# Clustering of Image Data Using K-Means and Fuzzy K-Means

Md. Khalid Imam Rahmani[1]
[1]Associate Professor, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

Naina Pal[2], Kamiya Arora[3]
[2,3]M.Tech. Scholar, Deptt. of Computer Sc. & Engg.
Echelon Institute of Technology
Faridabad, INDIA.

***Abstract*—Clustering is a major technique used for grouping of numerical and image data in data mining and image processing applications. Clustering makes the job of image retrieval easy by finding the images as similar as given in the query image. The images are grouped together in some given number of clusters. Image data are grouped on the basis of some features such as color, texture, shape etc. contained in the images in the form of pixels. For the purpose of efficiency and better results image data are segmented before applying clustering. The technique used here is K-Means and Fuzzy K-Means which are very time saving and efficient.**

***Keywords*—*Clustering; Segmentation; K-Means Clustering; Fuzzy K-Means***

## I. INTRODUCTION

Clustering is the unsupervised classification of patterns such as observations, data items, or feature vectors into groups named as clusters [1]. Applications of clustering is growing nowadays very rapidly because it saves a lot of time and the results obtained from the clustering algorithm is very suitable for the algorithms in the later stages of the applications. Clustering basically groups the data. The data in every group is similar to each other but quiet dissimilar to the data in different groups [5]. So, the data which are grouped together are similar to each other.

Clustering has very wide range of applications in the field of research & development like in medical science, where the symptoms and cures of diseases are grouped into clusters to save time and achieve efficient results [10]. It is applied in image processing, data mining and marketing etc. In information retrieval clustering can enhance the performance of retrieving of information from the Internet considerably. All pages are grouped into clusters and optimal results are achieved.

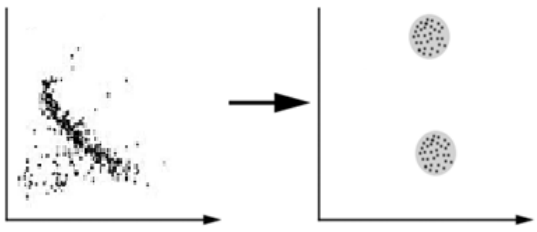

Fig. 1. Grouping of Similar Data Points

Clustering may also be used in marketing scenarios as it can segment the market into many profitable groups including advertising, promotions and follow ups etc [10]. Clustering can also be used in archeology where researchers are trying to discover stone tools, funeral tools etc. to save time in investigation surveys [10].

Image clustering can also be used in order to segment a movie [4]. Clustering is defined as unsupervised learning where user can randomly selects the data points without the help of a supervisor. There are huge applications of clustering as data clustering has proved a very powerful technique in classifying each application into clusters and sub-clusters for easy, quick and efficient results [11].

A brief description of the state of the art of clustering and various forms of clustering are given in section II. K-Means applied on image is described in section III. In section IV, an overview of existing methodologies has been described. Segmentation of images is being described in section V. In section VI, a proposed algorithm has been described. Section VII has been used for the conclusion and future direction of the research work.

## II. THE STATE OF THE ART

### A. Clustering

Clustering is a method which groups data into clusters, where objects within each clusters have high degree of similarity, but are dissimilar to the objects in other clusters. So, Clustering is a method of grouping data objects into different groups, such that similar data objects belong to the same cluster and dissimilar data objects to different clusters [9]. Clustering involves dividing a set of data points into non-overlapping groups or clusters of points where points in a cluster are "more similar" to one another than the points present in other clusters [2]. Clustering of images is done on the basis of the intra-class similarity. Target or close images can be retrieved a little faster if it is clustered in a right manner [8]. Data points in each cluster are calculated with a data points in the cluster, similar data points are brought in one cluster. So, each data points exhibits same characteristics present in one cluster.

So, a good clustering method would exhibit high similarity in a single cluster and a very less similarity with other clusters.

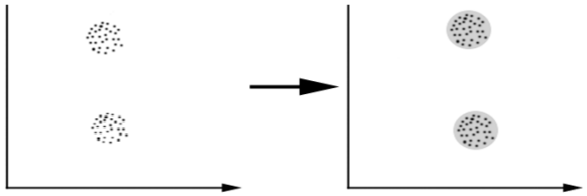

Fig. 2. Clustering of Data Points

In 1997, Haung brought the concept of k-modes which was the extension made on k-means algorithm. K-modes algorithm was introduced to cluster the numeric objects.

In 1999, Guha et al proposed a clustering algorithm of number of links between tuples. These links were used to captures the records and used to describe that which records are similar with each other. It gave satisfactory results.

In 2005, FUN and Chen presented KPSO clustering algorithm which merges some ideas of k-means and PSO. This was proposed to automatically detect the centroids of the cluster of geometric structure data sets.

