



A New and Efficient K-Means Clustering Algorithm

G.Kiran Kumar, T. Bala Chary

Department of CSE
MLR IT, Hyderabad
A.P., India

P.Premchand

Department of CSE
University College of Engineering
Hyderabad, A.P., India

Abstract: A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Since K-means is widely used for general clustering, its performance is a critical point. This performance depends highly on initial cluster centers since it may converge to numerous local minima. There are many variations of K-means algorithm such as Lloyd's K-means clustering algorithm, Grid based K-means algorithm, Hierarchical K-means algorithm etc.. This paper is going to review all existing K-means algorithms and to propose a new and efficient K-means algorithm for clustering.

Keywords: Clusters, K-means clustering algorithm, Euclidian Distance.

1. INTRODUCTION

Clustering is an essential task in data mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them[1]. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering has been a widely studied problem in a variety of application domains including neural networks[2]. Cluster analysis divides data into meaningful or useful groups or clusters. If meaningful clusters are the goal, then the resulting clusters should capture the natural structure of the data[3]. Clustering is an important area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics etc[4]. In this paper the cluster analysis is done with method K-means clustering algorithm. K-means is a clustering algorithm that deals with numerical attribute values (NAs) primarily, although it can also be applied to categorical datasets with binary values, by viewing the binary values as numerical. The K-means clustering algorithm for numerical datasets requires the user to specify the number of clusters to be produced and the algorithm builds and refines the specified number of clusters. But due to number of iterations in the loop, the basic K-means is computationally more time consuming and also it produces different results with different dataset [3]. So the proposed K-means clustering algorithm will reduce the number of iterations and the time complexity.

2. EXISTING K-MEANS ALGORITHM

K-means clustering algorithm is one of the most well known partitioning algorithms. In this algorithm we are taking the number of inputs, represented with the k, the k is called as clusters from the data set. The value of k will be defined by the user and each cluster having some distance between them, we calculate the distance between the clusters using the Euclidean distance formula.

Algorithm: K-means:

The K-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

K: the number of clusters.

D: a data set containing n objects.

Output: A set of K clusters.

Method:

1. Arbitrarily choose K objects from D as the initial cluster centers;
2. **Repeat**
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5. **Until** no change;

The algorithm Description of four steps:

1. Initialization

In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification

The distance is calculated for each data point from the centroid and the data point having minimum distance from the centroid of a cluster is assigned to that particular cluster.

3. Centroid Recalculation

Clusters generated previously, the centroid is again repeatedly calculated means recalculation of the centroid.

4. Convergence Condition

Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

In this K-means clustering algorithm if the number of clusters is more the complexity and number of iterations are more.

3. PROPOSED K-MEANS CLUSTERING ALGORITHM

In this paper we are going to propose A New and Efficient K-means clustering algorithm. K-means is a clustering algorithm that deals with numerical attribute values (NAs) primarily, although it can also be applied to categorical datasets with binary values, by viewing the binary values as numerical. The K-means clustering algorithm for numerical datasets requires the user to specify the number of clusters to be produced and the algorithm builds and refines the specified number of clusters. But due to number of iterations in the loop, the basic K-means is computationally more time consuming and also it produces different results with different dataset. So the proposed K-means clustering algorithm will reduce the number of iterations and the time complexity.

GOALS:

- To reduce the number of iterations in the K-means Clustering Algorithm.
- To reduce the time complexity.

Proposed Algorithm: A New and Efficient K-means Clustering Algorithm:

The K-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

K: the number of clusters.

D: a data set containing n objects.

Output: A set of K clusters.

Method:

1. Arbitrarily choose K objects from D as the initial cluster centers;
2. **Repeat**
3. (re)assign each object to the cluster;
 - i. **Repeat:**
 - ii. Calculate the distance for each centroids from a data point and the data point having minimum distance from the centroid of a cluster is assign to that particular cluster center and calculate the mean value for that cluster center.
 - iii. Update the cluster mean to that centroid.
 - iv. **Until** end of data points
4. **Until** no change in cluster values;

Description of Algorithm in Step wise:

This algorithm consists of four steps:

1. Initialization

In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification

The distance is calculated for each centroid from a data point and the data point having minimum distance from the centroid of a cluster is assigned to that particular cluster.

3. Centroid Recalculation

If the data point near to one of the centroid, then we need to calculate the mean value and change the previous centroid value with the mean value and again calculate the distance with another data point and updated cluster centers.

4. Convergence Condition

Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

In this proposed A New and Efficient K-means clustering algorithm is compare with the existing algorithm the number of iterations are less.

