Praca domowa nr 4

Anna Kozioł

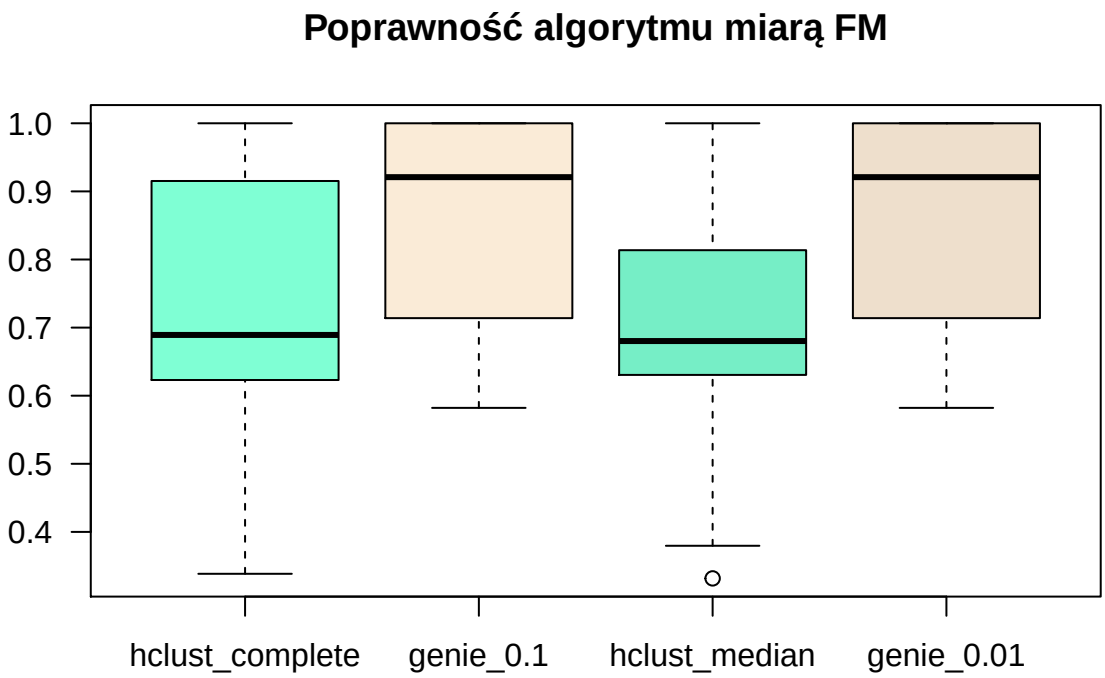
Cel zadania

- Dane przedstawiają wartości indeksów AR i FM dla funkcji hclust oraz genie testowania algorytmu skupień, dla różnych parametróworaz dla przeskalowanych danych. Zakładamy tutaj, ze algorytm analizy skupien jest dobry, jesli generuje podziały podobne do referencyjnych etykiet.
- Celem raportu jest zbadanie jakości działania każdego algorytmu dla wszytskich zbiorów oraz zmieniających się parametrów poprzez analizę indeksów , których wartości przyjmują wartości (0,1).
- Przykładowy wygenerowany plik dla ustalonego algorytmu zapisany został w poniższej postaci:

```
##          DANE        FM         AR  FM_stand      AR_stand
## 1          a1 0.9204150 0.91622843 0.6663222  0.6413114772
## 2          a2 0.9110338 0.90841548 0.9129788  0.9104304586
## 3          a3 0.9194996 0.91785563 0.9166449  0.9149388578
## 4 aggregation 0.8237319 0.77442001 0.7611451  0.6961027009
## 5        atom 0.6468033 0.08353119 0.6881696  0.3101118774
## 6   chainlink 0.6892174 0.31304549 0.5041759 -0.0009812061
```

