# Contribution Based Clustering Technique for Automatic Satellite Image Segmentation

Chandan Kumar Mohanty, Manish Pandey, Arun P. V.

*Department of computer science, Department of computer science, Department of GIS*
*Maulana Azad National Institute of Technology*
*Bhopal-462051, India*

***Abstract*-Clustering is an unsupervised classification that aims to classify an image into homogeneous regions. We have proposed a hierarchical content based image clustering algorithm to automatically cluster the remote sensing satellite image. The performance evaluation of this algorithm is done with reference to the LISS 4 sensor imagery of IRS-P6 satellite. Centroid of the clusters is uniformly distributed throughout the image and optimizes both inter-cluster and intra-cluster similarity metrics.**

**Keywords--Content based image clustering, remote sensing imagery, k-means algorithm.**

## I. Introduction

The huge amount of data available in living world has to be systematically organized so as to make the system able to interpret the information in a proper way. Classification in the context of image processing serves the purpose by categorizing the required information and avoiding unwanted information plays a vital role in human development [1]. Classification is broadly categorized as supervised and unsupervised classification among which the unsupervised approach is being discussed in this paper. The unsupervised classification approach investigated in this research work is called clustering, where no labeled data are available [2]. Clustering is the process of classifying the information into various groups where each group represents a cluster. The objective of clustering is to maximize intra-cluster similarity and minimize inter-cluster similarity [3] measures. Intra-cluster similarity denotes the closeness between the elements of a cluster and inter-cluster similarity denotes the similarity between all the clusters present in the image.

Cluster analysis tools are commonly used in diverse fields like Engineering( artificial intelligence, mechanical engineering, electrical engineering ), communication( mobile ad-hoc network , sensor tracking ), medical science(biology, microbiology, genetics, pathology , paleontology , psychiatry), social science( sociology , archeology , psychology, education ), economics(business, marketing ) , remote sensing and GIS [4][5]. Clustering algorithms are developed according to specific problems and it increases the probability of solving that problem.

Clustering is broadly classified into following types, as Partition clustering, Hierarchical clustering, Grid- based clustering, Density based clustering and Model based clustering [6]. Partition clustering divides the whole data set N elements into K non-empty clusters where K ≤ N. In hierarchical clustering either larger clusters are divided into smaller ones or smaller clusters are merged to larger ones. Density-based clustering adopts the splitting of elements based on their density by which high density regions are distinguished from low density ones. Grid based clustering holds both data and space around the data points where as in model based clustering some mathematical model enhance the relationship between the data.

In this paper we propose a hierarchical clustering algorithm where the data elements are partitioned at each level of hierarchical division We first consider the data set as a single cluster and moves down to divide the cluster into sub-clusters for analyzing the contribution of each data point towards newly formed centroids K-means [7, 8] algorithm considers only the intra-cluster similarity while our approach considers both intra-cluster and inter-cluster similarities.

Remaining sections is organized as follows. The next section discusses the related work and the notation used and section III outlined the proposed algorithm. Experimental results and comparison with k-means algorithm are analyzed in section IV. We conclude in section V by summarizing the investigation results and suggesting future works.

## II. Related Work

Hierarchical clustering classified into agglomerative and divisive [9] where in former the number of clusters goes on decreasing. Divisive approach increases the number of clusters at each step by splitting them according to their inter-cluster dispersion. The divisive approach is adopted in the proposed work.

Partition clustering divides the elements into clusters based on specific criteria [5]. Brute force method is computationally expensive for complete enumeration of data elements of an image. Thus heuristic approach is adopted to reduce the complexity and sum of the squared error is one of its most powerful criterion. Suppose we have N data elements, Xi ∈ I , i=1,2,...n which are divided into k clusters Cj, j=1,2,...k having individual cluster center at mj. The squared error criterion is given by

$$\text{Es}= \sum_{j=1}^{k}\sum_{i=1}^{n}(Xi-mj)^2$$

Centroid is represented as

$$\text{m}= \left(\frac{1}{n}\sum_{i=1}^{n}Xi\,1, \frac{1}{n}\sum_{i=1}^{n}Xi2\ \ldots, \frac{1}{n}\sum_{i=1}^{n}Xid\right)$$

where d is the dimension of each point.

