# BIRCH Algorithm

Aekansh 2016A7PS0127

# BIRCH Algorithm

- **BIRCH algorithm** (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining **algorithm** which is used to perform hierarchical clustering over particularly large data-sets.
- The BIRCH algorithm uses a tree structure to create a cluster. It is generally called the Clustering Feature Tree (CF Tree). Each node of this tree is composed of several Clustering features (CF).
- CFs of internal nodes have pointers to child nodes, and all leaf nodes are linked by a doubly linked list.

# CF Tree

A clustering feature (CF) is defined as follows: Each CF is a triplet, which can be represented by (N, LS, SS).

- Where N represents the number of sample points in the CF, which is easy to understand
- LS represents the vector sum of the feature dimensions of the sample points in the CF
- SS represents the square of the feature dimensions of the sample points in the CF.

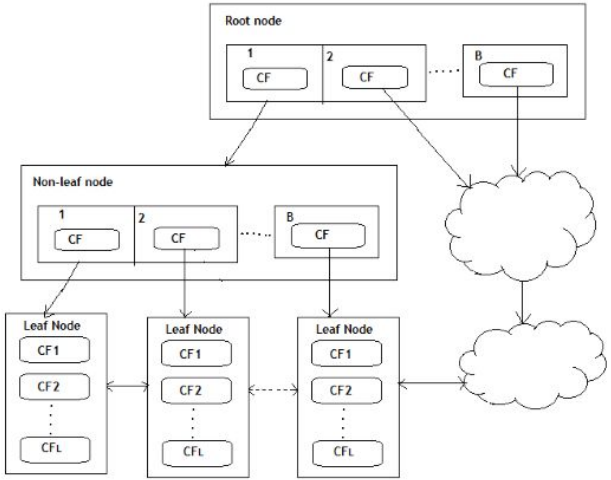
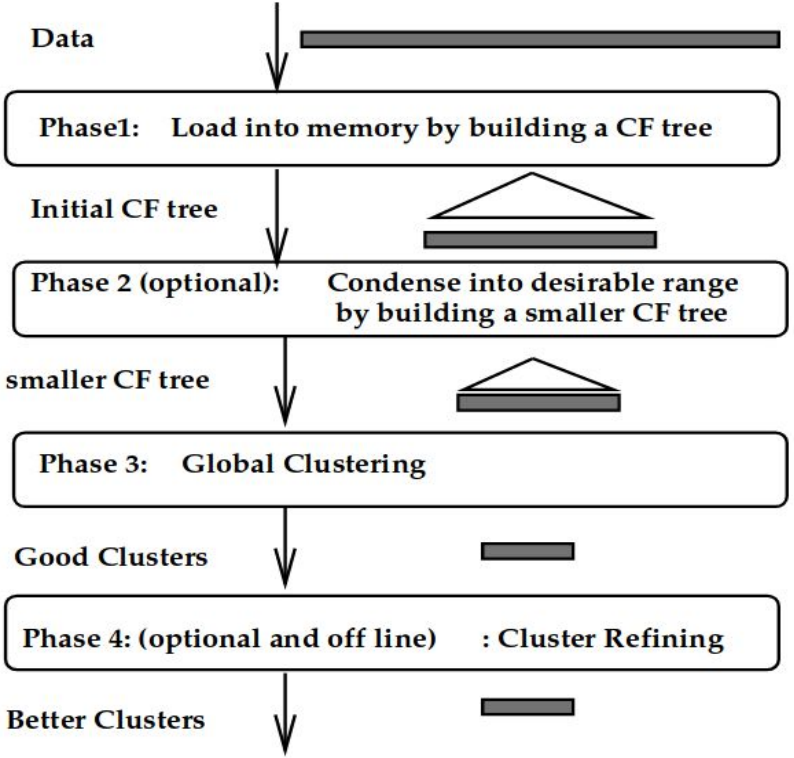


*Figure 2.* BIRCH Overview



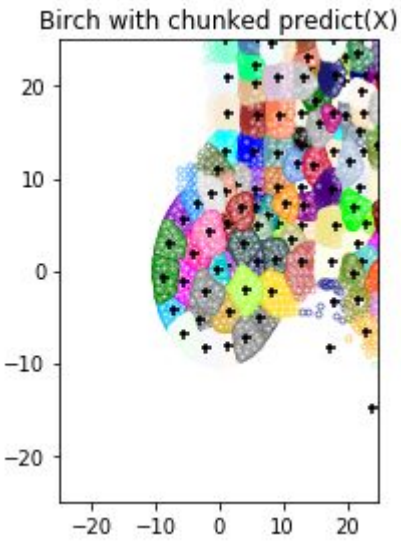
# Advantages and disadvantages

+ Finds a good clustering with a single scan and improves the quality with a few additional scans.
+ It can incrementally and dynamically cluster incoming, multi-dimensional metric data points in an attempt to produce the best quality clustering for a given set of resources (memory and time constraints).
+ The CF trees store pointers so they can be very memory efficient.
- It only works on numerical data.

# Results

- A total of 8386 clusters were formed

- Memory use was optimised by modifying the Predict() method of the conventional Birch algorithm to handle large matrix multiplications.



# Achievements of doing this project:

- Learnt to preprocess data
- Learnt to deal with huge number of data points.
- Learnt how to reduce load from the memory by using pointers.
- Learnt how to deal with dynamic data in clustering.
- Learnt how to deal with memory errore, what causes them and how to avoid it
- Learnt about modified BIRCH algorithms such as BirchChunked and NewBirch which improve the performance and avoid memory overloads.

# Shortcomings:

- Initial implementation could not handle the huge amount of data.

- Multiplication of huge matrices to find the clustering features were taxing on my computer which does not have enough memory to handle it.

- Memory errors were present which delayed the completion of project as I had to research more to find a solution.

# Thank You

https://github.com/FaceTheAce/DM_FinalAssign_Aekansh_2016A7PS0127H

:link to github repository