IJCTA, 10(8), 2017, pp. 779-792
© International Science Press
ISSN: 0974-5572

A Novel Approach to Image Clustering using Elitist Genetic Algorithm

M.S. Chelva* and S.V. Halse**

ABSTRACT

Genetic Algorithm (GA) is a stochastic randomized blind search and optimization technique based on evolutionary computing that has already been proved to be robust and effective from its outcome in solving problems from variety of application domains. Clustering is a vital technique to extract meaningful and hidden information from the datasets. Clustering techniques have a broad field of application including bioinformatics, image processing and data mining. In order to the find the close association between the densities of data points, in the given dataset of pixels of an image, clustering provides an easy analysis and proper validation. In this paper, we propose an evolutionary computing based approach for unsupervised image clustering using elitist GA (EGA) – a efficient variant of GA that segments an image into its constituent parts automatically. The aim of this algorithm is to produce precise segmentation of images using intensity information along with their neighbourhood relationships. Experimental results from simulation study reveal that the algorithm generates good quality segmented image.

Keywords: Image Clustering , Evolutionary Computing (EC), Genetic Algorithm (GA), Elitism, Image Segmentation.

1. INTRODUCTION

Clustering is practicable in various explorative pattern-analysis, grouping, decision-making, and machine learning circumstances, including data mining, document retrieval, image segmentation, and pattern classification [1]. Clustering a set of images into meaningful categories has many applications in the organization of image databases and the design of their human interface [13].

The cluster analysis has been widely used in many fields such as statics, marketing, engineering, medical and other social sciences, for clustering large data sets into natural groups. Likewise, there had been a number of clustering algorithms proposed for performing the clustering task [15].

Clustering is the search for distinct groups in the feature space. It is anticipated that these groups have different configurations and that can be clearly differentiated. The clustering task separates the data into number of partitions, which are volumes in the n-dimensional feature space. Therefore, clustering is basically a particular kind of NP-hard problem [4, 6, 26, 27, 28]. These partitions define a hard limit between the different groups and depend on the functions used to model the data distribution.

One of the main objectives of the clustering algorithms is to find the ‘natural’ groups in the dataset along with partitioning the data into those natural groups. But none of these clustering algorithms are efficient enough to discover ‘natural’ groups from all the input patterns, especially when the number of clusters included in the data set tends to be large. These algorithms also suffer from the problem of local convergence due to large clustering search space

In recent years, the GA has received a good deal of response as a robust stochastic search algorithm for numerous optimization problems. Extensive literature survey has revealed that researchers working in the area of clustering large data sets have shown interest in GA as an upcoming and promising tool based on

* Research Scholar, SVMV, SRTMU, Nanded, Maharastra, India

** Karnataka State Women’s University, Vijaypur, Karnataka, India

evolutionary computing (EC) technique and many promising solutions based on GA approach have been proposed and results reported for numerous image processing applications.

This paper will discuss the use of Genetic Algorithms (GAs) for the task of clustering image data. In particular, the application of GAs for clustering on very large data sets, such as image data sets, has already been proved to be successful. In this paper, presentation of GA for clustering data for image processing applications will be addressed. We propose an efficient genetic algorithm for clustering on very large image data sets.

The paper is organized as follows. In section II, we review the clustering problem and genetic algorithms. Section III presents a discussion on the nature of clustering problem. In Section IV, we detail our genetic algorithm for clustering on image data sets. In section V, we have outlined our proposed elitist GA (EGA) based image clustering algorithm. Experimental results on real image data sets are given in Section VI. Section VII concludes the paper.

2. SURVEY OF LITERATURE

There have been many works done in the area of image clustering by using varieties of methods targeting on different applications of image clustering techniques. K-means algorithm is the one of the most popular clustering algorithm and there are many methods implemented so far with different method to initialize the centroid [5]. Many researchers are also exploring the field to come up with new and more efficient methods of image clustering than the existing and currently practiced methods that shows better clustering performance. In this section, we discuss few existing and recent works going on in this field.

Gulhane *et al.* [15] have presented a comprehensive survey of clustering algorithm and claimed that K-means is one among the most popular and commonly used clustering algorithm, being first proposed over 50 years ago. Dhanachandra *et al.* [5] proposed a K-means based unsupervised clustering algorithm that involves subtractive data clustering method where the algorithm produces the centroid depending on the latent value of the data points. Goldberger *et al.* [7, 8] have an unsupervised hierarchical image-set clustering method based on the information–theoretic principle, i.e., the information bottleneck principle where the images are clustered such that the mutual information between the clusters and the image content is maximally preserved. The focus of the proposed approach was of particular emphasis placed on the application of the clustering technique for the efficient image search and retrieval. Hung and Yang [9] have proposed a modified fuzzy C-means (FCM) clustering algorithm with a claim of significantly reduced computation time and four times faster than the traditional FCM where the performance of the algorithm depends upon the choice of the initial cluster center and/or initial membership value. Chen *et al.* [11] proposed a novel image segmentation approach based on density peaks (DP) clustering algorithm with few advantages over the existing methods, resulting in directly giving the cluster number of the image based on the decision graph; correct identification of the cluster centers; performing hierarchical segmentation as per the applications requirement. The author claims the validity of the proposed segmentation algorithm through experimental findings. Seldin *et al.* [13] have presented an unsupervised content based images classification (clustering) algorithm built around the sequential Information Bottleneck algorithm. Tian *et al.* [16] proposes an automatic K-means clustering algorithm based on histogram analysis for processing CT scan images that contain several materials with similar gray-levels, which manages to overcome the intrinsic deficiency in K-means clustering where the choice of initial clustering centroids may highly influence the performance of the algorithm. Krinidis *et al.* [17] presents a novel unsupervised image clustering methodology based on the image histogram, which is processed by empirical mode decomposition (EMD) and claimed to exhibit good and robustness in the clustering of several real and synthetic images.

