PROJECT 2-Liu Liu

Using Ecoli dataset

This dataset has 336 datapoints in total, each data point has 7 dimensions and 1 name. This dataset comprises 8 different protein classes.

Plot 1

For the data points and clusters is relatively small, when clusters increase, parallel\_for each for loop no matter based on data points or the number of clusters would not improve the performance.

Using Red Wine dataset

This dataset has 1599 datapoints in total, each data point has 11 dimensions.

Plot 2

White Wine data points are relatively bigger than Ecoli dataset, and TBB can do better of saving execution time of this scale of dataset.

First, parallel\_for would bring better performance than using serial. Second, it would get better performance in parallel\_for based on data points than both parallel\_for based on clusters and parallel\_for on datapoints and clusters.

Using Page Blocks dataset

This dataset has 5473 datapoints in total, each data point has 10 dimensions. This dataset has 5 classes.

Plot 3

Page Blocks dataset is relatively big, of course, TBB also do better of saving execution time of this scale of dataset.

First, parallel\_for would bring better performance than using serial. Second, when cluster increases, parallel\_for based on data points and parallel\_for based on both data points and cluster would bring better performance than parallel\_for based on data points. Third, using parallel\_reduce on sum in the for-loop of recalculating the center of each clusters brings better performance than not using parallel\_reduce.

Using letter recognition dataset

This dataset has 20000 datapoints in total, each data point has 16 dimensions and 1 name. This dataset comprises 26 different plant species.

Plot 4

For the data points are so big and clusters is so big, parallel\_for would bring benefits for saving execution time.

First, parallel\_for would bring better performance than using serial. Second, it would get better performance in parallel\_for based on data points than parallel\_for based on clusters. Third, parallel\_for on clusters and data points would get better performance than just using parallel\_for on each one. Forth, setting grain size may get better performance than using default.