

Medical Image Segmentation using Improved Mountain Clustering Technique Version-2

Nishchal K. Verma¹, Abhishek Roy² and Shantaram Vasikarla³, SMIEEE

¹Department of Electrical Engineering, Indian Institute of Technology Kanpur, 208016 Kanpur, India

²Department of Electrical and Electronics Engineering, National Institute of Technology Karnataka, Surathkal, 575025 Mangalore, India

³School of Information Technology, American Intercontinental University, Los Angeles, CA 90066 USA

E-mail IDs: nishchal@iitk.ac.in, abhishekroyn@gmail.com, svasikarla@la.aiuniv.edu

Abstract: This paper proposes Improved Mountain Clustering version-2 (IMC-2) based medical image segmentation. The proposed technique is a more powerful approach for medical image based diagnosing diseases like brain tumor, tooth decay, lung cancer, tuberculosis etc. The IMC-2 based medical image segmentation approach has been applied on various categories of images including MRI images, dental X-rays, chest X-rays and compared with some widely used segmentation techniques such as K-means, FCM and EM as well as with IMC-1. The performance of all these segmentation approaches is compared on widely accepted validation measure, Global Silhouette Index. Also, the segments obtained from the above mentioned segmentation approaches have been visually evaluated.

Key Words: Clustering, Improved, Medical Image, Segmentation.

1. Introduction

Image segmentation is well-known for its applications in exploratory pattern analysis [1], grouping, decision-making and machine-learning situations for medical images. Though, information (e.g., statistical models) available about the data in such problems is generally very little, the decision-makers are expected to make as few assumptions about the data as possible. Under these restrictions clustering methodology is quite appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure. Using certain properties of an image such as color and texture, image segmentation can be accomplished for medical images.

Clustering is basically collection or grouping of similar

objects. Each cluster should be homogenous, that is, objects belonging to the same group are similar to each other. Also, each cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters. Clustering technique [2] can be hard or fuzzy. In a hard clustering algorithm, each object is allocated to a single cluster during its operation and in its output, whereas, in a fuzzy clustering method a degree of membership is assigned to each object depending on its association with several other clusters.

This paper is organized into 6 Sections. An overview of widely used clustering algorithms is mentioned in Section-2. The proposed technique is discussed in Section-3. Section-4, presents the results of all the clustering techniques including the results of comparison using other clustering techniques for medical image segmentation on the basis of cluster quality and validity index. Finally, the conclusions are drawn in Section-5 and the references are listed in Section-6.

2. An overview of some clustering based segmentation techniques

Here, we discuss some of the widely used clustering techniques, such as K-means Clustering [3], FCM Clustering [4], Gath-Geva Clustering [5], EM Clustering [6], Mountain Clustering [7], Modified Mountain Clustering [8] and IMC-1 [9, 10] (for convenience IMC will be referred to as IMC-1).

K-means is one of the simplest unsupervised clustering algorithms. It is proposed by Forgy and MacQueen in 1967. K-means clusters data into a fixed number of clusters and the centroid of one cluster is placed as far away as possible from another. Each data point is associated to the nearest centroid. Fuzzy C-Means (FCM) clustering, developed by Dunn in 1973 and improved by

J. C. Bezdek in 1981, allows one piece of data to be in two or more clusters. FCM is often used in pattern recognition. In this technique, data points are bound to each cluster by means of membership functions, which give degree of association to the clusters. In Gath-Geva Clustering, the sum of weighted squared distances between the data points and the cluster centers is minimized. In this Clustering method, the weighing component in the range (0, 1) determines the fuzziness of the resulting clusters. The Expectation Maximization (EM) clustering is the statistical model based algorithm that makes use of the finite Gaussian mixtures model. In this algorithm, initially a fixed number of clusters are obtained through K-means clustering and the cluster parameters (weights, means and co-variances) are re-computed until a desired convergence value is achieved. The Probabilistic Clustering technique [11] aims at assigning a fixed component Gaussian mixture model (GMM) to the data set. It employs the basic Expectation-Maximization Algorithm, thus evaluates the probability of the components of a GMM such that all the points in the data set can be categorized into components with high probability. Yager and Filev [7] proposed a simple and easily implementable, Mountain Clustering algorithm for estimating the number and location of cluster centers. This clustering algorithm is a grid based three-step procedure. In the first step, the grid points are obtained by discretizing the hyperspace with a certain resolution in each dimension. The second step uses the dataset to form the mountain function around all the grid points. In the third step the cluster centers are generated by an iterative destruction of the mountain function. Though this method is simple, the complexity increases as the computation grows exponentially with the dimension of hyperspace. In the n -dimensional hyperspace with m number of grid lines in each dimension, the number of grid points that must be evaluated is m^n . To solve the problem of computational complexity of this clustering technique, Azeem et al. [8, 12] presented the Modified Mountain Clustering technique which determines the cluster centers by an iterative destruction of the mountain function. Nischal and Hanmandlu proposed Improved Mountain Clustering (IMC-1) Technique in which the dataset is initially normalized to range [0, 1]. In this algorithm, for every potential cluster center determined, a cluster around the center is formed and the clustered points are removed from the dataset while determining next potential cluster center. Thus, it exhibits lower computational complexity.

