An Efficient K-means Clustering For Image Segmentation With An Application To MRI

S.P.Vimal K.Santle Camilus

Abstract

Segmentation means to classify the objects that exist in an image. One natural view of segmentation is an attempt to determine which components of data set naturally belong together. This is a problem known as clustering. Clustering techniques falls into several categories such as neural network based, center based, region growing etc. Among the existing techniques, Center based techniques is commonly used for segmentation. K Means Clustering and its variations such as Fuzzy K Means, K-Harmonic Means, and Relational K Means etc, are the examples of center based techniques. The variations are based on proper choice of cluster memberships, initial values for cluster center and the weights assigned to pixels for participating in the cluster. This paper proposes an efficient variant of k-means clustering algorithm which could produce quality segments with predictable performance. The proposed algorithm could cluster the input image in three scans. The algorithm is applied on color images for segmentation, also on MRI images to estimate the cortical gray matter volume on a semi-automatic basis. The results obtained are found to be better while comparing with results produced by the other variants of k-means, namely Hybrid 1 and Hybrid 2.

Key Words: Color Image Segmentation, k-means, MRI segmentation, Center based clustering, Relational k-means

1. INTRODUCTION

Clustering is a classification technique. Given a vector of N measurements describing each pixel or group of pixels in an image, a similarity of the measurement vectors and their clustering in the N-dimensional measurement space implies similarity of the corresponding pixels or pixel groups [3]. Clustering in measurement space may be an indicator of similarity of image regions and may be used for segmentation purposes. The vector of measurements describes some useful image feature and is also known as a feature vector. Similarity between image regions or pixels implies clustering the feature space. Once the clustering method suitable for an application is chosen, segmentation by clustering could be very useful.

Clustering techniques falls into several categories such as neural network based, center based, region growing etc. Among these methods, Center based techniques is most commonly used for segmentation purpose. K Means Clustering and its variations such as Fuzzy K Means (FKM), K-Harmonic Means (KHM) and Relational K Means (RKM) etc, are the examples of center based techniques. The variations are based on proper choice of Cluster Memberships (soft/hard) [5], Initial values for cluster center (random/mean based), Weights assigned to pixels for participating in the cluster (Fixed/varying) [1][8].

2. CENTER BASED CLUSTER ALGORITHMS

The term "Center Based Clustering" is used to refer the family of algorithms such as k-means and Gaussian expectation-maximization, since these algorithms use a number of "centers" to represent and/or partition the input

data [8]. Center based algorithms begin with a guess about the solution and refine the positions of centers until reaching a local optimum. These methods work well, but they often converge to a local minimum that is far from the global optimum. Convergence to bad local optima is related to sensitivity to initialization and is the primary problem [4].

Define a d-dimensional set of n data points $X=\{x_1,x_2,x_3,.....x_n\}$ as the data to be clustered. Define a d-dimensional set of k centers $C= \{c_1,c_2,c_3...c_k\}$ as the clustering solution that an iterative algorithm refines. A membership function $m(c_j/x_i)$ defines the proportion of data point x_i belonging to center c_j with constraints $m(c_j/x_i)>=0$ and sum of $m(c_j/x_i)$ for all c_j's must be one. Some algorithms use hard memberships meaning that $m(c_j/x_i) \in \{0, 1\}$, while the others use a soft memberships meaning that $0 \leq m(c_j/x_i) \leq 1$. One of the reasons that the k-means converge to poor solutions is due to its hard membership function [5]. However, the hard membership function makes it possible many computational optimizations that do not affect the accuracy of algorithm. A weight function $w(x_i)$ defines how much influence data point x_i has in recomputing the center parameters in the next iteration with constraint $w(x_i)>0$ [6].

Following steps mention the general framework for the family of center based clustering algorithms

Step 1

Initialize the algorithm with guessed centers C

Step 2

i. For each data point x_i, compute its membership $m(c_j/x_i)$ in each center c_j and its weight $w(x_i)$

ii. For each center c_j, recomputed its location from all data points x_i according to their memberships and weights as

$$Cj = \left(\sum_{i=1}^{n} m_{(c_j/x_i)}.w(x_i)\,x_i\right) / \left(\sum_{i=1}^{n} m_{(c_j/x_i)}.w(x_i)\right)$$

Step 3

Repeat **Step 2** until convergence

The proposed algorithms confirm to the following general framework.

