Performance Comparison of Centroid Based Clustering Algorithms

COMP-5704

Aagyapal Kaur

School of Computer Science

Carleton University

Ottawa, Canada K1S 5B6

aagyapalkaur@cmail.carleton.ca

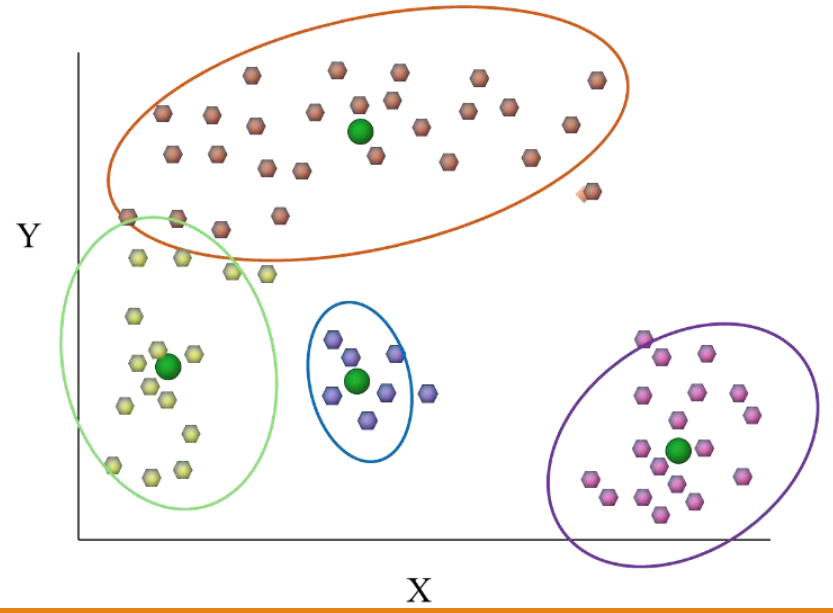
1. INTRODUCTION

- Enterprises today are dealing with the massive size of data.

- The need of the time is to extract, analyse, and process data in a timely manner.
- **Clustering** is an essential data mining tool for analysing the big data.

1.1 CLUSTERING

- It is one of the most popular unsupervised machine learning classification techniques.
- Dividing the data into clusters can be on the basis of **centroids, distributions, densities**, etc.
- **Centroid-based clustering** arranges the data into non-hierarchical clusters, in contrast to hierarchical clustering.

