Int. J. Math. And Appl., 9(1)(2021), 127–136
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Mathematical Approach to Representation of Locations Using K-Means Clustering Algorithm

N. Yogeesh[1,*]

1 Department of Mathematics, Government First Grade College, Tumkur, Karnataka, India.

Abstract: In artificial intelligence (AI), we will be able to handle a large amount of data without the need for human interaction in an efficient manner. The K-means clustering technique may be used to learn from unsupervised data in a straightforward manner. Machine learning and data mining both benefit from this method's ability to handle large datasets. K-means clustering uses t iterations to compute the outcomes of n items in k clusters. A variety of apps are employed in nearly every field of study to provide an uninterrupted, easy, and effective means of data learning. To illustrate how the K-means clustering method works, I'm going to use a mathematical way to describe the locations in a real-world dataset.

Keywords: K-means Clustering Algorithm, data learning, clustering problems, iteration, artificial intelligence.

1. Introduction

K-means unsupervised data learning algorithms such as the Clustering Algorithm can be used to handle well-defined data clustering challenges. There are K separate non-overlapping subgroups (clusters) that this method tries to divide the dataset into. Each data point fits to just one group. Data points inside clusters are made as similar as feasible, while the clusters themselves are kept different to the best of the algorithm's ability. Data points are placed in clusters (subgroups) based on the sum of squared distances from the data points to the cluster's centroid (the arithmetic mean of altogether the data points that belong to that cluster). The more homogenous (similar) the data points are inside a cluster, the less variance there is.

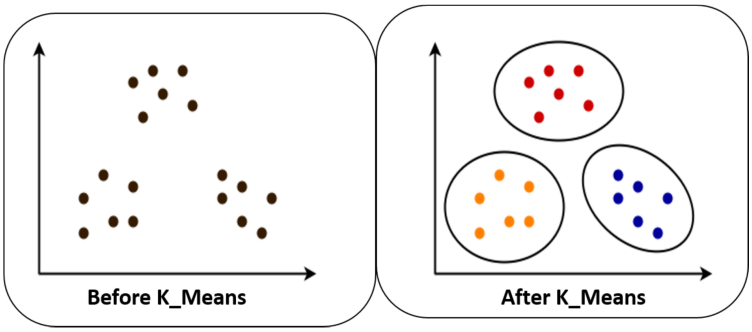
2. K-Means Clustering

- K-Means clustering is an iterative clustering method that does not require any prior knowledge.
- Clustering is done by dividing the data set into k different subsets.
- To put it another way, a cluster is a grouping of related data elements.

It divides the data collection into smaller subsets so that it is easier to analyze.

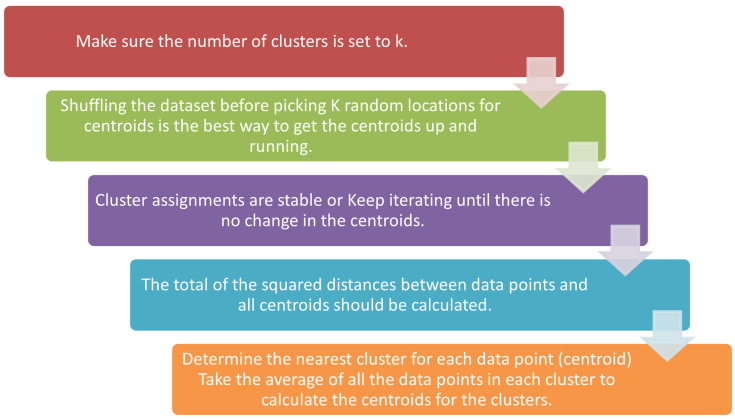
- Every single data point is clustered around the closest mean.
- There is a high degree of correlation between data points in a single cluster.
- The dissimilarity between data points in various clusters is rather large.