4. EXPERIMENTAL RESULTS

We have evaluated our proposed algorithm on several datasets. We have compared our results with the existing algorithm in terms of total number of iterations. A direct comparison with other algorithms is not feasible due to unavailability of their datasets and software. However, we present some qualitative comparisons.

All the experimental results are shown on the table-4.1 , this table shows the comparison between the existing algorithm and the proposed algorithm with ten data sets.

The data sets are:

1. A1(2,10), A2(2,5), A3(8,5), B1(5, 8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).
2. A1(2,1), A2(4,9), A3(1,0), A4(3,5), A5(2,6),A6(3,7), A7(8,2),A8(1,1)
3. A1(10,3),A2(6,7),A3(1,0),A4(2,4),A5(12,10)
4. A1(5,8),A2(3,0),A3(6,9),A4(1,9),A5(2,6),A6(0,1)
5. A1(13,3), A2(3,0), A3(9,1),A4(10,2),A5(1,2),A6(1,9),A7(12,3),
A8(0,3),A9(2,6)
6. A1(30,2), A2(1,10), A3(85,62), A4(12,3), A5(1,0), A6(101,28), A7(54,34), A8(4,9), A9(12,10), A10(74,21),
A11(33,8), A12(15,10), A13(40,2), A15(91,8), A16(13,9).
7. A1(1,1),A2(4,3),A3(10,2),A4(3,3),A5(12,10),A6(20,4),A7(1,7)
8. A1(1,6), A2(4,7), A3(2,9), A4(3,7), A5(12,0), A6(10,1), A7(1,9), A8(9,3)
9. A1(2,4), A2(3,7), A3(2,8), A4(20,13), A5(12,6),A6(25,31), A7(21,2), A8(1,23), A9(0,9), A10(5,30), A11(2,4),
A12(1,2).
10. A1(12,10), A2(2,10), A3(25,14), A4(12,3), A5(10,2), A6(14,3), A7(5,4), A8(1,9)

S.No	Data Sets	Centroids	# Iterations (Existing Algorithm)	# Iteration (Proposed Algorithm)
1	DS1	3	3	2
2	DS2	4	5	3
3	DS3	2	4	3
4	DS4	3	7	4
5	DS5	3	2	2
6	DS6	4	15	11
7	DS7	2	9	6
8	DS8	2	3	2
9	DS9	3	5	4
10	DS10	2	4	3

Table 4.1: Comparative study of existing algorithm vs. Proposed algorithm with ten different data sets

Example:

we take one dataset and take three centroids and we solve this problem using both the existing algorithm and the proposed algorithm, first we solve this problem with the existing algorithm.

Using Existing K-means Algorithm:

Data set: A1(2,10), A2(2,5), A3(8,5), B1(5, 8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

Centroids: A1(2,10), B1(5,8), C1(1,2).

Iteration: 1

We need to calculate the distance between each data points and the centroids using the Euclidean distance.

Two points (x1,y1), (x2,y2)

$$\text{Euclidean distance Formula: } = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

$$\text{or } = |x_2-x_1| + |y_2-y_1|$$

$$\text{Mean Formula: } ((x_1+x_2) / 2, (y_1+y_2) / 2).$$

1ST ROW:

Distance calculate between the A2 data point and the Centroids A1, B1, C1

$$\text{Distance between A2(2,5) \& A1(2, 10) } = |2-2| + |5-10| = 0+5 = 5$$

$$\text{Distance between A2(2,5) \& B1(5, 8) } = |2-5| + |5-8| = 3+3 = 6$$

$$\text{Distance between A2(2,5) \& C1(1, 2) } = |2-1| + |5-2| = 1+3 = 4$$

The A2 nearby Cluster Center is C1.

2ND ROW:

Distance calculate between the A3 data point and the Centroids A1, B1, C1

$$\text{Distance between A3(8,5) \& A1(2,10) } = 11$$

$$\text{Distance between A3(8,5) \& B1(5,8) } = 6$$

$$\text{Distance between A3(8,5) \& C1(1,2) } = 10$$

The A3 nearby Cluster Center is B1.

3RD ROW:

Distance calculate between the B2 data point and the Centroids A1, B1, C1

$$\text{Distance between B2(7,5) \& A1(2,10) } = 10$$

$$\text{Distance between B2(7,5) \& B1(5,8) } = 5$$

$$\text{Distance between B2(7,5) \& C1(1,2) } = 9$$

The B2 nearby Cluster Center is B1.