Mor and Gupta [4] have presented a comprehensive survey of various clustering techniques for image processing application employing GA, and in general, with reference to the optimality of the result obtained, claimed superiority of the GA based approach over the traditional approaches. Raposo *et al.* [2] have proposed a GA based approach for solving automatic (unsupervised) clustering problem with image data and through experimental validation revealed that the proposed approach outperformed the traditional K-means and FCM clustering approach. Ding and Gasvoda *et al.* [6] have proposed an efficient algorithm for clustering very image large data sets using GA. The simulation result of the proposed approach claims to have outperformed the k-means algorithm in terms of running time as well as the quality of the clustering. Venkatesh *et al.* [18] have proposed a GA based GA based approach for efficiently clustering image data. Their experimental results showed that GA gives better performance than the simulated annealing technique. Halder *et al.* [12] have presented a GA based clustering approach integrated in to the FCM clustering algorithm for unsupervised gray-scale image segmentation and their experimental result claims to have obtained good quality segmented image. Khashandarag *et al.* [3] have presented a novel approach for segmenting medical images by using variable string length genetic algorithm (VGA) integrated with K-means clustering algorithm and claimed to have obtained better accuracy in selecting the optimal cluster centres compared with their simple K-means clustering algorithm. Zanaty and Ghiduk [10] proposed a novel algorithm incorporating hybridization of GA and seed region growing to produce accurate segmentation of different MRI image. Experimental results show that the proposed method produces more correct and stable outcomes compared with other segmentation techniques such as fuzzy c-means (FCM) and hybrid GA and fuzzy clustering (GAFCM) methods. Kaur and Jindal [14] presented a segmentation technique hybridizing FCM and GA to produce quality results when applied to medical images for accurate tumour detection.

In this paper, we have presented a simple approach involving K-means based unsupervised clustering algorithm hybridized with elitist GA (EGA) to exhibiting fairly image clustering performance. The proposed algorithm is targeted for implementation in devices with low computation power and low memory foot print like microcontrollers and microprocessors.

3. IMAGE CLUSTERING PROBLEM

Image segmentation has been a major research topic among many image processing researchers. The reasons are obvious and applications are endless: most computer vision and image analysis problem involves a segmentation stage in order to detect objects or divide the image into regions which can be considered homogeneous with respect to a given criterion, such as colour, motion, texture, etc.

In broad sense, clustering algorithms can be grouped into two main classes of algorithms, namely supervised or manual clustering algorithm and unsupervised or automatic clustering algorithm.

Manual or supervised clustering method involves interactive role of the user for identifying pixels belonging to the same intensity range that are pointed out physically and then segmented. The major obstacle encountered in supervised segmentation is that it demands more time due to the user intervention into the process to influence the segmentation and becomes the worst when the image is of bigger size. Larger the image in size greater is the time demanded. With supervised clustering, the learning algorithm has an external teacher that indicates the target class to which a data vector should belong.

Automatic or unsupervised clustering method which is more complex and algorithms need some prior information such as the probability of the objects having a distinctive distribution to carry out the segmentation. Unsupervised segmentation automatically groups elements of an image agreeing to some criteria. Unsupervised methods have earned reputation as they agree to a great extend with the human perception. For unsupervised clustering, a teacher does not exist, and data vectors are grouped based on some measure of similarity, such as distance from one another.

The foremost disadvantage of supervised clustering is that it requires human intervention. In order to extract the cluster representation, the various approaches involved require a priori knowledge concerning the database content. This approach is, therefore, not appropriate for large unlabelled databases. A different set of techniques based on unsupervised clustering is adopted, where the clustering process is fully automated [8].

In works that use supervised clustering, the expert incorporates a priori knowledge, such as the number of classes present in the database and the illustrative signs for the different classes in the database [8].

In this work, we propose a novel unsupervised image clustering algorithm utilizing a hybrid of K-means clustering with elitist genetic algorithm [8].

Unsupervised image clustering is the process of search for the distinct groups in the feature space. It is expected that these groups have different arrangements and that can be clearly distinguished. The clustering task separates the data into number of partitions, which are bulks in the n-dimensional characteristic space. These partitions describe a hard limit between the different subsets or groups and depend on the functions used to model the data distribution.

Clustering is often based on some similarity or distance measure. The concept of similarity is invariably problem-dependent. The similarity (or dissimilarity) between the objects is typically computed based on the distance between each pair of the objects. These measurements comprise of the *Euclidean, Manhattan* and *Minkowski* distances [4]. The most popular distance measure is the Euclidean distance, which is defined as Eqn.(1) below:

$$D(i, j) = \sqrt{(\mathbf{X}_{i1} - \mathbf{X}_{j1}) + (\mathbf{X}_{i2} - \mathbf{X}_{j2}) + \cdots + (\mathbf{X}_{in} - \mathbf{X}_{jn})} \tag{1}$$

Where $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $\mathbf{X}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are two *n*-dimensional data objects [4].

Mostly, the objects are bunched on the base of Euclidean distance. The objects are bunched in such a way that each object belongs to the cluster whose centroid to object Euclidean distance is minimum [4].

The diversity of applications for clustering has lead to many problem definitions. Prime objective of all clustering algorithms is to split a set of data points into subgroup so that the entities within a subgroup are similar to each other and entities that are in different subsets have diverse qualities. For our research, we defined the clustering problem as the task of dividing an input dataset into a required number of subgroups so that the Euclidean distance between each data point and its corresponding cluster centre is minimized. This is a very common method of defining the clustering problem.

Clustering can be formally considered as a particular kind of NP-hard grouping as far as optimization perspective is concerned. Evolutionary algorithms (EAs), particularly genetic algorithms (GAs) are believed to be efficient on NP-hard problems. In this regard, GAs are tools capable of providing near-optimal solutions to such problems in a reasonable time. Under this postulation, a large number of GA based approach for solving clustering problems have been proposed in the literature as already discussed earlier in the paper. These algorithms are based on the approach of optimizing some objective function (i.e., also known as the fitness function) that escorts the evolutionary search process.

4. GENETIC ALGORITHM (GA)

The genetic algorithm based on the Darwinian principle of survival of the fittest developed by John Holland in 1975 [19] emulates nature's evolution process to solve various optimization problems. Genetic algorithm is a general purpose, random search and stochastic optimization techniques with domain autonomy based on the principles of genetic theory and natural selection.

GA starts with an initial set of random or arbitrary solutions creating a *population* of *individuals* of size N_p. Each individual in the population, called *chromosome*, represents a candidate solution to the problem in the solution space. These individuals evolve from population to population through consecutive iterations, called *generations*, keeping N_p fixed throughout the iterations. Each chromosome has a *fitness value* associated with it. During each successive generation, the chromosomes are *evaluated* with the help of the *fitness function* to assess their *fitness* to survive in the next generation. A successor population is formed by selecting the befitting individuals and enforcing the genetic operations to transform them: *mutation* of a single individual, or *crossover* among two individuals (where parents get children). Genetic operators are applied to the designated individuals to produce the new ones with characteristics inherited from their parents and the associated fitness function evaluates the extent to which these individual achieves the goal of optimization.

