British Journal of Mathematics & Computer Science

British Journal of Mathematics & Computer Science

22(6): 1-13, 2017; Article no.BJMCS.33729

ISSN: 2231-0851

Evaluation of Calinski-Harabasz Criterion as Fitness Measure for Genetic Algorithm Based Segmentation of Cervical Cell Nuclei

Caglar Cengizler[1*] **and M. Kerem Un**[1]

[1] *Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Cukurova University, Saricam, Adana 01330, Turkey.*

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2017/33729
Editor(s):
(1) Morteza Seddighin, Indiana University East Richmond, US.
Reviewers:
(1) S. Muthukumar, Indian Institute of Information Technology, Tamilnadu, India.
(2) Diana Bilkova, University of Economics, Prague.
Complete Peer review History: http://www.sciencedomain.org/review-history/19643

Received: 27th April 2017
Accepted: 11th June 2017
Published: 21st June 2017

Original Research Article

Abstract

In this paper, the classification capability of Calinski-Harabasz criterion as an internal cluster validation measure has been evaluated for clustering-based region discrimination on cervical cells. In this approach, subregions in the sample image are initially randomly constructed to be the individuals of the population. At each generation, individuals are evaluated according to their Accordingly a novel genetic structure for meta heuristic area isolation is proposed. Evaluation of proposed combination of genetic algorithm and Calinski-Harabasz measure is achieved by experiments, conducted on real cervical cell samples. We have used two separate cluster validity measures to evaluate the performance of the clustering approach. Jaccard index and F-score are utilized for objective comparison. Results shows that, Calinski-Harabasz criteria may have a better performance with proposed novel genetic structure and presented mechanism may have great potential on discrimination of specific regions.

Keywords: Cervical; nuclei; segmentation; evolutionary; Calinski-Harabasz; Davies-Bouldin.

*Corresponding author: E-mail: ccengizler@cu.edu.tr;

1 Introduction

Evolutionary computing is an unsupervised machine learning approach where system perform search for better solution to a complex problem [1]. Evolutionary computing is used in a wide variety of biomedical applications where automated characterization and classification of biological data is to be achieved by a meta-heuristic approach [2]. This approach allows us to mimic nature for automated solution search without any prior training stage [3][4]. In the conventional evolutionary approach, genetic operators are applied to modify the parameters ,the so-called *chromosomes*, of potential solution candidates in order to converge the "best" solution [5][6]. The fitness, a mathematical score indicating the success of the solution, is the key to carry the best individuals to the next generation [7][8]. For the overall success of the method, individuals that are strong solution candidates should be evolved towards increased fitness and passed to the consequent generations [9].

Meta-heuristic segmentation of regions of interest on microscopic images is also possible with genetic algorithms. In this work, the isolation of the nucleus from a cell image is formulated as a data clustering problem. To form the first generation, image subregions are created randomly to form the individual of the population. The pixels of nuclei and non-nuclei regions are grouped with respect to the features characterizing the subregion. The generations are let to evolve to eventually obtain an optimal subregion that overlaps the nucleus. The Calinski-Harabasz measure is taken as the fitness value of an individual, which, to the best of our knowledge, has not been tried in the evolutionary segmentation of cervical cells.

The proposed approach has been tested on real cervical cell images where manually segmented cytoplasmic area and nuclei regions are accepted as the ground truth. The Calinski-Harabasz measure is compared with the Davies-Bouldin measure by evaluating their F-score and Jaccard indices.

2 Structure of Individuals

In our approach, each individual of a population is a subregion defined on the image. Accordingly, random subregion centers are created first in the beginning of the algorithm. The subregion boundary is formed with a closed-contour polygon defined around each center (See Sections 2.1 and 2.2 for details). Each individual is assumed capture the local characteristics of the image (Fig. 1).

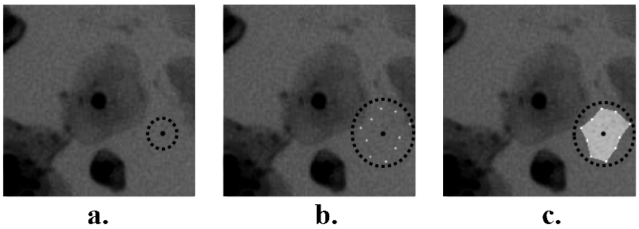


Fig. 1. Elements of an individual are visualised inside dotted circles
a) Center of the individual is indicated by black dot
b) Particles around the center
c) Closed contour shape formed by particles

There are four factors affecting the location and the shape of an individual subregion.

