ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Parallel Clustering of Gene Expression Dataset in Multicore Environment

Pranoti Kamble, Prof. Rakhi Wajgi

Dept. of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

Professor, Dept. of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

ABSTRACT: Clustering has become the powerful and widely used method in gene expression dataset analysis to obtain biological information. Clustering using sequential approach is time consuming task. So as to save time and to increase speedup we have applied parallel clustering on single machine utilizing computational power of multicore processors in the system. In this work, we have also done comparison of sequential clustering and the parallel clustering in terms of time consumed for clustering of yeast gene expression dataset. With the use of multicore processor the speedup gained from 5 to 7% on intel core 2 duo processor with 2Gb RAM.

KEYWORDS: Clustering, Gene expression dataset, Multicore.

I. INTRODUCTION

Data mining extracts knowledge from large amount of dataset. Clustering is the process of grouping data objects into a set of disjoint sets, called clusters So that objects within a class have high similarity to each other while objects in different classes are more dissimilar [1]. Clustering of gene expression data can be divided into two main categories: Gene-based clustering and Sample-based clustering. In gene based clustering, genes are treated as objects and samples are features or attributes for clustering. The goal of gene-based clustering is to identify differentially expressed genes and sets of genes with similar expression pattern or profiles, and to generate a list of expression measurements. In Sample based clustering, samples are treated as objects and genes are features for clustering. Sample based clustering can be used to reveal the phenotype structure or substructure of samples. The purpose of clustering gene expression data is to reveal the natural structured inherent data and extracting useful information from noisy data.

We are implementing clustering on the yeast gene expression dataset with 494 significantly identified genes. Yeast dataset contains information about bacterial microorganism. These organisms are useful in the development of foreign proteins. These proteins are useful in the production of insulin.

Clustering tells about the biological structure of the gene expression dataset. For the clustering process we have proposed the k-means algorithm with better centred method. The basic k-means clustering have some drawbacks: (1)selection of randomized k-value which sometime can form wrong cluster So to overcome this drawback we have proposed k-means centred algorithm. (2) Conventional k-means consumes lot of time.

In our approach we have assumed formation of a cluster as single task. In K-means algorithm each task is executed one after another. Applying parallelization to k-means algorithm in multicore environment exploits CPU power which almost goes waste in sequential execution. Fig. 1 shows the typical multicore architecture used for implementation. Parallelization is a process consists of breaking up large single process into multiple smaller independent tasks that can run in parallel on different cores and once finalized can be combined to obtain an overall improvement in performance.

This paper divides into following section. Section II deals with survey of existing clustering algorithms. In section III proposed methodology is discussed with respect to sequential and parallel approach. Section IV exhibited experimental results and paper ends with conclusion in section V.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

