A New Validity Index for Fuzzy C-Means for Automatic Medical Image Clustering

Sayed F. Bahght
Computer Science Department
College of Computers and Information Technology
Taif University, Saudi Arabia

Sultan Aljahdali
Computer Science Department
College of Computers and Information Technology
Taif University, Saudi Arabia

E. A. Zanaty
Computer Science Department
College of Computers and Information Technology
Taif University, Saudi Arabia

Ahmed S. Ghiduk
Computer Science Department
College of Computers and Information Technology
Taif University, Saudi Arabia

Ashraf Afifi
Computer Engineering Department
College of Computers and Information Technology
Taif University, Saudi Arabia

ABSTRACT
Many clustering and segmentation algorithms suffer from the limitation that the number of clusters/segments is specified by a human user. It is often impractical to expect a human with sufficient domain knowledge to be available to select the number of clusters/segments to return. Thus, the estimation of optimal cluster number during the clustering process is our prime concern. In this paper, we introduce a new validity index method based on multi-degree entropy algorithm. This multi-degree entropy algorithm combines a multi-degree immersion and entropy algorithm to partition an image into levels of intensity using multi-degree immersion processes. The output of the multi-degree immersion process is several regions which the interior does not contain any sharp grey value transitions, i.e. each level of intensity may contain one or more regions, connected points, or oversegmentation. These regions are passed to the entropy procedure to perform a suitable merging which produces the final number of clustering based on validity function criteria. Validity functions typically suggest finding a trade-off between intra-cluster and inter-cluster variability, which is of course a reasonable principle. The latter process uses a region-based similarity representation of the image regions to decide whether regions can be merged.

The proposed method is evaluated on a discrete image example to prove its efficiency. The existing validation indices like PC, XB, and CE and the proposed index are evaluated and compared on two simulation and one real life data. A direct benefit of this method is being able to determine the number of clusters for given application medical images.

General Terms
Image Processing.

Keywords
Clustering, Multi-Degree Immersion, Entropy, Validity Index.

1. INTRODUCTION
Clustering is one of the most popular classification methods and has found many applications in pattern classification and image segmentation [1]-[6]. Clustering algorithms attempt to classify a voxel to a tissue class by using the notion of similarity to the class. Unlike the crisp K-means clustering algorithm [4], the FCM algorithm allows partial membership in different tissue classes. Thus, FCM can be used to model the partial volume averaging artifact, where a pixel may contain multiple tissue classes [2]-[3]. The kernelized fuzzy C-means (KFCM) [6]-[8] used a kernel function as a substitute for the inner product in the original space, which is like mapping the space into higher dimensional feature space. Other approaches were used to incorporating kernels into fuzzy clustering algorithms for enhancing clustering algorithms designed to handle different shape clusters [8]. More recent results of fuzzy algorithms have been presented in [9] for improving automatic MRI image segmentation. They used the intra-cluster distance measure to give the ideal number of clusters automatically; more discussion can be found in [9]. Also, possibilistic clustering which is pioneered by the possibilistic c-means (PFCM) algorithm was developed in [10-12]. They had been shown that PFCM is more robust to outliers than FCM. The While PCM-based algorithms suffer from the coincident cluster problem, which makes them too sensitive to initialization [12]. The PCM-based algorithms suffer from the coincident cluster problem, which makes them too sensitive to initialization [12].

Most fuzzy methods have several advantages including yielding regions more homogeneous than other methods; reducing the spurious blobs; removing noisy spots; reduced sensitivity to noise compared to other techniques. However, they require prior knowledge about the number of clusters in the data, which may not be known for new data [13]. Many criteria have been developed for determining cluster validity [14-21], all of which have a common goal to find the clustering which results in compact clusters that are well separated. Now the challenge is to answer the two questions:" Can the appropriate number of clusters be determined automatically? And if the answer is yes, how?" [19]. To the best of our knowledge, however, faithful indexes for automatic fuzzy clustering algorithms have not been determined yet, i.e. to determine which validity indexes can achieve high accuracy segmentation when used with fuzzy algorithms.

In this paper, we seek the answer to the previous questions for exploring which indexes can achieve high accuracy segmentation. For that we introduce a new validity index based on multi-degree entropy and a new validity function to obtain the cluster validity in the domain of image

segmentation. The multi-degree entropy algorithm combines a multi-degree immersion and entropy algorithm. The proposed method begins to subdivide the data into fixed number of clusters called number of levels of intensity using multi-degree immersion processes. The multi-degree immersion results several regions. These regions are fed to the entropy procedure to perform a suitable merging which produces the final numbers of clustering based on validity function criteria. Validity function is used as pre-merge to find the final true number of clusters. The proposed method is tested with discrete grey image example to prove its efficiency. Also, it is applied to two simulation and one real life data. The obtained results are compared to those obtained from validation indices like PC, XB, and CE. It is shown that the proposed method produce accurate results. Furthermore, the proposed method is experimented on several brain images to show the applicability of this method in medical image segmentation.

The rest of this paper is organized as follows: Section 2 describes optimization of cluster number. Some well-known fuzzy clustering validity indexes are introduced in section 3. The proposed method steps are discussed in section 4. In section 5, the proposed algorithm is presented. The experimental results were performed in section 6. In section 7, we present the conclusion.

