International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 10, December- 2012

Introduction of Clustering by using K-means Methodology

Niraj N Kasliwal[1], Prof Shrikant Lade[2], Prof Dr. S. S. Prabhune [3]

M-Tech, IT *HOD,IT* *HOD,IT*
RKDF *RKDF* *SSGMCE*
Bhopal,(India) *Bhopal,(India)* *Shegaon(India)*

Abstract

This paper presents the integrated data mining processing technique to find appropriate initial centroids in data clustering process by k-means algorithm. The processes include data cleansing, preprocessing, and finding features relation to get appropriate features. Our clustering process compares different initial selection schemes: static selection and random selection. We propose the K-means that represents the processes for finding appropriate initial clustering centroids and selecting the most relevant features from datasets. we can get better clustering result with k-means clustering methodology.

Keywords—Data mining , K-means clustering

I. INTRODUCTION

The data mining is the automatic process of searching or finding useful knowledge. The process extracts data from database with mathematics-based algorithm and statistic methodology to reveal the unknown data patterns that can be useful information. The information got from data mining process is very important knowledge that help user in decision making concerned business strategies. These processes are also called Knowledge Discovery in Database (KDD) in that knowledge discovery and analysis can be performed from many information and raw data in databases. The knowledge can be used in decision support system or used to predict customer's behavior or predict product sale rate in the future.

This paper studies various techniques to adapt and improve the data clustering methodology of the k-means clustering. The problems in data clustering with k-means are the selection of initial centroids . The research has focused on the working of k-means clustering methodology for selecting the centroids.
In this paper, the main idea of data mining technique in data clustering from raw data with appropriate initial centroids selection is presented. The techniques used in this paper for clustering is k-means clustering methodology.

K-means Clustering Methodology

The data clustering is processing of raw data to find clusters or groups of similar data. In each cluster, members have of each group.
2) Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.

some similarity in type of data. The principles of data clustering are finding value of score in similarity, and assigning each member to be in the same group of other members that have similar or same score.
The data mining technique in finding data clusters is different from data classification in that user does not have to specify target feature for assigning each data record to the appropriate cluster. Data clustering is thus an unsupervised learning method. The clustering method relies on the similarity measurement to automatically from groups of relevant or similar data members as visually shown in figure. After the clustering process, user can apply some classification algorithm to extract data pattern in each cluster for a better understanding of cluster model
.

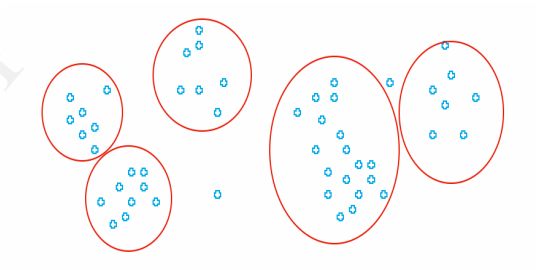


Fig: Clustering Visualization

K-means clustering algorithm is the most selected technique to cluster data. K-means is a nonhierarchical clustering and use looping to group data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The steps of k-means algorithm can be summarized as the following:
1) Specify group number and select initial centroid

3) Calculate distance's mean of every data member and own centroid to define new centroid in each group.

4) Repeat steps 2 and 3 until each group has stable centroid or same centroid.

Here will take a simple example for the clustering of datasets by using K means

Cluster the following eight points (with (x, y) representing locations) into three clusters 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). The distance function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as: $\rho(a, b) = |x2 – x1| + |y2 – y1|$.
Use k-means algorithm to find the three cluster centers after the second iteration.

Solution:

Iteration 1

		(2,10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

First we list all points in the first column of the table above. The initial cluster centers – means, are (2, 10), (5, 8) and (1, 2) - chosen randomly. Next, we will calculate the distance from the first point (2, 10) to each of the three means, by using the distance function:

point mean1
x1, y1 *x2, y2*
(2, 10) (2, 10)

$$\rho(a, b) = |x2 – x1| + |y2 – y1|$$

$$\begin{aligned}\rho(point, mean1) &= |x2 – x1| + |y2 – y1| \\ &= |2 – 2| + |10 – 10| \\ &= 0 + 0 \\ &= 0\end{aligned}$$

point mean2
x1, y1 *x2, y2*
(2, 10) (5, 8)

$$\rho(a, b) = |x2 – x1| + |y2 – y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 – x1| + |y2 – y1| \\ &= |5 – 2| + |8 – 10| \\ &= 3 + 2 \\ &= 5\end{aligned}$$

point mean3
x1, y1 *x2, y2*
(2, 10) (1, 2)

$$\rho(a, b) = |x2 – x1| + |y2 – y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 – x1| + |y2 – y1| \\ &= |1 – 2| + |2 – 10| \\ &= 1 + 8 \\ &= 9\end{aligned}$$

So, we fill in these values in the table:

		(2,10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 10) be placed in? The one, where the point has the shortest distance to the mean – that is mean 1 (cluster 1), since the distance is 0.

