Clustering Cells Shape Descriptors Using K-Means vs. Genetic Algorithm

Faten Abushmmala[1], Mohammed Alhanjouri[1*]

[1] Computer Department, Islamic University- Gaza, Gaza City, Gaza Strip, Palestine

Email Address

faten.fffa@gmail.com (Eng.Faten Abushmmala), malhanjouri@iugaza.edu.ps (Prof.Mohammed Alhanjouri)
*Correspondence: malhanjouri@iugaza.edu.ps

Received: 28 November 2017**; Accepted:** 27 December 2017

Abstract:

This paper interested in clustering red blood cells, these cells are in form of digital images of blood films, a comparison made between Genetic Algorithm (GA) and K-Means behavior/performance in clustering. The data set consists of shape descriptors of the cells shapes, the original number of samples are 100 samples. Each sample provided us with at least 10 cells (shape) with total number of 409 shapes (cells). The Genetic Algorithm shows better performance than K-Means in clustering these cells into two clusters (Normal and Abnormal) with success rate 99.48% where K-Means gave 83.16%. While K-Means shows a better performance in clustering the cells into four clusters (Burr, sickle, teardrop and normal cells) than GA where K-Means gave 86.74% and Genetic algorithm (GA) gave 83.2 %.

Keywords:

Cells Shape Descriptors, Cells Shapes, Clustering, K-Means, Genetic Algorithm

1. Introduction

Clustering is an important form of data mining. It can be used to extract useful and hidden information from different types and shapes of datasets [1].Clustering is considered unsupervised classifying technique known in Artificial Intelligence and pattern recognition fields, and considered extremely helpful in dealing with big data. Unsupervised learning requires no pre learning step before classifying. The Art of clustering/classifying images stems in finding the perfect features that represents these images/shapes uniquely; here in this paper we uses many shape descriptors, ten to be exact, discussed in the methodology section.

In general, shape descriptor is a set of numbers that are produced to represent a given shape feature [2]. A descriptor attempts to quantify the shape in ways that agree with human intuition (or task-specific requirements). Good retrieval accuracy requires a shape descriptor to be able to effectively find perceptually similar shapes from a database. Usually, the descriptors are in the form of a vector.

Many original literature employee shape descriptors to form more effective set of features such as in paper [3]. Shape features have been studied for several decades for their capability to describe the shape of an object, and are widely used in areas such as content-based image retrieval [4]; computer graphics [5]; and image registration [6].Images have different features associated to them. Shape is an important visual feature in an image. It plays a vital role in Content Based Image Retrieval (CBIR) [7]. Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is important application of computer vision used mainly in the image retrieval in big data sets. The advantages and disadvantages of these shape descriptors are also discussed in paper [7].

In this paper two algorithms were used to cluster these shapes (using their shape descriptors), K-Means and Genetic algorithm. First we will discuss the mechanism of K-Means algorithm:

The K-Means algorithm, first developed four decades ago [8], is one of the most popular center-based algorithms that attempts to find K clusters which minimize the mean squared quantization error, MSQE. The algorithm tries to locate K prototypes (centroids) throughout a data set in such a way that the K prototypes in some way best represent the data. A summarization of the K-Means algorithm through the following steps [9]:

1. Initialization
a) Define the number of prototypes (K).
b) Designate a prototype (a vector quantity that is of the same dimensionality as the data) for each cluster.
2. Assign each data point to the closest prototype. That data point is now a member of the class identified by that prototype.
3. Calculate the new position for each prototype (by calculating the mean of all the members of that class).
4. Observe the new prototypes' positions. If these values have not significantly changed over a certain number of iterations, exit the algorithm. If they have, go back to step 2.

The main problem of the K-Means algorithm [9] is its dependency on the prototypes' initialization. If the initial prototypes are not chosen carefully the computation will run the chance of converging to a local minimum rather than the global minimum solution. Thus initializing prototypes appropriately can have a big effect on K-Means. The performance function for K-Means may be written as

$$J_{km} = \sum_{i=1}^{N} {}_{j=1}^{k}\min \left\| x_i \ \ m_j \right\|^2 \tag{1}$$

The second algorithm discussed here is Genetic algorithm; which considered more smarter algorithm that require more parameters in order to suit the application in opposite of K-Means which consider blind clustering technique.