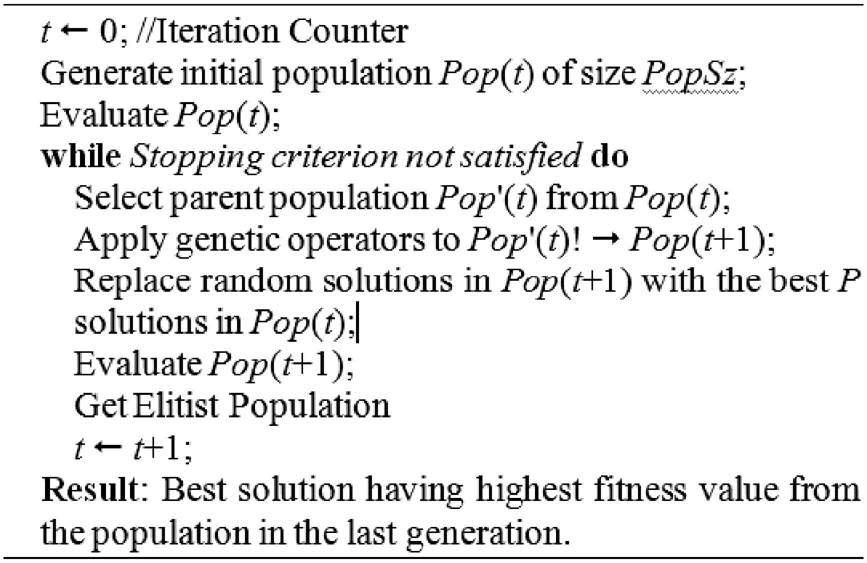
Normally, GA operates in two stages: in the first stage, the *initialization phase* through which the GA is initialised with N_p random individuals to form the starting population followed by the evaluation of fitness of each, and next, the *iteration phase* which starts the selection process during which the befitting individuals from the current population are selected stochastically [19] into the *intermediate population* from where the *reproduction* of the next generation through genetic alteration of individuals with the help of *genetic operators* is accomplished, again followed by the evaluation of fitness each individual. Steps in the iteration stage are repeated until some end point criteria are satisfied, resulting in the outcome of an optimal solution.

Modes of reproduction through genetic operators are primarily crossover and mutation [19]. In this paper, a 1-point crossover approach is adopted to produce two offspring from two parent chromosomes by exchanging genes [19] before and after the *locus* picked at random representing the *crossover point* or l_{cp}, $1\ l_{cp} < L_c$, in the parent chromosomes of length L_c. Mutation produces an offspring from a single parent chromosome by changing the allele of some randomly picked gene [19, 20]. Mutation adds diversity to the population [19]. A probability factor, called *crossover probability* (P_c), decides occurrence of crossover operation while another probability factor, called *mutation probability* (P_m), activates whether mutation to occur.

Function of the selection mechanism is to highlight healthier individuals in the population to take part in the reproduction process to produce children of even higher fitness. However, such individuals are sure to get lost if they are not nominated to participate in the reproduction process or if they are demolished by the crossover or mutation [21] mechanism. In order to avoid this loss, we have implemented *elitism* in our work along with pair-wise *tournament selection* without replacement, where the *tournament size* is 2 [20]. Elitism, initially introduced by Kenneth De Jong (1975), is a supplement to many selection procedures that forces the GA to retain some number of the best individuals from each generation [22]. Elitist selection approach ensures that the best individuals survive in the next generation with probability one [23] ensuring that the fitness of the best solution in a population does not deteriorate as the generation advances and helps GA converge to the global optimum [24]. GA in this work integrates *global elitism* in the selection mechanism instead of *simple elitism*. According to Vasconcelos *et al*. [25], simple elitism refers to the case where the best individual at generation k (the father) is maintained in the next generation $k + 1$ if its child has a performance inferior than its parent. Indisputably, without the elitism, the best individuals would have been lost during the process of selection, mutation and crossover operations. However, in the case of global elitism, each individual in the population of generation k+1 can replace its father of generation k, if it has a performance superior than him. Consequently, at a generation k+1, the individuals are better than the individuals at generation k.

Fig. 1 above presents the pseudo code of GA with elitism. The schematic diagram in Fig.2 summarizes the overall evolutionary process flow in GA where the elitism is applied to the same current population that also undergoes genetic alteration through crossover or mutation operations, the aggregation of which produces the next *generation*, i.e., *new current population*.

Effective implementation of a genetic algorithm for any problem demands attention into following distinct and yet related tasks:



```
t ← 0; //Iteration Counter
Generate initial population Pop(t) of size PopSz;
Evaluate Pop(t);
while Stopping criterion not satisfied do
    Select parent population Pop'(t) from Pop(t);
    Apply genetic operators to Pop'(t)! → Pop(t+1);
    Replace random solutions in Pop(t+1) with the best P
    solutions in Pop(t);
    Evaluate Pop(t+1);
    Get Elitist Population
    t ← t+1;
Result: Best solution having highest fitness value from
the population in the last generation.
```

Figure 1: Pseudo Code Algorithm for the Elitist GA (EGA)

1. Encoding or representation of the chromosome,
2. Implementation of the genetic operators, and
3. Implementation of fitness function

These factors are the deciding elements, greatly responsible for the quality of the solution obtained and thus the performance of the GA which are discussed in the subsequent section.

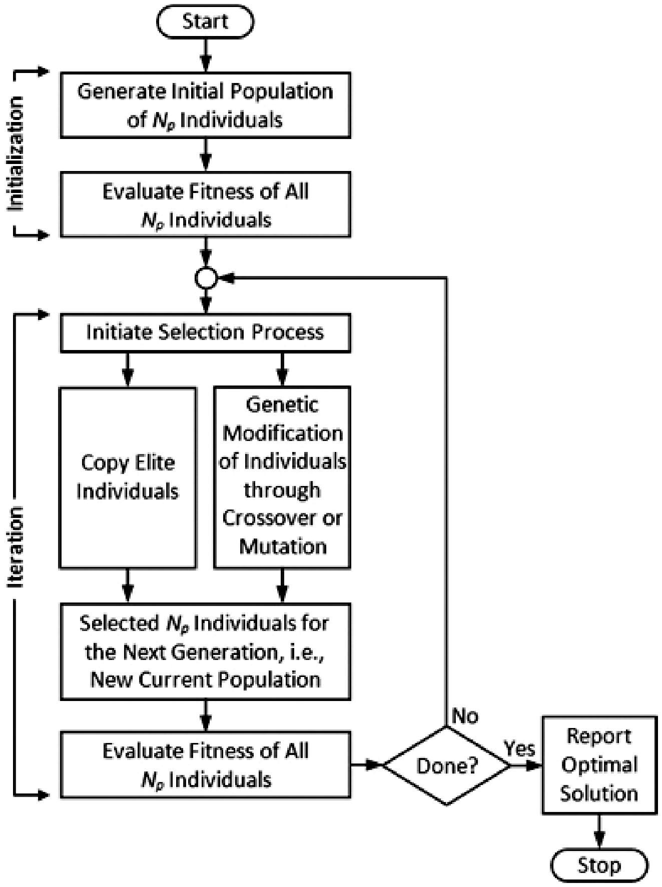

Start
Initialization
Generate Initial Population of N_p Individuals
Evaluate Fitness of All N_p Individuals
Iteration
Initiate Selection Process
Copy Elite Individuals
Genetic Modification of Individuals through Crossover or Mutation
Selected N_p Individuals for the Next Generation, i.e., New Current Population
Evaluate Fitness of All N_p Individuals
No
Done?
Yes
Report Optimal Solution
Stop