K-means is the best known widely used [5] squared error algorithm. In this paper we use the dispersion method [10] for calculation of error between cluster elements. For a cluster C having n data points with centroid at m, the intra-cluster dispersion is shown as

$$\text{Dispersion (Cj)} =\frac{1}{n}\sum_{i=1}^{n}(Xi-m)^2$$

The contribution [10] of individual points to a cluster is measured as the difference between the dispersion excluding that point and including that point in that cluster. The contribution is represented as

Contribution ( Cj , Xi )= dispersion( Cj-[Xi] ) – dispersion(Cj)

Any data point having negative contribution to its cluster should be shifted to a cluster where it has a better contribution preferably positive. At the same time even if the data point has positive contribution, we tried to move that point to any of the cluster where it gives maximum value.

Dispersion of all points in a cluster with respect to its centroid is known as intra-cluster dispersion [10], represented as

$$A=\frac{1}{n}\sum_{i=1}^{n}(Xi-m)^2$$

Dispersion between the clusters is known as inter-cluster dispersion [10] and is given as

$$B=\frac{1}{k}\sum_{j=1}^{k}(mj-m)^2$$

## III. Algorithm

The proposed methodology for hierarchical content based satellite image segmentation is as shown in figure 1.

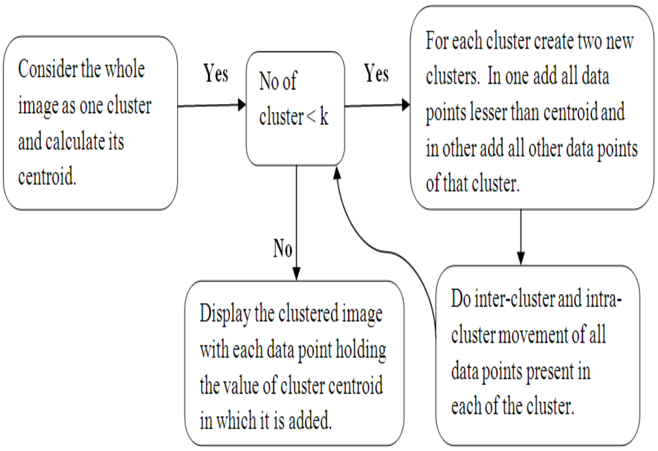

Consider the whole image as one cluster and calculate its centroid.
Yes
No of cluster < k
Yes
For each cluster create two new clusters. In one add all data points lesser than centroid and in other add all other data points of that cluster.
No
Display the clustered image with each data point holding the value of cluster centroid in which it is added.
Do inter-cluster and intra-cluster movement of all data points present in each of the cluster.

Fig1 : Automatic satellite Image Segmentation System

The detailed algorithm of the above said methodology is as given. In the algorithm the input is the data points and the number of cluster k, to which we want to divide the data points..

1. Read the image as single cluster.
2. Calculate the centroid m of the cluster.
3. Set number of cluster, O=1
4. While (O < k ) repeat steps 5 to 11
5. For P= 1 to O repeat steps 6 to 9
6. Form a temporary cluster T1 to keep all the data points below the centroid value.
7. Calculate the centroid of T1.
8. Form a temporary cluster T2 to keep all other data points of that cluster.
9. Calculate the centroid of T2.
10. Set O= 2*O
11. interdatarmovement( clusters , centroid of the cluster)
12. End

The interdatamovement function is used for inter cluster transfer of data points and the mechanism is as given

```
interdatamovement( all cluster , centroid of cluster )
    for each cluster Cj
        for each data point X present in the cluster
            if contribution( Cj , X ) < 0
                move the data point to that cluster Cnew where contribution (Cnew,X) is maximum.
                Update the cluster Cnew and centroid of Cnew.
            Else
                move that data point to a cluster Cnew where ( A-Anew ) / A + (Bnew-B)/Bnew is maximum. Update the cluster Cnew and centroid of Cnew.
            end if
        end for
    end for
```

The time complexity of interdatamovement function is O(kdN), where N is the total number of data points in the image and k is the number of clusters into which the image is segmented and d is the dimension of each point. The time complexity of our algorithm is O(ln k(Nkd)+2 $(lnk)^2$ (N+d)) and O(N+k) space complexity. The complexity nears to linear as the number of data points, N is very large compared to k and d [5].

