IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. II (May-Jun. 2016), PP 61-64
www.iosrjournals.org

A Study on Partition Based Clustering Image Segmentation

Pradeep Sharma[1], Gayatri Duwarah[2], Barnali Kalita[3], Sanjeev Shekhar Gogoi[4]
[1]*(Computer Science & Engineering, Assam Down Town University, India)*
[2]*(Computer Science & Engineering, Assam Down Town University, India)*
[3]*(Computer Science & Engineering, Assam Down Town University, India)*
[4]*(Computer Science & Engineering, Assam Down Town University, India)*

Abstract : *This paper presents a study on partition based clustering image segmentation algorithms such as K-means, K-medoids. Before describing image segmentation let us define what an image is – An image is a 2-dimensional function (x, y), where x and y are spatial co-ordinate and the amplitude of f at any pair of co-ordinates (x, y) is called the intensity of the image at that point. Image segmentation plays a major role in computer vision. It aims at extracting meaningful objects lying in the image. In image analysis, segmentation is the partitioning of a digital image into multiple regions (set of pixels), according to some homogeneity criterion. There are wide varieties of approaches that are used for image segmentation. Different approaches are suited to different types of images. Image Segmentation is widely used in image, video, medical field and computer vision applications. Researchers are still trying to create many approaches and algorithms, out of which it is very difficult to access which algorithm will produce the best segmentation results for a particular image or a whole image. Clustering is one of the powerful techniques that have been reached in image segmentation. The cluster analysis is to partition an image data set into number of disjoint groups or clusters. In this paper we have discussed about Cluster Based Image Segmentation using partition based algorithm.*

Keywords: *Image Segmentation, Clustering, Partition Based Methods, K-means, K- medoids, Euclidean distance, Mahalanobis distance.*

I. Introduction

This paper is aimed to give a correlative study of various partitioning based clustering methods of image segmentation. Image segmentation is one of most important area of research and has opened new research prospects in this field. Image processing is a very important area that can change the outlook of many designs and proposals. The task of image segmentation is to group pixels in homogeneous regions by using common feature approach. Features can be represented by the space of color, texture and gray levels, each exploring similarities between pixels of a region. The basic goal of image segmentation is to simplify the representation of an image and change it into something which is more meaningful and easier to analyze. Clustering is one of the methods to achieve image segmentation. Data are grouped into different clusters so that data of the same group are similar and those in other groups are dissimilar. Clustering is useful to obtain interesting patterns and structures from a large set of data. Various researchers have proposed different methods to achieve clustering. Among all these methods, this paper is aimed to examine on two methods such as K-means, K-medoids– which are partitioning based clustering methods. These methods are analyzed along with their flowcharts, algorithms, pros and cons and also applications of image segmentation. Some of the practical applications of image segmentation are medical application, remote sensing, face recognition, iris recognition, fingerprint recognition, industrial inspection, optical character recognition, computer guided survey, traffic control system, machine vision, agricultural imaging – crop disease detection, locate objects in satellite images (roads, forests, etc.).

II. Clustering Method

Clustering is the most important unsupervised learning problem in data mining. We can define clustering as the process of organizing objects into groups whose members are similar based on some characteristics. Therefore a clustering is a collection of objects which are similar in the same cluster and are dissimilar to the objects belonging to other clusters. The similar characteristics of an image are grouped to form a cluster by assigning value to each feature. Many clustering methods have been proposed such as partitioning based methods, hierarchical methods, density-based methods etc. But we will confine to partitioning based clustering method only.

III. Partitioning Based Clustering Method

One of the fundamental types of cluster analysis is partitioning, where it arranges the objects or datasets into a set of groups or clusters. The partition-based methods use an iterative method, and based on a distance measure it updates the cluster of each object. In this method, suppose we are given a database of 'n'

objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k<=n which means it will classify the data into k different groups where each group contains at least one object and each object must belong to exactly one group. A distance measure is one of the feature space used to identify similarity or dissimilarity of patterns between data objects. Some of the well known methods are K-means, K-medoids. The distance measure for partition based clustering method is Euclidean distance: (1) $D=\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$,where (x1,y1) & (x2,y2) are two pixel points or two data point and Mahalanobis distance: (2) D=(x – Ci) * Inverse(S) * (x – Ci), where x is a data point in the 3-D RGB space , Ci is the center of a cluster, S is the covariance matrix of the data points in the 3-D RGB and Inverse(S) is the inverse of covariance matrix S.

3.1 K-means

K-means is one of the simplest unsupervised learning algorithms among all partitioning methods that solved the well known clustering problem. The main idea is to define k centers, one for each cluster. K-means classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance. A centroid is needed for each cluster. In K-means the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. When the processing is completed all data objects, k-means, or centroid, are recalculated, and the entire process is repeated. As a result, k clusters are found representing a set of n data objects. A flowchart for K-means method is given below.