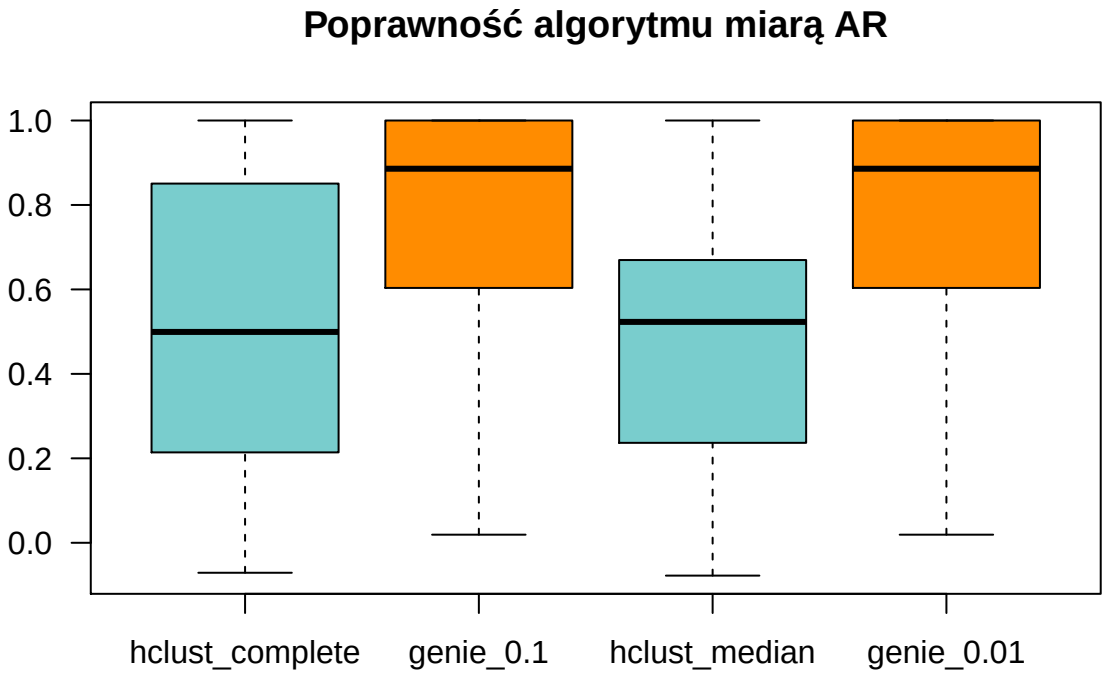
- Dla każdego rozważanego zbioru otzymaliśmy wartość indeksu.Zatem kolumnę takiego zbioru możemy potraktować jako próbkę, o długości równej ilości wczytanych zbiorów. Zbadajmy zachowane się funkcji indeksem FM

Jak "dobrze" zachowuje się algorytm mierzony indeksem FM?

