EFFICIENT CLUSTER CENTERS BY USING ENHANCED K-MEANS WITH MEDIAN UNIQUE VECTOR OPTIMIZATION ALGORITHM (MUVO)

R. Ranga Raj[1] and M. Punithavalli[2]
[1]Department Computer Science, Hindusthan College of Arts and Science, Coimbatore, India
[2]Department of Computer Applications, Sri Ramakrishna Engineering College, Coimbatore, India
E-Mail: rraj75@rediffmail.com

ABSTRACT

Clustering is the task of grouping a set of objects in such a way that objects in the same group or cluster are more similar to each other than to those in other groups. Clustering can be formulated as the multi objective optimization problem. The k-means algorithm is used to give a formal solution for optimization problem by assigning objects to the nearest cluster centers. But the main drawback is that, it depends on the initial starting condition of the cluster centers. Thus the quality of clustering is mainly depends on the initialization of clusters. To solve this problem this paper proposed a system named Median Unique Vector Optimization Algorithm. By using this algorithm it is used to sort out the correct selection of initial cluster centers for K-Means which is possibly used to avoid the local optimum problem and may reduce the number of iterations throughout the clustering process.

Keywords: clustering, initial cluster centers, K-means clustering algorithm, median unique vector optimization.

1. INTRODUCTION

Clustering is the method of grouping data objects into a set of disjoint classes is named as clusters. Clustering is an example of unsupervised classification. The clusters which are not labelled are termed as unsupervised cluster objects and it does not consist of predefined classes. Cluster analysis seeks to partition the given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [1, 2]. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters.

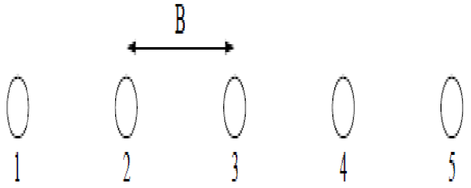
A local optimum is a kind of optimization problem which provides an optimal solution among all the possible clusters. Clustering is determined as an important criterion of research, which performs the process of optimization and finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc.

Clustering algorithms can be classified as, Hierarchical clustering algorithms and Partition clustering algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm makes the data partition set into desired number of sets in a single step [2].

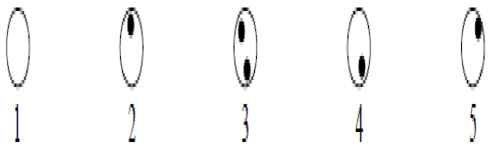
The k-means clustering algorithm [3] is a partitioning clustering method that separates data into k groups [4], [2]. The k-means clustering algorithm is more protuberant, since its intelligence to make data cluster rapidly and efficiently. However, k-means algorithm is highly stability for initial cluster centers. Because of the initial cluster centers produced arbitrarily, k-means algorithm does not promise to produce the specific or unique clustering results. Efficiency of the inventive k-means algorithm heavily depends on the initial centroids [4]. Initial centroids are also having ability on the number of iterations that are required while running the original k-means algorithm. The k-means algorithm is one of the local search procedures and the main drawback is that, it heavily depends on the initial starting condition.

2. EXISTING WORK

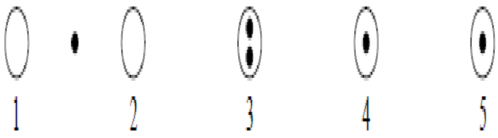
The k-means algorithm tends towards to a local optimum of its cost function. The quality of its final clustering depends heavily on the group of initialization. If this isn't done right, things could go disagreeable wrong. For example, suppose the data set consists of n points in five tight clusters that are arranged in a line, with some large distance B between them are showed clearly.



The optimal 5-clustering has cost roughly 2n. If we initialize k-means by choosing five centers at random from the given data, there may be a chance that the clustering process has no end up without centers from cluster 1, two centers from cluster 3, and one center each from clusters 2, 4, and 5:



In the first round of k-means, all points in clusters 1 and 2 will be assigned to the leftmost center. The two centers in cluster 3 will end up by sharing that cluster. And the centers in clusters 4 and 5 will be move roughly to the centers of those clusters.



Thereafter, no further changes will occur. This local optimum consists of cost (B2n). This local optimum is considered as an arbitrarily far away from the optimum cost and by setting B as large enough. Thus, good initialization is crucial.

This problem can be solved by various local optimization methods [5]. However, it must be specified that these techniques have not gained wide acceptance and in many practical applications. The clustering method that is used is the k-means algorithm with multiple restarts.

3. K-MEANS CLUSTERING ALGORITHM

One of the most admired clustering algorithms is k-means clustering algorithm [6], but in this method, the quality of the final clusters depends heavily on the initial centroids, which are based on the selection that are done randomly. Moreover, the k-means algorithm is computationally very expensive also. The proposed algorithm is found to be more accurate and efficient when compared to the original k-means algorithm.

