**An elitist approach to analyze breast cancer histology slides using genetic algorithm**

Rimpa Bairagi1, Shouvik Chakraborty2 (0000-0002-3427-7492)

1,2Women’s Polytechnic, Chandernagore

1rimpabairagi994@gmail.com, 2shouvikchakraborty51@gmail.com

Abstract:

**Keywords:** breast cancer, microscopic image segmentation, genetic algorithm, elitism, histology

**I. Introduction**

Image segmentation is the process of detecting multiple regions or segments from an image for analysis. With this segmented region, it is possible to analyze it further. Another way it can be viewed is that image segmentation sets the label for every pixel of an image. Image segmentation is used for different kinds of purposes, like object recognition, scene understanding, image editing, medical imaging, etc. By using segmentation, it is possible to easily extract objects, detect boundaries, classify pixels, etc. There are several methods available to perform image segmentation. The methods that are frequently used for image segmentation are thresholding, region-based segmentation, edge detection, clustering, contour-based segmentation, deep learning-based segmentation, etc. Every segmentation approach has some pros and cons, and the choice of a method depends on the image characteristics, task requirements, and available resources. Biomedical image segmentation is a very important part of medical imaging. Biomedical image Segmentation is the method of partitioning a medical image into various constituting parts depending on their diagnostic significance. Biomedical images contain different types of information in different modalities like X-rays, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and microscopy images.

Microscopy is an important diagnostic tool that helps in detecting the root cause of different diseases and possesses a significant impact on diagnosis and treatment planning. Breast cancer is one of the frequent diseases and many patients are suffering throughout the world. Microscopic images can be helpful in diagnosing this disease and also help in treatment planning. Segmentation of the breast histology slides can be helpful for domain experts in easy and quick understanding of the scenario. The images obtained at the tissue level and cellular level from tissue samples of the breast are obtained by biopsy or surgical procedure and analysed. Segmenting these images helps to detect the different tissue structures and irregularities that are key to the diagnosis and treatment planning of breast cancer. All breast cancers are not the same because of their cellular appearance and behavior. By segmenting microscopy images, one can easily see the patterns and structure of cells that determine the cancer subtype (how destructive the cancer cells look), which helps to identify whether some of the cancer might still be left in the body or not, and also ensure that the entire tumor has been removed or not. Clustering is the method of assembling the data points in a large data set. Clustering and assembling the data depend on the properties of the data. Each cluster contains some data points with the same properties as others. Labelling in clustering is performed without any guidance because it is an unsupervised machine-learning technique. Clustering is used in broad areas like marketing, biology, libraries, insurance, city planning, image processing, etc. If we talk about real-world examples of clustering, then we can say that when we go to restaurants for food, if we carefully observe the menu card of the restaurant, we can see the same type of food is organized in the same class like start-ups will be in the same class, the main course will be in another class, soft drinks will be in another class, etc. Exactly using this method, a clustering approach is used in image segmentation. By using distance matrices like Euclidean Distance, Hamming, Manhattan, etc., we measure the distance from one data point to another, and by using this measured result, we assemble data points. There are various clustering algorithms that we use for clustering, such as K-means clustering, hierarchical clustering, density-based spatial clustering (DBSCAN), Gaussian mixture models (GMM), etc. The use of this clustering algorithm depends on the data, the number of clusters, and the specific problem. We use clustering for different types of problems like image segmentation, image analysis, anomaly detection, etc. There are certain challenges associated with clustering in image segmentation. For clustering in image segmentation, a similar type of neighborhood data is required. Typically, an image contains overlapping background, variable contrast, different color, different shapes, etc. Because of this type of variability, the difficulty level is too high for clustering in image segmentation. Thereafter, during image acquisition, some noise from the environment may also get captured, which would affect the result of the clustering. If we talk about clustering in biomedical image segmentation, then it will be more difficult because biomedical images contain huge amounts of dimensional data and features for every pixel, which expands the estimation complexity of clustering algorithms. Many biomedical images consist of very complex structures. Because of the complexity of biomedical images, it is not possible always to achieve the desired accuracy of segmentation.

