IOP Conf. Series: Materials Science and Engineering **852** (2020) 012043 doi:10.1088/1757-899X/852/1/012043

Image clustering using genetic algorithm with tournament selection and uniform crossover

Gevin Valerian*, Tri Sutrisno and Dyah Erny Herwindiati
Informatics Department, Faculty of Information Technology
Universitas Tarumanagara, Jakarta 11440, Indonesia

*gevin.53515056@stu.untar.ac.id

Abstract. This study aims to test genetic algorithm with tournament selection and uniform crossover for clustering complicated images. Genetic algorithm can be used for searching the optimum centroid for clustering images. Images that used in this study is beach images, city images, traditional market images, and garden images. Based on the research result genetic algorithm with tournament selection and uniform crossover can be used for clustering images but there is some outlier in formed cluster. Based on trial the best parameters for image clustering with genetic algorithm with tournament selection and uniform crossover is population=200, iteration=200, and number of cluster=2. Fitness value on genetic algorithm is increase when population and iteration value are higher. The result of this study can be used as a reference in the development of images clustering.

1. Introduction

This paper aims to develop application to clustering images using genetic algorithm with tournament selection and uniform crossover. clustering analysis is a method, where a few groups are so automatically established based on the metrics of correlation between samples that the samples within the same group are similar to each other while the ones from different groups are different from each other [1]. This paper is made because genetic algorithm is a simple algorithm but powerfull enough to used as a solution on complex problem like seperating images. Genetic algorithm is used to identify the best center of the cluster that formed [2].

The most common clustering algorithm is k-means algorithm which has the advantages of simple and easy to use but there are some short comings with k-means algorithm which is it might fall into local optimum [3]. Aiming at these shortcomings, this paper proposed a new clustering method based on genetic algorithm. The focus of this research is to test the genetic algorithm with tournament selecion and uniform crossover for clustering complex images. Genetic algorithm can solve the local optimum problem due to improper selection of the intial center , with multivariable global optimization[4].

The images that used in this research is images of beach, images of traditional market, images of city and images of garden. These images choosen because each images have a different characteristics, e.g., beach images have ocean surface and sand surface on it, city images have road and skycrappers on it, traditional market images have lot of vegetables and people on it and garden images have plant and footpath on it. The images will clustered based on the colors similiarity between each images and based on similiarity based on texture of each images. In order to get the texture from images researcher use

Gray Level Coocurrence Matrix (GLCM) method, and in order to get color value from images researcher use RGB color in each pixel of images. After the images input into the application , the genetic algorithm with tournament selection and uniform crossover is start it process to find the best cluster center for each category of images, when the best cluster center founded the images clustered based on the distance between the average images color and texture with the center that founded.

The output of this application is folder of formed cluster, scatter plot distribution of cluster, and graphic plot of increase of fitness value. Based on test conducted on the application of images clusering using genetic algorithm with tournament selection and uniform crossover the application can clustered images using genetic algorithm with few outlier in cluster that formed.

2. Genetic Algorithm

Genetic algorithm is a self-organizing and adaptive AI technology. In the algorithm, the initial solution is called initial population, and usually the population is generated randomly according to certain constraints. The population can be further divided into independent individuals, also known as chromosomes, which is one solution to the problem, and all individuals constitute an understanding space. Individuals are usually represented by code strings consisting of binary character sets {0, 1}, which are used to describe various parameters in practical problems. In the iterative process of algorithm, the evolution of each generation is called heredity. The genetic process consists of three operation operators, namely, selection operator, crossover operator and mutation operator. In addition, the individual's good and bad in the genetic process is evaluated by the fitness, and the fitness is calculated by the adaptive function to everyone in the population. The greater the fitness of the individual, the more the solution it represents is closer to the optimal solution, the greater the probability of being selected to the next generation. The probability of elimination is greater, which is consistent with the idea of "survival of the fittest" in evolution theory [4]. Genetic algorithm is principally composed of the following five steps initialize population, fitness value evaluation, selection, crossover, and mutation [5].

2.1. Initialize population

The first step in genetic algorithm generated an initial population. Each member of this population served as a possible solution to a problem. Each individual is evaluated and assigned a fitness value according to the fitness function. Population is generated using random initialization. Random initial population seeding technique is the simplest and the most common technique that generates the initial population to genetic algorithm. This technique is preferred when the prior information about the problem or the optimal solution is trivial [6].

