Systemy Uczące Się
Grupowanie/klasteryzacja

Michał Zając
203229

30 listopada 2016

Spis treści

1 Wstęp

Zapoznanie się z systemem R wspierającym statystyczne obliczenia i metody uczenia maszynowego, na przykładzie zagadnienia klasteryzacji (czasem zwaną grupowaniem) danych.

2 Algorytm K-means

Algorytm K-means jest heurystyczną metodą wyznaczania klastrów, do których przydzielane są wszystkie elementy zbioru. Przebieg algorytmu wygląda następująco:

1. Wybieramy k losowych punktów jako centroidy.
2. Każdy element zbioru obserwacji przypisujemy do najbliższego centroidu na podstawie odległości euklidesowej.
3. Obliczamy nowy centroid bazując na średniej wartości poszczególnych atrybutów w centroidzie.
4. Kroki 2-3 powtarzamy dopóki nie nastąpi modyfikacja pomiędzy krokami, lub nie zostanie spełniony inny warunek stopu.

3 Algorytm PAM

W algorytmie PAM zamiast centroidów środek klastra(zwanych tutaj medoidami) zawsze jest jednym z obiektów, który znajduje się w zbiorze. Właściwy algorytm prezentuje się następująco:

1. Wybieramy k losowych punktów jako medoidy.
2. Każdy element zbioru obserwacji przypisujemy do najbliższego centroidu na podstawie odległości euklidesowej.
3. Dopóki odległość między punktami zmniejsza się:
   (a) Dla każdego medoida i każdego obiektu nie będącego medoidem zamień je funkcjami (uznaj medoid za zwykły obiekt a wybrany obiekt za medoid). Jeżeli odległość się zwiększy od poprzedniego kroku, cofnij zmianę.
   (b) Każdy obiekt z klastra potraktuj jako medoid. Jeżeli odległośćmiędzy obiektami nie zmniejszy się w stosunku do poprzedniego kroku to cofnij zmianę.

4 Wybrane zbiory danych

Do ćwiczenia użyto następujących zbiorów danych:

Iris - zbiór danych o irysach. Występują w nim trzy możliwe klasyfikacje, w zależności od gatunku kwiatu. Liczba cech: 4. Liczebność zbioru: 150

Pima Indians Diabetes - zbiór zawierający dane o osobach z USA, którzy chorują na cukrzycę. Występują w nim dwie możliwe klasyfikacje: osoba chora i osoba zdrowa. Liczba cech: 8. Liczebność zbioru: 768

Ionosphere - zbiór danych zawiera dane dotyczące jonosfery zebrane z 16 anten. Dwie klasy, liczba cech: 34, liczebność zbioru: 351.

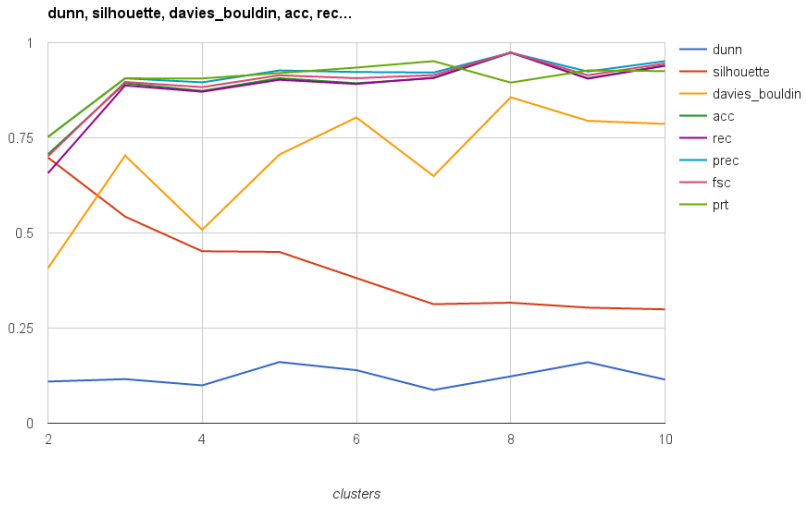
5 Implementacja

Do wykonania zadania napisano skrypt w języku R wykonujący potrzebne obliczenia.

