Result Analysis with Front Runners

The Idea of Front Runners

It would be inefficient to calculate the distance of the closest point to every point in a cluster. Instead, we calculate the distance to the outermost points of the cluster. So, they act as representers of the cluster. We call these representers as “front runners”. In earlier versions of Bacteria-Farm algorithm, the centroid used to represent the cluster. When the front runners are in action, the inner points are dormant / hibernating. As the cluster expands, we have to update the front runners. To achieve this, once the nearest point to the cluster is found, we replace the nearest front runner with the new point.

Parameters

1. Number of front runners
2. Percentage of noise

Performance

The algorithm has a performance of O (n*n*k). With some optimizations, this can be reduced to O (n*log(n)*k).

Advantages

1. Easy parameters in contrast with DBSCAN.
2. Noise is controlled. We can mention the rate of noise to exclude. This is unique to BFFR and is very useful for cluster estimation.

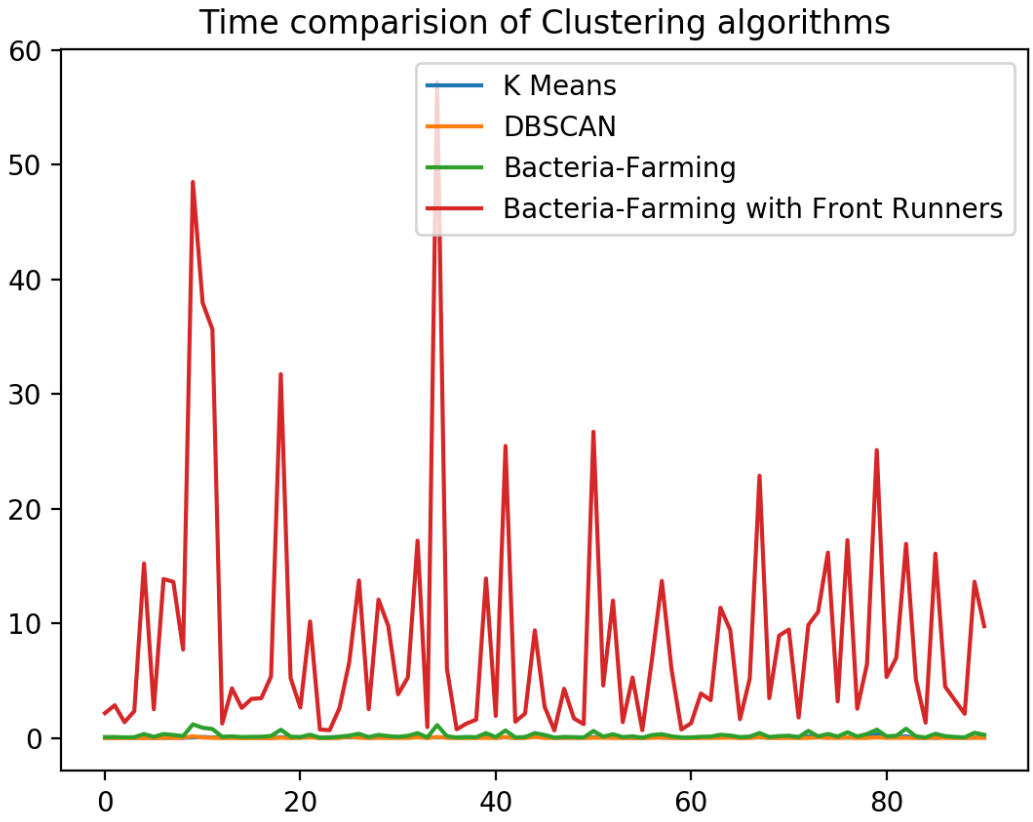
Disadvantages

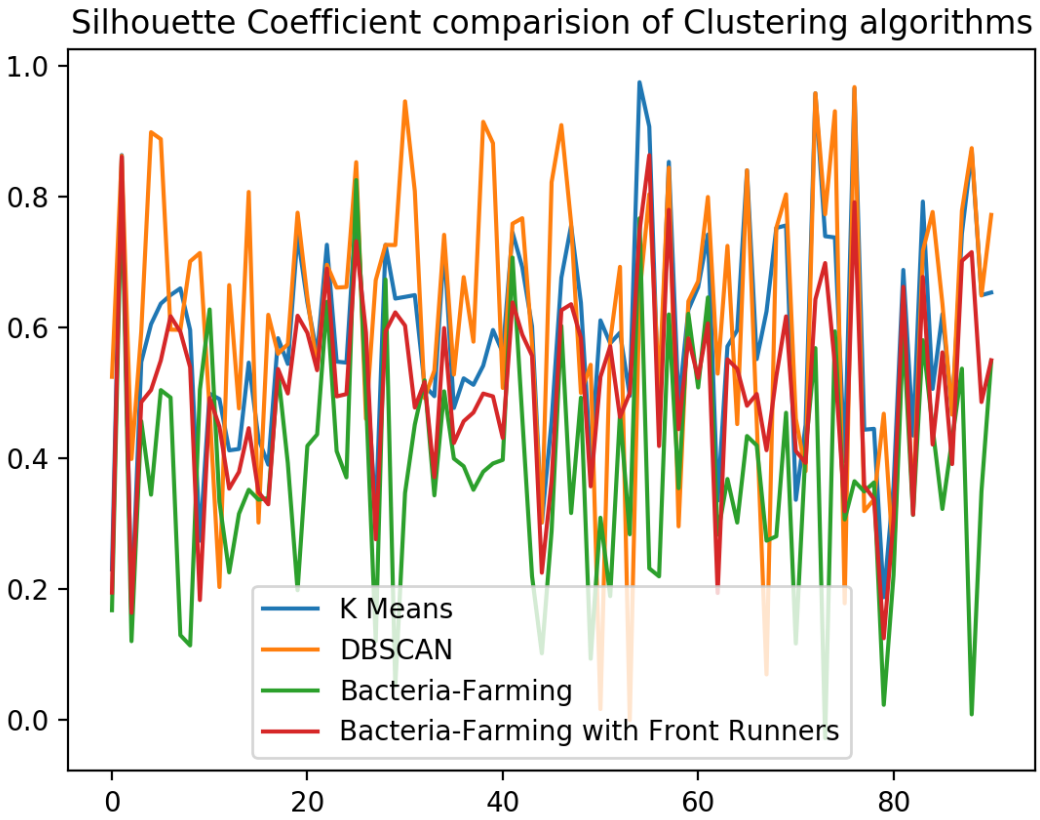
It is *m* times slower than DBSCAN where *m* is number of front runners.