2.2. Fitness evaluation

The critical thing in genetic algorithm is the choices of fitness function where it can influence the results of the population. For this paper the fitness function is based on distance between data and centroid . The distance is measure using euclidean distance. Fitness function[7].

$$F = \frac{1}{d+1} \tag{1}$$

Information:
F= Fitnesss value
d= distance

The Euclidean Distance method is used to calculate the distance between data and centroid. This measurement is based on the value of the object in each dimension in learning. Euclidean Distance can calculate the distance between data as much as two dimensions and more. Euclidean Distance formula [8]:

$$d = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2} \tag{2}$$

Information :
d= distance
X=the dimension of the first object
Y=the dimension of the second object
i= iteration
n=amount of data

2.3. Selection

Selection is a random process to select a parent's chromosome from a population based on fitness value everychromosome [9]. Selection process by a tournament is more efficient than other selection methods and leads to optimal solutions [9]. Before the selection process by tournament begins, it must define the number of participants . Than, tournament selection randomly selects candidate from the population. The winner of the tournament is the chromosome with the highest fitness value. This selection process is repeated until 2 chromosomes selected as a pair of the parent.

2.4. Crossover

Crossover is parents chromosomes combining process to get new individual. Used method is Uniform Crossover. This method defines biner zero and one. when the mask of locus has value 1, the first spring has gene compositions as same as the first parent, the second spring has gene compositions as same as the second parent. If the mask of locus has value 0, the first spring has the second parent gene compositions and the second spring has the first gene compositions[9].

2.5. Mutation

Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next gen. The mutation method that used in this paper is random mutation. In random mutation the gen that mutated defines by mutation rate and the gen that mutated is choice randomly[10].

3. Object classification

Object that used in this paper is 100 images of beach, 100 images of garden, 100 images of traditional market, 100 images of city. These images is selected because each images have a different characteristics, e.g., beach images have ocean surface and sand surface on it, city images have road and skycrappers on it, traditional market images have lot of vegetables and people on it and garden images have plant and footpath on it. Example of these images can be seen in figure 1 ,figure 2, figure 3 and figure 4.

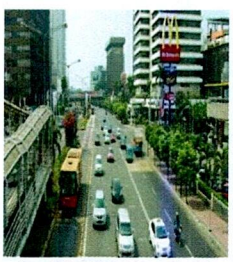

Figure 1. City image

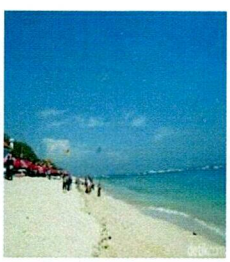

Figure 2. Beach image

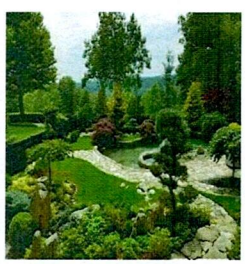

Figure 3. Garden image

Figure 4. Traditional Market images

The images color value taken using RGB of each pixel in the image.The texture value taken using GLCM feature. The images value later used to become the population in genetic algorithm

3.1. RGB images

In RGB color model, each colour appears in its primary spectral componets of red,green,and blue. The colour of a pixel is made up of three components; red, grenn, and blue(RGB), described by there corresponding intensities. Colour componenta are also known as colour channels or colour planes (components). In the RGB colour model, a colour image can be represented by the intensity function.

$$\mathrm{I\,RGB} = (\mathrm{FR}, \mathrm{FG}, \mathrm{FB}) \tag{3}$$

Where FR(x,y) is the intensity of the pixel (x,y) in the red channel, fG(x,y) is the intensitty of pixel (x,y) in the greenchannel, and fB(x,y) is the intensity of pixel (x,y) in the blue channel [11].

