Title: Reducing the time requirement of k-means Algorithm.

Author(s): Oyelade, O.J.

Outlet: PloS One Journal Vol. 7(12) **(IF: 4.25)**.

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0049946
http://www.ncbi.nlm.nih.gov/pubmed?term=oyelade

Date:

Abstract: Traditional k-means and most k-means variants are still computationally expensive for large datasets, such as microarray data, which have large datasets with large dimension size d. In k-means clustering, we are given a set of n data points in ddimensional space R^d and an integer k. The problem is to determine a set of k points in R^d, called centers, so as to minimizethe mean squared distance from each data point to its nearest center. In this work, we develop a novel k-means algorithm, which is simple but more efficient than the traditional k-means and the recent enhanced k-means. Our new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering. We provided the correctness proof for this algorithm. Results obtained from testing the algorithm on three biological data and six non-biological data (three of these data are real, while the other three are simulated) also indicate that our algorithm is empirically faster than other known k-means algorithms. We assessed the quality of our algorithm clusters against the clusters of a known structure using the Hubert-Arabie Adjusted Rand index (ARI_{HA}). We found that when k is close to d, the quality is good (ARI_{HA}>.0.8) and when k is not close to d, the quality of our new k-means algorithm is excellent ((ARI_{HA}>0.9).

In this paper, emphases are on the reduction of the time requirement of the k-means algorithm and its application to microarray data due to the desire to create a tool for clustering and malaria research. However, the new clustering algorithm can be used for other clustering needs as long as an appropriate measure of distance between the centroids and the members is used. This has been demonstrated in this work on six non-biological data.