15-440/640 Distributed Systems

Lab 4 Report: MPI: Clustering and DNA

Team:

Abhishek Bhowmik (abhowmi1),

Neil Rajesh Dhruva (ndhruva)

Date:

12/05/2014

**Contents**

**Problem Definition ………………………………………………………………….. 3**

**Solution Overview …………………………………………………………………… 4**

**Framework ……………………………………………………………………………… 6**

**Code Organization …………………………………………………………………… 11**

**Conclusion ………………………………………………………………………………. 13**

**Problem Definition**

Clustering is a task of grouping objects that are more similar to each other as compared to other objects in the dataset. Hence, the objects with the closest ‘distance’ to each other are put in one group. The goal of this project is derived from the idea of grouping similar objects using a distance metric in a parallel computation setting. The two main objectives of the project include:

1. Understand and implement a parallel computation algorithm for clustering in large datasets using the OpenMPI framework.
2. To analyze the performance of a parallel algorithm, and compare it with a sequential implementation of the same, for varying dataset sizes and varying degrees of parallelism.

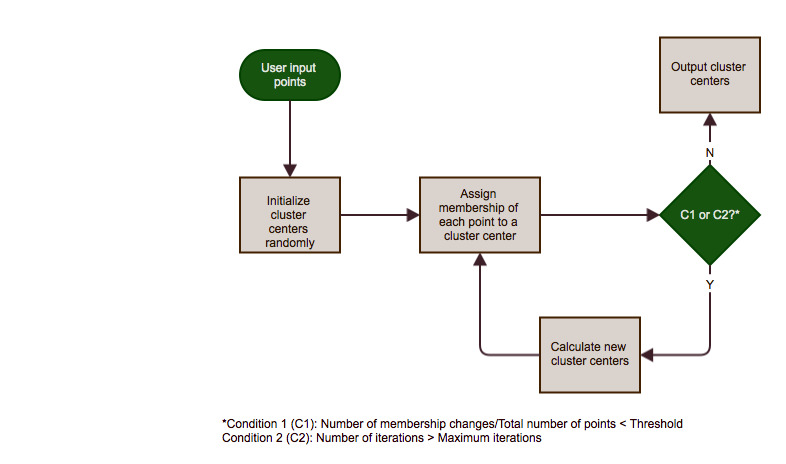
**Solution Overview**

For this project, we have created parallel algorithms for clustering points in 2D space using the K-means clustering technique, and generalized it for use in the DNA clustering application. In addition, we have provided implementations for generating random datasets of user-defined sizes for both these problems.

For solving the clustering problem using K-means, we use the Euclidean distance between data points to find the data center closest to a given point. The point is then a ‘member’ of the cluster centered at that data center. In an iterative manner, all points closest to a given cluster center are identified, and the cluster center is recalculated as an average of these points. This calculation is parallelized by distributing the dataset equally among all participating nodes, calculating the membership of points locally on those nodes, and recalculating cluster centers using OpenMPI for communication between the nodes.

This idea is then extended to calculate cluster centers in a dataset containing DNA strands (randomly generated using a script provided with the code). The closest center to each strand is calculated using the maximum frequency technique for each DNA base at a given position in the DNA strand. Using OpenMPI for communication, K-means is parallelized for better performance across several nodes.

**Clustering Overview**



**OpenMPI Overview**

time /usr/lib64/openmpi/bin/mpirun --mca btl\_tcp\_if\_include eth0 -np 2 -machinefile public/machines public/outmpi -n 5 -p 100000 -v1 -i public/cluster5\_100000.csv

--flowchart

**Implementation**

The implementation has been divided into two parts, one each for the corresponding applications of 2D data points and DNA strands. The major difference between these two implementations lies is the calculation of cluster centers from the points belonging to the cluster.

Both implementations are divided into the following phases:

1. Dataset generation: Dataset generation for the 2D dataset is done using the starter script provided with the handout. However, we have written a similar script to generate the dataset for DNA strands. We begin by choosing a certain (user supplied) number of centers, and generate equal number of data points close to these centers, as implemented in the 2D dataset generator. The key difference here, of course, is the representation of the DNA strands dataset. We use comma-separated strands, where is base has value ‘a’, ‘c’, ‘t’ or ‘g’.
2. User input: Invoking the scripts to generate the dataset and to run the sequential and MPI implementations of clustering is provided in the README file. We mainly require, from the user, the input file, number of points in the dataset, number of clusters to be generated, and the dimensions of the dataset (especially in the case of DNA strands). We assume, as mentioned in the handout, that the DNA strands are all of equal length.
3. Centroid Calculation:
   1. *2D Dataset:* For the 2D dataset, we represent each input point as an array of double values. Hence, for calculating centroids, we sum up all points that are members of a particular cluster, and take the average of these points by dividing the sum with the total number of points belonging to that center. This gives us the new center for that cluster.
   2. *DNA Strands Dataset:* For the DNA strands dataset, we use a different approach for centroid calculation. The distance function for this dataset is defined by the maximum number of bases in the DNA strand that are equal between the strand under consideration, and the centers. Hence, a new center (centroid) is calculated based on the maximum frequency of a base at a given position in the strand from among all strands belonging to a cluster.
4. Sequential KMeans: For sequential Kmeans calculation, the input is read from user file, and an initial set of centroids (randomly selected from the dataset), equal in number to the user supplied cluster number parameter, is generated. This is followed by the membership assignment phase, where each point in the dataset is assigned to its closest center. Now, the new centers are calculated based on the centroid calculation techniques mentioned in point 3 above. Finally, we make use of two stopping conditions, whichever is satisfied first:
   1. We use the ratio of number of points that change membership in an iteration, to the total number of points. If this ratio is low, then the centroid calculation is stopped.
   2. We also use a maximum iterations parameter. If the number of iterations it takes to converge to the actual centers reaches this maximum limit, then the process is stopped even if the threshold in (a) is not met.
5. Distributed KMeans: For distributed Kmeans calculation, we use the OpenMPI platform. We run our Kmeans calculations in parallel on several processors using this platform. The input is read from the user input file by the master in a manner similar to sequential Kmeans. However, in addition, the master then distributed equal number of points to all the processors (worker nodes) involved, including itself. Additionally, initial cluster centers are randomly chosen from the dataset provided by the user and are ‘broadcasted’ to the worker nodes by the master.

