Abstract

Objectives: The radical growth of brain MRI data demands faster and accurate processing. To meet these demands, it is necessary to develop a design in cloud platform using distributed platforms. Methods/Analysis: In this paper, we introduce an architecture developed for the cloud using Apache Hadoop to segment the brain MRI images. The scanned MRI images are uploaded through either through web interface or mobile app to the system in the public cloud. The Parallel Genetic Algorithm (PGA) in the cloud system enabled with Hadoop or Spark is used to segment the given MRI images. Findings: The processing time taken for different size of data varying from 2GB to 10GB in a different number of clusters varying from one to five are denoted. This process has been implemented in both Apache Hadoop and Apache Spark. The time ranges from 12 to 24 secs approximately in Hadoop whereas the processing time has come down from 4 to 7 secs in Spark. First of all, the results prove that the network based applications for Medical Image Processing are outperformed by the cloud platform applications. Novelty/Improvement: Distributed Platforms have been used in Cloud environment for Brain MRI segmentation using Parallel Genetic Algorithm.

Keywords: Apache Hadoop, Brain MRI Segmentation, Cloud Computing, Medical Image Processing, Parallel Genetic Algorithm, Spark

1. Introduction

Medical Image Processing is bottomless as well as the broader level of a research area that keeps the researcher especially those in the same domain always to bring out new findings to enhance the efficiency of medical image diagnostics. Even though there are new researches coming out every day on MIP, there is no end to say one has achieved 100% accuracy of results. There are more hidden parameters yet to be brought out to reach the desired accuracy. When a parameter causing a problem for attaining efficiency is addressed, other parameters emerge to be addressed. At the same time, the working pattern of algorithms, the direct and indirect variables, a combination of algorithms, the input produced by various machines, the number of inputs and the technology used are having a great impact in the preciseness of the output.

There are various factors hindering the quality process of segmentation of brain MR image. One of the critical problems in segmentation is time consumption. Accordingly, the research combs the basics of various brain segmentation algorithms employed with different methodologies by researchers. The different approaches like thresholding, region growing, clustering, neural network and genetic algorithms for segmentation are considered to find the optimum solution.

For any MIP algorithm to succeed in achieving efficiency, three critical mechanisms need to be fulfilled.

• focus and problem identification at the grassroots level.
• ensuring coverage, quality, relevance and reliability of data so that the information existing through MIP is effectively used for attaining the optimum result.

*Author for correspondence
• investigation, timely dissemination and feedback on
  the information available through the system which
  can be built by using the emerging technologies like
  Cloud Computing, Hadoop and Spark platforms.

Keeping these three key factors in mind, one could
identify that there is a continuous need for improving
the efficiency of the MRI processing to achieve maximum
accuracy to reach the right diagnosis. There are various
algorithms and methodologies used for processing MR
images. Each algorithm shows improvement in efficiency
in different conditions.

In the current days, there are various tools and
methods to analyze the medical images and to extract the
required and necessary information. Even though it is
evident that the research over the past century has taken
us to the advancement in MIP, one can accept that there is
a need to enhance the MIP with better tools and involve-
ment of cutting edge software like Cloud Computing,
Hadoop, and Spark platforms. Following are the concerns
even to obtain the imaging which is the first step of MIP.

Since the brain MR image data is huge in size and
the number of images are increasing every day, process-
ting time and memory utilization is a challenge before the
algorithm developers. As the earlier studies show that
Genetic Algorithms are strong enough and efficient with
large search space, it would be highly beneficial if GAs are
made as Parallel GAs.

The literature has shown that the accuracy in the
segmentation of brain MRI segmentation can be increased
by addressing the following three main problems found
in using GAs, especially in brain MR image processing.

1. The size and the number of demes.
2. Topology chose to interconnect the demes.
3. Migration rate to decide the number of demes can
   migrate among the populations at any given time.