* *E-mail: yogeesh.r@gmail.com*



3. K-Means Clustering Algorithm

The K-means algorithm's flowchart looks like what follows,



Each phase of our K-Means Clustering Algorithm is detailed below.

Step 1:

- Take the number of clusters K.

Step 2:

- Cluster centres can be chosen at random from any of the K data points.
- The farther apart the cluster centres may be from one another, the better.

Step 3:

- Find the distance between each cluster center and each data point.
- Either a predefined distance function or the Euclidean distance method can be used to determine the distance.

Step 4:

- Each data point should be placed in a cluster.
- To assign a data point to a cluster, look for the cluster with the closest center to that data point.

Step 5:

- If new clusters arise, recalculate their centers.
- The mean of altogether the data points in a cluster is used to discovery the cluster's center.

Step 6: You can keep going back and forth between Step 3 and Step 5 until any of the following conditions are met:

- Newly created clusters retain their original centers.
- In the same cluster, all the data points are still present.
- This is the end of the iterations.

Advantages

It has the following advantages: K-Means Clustering Algorithm

Point 1: It is comparatively efficient with time complexity O(nkt) where,

- n is the number of occurrences
- k is the number of clusters that are present.
- t iterations in the process.

Point 2:

- Local optimum is frequently reached.
- It is possible to determine the global optimum using procedures such as **simulated annealing** or **genetic algorithms.**

Disadvantages

Disadvantages of the K-Means Clustering Algorithm are as follows,

- The number of clusters (k) must be determined in advance.
- Noise and outliers can't be handled.
- You cannot use this method to identify non-convex clusters.

4. Problems Based on K-Means Clustering Algorithm

Problem 4.1. *These eight points* (x, y) *indicate places, therefore divide them into three groups:*

$$A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)$$

Initial cluster centers are: $A1(2, 10)$*,* $A4(5, 8)$ *and* $A7(1, 2)$*. Use K-Means Algorithm to calculate the three cluster centers after the second iteration.*

Solution. We know that, the function that calculates the distance between two places $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as $\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|$. We use the previously stated K-Means Clustering Algorithm. We can also use Matlab, Python, and Maxima etc. to find the K-Mean Clustering.

Iteration 1:

- Each point's distance from the center of the three clusters is calculated.

- The specified distance function is used to figure out the travel time.

The distance between point $A1(2, 10)$ and the center of each of the three clusters is shown in the following,

Distance between two points $A1(2, 10)$ and $C1(2, 10)$

$$\begin{aligned}\rho(A1, C1) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |2 - 2| + |10 - 10| \\ &= 0\end{aligned}$$

Distance between two points $A1(2, 10)$ and $C2(5, 8)$

$$\begin{aligned}\rho(A1, C2) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |5 - 2| + |8 - 10| \\ &= 3 + 2 \\ &= 5\end{aligned}$$

Finding the Distance Between $A1(2, 10)$ and $C3(1, 2)$

$$\begin{aligned}\rho(A1, C3) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |1 - 2| + |2 - 10| \\ &= 1 + 8 \\ &= 9\end{aligned}$$

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

Similar calculation is made for additional sites in relation to cluster centers. Next,

- We create a table that displays all of the data.
- With the assistance of the table, we can determine which points belong to which cluster.
- The supplied point is a member of the cluster whose center is closest to the given point.

From here, new clusters are

Cluster 1: First cluster holds points

- $A1(2, 10)$

Cluster 2: Second cluster holds points

- $A3(8, 4)$
- $A4(5, 8)$
- $A5(7, 5)$
- $A6(6, 4)$
- $A8(4, 9)$

Cluster 3: Third cluster holds points

- $A2(2, 5)$
- $A7(1, 2)$

Now,

- It's time to recalculate the new clusters.
- The mean of altogether the points in the new cluster is used to calculate the new cluster center.

For Cluster 1:

- Cluster 1 has only one point $A1(2, 10)$ in it.
- As a result, cluster center has not changed.