- **Factor 1** Center of the individual

- **Factor 2** Radius of the vertex distribution around the center
- **Factor 3** Number of vertices around the center
- **Factor 4** Elasticity of the vertices

These factors are embedded inside the genetic structure of the individual as explained in detail below.

2.1 Center of an individual

The center of a subregion, i.e. an individual in the population, is a point $C(x, y)$ in the two-dimensional space of the image. When creating the individual, the coordinates x and y are selected randomly between an upper and a lower bound value (indicated by the subscripts $_u$ and $_l$). Let$R$ indicate the random selection operator than:

$$x = R([x_l, x_u]) \tag{2.1}$$

$$y = R([y_l, y_u]) \tag{2.2}$$

Considering a population, x and y variables are uniformly distributed within the $[x_l, x_u]$ and $[y_l, y_u]$. Note that the upper and lower bounds could be modified during the iteration procedure.

2.2 Radius of the particle distribution

A radius value "ρ" defines how far a vertex of an individual can radiate from the center. This value directly affects the shape and the total area of an individual.In the beginning,ρ for each vertex is chosen as:

$$\rho = R([r_l, r_u]) \tag{2.3}$$

where r_l and r_u are the lower and upper bounds for the radius value, initially set to 5 and a value that equals[Window Size$*$0.9], respectively. It is possible to update these bounds to different values during iterative process to increase adaptation capability of the population. Effect of radius on the shape of an individual is illustrated in Figure 2.

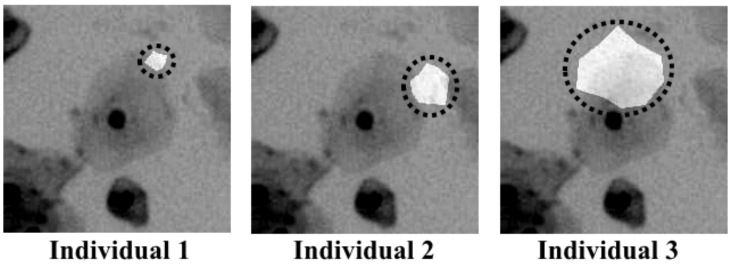


Fig. 2. Three individuals of different size and shape complexity (indicated with black dots)

2.3 Number of particles around the center

$N_p$vertices have been created for each individual around its center $C(x, y)$. The radial distribution of vertices around the center is uniform while their euclidean distances to the center vary according to Eq.2.3 and may be updated dynamically during iterative process. It is possible to assert that, particles around the center are sampling points for defining two dimensional complex geometries. It should be noted that, increasing number of vertices would generate more complex individuals with higher segmentation capabilities. Effect of vertexnumber on the shape of the individual is demonstrated in Fig. 3.

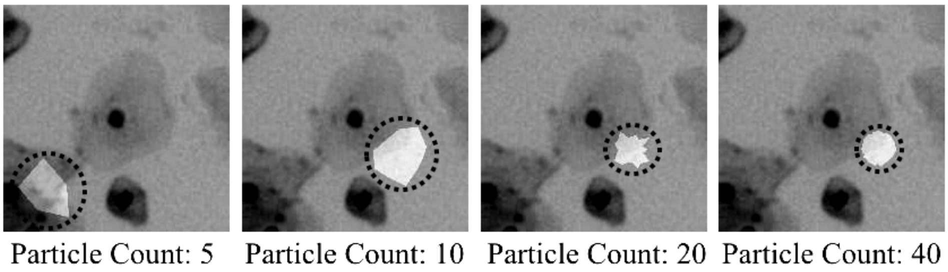


Fig. 3. Four individuals with different number of vertices(indicated with black dots

2.4 Elasticity of the particles

Elasticity, η defines the variation of the distance of particles to the center. Elasticity is a predefined coefficient between 0 and 1. Distance of each particle to the center is determined randomly in an elasticity band. However it is also possible to manipulate distance by an interaction.

