Breast Histopathology Image Clustering using Cuckoo Search Algorithm

Krishna Gopal Dhal¹, Iztok Fister Jr.², Arunita Da³, Swarnajit Ray⁴, Sanjoy Das⁵

¹Dept. of Computer Science and Application, Midnapore College (Autonomous), Paschim Medinipur, West Bengal, India. Email: krishnagopal.dhal@midnaporecollege.ac.in
²Faculty of Electrical Engg. and Computer Sc., University of Maribor, Slovenia, Email: iztok.fister1@um.si.
³Dept. of Information Technology, Kalyani Govt. Engineering College, Kalyani, Nadia, India. Email: arunita17@gmail.com.
⁴Skybound Digital LLC, Kolkata, West Bengal, India. Email: swarnajit32@gmail.com.
⁵Dept. of Engg. and Technological Studies, University of Kalyani, Kalyani, India, Email: dassanjoy0810@hotmail.com

ABSTRACT

Breast histopathological image segmentation is exigent due to the existence of imperceptibly correlated and indistinct multiple regions of concern. Clustering based segmentation is one of the most significant approaches to perform proper segmentation of such complex images. K-means is the well-known clustering techniques but very sensitive to initial cluster centers and easy convergences to local optima. Therefore, researchers are employing Nature-Inspired Optimization Algorithms (NIOA) in this domain. This study develops Cuckoo Search (CS) algorithm based image clustering model for the proper segmentation of breast histopathology images. Experimental results show that CS provides better-quality segmented images compare to classical K-means algorithm by considering the computational time, fitness values and the values of quality parameters.

KEYWORDS

Clustering, K-means, Image Segmentation, Optimization, Swarm intelligence, Histopathology image.

1 Introduction

Breast Cancer is the mainly widespread kind of cancer in women worldwide [1]. Present breast cancer clinical practice and treatment mostly depend on the evaluation of the disease’s diagnosis. Bloom-Richardson grading system [2] describes the scoring of three morphological features of the dubious tissue, which are fraction of tubule construction, amount of nuclear pleomorphism, and mitotic cell count. However, the scoring had been performed by pathologists based on the visual assessment of the tissue’s biopsy sample under the microscope [3]. Therefore, researchers concentrate and suggest the use of image analysis methods to mitigate the said issue [4]. Digital histopathology and microscopy images carry out an important role in decision making in disease diagnosis, as they could provide wide information for computer-aided diagnosis (CAD), which facilitates quantitative analysis of digital images with a high throughput processing rate [5, 6]. At present, analysis of digital histopathology through image processing significantly assists pathologists and has attracted many attentions in both research and clinical practice.

A critical requirement in computer-aided diagnosis is segmentation, which is typically measured as the basis of automated image analysis. It provides assistances for various quantitative analyses such as shape, size, texture, and other imagenomics [5, 6]. However, it is difficult to achieve stout and perfect pathological image segmentation as these images frequently reveal background clutter with many noises, artifacts such as blurred regions introduced during image acquisition, and poor contrast between the foreground and the background. Second, there exist significant variations on cell size, shape, and intracellular intensity heterogeneity [5, 6].

In this study, proper segmentation of breast histopathology images is the main aim. Many efforts have been performed to attain automated segmentation of breast histopathology images which includes thresholding [7, 8], watershed method [9, 10], Active Contour model [8, 11], edge based approach [14], neural network [15, 16] etc. A Particle Swarm Optimization (PSO) with Otsu criterion based multi-level thresholding technique was proposed by Jothi and Rajam [7] to automatically segment the nuclei from hematoxylin and eosin (H&E) – stained breast histopathology images. To remove noise, the input image filtered by 3x3 gaussian filter. Experimental result proved that this method automatically segmented the nuclei effectively. A Localized Active Contour Model (LACM) in conjunction with an automatic technique for optimal curve placement with Krill Herd Algorithm (KHA) was developed by Beevi et. al. [8] for the segmentation of nuclei from breast histopathology images. Based on Hausdorff (HD) and Maximum Address Distance (MAD) measures segmentation performance was investigated. The proposed segmentation approach provided superior results compared to GA, Bacterial Foraging Algorithm (BFA), Harmony Search (HS) algorithm and FCM clustering method by considering the convergence rate, objective function values, computational time, sensitivity, precision and F-score. Shen et. al. [9] developed one improved watershed algorithm by
incorporating opening-closing reconstruction and the distance transform with chamfer algorithm after color deconvolution, and H-minima. It was claimed that the proposed segmentation model accurately detect the nuclei and overcome the limitation of classical watershed algorithm like over-segmentation due to the sensitivity to noise.