1. Standard K-Means
2. Gaussian Expectation-Maximization
3. Fuzzy K-Means
4. K-harmonic Means

3. PROPOSED VERSION OF K-MEANS CLUSTERING

The K-Means algorithm remains one of the most popular clustering algorithms used in practice. K-Means is simple and fast. It works with variety of probability distributions. It has some drawbacks. The K-Means algorithm often converges to suboptimal solutions. It may take high number of iterations to converge such number of iterations cannot be determined beforehand and changes from run to run. Results may be bad with high dimensional data. The disk based implementation of K-Means, called as the RKM, overcomes these difficulties.

The algorithmic improvements by the RKM include: The centroids should be initialized using the global statistics such as global mean and co-variance of the data set [2][7]. This gives the better approximation of the centroids and avoids the sampling over the data set to pick the k random centroids. Sufficient statistics are combined with periodic M steps to achieve faster convergence. The algorithm can effectively handle transaction data by having special operations for sparse matrices. The algorithm requires only three scans over the dataset. The M step is run for every root (256x256) times per scan. The RKM is tailored

for the application on a color image for segmentation [10]. Pseudo-code for the modified version of RKM is provided below

Let nClusters denote the number of Clusters Required. The C matrix serves as the centroids matrix. The M matrix serves as the intermediate matrix, used for the purpose of frequent aggregation of the centroids.

The control routine is defined as

```
Begin
 Initialize ( );
L = √n ;
 For scan = 1 to 3 do
          For I = 1 to n do
     Estep ( );
     If (I mod (n/L) = 0 and scan = 1) then
     Mstep ( );
  End For
  Mstep ( );
 End For
End
```

Initialize ( ) function is defined as

initialize ( )

I. The C Matrix is initialized as
c[i].x=±r/5, c[i].y=±r/5, c[i].R=±r/5,
c[i].G=±r/5, c[i].B=±r/5

Where r is uniformly distributed random number in the range [0,1]

± Symbol can be chosen with 0.5 probability

II. The M matrix is initialized as m[i].x=0 , m[i].y=0 , m[i].R=0 , m[i].G=0, m[i].B=0

III. Precompute Delta as

$C[i]=c[i].x^2+ c[i].y^2+ c[i].R^2+ c[i].G^2+ c[i].B^2$

E-Step function is defined as

E-Step ( )

I. Get the current pixel in currPix

II. Calculate the distance between currPix to all the centers

III. Find the shortest distance, let it be minIndex

IV. Calculate

m[minIndex].x+=currPix.x
m[minIndex].y+=currPix.y
m[minIndex].R+=currPix.R
m[minIndex].G+=currPix.G
m[minIndex].B+=currPix.B

V. Increment n[minIndex] by one

VI. If this is the final scan, then fix the cluster membership for the current pixel as the minIndex

M-Step function is defined as

M-Step ( )

I. For all centers in the centroids matrix,calculate

i. [i].x=m[i].x/n[i]

ii. [i].y=m[i].y/n[i]

iii. [i].R=m[i].R/n[i]

iv. [i].G=m[i].G/n[i]

v. [i].B=m[i].B/n[i]

II. Precompute delta

The output image will have the number of clusters specified by the nClusters parameter.

4. EXPERIMENTAL RESULTS

In order to demonstrate the advantages of the proposed techniques, it was tested on the two applications: MRI cortical gray matter segmentation and color image segmentation. In the following sections, the results of tests are presented. The results demonstrate an accurate behavior of the algorithm on two different applications and suggest its applicability on several other medical image segmentation tasks.

A. MRI cortical gray matter segmentation

It is a hard task to estimate the volume of cortical gray matter by scanning the Magnetic Resonance Images (MRI) through naked eyes. Computer assistance is demanded in medical applications due to the fact that it

could improve the results of human interpretation in such a domain like MR scan image where the negative cases must be at a very low rate. Accurate segmentation of cortical gray matter is important for a study of central nervous system diseases such as multiple sclerosis and Alzheimer's diseases.