Figure 2: Flowchart for the Elitist GA (EGA)

5. PROPOSED EGA BASED IMAGE CLUSTERING

In this section, we present the elitist GA (EGA) based approach for solving the automatic (unsupervised) image clustering problem. A flow diagram of the proposed EGA based image clustering method is presented in Fig. 3 below:

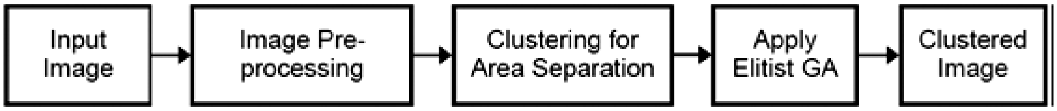


Figure 3: Phases in the EGA based Image Clustering

Broadly, the proposed system comprises of 3 steps: (1) the data cleaning (noise removal) phase where missing values, if any, from the data is replaced with mean, (2) the cleaned datasets is clustered using K-means to remove outliers, inconsistent and noisy data and the reduced data is used for selecting the optimal features with genetic algorithm, and finally, (3) the reduced dataset is classified using EGA as an image classifier [29].

Pre-processing of Input Image Data Set [6]: The input image can be any valid picture file format: jpg, bmp, etc. Image normalization image is carried out in order to remove possible presence of impurities and noise after the input from the system is taken. In this case, the normalized image implies transforming the image as per some defined specifications. During this image pre-processing phase, enhancements to the gray scale image are accomplished. Once the image pre-processing step is completed, the next task is to define this image as the initial population data set for the EGA. A very large input data set can be pre-processed to create a characteristic data set that can be used by the algorithm for better time and space efficiency. We implemented two alternate pre-processing methods for our clustering algorithm. The first pre-processing method used random sampling to obtain a data set with fewer points. This reduced data set was then used in evaluating the fitness of the chromosomes. The second pre-processing technique used is the summarization of the input data set and is based on the work presented in reference. For this method, a grid is first constructed and then the input data set is applied to this grid. A single point location and the corresponding weight are calculated for each region defined by the grid. The location of the representative point is chosen as the mean value of all the points in the neighbourhood and the weight of the characteristic point is equal to the number of points that it replaces.

K-Means Clustering [29]: The K-mean clustering algorithm is applied on the resulting image to accomplish the requirement for certain classification carried out in the second phase. The K-means clustering algorithm was developed in 1976 by MacQueen. It is an unsupervised clustering algorithm that generates a specific number of disjoint, flat (i.e., non-hierarchical) clusters. Basic purpose of image clustering is to categorize the image areas. As per the clustering approach followed here, it is required to define the number of clusters along with the centre of these groups. Once the job of defining the clusters or groups are done, the next task is to identify the Euclidean distance of each pixel from the corresponding centre points. As per the measure of these distances, the picture elements (i.e., pixels) are placed in the appropriately specified regions. When the clustering process is over, now we will consider the cluster that represents the central area of the image as the data for the new population set. This procedure adopts a simple and easy approach to classify a given data set through a specific number of clusters (assume *k* clusters) fixed apriori. K-means algorithms randomly choses *k* objects, representing the *k* initial cluster center. The next phase is to take each point under the scope of a given data set and associate it to the nearest cluster center based on the closeness of the object with cluster center by Euclidean distance. After all the objects have been distributed, compute a new *k* cluster centers. The process is repeated until there is no change in *k* cluster centers. The goal of this algorithm is to minimize an objective function known as squared error function given by the following.

$$f(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left\| x_i - v_j \right\| \qquad (2)$$

where,

‘$||x_i$ - $v_\mathrm{j}||$’ is the Euclidean distance between x_i and v_j,

‘c_i’ is the number of data points in i^{th} cluster, and

‘c’ is the number of cluster centers.

The steps of the K-means Algorithm are given below:

1. Randomly partition the dataset into k subsets.
2. For each data point in the dataset:
   - Calculate the distance from the data point to each cluster.
   - If the data points are close to their own cluster, leave it where it is. If the data point is not the closest to its own cluster, transfer it into the closest cluster.
3. Repeat the previous step until a complete walk through all the data points will result in no data point moving from one cluster to cluster. At this point the clusters get stablilized and the clustering or grouping process ends.
4. To be clear, the selection of the initial partition can greatly affect the final clusters that result, in terms of intra-cluster and inter-cluster distance and cohesion.

Elitist Genetic Algorithm: In the final stage, the implementation of the EGA is carried out on this resulting population set. The EGA procedure starts with some initialisation being carried out t6hat involves specification of input in terms of population set and the number of iterations to be carried out by the algorithm. After all these specifications, the EGA is commenced and is processed by the algorithm through steps of selection, crossover, mutation etc. in a repetitive manner.

The selection phase performs selection of any two random pixels to perform the comparative analysis. The crossover operation is being performed on these pixels to select the next elected pixel and therefore, it is followed by the mutation process as the election or the rejection of the specific pixel. This can carry out some alterations on the pixel data, if required. Once the GA steps are concluded, this will produce a valid threshold value that helps in the decision making process for selection of the pixel. This selected pixel area is now presented as the detected target image.

Different stages involved in the implementation of an elitist genetic algorithm are shown below in Fig.4 below:

Initial Population: We originally decide the size of the population, and then the method by which the members for the population are chosen. The primary idea of the size of the population always forms a trade-off between the efficiency and the effectiveness of GA implementation. In general, it appears that there should be some ‘ideal’ value for a given length or size of the population, on the grounds that too little a population would not permit adequate opportunity for exploring the search space effectively, while too big a population would damage the efficiency of the approach that no solution could be expected in an realistic amount of time limit [15].

