BIRCH ALGORITHM

BY – C. RACHITA

INTRODUCTION

- BIRCH – Balanced Iterative Reducing and clustering hierarchies
- Clustering algorithm
- Works well on large datasets because requires only one time scan
- Generally used by other clustering algorithms as a stepping stone.

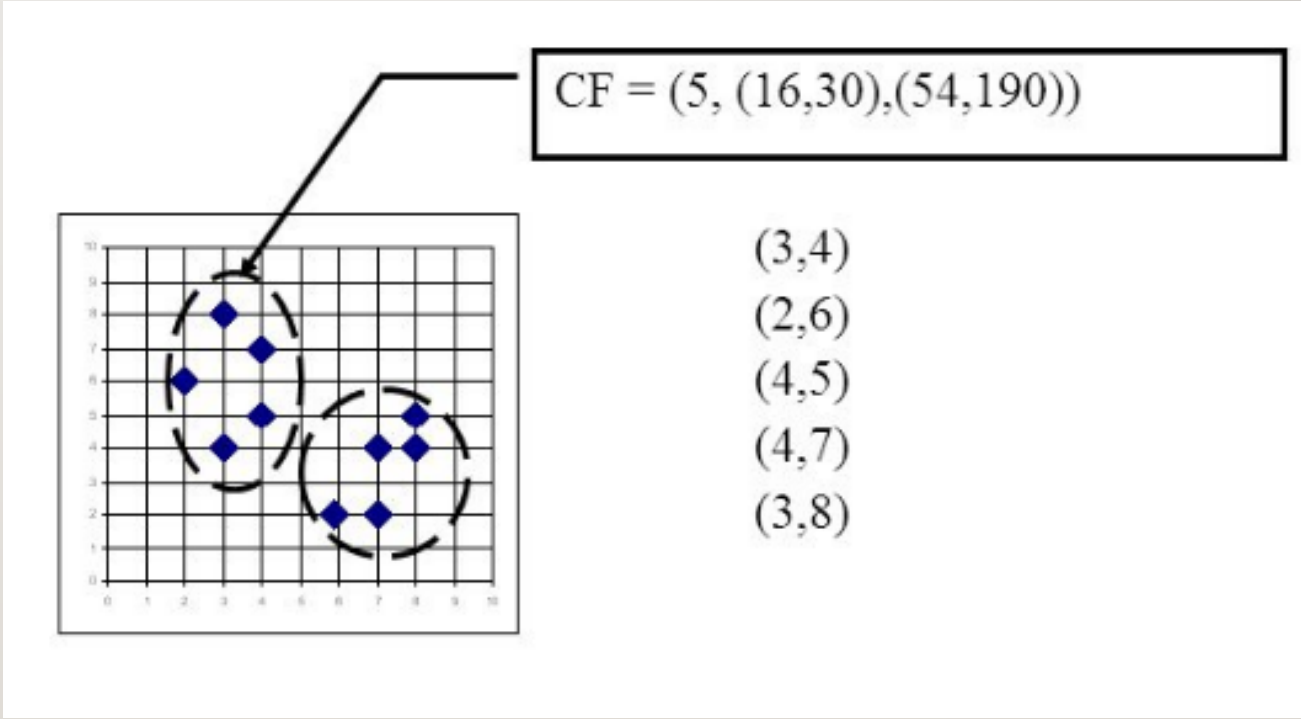
APPROACH

- Converts the data into a tree data structure
- The centroids made in each child node can either be the final cluster centroids, or an input for the next clustering algorithm
- To build this tree structure, it uses **Clustering Feature(CF) tree**
- CF Tree compresses data into sets of **Cluster Feature(CF) nodes.**
- **CF nodes** holds necessary information about data points.

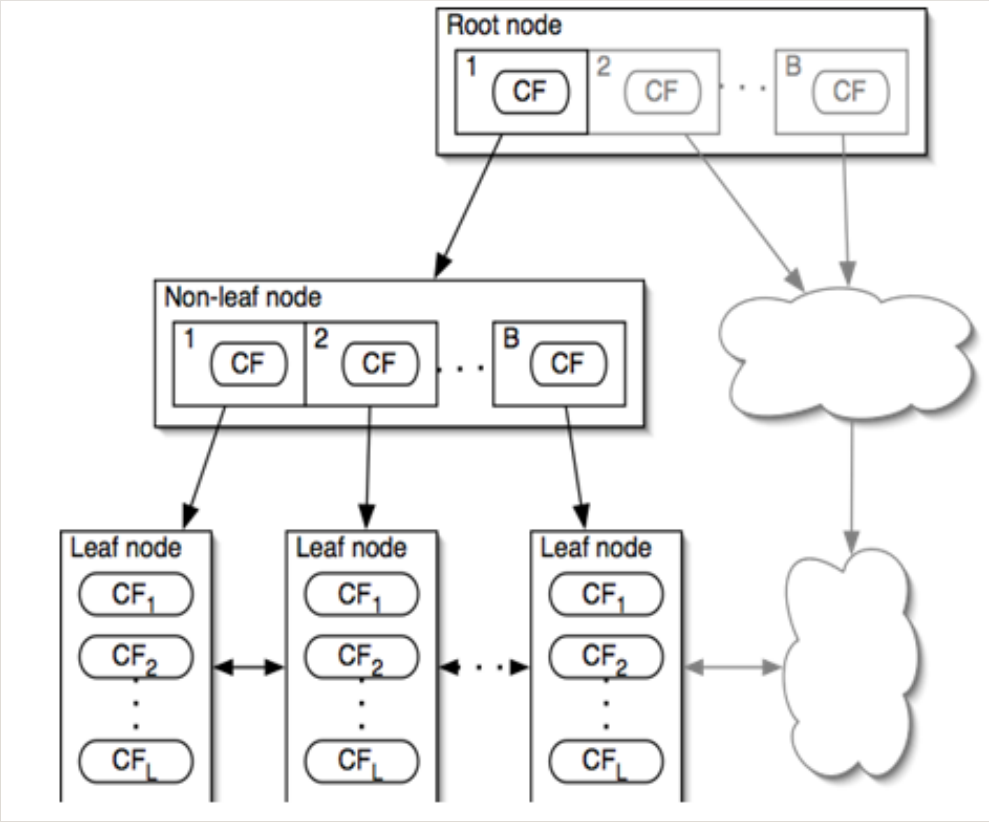
CF NODE

CF stores information about its data points using 3 factors:

1. N – no. of items in that cluster
2. LS- sum of their data points
3. SS – sum of squared data points



CF TREE


Root node
1
CF
2
CF
B
CF
Non-leaf node
1
CF
2
CF
B
CF
Leaf node
CF1
CF2
CFL
Leaf node
CF1
CF2
CFL
Leaf node
CF1
CF2
CFL

IMPLEMENTATION PARAMETERS

- Threshold - The maximum number of data samples to be considered in a subcluster of the leaf node in a CF tree.
- Branching_factor - It is the factor that is used to specify the number of CF sub-clusters that can be made in a node.
- n_clusters - number of clusters after the final clustering step(If set to None, final clustering step is not performed)