1.1 Parallel Genetic Algorithm for Image
Segmentation

Following a study of different works from various authors
exhibits that how Parallel Genetic Algorithms (PGAs) are
efficient in search techniques and to overcome the issues
arose in GAs.

There are two main types of PGAs namely coarse
grain and fine grain PGAs. Coarse grain PGAs are most
popular ones in the category.

Local optimizer was used by the authors in the
algorithm to make the PGA more efficient. But it is not
obvious that the results attained are due to the distributed
population or the local optimizer. Still it is realized that
the program they designed was efficient and influential.

It is an interesting characteristic of PGA that they
enable parallelism for sequential GAs along with exhibit a
sort of advanced competence and effectiveness in reaching
an optimum solution because of the structured population
and parallel execution.

PGAs are well required by the applications in the
fields of artificial intelligence, numeric and combinato-
rial optimization, business, engineering, etc. because of
their robust nature in dealing with highly complex prob-
lems to attain optimum results. Following are some of the
characteristics of a PGA gives advantages:

• Reliability in using depictions of the problem param-
eters instead of parameters themselves
• Robustness
• Flexibility in customization for a new problem
• Capable of multiple efficient solutions for a given
  problem.
• Faster.
• Less prone to finding only sub-optimal solutions.
• Accomplished by working with other search techniques
  in parallel.

PGAs have the advantage of making the procedures on
the demes independent from each other because of par-
allelism. At the Same time, PGAs are more efficient as
a result of the division of the whole population as sub-
populations to localize competitive selection among the
subpopulations. The literature proves that PGAs have the
additional features such as higher efficiency capable of
maintenance in the wider range and making memory
and CPU freer.

PGAs are not only sped up the execution but also give
excellent numerical performance even when these algo-
rithms are executed on a sole processor. Optimization
is achieved by the use of structured population which are
either formed as a set of islands or a diffusion grid because
of the numerical performance of PGAs. Therefore, it has
been proven that even without parallel machines, PGAs
can contribute better results as a result of structured-pop-
ulation models. It does not mean that parallel machines
have no influence in the result optimization. In the
meantime, it is experienced in the studies that hardware
parallelization is a supplement to accelerate the process,
and it can be achieved through several ways on a given
structured-population GA.
In13 implemented a parallel genetic algorithm for surface model fitting in 3D space. Still there is a need to address the issue of execution time in different parameters like a number of nodes and the topology used.

In14 shown in the experiment that parallel genetic algorithm can be a better optimizer than the classical GA. PGAs are made to coincide with the natural process of evolution by incorporating migration operation. In this scenario, PGAs produce numerous isolated subpopulations in parallel and the subpopulations from time to time exchange their best individuals based on the predetermined migration technique.

As it was mentioned, migration of individuals among the subpopulations plays a major role in the optimization of PGA. In this context, in15 proposed a reformed migration technique. In the given architecture, master-salve processors are used where GA on each subpopulation is run on the slave processor and the best partial demes will be sent to master process every so often. Now it is the job of the master process to choose the best-fit individual received from different slave process at a point of time and broadcast to the slaves. The authors have used six nodes on the network and could experience that there is a significant increase in the speed with the smaller number of nodes. The algorithm did not forecast the efficiency when the dataset increases extremely.

In16,17 proposed an active model using a parallel genetic algorithm to segment the lateral ventricles from brain MR images. They have defined an objective function to improve the performance of PGA. They have randomly chosen the number of subpopulations and the subpopulation from which the best fit individuals can migrate to others. The disadvantages of the island model and the stepping stone model are overcome in their proposed model by reducing the communication and make the parallel genetic algorithm behavior more close to nature18.

In19 a genetic algorithm has been used for feature selection after the segmentation process. They have used PCA also for feature reduction. Since the authors have worked more on the classification of normal and abnormal brains, the power of genetic algorithm could not be realized much in terms of efficiency.