For Cluster 2: Cluster 2 Center

$$= \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6)$$

For Cluster 3: Cluster 3 Center

$$= \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

This is the end of the first iteration.

Iteration 2:

- We figure out how far away each point is from the center of each of the three groups.
- The specified distance function is used to determine the distance.

The distance between point $A1(2, 10)$ and the center of each of the three clusters is shown in the following,

Finding the Distance between $A1(2, 10)$ and $C1(2, 10)$

$$\begin{aligned}\rho(A1, C1) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |2 - 2| + |10 - 10| \\ &= 0\end{aligned}$$

Finding the Distance between $A1(2, 10)$ and $C2(6, 6)$

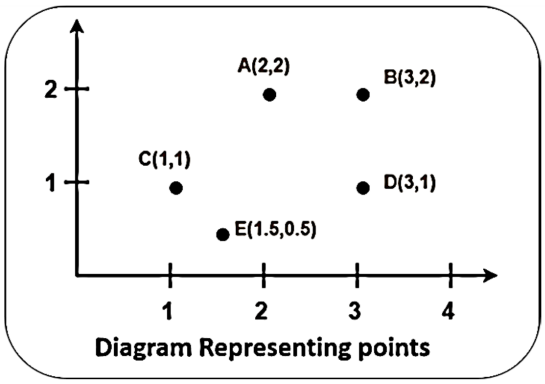
$$\begin{aligned}\rho(A1, C2) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |6 - 2| + |6 - 10| \\ &= 4 + 4 \\ &= 8\end{aligned}$$

Finding the Distance Between $A1(2, 10)$ and $C3(1.5, 3.5)$

$$\begin{aligned}\rho(A1, C3) &= |x_2 - x_1| + |y_2 - y_1| \\ &= |1.5 - 2| + |3.5 - 10| \\ &= 0.5 + 6.5 \\ &= 7\end{aligned}$$

Similarly, we determine the distances between additional sites and the cluster centers.

Problem 4.2. *Create two clusters using the K-Means Algorithm.*



Solution. Clustering is done using the K-Means algorithm that was previously mentioned. Think of the two clusters as having the centers $A1(2, 2)$ and $C(1, 1)$.

Iteration 1:

- We figure out how far apart the two clusters' centers are from one another.
- The Euclidean distance formula is used to figure out how far apart the two points are.

The following illustration shows the calculation of distance between point $A1(2, 2)$ and each of the center of the two clusters. An example of the computation of distance between points $A1(2, 2)$ and each of the two cluster centers is shown in this,

Finding the Distance between $A1(2, 2)$ and $C1(2, 2)$

$$\begin{aligned}\rho(A, C1) &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\ &= \sqrt{(2 - 2)^2 + (2 - 2)^2} \\ &= \sqrt{0 + 0}\end{aligned}$$

$$= 0$$

Finding the Distance between $A1(2, 2)$ and $C2(1, 1)$

$$\begin{aligned}\rho(A, C2) &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\ &= \sqrt{(1-2)^2 + (1-2)^2} \\ &= \sqrt{1+1} \\ &= \sqrt{2} \\ &= 1.41\end{aligned}$$

In the similar manner, each cluster's center is calculated as a distance from other places. Next,

- We create a table that displays all of the data.
- Determine which points belong in which cluster using the table.
- If the center of the cluster is closest to the supplied point, then it is part of that cluster.

Given Points	Distance from center (2, 2) of Cluster-01	Distance from center (1, 1) of Cluster-02	Point belongs to Cluster
A(2, 2)	0	1.41	C1
B(3, 2)	1	2.24	C1
C(1, 1)	1.41	0	C2
D(3, 1)	1.41	2	C1
E(1.5, 0.5)	1.58	0.71	C2

From here, New clusters are

Cluster 1: First cluster holds points

- $A(2, 2)$
- $B(3, 2)$
- $E(1.5, 0.5)$
- $D(3, 1)$

Cluster 2: Second cluster holds points

- $C(1, 1)$
- $E(1.5, 0.5)$

Now,

- It's time to recalculate the new clusters.
- The mean of all the points in the new cluster is used to calculate the new cluster center.