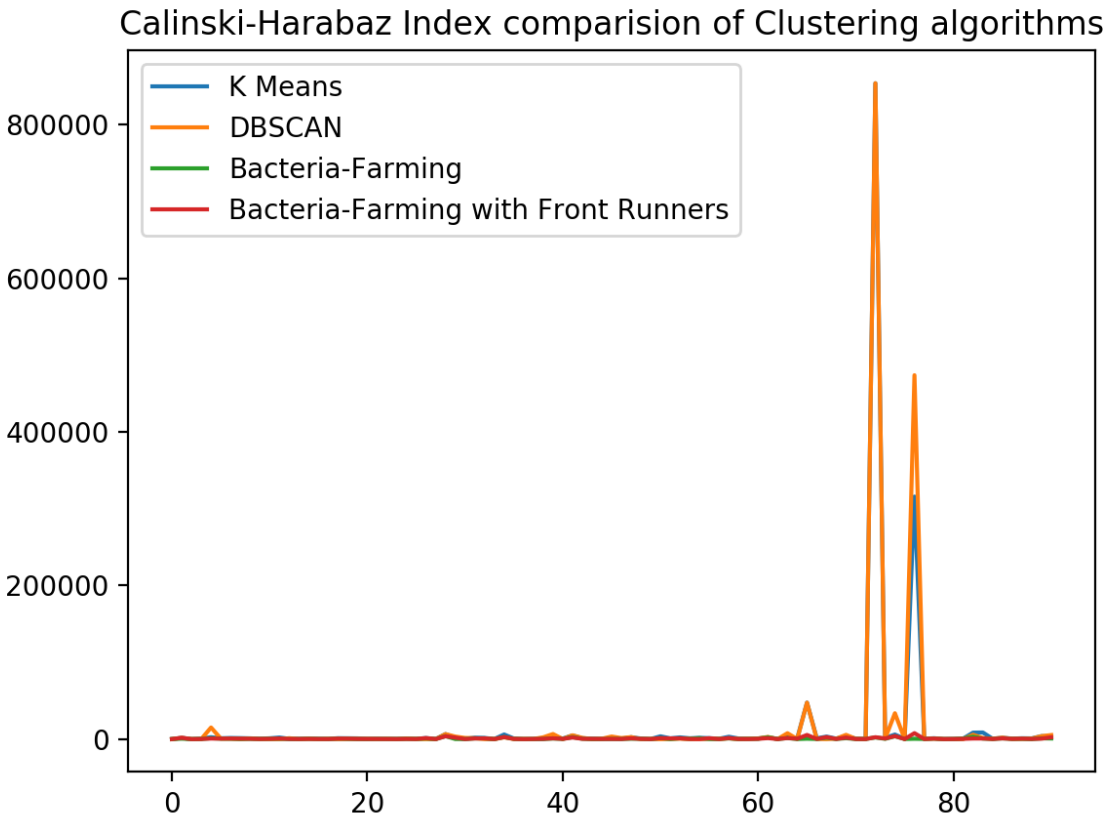
Inferences

1. DBSCAN has a high Silhouette Coefficient because it considers up to 90% of the data as noise, hence smaller clusters.
2. BFFR performs better than all its previous versions in almost all cases.
3. DBSCAN performs much faster than BFFR due to the presence of an optimization which reduces its complexity from O(n*n) to O(n*log(n)). A similar optimization can be used in BFFR to reduce its complexity from O(n*n*k) to O(n*log(n)*k).

Performance Comparison




Silhouette Coefficient comparision of Clustering algorithms
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
K Means
DBSCAN
Bacteria-Farming
Bacteria-Farming with Front Runners


Calinski-Harabaz Index comparision of Clustering algorithms
800000
600000
400000
200000
0
0
20
40
60
80
K Means
DBSCAN
Bacteria-Farming
Bacteria-Farming with Front Runners