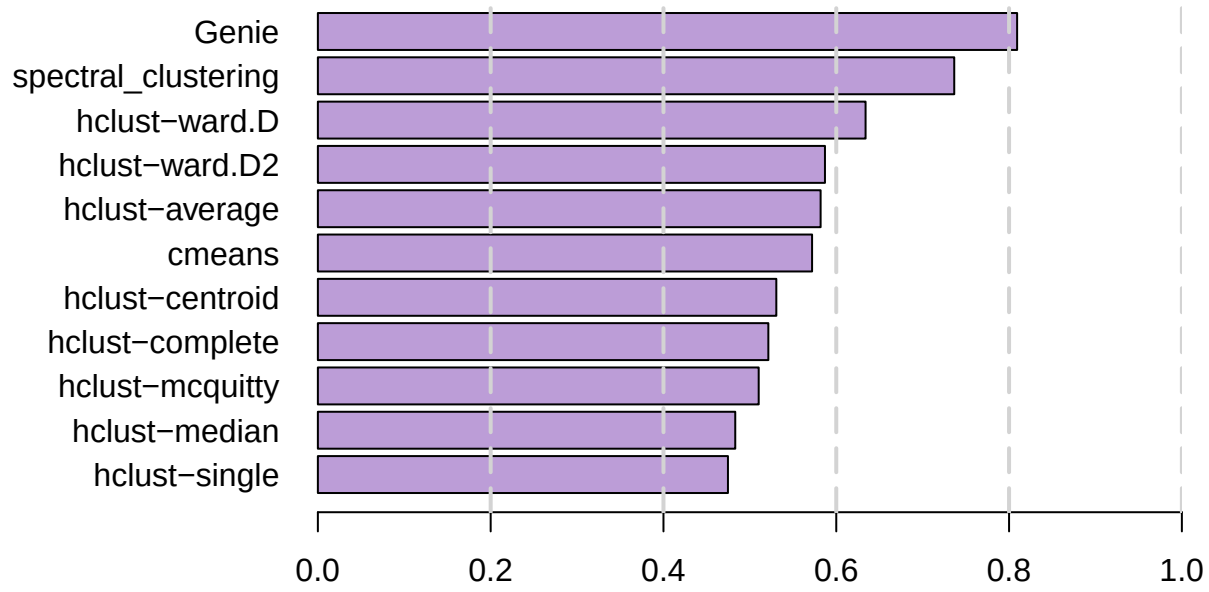
Oprócz badania surowych danych z opisanych powyżej 43 zbiorów, algorytmy zostały też sprawdzone na zbiorach po standaryzacji funkcją `scale`. Ma to sprawdzić wpływ standaryzacji na jakość otrzymanych wyników.

Wyniki

Średnia indeksów **FM** dla każdego algorytmu prezentuje się następująco:

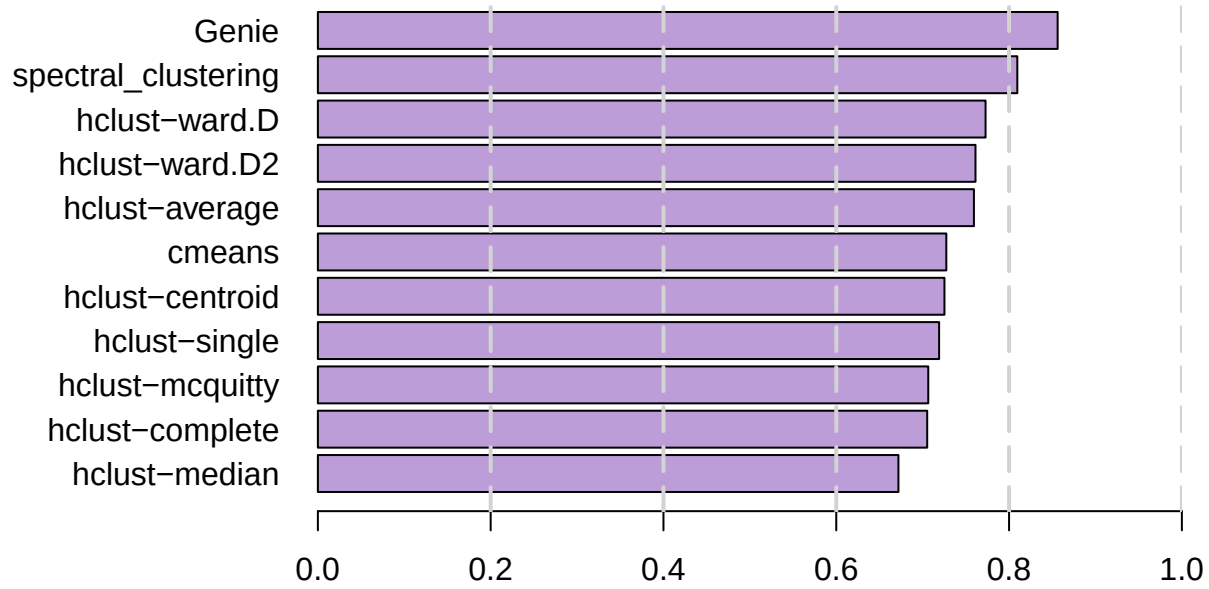


Z kolei średnia indeksów **AR** dla każdego algorytmu wygląda tak:

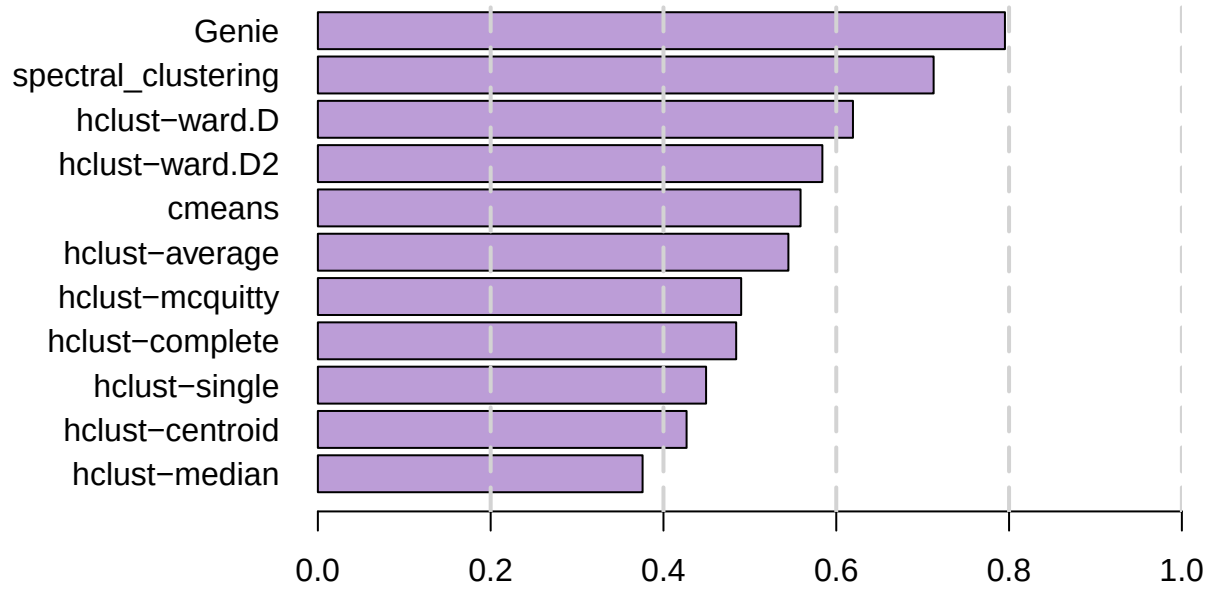


Wyniki (standaryzowane)

Średnia indeksów **FM** dla każdego algorytmu prezentuje się następująco:



Z kolei średnia indeksów **AR** dla każdego algorytmu wygląda tak:



Podsumowanie

Testy wskazują na to, że w dziedzinie analizy spektralnej niepodzielnie panuje algorytm *genie*. Jego średnia indeksów w każdym wypadku była najlepszą ze wszystkich w zestawieniu.

Wbrew moim oczekiwaniom, funkcja `spectral_clustering`, której napisanie było inną częścią tej pracy domowej, plasuje się na drugim miejscu w każdym rankingu. Spodziewałem się, że coś napisanego przez studenta pierwszego roku będzie czymś nieudolnie udającym porządny algorytm, ale miło się zaskoczyłem.

Najlepszym z algorytmów z rodziny `hclust()` jest `ward.D`, a najgorszy to `median`, który tylko raz nie był na ostatnim niejscu - wtedy był drugi, licząc od końca.

Z kolei `cmeans()` sprawdził się na poziomie przeciętnego algorytmu z `hclust`.

Standaryzacja danych nie wpłynęła w żaden pozytywny sposób na jakość otrzymanych podziałów.