This proposed work presents a semi-automated method to explore the cortical gray matter volume and its position. Post processing would necessary in order to completely isolate the cortical gray matter from other layers of MRI. The algorithms presented in this paper are sufficiently simple and they are applicable to T1-weighted, T2-weighted, PD weighted feature MRI of all the planes. Some assistance from the medical experts to check the importance of this work would be necessary. The segmentation results along with the patient's information could be used for better study of central nervous diseases like multiple sclerosis and Alzheimer's diseases.

Intel Pentium 4 Processor of speed 2.4 GHZ, 256MB RAM and VC++ language is used to develop this work. The software is tested with various 65KB (256x256) original MRI. For all the cases, the final segmentation result shows the accurate isolation of cortical gray matter and its volume in terms of pixels. A sample result is provided below. Figure 1 show the original PD weighted, axial plane MRI. The proposed algorithm is applied over the original MRI with number of cluster as three and the result is provided in Figure 2.

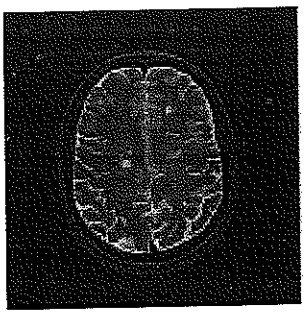

Figure 1. Original MRI

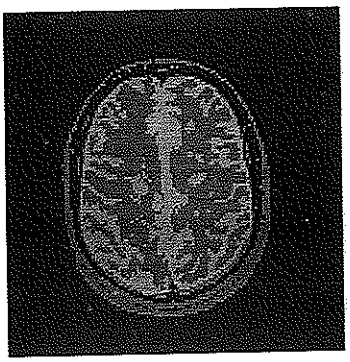

Figure 2. Segmented MRI

B. Color Image Segmentation

Color image segmentation is an important but still open problem in image processing. Considering that an image can be regarded as a dataset in which each pixel has a spatial location and a color value, color image segmentation can he obtained by clustering these pixels into different groups of coherent spatial connectivity and color.

The reason for the choice of scaling RKM towards color image segmentation is that the RKM proved to perform better for large datasets. It took only 3 scans to converge for large datasets. It is to be noted that an Image can also be considered as a five dimensional data space having the dimensions X-Coordinate, Y- Coordinate, R, G, and B. Hence a 256x256 image constitutes 256x256 points of 5 dimensions. In accordance with the dimensions the color image is segmented. A sample color image segmentation results for hybrid 1(H1), hybrid 2(H2) and scaled version of RKM are given in Figure 3a, Figure 3b and Figure 3d respectively.

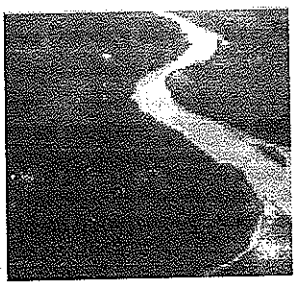

Figure 3a.Original Image

Figure 3b. Segmented image using Hybrid1

Figure 3c. Segmented image using Hybrid2

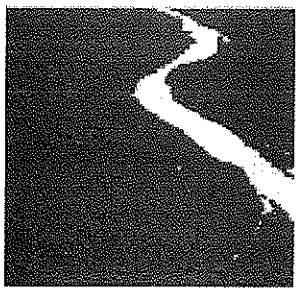

Figure 3d. Segmented image using scaled RKM

It is observed that the time taken to converge for varying cluster is having a linear relationship. The scaled version of RKM converges and produces comparable segmentation with H1 and H2. The Performance difference can be noted from Table 1.

