Doruk ÖZER
70192

Report

	LM Cost	MD Cost	Time
Random k=4	1532.49	29286	01.52 second
Random k=8	1776.97	33322	01.58 seconds
Random k=16	1889.86	36162	01.74 seconds
Random k=32	1950.39	37906	01.93 seconds
Random k=64	1980.799	38386	01.85 seconds
Random k= 128	2000	38642	02.02 seconds
Cluster k=4	341.90	10934	14.10 minutes
Cluster k=8	594.21	16130	10.16 minutes
Cluster k=16	881.38	20882	05.33 minutes
Cluster k=32	1216.12	26546	02.40 minutes
Cluster k=64	1506.35	30962	01.10 minutes
Cluster k= 128	1744.94	34802	34.71 seconds

I didn't put the experiments results for the Bottom-Up Approach. **My Bottom-Up algorithm works correctly. I checked it with 50 data** and even if with 50 data it took a lot of time for k = 8. I am able to say it produced better results in terms of LM Cost and MD cost. Using bottom up approach leads less utility loss.

In terms of randomized approach it is the most time efficient option. Because it randomly clusters data (k is the size of the clusters except the last cluster) and makes them qui-wise equivalent. There is less amount of computational complexity but it has the most utility loss. Since all of the data is randomly clustered some of the clusters contains data points which are really far away in terms of distance metrices and the algorithm drives those data points to generalize a lot. However, if the data points were more near or similar in terms of distance metrics the clusters can be made qui-wise equivalent with less amount of computation

For cluster based approach LM cost and MD cost increases as the K increases. Because we are allocating most similar k number of values in a cluster and making generalizations, so as the k increase number of data in cluster increase and amount of generalization increase. It is unique that we observe time decreases as the k increases. Because amount of data handled in one iteration is increases as the k increases (note handled data doesn't handled again after it marked) so it quits from while loop much more quickly for k =128 then k =4. Randomized approach is much faster than cluster based approach. Because cluster based approach tries to find k number of most near points in terms of LM cost to a data point in each iteration so it increases the complexity of each iteration.

Bottom-up approach leads to least utility loss in data. It provides the least amount of generalization and preserves the utility of the data. It takes a lot of time due to the its computational complexity. I was not able to make the experiments for the bottom up approach with 2000 data so I didn't put it in the table.

If I have time constraint then I will use the cluster based approach but If I don't have time constraint then no matter what I will prefer bottom up approach since it provides the least amount of utility loss.