ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Design of an Improved K-Means Algorithm for the Clustering Of Data

J. Jeya Caleb[1]* and M. Kannan[2]
[1]Dept. of ECE, Saveetha Engineering College, Chennai – 602105
[2]Dept. of ECE, MIT Campus, Anna University, Chennai – 600044
***Corresponding author: E-Mail: jeycaleb@gmail.com**, **Phone No: 9790719303**

ABSTRACT

Current technologies in many fields have been utilizing advanced methods for collecting data whose output contain immense data. Such data may be useless unless they are computationally processed to extract meaningful results. K-means clustering plays a vital role in extracting positive details with the help of a collective dataset.

The algorithm of K-means clustering is modified in many ways to improve its performance. The objective of this work is to design a better algorithm for K-means to analyze large datasets. The purpose is to bring in a better method of finding the nearest center for each pixel that is to be classified into different clusters. Then the improved algorithm of K-means is compared with the existing K-means algorithm.

KEY WORDS: Data Mining, K-means algorithm, Clustering, Improved K-means algorithm.

1. INTRODUCTION

Cancer is most widespread disease that involves mutation of cells in an irregular way. Medical imaging helps out to identify this in a better manner. Bioinformatics is defined as the science or techniques of organizing, storing, retrieving, and biological data analysis resulting from genomics and proteomics. These parameters contribute to an uncontrollable growth in the size of biological data.

As a consequence of data growth, scientists are facing computational errors that lead to higher execution times and larger power consumption. Additionally, the availability of the aforementioned molecular data has increased the difficulty of the biological questions that can now be asked by scientists which are calling for the improvement of new mathematical models or algorithms to answer such questions. These demands have led to the wider integration of principles of engineering and computer science into molecular biology to help in converting biological experimental data into biomedical knowledge and hypotheses, BCB has emerged to cater to these demands. A brief review of such methods is presented below.

Mike Estlick (2001), had conducted certain experiments to Map the K- means algorithm to FPGA hardware. They also explored the algorithm level transforms when mapping the algorithm to the hardware. Eventually, they stated that they have achieved 200 times of speed up in the hardware implementation when compared to the software implementation. Hanaa Hussain (2011), have realized clustering using K-means in FPGA. In this work, latest implementation on Graphics Processing Units (GPUs) and General Purpose Processors GPPs were weighed against the implementation of K-means algorithm for clustering in FPGA. It had been found that the improved speed played an essential part in the implementation and energy efficient. It was found that the speed had been boosted up by 15 times in a single core implementation.

Hanaa Hussain (2011), have performed the analysis of microarray data using K-means algorithm. By choosing a parallel architecture the complexity problems of the bio informatics were able to determine. Their experimental results have shown 51.7 times speed increase and 206.8 times more energy efficiency.

Prabhat Kumar (2011), had presented a frame work by implementing multiple kernel optimizations. They have also incorporated a concept of middleware which makes both the GPU and CPU work in unison. Thus an easy and the scalable environment for execution in the multicar CPUs were build up.

Prasath (2013), had used K-means algorithm for clustering to select noteworthy genes of leukemia cancer. It had been examined to cluster genes for K=5, 10 and 15. Here, the accuracy was calibrated to achieve the modification by proceeding with comparison using ground truth values. Nearly 114 genes had helped to detect abnormality out of 7000 genes. 97% accuracy was obtained for 434 genes.

Shailendra Singh Raghuwanshi (2012) compared K-means and Modified K-means Algorithms. A Modified K-means had been proposed. It had been given a provision in using more clusters in numbers and improved execution time when compared to K-means. It had been found that for large data set modified K-means proved to be a better one. The time taken for 600 records was 122 seconds in modified K-means whereas K means took 158 secs. Shanmugavadivu (2012), had provided a recent clustering method by clubbing the clustering of K-means and K mid-range (MkMC). Here images could be partitioned into clusters by using the Modified K mean value.