2. CLUSTER NUMBER OPTIMIZATION

The objective function of FCM can be formulated as follows [3]:

$$J_m = \sum_{i=1}^{C}\sum_{j=1}^{n} u_{ij}^{m}\left\|x_j - c_i\right\|^2$$

where C is the number of clusters, ci is the cluster centre of fuzzy group i, n is the number of data, and the parameter m is a weighting exponent on each fuzzy membership.

Where $u_{ij} = u_j(x_i)$ is the membership of the i-th object xi in the j-th cluster. In the commonly employed probabilistic version of fuzzy C-means, it is required that:

$$\sum_{j=1}^{C} u_{ij} = \sum_{j=1}^{C} u_j(x_i) = 1, \forall x_i, i = 1,2,\ldots,n$$

Fuzzy partitioning is carried out through an iterative optimization of the above objective function. Updating of membership uij and the cluster centers Ci is done as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_j - c_i\right\|}{\left\|x_j - c_k\right\|}\right)^{2/(m-1)}} \quad (1)$$

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \quad (2)$$

As mentioned before, the simple enumeration strategy for optimizing the cluster number, as outlined in the introduction, is not practicable in an online setting as it requires the consideration of too large a number of candidate values and, hence, applications of the clustering algorithm [22]. To minimize the effort , the idea of this paper is to pursue a local adaptation process that tries to adapt the cluster number C on the basis of a starting point Co in the style of a hill-climbing procedure. This strategy appears particularly appealing in an online setting where the optimal cluster number, C*, may "smoothly" change in the course of time. In fact, assuming, that C* does not make big jumps, the optimal number at time t+1. In other words, a local search is likely to succeed without getting trapped in local optima. Thus, staring with C=Co, each iteration of our method consists of a test that checks whether the cluster model can be improved by increasing or decreasing C. To this end, we make use of a suitable quality measure (validity function) Q(.) . Let Q(K) denote the quality of the cluster number K, that is, of the cluster model obtained for this number. In each iteration, K is then updated as follows:

$$C \leftarrow \arg\max\{Q(C-1), Q(C), Q(C+1)\}$$

This is repeated until C remains unchanged, i.e., Q(C)> max{Q(C-1),Q(C+1)}.Essentially, this approach requires two elements: Firstly, a suitable validity function Q(.), and secondly, a means for going from a clustering structure with C clusters to structures with C-1 and C+1 clusters, respectively.

3. FUZZY CLUSTERING VALIDITY INDEXES

Clustering analysis aims to place similar objects in the same groups. The purpose is to get an idea about the sample dispersions and about the correlations between variables in the samples which include huge data. However, many clustering algorithms necessitate pre-knowledge of the number of clusters. The fact that the researchers do not have pre-knowledge of the number of clusters in many studies make it impossible to know whether the end number of clusters is more or less than the actual number of clusters. If the end number of clusters turn out to be less than the actual number of clusters, then one or more of the present clusters will have to unite; if it turns out to be more, then one or more of the present clusters will be divided. The process of determining the optimal cluster number is called cluster validity in general. Thus, the accuracy of the end cluster number can be determined. When the data are in the two dimensional space, the number of clusters can be decided upon by commenting on the cluster results visually. However, as the number of dimensions increase in space, visually gets harder and there becomes a need for validity indexes. As a result, two criteria can be mentioned for value clusters and the most suitable cluster planning.

1. Density: It measures how close the group members are. The best example to this is variance.
2. Separation: It shows how two clusters are separated. It measures the distance between two different clusters.
3. Statistical: It adopted criteria for statistical model selection for determining the statistical behavior of the data.

In this paper, we focus on combining the density and separation methods to find the best cluster number. Thus, in this section, we evaluate the most well-known methods such as partition coefficient (PC), classification entropy (CE), and Xie-Beni index (XB) to compare with the results of the proposed method. These comparisons are necessary to prove the efficiency of the proposed method.

3.1 Partition Coefficient (PC)

This method proposed by Bezdek[23] and holds a value between 1/C and 1. Here, C is the number of clusters. If all membership values turn out to be equal as a result of fuzzy partition, uij= 1/C. This is the smallest degree of the PC. It is desirable that the value of the PC in the appropriate clustering process has a value close to 1. As the PC value gets closer to 1/C, clustering will become fuzzy. Besides, a value close to 1/C indicates that the clustering algorithm has failed.

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n} u_{ij}^{2} \quad (3)$$

3.2 Classification Entropy (CE)

This method has been proposed by Bezdek [23] as well.

$$V_{CE} = -\frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n} u_{ij}\log_a u_{ij} \quad (4)$$

In this equation, a logarithm is e base. CE value needs to be close to 0. The best number of clusters will be between the $2 \le C \le (n-1)$ ranges.

3.3 Xie-Beni Index (XB)

This index developed by Xie and Beni [24] is also known as the density and secession validity function and it is as follows:

$$V_{XB} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{n} u_{ij}^{m}\left\|x_k - c_i\right\|^2}{n\min_{ij}\left\|c_i - c_j\right\|^2} \quad (5)$$

Where $\min_{ij}\left\|c_i - c_j\right\|$ represents the shortest value between the i-th cluster ci in the j-th cluster cj.

4. THE PROPOSED APPROACH

In this section, we present a method for assessing cluster validity. This method combines with a proposed fuzzy clustering algorithm to yield an estimate of the data partition, namely, the number of clusters. This method is called multi-level entropy algorithm that combined multi-degree immersion and entropy algorithm. The algorithm can be described in the following steps:

1. Multi-degree immersion
2. Entropy procedure
3. Fuzzy validation function

These steps will be discussed in more details.