3. Improved Mountain Clustering Technique Version-2 (IMC-2)

In IMC-2, a heuristically determined factor ‘ α ’ is used in the threshold function for better distribution of

data points. As this factor ‘ α ’ is multiplied to the threshold function, the threshold value is optimally reduced, resulting in exclusion of unwanted data points from the clusters formed. Consequently we realized a substantial improvement in cluster quality for each of the successive clusters.

3.1 The algorithm

Step 1: Normalize each dimension of hyper-space, so that the data points are bounded by the unit hypercube.

We define the j^{th} data in \mathbf{x} hyperspace as:

$$\mathbf{x}^j = \{x_1^j, x_2^j, \dots, x_D^j\}.$$

The normalized data points $\bar{\mathbf{x}}^j$ are defined as:

$$\bar{\mathbf{x}}^j = \frac{\langle \mathbf{x}^j - (\mathbf{x}^j)_{\min} \rangle}{\langle (\mathbf{x}^j)_{\max} - (\mathbf{x}^j)_{\min} \rangle}, \quad (1)$$

$$\forall j = 1, 2, \dots, n;$$

where,

$$(\mathbf{x}^j)_{\min} = \left\{ \min_{j=1}^n x_1, \min_{j=1}^n x_2, \dots, \min_{j=1}^n x_D \right\}, \quad (2)$$

$$(\mathbf{x}^j)_{\max} = \left\{ \max_{j=1}^n x_1, \max_{j=1}^n x_2, \dots, \max_{j=1}^n x_D \right\} \quad (3)$$

and n is the total number of data points.

Step 2: Determine the threshold value d_1 for each window. d_1 is the positive constant defining the neighborhood of the data point. We compute these from the heuristics:

$$d_1 = \frac{1}{2n} \sum_j^n \left(\frac{\min(\mathbf{x}^j)}{\sum_{i=1}^D x_i^j} \right) \cdot (\alpha); \quad \alpha = \frac{M}{M+1} \quad (4)$$

where, M is the number of clusters.

Step 3: Calculate the potential value of each point using mountain function, which is a function of distance $d^2(\bar{\mathbf{x}}^r, \bar{\mathbf{x}}^j) = (\bar{\mathbf{x}}^r - \bar{\mathbf{x}}^j) Q (\bar{\mathbf{x}}^r - \bar{\mathbf{x}}^j)^t$,

between $\bar{\mathbf{x}}^r$ and all other data points.

$$P_1^r = \sum_{j=1}^n \exp \left[- \left(\frac{d^2(\bar{\mathbf{x}}^r, \bar{\mathbf{x}}^j)}{d_1^2} \right) \right], \quad (5)$$

$\forall r = 1, 2, \dots, n$.

Step 4: Select the first cluster center according to the highest value of P_1^r as:

$$\bar{c}_1 = \bar{x}^1 \Leftarrow P_1^* = \max_{r=1}^n (P_1^r) . \quad (6)$$

Step 5: Assign those data points to the first cluster whose Euclidean distance from the first cluster center is less than a threshold, d_1 i.e.

$$\text{If } d^2(\bar{x}^r, \bar{c}_1) \leq d_1, \quad \forall r = 1, 2, \dots, n ; \quad (7)$$

then \bar{x}^r is assigned to the first cluster.

