Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

International Conference on Mobility in Computing- ICMiC13, Organized by Mar Baselios College of Engineering and Technology during December 17-18, 2013 at Trivandrum, Kerala, India, pg.21 – 28

RESEARCH ARTICLE

Divisive Hierarchical Clustering for Random Data Points Based on Farthest Distance (DHCRF)

Raji.R[1], Philomina Simon [2]

Department of Computer Science, University of Kerala, Trivandrum

[1] rajichothish@gmail.com, [2] philomina.simon@gmail.com

Abstract— ***The term 'Clustering' means grouping of input datasets into subsets, which is commonly called 'clusters' .The elements in the clusters, are somewhat similar. Many clustering algorithms require the specification of the number of clusters to be produced, prior to execution of the algorithm. The specified method proposes a simple and efficient clustering method. This paper presents an overview of different pattern clustering methods from a statistical pattern recognition perspective. In this paper, we review the possibility of applying a variety of distance metrics and rely on Euclidean distance. The performance of the algorithm is evaluated based on time complexity, execution time and number of clusters formed. The attractive feature of the proposed method is that it solves 'local minima problem'.***

Keywords: Clustering; Farthest Distance; Hierarchical approach; Hard clustering; K means algorithm.

I. INTRODUCTION

Clustering is the organization of objects or data points into clusters (groups), so that each cluster contains approximate feature vectors according to some distance measure [1]. Clustering is an emerging research area in data mining and image processing fields. In general, clustering is a learning task as very little or no prior knowledge is given except the input data sets. Data clustering is a common technique for statistical data analysis, which is used in many fields: machine learning, data mining, pattern recognition,

image analysis, bioinformatics etc. Most of the clustering task requires iterative procedures to find local or global optimal solutions from high-dimensional data sets. Hence, they are computationally expensive .Most of the existing methods suffer from 'Local Minima Problem' .There are three major types of clustering processes according to the way they organize data: hierarchical, partitioning and mixture model methods. In this research work, we propose the Divisive Hierarchical Clustering for Random data points based on 'Farthest distance' (DHCRF).This is a novel clustering approach that avoids 'local minima problem' to great extent. Some important applications of this clustering method are in the areas of character recognition, image segmentation, object recognition, and information retrieval.

II. LITERATURE REVIEW

In the literature, there have been several clustering algorithms proposed for data streams [2], [3], [4], [5]. Earlier clustering algorithms for data streams used a single-phase model that treated data stream clustering as a continuous version of static data clustering. These algorithms used divide and conquer schemes. They partitioned data streams into segments and discovered clusters in data streams based on a k-means algorithm [2], [3] .In [5], a multistage random sampling method was proposed to speedup fuzzy c means. There were two phases in the method. In the first phase, random sampling was used to obtain an estimate of centroids and then fuzzy c means (FCM) was run on the full data with these centroids initialized. In [7], another method based on sampling for clustering large image data was proposed, where the samples were chosen by the divergence hypothesis test. Grid-based clustering algorithms divide up the data space into finite number of cells that form a grid structure and perform clustering on the grid structure. The main advantages of grid-based clustering are fast processing time, since it processes the grids and not all data points. Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. Fuzzy C-Mean (FCM) [1] is an unsupervised clustering algorithm that has been applied to wide range of problems involving feature analysis, clustering and classifier design.

III. FEATURES OF PROPOSED ALGORITHM

The proposed algorithm follows hard clustering approach. This method use Divisive hierarchical approach, a variation of Hierarchical Methods .This begins with all objects in one cluster initially. Groups are continually divided until there are as many clusters as objects. The proposed method handles the problem of 'local minima' efficiently. Compared to existing algorithms, the proposed method achieves random number of clusters based on seed point selected. The methods also yield efficiency in terms of execution time. The distance metric used here is 'Euclidean Metric'.

In Euclidean plane, if p = (p_1, p_2) and q = (q_1, q_2) then the distance is given by

$$\mathbf{d(p,q)} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

To identify farthest data point 'Chebyshev' distance can also be used instead of Euclidean Metric which also guarantees promising results.

$$D_{\mathbf{Chebyshev}}(p,q) := \max_i(|p_i - q_i|).$$

The proposed algorithm was applied on small number of data points and large data sets. The algorithm exhibits better performance for small number of data points and average performance for large data sets. The execution time was calculated in milliseconds.

