Using Parallel Computing to Performing the Cluster Analysis(K-means)

CS566 – Parallel Computing
Rajarajan Loganathan
Hood College
Department of Computer Science
Frederick, MD, USA

Abstract-Clustering is one of the most popular methods for exploratory data analysis, which is prevalent in many disciplines such as image segmentation, bioinformatics, pattern recognition and statistics etc. The most famous clustering algorithm is K-means because of its easy implementation, simplicity, efficiency and empirical success. However, the real-world applications produce huge volumes of data, thus, how to efficiently handle of these data in an important mining task has been a challenging and significant issue. In addition, MPI (Message Passing Interface) as a programming model of message passing presents high performances, scalability and portability. Motivated by this, a parallel K-means clustering algorithm with MPI, called M-Kmeans, is proposed in this paper. The algorithm enables applying the clustering algorithm effectively in the parallel environment. Experimental study demonstrates that M-Kmeans is relatively stable and portable, and it performs with low overhead of time on large volumes of data sets.

Keywords – Kmean Clustering model, MPI, parallel processing

I.Introduction:

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their

further processing. Data clustering has many engineering applications including the identification of part families for cellular manufacture. The K-means algorithm is a popular data clustering algorithm. To use it requires the number of clusters in the data to be pre-specified. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process made more difficult by the subjective nature of deciding what constitutes 'correct' clustering . This paper proposes a method based on information obtained during the K-means clustering operation itself to select the number of clusters, K. The method employs an objective evaluation measure to suggest suitable values for K, thus avoiding the need for trial and error. The remainder of the paper consists of five sections. Performing K-Mean algorithm using parallel processing or computing. Parallel data mining algorithms have been recently considered for tasks such as association rules and classification, see, for example, Agrawal and Shafer [6], Chattratichat et al. [7], Cheung and Xiao [8], Han, Karypis, and Kumar [9].

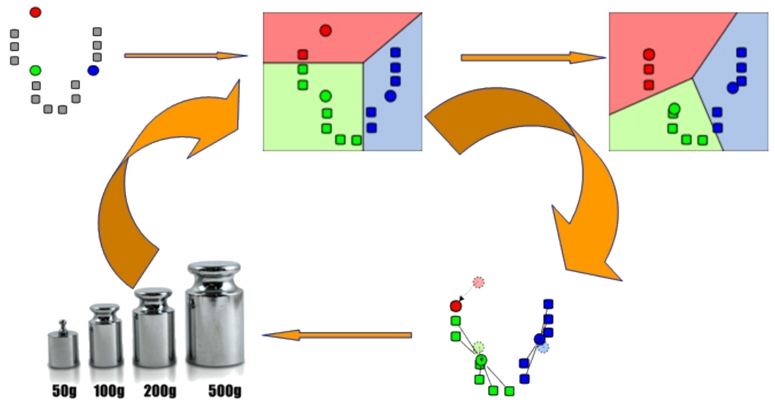
II.Parallel k-means :

Our parallel algorithm design is based on the Single Program Multiple Data (SPMD) model using message-passing which is currently the most prevalent model for computing on distributed memory multiprocessors; we now briefly review this model.

III.Message-Passing Model of Parallel Computing:

We assume that we have P processors each with a local memory. We also assume that these processors are connected using a communication network. We do not assume a specific interconnection topology for the communication network, but only assume that it is generally cheaper for a processor to access its own local memory than to communicate with another processor.

The message-passing model posits a set of processes that have only local memory but are able to communicate with other processes by sending and receiving messages. It is a defining feature of the message-passing model that data transfers from the local memory of one process to the local memory of another process require operations to be performed by both processes.



MPI, the Message Passing Interface, is a standardized, portable, and widely available message-passing system designed by a group of researchers from academia and industry [1, 2]. MPI is robust, efficient, and simple-to-use from JAVA.

From a programmer's perspective, parallel computing using MPI appears as follows. The programmer writes a single program in JAVA compiles it, and links it using the MPI library jar file. The resulting object code is

IV.Conclusion:

As data clustering has attracted a significant amount of research attention, many clustering algorithms have been proposed in the past decades. However, enlarging data in applications makes clustering of very large scale of data a challenging task. In this paper, we propose a fast parallel k-means clustering algorithm based on the above, which has been widely embraced by both academia and industry. We use speedup, scaleup and sizeup to evaluate the performances of our proposed algorithm. The results show that the proposed algorithm can process large datasets on commodity hardware effectively.