In20 a Distributed Parallel Genetic Algorithm oriented adaptive migration strategy has been proposed to improve the convergent speed without affecting the efficiency of the result by considering the benefits of parallel computer architecture and using PGA. There is a need for considering the parameter like communication topology which has greater influence in execution time.

In21,22 have used optimum population size, mutation rate and selection strategy which is parallelized with MapReduce architecture for finding the optimal conformation of a protein using the two-dimensional square HP model. They could find the result that convergence of GA is faster than the traditional to the optimum state.

2. Implementation of Image Processing using Parallel Genetic Algorithm

There are online databases from where the required real-time data for this research has been received. The brain MR image dataset was got from the Human Connectome Project and VIRTUALSKELETON database and they are cited in the bibliography.

Image features are commonly categorized into three core tissue types in brain MRI. They are,
1. White matter (WM),
2. Gray matter (GM), and
3. Cerebrospinal fluid (CSF).

To highlight the different aspects of normal and abnormal tissues in the brain, pulse sequences are used. The variations in the parameters such as time between the repetition of procedures (TR) and time consumed of echo (TE) are used to stress changes in anatomical images. For example, to stress on the contrast between gray and white matter, T1-Weighted images is scanned with short TR and short TE where in the case to stress on the contrast between brain tissue and cerebrospinal fluid, T2-Weighted is scanned with long TR and long TE. The sequence will vary based on the information provide as well as the time consumption to acquire the image.

The images received from online databases are processed in R for the conversion as a gray scale image of size 256*256 and then to a grainy image in which Signal-to-Noise Ratio (SNR) is less. The converted 2D images are retrieved from HDFS and displayed in a two dimensional matrices where the elements of matrices are the pixels of the images. The size is important to decide the processing time and it takes the vital role for effective processing. In gray scale image, the values of pixels start from 0 to 255 where 0 represents total black color and 255 represents pure white color as Mustaqeem, Javed and Fatima (2016) mentioned in the image Journal. The values in between the range display the concentrations of gray color.
In the research, the MR images are considered as 256 X 256 gray scale values and the images are stored in HDFS. There is a provision in Hadoop called Sequence Files to complete this procedure of storing files. Key/Value pairs in binary form are stored in a SequenceFile which is a flat file. There are classes in SequenceFile namely Writer, Reader and Sorter for writing, reading and sorting correspondingly. So, the file can be converted as a SequenceFile and stored it into the HDFS. Since the size of a brain MR image is 80 to 100 megabytes in general, there is no problem in storing the files in HDFS in either of the ways given.

Parallel Genetic Algorithm performs well with respect to efficiency especially the search space is complex and large where accuracy is more expected. PGA works more natural by integrating the migration operation of best fit individuals from different subpopulations generated in the various worker nodes.

The following is the basic flow of the parallel genetic algorithm.

1. Definition of genetic operators.
2. Random generation of a population of candidate solutions.
3. Partition the populations into several subpopulations.
   a) Apply the average migration strategy for individuals to flow amongst the subpopulations.
4. Execute the following steps (a) and (b) for each subpopulation.
   a) Depend on the selected genetic operators selection, crossover, and mutation, execute self-evolution.
   b) Based on the average migration technique, the best individuals are sent to the subpopulations, where the best ones replace the worst ones of the subpopulation.
5. Check whether the stopping conditions are satisfied. If fulfilled, break the iteration, else go to Step 3.

Performance improvement can be achieved by using the fitness valuation of parallel GA with respect to the data size and the number of clusters used. The efficiency of parallel GA can be enhanced with the fitness evaluation, migration rates, communication topology, or deme size. The time needed to reach the solution and the accuracy of the solution depends on the population size which is one of the important parameters in the proposed PGA. The research takes migration scheme as a vital parameter that decides which individual can migrate from one deme to another and which individuals are replaced Figure.1 to 3.
To enable the classification, the SVM algorithm uses the feature subset predefined in the research. There are three classes of images namely normal, benign and malignant. SVM applies the classification procedure using training set and test set data to evaluate the accuracy of segmentation by the proposed algorithm. In every processed image, individual subject is denoted by a vector.