We can triple the number of cluster at each iteration by taking a suitable variable var, and dividing the cluster into three parts. The extend of the new clusters will be from clusters lower bound to m-var/2 , m-var/2 to m+var/2 , m+var/2 to upper bound respectively. The proposed approach also adopts mixing of both double and triple cluster formation technique.

## IV. Experimental Results

The performance evaluation of above algorithm is carried out with the satellite imagery of the Bhopal city. In our experiment we divide the satellite image of Bhopal city into 2 clusters in the first iteration and each cluster to three new clusters in the second iteration.

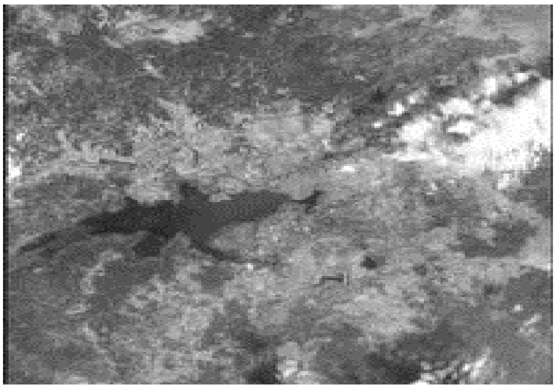


**Fig.2 Satellite image of Bhopal city**

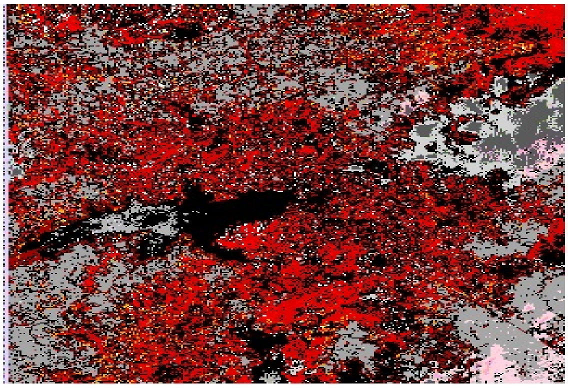
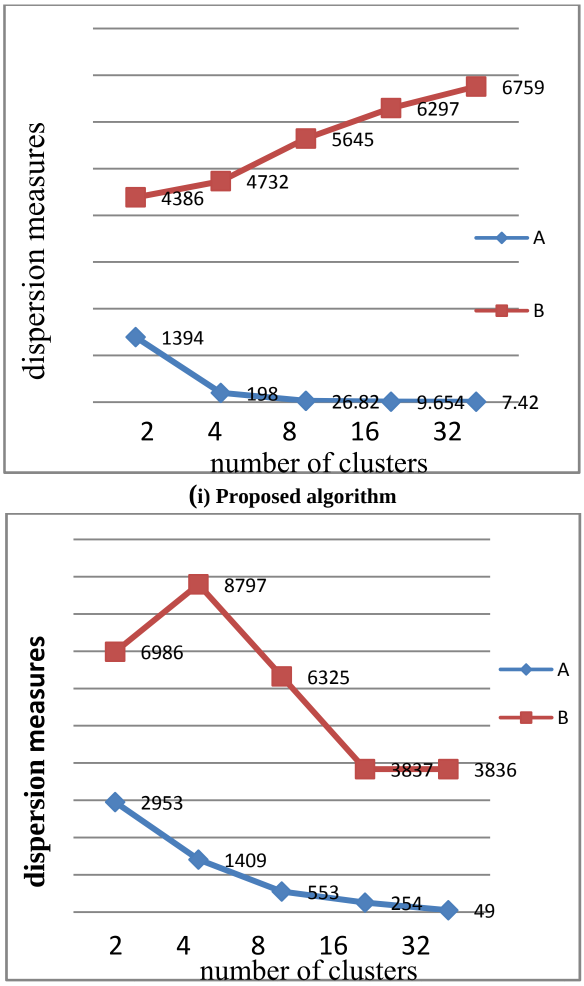


**Fig.3 Clustered image**



(i) Proposed algorithm

(ii) K-means algorithm

**Fig. 4. Value of A and B against cluster nos.**

Intra-cluster and inter-cluster dispersion values of our proposed algorithm and K-means algorithm are plotted against their cluster numbers and the results are as shown in figure[4]. From the figure we can conclude that as the number of clusters increases there is a clear trend of increase in inter-cluster dispersion and decrease in intra- cluster dispersion in our proposed algorithm. But the trend of increase in inter-cluster dispersion is not found in k-means algorithm and intra-cluster dispersion is lesser in our algorithm when compared to k-means.