IV. DIVISIVE HIERARCHICAL CLUSTERING FOR RANDOM DATA POINTS BASED ON FARTHEST DISTANCE (DHCRF) ALGORITHM

Divisive Hierarchical Clustering for Random data points based on Farthest Distance (DHCRF) is an unsupervised clustering algorithm that can be applied on wide range of problems involving feature analysis, clustering and classifier design. DHCRF has a wide domain of applications in areas such as pattern recognition, feature analysis, target recognition etc. In this method, a limiting factor (Lim) is computed which in turn controls the iterations of the loop. The basic structure of the DHCRF algorithm is discussed below. The Algorithm Divisive Hierarchical Clustering for Random data points based on Farthest distance (DHCRF) is a method of clustering random data points. This method is frequently used in any kind of pattern recognition tasks.

A. *Proposed Algorithm.* The algorithm consists of following steps:

Step 1: Accept random set of data points.

Step 2: Accept any of the data points available as seed point1.

Step 3: Calculate'd' as sum of Euclidean distance between all data points.

Compute Lim =1/2(d/n-1)

Step 4: Find farthest point among the data points w.r.to seed point. Calculate farthest distance 'fd'.

Step 5: The data point with farthest distance from seed point1 is considered as seed point 2.

Step 6: Calculate distance of each data point w. r. to seed points as'd1' and 'd2'.

Step 7: If d1<=fd/2, cluster data point in to group of seed point1, otherwise cluster data point in to other group of seed point2.

Step 8: i) If ((Lim<fd) && (size of left cluster>1)) continue to step 4 with same seed point.

ii) If ((Lim<fd) && (size of right cluster >1)) Continue to step 4 with seed point changed as farthest point calculated earlier.

Step 9.The process continues until last set of clusters are formed.

V. APPLICATION OF PROPOSED METHOD

The proposed algorithm is having wide applications in core areas of segmentation, pattern recognition, shape analysis [6] etc. One of the domains where this method was implemented and tested was 'character recognition' [2]. The Proposed method achieved an accuracy of 98.2% in case of printed character recognition and 96.7% in case of hand written character recognition.

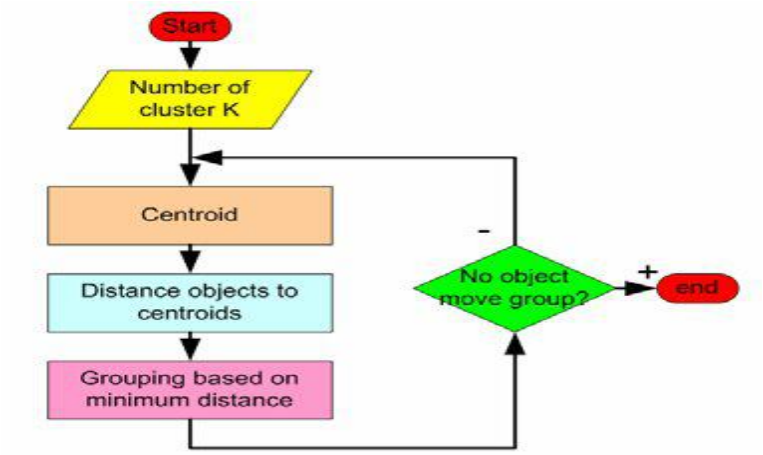
VI. RESULTS AND DISCUSSION

As stated, in this research work the data points and one seed point are randomly specified by user. The other seed point is automatically generated as farthest data point from selected seed point. Algorithm executes on formed clusters recursively. The proposed clustering technique was compared with K-Means algorithm.

A. *K-MEANS CLUSTERING*

K-Means is generally used partitioning algorithm. In this centroid of clusters are recomputed as soon as a sample joins a cluster.

Fig (1) K-means algorithm



To make comparisons simulations were done in MATLAB. We have randomly selected 50 data points and the output is plotted below. The time taken for cluster formation varies depending upon seed point selected.