On each worker node, of which the master is also a part now, the membership of each point assigned to that node is calculated by calculating the closest cluster centers. In addition, each node calculates the sum of all points belonging to a cluster for the 2D dataset, and the maximum frequency of all strands in the DNA dataset locally. Once all nodes finish, the ‘all reduce’ functionality provided by OpenMPI is used to get the global sum for each cluster (for the 2D dataset), and the global frequency count for each cluster in the DNA dataset. Each node also adds up the number of points belonging to each cluster center locally, and this sum is then added using the all reduce function to get a global sum of the total number of members per cluster.

Each node now has information on the total sum or frequency count for each cluster, and the total number of members in the cluster. Hence, each node can individually calculate the new cluster centers and proceed with the next iteration on the points assigned to it.

The stopping conditions are enforced by keeping track of the number of membership changes that take place in each iteration, globally, using the all reduce function. Hence, each node undergoes the same number of iterations, and exits the loop simultaneously when a stopping condition is met.

**Analysis**

For analyzing the performance of our distributed Kmeans implementation against the sequential implementation, we test for increasing number of data points for varying number of cluster centers, and increasing number of parallel nodes executing the algorithm. Since our algorithm selected random initial cluster centers, we ran the algorithm for each triple of (#points, #clusters, #processors) multiple times (typically, 3) and took the average run time for each such triple. Hence, it is a good approximation of the total time taken for each such triple. We ran our tasks on a total of 5 machines.

For the 2D dataset, the following tables show the run time in seconds (obtained through the time command in Unix). The graphs show the trends observed in the data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Num clusters = 5** | |  |  |  |  |  |
| **Num points** | **Seq** | **P=2** | **P=4** | **P=6** | **P=8** | **P=12** |
| **100000** | 3.552 | 4.039 | 4.098 | 4.725 | 4.946 | 4.631 |
| **500000** | 8.403 | 6.746 | 6.185 | 6.35 | 6.272 | 5.943 |
| **1000000** | 13.233 | 9.608 | 6.956 | 6.695 | 6.186 | 6.901 |
| **2000000** | 21.809 | 15.658 | 10.01 | 8.681 | 8.831 | 8.39 |
|  |  |  |  |  |  |  |
| **Num Clusters = 10** | |  |  |  |  |  |
| **Num points** | **Seq** | **P=2** | **P=4** | **P=6** | **P=8** | **P=12** |
| **100000** | 11.079 | 10.01 | 9.471 | 10.322 | 10.181 | 12.69 |
| **500000** | 26.613 | 16.322 | 12.782 | 14.198 | 12.164 | 11.923 |
| **1000000** | 48.103 | 28.283 | 19.827 | 16.169 | 16.832 | 16.904 |
| **2000000** | 55.882 | 36.784 | 24.333 | 20.18 | 17.497 | 21.249 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **Num Clusters = 15** | |  |  |  |  |  |
| **Num points** | **Seq** | **P=2** | **P=4** | **P=6** | **P=8** | **P=12** |
| **100000** | 21.261 | 15.539 | 17.982 | 17.435 | 19.357 | 17.48 |
| **500000** | 33.224 | 26.744 | 24.111 | 19.247 | 20.405 | 18.666 |
| **1000000** | 85.805 | 38.148 | 43.525 | 26.483 | 23.462 | 24.253 |
| **2000000** | 110.79 | 55.124 | 40.683 | 35.954 | 30.954 | 33.172 |

Figure 1: 5 Clusters

Figure 2: 10 Clusters

Figure 3: 15 Clusters

The tables and charts illustrate a clear trend of improved performance due to an increase in the number of processors. However, we observed the following details in the tables:

1. Due to the randomized initialization of cluster centers, we will get varying convergence rates. Some of the observations, like in Figure 3, P=8, and the line corresponding to 500000, we get a small increase in time which is unlikely. This is possibly a consequence of the random initialization of cluster centers.
2. Nonetheless, the general trend in almost all cases is that the amount of time taken for the entire process of IO and clustering decreases with increase in the number of parallel clustering jobs for a given dataset. We see that we achieve an increase of up to 4 times when using between 4-8 machines. But, we observe marginal diminishing returns after P=4, and the speedup does not scale with the number of worker nodes.
3. In fact, we observed that in many cases, that the time goes up for P=12, as compared to P=8. The reason for this could be the fact that our datasets are not large enough for the need to run them on 12 worker nodes. In this case, the communication time between the nodes is higher, without providing any time benefits for clustering.
4. This leads us to believe that, for the size of our 2D datasets, P=8 is kind of a sweet spot, and that we cannot expect much better performance with increase in P beyond 8.
5. In addition, we also observed that the more the number of clusters, the more the time it takes to converge. This is evident from the increase in the time for each execution for #Clusters=15 as compared 10, and the same for 10 compared to 5.
6. Though not evident form the dataset, we observed speedups due to reading a file from the cache. The first time a large input file is read, the IO time is quite large. However it significantly decreases on reading the same file immediately again.