Termination Condition: Traditional neighbourhood search methods are simple in the sense that they terminate when a local optimum is encountered. On the other hand, GAs are blind search stochastic optimization techniques that could in principle run for endlessly, but could converge fast to the solution with right termination criteria set. In practice, a termination criterion is mandatory to prevent the iterations get infinite; the common

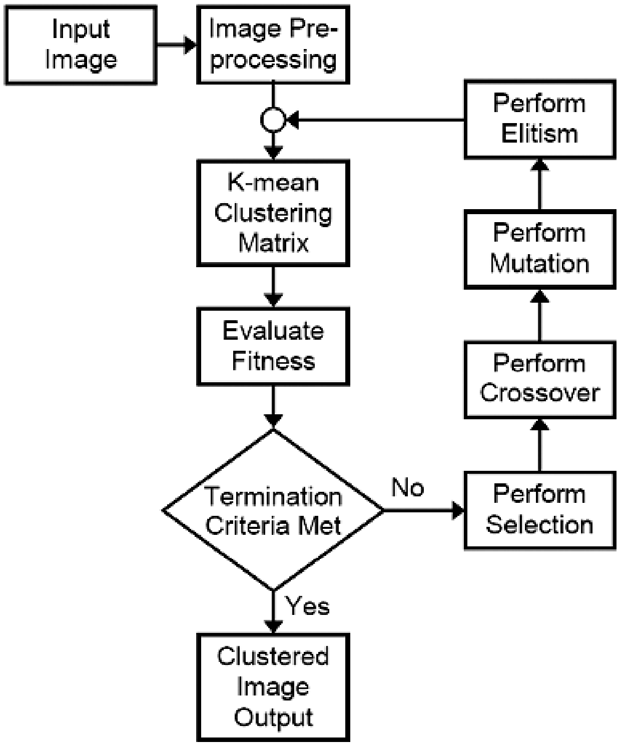


Figure 4: Flow graph of the EGA based Image Clustering

practices are to set a limit on the number of fitness evaluations or the computation time consumed, or to track the variability in the population and stop when this falls below a pre-set boundary value. The meaning of diversity in the latter case is not always apparent, and it could relate either to the genotype or the phenotype, but the most effective way is through a measure of the genotype statistics. For example, one could decide to break out of the iteration if at every locus the proportion of one particular allele rises above 90% [15].

Selection: There are many approaches to implement a genetic selection procedure. The key idea behind any selection method is that it should be related to fitness of the member in the population. Roulette-wheel scheme is a commonly known traditional approach to implement a selection scheme that makes use of a probability distribution function for the selection where the selection probability of a given string is proportional to its fitness value. Pseudo-random numbers are used one at a time to choose strings for the pool of parents [15].

Crossover: Crossover is simply a matter of replacing some of the genes in one parent by the matching genes of the other. Imagine two strings *a* and *b*, each consisting of six variables: $a = (a_1, a_2, a_3, a_4, a_5, a_6)$ and $b = (b_1, b_2, b_3, b_4, b_5, b_6)$ which represent two possible solutions to a problem. Two cross points are chosen randomly from the range 1…5, and a new solution created by merging the parts of the original ‘parents’. For instance, if the crossover points were 2 and 4, the ‘offspring’ solutions produced would be $a = (a_1, a_2, b_3, b_4, a_5, a_6)$ and $b = (b_1, b_2, a_3, a_4, b_5, b_6)$. The central argument is that there exits two sources of bias that need to be exploited in a genetic algorithm: the positional bias and the distributional bias. A simple crossover has substantial positional bias, in the sense that it relies on the building-block hypothesis, and if this is invalid, the bias may prevent the production of good solutions as desired.

Mutation: Mutation brings diversity in the population and adds new information in a random way to the genetic search process that eventually helps avoid getting trapped at local optima. Mutation is an operator that inserts variability in the population whenever the population gradients homogeneity due to repeated use of genetic operators: reproduction and crossover. Mutation may cause the chromosomes of entities to be completely different from those of their parent entities [3]. Mutation is the process of randomly disturbing and disturbing the genetic information. They operate at the bit level; when the bits are being taken from the current string to the new string. For example, the mutation operation can be illustrated as follows: let in the string (chromosome) 1011001, with genes 3 and 5 are mutated, resulting in a new chromosome 1001101. There involved probability P_m called the mutation probability that controls the way each bit may become mutated. In practice, P_m is usually a small rational number. A coin toss mechanism is employed; if the random number among zero and one is less than the mutation probability P_m, then the bit flip happens, so that the '0' becomes '1' and '1' becomes '0'. This helps in introducing a bit of diversity in the population by scattering the infrequent points. This random scattering would result in bringing out better optima, or even modify a part of the genetic code that will be beneficial in subsequent operations. Instead, it might also happen that the operation produces a weak individual that will never be selected for further operations [15].

Evolution – New Population with Elitism: Original GA implementation presumed to adopt a genetic evolution approach as per Darwinian Theory where the selection, recombination and mutation were used for a population of M chromosomes till a new set of M chromosomes had been produced. This set then forms the new population. However, from the optimization perspective this appears to be an abnormal thing to carry-on where we may have to spend considerable work in locating a good solution, only to run the risk of discarding it away and thus debarring it from taking part in further reproduction. Elitism and population overlaps are two approaches that are used to overcome such problems. An elitist scheme ensures that the objective of surviving the best fit individual found so far is accomplished by preserving it and replacing only the remaining (M–1) chromosome of the population with a new chromosome. On the other hand, the overlapping populations take this a step further by changing only a percentage G (the generation gap) of the population at each generation. Finally, taking this to its consistent conclusion produces so-called steady-state or incremental schemes, in which only one new chromosome is generated at each stage [15].

6. RESULTS ANALYSIS

The proposed algorithm for image clustering was studied through extensive simulation experiments. For the purpose of simulation of our proposed method, we have developed the simulation programs in our lab on Matlab platform using Intel Core i3 processor (2.20GHz) with 4GB RAM based computing system. In this section we first discuss the simulation approach and then present analytic thinking of the simulation results. Broadly, our simulation approach involved execution of the algorithms based on our novel idea presented in this paper on each input image with appropriate initialization done as per specifications. A flow graph model of the simulation program is presented in Fig. 2 and Fig. 4.

Simulation was carried out varieties of test image.