3. Implementation of brain MRI segmentation in Hadoop and Spark

The following architecture was implemented with single cluster Hadoop\textsuperscript{23, 24}, Multi Cluster Hadoop, and the recent in-built MapReduce Spark. The performance of the PGA was tested with approximately 2 GB, 4 GB, 6 GB, 8 GB and 10 GB brain MR images stored in HDFS.

The algorithm was tested with 8 GB RAM, 1 TB HDD, Intel Core i7 CPU with 2.50 GHz speed, Ubuntu 14.0 and Hadoop2.4.0 configuration for the single cluster Hadoop Table 1. The same configuration was maintained for the multi cluster Hadoop with five DataNodes and one NameNode Figure 4 and 5.

When the same setup was used for Spark1.4.0 to execute the PGA, there is a drastic change in reduction of execution time to less than half of Hadoop taken for execution Figure 6. Since Spark uses the memory for keeping intermediate reductions as discussed by this reduction of execution time could be realized. Whereas Hadoop uses the disk for writing intermediate results that increase latency time for to and fro communications with a disk which origins more execution time Table 2.

Table 1. Performance Evaluation in Single / Multi Cluster Hadoop

| Number of data nodes in a cluster | Process Throughput in GB /Secs |
|-----------------------------------|-------------------------------|
|                                   | 2    | 4    | 6    | 8    | 10   |
| 1                                 | 12   | 15   | 17.2 | 20   | 23.4 |
| 2                                 | 13   | 16   | 16.4 | 17   | 18   |
| 3                                 | 13   | 15   | 15.1 | 15   | 15.5 |
| 4                                 | 14   | 15   | 15   | 14.8 | 14.6 |
| 5                                 | 14   | 14.6 | 14.4 | 14.1 | 14   |

Figure 4. Proposed architecture for processing brain MRIs with multi cluster Apache Hadoop / Spark in cloud.

Table 2. Performance Evaluation in Multi Clusters Spark

| Number of data nodes in a cluster | Process Throughput in GB /Secs |
|-----------------------------------|-------------------------------|
|                                   | 2    | 4    | 6    | 8    | 10   |
| 1                                 | 4    | 5    | 5    | 5    | 6    |
| 2                                 | 5    | 6    | 6    | 7    | 7    |
| 3                                 | 5    | 6    | 6    | 5    | 5    |
| 4                                 | 6    | 6    | 5    | 5    | 4    |
| 5                                 | 6    | 7    | 5    | 4    | 3    |

Figure 5. Process throughput in GB/sec in Multi Cluster Hadoop.

Figure 6. Process throughput in GB/sec in Spark1.4.0.
4. Analysis and Discussion

The objective of was to enable the interaction of several parallel subcomponents of an evolving population in the brain MR image in the cloud environment using Hadoop/Spark.

4.1 Increase in Performance in Cloud Setup

One of the main objectives of the proposed work is to provide a cloud environment where the processing of brain MRI can be carried out anywhere from the globe. To meet the objective by achieving the best solution, we have run the proposed PGA in Hadoop as well as in Spark with single and multiples clusters in AWS. Instead of carrying out to evaluate the performance of proposed PGA for a single image, keeping the Big Data analytics in mind, the process was carried out for GBs of brain MRI data.

Performance evaluation was done for 2, 4, 6, 8 and 10 GB brain MRI data in a number of clusters starting from 1 to 5. Compared to Single cluster Hadoop, multi cluster Hadoop could perform faster with the size of data increases. In a single cluster Hadoop, 12 secs were taken to process 2GB of data whereas only 23.4 secs were taken to process 10 GB data. We are able to see that the processing time of PGA comes down even with the increase in data size in a single cluster setup. In Hadoop multi cluster setup, initially, there was either decrease or no change in the performance of proposed PGA with 2GB data. Since communication overhead between nodes increases due to the hybrid topology, there is a probability for either increase or no change in computation time.