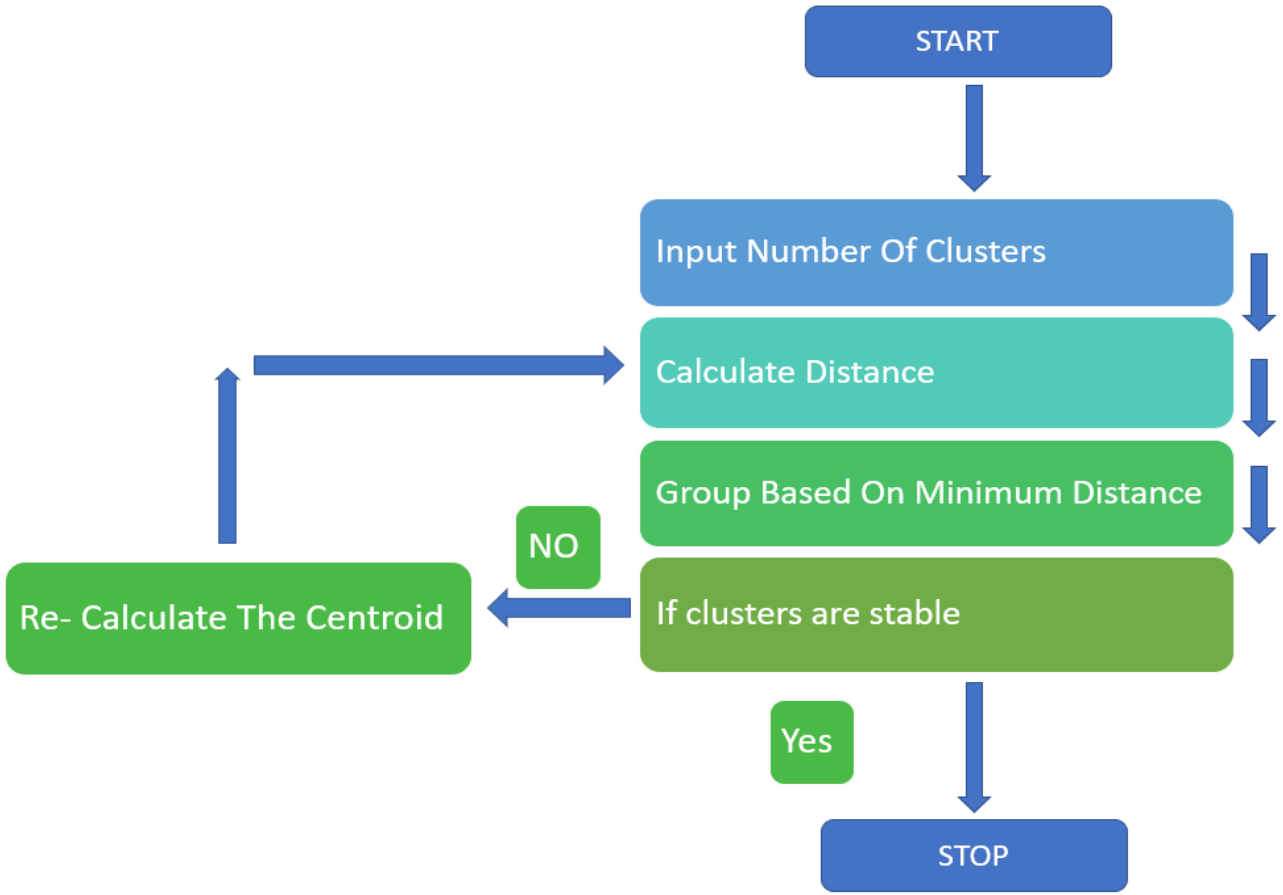


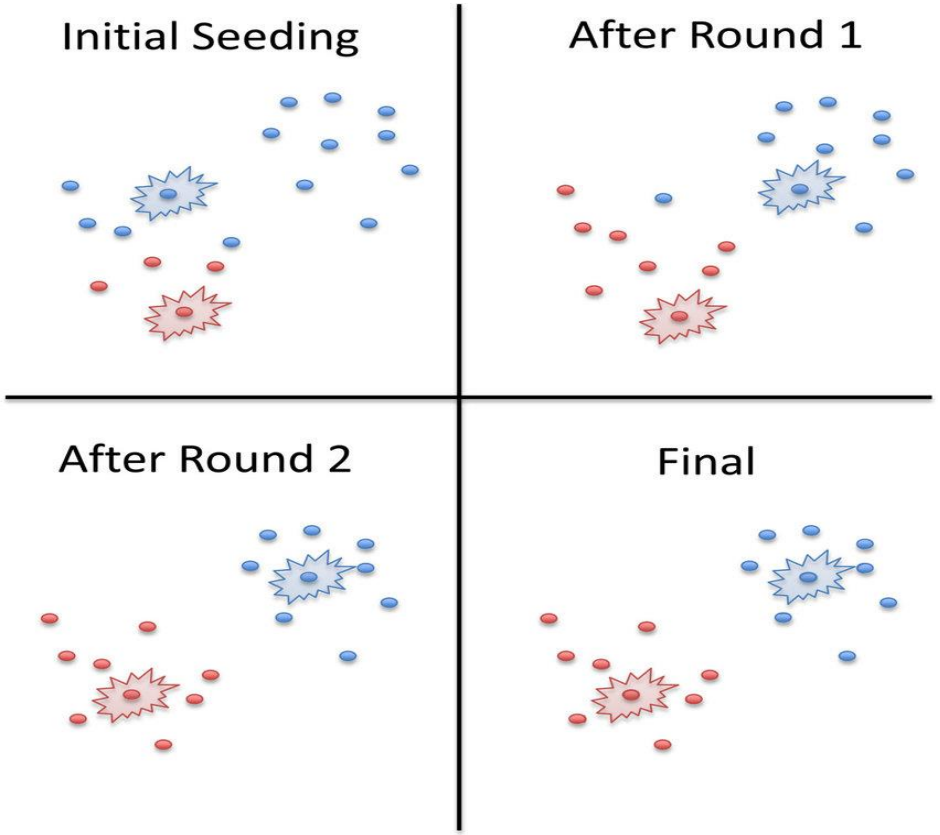
2. K-means Algorithm

- It is a well-known clustering algorithm for its simplicity and easy implementation.
- The objective is to partition N data objects into K clusters (K<N).
- The K-means algorithm requires three user-specified parameters:

1) number of clusters K,
2) cluster initialization,
3) distance metric

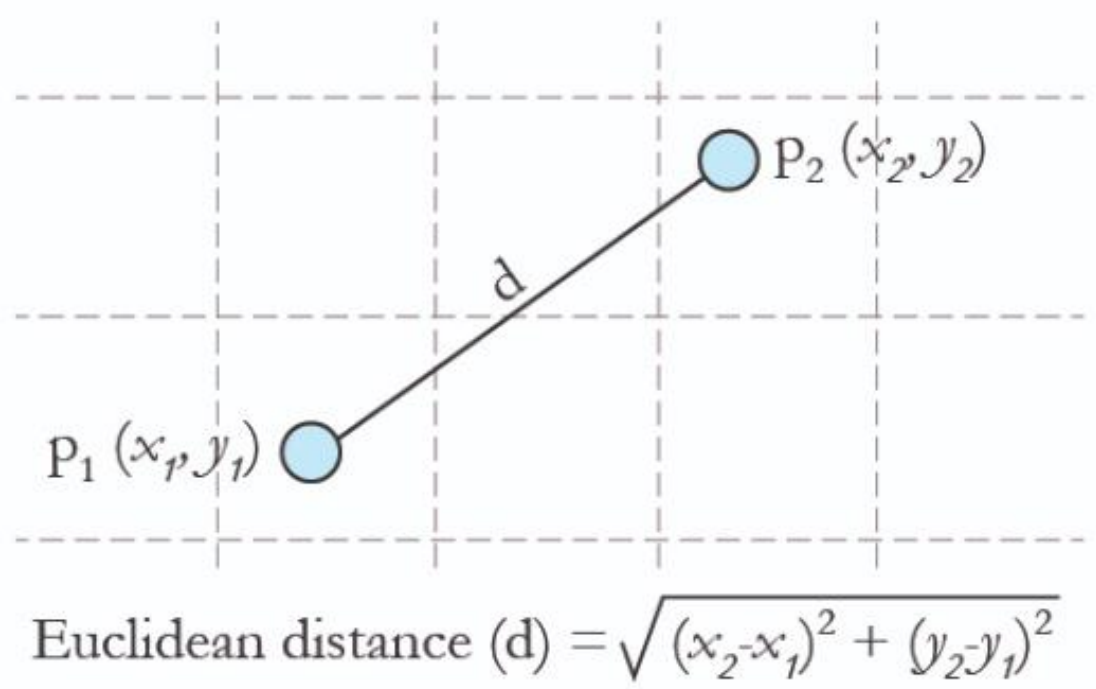
2.1 Flow chart: K-means clustering Algorithm





3. Distance Metric used in K-means Algorithm(Euclidean Distance)

- Euclidean distance formula can be used to calculate the distance between two data points in a plane.