3.2. Gray-Level Co-Occurrence Matrix (GLCM)

The texture analysis approach has been widely developed in recent decades, and can be classified into four categories such as statistical, geometrical, model-based and signal processing. Gray Level Co-occurrence Matrix (GLCM) has been shown to be one of the most efficient approaches to texture analysis among other statistical methods[12]. Texture analysis using Gray Level Co-occurence Matrix (GLCM) was introduced by Haralick[13]. GLCM is a matrix sized NxN where N is the grayscale intensity level. Each of matrix element value located on (i,j) is calculated from the number pattern from original grayscale image that contain a specified pattern. The specified pattern contains the gray pixel ineighboring with gray paxel j for distance d and angle θ. The parameter d represents the distance of two neigboring pixel in the grayscale image and θ is the size of the discrete angle (0° , 45° , 90°, 135°). In this paper the feature of GLCM that used is contrast ,correlation,energy, and homogenenity.

Contrast is a measure of intensity or gray level variations between the reference pixel and its neighbor. Large contrast reflects large intensity differences in GLCM:

$$contrast = \sum_i \sum_j (i-j)^2 \, Pd(i,j) \tag{4}$$

Homogeneity measures how close the distribution of elements in the GLCM is to the diagonal of GLCM. As homogeneity increases, the contrast, typically, decreases:

$$homogeneity = \sum_i \sum_j \frac{1}{1+(i-j)^2} Pd(i,j) \tag{5}$$

Correlation feature shows the linear dependency of gray level values in the cooccurrence matrix:

$$correlation = \sum_i \sum_j \frac{(i-\mu x)(j-\mu y)}{\sigma x \sigma y} \, Pd(i,j) \tag{6}$$

Energy is derived from the Angular Second Moment (ASM). The ASM measures the local uniformity of the gray levels. When pixels are very similar, the ASM value will be large :

$$ASM = \sum_i \sum_j (Pd(i,j))^2 \tag{7}$$

$$energy = \sqrt{ASM} \tag{8}$$

4. Classification& Evaluation

The Classification prosses begin when the application take the input of images that want to be classified, then the application will get the color and texture value of each images, the values of each images become the population that start in genetic algorithm, then measure the distance each value using euclidean distance that later used for fitness value, then the tournament selection begin to get the best parents and the crossover between the best parrent to get the better offspring, then repeat until get the best center of cluster. The flowchart of the classification prosses can be seen in **Figure 5.**