For Cluster 1: Cluster 1 Center

$$= \left(\frac{2+3+3}{3}, \frac{2+2+1}{3}\right) = (2.67, 1.67)$$

For Cluster 2: Cluster 2 Center

$$= \left(\frac{1+1.5}{2}, \frac{1+0.5}{2}\right) = (1.25, 0.75)$$

This is the end of the first iteration. Next, When the centers haven't changed any more, we go on to iteration 2, iteration 3.

In other way we can also express,

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

Cluster 1 has only one point $A1(2, 10)$, which was the original mean, hence the cluster center is unchanged.

For Cluster 2, we have $\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6)$.

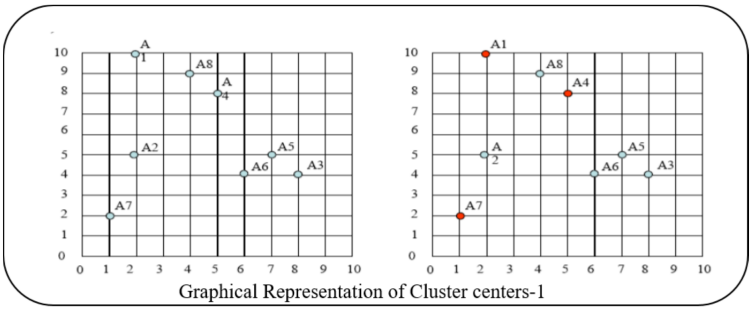
For Cluster 3, we have $\left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$.

New clusters: 1. $\{A1\}$, 2. $\{A3,\ A4,\ A5,\ A6,\ A8\}$, 3. $\{A2,\ A7\}$.

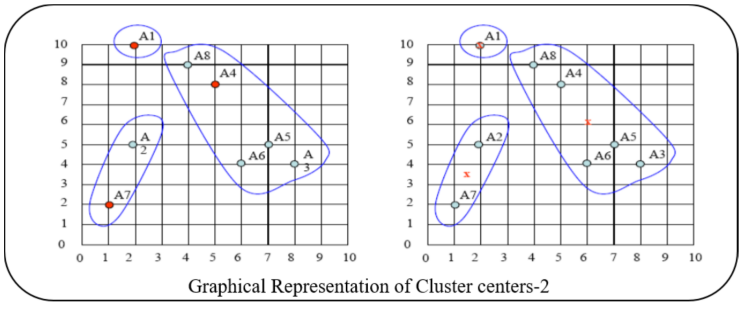
Centers of the new clusters:

$$C1 = (2, 10),$$
$$C2 = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6),$$
$$C3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

Red dots represent the cluster's original centers. X in red denote the location of the newest cluster centers.



Graphical Representation of Cluster centers-1



Graphical Representation of Cluster centers-2

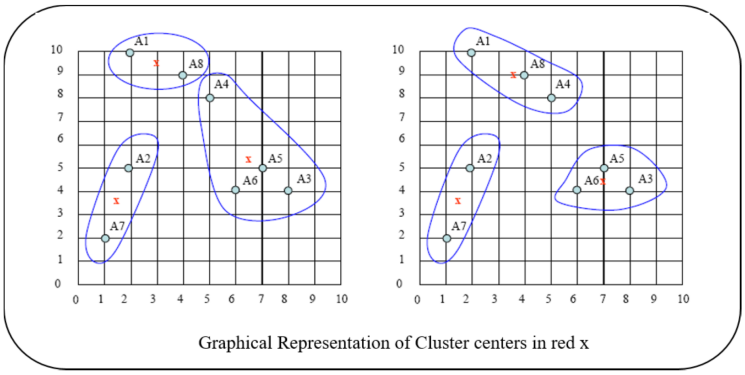
That concluded Iteration 1 (epoch1). Following Iteration 2 (epoch2) and so on, we'll continue to iterate until the means no longer change. Essentially, we repeat the procedure of Iteration 1 with a fresh set of methods. We would need two more epochs, after the 2^{nd} epoch the results would be:

1. $\{A1, A8\}$, 2. $\{A3, A4, A5, A6\}$, 3. $\{A2, A7\}$.