Cluster 1 Cluster 2 Cluster 3
(2, 10)

So, we go to the second point (2, 5) and we will calculate the distance to each of the three means, by using the distance function:

point mean1
x1, y1 *x2, y2*
(2, 5) (2, 10)

$$\rho(a, b) = |x2 – x1| + |y2 – y1|$$

$$\begin{aligned}\rho(point, mean1) &= |x2 – x1| + |y2 – y1| \\ &= |2 – 2| + |10 – 5| \\ &= 0 + 5 \\ &= 5\end{aligned}$$

point mean2
x1, y1 *x2, y2*
(2, 5) (5, 8)

$$\rho(a, b) = |x2 – x1| + |y2 – y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 – x1| + |y2 – y1| \\ &= |5 – 2| + |8 – 5| \\ &= 3 + 3\end{aligned}$$

= 6

point	mean3
x1, *y1*	*x2*, *y2*
(2, 5)	(1, 2)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\rho(point, mean2) = |x2 - x1| + |y2 - y1|$$
$$= |1 - 2| + |2 - 5|$$
$$= 1 + 3$$
$$= 4$$

So, we fill in these values in the table:

Iteration 1

		(2,10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 5) be placed in? The one, where the point has the shortest distance to the mean – that is mean 3 (cluster 3), since the distance is 0.

Cluster 1	Cluster 2	Cluster 3
(2, 10)		(2, 5)

Analogically, we fill in the rest of the table, and place each point in one of the clusters:

Iteration 1

		(2,10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2,10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

Cluster 1	Cluster 2	Cluster 3
(2, 10)	(8, 4)	(2, 5)
	(5, 8)	(1, 2)
	(7, 5)	
	(6, 4)	
	(4, 9)	

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

For Cluster 1, we only have one point A1(2, 10), which was the old mean, so the cluster center remains the same.

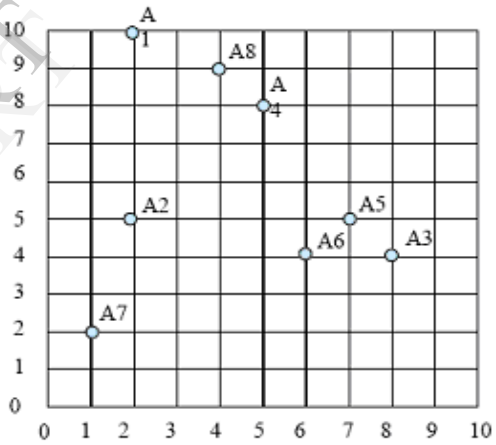
For Cluster 2, we have ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)

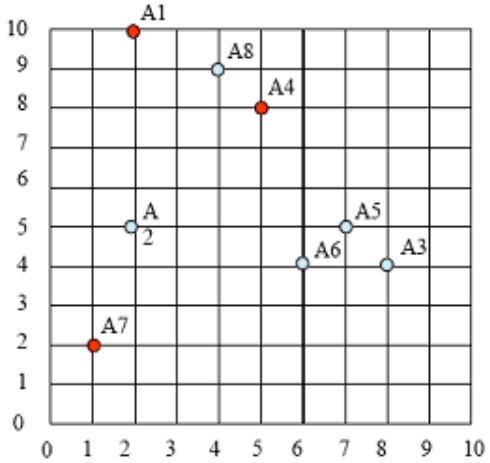
For Cluster 3, we have ((2+1)/2, (5+2)/2) = (1.5, 3.5)

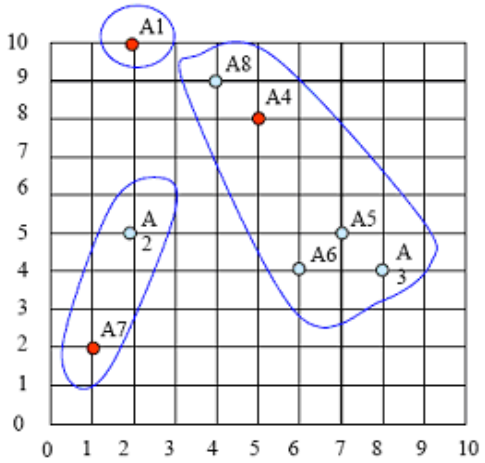
New clusters : 1: {A1}, 2 : {A3,A4, A5, A6, A8}, 3 : {A2,A7}