4.1 Multi-Degree Immersion

Now we summarize the definition of multi-degree immersion processes [25]. Let F: $D \to N$ be a digital grey value image, with $h_{\min}$ and $h_{\max}$ be the minimum and maximum values of F. Define an image with the grey level h increasing from $h_{\min}$ to $h_{\max}$, in which the basins associated with the minima of F are successively expanded. The multi-degree immersion implementation was introduced in [25] to resist the over segmentation problem. The threshold set of F is redefined at level h:

$$T_h = \left\{ p \in D \middle| F(p) - Diff(p) \le h \right\} \quad (6)$$

Here Diff (p) is a function which presents the immersion level when the flood procedure reaches pixel p. The segmentation results are sensitive to the value of this function. Generally, the greater value of Diff (p) means immersing more points, when the flood process goes to level-by-level, where:

$$Diff(p) = \sum_{q \in Neighbor(p)} \frac{\left|F(p) - F(q)\right|}{connectivity} \quad (7)$$

where the connectivity is a prescribed value. This shows that the phenomenon of over segmentation problem is still not enhanced since the connectivity of Diff (p) fails to merge more pixels.

Our algorithm is based on these definitions. Let we have the subset of points X_1 , X_2 , .., X_L corresponding to the thresholds T_h , T_{h+1}, …, $T_{h_{\max}}$ respectively, where X_1 , X_2 , .., X_L being connected components of $T_h(F)$ and L is the number of extracted subsets. Next, we calculate the entropy function based on these subsets X_1 , X_2 , .., X_L .

Algorithm 1
Input: digital grey scale image F matrix.
Output: subset X_i.
Procedure: SORT pixels in increasing order of grey values (minimum h_{min}, maximum h_{max})
i=1 (* Start Flooding *)
For $h = h_{min}$ to h_{max} Step T_h
Find X_h matrix (* $X_h = T_h(F)$ *)
Find the matrix Y_i which satisfy $X_h - Diff \le Y_i \le X_h + Diff$
Connect all pixels of Y_i to get connected regions X_i
i=i+1
End For (* End Flooding *)

4.2 Entropy Procedure

After performing multi-degree immersion processes at each level, sometimes there are segments that are difficult to merge due that the boundary of regions is disjoint. Thus, we apply entropy [26], in measuring the correlation between resultant regions. Here we treat segments as random variables. The most frequently used measure of information is the Shannon-Wiener entropy measure [27], the entropy H of a discrete random variable X with n values in the set [x1,x2,x3…….xn] with probabilities pri i=1,2,…,n can be defined as:

$$H(X) = -\sum_{i=1}^{n} p_{ri}\log p_{ri} \quad (8)$$

Where pri=pr[X=xi]. The image entropy, H(X) is usually estimated from:

$$p_{ri} \equiv \frac{g_i}{g_{total}}$$

Where gi is the number of pixels with the intensity i and gtotal is the total number of pixels. The joint entropy could be used as a similarity measure between two regions. Having two sets of pixels, one of X_i and another of X_j, and E_k $k=1,2,…M$ is the resultant of the intersection between two sets. The entropy Pri of pixels in E_k is computed corresponding to the union between pixels of X_i and X_j sets. The largest value of Pri shows that there is similarity between the two regions and then they must be merged in one segment. The algorithm can be described as follows:

Algorithm 2: Procedure E_i=Entropy(X_i, X_j), E_k=Entropy(X_i,X_j)
Input: X_i, X_j ; i , j=1,2,..,L.
Output: E_i=Entropy(X_i, X_j)

For i = 1 to L
For j = 2 to L-1
b= $X_i \cap X_j$
Calculate pr_i of matrix b corresponding to X_i
If $pr_i >$ a prescribe value then resultant segment is $R_i = X_i \cup X_j$.
Else
Return by $R_i = X_i$.
End IF
End For
End For
Delete the redundancy matrix R_i and the corresponding E_i .
End procedure

4.3 Fuzzy Validity Function

Regarding the evaluation of a cluster model (partition of the data into regions $R_K, K = 1,2,..,C$) in terms of a measure Q(.), several proposals can be found in the literature[1-8]. Unfortunately, most of these measures have been developed for the non-fuzzy case. Indeed, validity functions of that kind might still be (and in fact often are) employed, namely by mapping a fuzzy cluster model to a crisp one first (i.e. assigning each object to the cluster in which it has the highest degree of membership) and deriving the measure for this latter structure afterwards. However, our validity function can of course be criticized as it comes along with a considerable loss of information. On the other hand, many of the non-fuzzy measures can be adapted to the fuzzy case in a natural way. Validity functions typically suggest finding a trade-off between intra-cluster and inter-cluster variability, which is of course a reasonable principle. We can define the validity function as $V = Q(x_i, C, f_C, u_{ij}, R)$:

$$V = \frac{f\left\|c_K - c_{K+1}\right\|^2 \sum_{j=1}^{C}\sum_{i=1}^{f_C} u_{ij}^m \left\|x_j - c_j\right\|^2}{f_c\left\|\max(R_K) - \max(R_{K+1})\right\|^2 \sum_{i=1}^{f} u_i^m \left\|x_i - c_{K\cup(K+1)}\right\|^2}$$

$$\left\|c_K - c_{K+1}\right\|^2 \frac{1}{f_C}\sum_{j=1}^{C}\sum_{i=1}^{f_C} u_{ij}^m \left\|x_i - C_j\right\| \geq \left\|\max(R_K) - \max(R_{K+1})\right\|^2 \frac{1}{f}\sum_{i=1}^{f} u_i^m \left\|x_i - c_{K\cup(K+1)}\right\|^2 \quad (9)$$