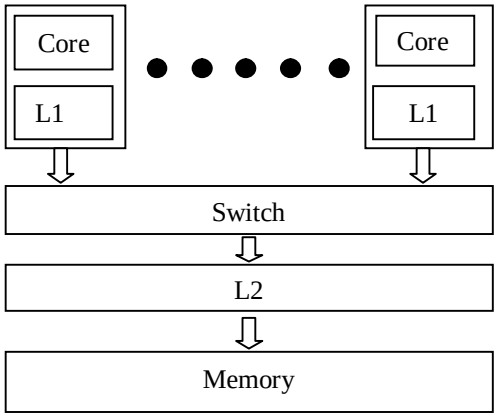

Fig. 1:Typical Multicore Architecture

II. RELETED WORK

Analysis of gene expression dataset is very important research topic in bioinformatics to identify diseases and other factors. Data mining have widely useful in gene expression dataset. The author Daxin Jiang et al. has analysed different clustering methods. The author categorised methods into gene based clustering consists of K-means and hierarchical clustering, model based clustering and density based hierarchical clustering. The author applied clustering methods to assess the quality and reliability of clustering results in [1]. Author T.Chandrasekhar et al. proposed new algorithm ECIA (Enhanced Centroid Initialization Algorithm) to overcome the drawback of k-means algorithm of randomized selection of K-values [9]. Author K.Y.Yeung et al. proposed method FOM (figure of merit) for assessing the quality of cluster results [10]. Soriful Horique et al. presents static methods to find regulation of gene from statistical gene expression data, comparison of genes having similar regulation and extraction of sub cluster from bib cluster and perform hierarchical analysis [8]. Author Sanjay Kumar Sharma et al. describes the parallel algorithms for computing the solution of dense system of linear equation using OpenMP interface [2]. The author Noha A. Yousri investigated the application of a novel validity measure to gene expression clustering. The leukaemia and breast cancer dataset are used to evaluate the applicability of the proposed validity to discover the natural connectedness of expression patterns [7].

III. METHODOLOGY USED

3.1 Sequential k-means Clustering:

K-means clustering algorithm have drawback of forming wrong cluster because of randomized selection of K-value i.e. cluster centroid. To overcome this drawback and to improve the quality of clusters we proposed better centred K-means clustering algorithm. In this, first we select initial random K- data points. Then again the K- data points will be selected from the formed clusters where we will assign the calculated data points again to the clusters centroid.

Euclidean distance:

$$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots . + (x_n - y_n)^2} \quad \text{.....equation (1)}$$

K-means Pseudocode:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

D= Gene expression dataset
K – Number of clusters
1) Enter the number of cluster
2) Choose randomized k value
3) Calculate Euclidean distances with respect to k value
4) Make clusters with minimum distances values
5) For the better clusters, choose again the pre-selected k- value centre from the clusters formed.
6) Again execute step 3) and step 4).
7) Exit

3.2 Multicore Clustering:

We have multicore processor in our system but the utilization of multicore system don't happen the way it should be. So in order to harvest the full power of a multi-core processor the software application must be able to execute tasks in parallel utilizing all available CPUs. Parallelization of a process consist of breaking up a large single process into multiple smaller tasks that can run in parallel and once finalized can be combined obtaining an overall improvement in performance. The result is the execution of a single task or process by multiple processors or CPUs "Parallel Processing", not to be confused with concurrency.

Executor framework (java.util.concurrent. Executor) was released with jdk 5 in package java.util.concurrent to achieve asynchronous parallelization, improve performance and take advantage of a multicore environment.

The introduction of the ExecutorService in the concurrent package make it easy to submit tasks and the service will execute them and hand us back a future object that we can use to test the progress and to obtain the result of a given executed task once it finishes. The ExecutorService will use threads internally to achieve parallelization and since the tasks are submitted we can also execute the tasks asynchronously, once all the tasks at hand are submitted we can move ahead in the main thread and do something else or just wait and monitor for the tasks that end to get the results making possible to achieve fork/join processing. With the executor class the gene expression dataset divides threads to multiple core. The core will execute threads internally as in clustering we have selected randomized K- value. so core will be allocated with the selected k-value Euclidean distances will be calculated parallel. And the formed cluster will be joined by fork/join method. In the end we get the combined clusters.