4TH ROW:

Distance calculate between the B3 data point and the Centroids A1, B1, C1

$$\text{Distance between B3(6,4) \& A1(2,10) } = 10$$

$$\text{Distance between B3(6,4) \& B1(5,8) } = 5$$

$$\text{Distance between B3(6,4) \& C1(1,2) } = 7$$

The B3 nearby Cluster Center is B1.

5TH ROW:

Distance calculate between the C2 data point and the Centroids A1, B1, C1

$$\text{Distance between C2(4,9) \& A1(2,10) } = 3$$

$$\text{Distance between C2(4,9) \& B1(5,8) } = 2$$

$$\text{Distance between C2(4,9) \& C1(1,2) } = 10$$

The C1 nearby Cluster Center is B1.

The above calculations are shown in the form of below table-4.2:

Data Sets	Centroids			Cluster
	A1(2,10)	B1(5,8)	C1(1,2)	
A2(2,5)	5	6	4	C1
A3(8,5)	11	6	10	B1
B2(7,5)	10	5	9	B1
B3(6,4)	10	5	7	B1
C2(4,9)	3	2	10	B1

Table 4.2: Iteration-1 of Existing algorithm

Then we need to calculate the cluster mean values

Cluster B1(5,8) nearby points are A3(8,5), B2(7,5), B3(6, 4), C2(4, 9)
 B1 Mean value = (6, 6.2)

Cluster C1(1,2) nearby points are A2(2,5)

C1 Mean value = (1.5, 3.5)

The updated Cluster points are : A1(2, 10), B1(6, 6.2), C1(1.5, 3.5)

Now we need to go for the next iteration with the updated cluster points

Iteration: 2

Now we need to calculate the distances between the each data points to centroids.

Again am not showing the entire calculation part just see the below table 4.3:

Data Sets	Centroids			Cluster
	A1(2, 10)	B1(6, 6.2)	C1(1.5, 3.5)	
A2(2,5)	5	5.2	2	C1
A3(8,5)	11	3.2	8	B1
B2(7,5)	10	2.2	7	B1
B3(6,4)	10	2.2	5	B1
C2(4,9)	3	4.8	8	A1

Table 4.3: Iteration-2 of Existing algorithm

So, after completion of the iteration 2 the cluster points are not equal to the iteration 1 cluster points, and then we need to go for the iteration 3 before that we need to calculate the cluster mean values.

Cluster A1(2, 10) nearby points are C2(4,9)
 A1 Mean value = (3, 9.5)

Cluster B1(6, 6.2) nearby points are A3(8,5), B2(7,5), B3(6,4)
 B1 Mean value = (6.7, 4)

Cluster C1(1.5,3.5) nearby points are A2(2,5)
 C1 Mean value = (1.7, 4.2)

The updated Cluster points are : A1(3, 9.5), B1(6.7, 4), C1= (1.7, 4.2)

Iteration: 3

Now we need to calculate the distances between the each data points to centroids. Again am not showing the entire calculation part just see the below table 4.4.

Data Sets	Centroids			Cluster
	A1(3, 9.5)	B1(6.7, 4)	C1=(1.7, 4.2)	
A2(2,5)	6.5	5.7	1.1	C1
A3(8,5)	9.5	2.3	7.1	B1
B2(7,5)	8.5	1.3	6.1	B1
B3(6,4)	8.5	0.7	4.5	B1
C2(4,9)	1.5	7.7	7.1	A1

Table 4.4: Iteration-3 of Existing algorithm

Here we see the cluster values are no change between the Iteration 2 and the iteration 3, then we stop the iteration/ this data set with the 3 centroids generates the cluster groups in 3 iterations.

Let us see with these data set values and the same centroids with our proposed system or algorithm what changes made happen.

Using proposed K-means Algorithm:

Data set: A1(2,10), A2(2,5), A3(8,5), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

Centroids: A1(2,10), B1(5,8), C1(1,2).

Iteration: 1

We need to Calculate the distance for each centroids from a data point and the data point having minimum distance from the centroid of a cluster is assign to that particular cluster center and calculate the mean value for that cluster center.

Two points (x1,y1), (x2,y2)

Euclidean distance Formula: $=\sqrt{(x2-x1)^2 + (y2-y1)^2}$

or $= |x2-x1| + |y2-y1|$

Mean Formula: $((x1+x2) / 2, (y1+y2) / 2)$.

1ST ROW:

Distance calculate between the A2 data point and the Centroids A1, B1, C1

Distance between A2(2,5) & A1(2,10) = $|2-2| + |5-10| = 0+5 = 5$

Distance between A2(2,5) & B1(5,8) = $|2-5| + |5-8| = 3+3 = 6$

Distance between A2(2,5) & C1(1,2) = $|2-1| + |5-2| = 1+3 = 4$

The A2 nearby Cluster Center is C1.