Where $f = \sum_{k=1}^{C} f_k$ is the summation number of points of in the regions R_K (f_K is the corresponding number of points of K- cluster, K=1, 2,..,C); and u_{ik}^m , u_i^m are the two memberships of two individual clusters R_K and R_{K+1} with two centers c_K, c_{K+1} and as one cluster $S = R_K \cup R_{K+1}$ with centre $c_{K\cup(K+1)}$ respectively.

$$\max(R_K) = \max|R_{K+1} - c_K|$$

$$\max(R_{K+1}) = \max|R_{K+1} - c_{K+1}|$$

If this validity function is true, two regions are one region else they are separated regions. Now we have fK the number of connected regions R_K , K=1,..,C, and the corresponding the entropy EK respectively.
For instance, if you have p-th and q-th regions with centres c_p, c_q , the validity criterion can be rewritten as:

$$V_{pq} = \left\|c_p - c_q\right\|^2 \frac{1}{f_C}\left|\sum_{i=1}^{f_p} u_{ip}^m \left\|x_i - c_p\right\|^2 + \sum_{i=1}^{f_q} u_{iq}^m \left\|x_i - c_q\right\|^2\right|$$

$$W_{pq} = \left\|\max(R_p) - \max(R_q)\right\|^2 \frac{1}{f}\sum_{i=1}^{f} U_i^m \left\|x_i - c_{p\cup q}\right\|^2$$

These two regions can be merged together if $V_{pq} > W_{pq}$ else the two regions still without merging.
This algorithm can be described as follows:

Algorithm 3: Optimization of cluster number.
Input: The connected regions R_K , *i=1,..,K*
Output: the entropy E_K for regions.
Labeling: E_K for each R_K.
Sort their regions R_K according to E_K.
Repeat
K=1
S=R_K U R_K+1
Estimate: the two centers and their memberships of S.
Evaluate $V_{K(K+1)}$
Evaluate $W_{K(K+1)}$
IF $V_{K(K+1)} > W_{K(K+1)}$
R_K and R_{K+1} are merged in R_{K+1} and delete R_K.
Else
Still without merging
K=K+1
End IF
End Repeat until checked all regions.
End

5. THE PROPOSED ALGORITHM DESCRIPTION

Determining the best cluster number in fuzzy clustering becomes more important especially if the clusters are not separated from each other significantly. In case of uncertainty, cluster validity indexes help the researcher in making definite decisions. Many cluster validity index in the literature give conflicting results about the cluster numbers with data in complicated form [19], [28]-[30]. After the application of fuzzy clustering method, each data is appointed to the cluster in which it has the highest membership degree. As a result of a classification done with these results any classification technique is expected to have high percentage of classification. In this technique, we used an alternative validity criterion based on validity function and entropy. If entropy method is used as a classification method, the input of this procedure will be the level of intensity and the output will be the initial cluster number and regions. These cluster number and regions are fed to the validity function which determines the true cluster number as a result of fuzzy clustering. After performing the proposed algorithm, we noted that this can give the high percentage of classification accuracy, where the most appropriate number of clusters can be determined in the fuzzy clustering. The proposed algorithm can be stated as follows:

Algorithm: the overall algorithm of our proposed approach.
Input: *F*: Image.
Output: *C*: number of clusters in the image *F*.
Begin:

1. **Sort** the image's pixels and identify the two pixels with minimum and maximum values.
2. **Divide** the image *F* into levels according a selected threshold.
3. **For** each level *F* (applying algorithm 1)
   a. Get the connected regions.
   b. Isolate the new regions (connected regions) in each level.

   End For
4. **For** each region (applying algorithm 2)
   a. Find the entropy of each region.
   b. Labeled the regions and the corresponding entropy value

   End For
5. **Sort** the connected regions according to their entropy.
6. **Merge** the regions which have a close entropy values to reduce the number of regions.
7. **For** each two adjacent regions (applying algorithm 3).
   a. Find the center and membership using fuzzy c-means
   b. Calculate the number of clusters using validity function.
   c. Merge two regions or not according to the value of validity function.
   d. Update the regions if the regions are merged together.
   e. Continue to check all regions.
   f. Count the number of resultant regions.

   End for

End

6. EXPERIMENTAL RESULTS

The algorithm is based on the definition given in section 3. We therefore start by partitioning the image into several levels of intensity using multi-degree immersion process which produces the initial partitioning of the image regions. We obtained the matrices X_1, X_2,..., X_L corresponding to the thresholds T_h, T_{h+1},..., $T_{h_{\max}}$. These subsets X_1, X_2 ..., X_L are fed to entropy function to decide if these regions can be merged or not. These output subsets are fed to the validity function which determines the true cluster number as a result of fuzzy clustering.