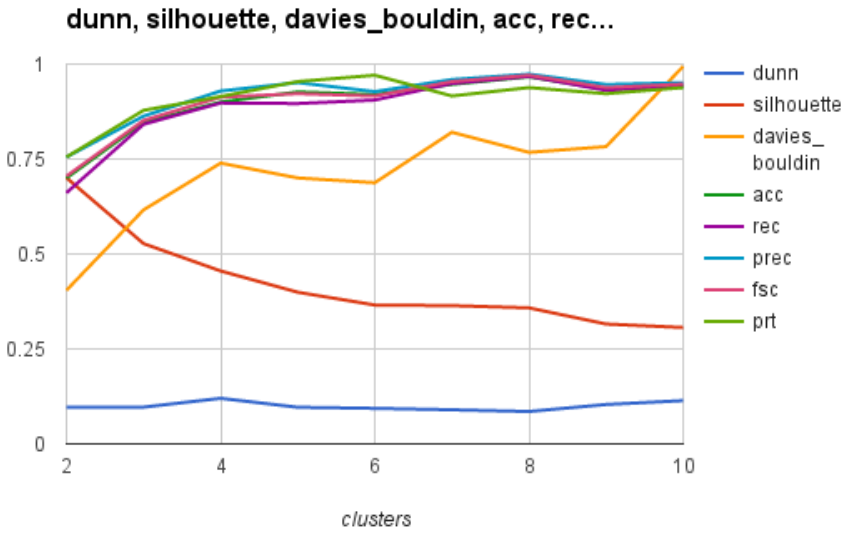
6 K-means

6.1 Iris

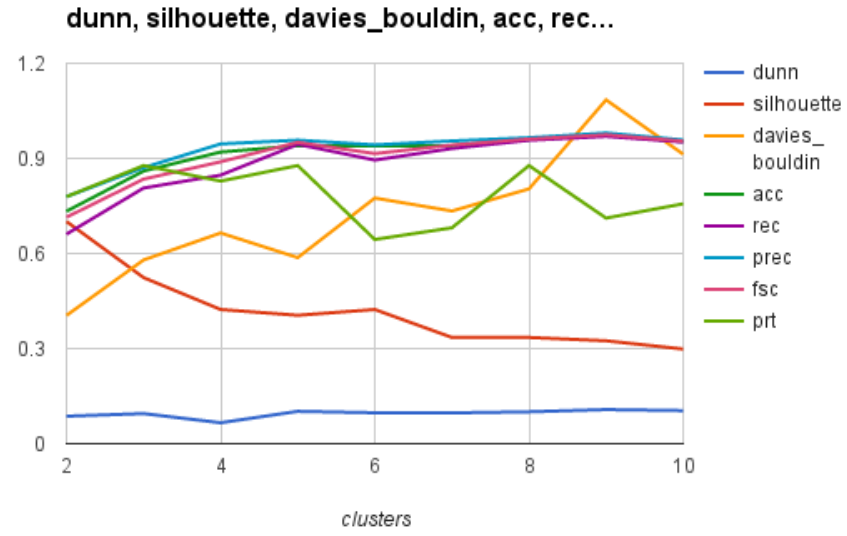
6.1.1 3-fold crossvalidation



6.1.2 5-fold crossvalidation

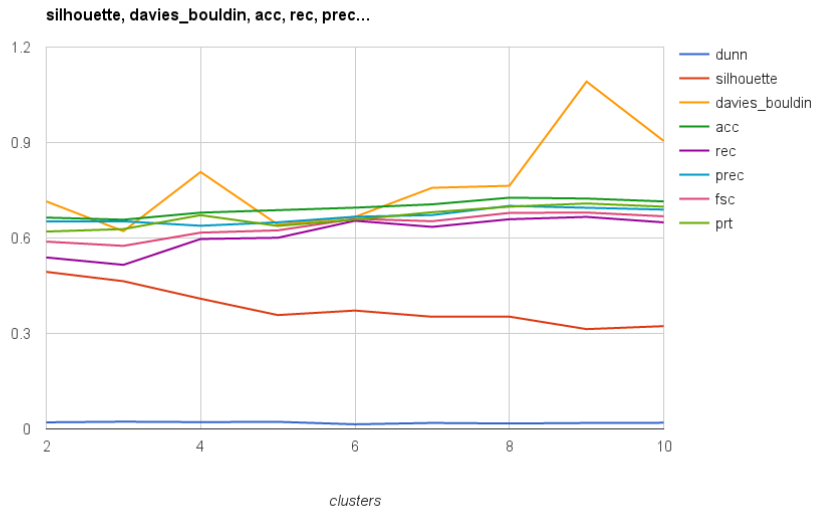


6.1.3 10-fold crossvalidation