Evolutionary approaches are proven to be effective in solving various real-world problems. It works on sophisticated problems and tries to find an optimal solution to a problem. The advantages of evolutionary approaches can be exploited in clustering-based biomedical image segmentation. Various techniques such as genetic algorithms, particle swarm optimization (PSO), and ant colony optimization (ACO), etc. can be exploited to determine the optimal clusters. Genetic algorithm is a very popular and frequently used approach and proves its effectiveness in solving various complex optimization problems. John Holland derives the idea of the genetic algorithm from the natural reproduction and survival of the fittest mimics various genetic operations. By using the metaheuristics method in the clustering of biomedical image segmentation it is possible to avoid the problem of the traditional clustering approach. Typically, the K-Means clustering approach suffers with the choice of the initial cluster centers. GA based solutions can eliminate this redundancy. Metaheuristics can maximize the accuracy of clustering-based image segmentation by exploring the search space effectively. So, in this work, an elitist genetic algorithm is used to get an optimal solution to the biomedical image segmentation problem. We can eradicate the effect of noise and intensity from an image by using the genetic algorithm.

The rest of the article is organized as follows: Section II illustrates the methodology including the proposed approach. Section III illustrates the experimental outcomes and their significance. Section IV concludes the article.

**II. Methodology**

**A. A brief overview of the K-Means clustering approach**

K-means is the most popular and frequently used unsupervised clustering approach. By using the K-means algorithm it is possible to perform image segmentation. Here the value of the K refers the how many clusters in which an image is to be divided i.e., how many groups are to be created from a large data set. The value of K depends on the problem and varies from image to image. The value of K is initialized manually. After setting the value of K the process begins by choosing center points from the data set where the center point is also chosen randomly. The distance of every point from the cluster centers is measured by using the Euclidean distance method. The class label is assigned according to the closest center point. The process gets repeated until the points in different clusters remain unchanged. The cluster centers get updated by computing the mean of the coordinate values.

**B. Genetic algorithm**

The idea of a genetic algorithm was invented by John Holland. It is really helpful to find out the optimal solution for a given problem. We all know Darwin’s theory. Sir Darwin invented the theory of evolution. Sir John Holland uses this idea in the Genetic Algorithm. Darwin says that species want to survive for the next generation but few species are only can survive, In GA the concept is the same as Darwin’s theory. In GA the possible solution set in the data set we can consider as population and the subset of a population is the chromosome which means among the many possible solutions one solution and gene are the subset of chromosome and population also. In GA we first find the fitness value of each chromosome by using the fitness function and the fitness value of each chromosome will be different. Which chromosome will have the best fitness value that chromosome will survive to create offspring there are 3 operators in the genetic algorithm. The first operator is a selection this operator selects randomly two chromosomes in a pair for crossover. There are three types of selection processes i.e., Roulette wheel selection, Event selection, and Rank-grounded selection. The second operator is cross-over where a gene is selected from each of the two parents' chromosome to exchange their information until the gene meets and create offspring. There are three types of crossover methods i.e., One-point crossover, Two-point crossover, Livery crossover, and Inheritable Algorithms crossover. We can use any method among them for crossover. The third operator is a mutation, it maintains the diversity in a population. In mutation, we assign the gene randomly in the chromosome to avoid premature convergence and diversification. We can mutate by three techniques i.e., Flip bit mutation, Gaussian mutation, and Exchange/Swap mutation. After that get the peak value in the population. When we use a genetic algorithm in clustering it is called GA Clustering. We use a genetic algorithm to minimize the cluster matrix M.