Fig (2) Output of K-means algorithm

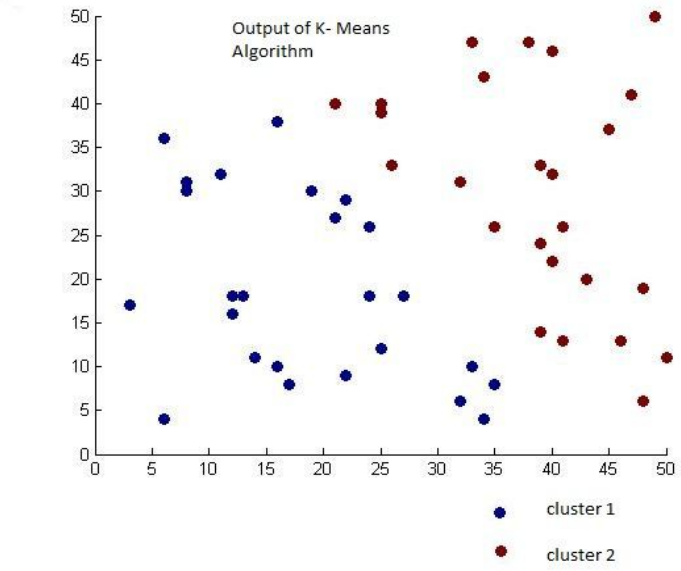


Fig (3) Output of proposed method

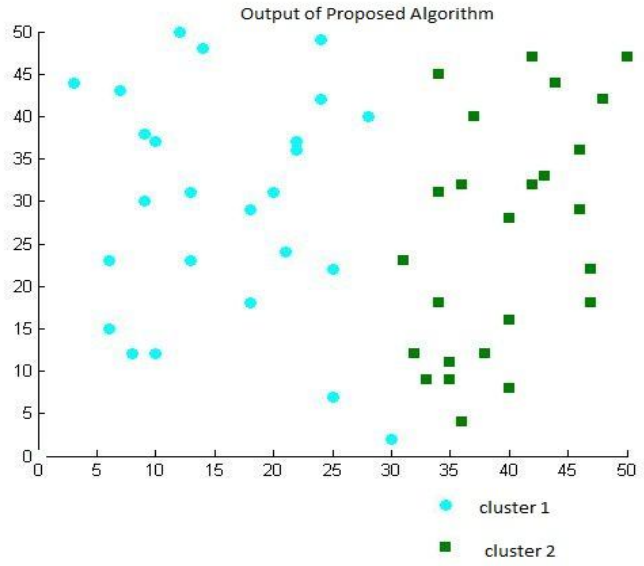


Table I: Result Comparison

Number of Data Points	50	100	
Number of Intermediate Clusters	6	9	This varies based on seed point selected.
No: of final clusters	4	6	"
Total Execution Time	110.5ms	139.5ms	"

Table I shows number of clusters formed for 50 and 100 data points. The proposed algorithm takes 110.5 milliseconds as average total execution time for 50 data points and 139.5ms for 100 data points.

Table II: Results of Different Runs

Number of data points		10		50	
Number of clusters		2	3	2	3
Average Time	TET	10.3	16.4	13.4	19.6
	ICT	6.1	12	8.6	15
Difference Time		4.2	4.4	4.8	4.6

In Table II, the total execution time (TET) is given in the row against 'TET' and the individual cluster time (ICT) is listed in second row. The difference in time between these two is given in third row. From table II, it is clear that the average time for clustering 10 data points were found to be 10.3 milliseconds when the number of clusters is 2. For the same data points, the execution time is increased when the number of cluster is increased.

VII. CONCLUSION

The experimental result of proposed algorithm shows that the DHCRF algorithm performs better for smaller data set and it gives better clusters. The proposed method also assures average performance for large data set. It can be noticed that the time taken for clustering given data points as well as effectiveness of formed clusters depends purely on seed point selected by the user. DHCRF method does not require specification of number of clusters to be produced, prior to the execution of algorithm. The experimental analysis with the proposed method shows that, it efficiently handles Local Minima Problem.

ACKNOWLEDGEMENT

The authors would like to thank Staffs and Students of Department of Computer Science, Karyavattom, Trivandrum for their constant encouragement and discussions.

REFERENCES

[1]Velmurugan.T and Santhanam,T, *Clustering of random data points using K-Means and Fuzzy-C Means Clustering Algorithms*, Proceedings of the IEEE International Conference on Emerging Trends in Computing, Virudhunagar, India, pp.177-180, 8-10,Jan - 2009.

[2] L. O'Callaghan, A. Meyerson, R. Motwani, N. Mishra, and S. Guha, *"Streaming-data algorithms for high-quality clustering."* Los Alamitos, CA, USA: IEEE Computer Society, 2002, p. 685.

[3] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "*Clustering data streams: Theory and practice," IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 515 – 528, June 2003.

[4] xiao-jun tong 1, shan zeng1,2, nong sang1,2, ling-hu zeng. *Hand-written numeral recognition based on fuzzy c-means algorithm* 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science Unrecognized Copyright Information

[5] Yong.Y, Z. Chongxun, L. Pan, A Novel Fuzzy CMeans *Clustering Algorithm for Image Thresholding,Measurement* Science Review, Volume 4(1), 2004.

[6] Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley and Sons, 1990.

[**7**] Cheng T.W., Goldgof D.B., and Hall L.O. Fast fuzzy clustering. *Fuzzy Sets and Systems*, pages 49–56, 1998.