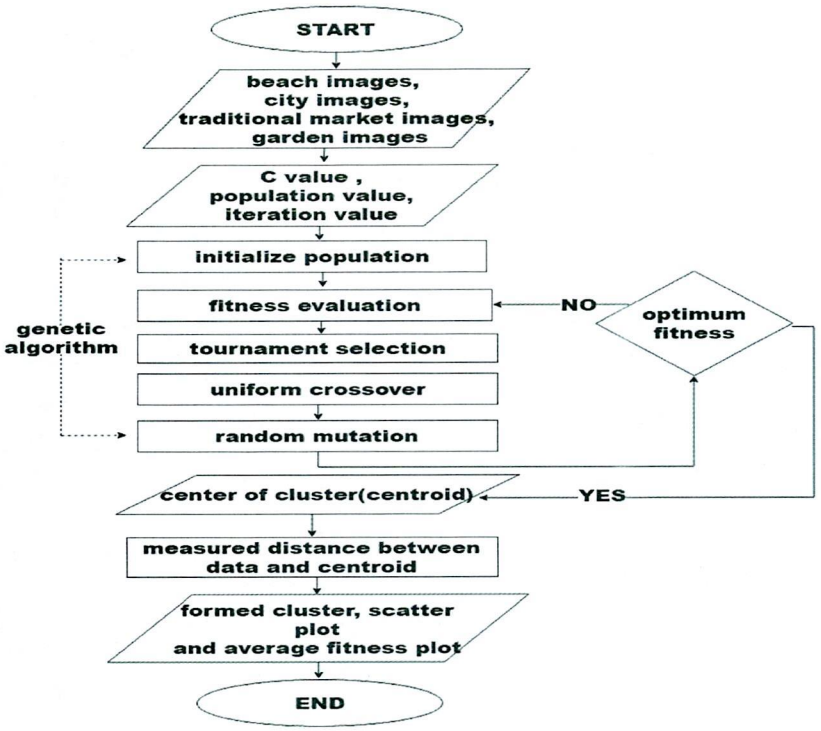

Figure 5. Flowchart process clustering

Test conducted on the application of images clustering using genetic algorithm with tournament seleciton and uniform crossover consist of testing the module and testing the data. Testing of the module aims to test wheter each module in application is running well. Testing of data is a test to find out the system runs according to the concept and to figure out the best parameters. Data testing consist of 286 test that combine between parameters that used in genetic algorithm such as combination of beach images and city images, combine with iteration value 200 , population value 200 and cluster =2. This method is used to find out the best parameters to make the best cluster as possible. Based on the test the best cluster formed is between beach images and garden images. The best parameters for image clustering using genetic algorithm with tournament selection and uniform crossover is cluster=2, population= 200 , and iteration=200. The higher the iteration and population the fitness value of centroid that formed is higher too. The best formed cluster in this test is between beach and garden images.The plot can be seen in figure 6.

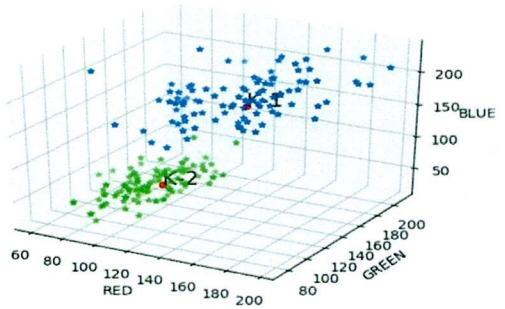

Figure 6. Plot 3D formed cluster and centroid between beach and garden

Figure 7. Cluster 1 beach-garden

Figure 8. Cluster 2 beach -garden

We are able to see that the images between beach and garden can be clustered in figure 7 & figure 8. The formed cluster have a high precision of 98%, altought there are some outlier that can be found in the formed cluster. In the formed cluster member of cluster-1 is 98 images of beach and 2 images of garden, member of cluster-2 is 98 images of garden and 2 images of beach. The images that become oulier placed in wrong cluster because the characteristic of the images more similar with the center of the cluster that images placed than the center of the cluster that the images should. With genetic algorithm the centroid that selected can be more optimum in each iteration because in genetic algorithm the best centroid will selected and combined in each process of iteration .Based on this test we can say that genetic algorithm can be use as alternative way to cluster a complex images altought there is some outlier in the formed cluster .

5. References

[1] John McCall, 2005, *Journal of Comutational and Applied Mathematics.* **205**-22
[2] Lewis R, Krawiec M, Confer E, et al. 2014 Data clustering: 50 years beyond K-means, Pattern Recognition Letters, Vol. 31, No. 8, pp.651-666
[3] Rahman MA, Islam MZ 2014 A hybrid clustering technique combining a novel genetic algorithm with KMeans, Knowledge-Based Systems, Vol. 71, No. 71, pp. 345-365
[4] EH Ruspini, 1970, *Numerical methods for fuzzy clustering.* Information Sciences, 2 (3):319-350.
[5] Konar, Amit. 2005. Computational Intelligence Principles, Techniques, and Applications. Springer: Calcutta, India
[6] Gong J, Haibin M, Yong D, et al. 2005 Redundant elimination of intrusion events with multi feature association, Journal of Southeast University (Natural Science), Vol.35, No. 3, pp. 366-371
[7] Lichman, M. 2013. UCI Machine Learning Repository
[8] Rabunal, J.R. & Dorado, J. 2006. Artificial Neural Networks in Real-Life Applications, Ideal Group Publishing: Hershey, United States of America
[9] Bai, L., J. Liang, & C. Dang. 2011. An Inititalization Method to Simultaneously Find Initial Cluster Centers and The Number of Clusters for Clustering Categorical Data. KnowledgeBased Systems24: pp. 785–795
[10] J P Dias and H S Ferreira Automating the Extraction of Static Content and Dynamic Behaviour from eCommerce Websites 297–304 ANT 2017 Procedia Computer Science 109C
[11] Miyazaki D, Kagesawa M and Ikeuchi K 2004 Transparent surface modeling from a pair of polarization images IEEE Trans. Pattern Anal. Mach. Intell. 26 73–82.
[12] Haralick, R. M., Shanmugam, K., & Dinstein, I. H. 1973. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on, (6), 610-621
[13] Hu, Y., Zhao, C. X., & Wang, H. N. (2008, December). Directional analysis of texture images using gray level co-occurrence matrix. In Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on (Vol. 2, pp. 277-281). IEEE..