**Elitism in Genetic Algorithm** An important thing in the genetic algorithm is elitism. It transfers the best individual chromosome from one generation to another. It Works like that when we find the fitness value then the best fitness value of a chromosome is stored in a variable and that best fitness value we do not pass from the selection, crossover, and mutation. Chromosomes pass through the selection, mutation, and crossover process among them one chromosome is directly changed with the chromosome that carries the best fitness value. We use elitism in the genetic algorithm to find the space where we can find the peak solution. By using it we can maintain a high-quality population, improve convergency, and especially find the optimal solution in a challenging fitness landscape.

**Propose Approach:-** Here we propose a biomedical image segmentation approach that is based on an elitism genetic algorithm. Here it updates the cluster center by using an elitism genetic algorithm. If the proper initial cluster center point is chosen incorrectly then its output becomes wrong which means the output of a clustering is indirectly dependent on the initial cluster point that is chosen manually. Otherwise, there are many problems like complete clustering iterations which will take more time because it wants to reach convergence, and also it is not a sufficient algorithm. Basically by using an elitism genetic algorithm, we try to solve those types of problems. If this approach is applied in clustering then the overcome will be possible. In K-means, the center point is by using the average value. But here the method is different from the means, here the cluster center has evolved by using an elitism genetic algorithm. after that, the cluster center will be updated by the evolutionary algorithm and then the output will be independent. How it has been done is discussed below:

1. **Data Set Preparation:-**  At first, we assume the number of pixels will be assigned as several rows and the number of dimensions will be assigned as several columns. If a picture considers 100\*100 then the number of rows will be 10000 and each pixel contains 3 dimensions (red value, green value, blue value) so the column number will be 3. The value of the columns will be the value of the 3 dimension value of each pixel. Here the three-dimensional data set is converted into the two-dimensional data set.
2. **Chromosome Encoding:**- In GA the chromosome is selected randomly from the data set. If an image has 100 pixels then there are a total of 100 values in the data set and each value contains a Red value, a green value, and a Blue value. To create a chromosome we randomly choose the value by using this formula 1+(rand\*maximum value). Here the gene number in a chromosome depends on the number of clusters which means the number of clusters is equal to the number of genes in a chromosome.

Population

Chromosome

Gene

K L M N O

F G H I J

A B C D E

1. **Selection:-** In selection, there are many processes i.e., Tournament selection, Roulette wheel selection, Event selection, and Rank-grounded selection. Among them, most of the most popular selection method is tournament selection. It works like a tournament competition. At first, the tournament size should be selected then choose randomly chromosomes from the population for selection. That means the selected chromosome is a subset of the population. After that, the fitness value of the two individual chromosomes is compared to which chromosome be the best fitness value that chromosome will be transferred to the next generation. Tournament selection has two types, One is k-way tournament selection and another is 3-way tournament selection. It has been explained fig in below:- Chromosome1

A B C D E

F G H I J

A B C D E

By

It contains best fitness value among these two chromosome for this it chosen by tournament selection

Chromosome2

**Fig: Tournament selection process**

4. **Crossover:-** There are many processes in crossover i.e., One-point crossover, Two-point crossover, Livery crossover, and Inheritable Algorithms crossover, etc. But today’s popular methods are one-point one-point crossover, Two-point crossover, and Uniform crossover. Another process is used according to the problem. In one point crossover, one crossover point is chosen randomly from two chromosomes, and all gene behind that point is exchanged their gene as a string. In the two-point crossover, two points are chosen randomly from two chromosomes among these two point genes, which are exchange genes as a string. In the Uniform method, one gene from a chromosome is selected randomly and the corresponding chromosome of that point of the gene is twice. In the uniform method mask is used, it looks like a chromosome but it only helps to do the crossover method. It is not exactly the parent chromosome because the mask does not participate as a parent chromosome in a population and it generates randomly. the chromosome of that point of the gene is twice from one value to another. Which chromosomes that participate in the crossover chromosomes are called parent chromosomes. After crossover which chromosomes are created that chromosome is known as offspring.

