International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Hybrid Clustering Approach in Data Mining

Vaishali[1], Shagun[2]
[1]M. Tech. Scholar, [2]Asstt. Professor
Department of Computer Science & Engineering, MITM Hisar Haryana, India

ABSTRACT

Analysis of cluster is a descriptive assignment that perceive homogenous group of objects and it is also one of the fundamental analytical method in facts mining. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. However, its performance in terms of global optimality depends heavily on both the selection of k and the selection of the initial cluster centres. Mean Shift clustering algorithm does not rely upon a priori knowledge of the number of clusters. Therefore, mean-shift can be utilized in initial phase for finding number of clusters and k-means in second phase for proper segmentation. In this paper, the importance of two-phase approach has been studied for images with non-uniform and noisy background like ultrasound images and Infrared images.

Keywords : Image Segmentation, Clustering, K-Means, Mean Shift.

I. INTRODUCTION

discovery and it is a multivariate statistical technique which identifies groupings of the information objects based totally at the inter-object similarities computed via a designated distance metric. These measures include the Euclidean, and Minkowski distance [7]. Clustering algorithms can be classified into categories: Hierarchical clustering and Partitioned clustering [8]. The partitioned clustering algorithms, which differ from the hierarchical clustering algorithms, are normally to create a few units of clusters at start and partition the statistics into similar groups after every new release. Partitioned clustering is extra used than hierarchical clustering because the dataset may be divided into more than two subgroups in a unmarried step however for hierarchy approach, always merge or divide into 2 subgroups, and don't need to finish the dendrogram[9].

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is the computational task to partition a given input into subsets of equal characteristics. These subsets are usually called clusters. It is a main task of exploratory data mining, and a common technique for analysis of statistical data, used in many fields including image analysis, pattern recognition, machine learning, information retrieval and bioinformatics. Many variations of k-means clustering algorithm [6] have been developed recently.

The procedure of k means algorithm follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The first step is to find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. K centroids will be defined, one for each cluster. These centroids should be placed in such a way that cluster label of the images does not change anymore. Variations of k-means often include such

CSEIT1835176 | Received : 01 June 2018 | Accepted : 08 June 2018 | May-June-2018 [(3)5 : 695-699]

optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means)[7].

The mean shift procedure was originally presented in 1975 by Fukunaga and Hostetler. Basic mean-shift clustering algorithms maintain a set of data points the same size as the input data set. Initially, this set is copied from the input set. Then this set is iteratively replaced by the mean of those points in the set that are within a given distance of that point. On the other hand, *k*-means restricts this updated set to *k* points usually much less than the number of points in the input data set, and replaces each point in this set by the mean of all points in the *input set* that are closer to that point.

II. Hybrid Clustering Approach

The benefit of mean Shift over k-mean is that imply Shift clustering does no longer rely on a priori knowledge of the wide variety of clusters. Consequently, we are able to utilize suggest-shift in initial section for finding wide variety of clusters and k-means in 2nd segment for proper segmentation. On this paper, the significance of two-segment method has been studied for photos with non-uniform and noisy historical past like nano photos, ultrasound pictures and IR images.[1]

Suggest shift algorithm clusters an n-dimensional statistics sets. For every point, imply shift computes its associated height via first defining a spherical window on the statistics point of radius r and computing the imply of factors that lie inside the window. Set of rules then shifts the window to the imply and repeats till convergence [8]. At every iteration, the window will shift to a greater densely populated portion of statistics set until peak is reached where records is similarly dispensed. Mean Shift is an iterative method consists of the following steps:

1. Initialise estimate *d*.
2. Initialise $K(d_i - d), K(d_i - d) = e^{c\|d_i - d\|}$ be a given Kernel function. The weighted mean of the density in the window is determined by *K*.

$$m(d) = \frac{\sum_{d_i \epsilon N(d)} K(d_i - d) d_i}{\sum_{d_i \epsilon N(d)} K(d_i - d) d_i} \qquad (1)$$

is where *N* (*d*) is the neighbourhood of *x*, a set of points for which *K (d) ≠0.*

3. Now, set $d_i \leftarrow m(d)$, and repeats the estimation until *m(d)* converges.

The k means algorithm takes the input parameter okay, and partitions a fixed of n objects into okay clusters so that the resulting intra-cluster similarity is high while, the inter-cluster similarity is low. Cluster similarity is measured with the aid of the suggest price of the items in a cluster.