The k-means algorithm [6] finds the optimal solutions locally with respect to the clustering errors which occur during clustering. It is a fast clustering and iterative algorithm that has been used in many clustering applications. It is a point-based clustering method that starts with the cluster centers initially placed at an arbitrary positions and proceeds by moving at every step so that the cluster centers recognized in order to minimize the clustering error. The main disadvantage of the method is that it lies in sensitivity to initial positions of the cluster centers. The local k-means clustering algorithm is proposed, which constitutes a deterministic local optimization method that does not depend on any initial parameter values and employs the k-means algorithm as a local search procedure [7]. Instead of randomly selecting the initial values for all cluster centers which makes the most local clustering technique thus the proposed technique proceeds in an incremental way to attempt the optimally added value of one new cluster center at each stage.

There are some studies on the choices of initial centers that determined to avoid those local optimums but these methods do not show too much advantage over the simple random selection in the data set.

4. MEDIAN UNIQUE VECTOR OPTIMIZATION ALGORITHM (MUVO)

By using this algorithm, the sorting out the correct selection of initial cluster centers for K-Means which is possibly avoid the local optimum problem and reduce the number of iterations throughout the clustering process.

The algorithmic Procedure that involves the optimization for selecting initial cluster centers that is the main responsible for solving the global optimum problem. In order to do that, the proposed method called MUVO is used.

Step-1: First thing of the process is to check the dataset type i.e., find whether the given dataset is in 2-Dimensional or Vector. Vectorized the data, when the given data set is in 2-D.

Step-2: Since the Cluster center value does not depends on the number of element that determine the number of occurrences in the vector. If the value occurred more than once are eliminated and called as Unique Occurrence.

For example

Let it X be our dataset

2	3	4	5	5	5	6	7	6	8	9	8	7	6	5

Sorting process either by ascending or descending order

2	3	4	5	5	5	5	6	6	6	7	7	8	8	9

Finding unique occurrence

2	3	4	5	6	7	8	9

Step-3: The cluster centers always takes its center by assigning range of values to maintain distance between one cluster and another cluster. Since to derive a good optimized structure, the clusters makes to divide the data value as a unique occurrence vector in to number of section depends on number of clusters.
For example: for 3 clusters i.e.

2	3	4

a)

5	6	7

b)

8	9

c)

Finally the cluster centers are selected from the section made above

For cluster-1: the Median of section (a) is chosen and divided by 2.
For cluster-2: the Median of section (b) is chosen and divided by 2.
For cluster-3: the Median of section (c) is chosen and divided by 2.

To improve the accuracy of the process, only half of median value is used.

5. PROPOSED WORK

Median Unique Vector Optimization (MUVO) sort out the correct selection of initial cluster centers for K-Means which is possibly avoid the local optimum problem and reduce the no of iterations throughout the clustering process. Experimental results in Table-1 shows that the proposed algorithm produces better clusters in less computation time and also when the dataset size increases the number of iterations is also reduced.

Table-1. Experimental result.

Data set size	No. of iterations		Total time taken		Average time taken for iteration	
	EK-means	Proposed EK-means	Ek-means	Proposed EK-means	EK-means	Proposed EK-means
64×64	14	11	0.3503	0.2858	0.0250	0.0260
128×128	11	10	1.5579	1.2738	0.1416	0.1274
256×256	16	8	22.2739	11.0209	1.3921	1.3776
512×512	31	9	800.6809	196.0599	25.8284	21.7844

Table-2. Clusters center details of enhanced K- Means.

Data set size	Clusters center details of enhanced K- means							
	Initial clusters				Final clusters			
N of cluster center = 4	C1	C2	C3	C4	C1	C2	C3	C4
64 × 64	2	39	14	108	9.99	82.13	43.38	178.63
128 × 128	192	168	178	10	178.81	49.58	86.56	9.33
256 × 256	130	26	67	75	179.04	7.45	51.23	87.51
512 × 512	67	195	161	88	6.78	179.36	90.11	53.88

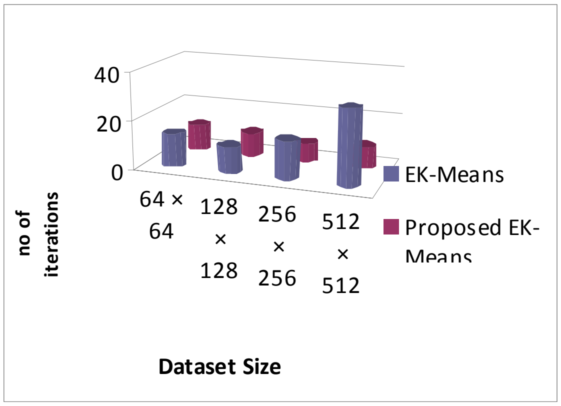


Figure-1. No. of Iterations.