Distance of any $particle_n$ is determined by:

$$\text{Distance}_n = R([\rho - (\frac{\rho * \eta}{100}), \rho]) \tag{2.4}$$

It should be noted that, η value is also embedded inside the gene. Effect of η to the shape coverage of any individual is shown in Fig. 4.

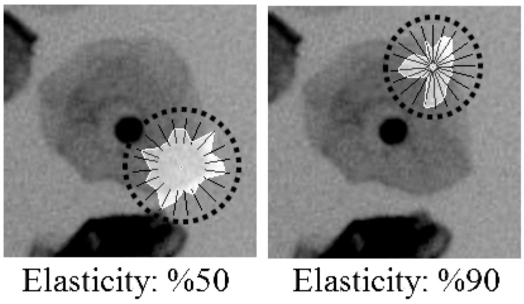
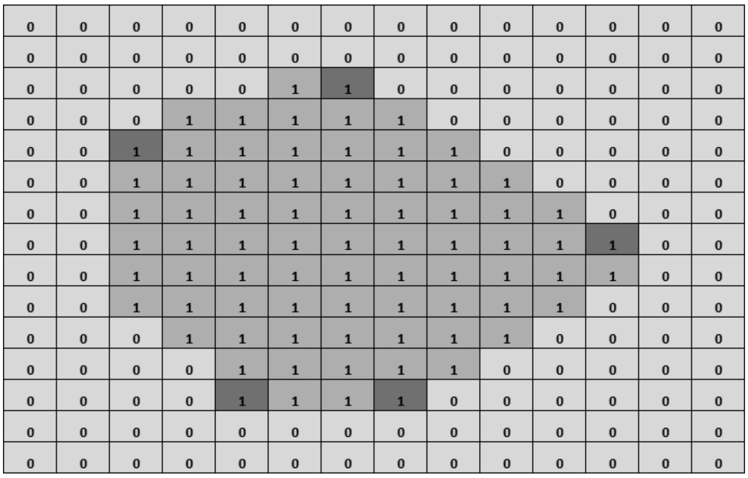


Fig. 4. Two individuals with η value of %50 and %90 are shown in dotted circles respectively. Black lines indicates the elasticity band where particles are located randomly

3 Pixel Clustering Approach

Each individual in a population is considered as a solution to a clustering problem. In that sense, a class (or cluster) is simply a set of pixels showing similar characteristics according to their features [10][11][12][13][14]. An individual discriminates all pixels lying within its contour as nuclei class pixels and those remaining outside as belonging to the non-nuclei class (Fig. 5).



0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	0	0	0	0
0	0	1	1	1	1	1	1	1	1	1	0	0	0
0	0	1	1	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	1	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 5. Pixel discrimination of an individual where pixels in the darker colored area belongs to the nuclei class and pixels in the lighter colored area belongs to the non-nuclei class

According to this approach, N classification objects (pixels in our case) are to be grouped into$K$ clusters (with nuclei and non-nuclei pixels, K=2 in our case). Accordingly, in an image of size $[X_w \times Y_w]$, $N = X_w \times Y_w$ and each observation in a feature space of, say, three features$F(N, 3)$ represents a pixel and is defined as $X = \{x_i, i = 1, ..., N\}$. Also, clusters expressed as $\{C_k, k = 1, ..., K\}$.

Discrimination result of an individual is visualised on feature space in Fig. 6.

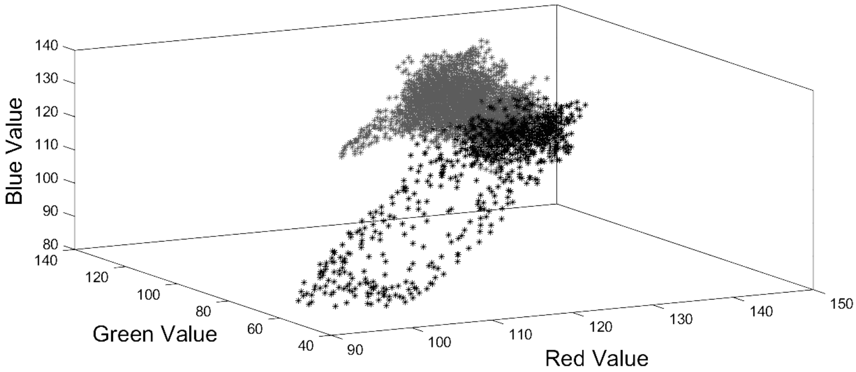


Fig. 6. Discrimination result of an individual is shown on a three axis feature space where there are two classes formed by individual