4. New Approach to K-means i.e. K-means++

- The limitation of K-means algorithm is that it might be blocked locally based on the initial random chosen centers.

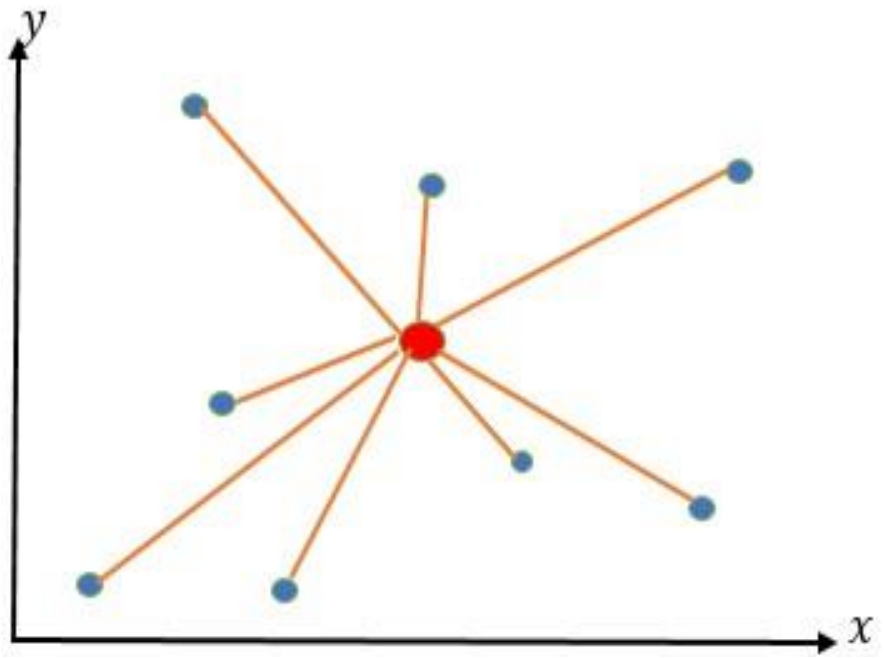
- K-means++ tries to choose a set of carefully selected initial centers instead of random initialization.
- This algorithm ensures a smarter initialization of centroids and ameliorate the quality of clustering.

4.1 Steps of K-means++ Algorithm

- The first centroid C_1 is selected randomly.
- Choose the next centroid C_2 , with probability proportional to

$$\frac{D\prime(m)^2}{\sum_{m\in M} D(m)^2}$$

- Repeat Step (2) until we have chosen a total of k centers
- Proceed as with the standard K-means algorithm



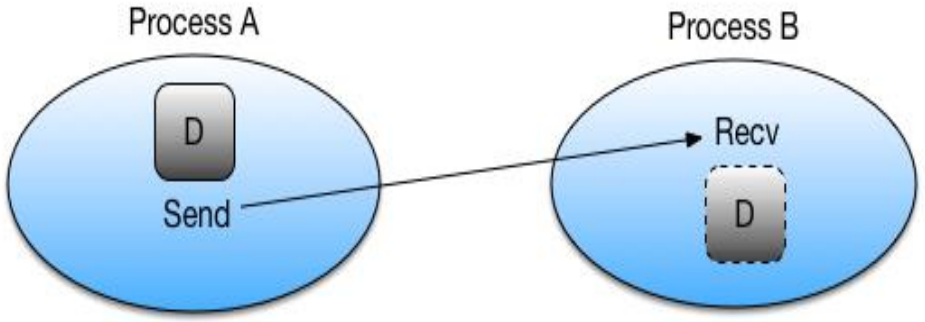
5. Need of Parallelism

- Performing operations on huge datasets to develop various different machine learning models, it takes a lot of time due to lack of parallelism.
- Hence, performing parallelism using the MPI is an effective way to get desired results in a shorter span of time.