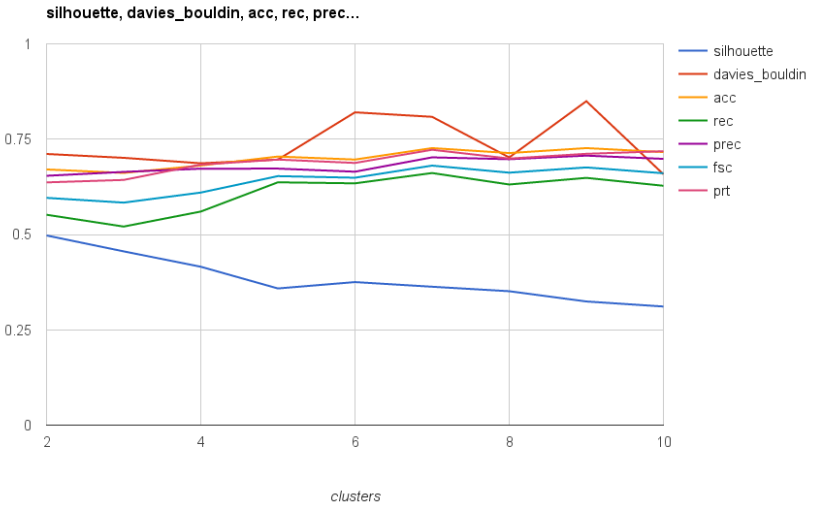


6.2 Diabetes

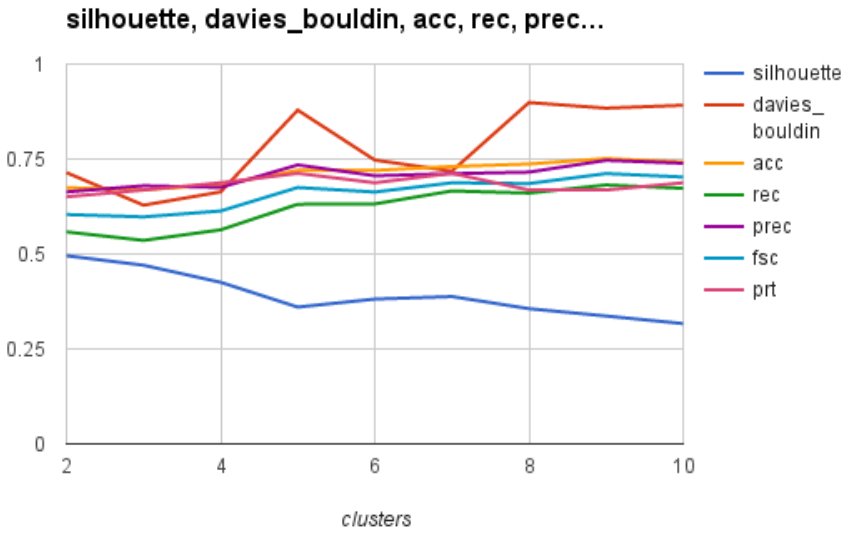
6.2.1 3-fold crossvalidation



6.2.2 5-fold crossvalidation

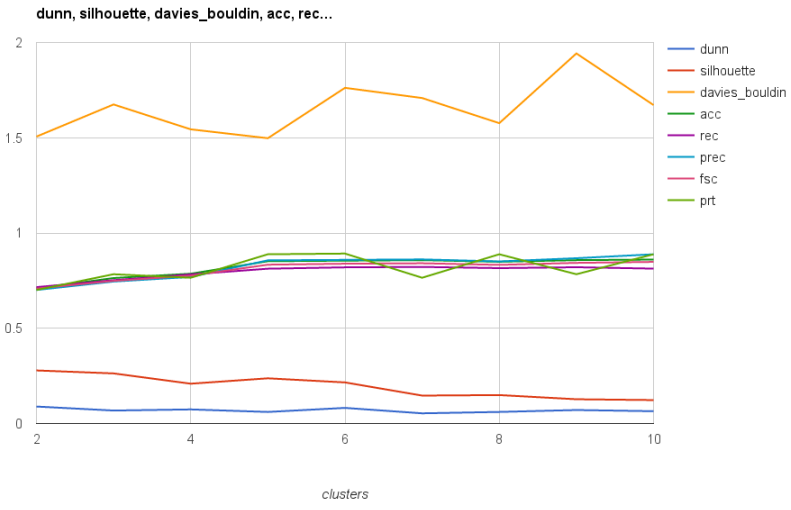


6.2.3 10-fold crossvalidation