Association of each pixel to a cluster is unique in this study, meaning that there is no degree of belonging as in fuzzy clustering approaches. Association $w_{k,i}$ may be expressed as

$$w_{k,i} = \begin{cases} 1, \text{ if pattern } x_i \in \text{cluster } C_k \\ 0, \text{ Otherwise} \end{cases} \tag{3.1}$$

Additionally, number of pixels belonging to a cluster Ck is denoted by:

$$|C_k| = \sum_{i=1}^{N} w_{k,i} \tag{3.2}$$

3.1 Extracted features

In this work, three features are chosen to characterize a pixel. These are mean gray, mean gradient values and entropy. It should be noted that these values are averagedover a fixed window size of2x2 in our study.

Mean gray value is given by

$$Gray_{mean} = \frac{1}{K * M} \sum_{x=1}^{K} \sum_{y=1}^{M} I(x, y) \tag{3.3}$$

where I(x,y) is the intensity value of grayscale image. This quantity would vary significantly between nuclei and non-nuclei regions.As mentioned before, K and M are 2 in our case.

With the intensity gradient at a pixel, it is possible to locate regions of sudden change in the image which could be important in identifying the nuclei boundaries[15]. The magnitude of the gradient is defined as:

$$|\triangledown G| = \sqrt{(\frac{\partial I}{\partial x})^2 - (\frac{\partial I}{\partial y})^2} \tag{3.4}$$

where

$$\frac{\partial I(x,y)}{\partial x} = \frac{I(x+1,y) - I(x-1,y)}{2} \tag{3.5}$$

$$\frac{\partial I(x,y)}{\partial y} = \frac{I(x,y+1) - I(x,y-1)}{2} \tag{3.6}$$

The mean value of the gradient magnitude is given as:

$$|\triangledown G|_{mean} = \frac{1}{K * M} \sum_{x=1}^{K} \sum_{y=1}^{M} |\triangledown G| (x, y) \tag{3.7}$$

Entropy, defining the randomness in a given frame, is the third feature extracted. It would capture textural changesin the image,which may be important for locating the nuclei contours. Average entropy is defined as:

$$Entropy = -\sum p * log_2(p) \tag{3.8}$$

p value indicates the number of histogram counts of given region above.

4 Fitness of an Individual

It is planned to utilize genetic operators for evolving generations to locate the best solution. Therefore a fitness criteria is necessary to determine best fitting individuals in a generated population. Hence, a quality measure for formed clusters should be chosen as fitness function.

In this study Calinski-Harabasz (CH) index is utilized as an internal cluster validity measure which grades clusters created by each individual. It is described by:

$$CH(k) = \frac{B_c(k)}{(k-1)} / \frac{W_c(k)}{(n-1)} \tag{4.1}$$

where n stands for number of the clusters and k stands for class. B_c and W_c denotes between and within cluster sums of squares respectively, given by:

$$B_c = \sum_{k=1}^{K} |C_k| \left\| \overline{C_k} - \overline{x} \right\|^2 \tag{4.2}$$

$$W_c = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{k,i} \left\| x_i - \overline{C_k} \right\|^2 \tag{4.3}$$

Given criteria in equation (4.1) judges each possible cluster solution by it's quality which is dependent on how large inter-cluster distances and proximity of intra-cluster distances [16].

5 Applied Genetic Operators

A geometrical genetic structure is proposed in the study. It was aimed the observe the capability and compatibility of given fitness criteria with proposed geometrical structure. Additionally genetic operators are applied with an iterative algorithm to populations. Which allowed us to observe and compare efficiency of fitness criteria while populations are evolving. Proposed proof of concept evolutionary mechanism includes three basic genetic operators which are Selection, crossover and mutations.

