CS546 Project Report

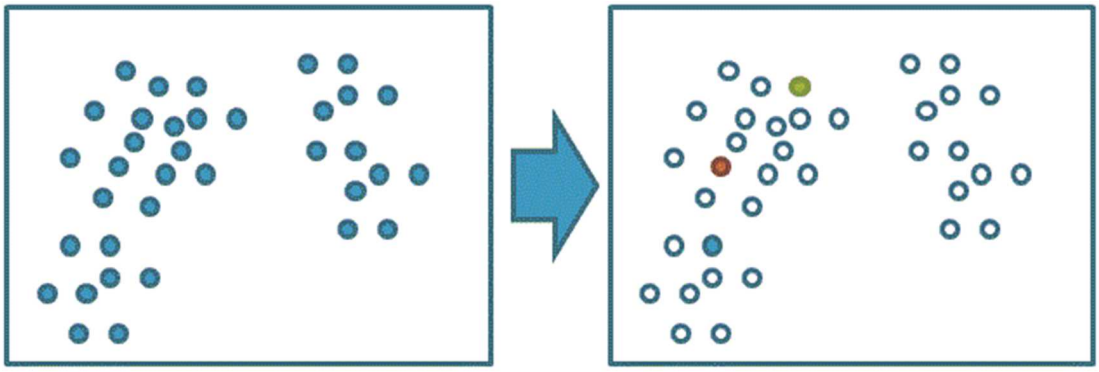
Yedong Liu

A20344124

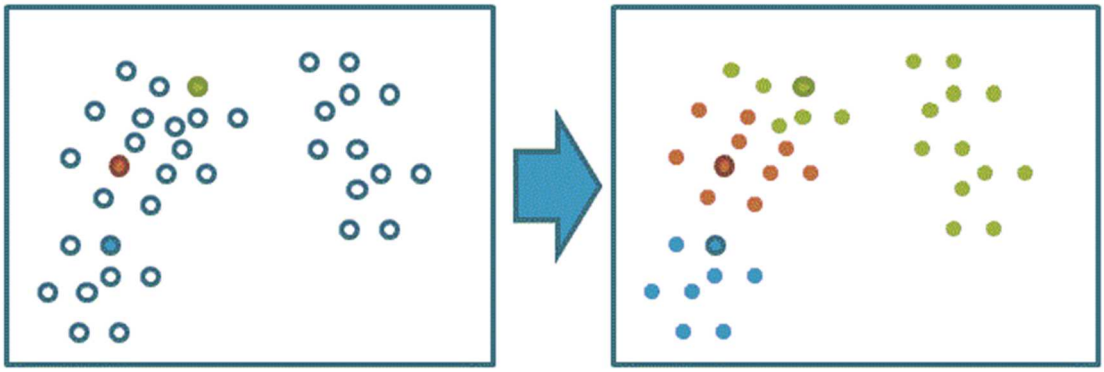
For k means clustering, I wrote a serial code together with a MPI version.

For the serial code:

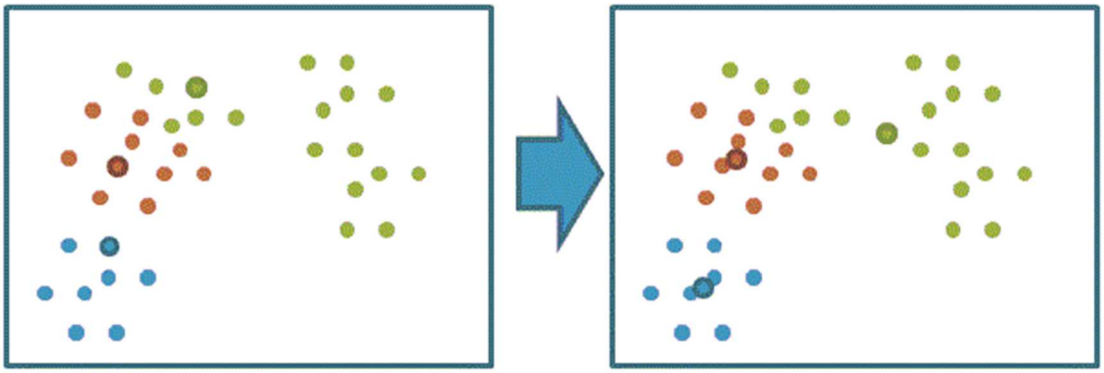
1. Randomly select k data points as the initial cluster centers



2. Calculate the distance between every data point and all cluster centers to determine which cluster this data point belongs to



3. Re-calculate the cluster centers based on the pivot point of every cluster



4. Repeat 2 and 3 until the error is small enough.

Progress:

Progress	Done or not
Algorithm design	Y
Coding structure	Y
Input format	Y
Random Input	N
Coding	Y
Testing	Half done
Modification	N, in doubt

For the MPI code:

Using master-slave structure, rank 0 will assign work to every other rank, and after every rank finished its calculation, it will send data back to rank 0.

Algorithm design:

1. Rank 0 is the master, reading input from the file and assign work to other ranks.
2. Rank 0 will calculate the cluster centers and send to other ranks.
3. Other ranks calculate the distance between cluster centers and all data points, identify the cluster the data point belongs to, then send back to rank 0.
4. Rank 0 re-calculate the centers and send to other ranks, then calculate the total distance of all data points to their cluster center.
5. Repeat 3 and 4 until the error is small enough.

Progress	Done or not
Algorithm design	Y(based on serial)
Coding structure	Y(based on serial)
Input format	Y(based on serial)
Random Input	N
Coding	Half done
Testing	N
Modification	N, in doubt