BIRCH: An Efficient Data Clustering Method for Very Large Databases

Presentation By Nathan Bullock

For CMPUT 695: Principles of Knowledge Discovery in Databases

Discussion Points

- Points can't be removed from a cluster.
- Rate at which T_i is increased.
- Leaf entry is considered an outlier if it has "Far fewer" data points than the average.
- Only finds spherical clusters.
- Node size is limited by memory, therefore doesn't always represent a natural cluster.

Overview of BIRCH

1. Build CF tree from the data.

2. (*optional*) Condense CF tree to desirable size.

3. Global clustering.

4. (*optional*) Cluster refinement.

CF vector

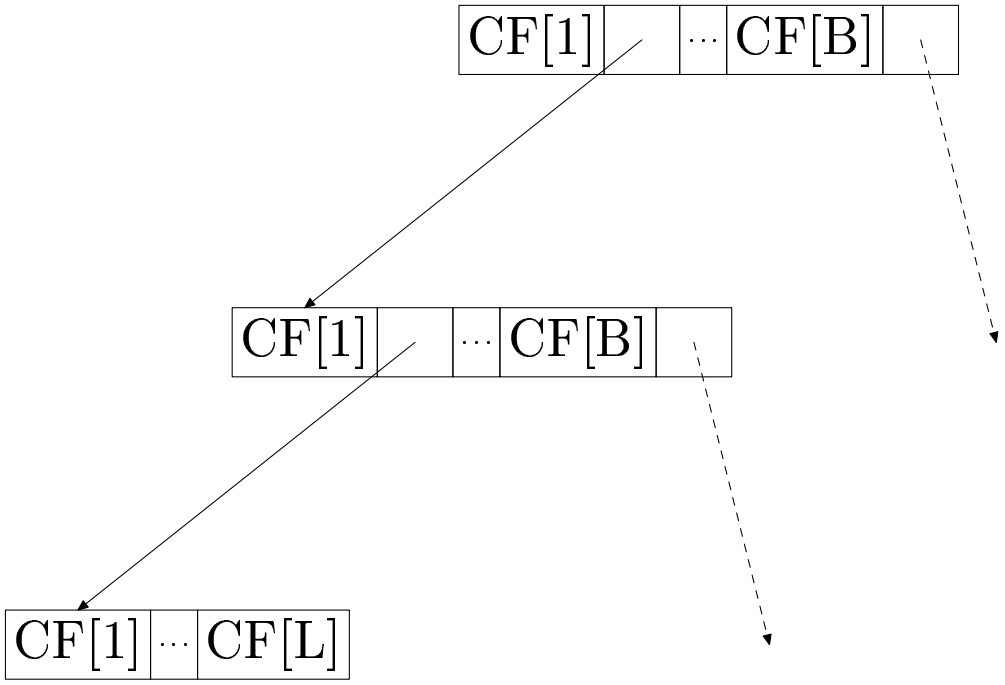
- $CF = (N, LS, SS)$
- summary of a cluster.
- It is all that is needed for the calculation of measurements.

$$CF_1 + CF_2 = (N_1 + N_2, LS_1 + LS_2, SS_1 + SS_2)$$

Cluster Metrics

$$
\begin{aligned}
X0 &= \frac{\sum_{i=1}^{N} X_i}{N} \\
&= \frac{LS}{N} \\
R &= \left(\frac{\sum_{i=1}^{N}(X_i - X0)^2}{N}\right)^{1/2} \\
&= \left(\frac{SS + N*X0^2 - 2*LS*X0}{N}\right)^{1/2} \\
D &= \left(\frac{\sum_{i=1}^{N}\sum_{i=1}^{N}(X_i - X_j)^2}{N(N-1)}\right)^{1/2} \\
&= \left(\frac{2*N*SS + 2*LS^2}{N(N-1)}\right)^{1/2}
\end{aligned}
$$

CF tree



note: parameters L and B are derived from P and all entries in the leafs must satisfy T.

Global Clustering

- Use existing clustering algorithm on sub-clusters.
  - Use the centroid of the CF vectors.
  - Use the count and the centroid of the CF vectors.
  - Use the CF vector directly.

Overview of BIRCH

1. Build CF tree from the data.

2. **(optional) Condense CF tree to desirable size.**

3. Global clustering.

4. **(optional) Cluster refinement.**

Advantages of BIRCH

- Clustering decisions are local.
- It can be incremental.
- Exploits the idea that not all data points are equally important.
- Full use of memory to derive finest possible sub-clusters.

Contributions of BIRCH

- They focus on a finite amount of memory.
- Try to minimize the amount of I/O.
- First algorithm to address outliers and proposes a plausible solution.
- Comes up with the idea that all points are not equally important. Data points which are close and dense should be considered collectively.

QUESTIONS?

and

DISCUSSION