Another novel approach is to segment the nuclei regions and then resolve the overlapping or clump nuclei separation through heuristic approaches like the Concave Point Detection [11]. Wienert et. al. [11] presented novel contour-based “minimum-model” cell detection and segmentation approach that used minimal a priori information and detects contours independent of their shape. Experimental results proved that the proposed segmentation model capable to avoid the segmentation bias with respect to shape features and allows for an accurate segmentation with high precision (0.908) and recall (0.859).

A few researchers also presented different voting algorithms, which cast votes along gradient directions amplifying votes inside the centre of nuclei thereby locating the seed points as ones having maximum votes [12, 13, 14]. Consequently, the detected nuclei seed points had been either utilized to initialize active contours [12, 13] or an edge grouping algorithm [14]. Recently, deep neural network proved its effective performance in breast histopathology image segmentation field [15, 16]. Su et. al. [15] employed one fast Deep convolutional neural network (CNN) for pixel-wise region segmentation of breast histopathology images. Experimental results proved that the proposed segmentation model gave superior performance over both the LBP feature-based and texton-pixel-wise methods within less computational time. Naylor et. al. [16] developed one hybrid nuclei segmentation model by combining deep learning and mathematical morphology. Test results showed the promising performance of the proposed model.

Although, clustering based segmentation which shows its effective performance in hematopathology [17] or other histopathology [18] image segmentation domain, but have not been used in breast histopathology image field according to best of the knowledge. Therefore, this study concentrates to apply the clustering based segmentation to segmentize the different regions of the breast histopathology images. K-means is the well-known clustering techniques but sensitive to initial cluster centres and easy convergences to local optimization. Therefore, Nature-Inspired Optimization Algorithms (NIOA) are successfully employed to overcome the problems of K-means in image clustering domain [19-23]. For example, Orman et. al. [20] developed Particle Swarm Optimization (PSO) based satellite and MRI image clustering model and claimed that it outperformed some state-of-the-art methods for image classifier such as K-means, Fuzzy C-means, K-Harmonic means and Genetic Algorithm based model. In this study, Cuckoo Search (CS) algorithm has been employed and compared with classical K-means algorithms. Experimental results show that CS provides better-quality segmented images compare to classical K-means by considering the values of objective (fitness) function, computational time and quality parameters.

The paper is organized as follows: section 2 demonstrates the problem formulation and brief implementation of CS algorithm. Section 3 describes the experimental results and the paper is concluded in section 4.

2. Image Clustering using Cuckoo Search (CS) Algorithm

Clustering is a process of organizing data into clusters that have high intra-cluster and low inter-cluster similarity. It is clear that intra-cluster similarity should be maximized and inter-cluster similarity should be minimized. Based on this idea, objective functions are defined [24]. The best partitioning of a given data set can be attained by minimizing/maximizing one or more objective functions. The objective functions can be formed by capturing a certain statistical–mathematical relationship among the individual data items and the candidate set of representatives of each cluster (also known as cluster centroids) [25].