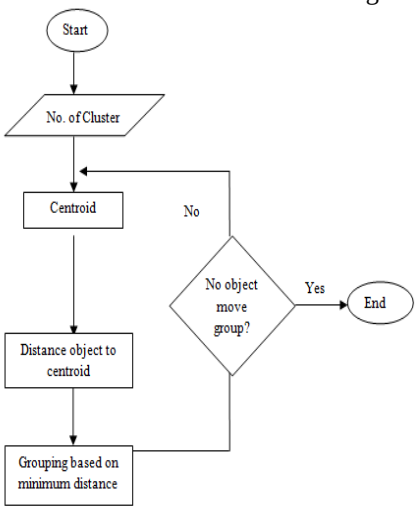

Fig1. Flowchart of K-means

Algorithm for K-means method is given below.
Input: k, D; where k is the number of clusters to be partitioned and D is a dataset containing n objects.
Output: A set of 'k' clusters based on given similarity function.

Steps:

i) Randomly choose 'k' objects from D as the initial cluster centres;
ii) Repeat,

a. (Re) assigns each object to the cluster to which the object is the most similar; based on the mean value of the objects in the cluster;
b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster;
c. iii) Until no change.

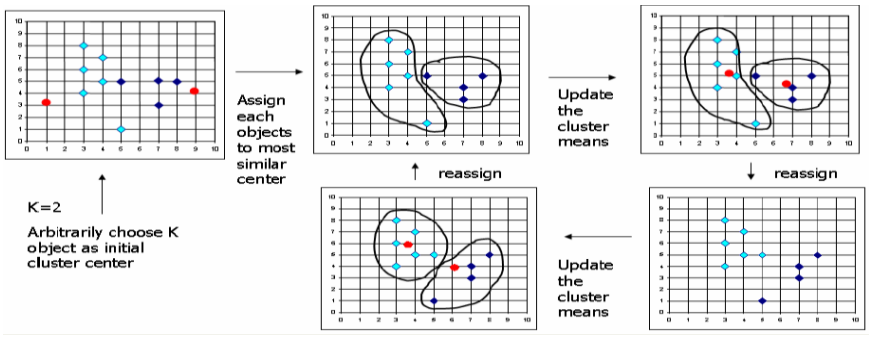

Fig2. Working of K- means algorithm.

Pros:

1. Fast, robust and easier to understand.
2. Relatively efficient; complexity is O (i*k*n), where n is total number of objects, k is total number of clusters, and i is total number of iterations. Normally k, i << n.
3. Gives best result when data set are distinct or well separated from each other.

Cons:

1. Applicable only when the mean of a cluster is defined.
2. Need to specify *k*, the total number of clusters in advance.
3. Unable to handle noisy data.
4. Result and total time depends on initial partition.

3.2 K–medoids

The *K*-means and *K*-medoids algorithms are partitioned algorithms and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. K-medoids method use medoid to represent the cluster rather than centroid. In this approach medoid is the most centrally resided data object in a cluster. The basic idea of K-medoids algorithms is to find out k clusters in n objects by finding the medoids arbitrarily for each cluster then each remaining object is clustered with the medoid based on similarity measure. In K-medoids approach representative objects are chosen as reference points instead of taking the mean value of the objects like K-means. A flowchart for K-means method is given below.

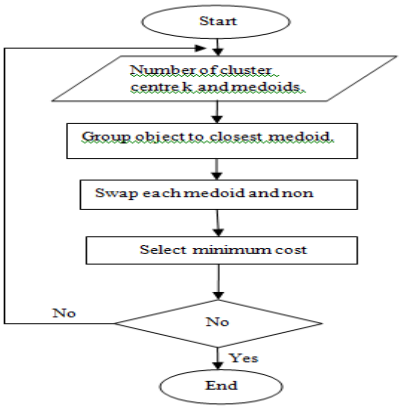

Fig3. Flowchart of K-medoids

The algorithm for K-means method is given below:

Input: k and D, where k is the number of clusters to be partitioned and D is a dataset containing n objects.

Output: A set of 'k' clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Steps:

i) Randomly choose 'k' objects as the initial medoids from dataset D;

ii) Repeat the following steps (a to d)

a. Assign each remaining object to the cluster with the nearest medoid;

b. Randomly select a non-medoid object;

c. Compute the total cost of swapping old medoid object with newly selected non medoid object.

d. If the total cost of swapping is less than zero, then perform that swap operation to form the new set of k-medoids.

iii) Until no change.