Step 6: Remove all those data points from the total dataset which are assigned to the clusters formed.

Step 7: Repeat Steps 2 to 5 for the remaining data to make successive clusters. Similarly for selection of m^{th} cluster center, revision of potential value is done for the reduced dataset and m^{th} cluster center is selected with the highest value of P_m^r as under:

$$\bar{c}_m = \bar{x}^{m*} \Leftarrow P_m^* = \max_{r=1}^n (P_m^r) . \quad (8)$$

Step 8: Determine the optimum number of clusters, N_o using cluster validity measure Global Silhouette Index (GS) [13, 14] defined by:

$$GS = \frac{1}{M} \sum_{m=1}^M S_m . \quad (9)$$

S_m is calculated as:

$$S_m = \frac{1}{N_m} \sum_{i=1}^{N_m} s(i) , \quad (10)$$

where N_m is number of data points in m^{th} cluster and $s(i)$ is :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} , \quad (11)$$

where $a(i)$ is the average distance between the i^{th} data point and all of the data points included in X_m (m^{th} cluster); ‘max’ is the operator, and $b(i)$ is the minimum average distance between the i^{th} data point and all of the data points clustered in X_k ($k = 1, \dots, M ; k \neq m$). From this formula it follows that $-1 \leq s(i) \leq 1$. The value of M corresponding to the maximum value of GS obtained will be assigned as the optimum number of clusters, N_o .

Step 9: Form an optimum number of clusters N_o , using the Steps 2 to 7 and then separate out these clusters from the whole dataset. Rest of the data points are distributed among the formed clusters depending upon their Euclidean distance, i.e. nearness to the respective clusters.

3.2 Global silhouette index (GS) as a quality measure

GS can be evaluated using (9). While comparing different segmentation approaches for better quality segments, GS value should approach near to 1. For the expression (11), when $s(i)$ is close to 1, one may infer that the i^{th} data point has been “well-clustered”, or placed in an appropriate cluster. Whereas if $s(i)$ is close to zero, it suggests that the i^{th} sample could also be assigned to the nearest neighbor cluster.

4. Results

The segmentation approaches discussed in this paper have been applied on various categories of medical images, such as MRI images, dental X-rays, chest X-rays, etc. Simulations are performed on a PC with 3.00 GB RAM, using a 2.99 GHz processor. Prior to the experiments, no pre-processing is done on these images. RGB features are used in clustering. The effectiveness of clusters is compared in terms of GS values of clusters formed.

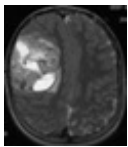
Example 1: In this example an MRI image of a person having brain tumor is segmented. Tumor is shown in the Table 1 as whitish patch in the MRI image. On performing Step 8, as shown in Figure 1, the optimum number of clusters comes out to be four. Table 1 shows the performance in terms of GS for K-means, FCM, EM, IMC-1 and IMC-2 clustering techniques. Table 2 shows the clusters formed by the above mentioned clustering techniques. From Table 2 we can easily visualize that the tumor area is clearly retrieved as a separate segment with IMC-2. EM was unable to separate the tumor. GS values show that segments formed by IMC-2 are of better quality than other segmentation approaches.

Example 2: Here we are concerned with an X-ray image showing tooth decay. The decay is shown as whitish region in the X-ray image. Using Step 8, as shown in Figure 2, the optimum number of clusters has been found to be four. Table 3 shows the performance in terms of GS for K-means, FCM, EM, IMC-1 and IMC-2 clustering techniques. Table 4 shows the segments formed by the above mentioned segmentation approaches (for proper visualization cyan color is used as background color in segments formed). From this table we can see that the decay part is properly segmented only in IMC-1 and IMC-2. Here, GS values confirm IMC-2 segments as the best, followed by FCM.

Example 3: In this example we consider segmentation of an X-ray image showing complicated pneumoconiosis in coal workers. Implementing Step-8, as shown in Figure 3 optimum number of clusters comes out to be two. The performance in terms of *GS* for K-means, FCM, EM,

IMC-1 and IMC-2 segmentation approaches are shown in Table 5. Table 6 demonstrates the clusters formed by above mentioned clustering techniques. *GS* values show that segments formed by IMC-2 are better than other approaches, closely followed by IMC-1.