2.1 Problem Formulation

Suppose one specific dataset consists of C classes (i.e. C₁, C₂, ..., Cₖ) and N features. Therefore, the clustering problem is the finding of the optimal position of C centroids in an N-dimensional space i.e. each centroid is an N-dimensional vector. With this premises, the kth individual or solution of the applied optimization algorithm is a vector with N. C components which can be denoted as follows [19, 26]:

\[ X_i = (x_i^1, x_i^2, \ldots, x_i^C) \] (1)

Where, \( x_i^j = (x_{ij1}, x_{ij2}, \ldots, x_{ijk}) \)

So any solution in the population of the employed optimization algorithm consists of N. C components, each of which can take any real value.

The fitness function \( Fit \) has been calculated by summing out the Euclidean distance between the data vector instance \( b_k \) and the centroid of class \( CL \) it belongs to according to minimum distance criterion to the corresponding centroid i.e. \( (x_k^{CL_{min}(b_k)}) \) as in K-means.

\[ Fit (X_i) = \sum_{k=1}^{D_v} d_{(b_k,x_i^{CL_{min}(b_k)})} \] (2)

\( D_v \) is the number of data vectors to be clustered. \( d(\cdot, \cdot) \) is the Euclidean distance, \( X_i \) is the ith solution of the population. So by choosing the discussed fitness function, the problem can be considered as a minimization problem which is defined as follows:

\[ X_o = \text{Arg}[\text{Min}_{X_i}(Fit(X_i))] \] (3)

\( X_o \) is the optimal set of centroids. In the case of image clustering, \( C \) depends on user, \( N = 3 \) for RGB colour image, \( D_v \) is equal to
image size. The partitions into \( C \) classes should maintain the following properties:

1. Each cluster should have at least one data vector assigned i.e., \( C_i \neq \emptyset \) \( \forall i \in \{1, 2, \ldots, c\} \).
2. Two different clusters should have no data vector in common i.e., \( C_i \cap C_j = \emptyset, \forall i \neq j \) and \( i, j \in \{1, 2, \ldots, c\} \).
3. Each data vector should definitely be attached to a cluster i.e., \( U_{i-1}^i = D_{x} \).

### 2.2 Cuckoo Search (CS) Algorithm

Cuckoo search (CS) algorithm is a powerful optimization algorithm proposed by Xin-she Yang and Suash Deb in 2009 under the inspiration of the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of other host birds [27, 28].

A solution in the original cuckoo search algorithm corresponding to cuckoo nests represents the position of the cuckoo egg within the search space. Mathematically, this position is defined as:

\[
x^i_t = \{x^i_{t,j}\}, \text{ for } i = 1, \ldots, n \text{ and for } j = 1, \ldots, d \quad (4)
\]

Where, \( n \) denotes the number of cuckoo nests in the population, \( d \) is the dimension of the problem to be solved, and \( t \) the generation number. Generation of new solutions signifies the exploitation of the current solutions is carried out by using the Lévy flight distribution expressed as:

\[
x^{i+1}_t = x^i_t + \alpha \text{ Lévy}(\cdot) \quad (5)
\]

\( \alpha > 0 \) represents a scaling factor of the step size drawn from Lévy distribution i.e. \( \text{Lévy}(\cdot) \). Lévy distribution has the ability of exploring a large amount of search space. In this study, Mantegna’s algorithm [28] has been used to generate Lévy distribution. It produces random numbers according to a symmetric Lévy stable distribution as described below—

\[
\sigma = \left[\Gamma(1 + \beta) \sin(\pi \beta / 2) / \Gamma((1 + \beta) / 2) \beta 2^{((\beta - 1) / 2)}\right]^{1/\beta} \quad (6)
\]

Where, \( \Gamma \) is the gamma function, \( 0 < \beta \leq 2 \). \( \sigma \) is the standard deviation. As per Mantegna’s algorithm, the step length \( v \) can be calculated as,

\[
v = x / \left| y \right|^{1/\alpha} \quad (7)
\]

Here, \( x \) and \( y \) are taken from normal distribution and \( \sigma_x = \sigma_y = 1 \). Where \( \sigma \) is the standard deviation. The resulting distribution has the same behavior of Lévy distribution for large values of the random variables. Mantegna’s algorithm associates with faster computational speed in the range \( 0.75 \leq \alpha \leq 1.95 \) [28].