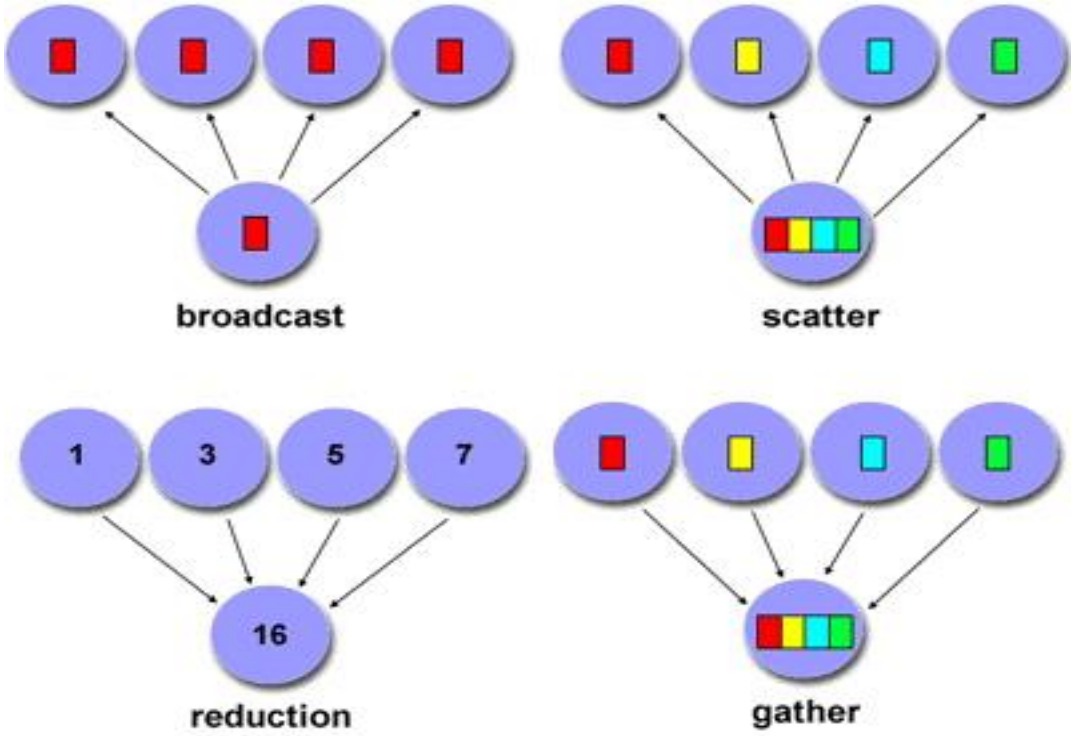
Message Passing Interface(MPI)

- *Message Passing Interface*, is a standardized and portable message-passing system designed to function on a wide variety of parallel computers.
- MPI support both point to point and collective communication.

5.1 MPI Framework



Point to Point communication



Collective communication

6. Parallelise the K-means using MPI

Parallel K-Means Algorithm

Input: Data, K Cluster

Output: K Centroid

1: MPI_INIT// start MPI Procedure

2: Read N object from file

/start parallel proccess by divide same amount of object to each processes/

3: **repeat**

4:Choose K point as intial centroid randomly

5: Initiate each object to the closest centroid by using Euclidean Distance Formula