Huanyi Yang (2012), compared the routine of K-means clustering with Expectation Maximization with Maximization of Posterior Marginals (EM-MPM) and found that EM-MPM performed better than K-means when it comes to clustering highly dense tissues. Siti Noriai Sulaiman (2012), had introduced diagnostic systems that used the HRBF (Hybrid Radial Basis Function) network with AFKM (Adaptive Fuzzy K-Means Clustering). By doing so, they attained enhanced accurateness in classifying the images.

2. MATERIALS AND METHODS

The major disadvantage of all the existing clustering techniques is that the initial centroids must be chosen correctly. Otherwise, the performance will be degraded and the iteration count will be increased.
The proposed clustering technique introduces an enhanced procedure of finding the nearest center for each pixel to be classified into different clusters.

To work out the modified mean, the mean value and the standard deviation of the pixel is computed first. Then half of the standard deviation is added to the actual mean value. This gives the new modified mean value.

The algorithmic description of the devised mechanism is given as:

Step 1: Read the input image.
Step 2: Convert the image into its equivalent matrix I.
Step 3: Compute mean value for the matrix given.
Step 4: Calculate the standard deviation value for the matrix given.
Step 5: Calculate the modified mean value for the matrix as
Mid = Mean (I) + ½ (Standard Deviation (I))
Step 6: The novel mean is calculated based on the observations on centroid in the cluster
Centroid = Mean + (Standard Deviation/2)

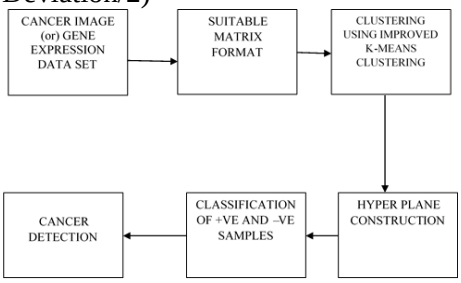

Figure.1. Block Diagram for clustering technology

In the proposed system, the inputs can be either in the type of image or gene expression dataset. Initially, the input image or dataset is converted into a suitable matrix format. Then the matrix is subjected to improved K-means clustering and cluster of data or image parts are obtained. A hyper plane is constructed to categorize the obtained clusters. With the help of the hyper plane the image parts or the dataset containing Cancer contents are labeled as positive samples and the rest are labeled as negative samples. The positive samples account for Cancer detection. This flow is clearly explained in Fig.1. Improved K-means Algorithm is developed using MATLAB R2010a software version (7.10.0.499).

3. RESULTS

The snapshots of the experimental results obtained for K-means clustering and improved K-means clustering are shown below. Fig.2, shows the un clustered input Cancer image. Fig.3, shows the clustered output images obtained using K-means clustering. Fig.4, shows the clustered output images obtained using proposed improved K-means clustering algorithm.