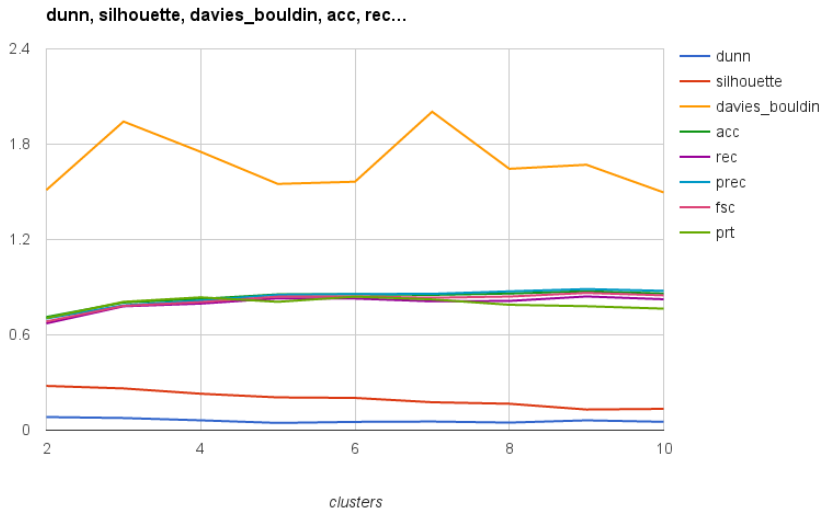


6.3 Ionosphere

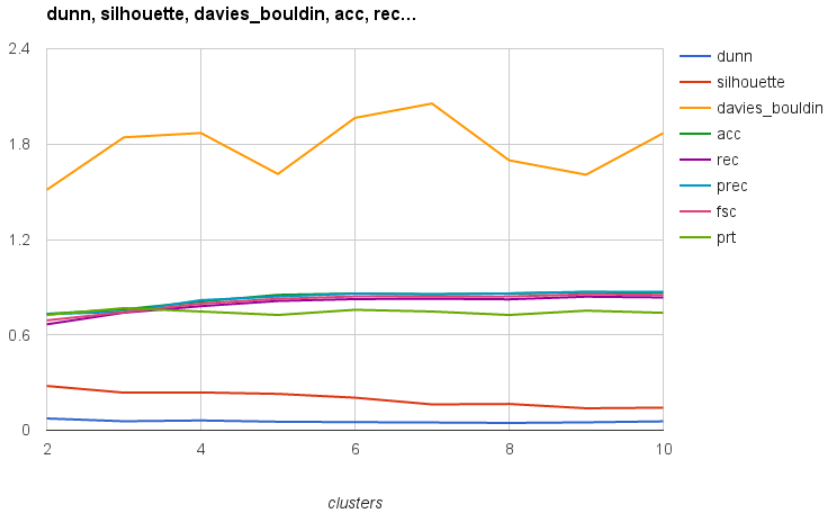
6.3.1 3-fold crossvalidation



6.3.2 5-fold crossvalidation



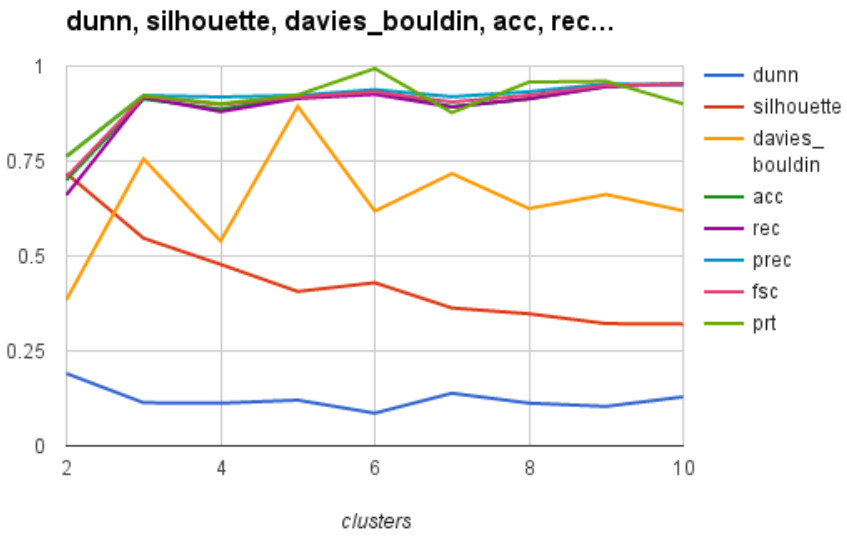
