2010 International Conference on Computational Intelligence and Communication Networks

Clustering Of Image Data Set Using K-Means And Fuzzy K-Means Algorithms

Vinod Kumar Dehariya* Shailendra Kumar Shrivastava* R. C. Jain*
I.T dept. S.A.T.I I.T dept. S.A.T.I I.T dept. S.A.T.I
Vidisha (M.P), India Vidisha (M.P), India Vidisha (M.P), India
vkdworld@yahoo.com, shailendrashrivastava@rediffmail.com, dr.jain.rc@gmail.com

Abstract - **Clustering or data grouping is a key initial procedure in image processing. In present scenario the size of database of companies has increased dramatically, these databases contain large amount of text, image. They need to mine these huge databases and make accurate decisions in short durations in order to gain marketing advantage. As image is a collection of number of pixels. It is difficult to take account of all pixels for clustering. So the concept of image segmentation play very useful role in clustering as it save times and it is efficient too. With the use of k-mean and it's variant fuzzy k-means algorithm clustering of these large data become easy and time saving.**

This paper deals with the application of standard k-means and fuzzy k-means clustering algorithms in the area of image segmentation. In order to assess and compare both versions of k-means algorithm and fuzzy k-means, appropriate procedures implemented. Experimental results point that fuzzy logically optimized k-means algorithms proved their usefulness in the area of image analysis, yielding comparable and even better segmentation results.

Keywords-clustering, k-means, fuzzy k-means, image segmentation, fuzzy logic, unsupervised learning.

1. INTRODUCTION

In recent scenario, growing attention has been put on data clustering as robust technique in data analysis. Clustering or data grouping describes important technique of unsupervised classification that arranges pattern data (most often vectors in multidimensional space) in the clusters (or groups). Patterns or vectors in the same cluster are similar according to predefined criteria, in contrast to distinct patterns from different clusters [1,2].

Possible areas of application of clustering algorithms include data mining, statistical data analysis, compression, vector quantization and pattern recognition [1,2]. Image analysis is the area where grouping data into meaningful regions (image segmentation) presents the first step into more detailed routines and procedures in computer vision and image understanding.

This paper in Section 2 briefly reviews clustering techniques. K-means algorithms described in section 3. Fuzzy logic is outlined in Section 4. The fuzzy k-means algorithms briefly described in section 5. Section 6 briefly describe Image segmentation and section 7 describe performed experiments and obtained results. Section 8 concludes the paper.

2. CLUSTERING

Clustering[1,2] is an unsupervised learning task where one seeks to identify a finite set of categories termed clusters to describe the data .Unlike classification that analyses class-labelled instances, clustering has no training stage, and is usually used when the classes are not known in advance. A similarity metric is defined between items of data, and then similar items are grouped together to form clusters. Often, the attributes providing the best clustering should be identified as well. The grouping of data into clusters is based on the principle of maximizing the intra class similarity and minimizing the inter class similarity. Properties about the clusters can be analysed to determine cluster profiles, which distinguish one cluster from another. A new data instance is classified by assignment to the closest matching cluster, and is assumed to have characteristics similar to the other data in the cluster.

A good clustering method[4] will produce high quality clusters with high intra-class similarity - Similar to one another within the same cluster low inter-class similarity - Dissimilar to the objects in other clusters The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

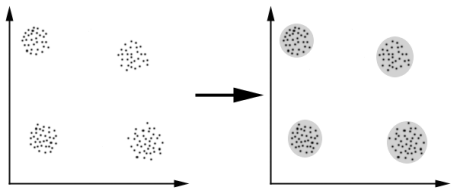


Figure: clustering

3. K-MEANS ALGORITHM

K-means (MacQueen[12][4]) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to recalculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

978-0-7695-4254-6/10 $26.00 © 2010 IEEE
DOI 10.1109/CICN.2010.80

Finally, this algorithm aims at minimizing an objective function; in this case a sum squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where $\| x_i^{(j)} - c_i \|^2$ is a chosen distance measure between a data point $x_i^{(j)}$and the cluster centre Cj, is an indicator of the distance of the n data points from their respective cluster centres.The algorithm is composed of the following steps:

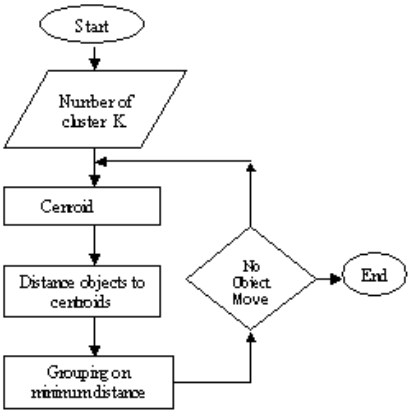


Figure : Flow-Chart of K-mean Algorithm

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect.
The algorithm is composed of the following steps:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