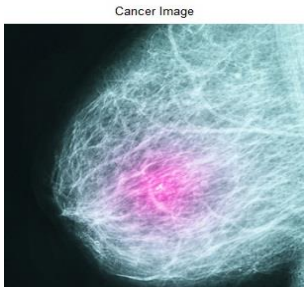

Figure.2. Cancer Image

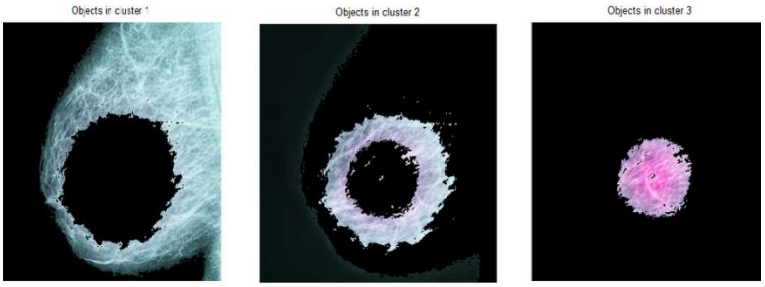

Figure.3. Output of K-means clustering

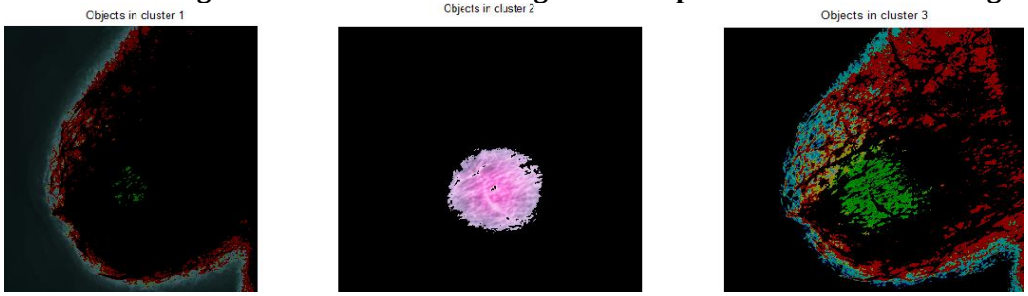

Figure.4. Output of Improved K-means clustering

4. CONCLUSION

A novel clustering technique improved K-means clustering for the segmentation of masses in digital mammograms is presented in this paper. Moreover, it is evident from the obtained results that the proposed technique using improved K-means clustering distinctly carves out the masses/micro calcifications/ macro calcifications present in the test digital mammogram images, than the conventional clustering techniques taken for comparison. The proposed system has an added advantage of identifying the micro calcifications present in the images from the first level cluster itself and hence the computation time and convergence time are drastically reduced. Also, the degree of accuracy of identification of microcalcification using the proposed system is comparatively higher as evidenced by the clustered images. The segmented extraction can be subjected to texture identification in order to analyze the nature of the segmented masses.

REFERENCES

Estlick M, Leeser M, Theiler J, Algorithmic Transformations in the Implementation of K-means Clustering on Reconfigurable Hardware. International Symposium on Field Programmable Gate Arrays, 2001, 103–110.

Hanaa Hussain M, Khaled Benkrid, Ali Ebrahim, Ahmet Erdogan T and Huseyin Seker, Novel Dynamic Partial Reconfiguration Implementation of K-Means Clustering on FPGAs, International Journal of Reconfigurable Computing, 2012, 1-14.

Huanyi Yang, Lauren A. Christopher, Nebojsa Duric, Erik West, Predrag Bakic, Performance Analysis of EM-MPM and K-means Clustering in 3D Ultrasound Image Segmentation, IEEE International Conference on Electro/Information Technology, 2012, 1-4.

Hussain HM, Benkrid K, Seker H and Erdogan AT, FPGA Implementation of K-means algorithm for Bioinformatics Application: An Accelerated Approach to Clustering Microarray Data, NASA/ESA Conference on Adaptive Hardware and Systems, 2011, 248–255.

Kumar P, Ozisikyilmaz B, Liao WK, Memik G. and Choudhary A , High Performance Data Mining using R on Heterogeneous Platforms, IEEE International Parallel and Distributed Processing Symposium, Workshops and PhD Forum, 2011, 1720–1729.

Prasath, Palanisamy Perumal, Thangavel K, and Manavalan R, A Novel Approach to Select Significant Genes of Leukemia Cancer Data Using K-Means Clustering, International Conference on Pattern Recognition, Informatics and Mobile Engineering, 2013, 104-108.

Rafael Gonzales C, Richard E. Woods, Digital Image Processing, Prentice Hall, 2, 2002.

Shailendra Singh Raghuwanshi and Prem Narayan Arya, Comparison of K-means and Modified K-mean algorithms for Large Data-set, International Journal of Computing, Communications and Networking, 1, 2012, 106-110.

Shanmugavadivu P and Santhini Rajeswari R, Identification of microcalcifications in Digital Mammogram using Modified K-Mean Clustering, International Conference on Emerging Trends in Science, Engineering and Technology, 2012, 216-221.

Siti Noraini Sulaiman, Khairul Azman Ahmad, Rohaiza Baharudin, Azizah Ahmad, Nur Athiqah Harron and Aini Hafizah Mohd Saod, Nor Ashidi Mat Isa, Intan Aidha Yusoff, Performance of Hybrid Radial Basis Function Network: Adaptive Fuzzy K-Means versus Moving K- Means Clustering As Centre Positioning Algorithms on Cervical Cell Pre-cancerous Stage Classification, IEEE International Conference on Control System, Computing and Engineering, 2012, 607-611.