6.1 Numerical Results

For example, if one has a 7x5 discrete image F on the square grid with 4-connectivity (see Fig.(1a)). The local minima $X_{h_{\min}}$ which belong to the minima of lowest altitude $h_{\min}=1, T_h=30$ a multi-level by immersions are applied on the 4-connected grid, and define Diff=4, connectivity is equal to 2.We can apply our algorithm as the following:

Step-1: According to algorithm 1, we can divide the image F into three levels according to

$$0 \le h \le 30,\ 30 \le h \le 60, and\ 90 \le h \le 120.$$

For the first level, we obtained three regions (X1, X2, X3). Similarly, we get one for the second level (X4), and two regions for the third level (X5, X6); the set of all regions as shown in figure (1-b, c, d) are $X_1=(1,3,2,3,5,2,1,4,8)$, $X_2=(11,10,7,6)$, $X_3=(1,5,2,3,6,3)$, $X_4=(56,50,59,54,51,58)$ $X_5=(90,\ 100,\ 95,\ 98,\ 93,\ 102)$, $X_6=(100,\ 105,\ 106,\ 107)$.

1	3	2	90	100
3	5	2	95	98
1	4	8	93	102
56	50	59	11	10
54	51	58	7	6
1	5	2	100	105
3	6	3	106	107

(a)

1	3	2	0	0
3	5	2	0	0
1	4	8	0	0
0	0	0	11	10
0	0	0	7	6
1	5	2	0	0
3	6	3	0	0

(b)

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
56	50	59	0	0
54	51	58	0	0
0	0	0	0	0
0	0	0	0	0

(a)

0	0	0	90	100
0	0	0	95	98
0	0	0	93	102
0	0	0	0	0
0	0	0	0	0
0	0	0	100	105
0	0	0	106	107

(b)

Fig. (1): multi-degree immersion process on the 4-connected grid, (a) Original image, b, c, d are segmented using Eq. (8).

Step 2: according to algorithm 2, we compute E_1 for R_1 as:

$$E_1=\frac{-1}{9}\times\left\{\left[\frac{1}{1407}\log_{10}\frac{1}{1407}\right]+\left[\frac{3}{1407}\log_{10}\frac{3}{1407}\right]+\left[\frac{2}{1407}\log_{10}\frac{2}{1407}\right]+\left[\frac{3}{1407}\log_{10}\frac{3}{1407}\right]+\left[\frac{5}{1407}\log_{10}\frac{5}{1407}\right]+\left[\frac{2}{1407}\log_{10}\frac{2}{1407}\right]+\left[\frac{1}{1407}\log_{10}\frac{1}{1407}\right]+\left[\frac{4}{1407}\log_{10}\frac{4}{1407}\right]+\left[\frac{8}{1407}\log_{10}\frac{8}{1407}\right]\right\}=5.86*10^{-3}$$

Similar,

$E_2=13.34*10^{-3}, E_3=6.08*10^{-3}, E_4=54.77*10^{-3}$

$E_5=79.7*10^{-3}$, *and* $E_6=83.8*10^{-3}$ are computed according to regions R_1, R_2, R_3, R_4, R_5,R_6 respectively.

Step 3: The regions R_1, R_2, R_3, R_4, R_5 and R_6 are sorted according to their entropy values E_6, E_5, E_4, E_3, E_2, E_1. Then, we merge the similar regions according to their entropy values. In this example, there are similar entropy values between (X_1, X_3) and (X_5, X_6) to get R_1 and R_4 regions and pixels of (X_2, X4) corresponding regions R_2, R_3 as shown in figure (2b).

1	3	2	90	100
3	5	2	95	98
1	4	8	93	102
56	50	59	11	10
54	51	58	7	6
1	5	2	100	105
3	6	3	106	107

(a)

1	3	2	90	100
3	5	2	95	98
1	4	8	93	102
56	50	59	11	10
54	51	58	7	6
1	5	2	100	105
3	6	3	106	107

(b)

Fig. 2: (a) Original image, (b) Clustering image after entropy

Step 4: For each region R_i, we calculate the center and memberships by applying the fuzzy-c means algorithm in Eq. (2) and Eq. (1) respectively. Consequently, we obtain the following center and memberships as:

$C_1 = 3.27 \quad , C_2 = 8.5 \quad , C_3 = 54.67 \quad , and \quad C_4 = 99.6.$

$u_1 = (3.23*10^{-3},\ 0.2285,\ 0.01,\ 0.229,\ 5.57*10^{-3},\ 0.01,\ 3.23*10^{-3},$
$0.031,\ 7.3*10^{-4}\ 3.23*10^{-3},\ 5.57*10^{-3},\ 0.01,\ 0.229,$
$2.23*10^{-3},\ 0.229).$

$u_2 = (0.133,\ 0.37,\ 0.37,\ 0.133).$

$u_3 = (0.185,\ 0.015,\ 0.03,\ 0.728,\ 0.024,\ 0.0294).$

$u_4 = (8.19*10^{-4},\ 0.465,\ 3.495*10^{-3},\ 0.026,\ 1.71*10^{-3},$
$1.71*10^{-3},\ 0.013,\ 0.465,\ 2.53*10^{-3},\ 1.787*10^{-3},\ 1.34*10^{-3}).$

$V_{13} = (0.185)^2\|56-54.67\|^2 + (0.015)^2\|50-54.67\|^2 + (0.03)^2\|59-54.67\|^2 +$
$(0.728)^2\|54-54.67\|^2 + (0.024)^2\|51-54.67\|^2 + (0.0294)^2\|58-54.67\|^2$
$= 2.006.$