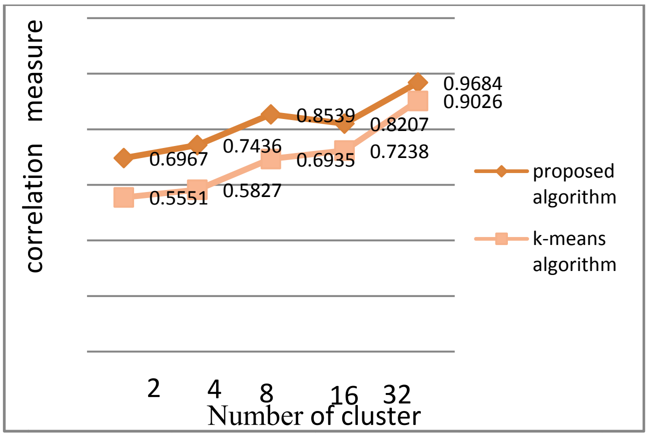


**Fig. 5 Correlation measure**

The correlation [11] measure of our algorithm and k-means algorithm with the original data set is plotted against different clusters in figure [5]. The graph shows that the image segmented by our method is more correlated to the original image when compared to the k-means generated image.

## V. Conclusion

Most of the clustering algorithms are biased on the basis of initial cluster centroid selection, as they do not have the prior knowledge about the image. In the above proposed algorithm the centroid points are selected according to the mean of the cluster and the points are uniformly distributed all over the image. At each stage of clustering special care had been taken for both inter-cluster and intra-cluster movements.

The proper selection of the cluster numbers( [12], [13] ) play a vital role in the clustering of the image. A limitation with our algorithm is that it is biased towards the more number of data points. If proper numbers of clusters are not selected then points covering fewer regions may not be properly recognized. Computational complexity of clustering is heavily dependent on the number of data points and hence our future work will be towards the parallelization of this approach. The results of the clustering can be improved by extending the work into shape, texture and other distinguished visual effects of image [14].

## REFERENCES

[1] M. Anderberg, *Cluster Analysis for Applications*. New York: Academic, 1973.

[2] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[3]. Y. Liu, D. Zhang, G. Lu, and W. Ma, "*A survey of content based image retrieval with high-level semantics*," The Journal of the Pattern Recognition Society, vol. 40, pp. 262-282,2007.

[4] B. Everitt, S. Landau, and M. Leese, Cluster Analysis. London: Arnold, 2001.

[5] R. Xu and D. Wunsch, "*Survey of clustering algorithms*," IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp. 645–678, May 2005.

[6] J. Han and K. Micheline, "*Data mining concepts and techniques*," Morgan Kauffman, 2006.

[7] E. Forgy, "Cluster analysis of ultivariate data: Efficiency vs.interpretability of classifications," Biometrics, vol. 21, pp. 68–780,1965.

[8] J. MacQueen, "*Some methods for classification and analysis of multivariate observations*," in Proc. 5th Berkeley Symp., vol. 1, 1967,pp.281–297.

[9] S. Theodoridis, K. Koutroubas. Pattern recognition, Academic Press, 1999

[10] Narasimhan, H.; Ramraj, P.; , "*Contribution-based clustering algorithm for content-based image retrieval*," *Industrial and Information Systems (ICIIS), 2010 International Conference on* , vol., no., pp.442-447, July 29 2010-Aug. 1 2010

[11] http://en.wikipedia.org/wiki/correlation

[12] V.K. Garg, "*Pragmatic data mining: Novel paradigms for tackling key challenges*," Project Report, Computer Science & Automation (CSA), Indian Institute of Science, 2009.

[13] D. Pelleg and A. Moore, "*X-means: Extending k-means with efficient estimation of the number of clusters*," in Proc. 17th Int. Conf. Machine Learning (ICML'00), 2000, pp. 727–734.

[14]F.H. Long, H.J. Zhang, and D.D. Feng, "*Fundamentals of content-based image retrieval*," in D.D. Feng, W.C. Siu, and H.J. Zhang (Eds), 'Multimedia information retrieval and management technological fundamentals and applications', Springer-Verlag, New York, 2003, pp.1–26