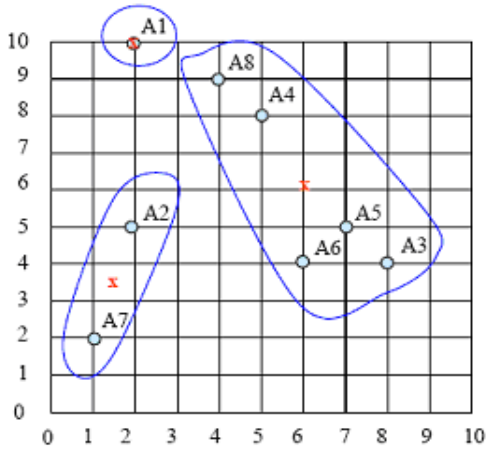
b) Centers of the new clusters :
c1=(2,10) , C2=((8+5+7+6+4)/5,(4+8+5+4+9)/5)=(6,6) , C3=((2+1)/2,(5+2)/2)=(1.5,3.5)

C)







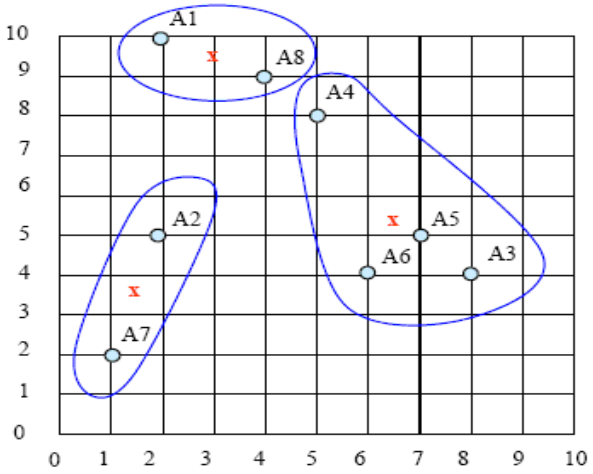


The initial cluster centers are shown in red dot. The new cluster centers are shown in red x.

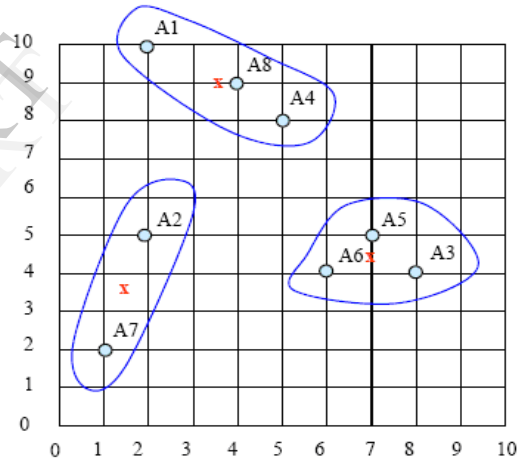
That was Iteration1 (epoch1). Next, we go to Iteration2 (epoch2), Iteration3, and so on until the means do not change anymore.
In Iteration2, we basically repeat the process from Iteration1 this time using the new means we computed.

d) We would need to more epochs . After the 2nd epoch the result would be :

1: {A1,A8}, 2 : {A3,A4, A5, A6}, 3 : {A2,A7} with centers C1=(3 , 9.5) C2= (6.5 , 5.25) C3= (1.5 , 3.5)



After the 3rd epoch , the result would be :
1: {A1,A4,A8} 2 : {A3, A5, A6,} 3 : {A2,A7}
with centers C1=(3.66 , 9) C2= (7 , 4.33) C3= (1.5 , 3.5)



2 CONCLUSION

The data mining has many techniques available for users to apply to suitable data types and usage. From this research we present one of unsupervised data mining technique called data clustering that integrated other mining technique.

3 REFERENCES :

[1] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE VOL. 24, NO. 7, JULY 2002

[2] Tan , Steinbach and Kumar , on "Introduction to Data Mining" Apr 2004.
[3]Tan,Steinbach and Ghosh Kumar ,on " Top Ten Data Mining Algorithms" Dec 2006.
[4] Pradeep Rai & Shubha Singh," A Survey of Clustering Techniques",IJCA, *Volume 7– No.12, October 2010*
[5] *S. Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama on "*A Survey on Partition Clustering Algorithms*"* IJECBS
Vol. 1 Issue 1 January 2011
[6] Noppol Thangsupachai, Phichayasini Kitwatthanathawon, Supachanun Wanapu, and Nittaya Kerdprasop , on " Clustering Large Datasets with Apriori-based Algorithm and Concurrent Processing" ,IMECS 2011, March 16-18 , 2011,Hong Kong.
[7] www.faculty.uscupstate.edu/atzacheva/SHIM450/KMeansExample.doc