6: **until** centroid don't change

/merge centroid procedure /

7: Generate cluster id to each object

8: Generate new centroid cluster by centroid result in each processes

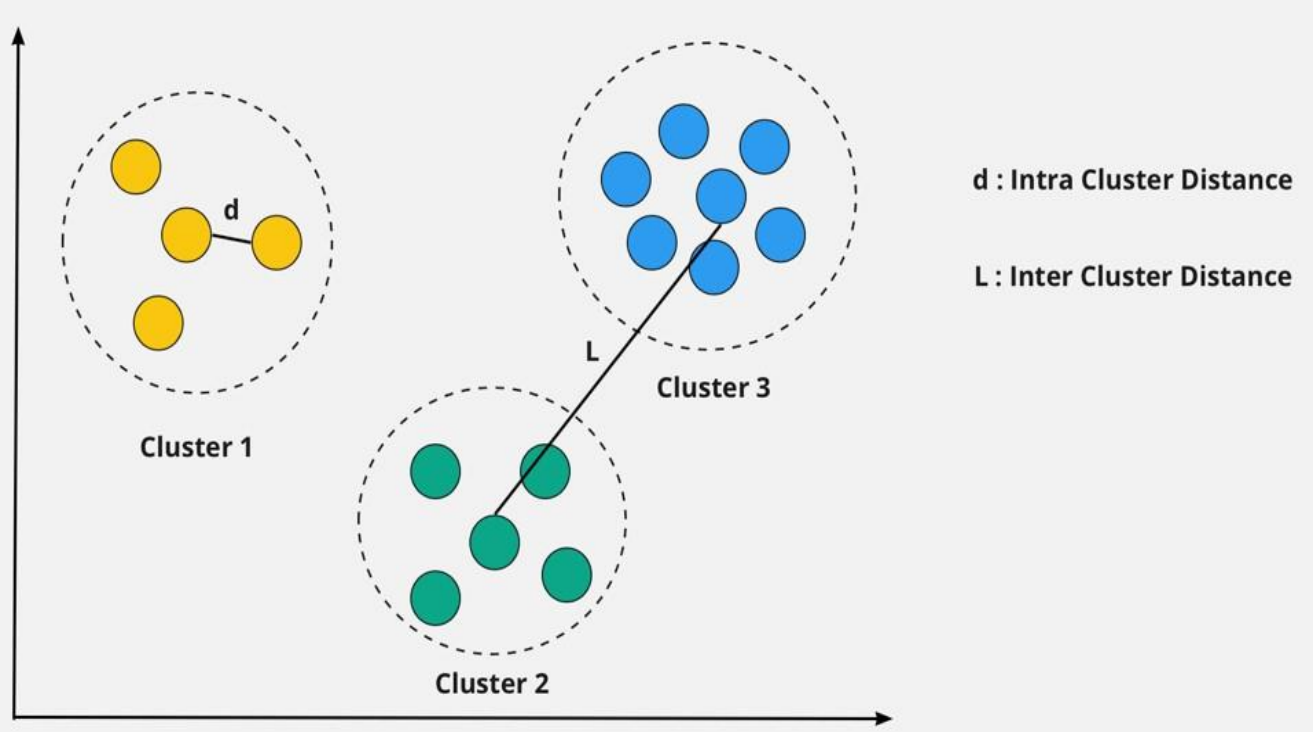
9: Generate final centroid

10: MPI_Finalize() // Terminate MPI Process

7. Performance metric (Sum of Squared Errors)

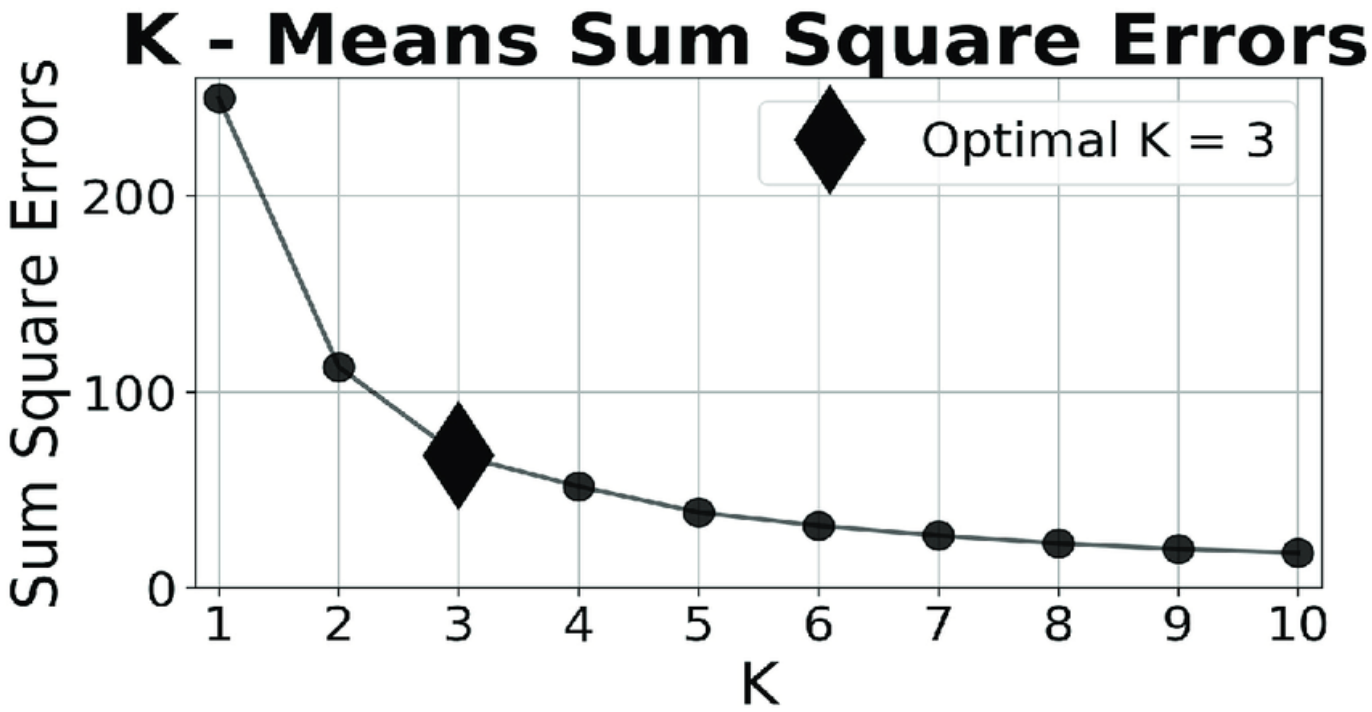
- The objective function that can be used for measure the quality of cluster is Sum Squared of Error (SSE).

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2$$



8.Using the elbow method to determine the optimal number of clusters

- The sum of squared errors(SSE) is used as a performance indicator.
- It Iterate over the K-value and calculate the SSE.