In 2006, Csorba and Vajk introduced a document clustering method in which there was no need to assign all the documents to the cluster, only relevant documents were being assigned to the cluster. So, it leads to the cleaner results.

In 2007, Jing et al introduced a new k-means technique for the clustering of high dimensional data. So, different topic documents are placed with the different keywords.

### *B. K-Means Clustering*

K-Means algorithm is the most popular partitioning based clustering technique. It is an unsupervised algorithm which is used in clustering. It chooses the centroid smartly and it compares centroid with the data points based on their intensity and characteristics and finds the distance, the data points which are similar to the centroid are assigned to the cluster having the centroid. New 'k' centroids are calculated and thus k-clusters are formed by finding out the data points nearest to the clusters.

Steps of the K-Means [10] algorithm can be outlined as mentioned below:

1. Choose k number of points randomly and make them initial centroids.
2. Select a data point from the collection, compare it with each centroid and if the data point is found to be similar with the centroid then assign it into the cluster of that centroid.
3. When each data point has been assigned to one of the clusters, re-calculate the value of the centroids for each k number of clusters.
4. Repeat steps 2 to 3 until no data point moves from its previous cluster to some other cluster (termination criterion has been satisfied).

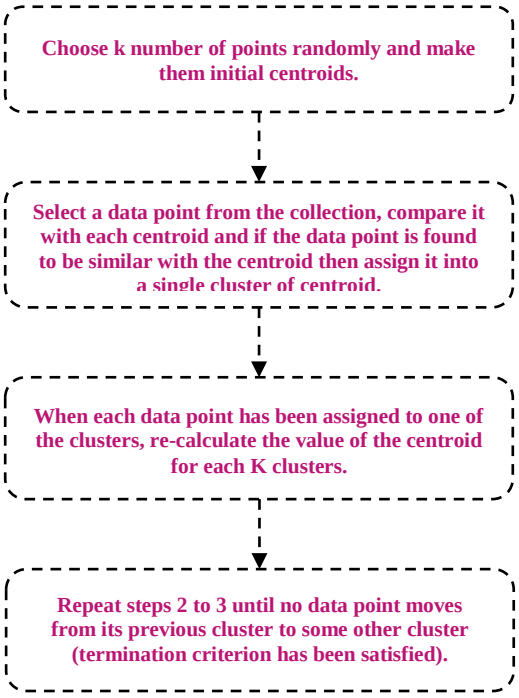


Fig. 3. K-Means Clustering Algorithm

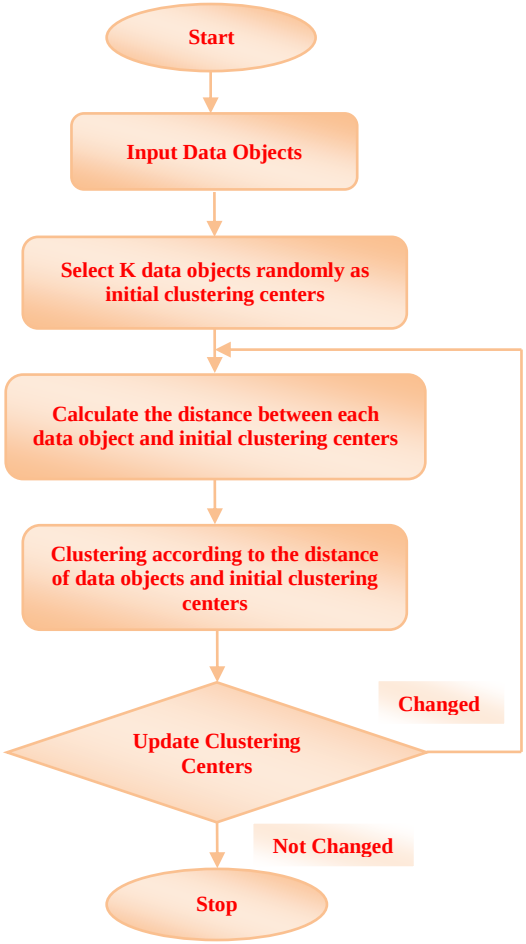


Fig. 4. K-Means Flow Chart

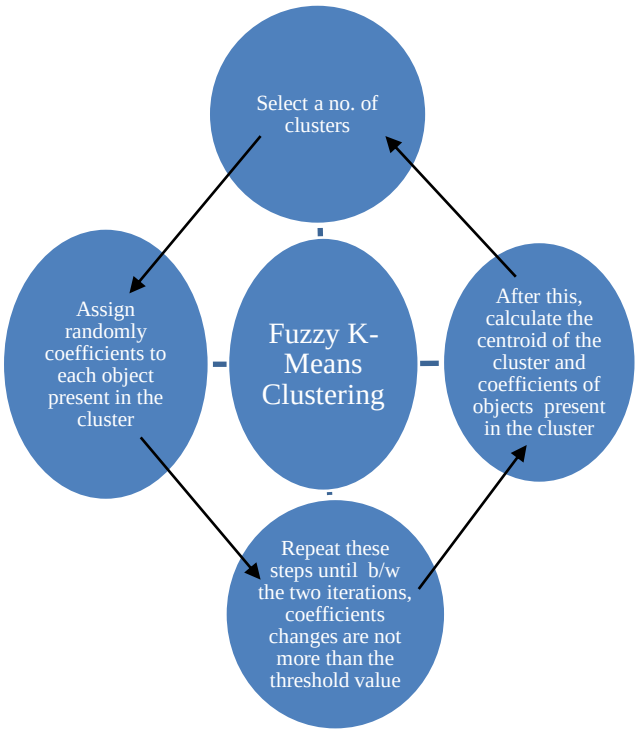


Fig. 5. Fuzzy K-Means Flow Chart

### C. Fuzzy K-Means Clustering

Each object in the fuzzy clustering has some degree of belongingness to the cluster .So, the objects that are present on the edge of the cluster are different from the objects that are present in the centroid i.e. objects on edge have lesser degree than the objects in the center. Any object p has assigned a set of coefficients that are present in the k-1th cluster $w_{k-1}(p)$. In the fuzzy c-means [1], the centroid of a cluster is the mean of all objects present in the cluster, measured by their degree of belonging of points to the cluster:

$$C_{k-1} = \frac{\sum_p w_{k-1}(p)^m p}{\sum_p w_{k-1}(p)^m}$$

## III. K-Means Applied on Image Data

Let us elaborate this by taking a example, DSR (Dynamic Spatial Reconstructor) scans left atrium and also there is a left ventricle which includes aorta and chamber [1]. Although there are valves separating the left ventricle chamber from left atria chamber and aorta still visibility is diminished because of the limitation of DSR. This disadvantage of DSR in medical image system is removed by K-means as K-means calculates the intensity of every pixel and then makes clusters. So, K-means proposed a cluster corresponding to the brightest regions would represent the left ventricle chamber and left ventricle chamber visibility becomes bright [1].

The fuzzy k-means algorithm is very similar to the k-means algorithm as depicted in figure 5 in the previous column.