Then we need to calculate the mean value between the C1 and A2.

C1 Mean value = (1.5,3.5)

Then we need to update the centroid C1 value as (1.5, 3.5).

2ND ROW:

Distance calculate between the A3 data point and the Centroids A1, B1, C1

Distance between A3(8,5) & A1(2,10) = 11

Distance between A3(8,5) & B1(5,8) = 6

Distance between A3(8,5) & C1(1.5,3.5) = 8

The A3 nearby Cluster Center is B1.

Then we need to calculate the mean value between the A3 and B1.

B1 Mean value = (6.5, 6.5)

Then we need to update the centroid B1 value as (6.5, 6.5).

3RD ROW:

Distance calculate between the B2 data point and the Centroids A1, B1, C1

Distance between B2(7,5) & A1(2,10) = 10

Distance between B2(7,5) & B1(6.5,6.5) = 2

Distance between B2(7,5) & C1(1.5, 3.5) = 7

The B2 nearby Cluster Center is B1.

Then we need to calculate the mean value between the B2 and B1.

B1 Mean value = (6.7, 5.7)

Then we need to update the centroid B1 value as (6.7, 5.7).

4TH ROW:

Distance calculate between the B3 data point and the Centroids A1, B1, C1

Distance between B3(6,4) & A1(2,10) = 10

Distance between B3(6,4) & B1(6.7, 5.7) = 2.5

Distance between B3(6,4) & C1(1.5,3.5) = 5

The B3 nearby Cluster Center is B1.

Then we need to calculate the mean value between the B3 and B1.

B1 Mean value = (6.3, 4.8)

Then we need to update the centroid B1 value as (6.3, 4.8).

5TH ROW:

Distance calculate between the C2 data point and the Centroids A1, B1, C1

Distance between C2(4,9) & A1(2, 10) = 3

Distance between C2(4,9) & B1(6.3, 4.8) = 6.5

Distance between C2(4,9) & C1(1.5, 3.5) = 8

The C2 nearby Cluster Center is A1.

Then we need to calculate the mean value between the C2 and A1.

A1 Mean value = (3, 9.5)

Then we need to update the centroid A1 value as (3, 9.5).

The updated Cluster points are : A1(3, 9.5), B1(6.3, 4.8), C1(1.5, 3.5)

The above calculations are shown in the form of below table:

Data Sets	Centroids			Cluster	Mean value
	A1(2, 10)	B1(5, 8)	C1(1, 2)		
A2(2,5)	5	6	4	C1	(1.5, 3.5)
	A1(2, 10)	B1(5, 8)	C1(1.5, 3.5)		
A3(8,5)	11	6	8	B1	(6.5, 6.5)
	A1(2, 10)	B1(6.5, 6.5)	C1(1.5, 3.5)		
B2(7,5)	10	2	7	B1	(6.7, 5.7)
	A1(2, 10)	B1(6.7, 5.7)	C1(1.5, 3.5)		
B3(6,4)	10	2.5	5	B1	(6.3, 4.8)
	A1(2, 10)	B1(6.3, 4.8)	C1(1.5, 3.5)		
C2(4,9)	3	6.5	8	A1	(3, 9.5)

Table 4.5: Iteration-1 of proposed algorithm

Now we need to go for the next iteration with the updated cluster points

Iteration: 2

Again am not showing the entire calculation part just see the below table 4.6:

Data Sets	Centroids			Cluster	Mean value
	A1(3, 9.5)	B1(6.3, 4.8)	C1(1.5, 3.5)		
A2(2,5)	5.7	5.1	1	C1	(1.8, 4.6)
	A1(3, 9.5)	B1(6.3, 4.8)	C1(1.8, 4.6)		
A3(8,5)	8.7	1.8	6.7	B1	(7.4, 4.3)
	A1(3, 9.5)	B1(7.4, 4.3)	C1(1.8, 4.6)		
B2(7,5)	7.7	1	5.5	B1	(7.2, 4.6)
	A1(3, 9.5)	B1(7.2, 4.6)	C1(1.8, 4.6)		
B3(6,4)	7.7	18	4.7	B1	(6.6, 4.3)
	A1(3, 9.5)	B1(6.6, 4.3)	C1(1.8, 4.6)		
C2(4,9)	0.7	7	6.5	A1	(3.7, 9.1)

Table 4.6: Iteration-2 of proposed algorithm

Here we see the cluster values are no change between the Iteration 1 and the iteration 2, then we stop the iteration. This data set with the 3 centroids generates the cluster groups in 2 iterations.