5. CONCLUSION

Clustering in high dimensions has been an open problem for many years. Recent research shows that it may be preferable to use dimensionality reduction technique before clustering and use a low dimensional clustering algorithm such as k-harmonic means rather than clustering in the high dimension directly. The algorithms such as KHM, H1 and H2 proved to cluster well. But with the image being considered as a five dimensional data set, their performance in terms of time taken to converge is unacceptably low. The initialization of centroids has got some little influence on the convergence as well. On scaling the RKM towards Image segmentation, the convergence could be assured to be in three scans over the image, whose performance is linear in the number of points, dimensionality and desired number of clusters. The quality of clustering is also comparable to that GEM, KHM, H1 and H2. Hopefully the scaled version of RKM will offset the low performance with the most other center based techniques.

Algorithm	Number Of clusters(k)					
	2	**4**	**8**	**16**	**32**	**64**
Hybrid 1	220.6	221.3	222.2	225.9	232.0	240.0.
Hybrid 2	192.3	193.4	195.5	198.0	199.0	199.9
Scaled RKM	23.3	23.7	24.1	24.1	24.1	24.7

Table 1: Performance table in terms of seconds for Random centers

The Relational K Means Clustering Algorithm, on scaling for images, proved to perform well, producing acceptable segmentation on images. It could converge in three scans. The basic framework of k-means says, "Repeat grouping until convergence". The RKM shifts this basic framework into "Do the grouping for three scans". This shift in framework assures predictable performance. There are lot other areas where the segmentation task remains the primary activity, especially in medical diagnostics. It would be the better direction exploring the possibilities of extending this work towards brain tumor detection, by making an exclusive scaling.

References

[1] Paul S. Bradley , Usama M. Fayyad, *Refining Initial Points for K-Means Clustering*, Proceedings of the Fifteenth International Conference on Machine Learning, p.91-99, July 24-27, 1998

[2] Sanjoy Dasgupta, *Experiments with Random Projection*, Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, p.143-151, June 30-July 03, 2000

[3] P. Drineas , Alan Frieze , Ravi Kannan , Santosh Vempala , V. Vinay, *Clustering in large graphs and matrices*, Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms, p.291-299, January 17-19, 1999, Baltimore, Maryland, United States .

[4] Annaka Kalton , Pat Langley , Kiri Wagstaff , Jungsoon Yoo, *Generalized clustering, supervised learning, and data assignment*, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, p.299-304, August 26- 29, 2001, San Francisco, California

[5] M. Kearns, Y. Mansour, and A. Y. Ng. *An information-theoretic analysis of hard and soft assignment methods for clustering*. In Proceedings of Uncertainty in Artificial Intelligence, pages 282—293. AAAI, 1997.

[6] A. Likas, N. Vlassis, and J. Verbeek. *The global k-means clustering algorithm*. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands, February 2001. IAS-UVA-01-02.

[7] J. MacQueen. *Some methods for classification and analysis of multivariate observations*. In L. M. LeCam and J. Neyman, editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281—297, Berkeley, CA, 1967. University of California Press.

[8] Greg Hamerly, Charles Elkan, *Clustering algorithms: Alternatives to the k-means algorithm that find better clusterings*, ACM November 2002 Proceedings of the eleventh international conference on Information and knowledge management

[9] Carlos Ordonez and Edward Omiecinski, *Efficient Disk-Based K-Means Clustering for Relational Databases*,IEEE Transactions on knowledge and data engineering, vol. 16, no. 8, August 2004

[10] P. Bradley, U. Fayyad, and C. Reina, *Scaling Clustering Algorithms to Large Databases*, Proc. ACM KDD Conf., 1998.

Vimal received the B.E and M.E degrees in Computer science and Engineering from the Manonmaniam Sundaranar University of India in 2001 and 2005 respectively. He is now working as a senior lecturer in Narayanaguru College of Engineering, Tamilnadu, India. His research interests include Digital Image Processing and Data mining.

E-mail:vimalsp@engineer.com

Camilus received the B.Tech degree in Information Technology from the Madras University of India in 2003 and M.E degree in Computer Science and Engineering from Manonmaniam Sundaranar University of India in 2005. He is currently working as a lecturer in Department of Information Technology in National College of Engineering, Tamilnadu, India. His research interest includes Data mining and Medical Image Processing.

E-mail: Camilus@mail.com