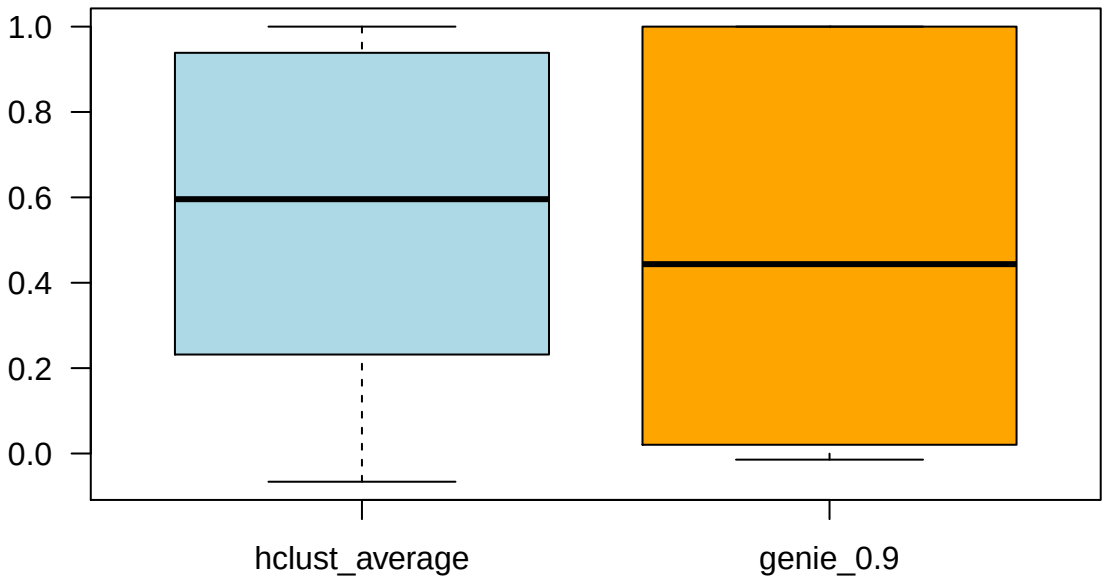


- Widać że średno algorytm z pakietu Genie przy współczynniku 0.1 oraz 0.01 działa lepiej niż funckja hclust przyjmująca parametry odpowiednio complete i median.
- Średnia poprawność algorytmu 'genie'mierzona indeksem AR wynosi ok 90 %.
- W przypadku funkcji 'hclust' wyniki są poprawne w średnio 60% - 70%

- Widać też że algorytm hclust jest też mniej stabilny, o czym świadczy rozrzut punktów.

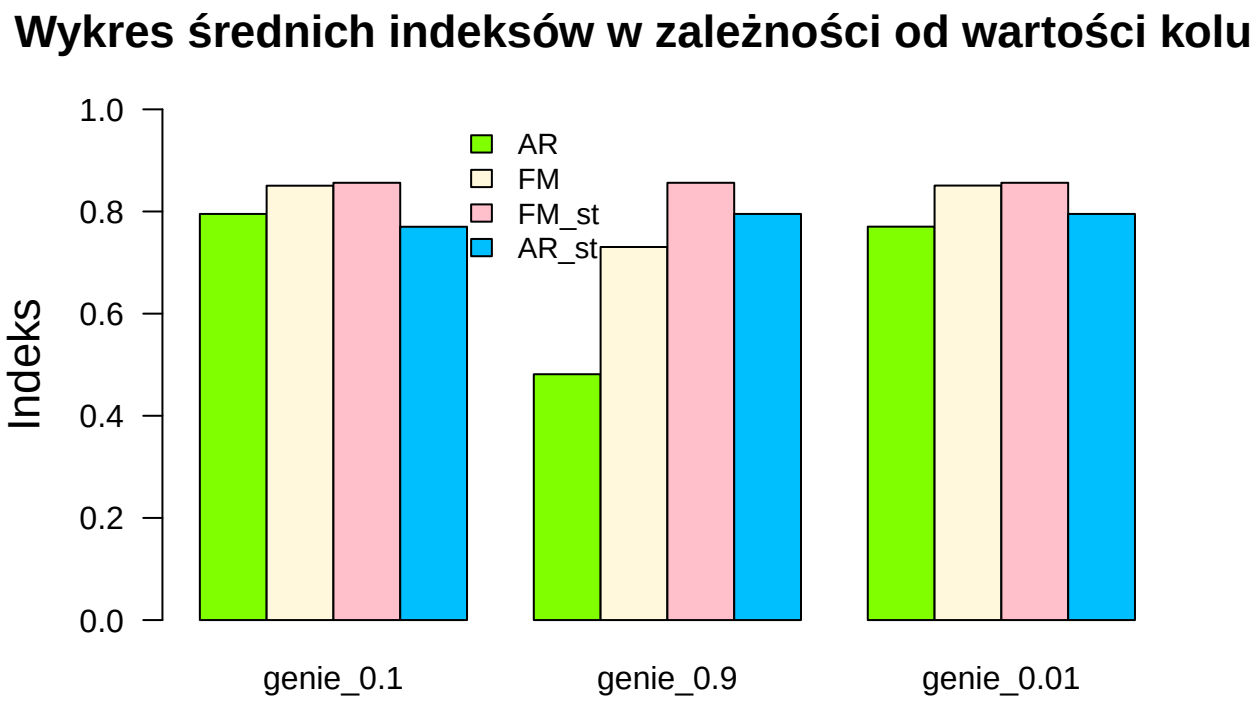


- Na podtsawie wykresów skrzynkowyh możemy stwierdzić zdecydowaną przewagę algorytmu 'genie' nad 'hclust' względem miary FM.
- Poprawność działania algorytmu 'hclust' określa się średnio dla 50% rozważanych przypadków.
- Porównując wyniki możemy również stwierdzić że dla parametrów 'median' i 'complete' indeks AR wskazuje wyższą poprawność algorytmu w porównaniu do indeksu FM.