All the experimental results are shown on the Figure-4.1, this graph shows the comparison between the existing algorithm and the proposed algorithm with ten data sets.

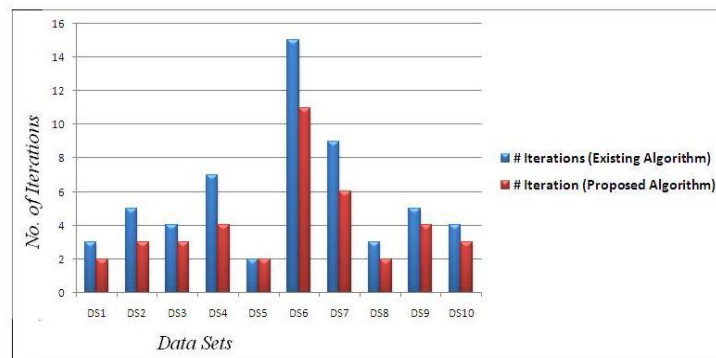


Fig. 4.1: The number of iterations between existing algorithm vs proposed algorithm

5. CONCLUSION

This paper is going to review all existing K-means algorithms and propose a new and efficient K-means algorithm for clustering with less number of iteration and time complexity compare with the existing algorithm. The future enhancement in this paper is to develop this paper in the GUI based system. This can be used for spatial data in finding spatial co-location patterns [9]. In this paper, we present a New and Efficient K-means clustering algorithm for performing K-means clustering algorithm. Our experimental results demonstrate that our scheme can improve the direct K-means algorithm by few Iterations.

REFERENCES

1. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur. "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining." *Volume 1, Issue 3, May2012*
2. Khaled Alsabti, Syracuse University, "An Efficient K-Means Clustering Algorithm",
3. Neha Aggarwal, Kirti Aggarwal, Kanika gupta "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining", Volume 3, Issue 3, August-2012 ISSN 2229-5518
4. G Pradeepini1, S. Jyothi2., "An Improved K-Means Clustering Algorithm With Refined Initial Centroids, Vol 04, Special Issue01; 2013
5. Neha Aggarwal "A Mid - Point based k-mean Clustering Algorithm for Data mining", Neha Aggarwal et al. / *International Journal on Computer Science and Engineering (IJCSSE)*
6. Tapas Kanungo, Senior Member, IEEE, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", VOL. 24, NO. 7, JULY 2002
7. "k-Means: A new generalized k-means clustering algorithm", Yiu-Ming Cheung, Department of Computer Science, Hong Kong Baptist University, Received 23 July 2002; received in revised form 11 April 2003
8. Jiawei Han and Micheline Kamber- "Data Mining Concepts and Techniques"
9. G.Kiran Kumar, P. Premchand, T. Venu Gopal, "Mining of spatial co-location pattern from spatial data sets", *International Journal of Computer Applications*, Vol 42, Number-2.1, 2012.
10. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
11. G. Hamerly and C. Elkan, "Learning the k in k-means," *Neural Information Processing Systems*, 2003.
12. T. Kanungo, D.M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "A local search approximation algorithm for k-means clustering," *Computational Geometry: Theory and Applications*, 28 (2004), pp. 89-112, 2004.
13. L. Kaufman and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, 1990.
14. D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," *International Conference on Machine Learning 2000*, pp. 727-734, 2000.
15. M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques", *Proc. KDD Workshop on Text Mining*, 2000.
16. P.K. Agarwal and C.M. Procopiuc, "Exact and Approximation Algorithms for Clustering," *Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, pp. 658-667, Jan. 1998.
17. P.S. Bradley and U. Fayyad, "Refining Initial Points for K-means Clustering," *Proc. 15th Int'l Conf. Machine Learning*, pp. 91-99, 1998.
18. Q. Du, V. Faber, and M. Gunzburger, "Centroidal Voronoi Tessellations: Applications and Algorithms," *SIAM Rev.*, vol. 41, pp. 637-676, 1999.
19. M. Ester, H. Kriegel, and X. Xu, "A Database Interface for Clustering in Large Spatial Databases," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD-95)*, pp. 94-99, 1995.
20. M. Inaba, H. Imai, and N. Katoh, "Experimental Results of a Randomized Clustering Algorithm," *Proc. 12th Ann. ACM Symp. Computational Geometry*, pp. C1-C2, May 1996.