lab3

Omer Ronen

10/19/2020

1 Parallelizing k-means

1.

(a). Code

(b). Code

(c). Obviously the cpp code is much faster

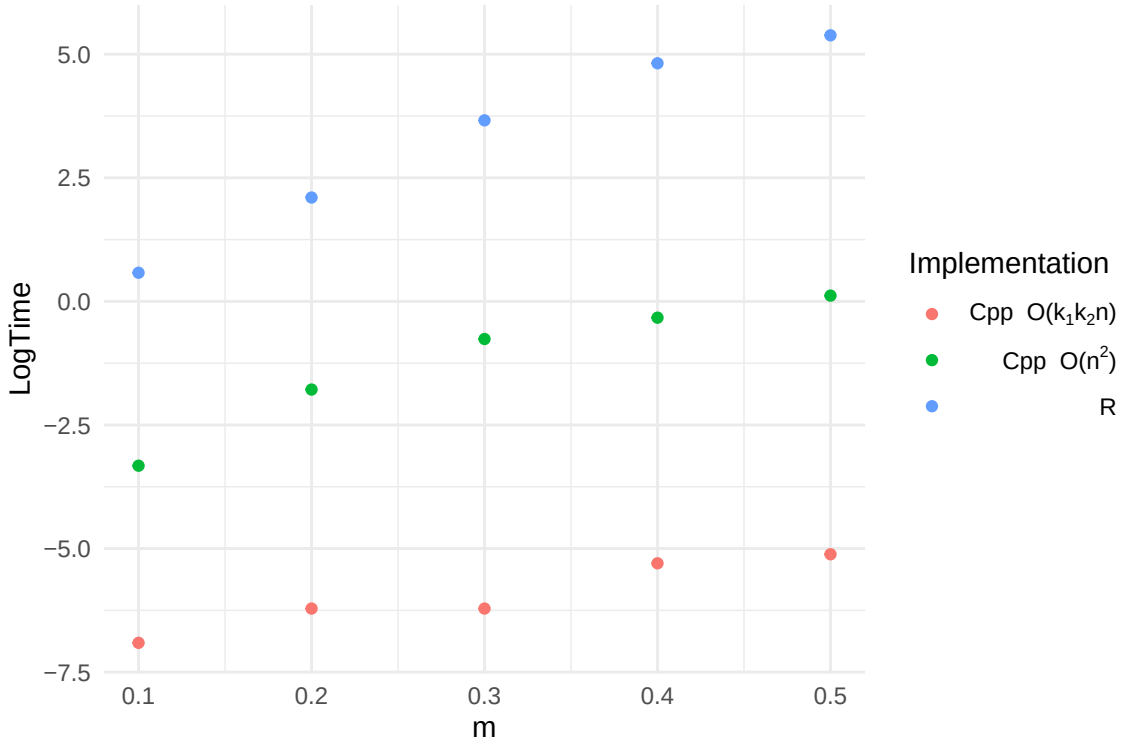

Figure 1: Run time (log scale) for R and Cpp implementations of similarity (correlation) score

2. code

3.

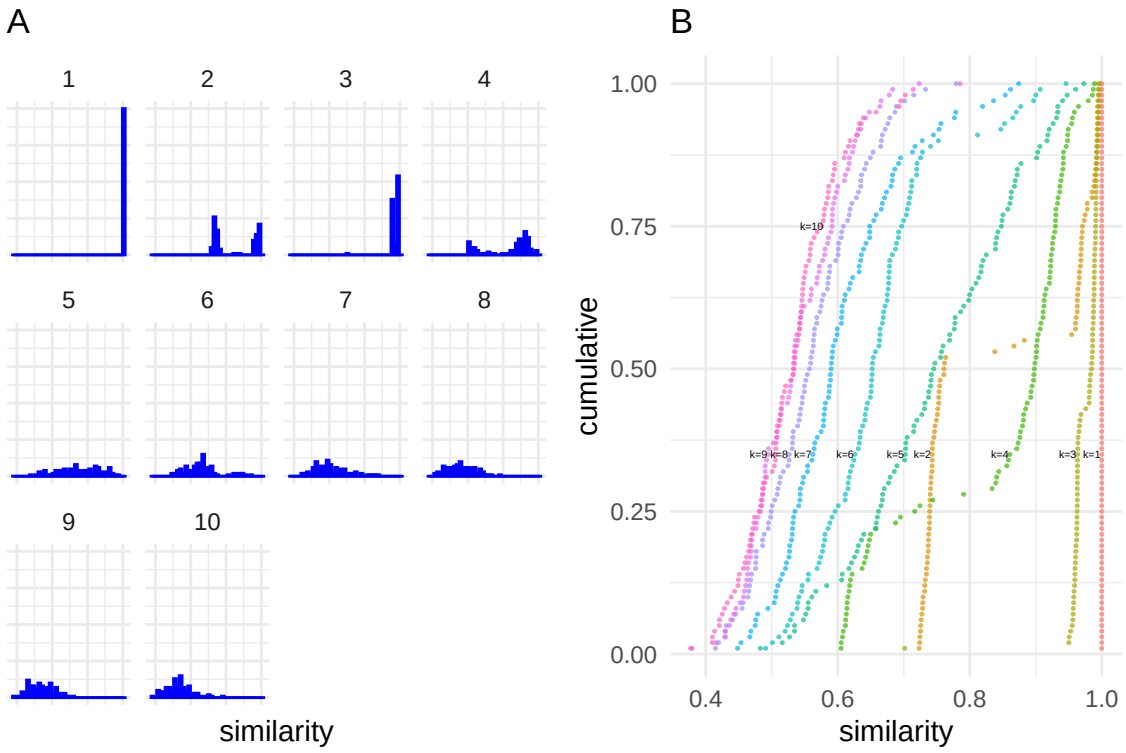

Figure 2: Stability analysis for choice of K (number of clusters). In A we can see the similiraity histograms for various choices of k. In B we show the empirical CDF obtained from calculating the similarity over 100 sub-samples of the dataset

It is clear to see that 3 cluster is the most stable (excluding 1 of course which is shown as a sanity check).