Figures below shows varieties of source (Figures 5 and 6):

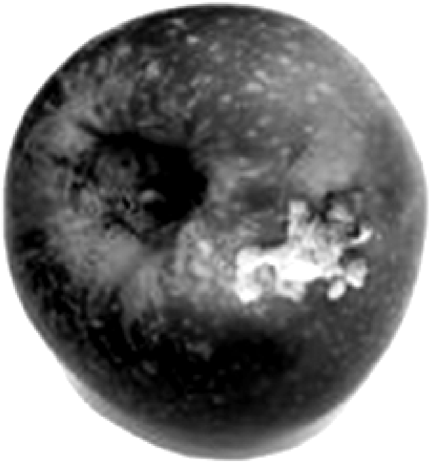


Figure 5: Source Image – 1

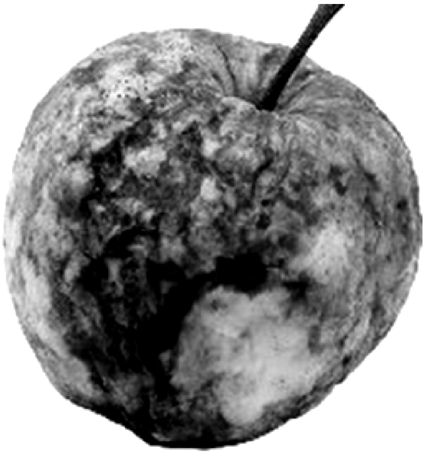


Figure 6: Source Image – 2

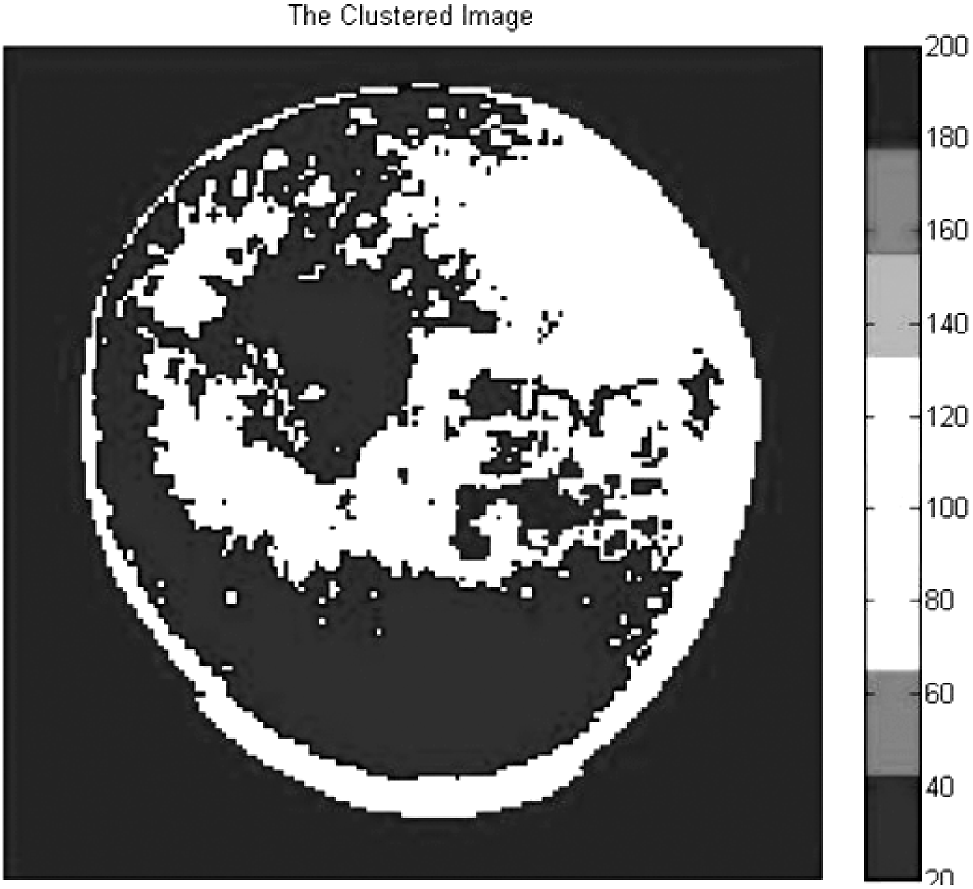


Figure 7: Clustered Output of Image-1

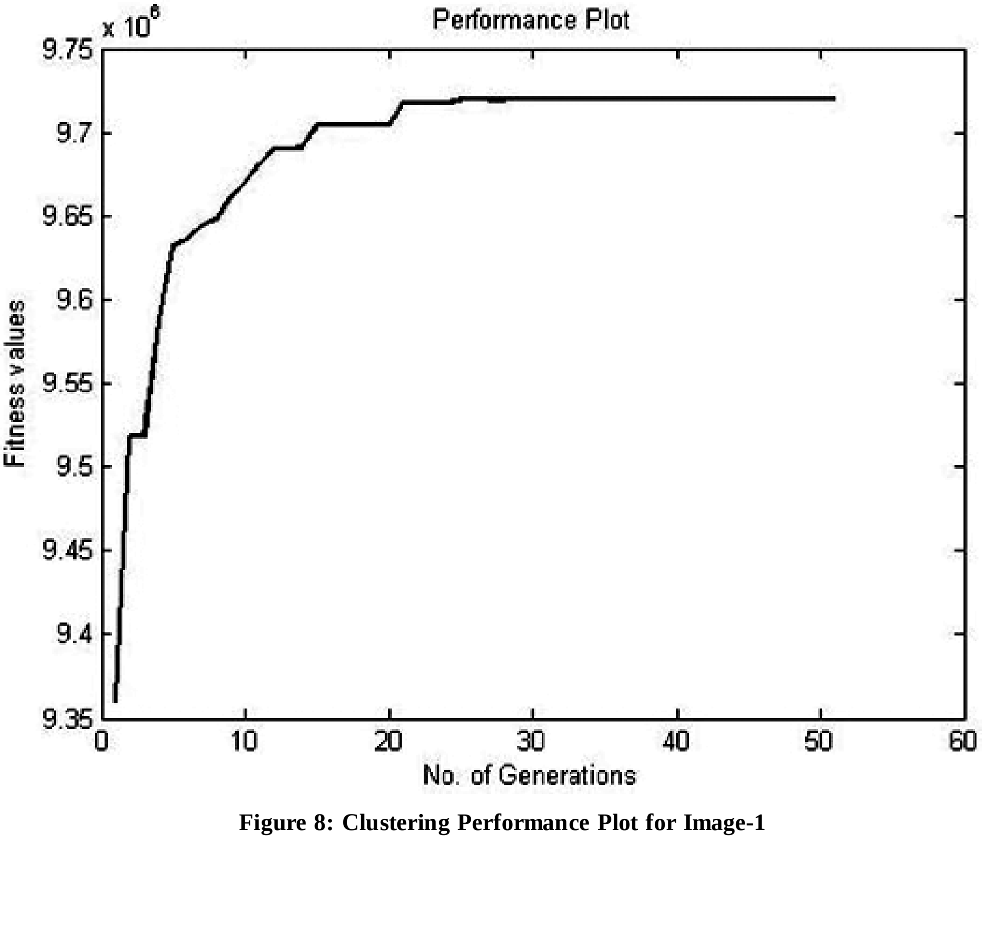


Figure 8: Clustering Performance Plot for Image-1

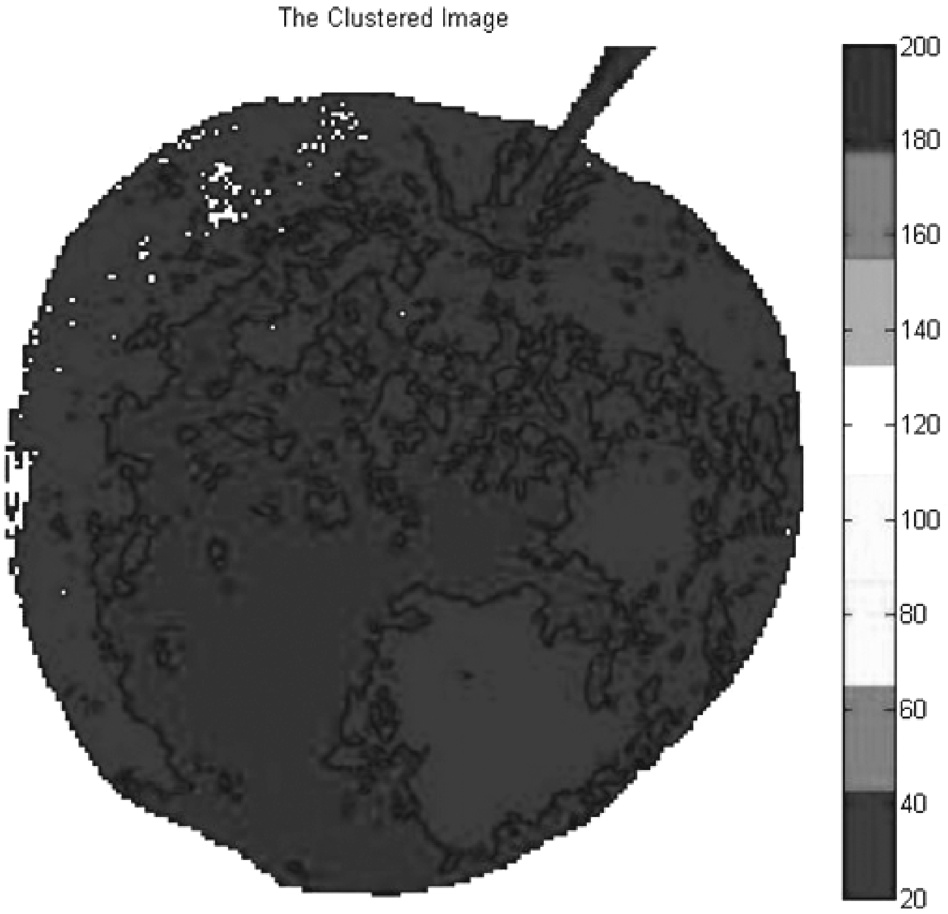


Figure 9: Clustered Output of Image-2

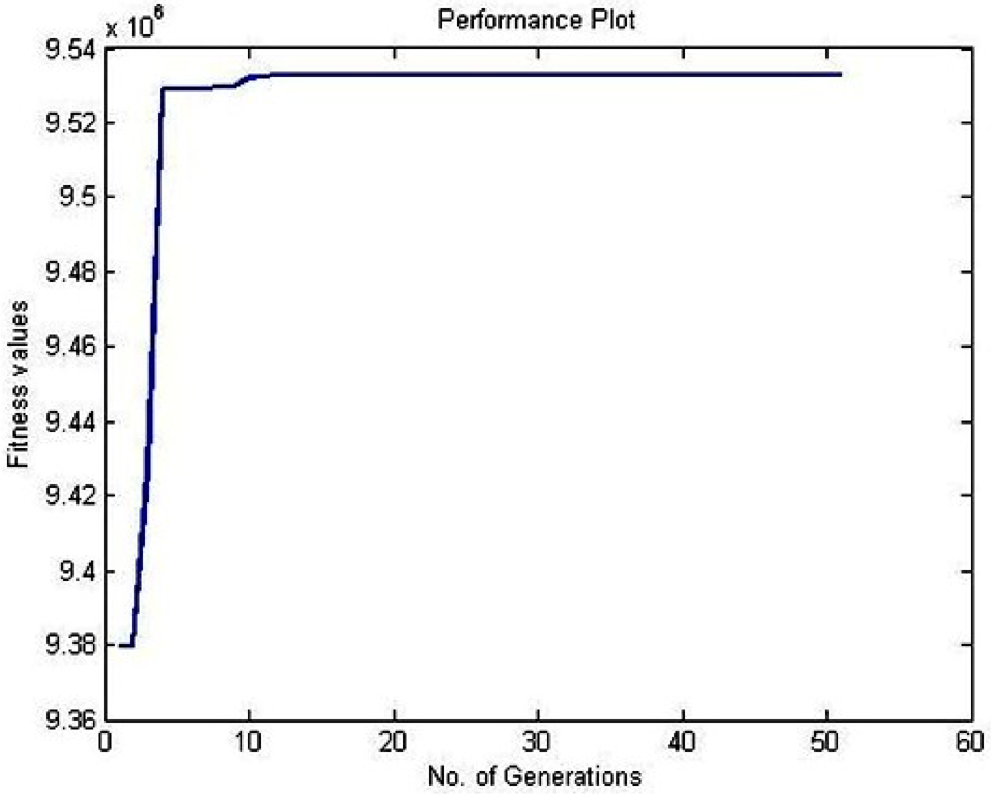


Figure 10: Clustering Performance Plot for Image-2

Simulation was carried out average populations size of 50 with 60 generations and 300 iterations to execute the operation. The source images are of different in size. Fig.7 and 9 shows clustered image output while Fig.8 and 10 shows the corresponding performance plots.

The result shows the clustered image as well as the performance plot. The performance plot represents the number of generations in *x* axis and the fitness values in *y* axis. As we have considered the solution to be a minimization problem so the performance plots show the decrement. Convergence pattern achieved in the simulation study plots establishes the truthfulness of the proposed approach.

7. CONCLUSION

Clustering is an important task having applications in many fields. Heuristic algorithms are used for this task in an attempt to provide acceptable results, both in terms of solution quality and running time. GA has been applied to the clustering problem for many applications. For clustering on very large data sets, such as image data sets, the size of the related databases makes it necessary to modify the traditional GAs because of their slow run-time behaviour and convergence. In this paper we proposed an Elitist GA (EGA) approach with efficient encoding technique and GA operators along with input set pre-processing. The experimental results here show steady and promising performance.

REFERENCES

[1] Metin KAYA, An Algorithm for Image Clustering and Compression, Turk J Elec Engin, Vol.13, No.1, pp.79-91, 2005.