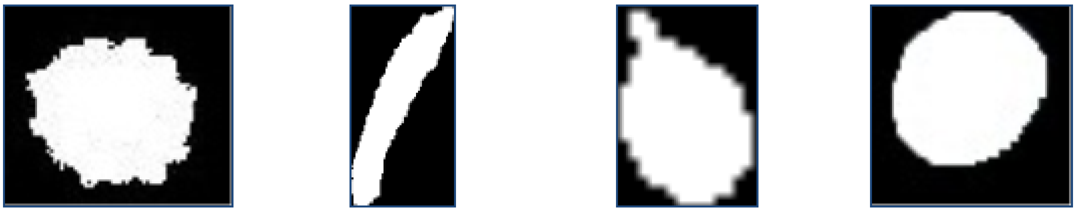
Genetic algorithms are a stochastic search algorithm, which uses probability to guide the search. It was first suggested by John Holland [10]; in the seventies. Over the last thirty years, it has been used massively in many application areas, such as image processing, pattern recognition, feature selection, and machine learning [11].It is a powerful search mechanism that emulates natural selection and genetic operators. Its power comes from its ability to attach good pieces from different solutions and combine them into a single strong solution [11].Genetic algorithms are initial population of solution called individuals is (randomly) generated, the solutions are evaluated. The algorithm creates new generations of population by genetic operations, such as reproduction, crossover and mutation. The next generation consists of the possible survivors (i.e. the best individuals of the previous generation) and of the new individuals obtained from the previous population by the genetic operations. The best source of information about GA is Holland's adaptation in natural and artificial systems; Holland uses terms borrowed from mendelian genetics to describe the process: each position in the string is called a gene. The possible values of each gene are called alleles. A particular string is called a genotype. The population of strings also called the gene pool. The organism or behavior pattern specified by a genotype is called a phenotype. If the organism represented is a function with one or more inputs [12] these inputs are called detectors. The algorithm (pseudo code) of simple GAs next [13, 14]. The pseudo code illustrates the main steps that should be performed to produce the required solution. Many literature discuss improving Genetic algorithm by manipulation it to increase the efficiency, here is a literature to improve the effectiveness of crossover and mutation [15-17]. Some other literature discusses improving Genetic Algorithm through its Initial Population such as [18]. The algorithm of simple GA are:

```
1.  Initialization [population];
2.  Evaluation [population];
Generation: =0;
-do
Selected-parents: = selection [population];
Created-offspring: =recombination [selected-parents];
Mutation [created-offspring];
Population: =created-offspring;
3.  Evaluation [population];
Generation: = generation+1;
4.  UNTIL stop-criterion;
```

2. Methodlogy

First, preprocessing steps needed to prepare the images for the feature extraction phase, the preprocessing contain cleaning, de-noising and segmentation of the cells. Check the following papers for more details [19 - 21]. These papers suggested many methodologies to accomplish the needed task. As a result we need to get binary images of the cells shapes alone, isolated from the background and away from the images boundary.

After the preprocessing phase, the result obtained would be binary images of cells shapes, all images must be in one size.



(a) Burr *(b) Sickle* *(c) Teardrop* *(d) Normal cells*

Figure 1 (a), (b), (c) Abnormal cells where (d) Normal

***Figure1.** Blood Cells.*

Figure 2. Shows the obtained results after the preprocessing phase completed, next the feature extraction phase. The features will be shape descriptors, the definitions of the used shape descriptors are:

1- Solidity: Scalar specifying the proportion of the pixels in the convex hull that are also in the region. Computed as

$$\frac{Area}{ConvexArea} \tag{2}$$

2- Eccentricity: Scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases; an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.)

Eccentricity's the measure of aspect ratio. It is the ratio of the length of major axis to the length of minor axis. It can be calculated by principal axes method or minimum bounding rectangle method.

3- Circularity ratio represents how much a shape is similar to a circle. There are 3 definitions:

Circularity ratio is the ratio of the area of a shape to the area of a circle having the same perimeter.

$$\mathrm{C1}=\frac{As}{Ac} \tag{3}$$

Where *As* is the area of the shape and Ac is the area of the circle having the same perimeter as the shape. Assume the perimeter is O, so *As*=O^2/4π. Then

$$\mathrm{C1}=4\pi.\mathrm{As}=O^2 \tag{4}$$

As 4π is a constant, we have the second circularity ratio definition.