## IV. Overview of Methodologies Used

Hartigan (1975) defines clustering as the group of similar data objects. The goal of clustering is to partition the data sets into several groups based on its similarities or dissimilarities i.e. entities that belong to a single group are considered to be similar to each other.

There has been much advancement done in clustering using k-means. Some of the advancements are given below in various fields of clustering:

Hierarchical clustering was introduced which is the one which makes hierarchies of cluster and these hierarchies of clusters are made to form a tree of clusters are known as dendrograms. There are two types of hierarchical clustering one is agglomerative method and the second one is divisive method. In agglomerative method, each object makes a cluster and the two most similar clusters are merged with each other and they merges iteratively up to a single cluster with the objects has been formed.

Agglomerative is based on the inter cluster similarity whereas in Divisive method, cluster is selected and slits up into many smaller clusters recursively until some termination criterion has been obtained.

Partitioning clustering introduced which splits objects into many subsets. It uses some greedy heuristics. Partitioning clustering has a drawback that many output clusters are being formulated. Berkhin (2006) describes it as a major advantage of partitioning clustering, the fact that iterative optimization may gradually improve clusters. This would results in high quality clusters. This is unlike hierarchical clustering, as algorithms in which class do not feature re-visits to clusters.

K-means clustering introduced K-Means is also known as straight K-means originated independently in the works of MacQueen (1967) and Ball and Hall (1967). Clustering came in the research since the 1960s. Factor analysis was the first related work took place by scholars (Holzinger, 1941), numerical taxonomy (Sneath and Sokal, 1973), and unsupervised learning in pattern recognition (Duda and Hart, 1973). Nowadays, clustering is used in many fields. In bioinformatics, Clustering is widely used in microarray data analysis (Allison et al., 2006) also bioinformaticians mostly use clustering because of this reason researchers have compared clustering algorithms within the field (Yeung et al., 2001), including the problem of K-means initialization. One who is working within the field of computer vision is also become the keen user of K-means, an example of the use of K-Means in this context would be to cluster the entities in an image (Szeliski, 2010) based on each pixel's features: normally their color and position. K-Means is the most popular clustering algorithm, which generates the non-overlapping clusters. It is more efficient than the hierarchical algorithms (Manning et al., 2008). K-means has been used to solve much number of problems since the 1960s. Each cluster has a centroid which is used to represent the general features of the cluster, it basically chooses any random centroid and assigns data points to the centroid by comparing distance of the data points with the centroid ,the data points which has least distance with the centroid are made to form one cluster. It

computes k-centroids by using this process and changes k-centroids values iteratively until some termination criterion has been obtained.