[2] Carolina Raposo, Carlos Henggeler Antunes, and Joao Pedro Barreto, Automatic Clustering using a Genetic Algorithm with New Solution Encoding and Operators, Proceedings of the 14th International Conference ICCSA 2014, Lecture Notes in Computer Science, Computational Science and Its Applications – ICCSA 2014, Series Volume 8580, Part II, pp 92-103, June 30 – July 3, 2014.

[3] Akbar Shahrzad Khashandarag, Mirkamal Mirnia and Aidin Sakhavati, A New Method for Medical Image Clustering Using Genetic Algorithm, International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, pp. 551-557, January 2013.

[4] Mamta Mor and Poonam Gupta, A Review on Clustering with Genetic Algorithms, International Journal of Computer Science & Communication Networks, Vol. 4, No. 3, pp.94-98

[5] Nameirakpam Dhanachandra, Khumanthem Manglem and Yambem Jina Chanu, Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm, Proceedings of the 11th International Multi-Conference on Data Mining and Warehousing, ICDMW 2015, August 21-23, 2015, 11th International Conference on Image and Signal Processing, ICISP 2015, August 21-23, 2015, Volume 54, pp. 764 – 771, 2015.

[6] Qin Ding and Jim Gasvoda, A Genetic Algorithm for Clustering on Image Data, World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol. 1, No. 5, pp. 1506-1511, 2007.

[7] Jacob Goldberger, Hayit Greenspan and Shiri Gordon, Unsupervised Image Clustering using the Information Bottleneck Method, Proceedings of the 24th DAGM Symposium Zurich, Switzerland, Lecture Notes in Computer Science, Pattern Recognition, Series Volume 2449, pp. 158-165, September 16–18, 2002.

[8] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, Unsupervised Image-Set Clustering Using an Information Theoretic Framework, IEEE Transactions on Image Processing, Vol. 15, No. 2, pp. 449-458, February 2006.

[9] Ming-Chuan Hung and Don-Lin Yang, An efficient Fuzzy C Means clustering Algorithm, Proceedings of the IEEE International Conference on Data Mining (ICDM 2001), pp.225-232, 2001.

[10] Elnomery A. Zanaty, and Ahmed S. Ghiduk, A Novel Approach Based on Genetic Algorithms and Region Growing for Magnetic Resonance Image (MRI) Segmentation, International Journal of Computer Science and Information Systems, Vol. 10, No. 3, pp. 1319-1342, June 2013.

[11] Zhensong Chen, Zhiquan Qi, Fan Meng, Limeng Cui and Yong Shi, Image Segmentation via Improving Clustering Algorithms with Density and Distance, Proceedings of the International Conference on Information Technology and Quantitative Management (ITQM 2015), Vol. 55, pp. 1015–1022, 2015.

[12] Amiya Halder, Soumajit Pramanik and Arindam Kar, Dynamic Image Segmentation using Fuzzy C-Means based Genetic Algorithm, International Journal of Computer Applications, Volume 28, No.6, pp..15-20, August 2011.

[13] Yevgeny Seldin, Sonia Starik and Michael Werman, Unsupervised Clustering of Images using their Joint Segmentation, Proceeding of the 3rd International Workshop on Statistical and Computational Theories of Vision (SCTV 2003), pp.1-24, November 2003.

[14] Amanpreet Kaur and Gagandeep Jindal, Overview of Tumor Detection using Genetic Algorithm - Clustering Flowchart, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2, No. 2, pp.348-352, April 2013.

[15] Ashwini Gulhane, Prashant L. Paikrao and D. S. Chaudhari, A Review of Image Data Clustering Techniques, International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, No. 1, pp.312-315, March 2012.

[16] Mengqiu Tian, Qiao Yang, Andreas Maier, Ingo Schasiepen, Nicole Maass and Matthias Elter, Automatic Histogram-Based Initialization of K-Means Clustering in CT, Proceedings of the Workshop ASA-2013, Vol. 3, No. 5, pp. 277-282, March 2013.

[17] Stelios Krinidis, Michail Krinidis and Vassilios Chatzis, An Unsupervised Image Clustering Method Based on EEMD Image Histogram, Journal of Information Hiding and Multimedia Signal Processing, Volume 3, Number 2, pp.151-163, April 2012.

[18] Koreddi Venkatesh, G.Sudhakar, A.Prakashni and P.Shyamala Madhuri, Automatic Image Pixel Clustering Using Genetic Algorithms, International Journal of Computer Science And Technology, Vol. 4, Issue Spl - 4, pp. 250-251, Oct-Dec 2013.

[19] David E. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning, Pearson Education, 2006.

[20] Sridharan B. Modifications in Genetic Algorithm using Additional Parameters to make them Computationally Efficient. Proceedings of the IEEE 2nd International Advance Computing Conference; pp. 55-59, Feb 19-20, 2010.

[21] Cheng H, Yang S. Genetic Algorithms with Elitism-based Immigrants for Dynamic Load Balanced Clustering Problem in Mobile Ad hoc Networks. IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments, pp. 1-7, April 15-17, 2011.

[22] Melanie M. An Introduction to Genetic Algorithms. The MIT Press, Massachusetts. 1999.

[23] Rudolph G. Convergence Analysis of Canonical Genetic Algorithms. IEEE Transactions on Neural Networks. 1994 Jan; 5(1):96-101.

[24] King RTFA, Rughooputh HCS. Elitist Multiobjective Evolutionary Algorithm for Environmental/Economic Dispatch. Proceedings of The 2003 Congress on Evolutionary Computation; 2003 Dec 8-12; pp. 1108-14, Vol.2.

[25] Vasconcelos JA, Ramírez JA, Takahashi RH, Saldanha RR. Improvements in Genetic Algorithms. IEEE Transactions on Magnetics. 2001 Sep; 37(5):3414-7.

[26] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan, The Planar k-Means Problem is NP-Hard, In Proceedings of the 3rd International Workshop on Algorithms and Computation (WALCOM ’09), Springer-Verlag, Berlin, Heidelberg, pp. 274-285, 2009.

[27] Amir Alush and Jacob Goldberger, Hierarchical Image Segmentation using Correlation Clustering, IEEE Transactions on Neural Networks and Learning Systems (Volume: 27, Issue: 6, June 2016), Page(s): 1358 – 1367,

[28] Julian Yarkony, Chong Zhang and Charless C. Fowlkes, Hierarchical Planar Correlation Clustering for Cell Segmentation, Proceedings of the 10th International Conference, EMMCVPR 2015, Hong Kong, China, January 13-16, 2015.

[29] T. Santhanam and M.S Padmavathi, Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis, Proceedings of the International Conference on Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014), Procedia Computer Science, Vol. 47, 2015, pp. 76–83