Step 5: According to algorithm 3, we select first two regions R_4 and R_3 and compute

$V_{14} = (8.19*10^{-4})^2\|90-99.6\|^2 + (0.465)^2\|100-99.6\|^2 +$
$(3.495*10^{-3})^2\|95-99.6\|^2 + (0.026)^2\|98-99.6\|^2 +$
$(0.013)^2\|102-99.6\|^2 + (1.71*10^{-3})^2\|93-99.6\|^2 +$
$(0.465)^2\|100-99.6\|^2 + (2.53*10^{-3})^2\|105-99.6\|^2 +$
$(1.787*10^{-3})^2\|106-99.6\|^2 + (1.34*10^{-3})\|107-99.6\|^2$
$= 0.0818$

$R_{43} = (90,\ 100,\ 95,\ 98,\ 93, 102,\ 100,\ 105,\ 106,\ 107,\ 56,\ 50,\ 59,\ 54,\ 51,\ 58)$

$C_{43} = 82.75.$

$u_{43} = (0.306,\ 0.054,\ 0.1074,\ 0.0693,\ 0.154,\ 0.0433,$
$0.054,\ 0.0325,\ 0.029,\ 0.027,\ 0.02,\ 0.15,$
$0.028,\ 0.0195,\ ,\ 0.016,\ 0.0263)$

$V_{43} = (0.306)^2\|90-82.75\|^2 + (0.054)^2\|100-82.75\|^2 +$
$(0.1074)^2\|95-82.75\|^2 + (0.069)^2\|98-82.5\|^2 +$
$(0.154)^2\|93-82.75\|^2 + (0.043)^2\|102-82.75\|^2 +$
$(0.054)^2\|100-82.75\|^2 + (0.0325)^2\|105-82.75\|^2 +$
$(0.09)^2\|106-82.75\|^2 + (0.027)^2\|107-82.75\|^2 +$
$(0.02)^2\|56-82.75\|^2 + (0.15)^2\|50-82.75\|^2 +$
$(0.029)^2\|59-82.75\|^2 + (0.0193)^2\|54-82.75\|^2 +$
$(0.06)^2\|51-82.75\|^2 + (0.0263)^2\|58-82.75\|^2 = 43.343$

$V_1 = \|99.6-54.67\|^2 * (8.18*10^{-3} + 0.354) = 691.43$

$V_2 = \frac{1}{16}\|107-59\|^2 * 43.343 = 6241.45$

From the previous calculation, we note that $V_2 > V_1$. Therefore, the regions R_4 and R_3 cannot be merged. Repeat step 5 for the new regions $R_3$and R_2. For the two regions R_3 and R_2, $V_1 = 1160.339$, *and* $V_2 = 13146.624$. From the previous calculation, we note that the two regions R_3 and R_2 cannot be merged, similarly, the algorithm selects the next region R_1 and ignore R_3. Repeat step 5 for the new regions $R_2$and R_1. For the two regions R_2 and R_1, $V_1 = 5.77$ and $V_2 = 0.641$. We note that $V_1 > V_2$. from calculations and the two regions R_2 and R_1can be merged into R_{21}. Finally the image F are segmented into three regions and can be highlighted into three different colors as shown in fig.(3b).

1	3	2	90	100
3	5	2	95	98
1	4	8	93	102
56	50	59	11	10
54	51	58	7	6
1	5	2	100	105
3	6	3	106	107

(a)

1	3	2	90	100
3	5	2	95	98
1	4	8	93	102
56	50	59	11	10
54	51	58	7	6
1	5	2	100	105
3	6	3	106	107

(b)

Fig. 3: (a) Original image. (b) Final clustering.

6.2 Experimental with Real Data Set

In this section, the proposed algorithm and the existing methods (such as PC, CE, and XB) have been applied to two simulation and one real life data. First, the proposed method has been applied to the data which has three clusters in real. Moreover, the values of the other cluster validity indexes are obtained and the results are presented in Table (1). When Table (1) is studied, the most appropriate number of clusters for PC, CE and XB criteria is 3. When the results obtained by the proposed method column are studied, it can be seen that the output number of clusters is 3.

Second, the proposed method for the simulation value with four clusters in real is applied. Moreover, the values of the other cluster validity indexes are obtained and the results are presented in Table (2). According to Table (2), the most appropriate number of clusters for PC, CE and XB criteria is 4. Of course the proposed result is obtained 4 clusters.

Lastly, the proposed method is applied to the synthetic data with five clusters which is a real life data. Besides, the values of the other cluster validity indexes are obtained and the results are summarized in Table 3. According to Table (3), the most appropriate number of clusters is 18 for PC criterion, seven for CE criterion, and four for XB criterion. When the proposed method appoints the number of clusters correctly, PC, CE and XB criteria make the wrong choice. The proposed method appoints the most appropriate cluster number correctly. It is shown that the proposed method gives true number of clusters in nearly all the data sets, especially those of high number of clusters.

Table 1. Results of simulation data with three clusters.

Number of clusters	Cluster validity indexes			
	PC	CE	XB	The proposed method
2	0.8167	0.2765	1.4454	3
3	0.9992	0.0034	49.2682	3
4	0.8996	0.1562	43.2621	3
5	0.8808	0.1954	24.8372	3

Table 2. Results for the simulation data with four clusters.