9. Result and Discussion

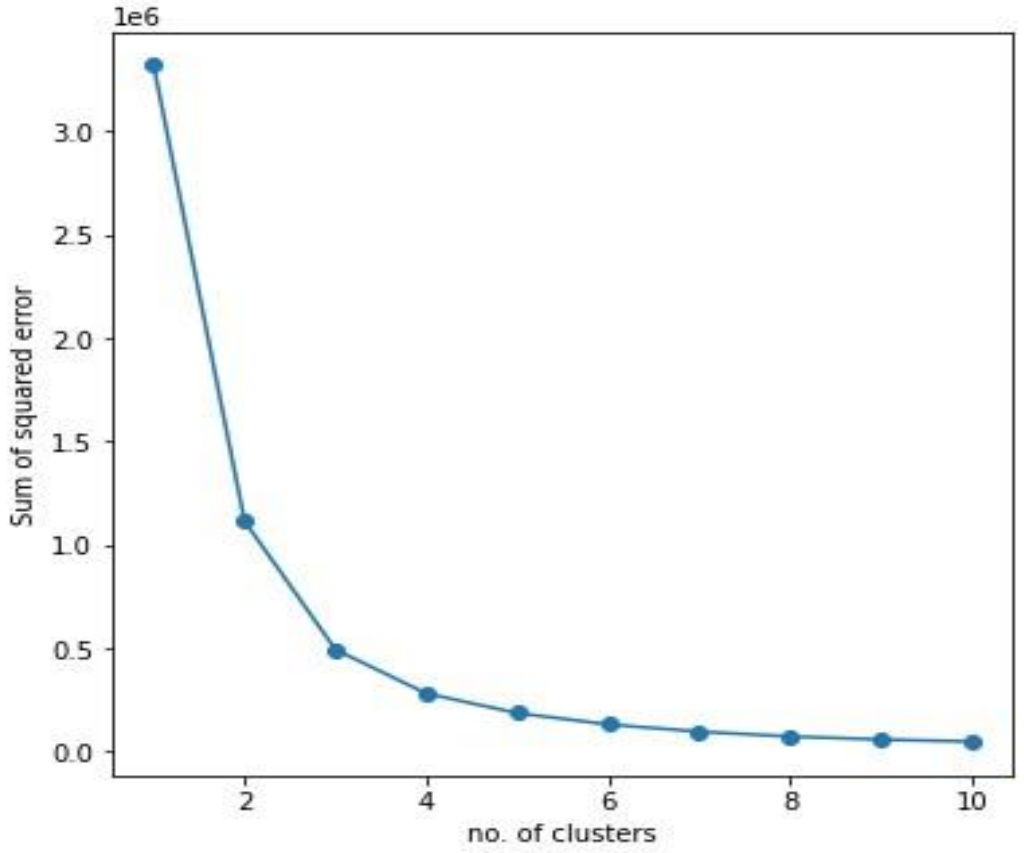
No. of Clusters	Np-2	Np-3	Np-4	Np-5	Np-6	Np-7	Np-8	K Means	K Means ++
K=2	5.954	4.040	3.639	4.686	5.788	5.992	6.502	11.499	8.359
K=3	10.336	7.644	7.324	7.865	8.398	9.134	9.654	16.931	12.426

Table 1: The run-times (in seconds) of parallel, sequential K-means and K-means++ clustering algorithm on 3D Road Network dataset

3D Road Network (North Jutland, Denmark) Data Set: This dataset was constructed by adding elevation information to a 2D road network in North Jutland, Denmark (covering a region of 185 x 135 km2) which contains 434874 samples and 4 features.

9.1 SSE(Sum of Squared Errors) Result

No. of Clusters	SSE
1	3319150.878768924
2	1115658.3960084815
3	494782.04378315335
4	282041.88108420983
5	187139.21500579204
6	132159.84742764695
7	97597.23462518021
8	73882.71137315751
9	59544.59791930749
10	49553.67750304196



10. Conclusion

- Experimental results of two clusters and three clusters formation show that K-means using parallel configuration is more faster, stable and portable, and it is efficient in the clustering on large data sets as compared to K-means and K-means++

References

1. https://www.visualcapitalist.com/wp-content/uploads/2021/11/data-never-sleeps-prev.png
2. https://editor.analyticsvidhya.com/uploads/36438flow2.PNG
3. https://www.researchgate.net/profile/Justin-Page/publication/268880805/figure/fig3/AS:282625324404757@1444394536795/k-meanss-clustering-algorithm-An-example-2-cluster-run-is-shown-with-the-clusters.png
4. https://www.tutorialexample.com/wp-content/uploads/2020/05/Euclidean-distance-in-tensorflow.png
5. https://nyu-cds.github.io/python-mpi/fig/02-send-recv.png
6. https://pages.tacc.utexas.edu/~eijkhout/pcse/html/graphics/collective_comm.jpg
7. https://dataaspirant.com/wp-content/uploads/2020/12/8-Intra-Cluster-Distance-and-Inter-Cluster-Distance.png
8. https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29

Questions

1. What is Clustering and why it is so popular?
2. What is SSE ?
3. Why we need parallelism ?

Thank You