Selection operator is utilized for eliminating weakest genes while passing best ones to next generations [17]. A certain amount of individuals with low fitness score are killed with each iterations. Following, crossover operator is utilized for reproducing new populations. In the study crossover operator is applied to best living individuals after elimination of weakest individuals. Than next generation is populated inside the region formed by centres of selected individuals. Selection is performed on sorted remaining individuals according to their fitness where better fitting means higher chance of selection [18][19]. Finally mutation operator is applied for expanding search capability and rich diversity. Mutation operator is applied to randomly selected individuals. It changes a single random gene of selected individual which means re-calculation of position of the selected point in out case. Basic flow of the experimental mechanism is shown in Fig. 7.

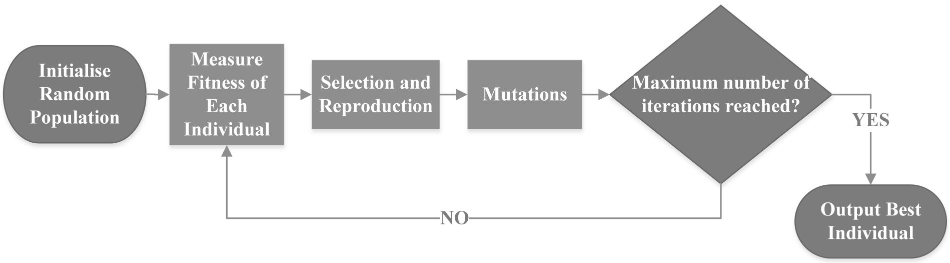

Initialise Random Population
Measure Fitness of Each Individual
Selection and Reproduction
Mutations
Maximum number of iterations reached?
YES
NO
Output Best Individual

Fig. 7. Applied genetic operators are shown in a basic flow chart of proposed experimental isolation mechanism

6 Results

6.1 Experimental data

Data set consists of 300 specimens obtained from the Department of Pathology at Cukurova University, Adana-Turkey. Experiments are conducted on randomly selected samples. All image samples are obtained from slides which have been processed using Papanicolaou staining. A Nikon microscope equipped with 100x magnification is used for taking sample images which are down-sized from 2560 x 1920 pixel to 1280x960 pixel resolution. Sample images were stored in RGB color space in JPEG format. Contours of the nucleus and cytoplasm of each specimen are segmented and examined by a pathologist in the Department of Pathology of Cukurova University. 2 Sample images from data set are displayed in Fig. 8.

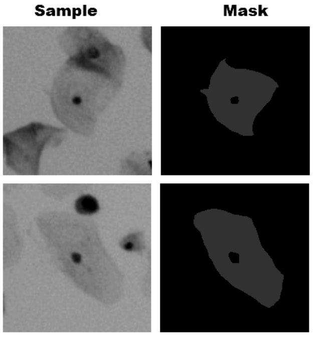


Fig. 8. A sample image and it's empiric area

6.2 Performance of fitness criteria

Introduced methods are applied in the MATLAB environment. It is aimed to observe if genetic structure is capable of generating diverted individuals of different characteristics and if given criteria is performing an efficient discrimination of best fitting individuals on nuclei regions. During experiments the algorithm is expected to sort individuals according to their fitness. A visual result of sorting process is shown in Fig. 9.

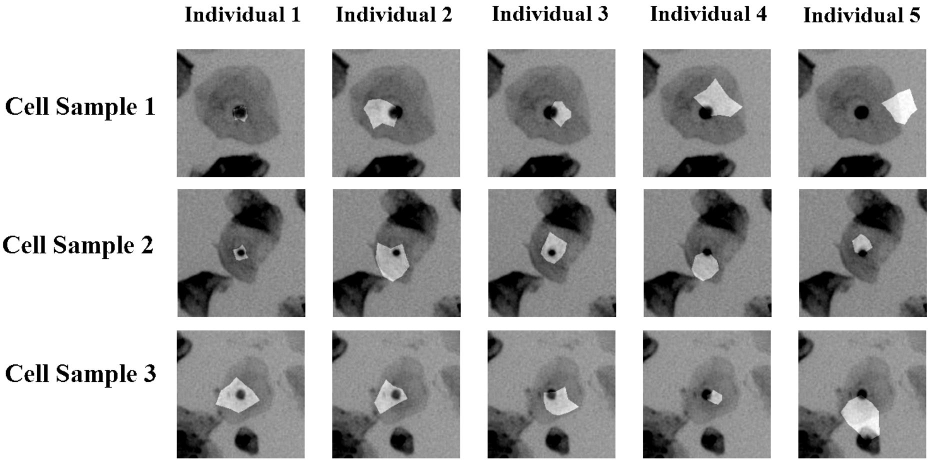


Fig. 9. Sorting process is visualised for 3 different populations on different samples where marked areas indicates each individual's region. First 5 best individuals are visualised for each population