Circularity ratio is the ratio of the area of a shape (A) to the shapes perimeter (O) square:

$$\mathrm{C2}=\frac{As}{O^2} \tag{5}$$

Circularity ratio is also called circle variance and defined as:

$$\mathrm{C3}=\frac{\delta R}{\mu R} \tag{6}$$

Where μ_r and σ_r are the mean and standard deviation of the radial distance from the centroid (g_x ,g_y) of the shape to the boundary points (x_i ,y_i), i € [0, N-1]. They are the following:

$$\mu_r=\frac{1}{N}\sum_{i=1}^{N-i} di \qquad \text{and} \qquad \delta R=\sqrt{\frac{1}{N}\sum_{i=1}^{N-i}(di-\mu R)^2} \tag{7}$$

$$\text{Where} \quad di=\sqrt{(x_i-g_x)^2+(y_i-g_y)^2} \tag{8}$$

4- Rectangularity: the Area of the shape divided by the area of the minimum bounding box Rectangularity represents how rectangular a shape is, i.e how much it fills its minimum bounding rectangle:

$$\mathrm{R}=\frac{As}{AR} \tag{9}$$

Where *As* is the shape area, *AR* is the area of the minimum bounding rectangle.

5- Convexity: is defined as the ratio of perimeters of the convex hull 当前无法显示此图像。 over that of the original contour O:

$$\text{Convexity} = \frac{O_{ConvexHull}}{O} \tag{10}$$

The Region R^2 is convex if and only if for any two points P_1 ,P_2 €R^2 , the entire line segment P_1 ,P_2 is inside the region. The convex hull of a region is the smallest convex region including it.

6- Ellipse Ratio: with knowing the next two concepts:

A- Major Axis Length: Scalar specifying the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region.

B- Minor Axis Length: Scalar; the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region.

Calculating the area of an ellipse knowing the major axis length and the minor axis we get Ae.

Ellipse Ratio is:

$$\frac{As}{Ae} \tag{11}$$

Where As is the shape area and Ae ellipse area with the same minor axis and major axis of the shape.

7. Extent: Scalar that specifies the ratio of pixels in the region to pixels in the total bounding box. Computed as the Area of the shape divided by the area of the bounding box.

8. Ecc: first let's understand the following definition:

Longest Length: is a variable we made to measure the distance between farthest two points at the boundary of a shape.

So,

$$\text{Ecc} = \frac{Minor\ Axis\ Length}{Longest\ Length} \tag{12}$$

First we calculate distance of the farthest point on the boundary and the closest point on the boundary in relative to the centroid.

D1= the longest distance (Centroid to boundary).

D2= the shortest distance (Centroid to boundary).

9. R2 is the ratio of longest distance of the boundary to the centroid to the major axis length of the shape

$$\text{R1}=\frac{D_1}{Major\ Axis\ Length} \tag{13}$$

10. R2 is the ratio of shortest distance of the boundary to the centroid to the major axis length of the shape

$$\text{R2}=\frac{D_2}{Major\ Axis\ Length} \tag{14}$$

These descriptors are in the features matrix set in follwoing order:

Features (Shape Descriptor (SD)) = [Solidity, E, Circle Ratio, Rectangularity, Convexity, Ellipse Ratio, Extent, Ecc, R1, R2];

The clustering phase (final phase) is about giving these features to the both algorithms, where the algorithms will cluster the data first into two clusters, normal and abnormal cells. And secondly into four clusters (Burr, sickle, teardrop and normal cells) and comparison made. As mentioned before K-Means need no parameters to accomplish this mission while Genetic algorithm need more work. K-Means must runs three times to avoid local minima.