In CS, exploration of the search space has been done by the following expression:

\[
x^i_{t,j} = (U b_j - L b_j) * U_j(0,1) + L b_j \quad (8)
\]

Where, \( U b_j \) and \( L b_j \) are the upper and lower bound of the specific variable. \( U_j(0,1) \) is the random variable drawn from the uniform distribution. But, this exploration ability of the CS algorithm crucially depends on the probability \( p_a \in [0,1] \) as this fraction of nests is abandoned.

Pseudo Code of the traditional CS is given as Algorithm 1.

#### Algorithm 1: Cuckoo Search Algorithm

1. **Step 1:** Take objective function and generate initial population of \( n \) solution randomly using Eq.8.
2. **Step 2:** While termination condition does not meet Do
   1. **Step 3:** for \( i=1 \) to \( n \) Do
   2. **Step 4:** Generate new solutions around \( x_t \) with Lévy flight as per Eq.5
   3. **Step 5:** \( f_t = \text{Suppose the new solution is } u_t \) and find the fitness values of \( u_t \)
   4. **Step 6:** \( j = [\text{rand} (0,1) \ast n + 1] \)
   5. **Step 7:** if \( f_t < f_j \) then
   6. **Step 8:** \( x_j = u_t; f_j = f_t \)
   7. **Step 9:** end if
   8. **Step 10:** if \( \text{rand} (0,1) < P_a \) then
   9. **Step 11:** Do the initialization of worst nest according to Eq.8 and \( P_a \)
   10. **Step 12:** end if
   11. **Step 13:** if \( f_t < f_{\text{min}} \)
   12. **Step 14:** \( x_{\text{best}} = u_t; f_{\text{min}} = f_t; \) replace the global best solution
   13. **Step 15:** end if
   14. **Step 16:** end for
   15. **Step 17:** end While

### 3. Experimental Results

The experiment has been performed over 40 colour breast histopathology images with MatlabR2012b and Windows-7 OS, x64-based PC, Intel(R) Pentium(R)-CPU, 2.20 GHz with 4 GB RAM. The proposed methods are tested on images taken from UCSB Bio-Segmentation Benchmark dataset [29, 30]. Fig.1 represents the original images of different breast histopathology images.

#### 3.1 Parameter Setting

Parameter setting is very crucial for any Nature-Inspired Optimization Algorithm (NIOA) and most of the cases it is performed from experience. The parameter setting of CS algorithm is as follows: \( p_a = 0.2, \beta = 1.5, \alpha = 0.01 \), population
size \( n \) = 50. The Stopping Criterion for both CS and K-means \cite{22} is the number of Fitness Evaluations (FEs), and the maximum number of FEs (i.e. MAX_FE) has been taken as 1000 \( \times D \). Where, \( D \) is the number of clusters. In K-means, new centroid calculation has been considered as Fitness Evaluations (FEs).

### 3.2 Segmentation Quality Parameters

Segmentation efficiency of the employed methods are judged with the help of three well-known image quality assessment parameters namely Peak Signal to Noise Ratio (PSNR), Quality Index based on Local Variance (QILV) and Feature Similarity Index (FSIM).

QILV is a performance measurement technique proposed by Santiago Aja-Fernandez \cite{31, 32}. QILV is used to measure the structural information of the image. A great amount of the structural information of an image is coded in its local variance distribution. Local variances features of an image can assist to compare two images properly. Greater QILV values indicate better segmentation quality.

PSNR \cite{33, 34} is a pixel difference measurement technique. This is computed by averaging the squared intensity of original image and output image. Large value of PSNR demonstrates better-segmented image.