F-score and jaccard index is accepted as objective criteria for evaluating actual segmentation success of each individual. They are described by:

$$Accuracy = \frac{Tp + Tn}{(N)} \tag{6.1a}$$

$$Precision = \frac{Tp}{(Tp + Fp)} \tag{6.1b}$$

$$Recall = \frac{Tp}{(Tp + Fn)} \tag{6.1c}$$

$$Fscore = 2\frac{Precision * Recall}{(Precision + Recall)} \tag{6.1d}$$

$$J(A \cap B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{6.1e}$$

Where Tp indicates number of true positives, Fp indicates number of false positives and Fn indicates number of false negatives. Jaccard measure gives a similarity score for judging how similar two binary images A and B [20]. It is calculated by division of number of common 1 valued pixels to number of total 1 valued pixels which is expressed above.

Experiments are conducted on randomly selected sample images. Each sample is image of a cervical cell smeared on thin glass. These cell images are previously segmented by an operator and serve as an empirical ground truth. Empirical mask of nuclei is accepted as actual value for measuring F-Score of each individual. Thus the region formed by individual is accepted as prediction mask.

Fitness criteria performance results given in Table 1 are obtained by 10 repetition for each sample with the following parameters and results are given for first 5 best fitting individuals.

- **Number of particles** $N_p = 10$

- **Elasticity** $\eta = 0.5$
- **Number of Individuals** $N_i = 100$
- **Radius Margin** $r_l = 5$ and $r_u =$ [Window Size $*$ 0.9]

F-score value is measured for observing actual sorting capability of fitness criteria. Numbers given in Table 1 indicates the percentage of the total score of 5 best fittest samples. It should be noted that given values are calculated for a single generation without application of any genetic operators.

Table 1. Fitness criteria performance results where individuals are sorted according to their CH and DB Scores

	CH Index				
	Individual 1	**Individual 2**	**Individual 3**	**Individual 4**	**Individual 5**
Sample 1	%35,98	%23,19	%20,21	%12,72	%7,90
Sample 2	%44,10	%24,98	%10,04	%13,66	%7,21
Sample 3	%45,64	%21,44	%17,02	%13,85	%2,05
Sample 4	%43,04	%28,86	%17,91	%4,53	%5,65
Sample 5	%40,87	%28,60	%15,87	%8,28	%6,39
	DB Index				
	Individual 1	**Individual 2**	**Individual 3**	**Individual 4**	**Individual 5**
Sample 1	%25,13	%8,04	%7,25	%25,02	%24,56
Sample 2	%20,08	%17,50	%29,07	%2,31	%11,04
Sample 3	%46,94	%10,63	%21,42	%2,66	%8,34
Sample 4	%12,26	%20,58	%14,94	%26,52	%15,71
Sample 5	%34,95	%7,90	%14,94	%7,53	%24,69

Moreover, genetic algorithm based experimental mechanism is set up for comparing evolutionary compatibility of both indices with proposed geometric genetic structure. Results are given as Jaccard similarity score. Genetic operators are applied to each generation and best fitting individual in each generation is accepted as solution. Results are given in Fig. 10.

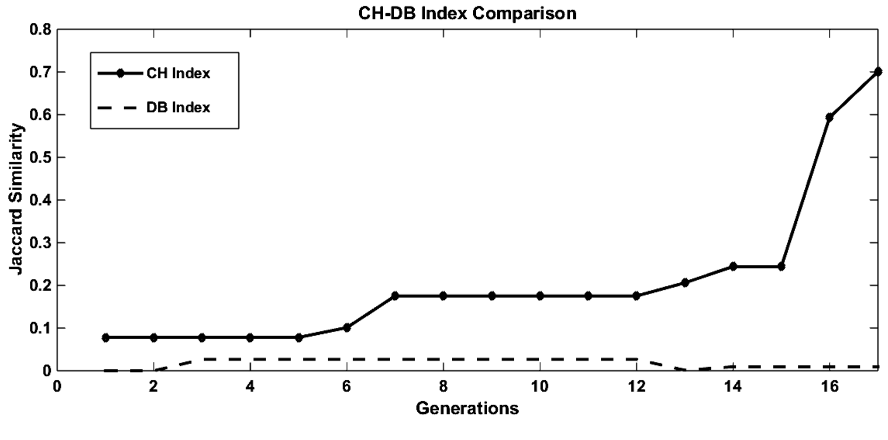


Fig. 10. Jaccard similarity score of best individual in each generation for CH and DB indices