Number of clusters	Cluster validity indexes			
	PC	CE	XB	The proposed method
5	0.931	0.1129	6.7532	4
4	0.9988	0.0053	7.5563	4
3	0.8034	0.3597	1.9668	4
2	0.7121	0.4626	0.08065	4

Table 3. Results for the synthetic data with five clusters.

Number of clusters	Cluster validity indexes			
	PC	CE	XB	The proposed method
18	0.999	0.021	10.6709	5
7	0.9262	0.0002	25.4019	5
6	0.966	0.0555	14.7612	5
5	0.998	0.00049	26.0438	5
4	0.9265	0.1702	35.9476	5
3	0.9	0.2286	1.1935	5
2	0.6949	0.471	0.99	5

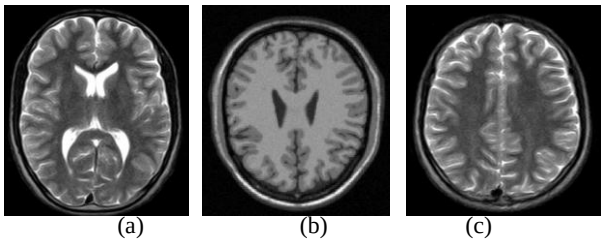
6.3 Experimental with Medical Images

The experiments were performed on medical data such as data1, data2, and data3 while the segmentation of such images is the challenge. The image size of these data is 384×512 pixels, as shown in Fig. 4(a). We used a high-resolution T1-weighted MR phantom with slice thickness of 1mm, 3% noise and 20% inhomogeneity, obtained from the classical simulated brain database of McGill University Brain Web. The parameters of these algorithms are presented as follows; for $T_{h_{\min}}$ =20, $h_{\min}$ =1, h_{max}= 255, mask 3X3, *diff* = 2. A multi-level by immersion is applied on the 4-connected grid. The quality of the segmentation algorithm is of vital importance to the segmentation process. The comparison score *S* for each algorithm is proposed in [6], which defined as:

$$S = \left| \frac{A \cap A_{ref}}{A \cup A_{ref}} \right|$$

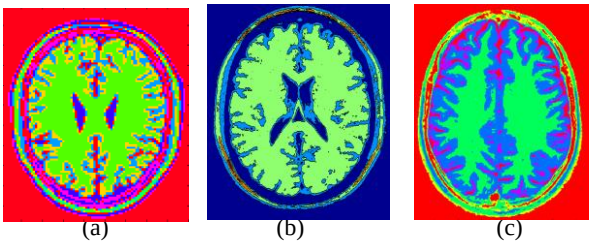
where *A* represents the set of pixels belonging to a class as found by a particular method and A_{ref} represents the set of pixels belonging to the very same class in the reference segmented image (ground truth).

The proposed algorithm is performed for each data image using iterative fuzzy c-means algorithm in Eqs.(1), (2). The number of clusters are obtained six clusters for data1 image with accuraccy 0.823 as shown in Fig.(5a). Seven clusters are got for the data2 and data3 images with accuracy 0.669 and 0.743 respectivelly as shown ing Figs. (5b)-(5c). These results prove that the proposed method achieved highly accurate results in medical image segmentation.



(a) (b) (c)

Fig. 4. (a) Data1, (b) Data2, and (c) Data 3 are MRI images.



(a) (b) (c)

Fig. 5. Segmented image. (a) Six clusters for data1, (b) and (c) Seven clusters for data1 and data2 images.

7. CONCLUSION

In clustering analysis, determining the most appropriate number of clusters in order to reach accurate and sound results is an important problem. In some complicated data, because of the uncertainty of some cluster members, cluster validity indexes can give conflicting results in determining the most appropriate number of clusters. In this study, an alternative reliable validity index algorithm has been proposed that could improve the image clustering. The proposed method has been tested with discrete image example to show the applicability of this method. Also, it has compared with the results obtained from cluster validity indexes such as PC, CE, and XB. The proposed method is applied to two simulation and one real life data. In the results obtained for the simulation data, the criteria which are PC, CE, XB and the proposed method is appointed the appropriate number of clusters correctly. For the real life data called synthetic data, it is shown that only the proposed method appoint the most appropriate number of clusters correctly. Furthermore, the proposed method has been experimented with different brain images. The accuracy of the obtained clusters is good and encouraging. As a result of the applications, it can be seen that the most appropriate number of clusters can be appointed in fuzzy clustering with the proposed method.

Overall, the proposed method has given more stable results in all tests and yielded satisfactory results, which are more compatible with medical image segmentation perception.

8. REFERENCES

[1] J.C. Bezdek, “Pattern recognition with fuzzy objective function algorithms”, Plenum Press, New York, 1981.

[2] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag, T. Moriarty, “A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data”, IEEE Trans. Med. Imag. 21, pp. 193-199, 2002.

[3] U. Maulik and S. Bandyopadhyay, “Fuzzy partitioning using a real coded variable length genetic algorithm for pixel classification”, IEEE Transactions Geoscience and Remote Sensing, vol. 41, no. 5, pp. 1075– 1081, 2003.

[4] S .Wu, AWC. Liew, H. Yan, “Cluster analysis of gene expression data based on self-splitting and merging competitive learning”, IEEE Trans. on Information Technology in Biomedicine, Vol. 8, pp. 5-15, 2004.