4. FUZZY-LOGIC

In this context, Fuzzy Logic [13] is a problem-solving control system methodology that lends itself to implementation in systems ranging from simple, small, embedded micro-controllers to large, networked, multi-channel PC or workstation-based data acquisition and control systems. It can be implemented in hardware, software, or a combination of both. Fuzzy Logic provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL's approach to control problems mimics how a person would make decisions, only much faster.

Fuzzy logic incorporates a simple rule based if X and Y then Z to a solving control problem rather than attempting to a model a system mathematically. The FL-model is empirically based, relying on an operators experience rather than their technical understanding of the system. Rather than dealing with the temperature control in terms such as “SP = 500F”, "T <1000F", or "210C <TEMP <220C", terms like “ if(process is too cool) and (process is getting colder) then (add heat to the process)” or “ if (process is too heat) and (process is heating rapidly) then (cool the process quickly)” are used. FL is capable of mimicking this type of behavior but at very high rate.

5. FUZZY K-MEANS ALGORITHM

In fuzzy clustering [15], each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. For each point x we have a coefficient giving the degree of being in the kth cluster $u_k(x)$. Usually, the sum of those coefficients is defined to be

$$\forall x \quad \sum_{k=1}^{\text{num. clusters}} u_k(x) = 1.$$

With fuzzy k-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$\text{center}_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}.$$

The degree of belonging is related to the inverse of the distance to the cluster center:

$$u_k(x) = \frac{1}{d(\text{center}_k, x)},$$

then the coefficients are normalized and fuzzyfied with a real parameter m > 1 so that their sum is 1. So

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)} \right)^{2/(m-1)}}.$$

For m equal to 2, this is equivalent to normalising the coefficient linearly to make their sum 1. When m is close to 1, then cluster center closest to the point is given much more

weight than the others, and the algorithm is similar to k-means.

The fuzzy k-means algorithm is very similar to the k-means algorithm:

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ε, the given sensitivity threshold)

Compute the centroid for each cluster, using the formula above. For each point, compute its coefficients of being in the clusters, using the formula above.

6. IMAGE SEGMENTATION

Image segmentation [8][9] refers to the process of partitioning a dig-ital image into multiple regions (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images.The result of image segmentation is a set of regions that collectively cover entire image, or a set of contours extracted from the image. Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture[11]. Adjacent regions are significantly different with respect to the same characteristics.

7. EXPERIMENTAL RESULTS

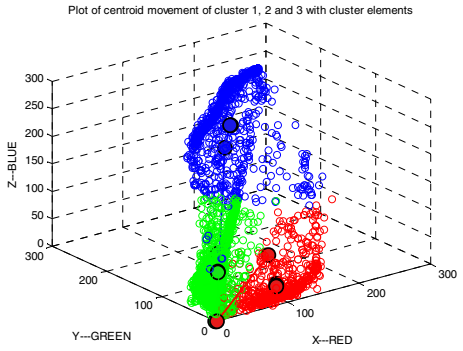


Fig-1: k-means algorithms result of image Tree1.

1(R)	0(G)	0(B)
106.15	101.63	92.632
76.156	51.76	34.772
71.83	45.797	29.366
71.224	44.47	28.602
71.131	44.195	28.476
71.043	44.113	28.437
71.043	44.113	28.437

Table-1: Centroid movement of cluster 1

0(R)	1(G)	0(B)
72.517	91.14	57.585
52.785	66.243	39.795
54.102	67.25	40.623
54.517	67.709	40.872
54.654	67.793	40.961
54.687	67.841	41.043
54.687	67.841	41.043

Table-2: Centroid movement of cluster2

0 (R)	0 (G)	1(B)
142.52	169.08	188.38
168.39	193.87	209.17
169.1	193.88	208.92
169.21	193.82	208.91
169.19	193.88	209.02
169.2	193.93	209.12
169.2	193.93	209.12

Table-3: Centroid movement of cluster3

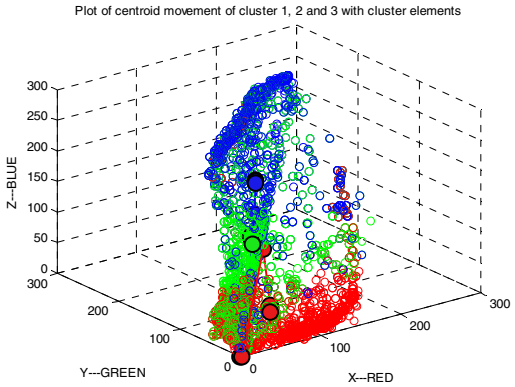


Fig-2: fuzzy k-means with membership function 0.5 of image Tree1.