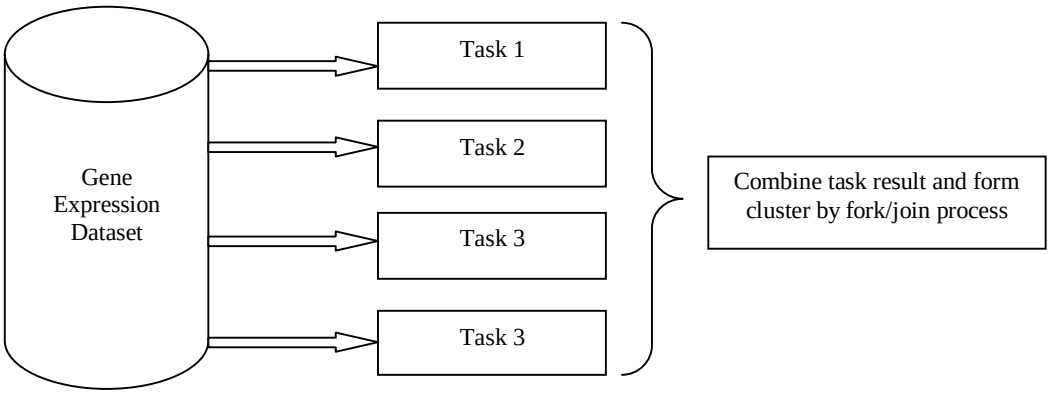

Fig. 2: Parallel clustering concept

IV. EXPERIMENTAL RESULTS

Fig. 3 shows the working of sequential clustering. We have formed 5 cluster in 3.042 sec while the time consumed in parallel rapidly decreases to 0.593 sec. so with the utilization of multicore architecture of system we saved upto 3 sec(3.042s-0.593s) with minimum speedup of 5.1%

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

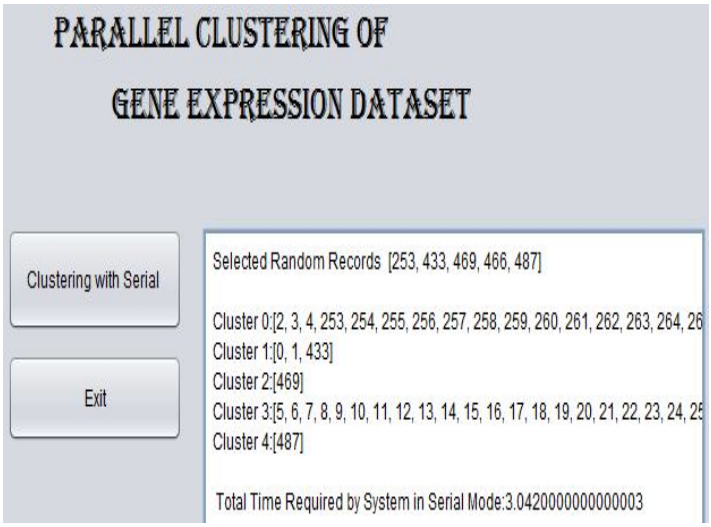

Fig. 3: Serial Clustering

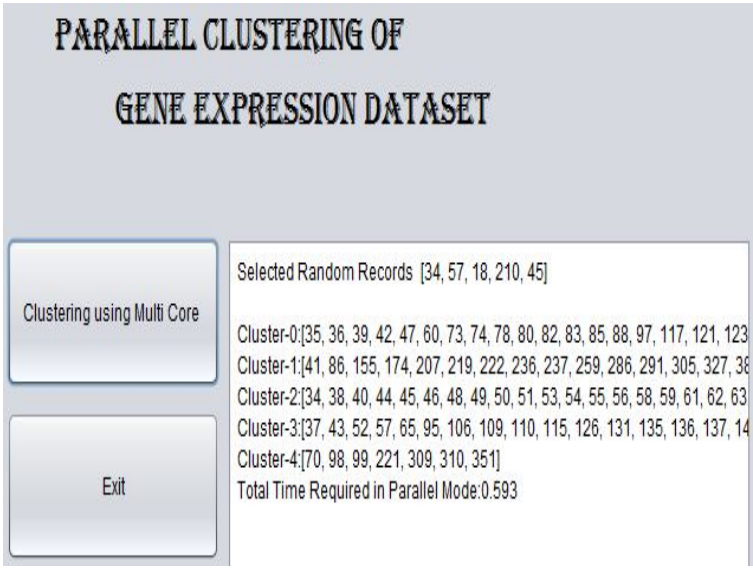

Fig. 4: Parallel Clustering

I am running the clustering application in core 2 duo processor which have two CPU's. The performance view under the windows task manager shows the CPU usage of only 4% of the total CPU power combined as shown in fig. 5. Using the ExecutorService to process the clustering task in parallel utilizes 19% of the total CPU power which results in 15% increased utilization as shown in fig. 6.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Fig. 5: CPU utilization when serial processing going on

Fig. 6: CPU utilization when parallel processing going on

Comparison Of Sequential Clustering and Multicore Clustering:

I have core 2 duo Intel processor. The speedup will increase with i3, i5 and i7 processor.