6.3.3 10-fold crossvalidation



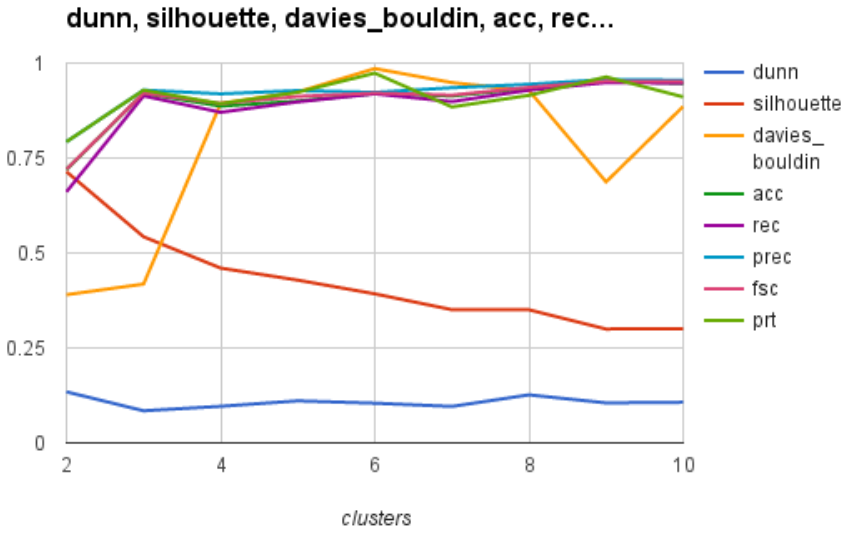
7 PAM

7.1 Iris

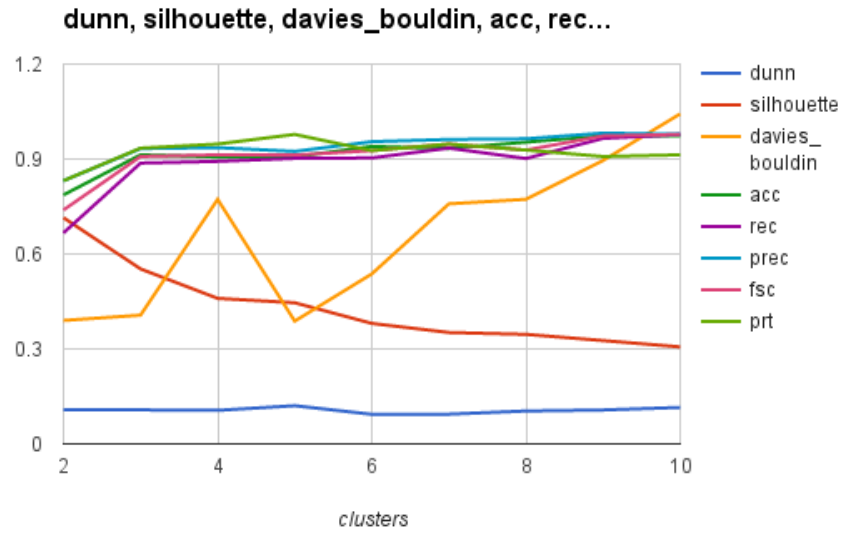
7.1.1 3-fold crossvalidation



7.1.2 5-fold crossvalidation

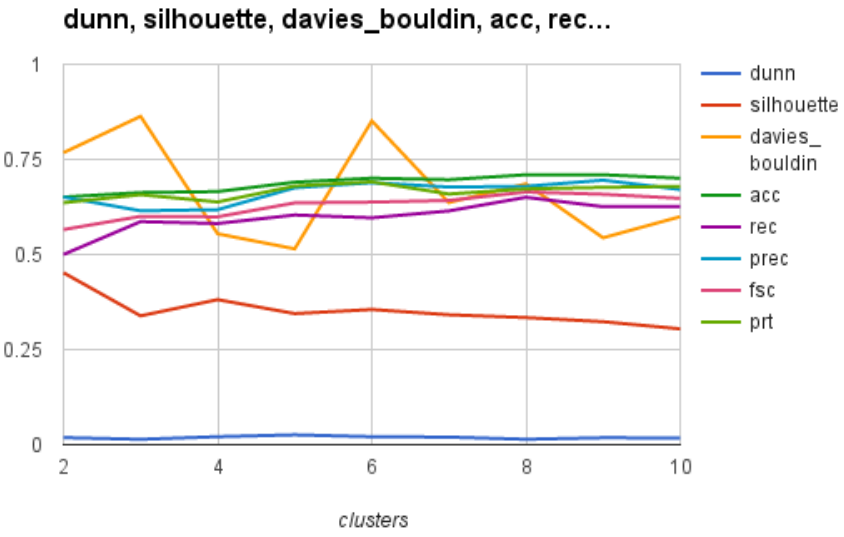