1(R)	0(G)	0(B)
119.78	60.712	51.367
105.9	28.359	18.498
104.34	25.665	16.819
103.63	25.054	16.465
103.53	24.935	16.409
103.53	24.935	16.409
103.53	24.935	16.409

Table-4: Centroid movement of cluster1

0(R)	1(G)	0(B)
103.15	118.55	107.54
102.34	117.3	106.25
100.49	112.1	100.61
99.938	111.02	98.939
99.804	110.89	98.609
99.824	110.89	98.485
99.824	110.89	98.485

Table-5: Centroid movement of cluster2

0(R)	0(G)	1(B)
117.61	130.15	133.38
137.44	155.5	166.05
138.76	156.84	168.44
138.58	156.74	168.3
138.21	156.68	168.3
138.21	156.68	168.3
138.21	156.68	168.3

Table-6: Centroid movement of cluster3

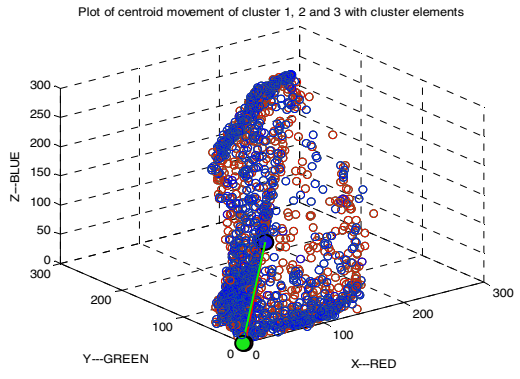


Fig-3: fuzzy k-means with membership function 1.0 of image Tree1.

1(R)	0(G)	0(B)
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521

Table-7: Centroid movement of cluster1

0(R)	1(G)	0(B)
104.19	100.87	90.59
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521

Table-8: Centroid movement of cluster2

0(R)	0(G)	1(B)
104.19	100.9	90.63
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521
104.15	100.79	90.521

Table-9: Centroid movement of cluster3

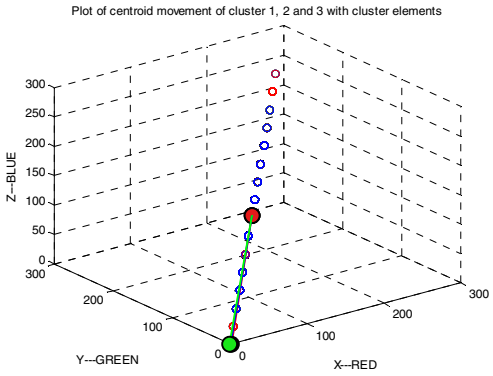


Fig-4: k-means algorithms result of image Lena.

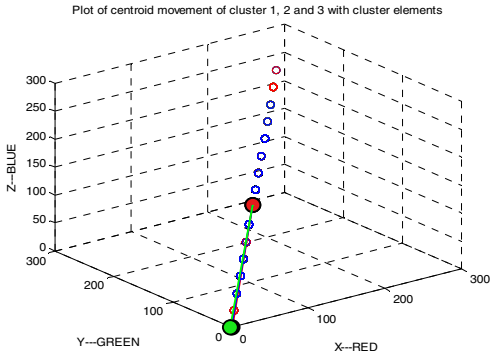


Fig-5: fuzzy k-means with membership function 0.5 of image Lena.

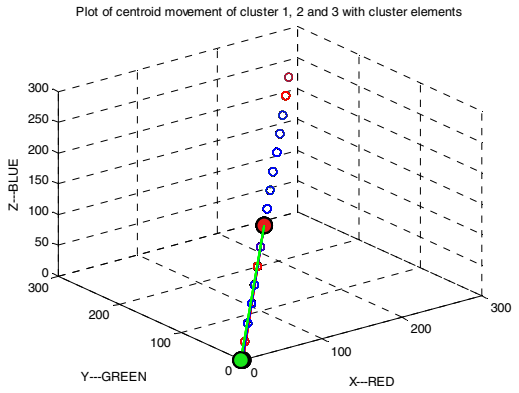


Fig-6: fuzzy k-means with membership function 1.0 of image Lena

1(R)	0(G)	0(B)
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78

*Table-10: Centroid movement of cluster1

*Table-11: Centroid movement of cluster2

0(R)	0(G)	1(B)
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78

*Table-12: Centroid movement of cluster3

*-Table –10,11,12 are common 4 fig.-4,5,6 as result obtain would be same for 2-grey level (black & white) image.