Zhang et al. proposed a new human vision measurement technique in 2011 called Feature Similarity Index (FSIM) \cite{35}. FSIM is mainly designed for gray scale images or luminance component of colour images. This method is a combination of two low level features namely Phase Congruency (PC) and Gradient Magnitude (GM). Greater value of FSIM represents better-segmented image.

![Fig.1 Original breast cancerous histopathology images](image1)

No. of Clusters = 4  No. of Clusters = 6  No. of Clusters = 8

![Fig.2. Results of clustering for Fig.1(a): (a)–(c) represent results of CS using cluster number 4, 6 and 8 respectively; (d)–(f) represent results of K-means using cluster number 4, 6 and 8 respectively.](image2)

![Fig.3. Results of clustering for Fig.1(d): (a)–(c) represent results of CS using cluster number 4, 6 and 8 respectively; (d)–(f) represent results of K-means using cluster number 4, 6 and 8 respectively.](image3)

Table 1. Average Quality parameters and other numerical results over 40 images

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Clus.</th>
<th>PSNR</th>
<th>QILV</th>
<th>FSIM</th>
<th>Fitness</th>
<th>Time</th>
<th>STD.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>4</td>
<td>25.69</td>
<td>0.98</td>
<td>0.95</td>
<td>0.097</td>
<td>2.003</td>
<td>3.31 E-17</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>28.82</td>
<td>0.99</td>
<td>0.96</td>
<td>0.072</td>
<td>2.665</td>
<td>8.37 E-14</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>30.72</td>
<td>0.99</td>
<td>0.97</td>
<td>0.068</td>
<td>3.108</td>
<td>1.23 E-14</td>
</tr>
<tr>
<td>K-means</td>
<td>4</td>
<td>24.81</td>
<td>0.97</td>
<td>0.90</td>
<td>0.189</td>
<td>45.14</td>
<td>............</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>26.11</td>
<td>0.97</td>
<td>0.91</td>
<td>0.127</td>
<td>50.99</td>
<td>............</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>27.80</td>
<td>0.98</td>
<td>0.93</td>
<td>0.102</td>
<td>56.84</td>
<td>............</td>
</tr>
</tbody>
</table>

/* Best results obtained are given in bold*/

### 3.3 Analysis of the experimental results

Performance analysis of the CS and classical K-means algorithms has been performed in the breast histopathology image clustering...
domain by computing their objective (fitness) function minimization ability, robustness in term of standard deviation (STD.) and computational time. Segmentation quality parameters like QILV, PSNR and FSIM are also calculated to judge the segmentation accuracy of the clustering models over cluster number 4, 6 and 8. CS algorithm based clustering model has been run 40 times for each image. The segmented images corresponding to Figs.1(a)-(b) are given as Figs.2-3 respectively. On the other hand, Table 1 represents the average values of fitness, standard deviation (STD.), computational time and quality parameters over 40 images. Table 1 shows that CS based clustering model provide minimized fitness value within less computational time. The standard deviation also proves the significant robustness of CS algorithm. The resultant segmented images of CS based model associated with greater QILV, FSIM and PSNR over cluster number 4, 6 and 8. Therefore, it can be said that CS is the better than traditional K-means algorithm with the MAX_FE based termination condition.

Conclusion

A Cuckoo Search (CS) algorithm based clustering model has been proposed for the proper segmentation of breast histopathology images. The performance of the CS algorithm based clustering model has been compared to K-means algorithms in terms of fitness, computational time and quality parameters. Values of quality parameters indicate that CS based clustering model provide segmented images with higher QILV, FSIM and PSNR compares to K-means based clustering model. Analysis of the results also shows that K-means also associates with huge computational time when number of clusters increases and this time consume problem has been successfully surmounted by using CS algorithm. Several future directions exist of this study such as use of fuzzy logic based clustering, rough set based clustering, formulation of the multi-objective clustering models, and use these clustering models for different kinds of images.

REFERENCES