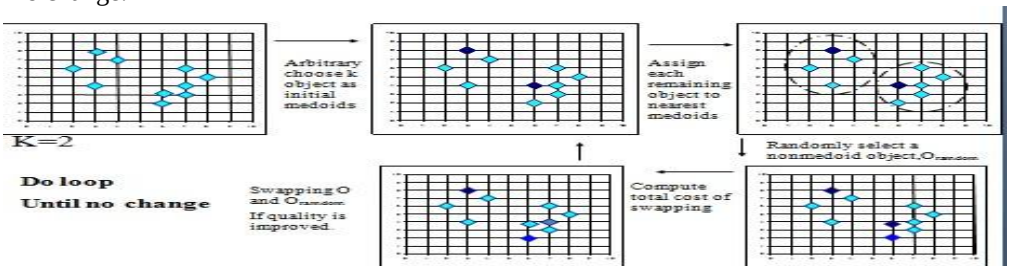

Fig4. Working of K- medoids algorithm.

Pros:

1. It is less sensitive to noise and outliers as compared to K-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances. A medoid is less influenced by outliers or other extreme values than a mean.

Cons:

1. Relatively more costly; complexity is O (i*k*(n-k) ²), where i is the total number of iterations, is the total number of clusters, and *n* is the total number of objects.
2. Need to specify *k*, the total number of clusters in advance.
3. Result and total run time depends upon initial partition.

3.3 Differences between K-means and K-medoids

Following are some of the differences between K-means and K-medoids:

- **Complexity:** computational complexity theory is a part of computer science. It looks at algorithms, and tries to say how many steps or how much memory a certain algorithm takes for a computer to do. K-means complexity is O (i*k*n) and that of K- medoids is O (i*k*(n-k) ²). The complexity of k-means depends on the dimension of the data and it can be NP-Hard problem.
- **Distance:** distance measure in K-medoids is better than K-means there is an advantage to using the pair wise distance measure in the K-medoids algorithm, instead of the more familiar sum of squared Euclidean distance-type metric to evaluate variance that we find with K-means. It reduces noise and outliers.
- **Flexibility:** K-medoid is more flexible than K-means. We can use K-medoids with any similarity measure. K-means however, may fail to converge - it really must only be used with distances that are consistent with the mean.
- **Efficiency:** K-means efficiency is more than compare to K-medoids.
- **Outliers:** K-means is sensitive to outliers while K-medoids is not. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
- **Implementation:** It is easy to implement K-means than K-medoids in any platform (e.g. MATLAB, Java etc.).

IV. conclusion

From the above study, it can be concluded that partitioning based clustering methods are suitable for clusters in small to medium sized data sets. This paper mainly focuses on the study of two different partition based clustering methods (i.e. K-means and K-medoids) for image segmentation. These methods require specifying *k*, no of desired clusters, in advance. Result and runtime depends upon initial partition for these methods. Through clustering algorithms, image segmentation can be done in an effective way. Clustering techniques also helps to increase the efficiency of the image retrieval process .The advantage of K-means is its low computation cost, while drawback is sensitivity to noisy data and outlier as mean is very much influenced by noise and outlier since in K-mean we have to calculate the mean value every time and mean value can be any value (either a data point or any value) in the given cluster. Compared to this, K-medoid is not sensitive to noisy data and outliers (since here we take a medoid that is a representative object or the main data point and medoid is less influenced by outlier), but it has high computation cost.

References

[1]. Krishna Kant Singh, Akansha Singh, A Study of Image Segmentation Algorithms for Different Types of Images, International Journal of Computer Science Issues, 7(5), 2010, 1694-0784.

[2]. M.N. Murty, A.K. Jain, P.J. Flynn, Data clustering: a review, ACM Computing Surveys. 31(3), 1999, 0360-0300.

[3]. T. Velmurugan, T. Santhanam, A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach, Information Technology Journal, 10(3), 2011, 1812-5638.

[4]. A.K. Jain, R.C. Dubes, Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice Hall, 1988).

[5]. J.Han and M.Kamber, Data Mining: Concepts and Techniques (Morgan Kaufmann Publishers, 2000).

[6]. Shalini S Singh, N C Chauhan, K-means v/s K-medoids: A Comparative Study, National Conference on Recent Trends in Engineering & Technology, 2011.

[7]. R.T. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, Proc. 20th International Conf. on Very Large Databases, Santiago, Chile,1994, 144-155620.

[8]. A.K. Jain, Data Clustering: 50 Years beyond K-Means, Pattern Recognition Letters, 31(8), 2010, 651-666.

[9]. F.MarquCs B.Marcotenui, F.Zanoguera, Partition-based image representation as basis for user assisted image segmentation, Proc. 2000 International Conf. on Image Processing, Canada, 2000. 1522-4880.