7.1.3 10-fold crossvalidation

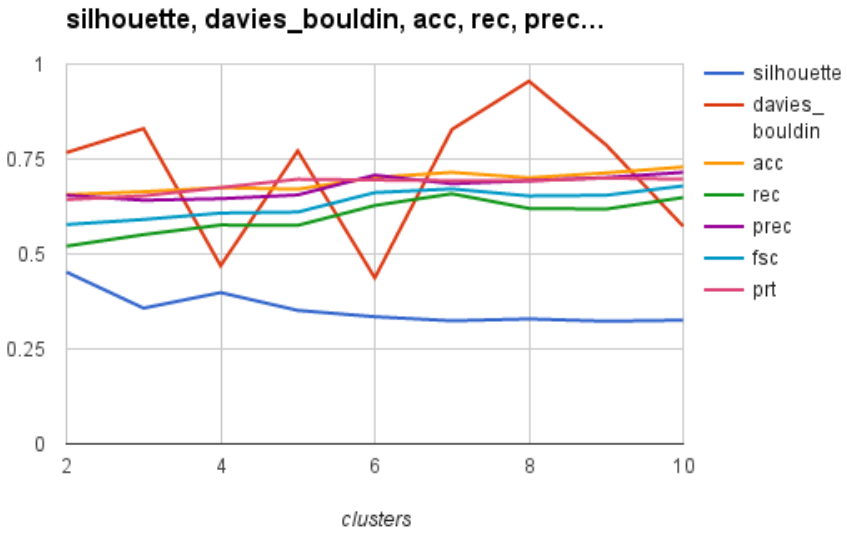


7.2 Diabetes

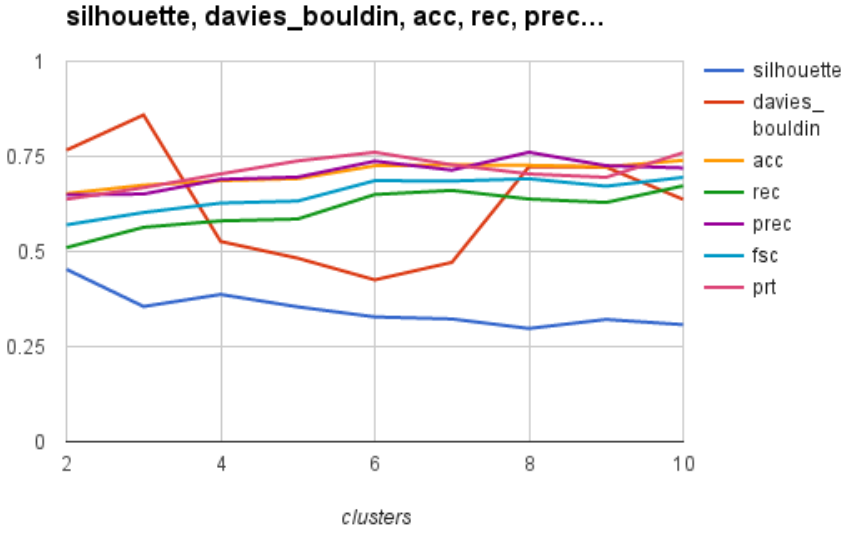
7.2.1 3-fold crossvalidation



7.2.2 5-fold crossvalidation

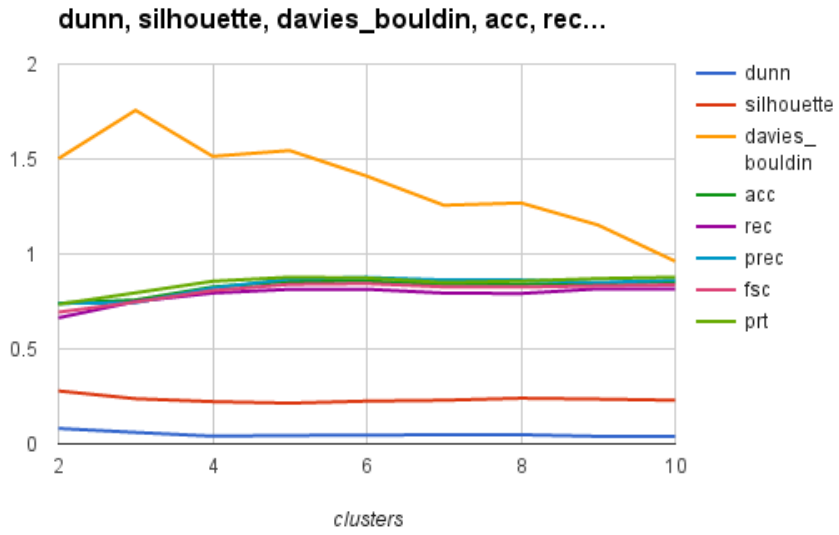