Additionally euclidean distance of the cluster centroids are observed with CH index to support that form of the clusters are getting better with consequent generations. Fig. 11 indicates the increase on euclidean distance while enhancement on accuracy of the classification occurs.

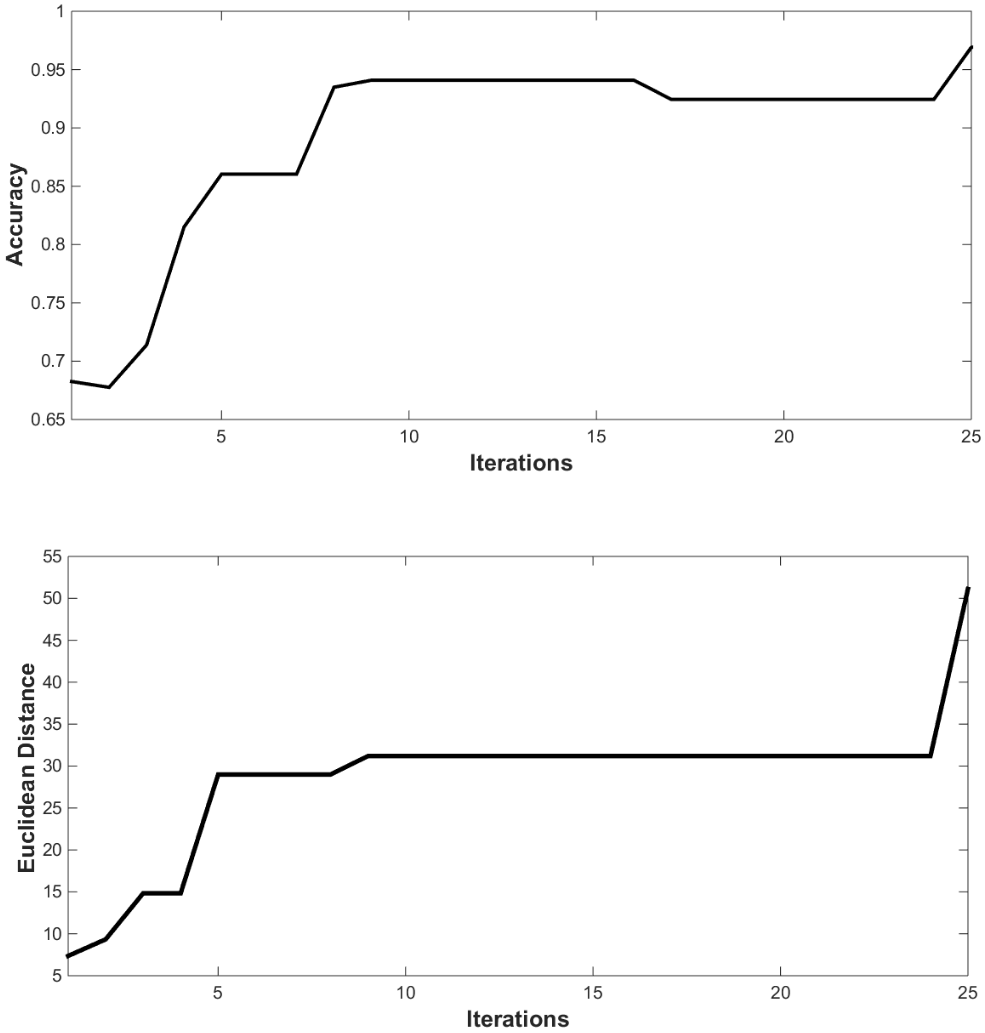


Fig. 11. Increase on accuracy and euclidean distance with several consequent populations is plotted

7 Discussions and Conclusion

In this study isolation of nuclei region from cytoplasmic area of cervical cell images is accepted as a data clustering problem which is desired to be solved by a genetic meta-heuristic algorithm. It was aimed to evaluate the effectiveness and efficiency of CH index as fitness criterion for proposed approach. Accordingly, data clusters are formed with a genetic structure where each individual in a generation represents a clustering result where nuclei and non-nuclei pixels are distinguished with binary values.