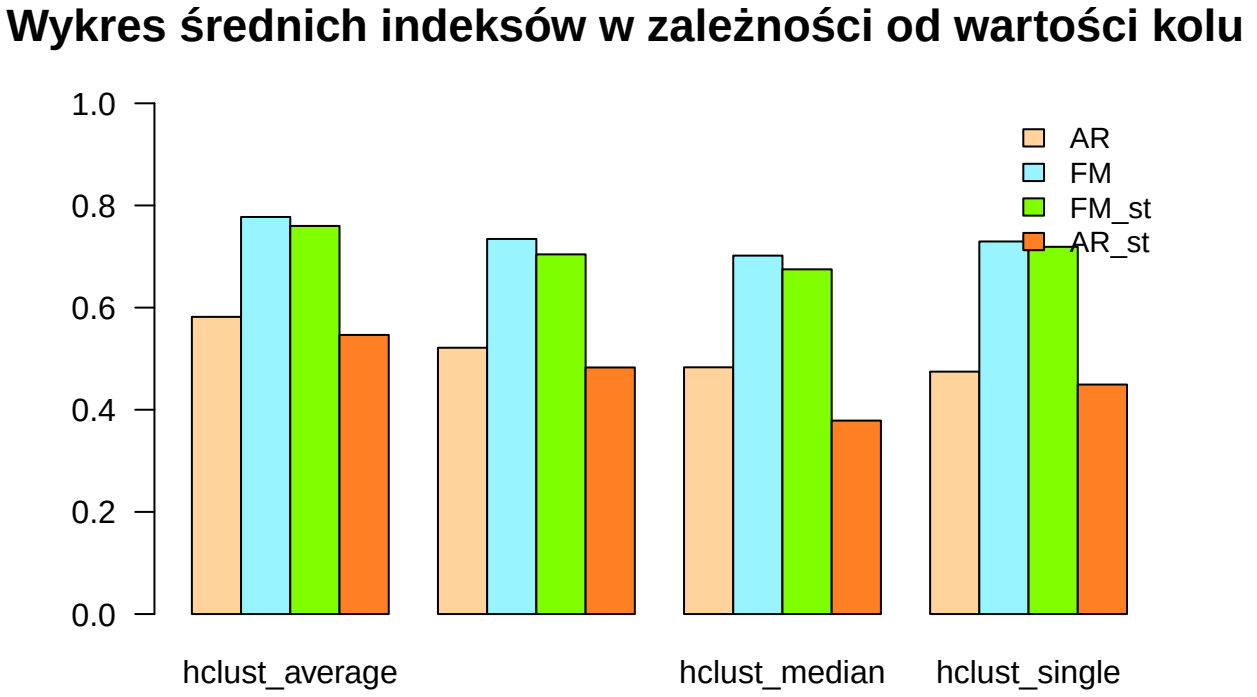
- Algorytm 'genie' daje średnio lepsze wyniki, jednakże przy parametrze współczynnika Giniego = 0.9 jest mało stabliny, o czym świadczy duży rozrzut punktów. Na wykresie porównanie z funkcja hclust z parametrem średniej. Przy takim porównaniu, funkcja hclust wypada trochę lepiej.

Badanie wpływu standaryzacji zmiennych na jakość działania algorytmu 'genie'



- W większości rozważanych algorytmów standaryzacja zmiennych obniża jakość używanych funkcji. Jedynie dla funkcji genie z parametrem 0.9 stadndaryzacja kolumn poprawiła wyniki indeksów.
- Zauważalne jest również że FM średnio wskazuje wyższą poprawność algorytmu niż AR.

Badanie wpływu standaryzacji zmiennych na jakość działania algorytmu 'hclust'



- Standaryzacja kolumn nie wpływa pozytywnie na wyniki algotytmów analizy skupień.