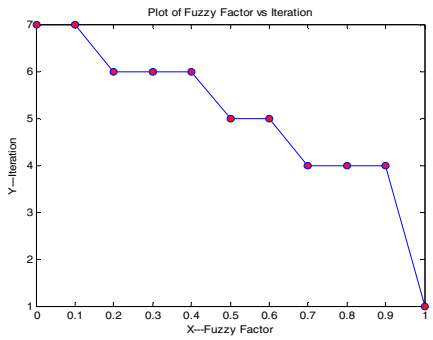
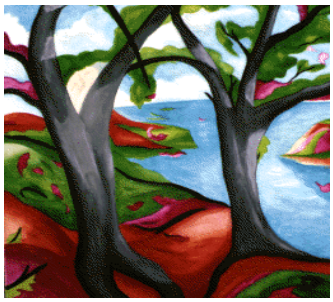


Fig-7: Shows the plot Fuzzy Factor Vs No. of Iteration



0(R)	1(G)	0(B)
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78
121.78	121.78	121.78

Fig-8:Input Image Data-Set Tree1

Fig-9: Input Image Data-Set Lena.

8. CONCLUSION AND SUMMARY

Result obtained in the performed experiments suggests that the performance of fuzzy K-means algorithms is better than the performance of K-means algorithms. As we increase the value of fuzzy factor of fuzzy k-means algorithms we get better results (show in fig-7). The result of k-means algorithms and fuzzy k-means algorithms applied to input image data set (fig-8 and fig-9) are shown in above figures and tables shows that the fuzzy k-means algorithms take less time than the k-means to cluster the image dataset ,hence performed better.

REFERENCES

[1] Jiawei Han and Michheline Kamber, “Data mining concepts and techniques”-a reference book

[2] Arun K. Pujari, “Data mining techniques”-a reference book.

[3] .Dariusz Mayszko, Sawomir T Wierzchon”Standard and Genetic k-means Clustering Techniques in Image Segmentation.” (CISIM'07)-2007.

[4] R.Xu, D.Wunsch A.Jain, M. Murty, P. Flynn, "Data clustering: A review", ACM Computing Surveys, 31, 1999,pp- 264

[5] A.Jain, M. Murty, P. Flynn, "Data clustering: A review", ACM Computing Surveys, 1999,pp- 264-323.

[6] J. Han, M. Kamber, A. Tung. “Spatial clustering methods in data mining: A survey”.,pp-188–217, 2001

[7] G.Hamerly, C. Elkan, "Alternatives to the k-means algorithm that find better clusterings", Proc. of the ACM Conference on CIKM-2002,pp-600-607.

[8] G.H.Omran, A.Salman, A.P. Engelbrecht, "Dynamic clustering using particle swarm optimization with application in image seg-mentation, Pattern & Application”2006,pp-332-344.

[9] Y. Chang, and X. Li, “Adaptive image region-growing,” IEEE Trans. On Image Processing, vol. 3, no. 6, pp. 868-872, 1994.

[10] M. Halkidi, M. Vazirgiannis, I. Batistakis, "Quality scheme assessment in the clustering process". In Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge

[11] R.H.Turi,"Clustering-based color image segmenta-tion", PhD Thesis, Monash Univ, Australia 2001.

[12] J. MacQueen.”Some methods for classification and analysis of multivariate observations”. Proc. of Berkeley Sym. on Math. and Prob., pp-281–297, 1967.

[13] Y. Wang and J. Mo, “Fuzzy logic applied in remote sensing image classification,”in *Proc. Int. Conf. Syst., Man Cybern.*, 2004, pp. 6378–6382.

[14] M.Ryoke,H.Tarnura and Y.Nakarnori.Fuzzy Rule Generation by Hyperellipsoidal Clustering. Methodologies for the Conception, Design and Application of Intelligent Syst.,pp.86-89, World Scientific,1996.

[15] YE Ping, Fuzzy K-means algorithms based on membership function improvement[J].Changchun Institute of Technology(Natural Sciences Edition),,2007,(01)