Dara Lim
CMPSCI 4370
16 May 2019

Improving the Time Complexity And Accuracy of K-Means Clustering Algorithm

The K-means algorithm is effective in producing clusters for many practical applications. However the computational complexity of the original k-means algorithm is very high, particularly for a large data sets. In addition, the algorithm outputs in different types of clusters base on the random choice of the initial centroids. For the semester project, my team decide to deal with the method for improving the accuracy and efficiency of the k-means algorithm. I would like to share the experience through the project as the following paragraph.

I work on implementing the efficiency of the k-mean algorithm while my team member working on the improving the accuracy. I have the difficulty with the linked list data structure, so for this project I choose to use the two dimensional array data structure instead. I use the Euclidean distance formula to find the path from one vertex to another vertex. The other thing is that only the numerical data value from the provided data set is taken to test on the algorithm. Due to the project hardness, my team does not meet the requirements, but I have learnt from it.

Through out the project process, I learnt the k-means clustering algorithm is powerful in arranging data in its relative group. With the k-means clustering, the data can be grouped into different amount of cluster. The standard algorithm of k-means clustering does not always ensure good outputs as the accuracy of the clusters depend on the selection of the centroids. Therefore, the project also implement the code to find the accuracy of the cluster.

Overall, the course is hard for me despite it is designed as an introduction. At the beginning, there are some algorithm to learn and these algorithm are difficult but useful for the data scientist. Especially, from the start of the biology topics, there are a lot of terminology, concept, and so on which are new to me and it is time consuming. At least, I know how the computer science field work together with biology data.