K-Mean approach set of rules is an unsupervised clustering set of rules that classifies the input records factors into more than one training primarily based on their inherent distance from each other[9]. The algorithm assumes that the information functions form a vector area and attempts to find herbal clustering in them. The points are clustered around centroids μ_i ¥ $i = 1 \dots k$ which are obtained by minimizing the objective

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i \right)^2 \qquad (2)$$

where there are k clusters S_i, i = 1,2,…, k and μ_i is the centroid or mean point of all the points $x_j \in S_i$ [4].

In this study, an iterative version of the algorithm is presented for image segmentation. The algorithm takes a 2 dimensional image as input. Various steps used in the algorithm are as follows:

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters which minimises the square error criterion.

Method:

1. Take K as the initial cluster centers as calculated by the mean shift algorithm.
2. Repeat the following steps 3 and 4 until the cluster labels of the image do not change anymore.
3. Cluster the points based on distance of their intensities from the cluster center using distance formula.

$$c_{(i)} := \arg \underset{j}{min} \, ||x^{(i)}-\mu_j|| \qquad (3)$$

4. Compute the new centroid for each of the clusters.

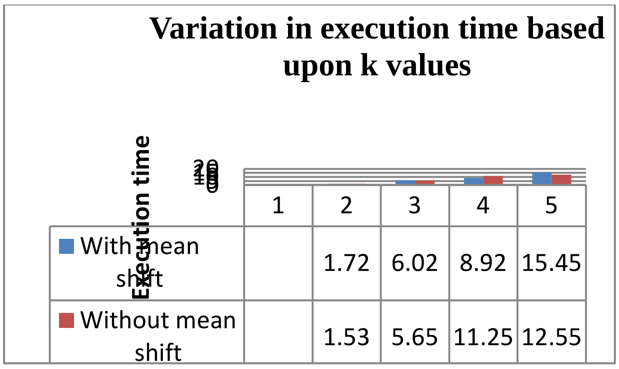
$$\mu_i := \frac{\sum_{i=1}^{m} 1\{c_{(i)} = j\}x^{(i)}}{\sum_{i=1}^{m} 1\{c_{(i)} = j\}} \qquad (4)$$

where k is a parameter of the algorithm (the number of clusters to be found), i iterates over the all the intensities, j iterates over all the centroids and μ_i are the centroid intensities.

5. The resultant new values are used to find out again the mean value until convergence.

III. Results and Analysis

Dependent materials appeal to a growing interest because of their superior mechanical and bodily properties. Such properties are inherently related to the particular shape which is controlled on the micro-scale. The ultra-high decision pics may be correctly processed to gain quantitative description of the micro-debris of cellular segmentation.



	1	2	3	4	5
🟦 With mean shift		1.72	6.02	8.92	15.45
🟥 Without mean shift		1.53	5.65	11.25	12.55

Fig. 1 Variation in execution time based upon k values.

IV. Ultrasound Image

This image represents the cancer in the bile duct of human body. Bile *duct cancer* begins near the liver and is an aggressive type of *diesis.*

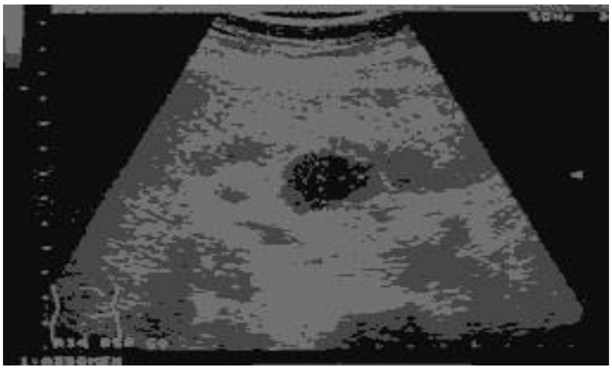
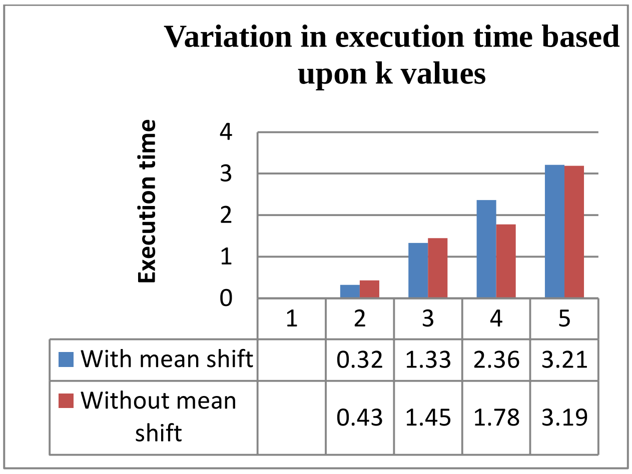