With centers $C1 = (3, 9.5)$, $C2 = (6.5, 5.25)$ and $C3 = (1.5,\ 3.5)$. After the 3^{rd} epoch, the results would be:

1. $\{A1, A4, A8\}$, 2. $\{A3, A5, A6\}$, 3. $\{A2, A7\}$.

With centers $C1 = (3.66, 9)$, $C2 = (7, 4.33)$ and $C3 = 1.5, 3.5)$.



Graphical Representation of Cluster centers in red x

5. Conclusion

A clustering technique known as K-means is widely utilized in a wide range of industries, and it's easy and effective. In computer applications, cluster analysis makes it simpler to comprehend the nature of data learning through the use of mathematical approaches/tools. K means clustering is a simple and effective approach for unsupervised data learning in artificial intelligence (AI) (AI). Fast and efficient: It uses the simple formula O(tkn) to compute the results, with n points or objects and k clusters, for t iterations, with n points or objects.

References

[1] J. O. Omolehin, J. O. Oyelade, O. O. Ojeniyi and K. Rauf, *Application of Fuzzy logic in decision making on students' academic performance*, Bulletin of Pure and Applied Sciences, 24E(2)(2005), 281-187.

[2] Susmita Datta and Somnath Datta, *Comparisons and validation of statistical clustering techniques for microarray gene expression data*, Bioinformatics, 19(2003), 459-466.

[3] S. Nittel, K. T. Leung, A. Braverman, U. Dayal, K. Ramamritham and T. M. Vijayaraman, *Scaling clustering algorithms for massive data sets using data stream*, Proceedings of the 19th International Conference on Data Engineering, Bangalore, (2003).

[4] P. J. Rousseeuw, *A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational Appl Math, 20(1987), 53-65.

[5] N. V. Anand Kumar and G. V. Uma, *Improving Academic Performance of Students by Applying Data Mining Technique*, European Journal of Scientific Research, 34(4)(2009).

[6] E. Januzaj, H. P. Kriegel and M. Pfeifle, *Scaling clustering algorithms for massive data sets using data stream*, Proceedings of the Workshop on Clustering Large Datasets (ICDM '03), (2003).

[7] R. Sharmir and R. Sharan, *Algorithmic approaches to clustering gene expression data*, In current Topics in Computational Molecular Biology, (2002), 53-65.

[8] H. J. Mucha, *Adaptive cluster analysis, classification and multivarite graphics*, Weirstrass Institute for Applied Analysis and Stochastics, (1992).

[9] P. Berkhin, *A Survey of Clustering Data Mining Techniques*, Springer, USA, (2006).

[10] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, *An efficient enhanced k-means clustering algorithm*, Journal of Zhejiang University Science A., (2006), 1626-1633.

[11] P. Varapron, *Using Rough Set theory for Automatic Data Analysis*, 29th Congress on Science and Technology of Thailand, (2003).

[12] S. Sujit Sansgiry, M. Bhosle and K. Sail, *Factors that affect academic performance among pharmacy students*, American Journal of Pharmaceutical Education, (2006).

[13] D. K. Girija, *Data mining approach for prediction of fibroid disease using neural networks*, https://ieeexplore.ieee.org/document/6749370.

[14] D. K. Girija, *Data mining techniques used for uterus fibroid diagnosis and prognosis*, https://ieeexplore.ieee.org/document/6526439.