## V. SEGMENTATION OF IMAGES

Image segmentation has attracted considerable attention for the last few years, due to the advances in multidimensional image acquisition techniques [3].

Image segmentation is used to represent some characteristics, features from images. Segmentation operated on the images segments the images, extracts some of its important features and matches these features and matches these features with the pixels of the images. Efforts have been made to segment an entire volume (rather than merging a set of segmented slices) using supervised pattern recognition techniques or unsupervised fuzzy clustering [6]. The similar one makes one cluster and the similar cluster is dissimilar from another clusters.

## VI. PROPOSED WORK

The work which has to be done combines some ideas of image segmentation into content based image classification. In this work, the concept of image segmentation for medical images using techniques of clustering is being proposed. Retrieval of images based on segmentation and clustering of images gives better results. Here, in this it is being proposed to focus on some feature selection and moreover on classification of medical image data which is based on some of the feature selection algorithm**.**

## VII. CONCLUSION

Fuzzy k-means is better than k-means by many factors like first, it give better results when compared with k-means algorithm by increasing the fuzzy factor. Secondly, Fuzzy K-means takes lesser time to cluster the images than K-means. Thirdly, K-means is considered to be a hard clustering and in hard clustering, after some iteration most of the centers are converged to their final positions and the majority of data points has only few candidates to be selected as their closest centers where as Fuzzy K-means is known as soft clustering in which the data points which are present in the fuzzy K-means can belong to more than one cluster with having certain probability. Moreover the distance measured by the k-means is considered to be a distortion measured and the distance measured has been extended to the fuzzy K-means.

## REFERENCES

[1] A.K. Jain, M.N. Murty and P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31(3), September 1999, pp. 264-323.

[2] Vance Faber, “Clustering and the Continuous k-Means Algorithm”, Los Alamos Science, Vol. 22, 1994, pp. 138-144.

[3] Chang Wen Chen, Jiebo Luo and Kevin J. Parker, “Image Segmentation via Adaptive K-Mean Clustering and knowledge-Based Morphological Operations with Biomedical Applications”, IEEE Transactions on Image Processing, ISSN: 1057-7149, Vol. 7(12), December 1998, pp. 1673-1683.

[4] Yevgeny Seldin Sonia Starik Michael Werman, “Unsupervised Clustering of Images Using their Joint Segmentation”, pp. 1-24.

[5] Madhuri A. Tayal,M.M.Raghuwanshi, “ Review on Various Clustering Methods for the Image Data” , Journal of Emerging Trends in Computing and Information Sciences, ISSN: 2079-8407, Vol. 2, 2011, pp. 34-38.

[6] Clark MC, Hall LO, Goldgof DB, Clarke LP, Velthuizen RP, Silbiger MS, “MRI Segmentation Using Fuzzy Clustering Techniques”, IEEE Engg Medicine and Biology, ISSN: 0739-5175, Vol. 13(5), Dec 1994, pp. 730–742.

[7] Pham DL, Prince JL., “Adaptive Fuzzy Segmentation of Magnetic Resonance Image”, Vol. 18(9), Sep 1999, pp. 737-752.

[8] A.Kannan, Dr.V.Mohan, Dr.N.Anbazhagan, “Image Clustering and Retrieval using Image Mining Techniques”, 2010 IEEE International Conference on Computational Intelligence and Computing Research, 2010.

[9] Dr. Sanjay Silakari,Dr. Mahesh Motwani,Manish Maheshwari, “Color Image Clustering using Block Truncation Algorithm”, International Journal of Computer Science Issues, ISSN: 1694-0784, Vol. 4(2), 2009, pp. 31-35.

[10] P. Bradley, and U. Fayyad, “Refining Initial Points for K-Means Clustering,” In Proceeding of 15th International Conference on Machine Learning, Jan 1998, pp. 91-99.

[11] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, 2006.

## AUTHORS’ BIBLIOGRAPHY

**Md. Khalid Imam Rahmani** is an Associate Professor in the Department of Computer Science & Engg. of a very reputed NBA accredited Engineering College, Echelon Institute of Technology, Faridabad, India. He is having about 17 years of teaching, industry and administrative experience. He has done B.Sc. Engg. in Computer Engineering from A.M.U., Aligarh, M.Tech. in Computer Engineering from M.D.U., Rohtak and is pursuing Ph.D. in Digital Image Retrieval Algorithms. Digital Image Processing, Innovative Programming techniques, Mobile Computing, Algorithms Design and Internet & Web Technologies are his research areas.

**Naina Pal** has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Classification of data using Clustering and Data mining.

**Kamiya Arora** has earned her M.Tech. degree in Computer Science & Engg. from Echelon Institute of Technology under Maharshi Dayanand University, Rohtak. Her research interests include Image Processing, Steganography, Data mining and Cryptography.