Fig. 9 visually shows that best fitting individual according to CH index is the geometrically best nuclei covering individual which would mean, in case of approaching to the actual contours of nuclei causes higher CH score.

In addition to visual results, Table 1 objectively reveals that, CH index may be an compatible fitness criteria for introduced segmentation approach. CH index gives better scores to better fitting regions on nuclei on most of the cases. Also according to results, it is possible to conclude that CH is more accurate then DB index on estimating which individual is fitting best.

Two different intra-cluster validation indices are observed while genetic operators are applied through iterations. Results are given as jaccard similarity index in Fig. 10 which reveals that CH index is functioning effectively with experimental genetic algorithm set-up where it is possible to observe that algorithm increases the total success of best fitting individual when CH index is preferred. It was observed that DB index is not capable of leading algorithm to better generations with same parameters.

Moreover Fig. 11 indicates that CH index is effectively discriminating best cluster forming individuals which leads algorithm to generate better individuals with each consequent generation.

According to presented results it would be possible to conclude that CH index as fitness criteria for proposed combination of genetic operators and data clustering would have a great potential on discriminating pixels belonging to cervical cell nuclei.

Acknowledgement

We would like to thank Seyda Erdogan and Aysun Uguz of Cukurova University Faculty of Medicine. They provided insight and expertise which improved the quality and expanded the scope of the research.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. Pattern Recognition. 2000;33(9):1455-65.

[2] Marinakis Y, Marinaki M, Dounias G. Particle swarm optimization for pap-smear diagnosis. Expert Systems with Applications. 2008;35(4):1645-56.

[3] Jiang T, De Ma S. Cluster analysis using genetic algorithms. In Signal Processing, 1996., 3rd International Conference, IEEE. 1996;2:1277-1279.

[4] Deb Kalyanmoy, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6.2. 2002;182-197.

[5] Murthy CA, Chowdhury N. In search of optimal clusters using genetic algorithms. Pattern Recognition Letters. 1996;17(8):825-32.

[6] Segura, Carlos, et al. Improving diversity in evolutionary algorithms: New best solutions for frequency assignment. IEEE Transactions on Evolutionary Computation; 2016.

[7] Scheunders P. A genetic c-means clustering algorithm applied to color image quantization. Pattern Recognition. 1997;30(6):859-66.

[8] Hornby Gregory et al. Automated antenna design with evolutionary algorithms. Space; 2006.

[9] Bandyopadhyay Sanghamitra, Ujjwal Maulik. Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognition 35.6. 2002;1197-1208.

[10] JainAK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010;31(8):651-66.

[11] Guven Mustafa, Caglar Cengizler. Data cluster analysis-based classification of overlapping nuclei in Pap smear samples. Biomedical Engineering; 2014;159.

[12] Talukdar J, Nath CK, Talukdar PH. Fuzzy clustering based image segmentation of pap smear images of cervical cancer cell using FCM algorithm. Markers. 2013;3(1).

[13] Cowgill MC, Harvey RJ, Watson LT. A genetic algorithm approach to cluster analysis. Computers & Mathematics with Applications. 1999;37(7):99-108.

[14] Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. Signal processing. 2003;83(4):825-33.

[15] Cengizler Caglar, Mustafa Guven, Mutlu Avci. A fluid dynamics-based deformable model for segmentation of cervical cell images. Signal, Image and Video Processing 8.1. 2014;21-32.

[16] Caliski T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods. 1974;3(1):1-27.

[17] Hruschka ER, Campello RJ, Freitas AA. A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2009;39(2):133-55.

[18] Schaffer J. David, Amy Morishima. An adaptive crossover distribution mechanism for genetic algorithms. Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1987.

[19] Li Ke et al. Adaptive operator selection with bandits for a multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on Evolutionary Computation 18.1. 2014;114-130.

[20] Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R, Vanhoutte A. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management. 1989;25(3):315-8.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/19643