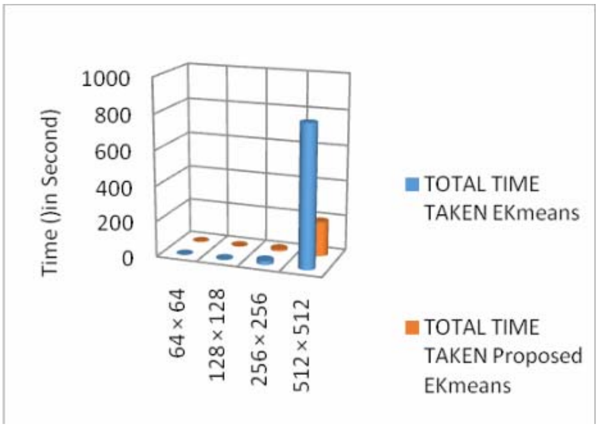


Figure-2. Total time taken (in seconds).

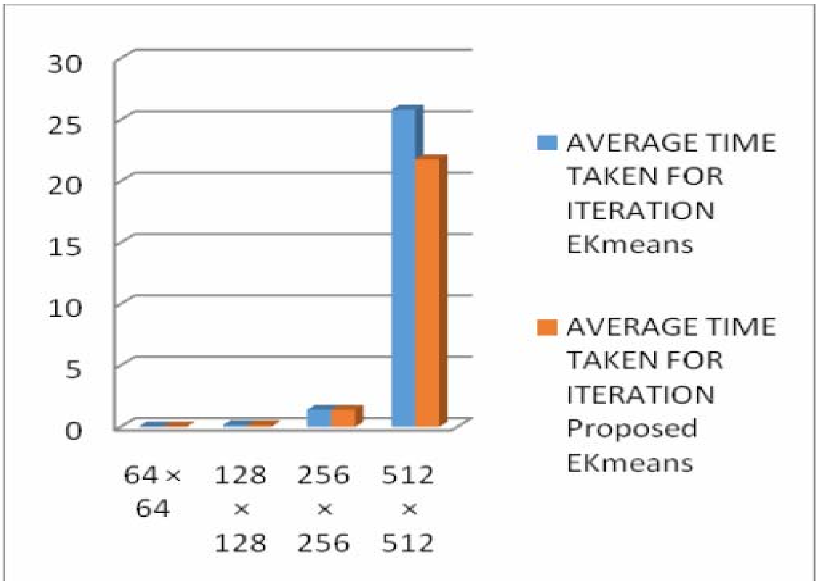


Figure-3. Average time taken (in seconds).

Table-3. Clusters center details of enhanced K- means optimized by MUVO.

Data set size	**Clusters center details of enhanced K- means optimized by MUVO**							
	Initial clusters				**Final clusters**			
	C1	C2	C3	C4	C1	C2	C3	C4
64 × 64	12.5	38	63.5	90	9.99	43.38	82.13	178.63
128 × 128	13.75	41.75	69.75	97.75	8.51	47.43	84.33	178.71
256 × 256	15.5	47	78.5	110	6.77	84.81	48.87	178.74
512 × 512	15.75	47.75	79.75	111.75	6.18	51.12	86.25	179.03

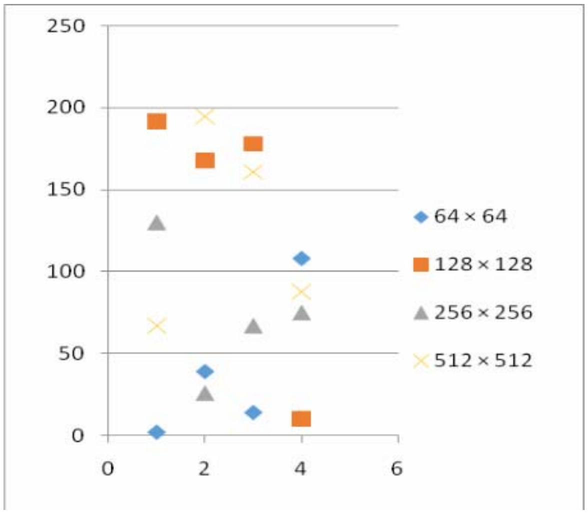


Figure-4. Initial clusters centers of E K-Means.

