BIRCH clustering on benchmark Artificial Datasets

Github link : https://github.com/grabcollab18/BIRCH-implementation-from-scratch-on-artificial-datasets

INALA VIVEK VAMSI
2016AAPS0230H

BIRCH - Motivation

- Hierarchical clustering has its own advantages when compared to its other types
- Agglomerative clustering (one of the Hierarchical clustering methods) though is flexible, it is not scalable to the data set fed
- It also lacks the ability to undo what was done in the previous step
- To avoid these we need an extension to this method which can cater to the space and time complexity issues
- BIRCH is an unsupervised data mining technique and is like an extension to Agglomerative method, proposed to deal with these issues

About BIRCH

- It is a multi phase algorithm which consists of two(2) main phases
- introduces concept of Cluster Frequency(CF) tuple – (N,LS,SS)
- minimizes several data points of a sub cluster into a single CF tuple
- constructs CF tree with each node containing CF s of sub clusters
- leaf nodes contain the final sub clusters whose centroids serve as reduced data set points
- **Note** that reduced here doesn’t mean that data is lost - as all important data statistics still remain with corresponding CF

Major Achievements

- Used only numpy, pandas and matplotlib libraries throughout the project though I was told scikit(sk)learn could be used
- learnt whole object oriented python from scratch for the sake of implementation of BIRCH algorithm
- implemented algorithm forms the final sub clusters very accurately as we expect
- ran the implemented algorithm on all the artificial data set files and concluded that it performs very well in terms of time as well as space complexity
- had put great amount of effort and time to code as well as comment it at each significant juncture or line
- Overall it was a comprehensive learning experience, both practically as well as theoretically

Shortcomings

- Distance metrics other than ‘euclidean’ can be used as per application and data
- Another shortcoming or rather we can call it a scope for extension as explained below
- Apropos to the 3rd slide, BIRCH in whole has 2 main phases :
- one is building the CF tree and the other is applying Agglomerative clustering on centroids of sub clusters in leaf nodes
- As the main goal of this assignment was to implement BIRCH from scratch, this implementation only confines to phase 1 of the complete clustering using BIRCH
- However implementation of Agglomerative clustering algorithm can also be done as a separate task to complete both clustering algorithm as well as its extension(BIRCH) as whole

Acknowledgment

- As soon as I received this assignment on mail, I was totally confused as in how to start because I was completely new to Object oriented python. I acted in haste and drafted a mail to Manik ma’am and Deepak Sir asking for help or change of assignment. But then I realized that I had enough time to learn, understand and implement. From that day I spent time everyday till today (last date of submission) to do my best by going through lectures, various online resources and research papers. I am very pleased to complete this assignment on time with my best efforts and it proved to be a great learning curve overall. This will definitely add more to my resume worthy projects bucket. So I would like to thank **Dr. Manik Gupta** ma’am very much for her knowledge transfer and guidance through out the course.