Fig. 2 Without Mean Shift



	1	2	3	4	5
🟦 With mean shift		0.32	1.33	2.36	3.21
🟥 Without mean shift		0.43	1.45	1.78	3.19

Fig. 3 Variation in execution time based upon k values

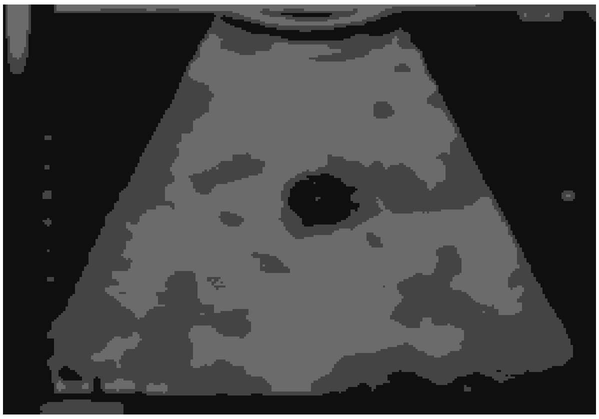


Fig. 4 With Mean Shift at k=3

V. Infrared Image

Many surveillance structures are used with human as an item. This photo is an Infrared picture. So it's miles crucial and useful to faze the human body from a photograph.

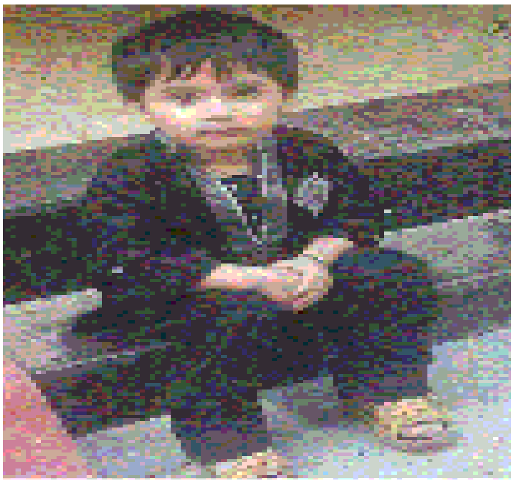


Fig. 5 Without Mean Shift

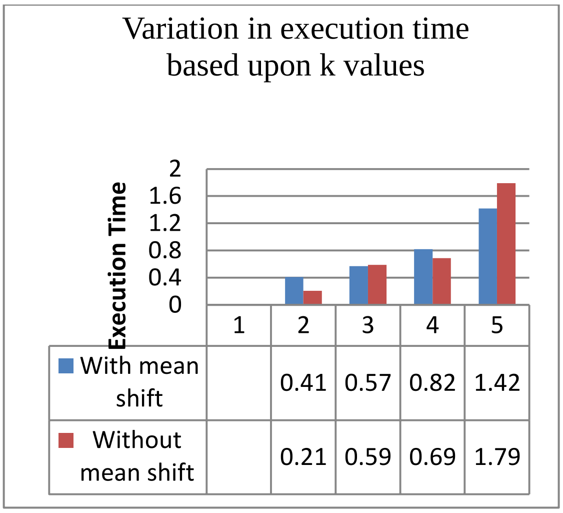


Fig. 6 Variation in execution time based upon k values

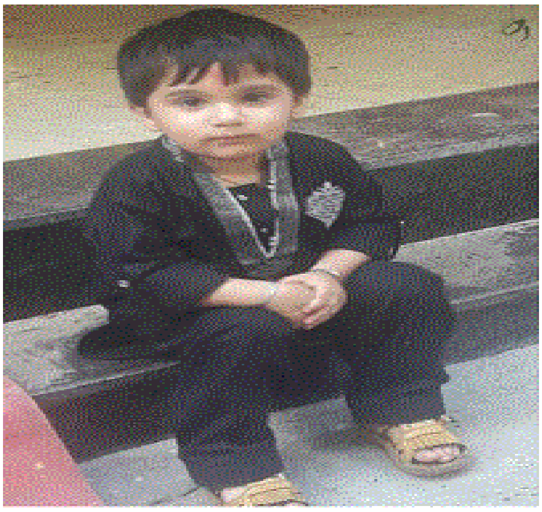


Fig. 7 With Mean Shift at k=3

VI. Conclusion

This chapter presents the result of mean shift and k-means clustering algorithm. K-means clustering is a common way to define classes of jobs within a Data set. Graphs show that with increasing fee of okay, there is not a good deal huge variant in execution time. Therefore, phase technique outperforms the k-way set of rules and this may be visible within the outcomes shown above for the 3 domains. Consequently, mean-Shift can be used for initialization purpose in ok-means and effects can be studied for these domain names where execution time is important other than segmentation.