To use genetic algorithm, we must represent a solution to our problem as a genome (or chromosome). The genetic algorithm creates a population of solutions and applies genetic operators such as mutation and crossover to evolve solutions in order to find the best (optimum) solutions. The most important aspects of using genetic algorithms are: (1) definition of the objective function, (2) definition and implementation of the genetic representation, and (3) definition and implementation of the genetic operators. The Genetic algorithm used here in unsupervised learning (Clustering) of the data. Centroid of the cluster i is (x_i, y_i), Centroid of the cluster j is (x_j, y_j).The objective function (18):

$$Sim_i = \left(\frac{1}{|C_i|}\sum_{x\in c_i}\left\|(x_j,y_j)-(x_i,y_i)\right\|_2\right) \tag{15}$$

$$Re_i = Max\left\{\frac{Sim_i+Sim_j}{d_{ij}}\right\}_{j,i\neq i} \tag{16}$$

Where $$d_{ij} = d(C_i,C_j) = \left\|(x_j,y_j)-(x_i,y_i)\right\|_2 \tag{17}$$

$$D = \frac{1}{K_r}\sum_{i=1}^{K_r} Re_i \tag{18}$$

$$\text{Fitness (x)} = \frac{1}{D} \tag{19}$$

Best results get by maximize the difference between clusters and minimize the difference between elements among one cluster and this is accomplished by minimizing the Re and maximizing the Sim meaning maximizing the Fitness function (18).

Genetic Algorithm parameters chosen are:

1- Initialization of population is random.
2- Population size is 20.
3- Maximum number of generation is 100.
4- Maximum time limit is infinity (no limit).
5- Crossover fraction percentage is 80%.
6- Mutation Function Fraction is: 0.2.
7- Crossover Function: scattered.
8- Migration Direction: Forward.
9- Function tolerance: 1e-6.
10- Selection function is: stochastic.
11- Scaling function is: Rank.

3. Results and Discussion

The features where taken from abnormal cells (Burr, Sickle and teardrop) and of normal cells, this table (Table 1) describes these features:

Table 1. *Data Set Description for Shape Descriptor (SD) features for each type of cells.*

Type of The Cell	Max.Value	Mini. Value	Features Length	# of Cells
Burr (Abnormal)	0.9912	0.1754	12	114
Sickle (Abnormal)	1	0	12	118
Tear drop (Abnormal)	1	0.1485	12	59
Normal	1	0.1630	12	128

This data set of features used here to accomplish the clustering phase, where fed to the two clustering algorithms, K-Means and GA.

When cluster these data set into two clusters (normal and abnormal) the results for K-Means were 83.16% success rate while Genetic algorithm 99.48%.

Clustering these data into four clusters (Burr, sickle, teardrop and normal cells) K-Means gave 86.74% success rate while Genetic algorithm gave 83.2%.
An attempt done to enhance the Genetic algorithm performance in clustering the data into four clusters is by making the initial population not random as before but taken out from the K-Means after running it, the result was 89.2% which consider higher than both algorithms performance. For the same attempt when changing population size from 20 to 200 (20, 50, 100 and 200) the results do not change significantly, for changing the selection function the Roulette and stochastic both of them gave the best results which was (89.2%). Check the table (Table.2) below for more.

Table 2. *Scaling function manipulations.*

Result	Selection Function
89.20%	Roulette
89.20%	Stochastic
88.00%	Reminder
89.10%	Uniform
88.60%	Tournament

4. Conclusions

In this paper red blood cells shapes are extracted and converted to binary images through a series of four phases, each phase contain several procedures. These images taken from digital images of blood films under microscope. The literature that discuss these procedures are in [19-21]. Most popular red blood cells shapes are burr, sickle and teardrop for abnormal cells and circular for normal cells. Many published papers discuss segmenting red blood cells and locating these cells [19-23], while identifying the shapes of these cells such as our paper is considered rare to none.

After the feature extraction phase shape descriptors set taken from these cells as discussed. Then these descriptors where clustered using K-Means and Genetic Algorithm (GA).

For clustering the shape descriptors of the cells into two clusters (normal and abnormal) the K-Means alone gave 83.16% success rate, but the GA gave higher success rate which was 99.48%. Genetic Algorithm shows more abilities in clustering the data sets in hand in appose to K-Means. Shape descriptors were extremely informative in representing the shapes in clustering these cells into normal and abnormal cells.

In trying of clustering the data sets into four clusters (Burr, sickle, teardrop and normal cells) K-Means gave 86.74% but GA gave 83.2%. Where K-Means shows better ability in distinguishing and clustering the shapes a part than GA. To enhance GA performance we take the initial population from what came out of the K-Means the results was 89.2%, this result considered the highest of both K-Means and GA results in clustering these descriptors into four clusters.