Table 1: Comparison of Clustering on different platforms

Core 2 duo processor:

Dataset used	Data Points	No. of Clusters	Time(in sec)	
			Sequential	Multicore
Yeast	419	5	3.042	0.593

Minimum Speedup=5.1%

V. CONCLUSION

We have executed clustering on dual core processor in which we got increased speedup of 5.1. With the execution of java utility provided concurrent package ExecutorService we have successfully increased the utilization of processors from 4% to 19%.

REFERENCES

1. Daxin Jiang, Chun Tang, and Aidong Zhang, "Cluster analysis for gene expression data: a survey", IEEE Transactions on knowledge and data engineering, vol. 16, no. 11, November 2004.
2. Sanjay Kumar Sharma and Dr. Kusum Gupta, "Performance analysis of parallel algorithms on multi core system using OpenMP", International journal of computer science, engineering and information technology (IJCSEIT), vol.2, no.5, October 2012.
3. Kittisak Kerdprasop and Nittaya Kerdprasop, "Parallelization of K-means clustering on multi-core processors", ACS'10 Proceedings of the 10th WSEAS international conference on applied computer science pages 472-477 2010.
4. S. N. Tirumala Rao, E. V. Prasad and N. B. Venkateswarlu, "A critical performance study of memory mapping on multi-core processors: an experiment with k means algorithm with large data mining data sets", International journal of computer applications pages 0975 – 8887 volume 1, no. 9, 2010.
5. Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, SeungHee Bae, George Chrysanthakopoulos and Henrik Nielsen, "parallel data mining on multicore clusters", 7th International conference on grid and cooperative computing, pages 41-49, Oct 2008.
6. Ping Guo and Xiaoyan Deng, "Gene Expression Data Cluster Analysis", IEEE WASE International conference on information engineering 2009.
7. Noha A. Yousri, "On the Validation of Gene Expression Clusters", 5th Cairo International biomedical engineering conference, december 2010.
8. Sushmita Mitra and Sampreeti Ghosh, "Feature selection and clustering of gene expression profiles using biological knowledge", IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 42, no. 6, November 2012.
9. Soriful Hoque, Salim Istyaq and Md. Mushir Riaz, "*A* hierarchical approach for clustering and pattern matching of gene expression d*ata"*, 6th International conference on genetic and evolutionary computing, 2012.
10. T.Chandrasekhar, K.Thangavel, E.Elayaraja and E.N.Sathishkumar, "unsupervised gene expression data using enhanced clustering method", IEEE International conference on emerging trends in computing, communication and nanotechnology, 2013.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

11. K.Y.Yeung, D. R. Haynor and W. L. Ruzzo, "Validating clustering for gene expression dataset", Oxford university press, vol. 17, pages 309-318, 2001.

BIOGRAPHY

Pranoti Kamble is a student of M.Tech Computer Science Final Year. She has received B.E. (Computer Science) from P. R. Pote College of engineering, Amravati. Her research interests are bioinformatics, data mining, parallel computing.

Rakhi Wajgi received her Bachelor of Engineering degree from Pune University in 2004. She has completed her M.E. in Computer Science and Engineering from BITS Pilani, Rajasthan in 2008. She is an Asst. Professor in Yeshwantrao Chavan College of Engineering, Nagpur. She has around 6 Yrs of teaching experience. Currently she is pursuing her PhD in Gene Regulation from Nagpur University. Her area of research includes Data Structures, Operating Systems, Parallel Programming and Bioinformatics.