1 0 1 1 0 1 1 1 0 1

p1

0 1 1 0 1 0 1 1 0 1

p2

1 0 1 1 0 1 0 1 0 1

Parents Mask

1 0 1 1 0 0 1 0 1 0 1

O1

O2 1 0 1 1 0 1 1 1 0 1

Offspring

5. **Mutation**:-In Darwin’s theorem Sir Darwin states that mutation possibility in a gene is very low. In mutation, there are many methods for mutating an offspring i.e., Flip bit mutation, Gaussian mutation, and Exchange/Swap mutation. The mutation method is used in genetic algorithms to overcome the premature convergence problem. In the mutation process, one gene is selected randomly for mutation. Mutation happens by flipping selected genes. In this process, one or more genes can be flipped in individual offspring.

**Before Mutation After Mutation**

75 29 31 58 35 98 46 65 53 13

100 54 98 17 26 23 54 97 46 79

75 29 31 58 35 98 46 64 53 13

100 54 98 16 26 23 54 97 46 79

6. **Fitness Function**:- For calculating the fitness value here Davies Bouldin Index algorithm and Silhouette algorithm will be used. Davies Bouldin Index is also called DBI. Get a result after doing clustering how will be decided the cluster quality is best? For deciding the quality of the cluster there is used the DBI algorithm. In DBI, calculating the average distance between each point and center point in a cluster is called within cluster distance, and calculating the average distance from the center point of a cluster to each center point of another cluster are there it is called outer cluster distance. It happens that each of the clusters will be there. The implementing process of DBI is given below:

1. Calculate the within-cluster distance, and the formula is given below:

Si = (1/Ti ∑||X j -A i ||qp) 1/q   Where Ai = Centroid of Ci, Ti = is the size of the cluster i Si is the qth root of the qth moment of the points in cluster i about the mean. If then is the average distance between the feature vectors in cluster I and the centroid of the cluster. Usually, the value of p is 2 which makes the distance a Euclidean distance function.

1. Calculate the outer cluster distance, and the formula is given below:

Mij = ||Ai – Aj||p = ∑|ak, i – ak, j|p )1/p

6. **Fitness Function**:- For calculating the fitness value here Davies Bouldin Index algorithm and Silhouette algorithm will be used. Davies Bouldin Index is also called DBI. Get a result after doing clustering how will be decided the cluster quality is best? For deciding the quality of the cluster there is used the DBI algorithm. In DBI, calculating the average distance between each point and center point in a cluster is called within cluster distance, and calculating the average distance from the center point of a cluster to each center point of another cluster are there it is called outer cluster distance. It happens that each of the clusters will be there. The formula for calculating the Davies value is:- (within cluster distance \_ i+ within cluster distance \_ j) / outer cluster distance \_ i, j. where within cluster distance \_ i and j is the distance value in a cluster and outer cluster distance \_ i, j is the distance between two center points. If the value of the DBI contains 0 it means clustering quality is best and it separates perfectly. If the value of the DBI contains a lower value it indicates clustering quality is better and if the value of the DBI contains a higher value then the clustering quality is very poor.

543 2 1

C1

C2

0

1 2 3 4 5 6 7 8 9 10

Fig: **Davies Bouldin**

Silhouette measures the similarity and dissimilarity between a cluster point and another point in the same cluster and also between a cluster point and the cluster point of another cluster. Its value occurs between +1 to -1. If Silhouette results get a high value then its similarity rate is maximum with its own cluster and dissimilarity rate is maximum with the neighbor cluster point which means the points are well grouped from the large data set. If Silhouette results get a low value then its dissimilarity rate is maximum with its cluster and the similarity rate is maximum with the neighbor cluster point which means the points are very poor grouping from the large data set. In the Silhouette algorithm, it defines the average distance from one data point to all data points within the same cluster it’s called cohesion value(a\_i) and it also defines the average distance from one data point to all neighbor cluster data points it is called separation value(b\_i). To evaluate the silhouette value the formula is:- s\_ i= (b \_i – a \_i)/ max(a \_ i, b \_i).

Fig: **Silhouette**

0

1 2 3 5 6 7 8 9 10

1 2 3 4 5