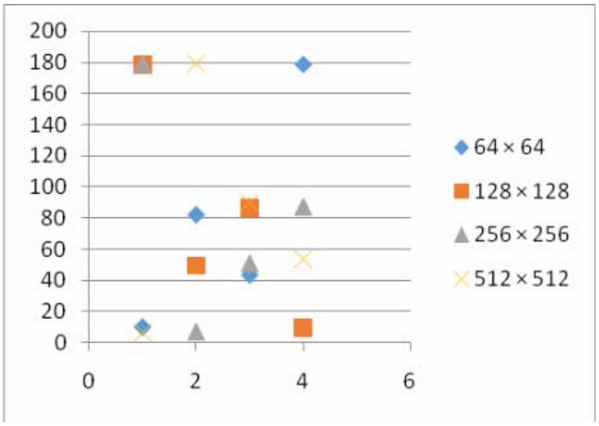


Figure-5. Final clusters centers of E K-Means.

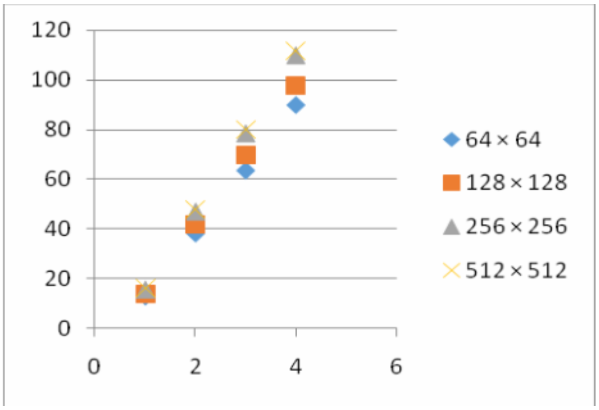


Figure-6. Initial clusters centers of E K-Means with MUVO.

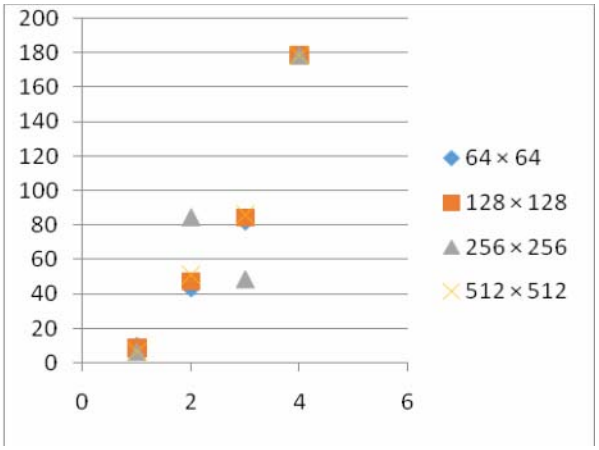


Figure-7. Final clusters centers of E K-Means with MUVO.

6. CONCLUSIONS

K-Means algorithm is highly precarious in finding the solution for initial cluster centers. Initial centroids also have an influence on the number of iterations that are required while running the original k-means algorithm. The k-means algorithm is a local search procedure and the main drawback is that, it heavily depends on the initial starting condition. The quality of its final clustering depends heavily on the manner of initialization. To solve this problem, this paper proposed a system named as Median Unique Vector Optimization Algorithm. By using this algorithm the sorting of the correct selection of initial cluster centers for K-Means which is possibly avoid the local optimum problem and reduce the number of iterations throughout the clustering process. This algorithm is found as more accurate and efficient compared to the original k-means algorithm [6]. Hence this proposed algorithm provides better method of finding the initial centroids accurately.

REFERENCES

[1] S. Deelers and S. Auwatanamongkol. Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance. International Journal of Computer Science. 2(4).

[2] Margaret H Dunham. 2006. Data Mining-Introductory and Advanced Concepts. Pearson Education.

[3] K A Abdul Nazeer, S D Madhu Kumar and M. P Sebastian. 2011. Enhancing the k-means clustering algorithm by using a O (n logn) heuristic method for finding better initial centroids. 2nd International Conference on Emerging Applications of Information Technology.

[4] K. A. Abdul Nazeer and M. P. Sebastian. 2009. Improving the accuracy and efficiency of the k-means clustering algorithm. In: International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009). Vol. 1, July, London, UK.

[5] Rutvik Desai and Rajendra Patil. 1996. SALO: Combining Simulated Annealing and Local Optimization for Efficient Global Optimization. In: Proceedings of the 9th Florida AI Research Symposium (FLAIRS-'96), Key West, FL. pp. 233-237, June.

[6] M.V. B. T. Santhi, V.R.N.S.S.V. Sai Leela, P.U. Anitha and D. Nagamalleswari. 2011. Enhancing K-Means Clustering Algorithm. ISSN: 0976-8491(Online) | ISSN: 2229-4333 (Print) IJCST. 2(4), October - December.

[7] Changseok Bae, Wei-Chang Yeh, Noorhaniza Wahid, Yuk Ying Chung and Yao Liu. 2012. A New Simplified Swarm Optimization (Sso) Using Exchanges Local Search Scheme. International Journal of Innovative Computing, Information and Control ICIC International c 2012 ISSN 1349-4198. 8(6), June.