Some of the shape descriptors (SD) discussed in this paper are known (widely used) and others created to fit our application specifications (these SD are Ecc, R1 and R2), the SD was extremely useful to this paper, it's also was and still useful for image retrievals in big data. That's why SD is important subject to be further studied and developed.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] H. Parmar; B. Limbasiya. A Review on Genetic Algorithm-based Text Clustering Technique. *International Journal of Advance Research in Computer Science and Management Studies,* 2015, 3(2), 81-85.

[2] Peng-Yeng Yin. Pattern Recognition Techniques, Technology and Application. Veinna, Austria, 2008, 626.

[3] Liu Z.; Zhao C.; Wu X.; Chen W. An Effective 3D Shape Descriptor for Object Recognition with RGB-D Sensors. *Sensors (Basel),* 2017, 17(3), 451.

[4] M. Eitz; R. Richter; T. Boubekeur; K. Hildebrand; M. Alexa. Sketch-based shape retrieval. *ACM Trans. Graph.* 2012, 31(4).

[5] M.E. Yumer; S. Chaudhuri; J.K. Hodgins; L.B. Kara. Semantic shape editing using deformation handles. *ACM Trans. Graph.* 2015, 34(4), 86.

[6] M. As'ari; U.U. Sheikh; E. Supriyanto. 3D shape descriptor for object recognition based on kinect-like depth image. *Image and Vision Computing,* 2014, 32(4), 260-269.

[7] P. D'Silva1; P. Bhuvaneswari. Various Shape Descriptors in Image Processing – A Review. *International Journal of Science and Research (IJSR),* 2015, 4(3), 2338-2342.

[8] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* 1967, 1(14), 281-297.

[9] W. Barbakh; Ying Wu; Colin Fyfe. Non-standard parameter adaptation for exploratory data analysis, University of the west of Scotland, Scotland, 2009. ISBN: 978-3-642-04004-7.

[10] J.H. Holland. Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, 1975, 15.

[11] P. Kohn. Combing Genetic Algorithm and Neural Networks. M.Sc. thesis, Tennessee Unive, 1994.

[12] Sh.A. Rasheed. Genetic Algorithms Application in Pattern Recognition. M.Sc. thesis, National Computer Center Higher Education Institute, 2000.

[13] R.L. Wainwright. Introduction to Genetic Algorithms Theory and Applications. Addison-Wesley, 1993.

[14] M. Schamidt; T. Stidsen. Hybrid Systems: Genetic Algorithms, Neural Network and Fuzzy logic. Univ. Aarhus. Denmark, 1997.

[15] N.K. Pareek; V. Patidar. Medical image protection using genetic algorithm operations. *Soft Computing*, 2016, 20(2), 763-772.

[16] F. Liu; G. Zeng. Study of genetic algorithm with reinforcement learning to solve the TSP. *Expert Systems with Applications*, 2009, 36(3), 6995-7001.

[17] S.M. Elsayed; R.A. Sarker; D.L. Essam. A new genetic algorithm for solving optimization problems. *Engineering Applications of Artificial Intelligence*, 2014, 27, 57-69.

[18] Y. Deng; Y. Liu; D. Zhou. An Improved Genetic Algorithm with Initial Population Strategy for Symmetric TSP. *Mathematical Problems in Engineering*, 2015, 3, 1-6.

[19] F. Abushmmala; F. Abushmmala. Processing Overlapped Cells Using K-Means and Watershed. *International Journal of Intelligent Information Systems*, 2014, 3(1), 8-12.

[20] F. Abushmmala; M. Alhanjouri. Colour Based Segmentation of Red Blood Cells using K-means and Image Morphological Operations. *Journal of Advanced and Innovative Research*, 2013, 2(11), 344-350.

[21] F. Abushmmala; W. Barbakh. Color Based Segementation using different versions of K-means in two Spaces. *Global Advanced Research Journal of Engineering, Technology and Innovation*, 2013, 1(9), 030-041.

[22] G. Karkavitsas; M. Rangoussi. Object Localization in medical images using genetic algorithm. *World academy of Science, Engineering and Technology*, 2007, 1(2), 72-75.

[23] P.J.H. Bronkorsta; M.J.T. Reinders; E.A. Hendriks; J. Grimbergen; R.M. Heethaar; G.J. Brakenhoff. On-line detection of red blood cell shape using deformable Templates. *Pattern Recognition Letters*, 2000, 21(5), 413-424.