VII. REFERENCES

1. F Gibou and R Fedkiw," A Fast Hybrid k-Means Level Set Algorithm for Segmentation", 4th Annual Hawaii International Conference on Statistics and Mathematics, pp. 1-11; 2002.
2. K Sravya and S.V. Akram, "Medical Image Segmentation by using the Pillar K-means Algorithm", International Journal of Advanced Engineering Technologies Vol. 1-Issue 1, pp. 1080-1087; 2013.
3. G Pradeepini and S. Jyothi, "An improved k-means clustering algorithm with refined initial centroids", Publications of Problems and Application in Engineering Research, Vol 04, Special Issue 01; 2013.
4. G Frahling and C Sohler, "A fast k-means implementation using coresets", International Journal of Computer Geometry and Applications 18(6): pp. 605-625; 2008.
5. Charles Elkan, "Using the Triangle Inequality to Accelerate k-means", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, pp. 147-153; 2003.
6. K Wagsta, C Cardie, S Rogers and S Schroedl, "Constrained K-means Clustering with Background Knowledge". Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584; 2001

7. M Alata, M Molhim and A Ramini, " Using GA for Optimization of the Fuzzy C-Means Clustering Algorithm", Research Journal of Applied Sciences, Engineering and Technology 5(3): pp.695-701; 2013.
8. V Sumant, A. Joshi and N Shire, "A review of an enhanced algorithm for color Image Segmentation", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, pp. 435-440; 2013.
9. Suman Tatiraju, Avi Mehta ,"Image segmentation using K-means clustering, EM and Normalized Cuts", 2008.
10. J. Gu; J. Zhou and X. Chen, "An Enhancement of K-means Clustering Algorithm". In proceedings of Business Intelligence and Financial Engineering (BIFE '09); 2009.
11. D. Mal yszko and S. T. Wierzchon "Standard and Genetic K-means Clustering Techniques in Image Segmentation", (CISIM'07) 0-7695-2894-5/07 IEEE 2007.
12. Sumant V. Joshi and Atul. N. Shire, "A Review of an enhanced algorithm for color Image Segmentation", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, pp. 435-440; 2013.
13. J. Wang and X. Su; "An improved K-means Clustering Algorithm". In proceedings of 3rd International Conference on Communication Software and Networks (ICCSN), 2011.
14. C. Kearney and Andrew J. Patton, "Information gain ranking". Financial Review, 41, pp. 29-48; 2000.
15. R.V. Singh and M.P.S. Bhatia, "Data Clustering with Modified K-means Algorithm". In proceeding of International Conference on Recent Trends in Information Technology (ICRTIT), 2011.
16. Li Xinwu, "Research on Text Clustering Algorithm Based on Improved K-means". In proceedings of Computer Design and Applications (ICCDA), Vol. 4, pp.: V4-573 - V4-576; 2010.
17. F Gibou and R Fedkiw," A Fast Hybrid k-Means Level Set Algorithm for Segmentation", 4th Annual Hawaii International Conference on Statistics and Mathematics, pp. 1-11; 2016.
18. K Sravya and S.V. Akram, "Medical Image Segmentation by using the Pillar K-means Algorithm", International Journal of Advanced Engineering Technologies Vol. 1-Issue 1, pp. 1080-1087; 2013.
19. G Pradeepini and S. Jyothi, "An improved k-means clustering algorithm with refined initial centroids", Publications of Problems and Application in Engineering Research, Vol 04, Special Issue 01; 2013.
20. G Frahling and C Sohler, "A fast k-means implementation using coresets", International Journal of Computer Geometry and Applications 18(6): pp. 605-625; 2015.
21. Charles Elkan, "Using the Triangle Inequality to Accelerate k-means", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2013), Washington DC, pp. 147-153; 2013.
22. K Wagsta, C Cardie, S Rogers and S Schroedl, "Constrained K-means Clustering with Background Knowledge". Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584; 2016
23. M Alata, M Molhim and A Ramini, " Using GA for Optimization of the Fuzzy C-Means Clustering Algorithm", Research Journal of Applied Sciences, Engineering and Technology 5(3): pp.695-701; 2013.
24. V Sumant, A. Joshi and N Shire, "A review of an enhanced algorithm for color Image Segmentation", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, pp. 435-440; 2013