7.2.3 10-fold crossvalidation

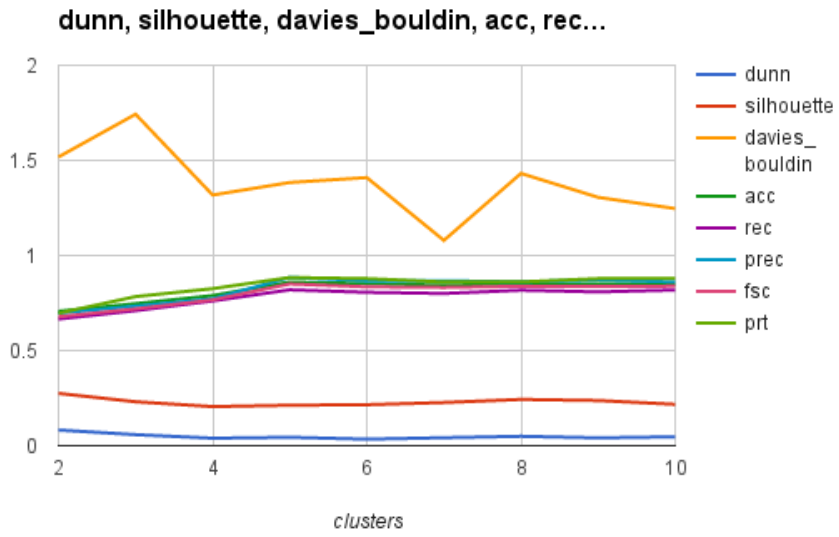


7.3 Ionosphere

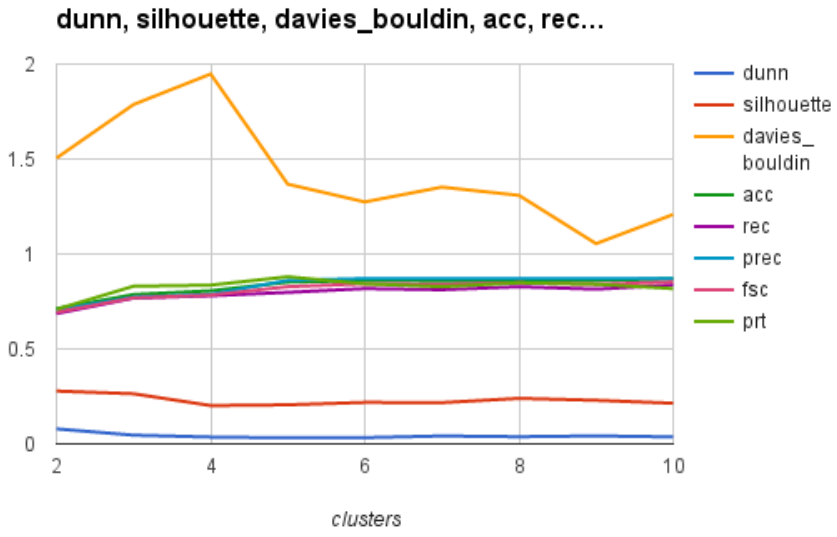
7.3.1 3-fold crossvalidation



7.3.2 5-fold crossvalidation



7.3.3 10-fold crossvalidation



8 Wnioski

1. Kroswalidacja w przypadku zadania klasteryzacji ma niewielki wpływ
2. Sam wskaźnik DBI nie jest najlepszym wyznacznikiem jakości klastra
3. R ląduje na drugim miejscu pod względem toporności wśród języków programowania