But we are able to meet the objective with the increase of data size by experiencing the reduction in execution time in multi cluster Hadoop. As it is mentioned earlier, Hadoop works with MapReduce technique and meanwhile PGA also executes with parallel processing and reduction with migration topology, it is double benefited with respect to the huge amount of data in terms of execution time.

It is similar to the case of Apache Spark in which the intermediate storage takes place in the memory itself whereas it happens with a disc in Apache Hadoop. We could understand the power of Spark in terms of drastic reduction in execution time. In Spark, time taken for processing 10 GB of data was 6 secs which are only 25% of the computing time taken by Hadoop to process the same volume of data. Even though the variation in processing time is less with respect to multiple clusters in Spark, as the data size increases, computing time also decreases radically. When we compare the processing time 14 secs and 3 secs in five clusters setup for 10 GB of data in Hadoop and Spark respectively, Spark consumes approximately 20% of the computing time taken by Hadoop.

So we are able to realize the power of parallelism with average migration technique integrated into Genetic Algorithm processed either in Hadoop or Spark which can be extended to support huge brain MRI data set termed as “Big Data” available globally in the cloud to be accessed from anywhere.

5. Conclusions

We have articulated the need for medical image processing for the healthcare system. We have incorporated the technologies like Hadoop and Spark which can support medical image processing in the cloud. It is the motivation of the study as well as the research gap found which needs to be addressed in the current fast growing technologies.

The research has contributed in two major areas. Firstly parallel genetic algorithm for brain MRI segmentation has been analyzed with the various brain MRI processing algorithms and their advantages and disadvantages were discussed. Secondly, to support the novel approach, the medical image processing in the cloud with Hadoop and Spark has been analyzed to fill the research gap.

Since the implementation of the proposed methodology is in the cloud, transferring of data needs efficient lossless compression techniques for to and fro communication with the cloud. The novel approach of introducing average migration technique in the parallel genetic algorithm for segmentation of brain tumour is carried out and discussed in detail. Medical Image Processing does not stop only with the processing algorithms. PACS and DICOM are enabling us to go beyond processing and make the medical image available for various users within the legal limitations. Since the performance evaluation proves that a huge amount of brain image data can be processed in few seconds, the future work can be to derive effective mechanisms to store the data on the remote machine in the cloud to save the problems in cloud storage since the medical image data is significantly increasing in volumes every day.

6. Acknowledgments

The author thank Dr. Pethuru Raj for providing various resources and for Dr. Pethuru Raj in developing and
implementing Cloud applications, Big Data, and Hadoop applications

