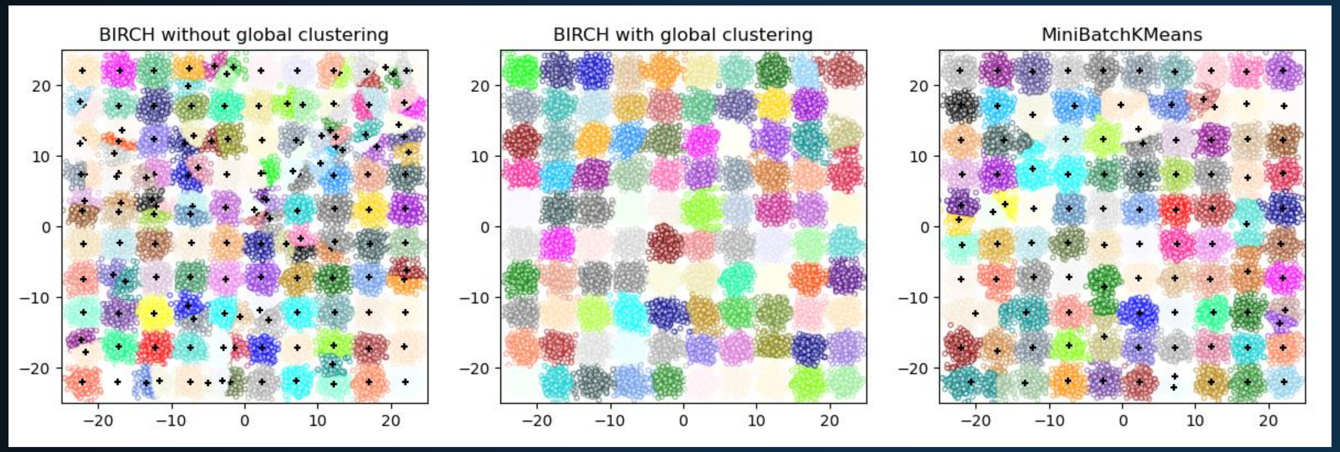
BIRCH Clustering

MACHINE LEARNING CLUSTERING

BIRCH Clustering

Introduction to BIRCH Clustering

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a hierarchical clustering algorithm designed for large datasets. Constructs a Clustering Feature (CF) tree to summarize the dataset efficiently and perform clustering in a single scan.



Key Concepts:

- **Clustering Feature (CF):** A compact representation of a cluster, including the number of points, linear sum, and square sum.
- **CF Tree:** A height-balanced tree structure that stores the CFs.
- **Threshold (T):** Maximum diameter of a sub-cluster to control the growth of the CF tree.

Mechanics of BIRCH Clustering

1. **Initial Scan:**
    - Insert data points into the CF tree, updating CFs and splitting nodes as necessary based on the threshold (T).
2. **CF Tree Construction:**
    - Build the CF tree dynamically as data points are added, ensuring the tree remains balanced and within the specified threshold.
3. **Clustering:**
    - After the CF tree is built, perform an optional global clustering algorithm (e.g., K-means) on the leaf entries of the CF tree to refine clusters.
4. **Refinement:**
    - Optionally, refine the clusters by rebuilding the CF tree with a smaller threshold and reapplying the clustering algorithm.

Application and Evaluation

Application:

- Suitable for large datasets, such as customer segmentation and image analysis.
- Steps:
  - ✓ Build the CF tree from the dataset.
  - ✓ Apply a global clustering algorithm to the leaf entries.

Advantages:

- Efficient for large datasets due to its single scan and incremental nature.
- Handles noise effectively by summarizing data compactly.
- Can be applied to dynamic data, allowing incremental updates.

Disadvantages:

- Sensitive to the choice of threshold (T).
- The initial clustering quality depends on the CF tree construction.
- May not perform well with non-spherical clusters.

Thank You