Table 1. Performance of the different clustering techniques

Original Image	Clustering Method	GS
Optimum no. of Clusters=4		
	K-means	0.7859
	FCM	0.7905
	EM	0.7333
	IMC-1	0.7478
	IMC-2	0.8077

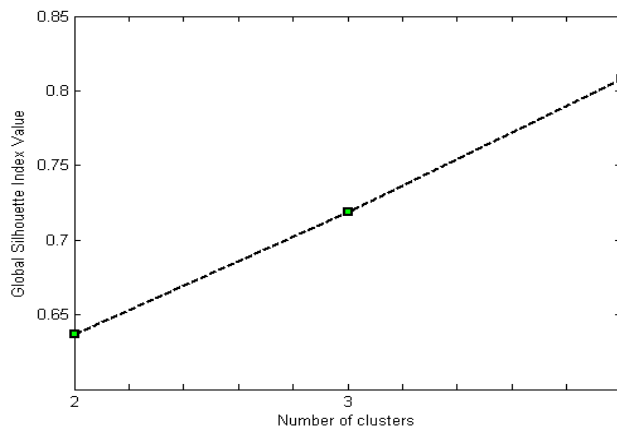


Figure 1. Variation of GS with number of clusters

Table 2. Clusters of MRI image (showing brain tumor) formed by the different clustering techniques













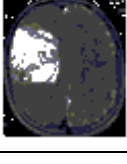


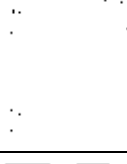





Clustering Method	1 st Cluster	2 nd Cluster	3 rd Cluster	4 th Cluster
K-means				
FCM				
EM				
IMC-1				
IMC-2				

Table 3. Performance of the different clustering techniques

Original Image	Clustering Method	GS
Optimum no. of Clusters=4		
	K-means	0.7238
	FCM	0.7344
	EM	0.6193
	IMC-1	0.7173
	IMC-2	0.7366

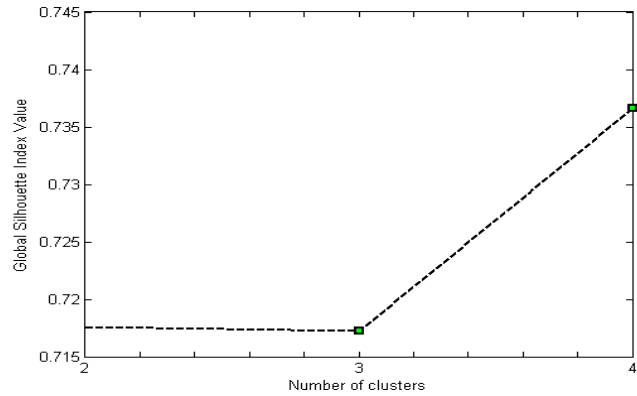


Figure 2. Variation of GS with number of clusters

Table 4. Clusters of X-ray image (showing tooth decay) formed by the different clustering techniques






















Clustering Method	1 st Cluster	2 nd Cluster	3 rd Cluster	4 th Cluster
K-means				
FCM				
EM				
IMC-1				
IMC-2				

Table 5. Performance of the different clustering techniques

Original Image	Clustering Method	GS
Optimum no. of Clusters=2		
	K-means	0.8215
	FCM	0.8260
	EM	0.7362
	IMC-1	0.8470
	IMC-2	0.8509

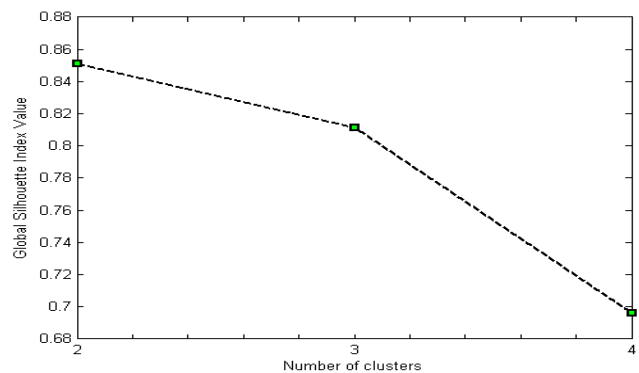












Figure 3. Variation of GS with number of clusters

Table 6. Clusters of X-ray image (showing pneumoconiosis) formed by the different clustering techniques