7. References

1. Application on Reinforcement Learning for Diagnosis Based on Medical Image. Available from: http://www.intechopen.com/books/reinforcement_learning/application_on_reinforcement_learning_for_diagnosis_based_on_medical_image. Date Accessed.01/01/2008
2. Grefenstette JJ. GENESIS: A System for Using Genetic Search Procedures. Proceedings of the 1984 Conference on Intelligent Systems and Machines.1984. p. 161–65.
3. Muhlenbein H, Schomisch M, Born J. The Parallel Genetic Algorithm as Function Optimizer. Genetic Algorithms. 1991; 17(7):619–32.
4. Kaghed HN, Al–Shamery SE, Al-Khuzaie FE. Multiple Sequence Alignment based on Developed Genetic Algorithm. Indian Journal of Science and Technology. 2016 Jan; 9(2):1–7.
5. Peteye CC, Leuze MR, Grefenstette J. A Parallel Genetic Algorithm. Proceedings of the ICGA Symp.1987; 155–61.
6. Alba E, Troya JM. A survey of parallel distributed genetic algorithms. John Wiley & Sons, Inc: USA. 1999; 4(31):31–52.
7. Tanese R. Distributed Genetic Algorithms. Proceedings of the 3rd International Conference on Genetic Algorithms. 1989. p. 434–39.
8. Shenbagarajan A, Ramalingam V, Balasubramanian C, Palanivel S. Tumor Diagnosis in MRI Brain Image using ACM Segmentation and ANN-LM Classification Techniques. Indian Journal of Science and Technology. 2016 Jan; 9(1):1–12.
9. Lin SC, Punch WF, Goodman ED. Coarse-Grain Parallel Genetic Algorithms: Categorization and New Approach. Proceedings of 46thIEEE Symposium on Parallel and Distributed Processing, Dallas, TX. 1994; 28–37.
10. Spiessens R, Manderick B. A Massively Parallel Genetic Algorithm. Proceedings of the 4th International Conference on Genetic Algorithms. 1991. p. 279–86.
11. Gordon VS, Darall Whitley L. Serial and Parallel Genetic Algorithms as Function Optimizers. Proceedings of the 5th International Conference on Genetic Algorithms. 1993. p. 177–83.
12. Lozano M. Application of Fuzzy Logic Based Techniques for Improving the Behavior of GAs with Floating Point Encoding, Univ Granada. 1996.
13. Manderick B, Spiessens P. Fine-Grained Parallel Genetic Algorithms. Proceedings of the 3rd International Conference on Genetic Algorithms. 1989. p. 428–33.
14. Mejia-Olvera M, Cantu-Paz E. Dgenesis-Software for the Execution of Distributed Genetic Algorithms. Proceedings of the Latinoamericana de Informatica. 1994. p. 935–46.
15. Marin FJ, Trelles-Salazar O, Sandoval F. Genetic algorithms on LAN message passing architectures using PVM: Application to the routing problem. Proceedings of the International Conference on Evolutionary Computation. 3rd International Conference on Parallel Solving from Nature Parallel Problem Solving from Nature. 1994. p. 534–43.
16. Fan Y, Jinag T, Evans DJ. Volumetric segmentation of brain images using parallel genetic algorithms. IEEE Transactions on Medical Imaging. 2002; 21(8):904–09.
17. Baraiya N, Modi H. Comparative Study of Different Methods for Brain Tumor Extraction from MRI Images using Image Processing. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–5.
18. Chipperfield A, Fleming P. Parallel Genetic Algorithms, Parallel and Distributed Computing Handbook, MacGraw-Hill: USA. 1996; 1118–43.
19. Jafari M, Shafaghi R. A Hybrid Approach for Automatic Tumor Detection of Brain MRI Using Support Vector Machine and Genetic Algorithm. Global Journal of Science, Engineering and Technology. 2012; 1(3):1–8.
20. Bhatia M, Bansal A, Yadav D, Gupta P. Proposed Algorithm to Blotch Grey Matter from Tumored and Non Tumored Brain MRI Images. Indian Journal of Science and Technology. 2015 Aug; 8(17):1–10.
21. Li W, Huang Y. A Distributed Parallel Genetic Algorithm oriented adaptive migration strategy. 2012 Eighth International Conference on Natural Computation (ICNC). 2012. p. 592–95.
22. Sasirekha N, Kashwan KR. Improved Segmentation of MRI Brain Images by Denoising and Contrast Enhancement. Indian Journal of Science and Technology. 2015 Sep; 8(22):1–7.
23. Narayana AGH, Krishnakumar U, Judy MV. An Enhanced Map Reduce Framework for Solving Protein Folding Problem using a Parallel Genetic Algorithm. ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India. 2014; 1:241–50.
24. Eizadpanah E, Koroupi F. Timing of Resources in Cloud Computing by using Multi-Purpose Particles Congestion Algorithm. Indian Journal of Science and Technology. 2015 Apr; 8(S8):1–10.
25. Peter DA. Enhancing the Efficiency of Parallel Genetic Algorithms for Medical Image Processing with Hadoop. International Journal of Computer Applications. 2014; 108(14):11–6.