[5] L. Zhu, F. L. Chung, S. Wang, "Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions ", IEEE Transactions on, vol. 39, no. 3., pp. 578-591, 2009.

[6] Dao-Qiang Zhang, Song-Can Chen, "A novel kernelized fuzzy C-means algorithm with application in medical image segmentation", Artificial Intelligence in Medicine Vol. 32, pp. 37-50, 2004.

[7] Jiayin Kang, Lequan Min, Qingxian Luan, Xiao Li, Jinzhu Liu, " Novel modified fuzzy C-means algorithm with applications", Digital Signal Processing 19, 309–319, 2009.

[8] D. W. Kim, K. Y. Lee, D. Lee, K. H. Lee, "A kernel-based subtractive clustering method", Pattern Recognition Letters vol.26(7), pp. 879-891, 2005.

[9] E.A. Zanaty, Sultan Aljahdali, Narayan Debnath, "A kernelized fuzzy C-means algorithm for automatic Magnetic Resonance Image Segmentation", Journal of Computational Methods in Science and engineering (JCMSE), pp. 123-136, 2009.

[10] H. Timm, C. Borgelt, C. Doring, R. Kruse, "An extension to possibilistic fuzzy cluster analysis", Fuzzy Sets and Systems", vol. 147, no. 1, pp. 3–16, 2004.

[11] J. S. Zhang, Y. W. Leung, "Improved possibilistic c-means clustering algorithms", IEEE Transactions On Fuzzy Systems, vol. 12, no. 2, pp. 209–17, 2004.

[12] Ze-Xuan Ji, Quan-SenSun, De-ShenXia, "A modified possibilistic fuzzy c-means clustering algorithm for biasfield estimation and segmentation of brain MR image", Computerized Medical Imaging and Graphics, Computerized Medical Imaging and Graphics xxx (2010) xxx–xxx.

[13] G. Yuhua, O. H. Lawrence, "Kernel based fuzzy ant clustering with partition validity", IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[14] Z. Volkovich, Z. Barzily, L. Morozensky, "A statistical model of cluster stability", Pattern Recognition 41, pp. 2174 – 2188, 2008.

[15] M. K. Pakhira, S. Bandyopadhyay, U. Maulic, "Validity index for crisp and fuzzy clusters", Pattern Recognition, vol.37, pp.487–501, 2004.

[16] K. Malay, Pakhiraa, B. Sanghamitr, U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification", Fuzzy Sets and Systems, vol.155, pp.91–214, 2005.

[17] L.Jegatha Deborah, R.Baskaran, A.Kannan,"Survey on internal validity measure for cluster validation", International Journal of Computer Science and Engineering Survey (IJCSES) Vol.1, No.2, 2010.

[18] N. Alp Erilli, Ufuk Yolcu, Erol Egˇriogˇlu , Ç. Hakan Aladag, Yüksel Öner, "Determining the most proper number of cluster in fuzzy clustering by using artificial neural networks", Expert Systems with Applications 38, pp. 2248–2252, 2011.

[19] M.T.El-Melegy, E.A.Zanaty, Walaa M.Abd-Elhafiez, Aly Farag, "On cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation", IEEE international conference on computer vision, vol. 6, VI 5-8, 2007.

[20] Y. Xu, G. Richard, and A. Brereton, "A comparative study of cluster validation indices applied to genotyping data", Chemometrics and Intelligent Laboratory Systems, vol. 78, pp. 30–40, 2005.

[21] K.L. Wu, and M.S. Yang, "A cluster validity index for fuzzy clustering", Pattern Recognition Letters, vol. 26, pp.1275–1291, 2005.

[22] Do-Jong K.,Young-woon P.,and Dong-Jo P.,"A noval validity index for determination of the optimal number of clusters". IEICE Trans. Inf.&Syst. ,Vol. E84-D,No.2,pp. 281-285,2001.

[23] Bezdek." Numerical taxonomy with fuzzy sets". Journal of MathematicalBiology, Vol.1,pp. 57–71,1974.

[24] Xie, L., &Beni, G. ,"A validity measure for fuzzy clustering". IEEE Transactions Pattern Analysis and Machine Intelligence, Vol.13(4),pp. 841–846, 1991.

[25] Maria F., Gabriella S. B.,."Oversegmentation reduction in watershed-based gray-level image segmentation." Int. J. Signal and Imaging Engineering. Vol.1,pp. 4-10,2008.

[26] Kaczynski,K.,Mikolajczak,P.." Information theory based medical image processing". OPTO-Electronics Review. Vol.11,pp.253-259,2003.

[27] Shaannon,C.E.,."A mathematical theory of communication". The Bell system Technical Journal. Vol. 27,pp. 379-423,623-656,1948.

[28] Cho, S. B., & Yoo, S. H. "Fuzzy Bayesian validation for fuzzy clustering of yeast cell-cycle data". Pattern Recognition,2005.

[29] Rezaee, M. R., Lelieveldt, B. P. F., & Reiber, J. H. C. " A new cluster validity index for the FCM". Pattern Recognition Letters, Vol.19,pp. 237–246,1998.

[30] Rhee, N. S., & Oh, K. W. " A validity measure for fuzzy clustering and its use in selecting optimal number of clusters". IEEE International Conference on Fuzzy Systems, Vol. 2, pp.1020–1025,1996.