Algorithm

BIRCH is a two step clustering process. The process is listed below

* First, creation the CF Tree
* Second, Global clustering(existing clustering algorithm) is applied to leaves of the CF Tree

**Creating CF**

BIRCH clustering effectiveness is because of the employment of a small set of summary statistics to represent a larger set of data points. For clustering purposes, these summary statistics constitute a CF, and represent a sufficient substitute for the actual data.

A CF is a set of three summary statistics that represent a set of data points in a single cluster.

These statistics are *count, Linear sum, and Squared sum*.

Listed below is an example for better understanding of how elements of a CF Tree is created

Considering two clusters: Clusters 1 and 2.

Cluster 1 contains data values (1, 1), (2, 1), and (1, 2)

Cluster 2 contains data values (3, 2), (4, 1), and (4, 2).

CF1, the CF for Cluster 1, consists of the following:

CF1 = {3, (1 + 2 + 1, 1 + 1 + 2), (12 + 22 + 12, 12 + 12 + 22)}

= {3, (4, 4), (6, 6)}

For Cluster 2, the CF is

CF2 = {3, (3 + 4 + 4, 2 + 1 + 2), (32 + 42 + 42, 22 + 12 + 22)}

= {3, (11, 5), (41, 9)}

Also, we know that BIRCH calls for merging of cluster under conditions, described later, for effective clustering of large data sets. Here, we use Additive theorem to combine the CFs. Thus, merging of Clusters 1 and 2 would result a CF having following statistics.

CF12 = {3 + 3*,* (4 + 11*,* 4 + 5)*,* (6 + 41*,* 6 + 9)} = {6*,* (15*,* 9)*,* (47*,* 15)}

**Creating CF Tree**

Since now we have understanding of CF, we move ahead to create the CF Tree. Creating a CF tree depends on 3 parameters, as described below

• **Branching Factor *B***. Determines the maximum children allowed for a non-leaf node.

• **Threshold *T***. *it* is an upper limit to the radius of a cluster in a leaf node.

• **Number of Entries in a Leaf Node *L***.

Creating of CF Tree is done using a sequential clustering approach, whereby the algorithm scans the data one record at a time, and determines whether a given record should be assigned to an existing cluster, or a new cluster should be constructed.

The CF tree building process consists of four steps, as follows[2]:

**1.** For each given record, BIRCH compares the location of that record with the location

of each CF in the root node, using either the linear sum or the mean of the

CF. BIRCH passes the incoming record to the root node CF closest to the incoming

record.

**2.** The record then descends down to the non-leaf child nodes of the root node CF

selected in step 1. BIRCH compares the location of the record with the location of

each non-leaf CF. BIRCH passes the incoming record to the non-leaf node CF closest

to the incoming record.

**3.** The record then descends down to the leaf child nodes of the non-leaf node CF

selected in step 2. BIRCH compares the location of the record with the location

of each leaf. BIRCH tentatively passes the incoming record to the leaf closest to the

incoming record.

**4.** Perform one of (a) or (b):

**a.** If the radius (defined later in the paper) of the chosen leaf including the new record does

not exceed the Threshold *T*, then the incoming record is assigned to that leaf.

The leaf and all of its parent CFs are updated to account for the new data

point.

**b.** If the radius of the chosen leaf including the new record does exceed the Threshold

*T*, then a new leaf is formed, consisting of the incoming record only. The parent

CFs are updated to account for the new data point.

**Global Clustering**

Once we are done creating the CF tree, any existing standard clustering algorithm(we used k-means) may be applied to the sub-clusters (the CF leaf nodes), to combine these sub-clusters into clusters. Because there are many fewer sub-clusters than data records, the task becomes much easier(reduced computational expense) for the clustering algorithm in the cluster step.