Clustering Method	1 st Cluster	2 nd Cluster
K-means		
FCM		
EM		
IMC-1		
IMC-2		

4.1 Comparison of performance

The results of medical image segmentation indicate that IMC-2 technique shows improved results over IMC-1. The Global Silhouette Index (GS) values for IMC-2 are the best in most of the cases. A good number of medical image segmentation examples with IMC-1 having low GS values as compared to other segmentation approaches, attains superior GS values when segmentation technique used is IMC-2 and thus can separate the disease-infected region from medical images quite clearly. K-means has a disadvantage of random selection of cluster centers which may consume more time while dealing in large medical images. It is found that FCM and IMC-1 trails behind IMC-2 performance wise though it is equally competitive at times. Most of the segmentations show the lowest GS value for EM.

5. Conclusion

This paper presents IMC-2 based medical image segmentation. We have applied some well known clustering techniques, viz., K-Means, Fuzzy C-Means, EM, IMC-1 and IMC-2 for the segmentation of medical images and compared their performance in terms of Global Silhouette Index that they achieve. We have found IMC-2 showing noticeable improvement over IMC-1. Also, it showed appreciably better results over the other techniques as well in most of the cases in view of its cluster quality and less computational complexity. Visual assessment has confirmed our findings through experiments on medical images.

6. References

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Second Edition, John Wiley & Sons Inc., 2001.

[2] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, September 1999, pp. 264-323.

[3] J.A. Hartigan, and M.A. Wong, "A K-means clustering algorithm", *Appl. Stat.*, Vol. 28, 1979, pp. 126-130.

[4] M. Razaz, "A Fuzzy C-Means Clustering Placement Algorithm", *IEEE International Symposium on Circuits and Systems, ISCAS '93*, Vol. 3, No. 6, May 1993, pp. 2051-2054.

[5] J. Abonyi, R. Babuska, and F. Szeifert, "Modified Gath-Geva Clustering for Identification of Takagi-Sugeno Fuzzy Models", *IEEE Transactions on Systems, Man, and Cybernetics-part B: cybernetics*, Vol. 32, No. 5, October 2002, pp. 612-621.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc. B*, Vol. 39, 1977, pp.1-38.

[7] R.R. Yager, and D.P. Filev, "Approximate Clustering via the Mountain Method", *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 24, No. 8, August 1994, pp. 1279-1284.

[8] M.F. Azeem, M. Hanmandlu, and N. Ahmad, "Modified Mountain Clustering and Dynamic Fuzzy Modeling", *2nd Int. Conf. on Inform. Techno.*, Bhubaneswar, India, 1999, pp. 61-65.

[9] N.K. Verma, and M. Hanmandlu, "Color Segmentation via Improved Mountain Clustering Technique", *International Journal of Image and Graphics*, Vol. 7, No. 2, April 2007, pp. 407-426.

[10] N.K. Verma, P. Gupta, P. Agarwal, M. Hanmandlu, S. Vasikarla, and Y. Cui, "Medical Image Segmentation Using Improved Mountain Clustering Approach", *6th Int. Conf. on Inform. Techno.: New Generations, ITNG'09*, Las Vegas, USA, 2009, pp. 1307-2312.

[11] M.T. Gan, A. Tan, and M. Hanmandlu, "A new Model-Selection Algorithm for a Rule-based Fuzzy System", submitted to *IEEE Trans. on Fuzzy Systems*.

[12] M.F. Azeem, M. Hanmandlu, and N. Ahmad, "Structure Identification of Generalized Adaptive Neuro-Fuzzy Inference Systems", *IEEE Trans. on Fuzzy Systems*, Vol. 11, No. 5, October 2003, pp. 666-681.

[13] P.J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *J. Comput. Appl. Math.*, Vol. 20, 1987, pp. 53-65.

[14] N. Bolshakova, and F. Azuaje, "Clustering Validation Techniques for Genome Expression Data", *Genomic signal processing*, Vol. 83, Issue 4, April 2003, pp. 825-833.