"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 8, No. 2, pp. 163–167
P-ISSN: 2069-0932, E-ISSN: 2066-1061
© 2016 Pro Universitaria
www.orizonturi.ucdc.ro

EXPANDING K-MEANS ALGORITHM FOR ABSOLUTE DATA

Ana-Maria Ramona STANCU[1], **Mihaela MOCANU**[2]

[1] *Academy of Economic Studies, Romania, e-mail: ana_maria_ramona@yahoo.com*
[2] *Dimitrie Cantemir "Christian University",Romania, e-mail: rmocanu99@yahoo.fr*

Abstract *In the majority of works published so far on k-means algorithm, the study was performed on numerical data and functions with which the distance between the data points can be calculated. Recently, as far as the clustering issue is concerned, the problem of using absolute data has also been raised, and the algorithms used so far have been considered unacceptable for their implementation in large databases. This article aims to apply accurately the "notion of the cluster center" on a set of absolute objects and how it is used in issues related to absolute objects grouping.*

Key words:
algorithm, center, cluster, data mining, distant, object

JEL Codes:
L8

1. Introduction

This algorithm used in cluster analysis based on partitioning methods was first proposed in 1957 by Stuart Lloyd, developed in 1967 by MacQueen, being followed by a more efficient version in 1975/1979 which was proposed and published by Fortran Hartigan and Wong. The K-Means algorithm is a partitioning-based method consisting of spherical groups. It makes use of statistical methods used in grouping attributes. The K-Means algorithm is a numerical algorithm aimed to find all the positions that clusters occupy minimizing the distance between them and the data points. The algorithm divides the n objects into k partitions (clusters), and each partition represents a group, being easy to implement and applicable to large data sets.

In order to achieve this algorithm, the following steps must followed:

1. K clusters must be created by choosing them out of a number of random data
2. The arithmetic average must be calculated for each group format
3. Each entry is assigned to the nearest cluster using the similarity calculus formula, namely:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)}$$

Formula 1 - Euclidean distance

Each attribute is assigned a cluster, then the arithmetic average of all groups from the data set is recalculated. This process continues with step 3 until no point can be moved and the procedure is complete.

Namely, the algorithmcan can be written as follows:

```
Algorithm k-means(k, D)
Input
k (no of clusters)
Choose k as initial centroids (clusters centers)
repeat
        for each data x do
                distance (x, centroid); // Calculate the distance from x to each centroid
                Associate x to the nearest centroid // the centroid represents a cluster;
        endfor
        the centroid is associated with the help of the current members of the cluster
until the stopping criterion is fulfilled.
```

In 2009, Pachira modified the K-Means algorithm so that he moved the center of each group, making sure there were no bare clusters. The comparison between the original and modified algorithm has shown that the original algorithm has a greater number of iterations than the one which was modified. For the numerical examples that produce bare clusters, the method proposed by him cannot be compared with any other method because the algorithm avoids the bare clusters [2]. Methods for achievieng the clustering algorithm were proposed by Bradley and Fayyad in 1998, by Wu in 2008, and in 2004 the Khan and Ahmed algorithm was proposed, which creates complex procedure and it is quite expensive [3],[4],[5].

Consider $T \subset R^d$ data set of *d* dimension with *n* points and consider *k* no of clusters with $k \in Z$ and *k>1*. The clustering procedure is given by $S = \{s_1, s_2, \ldots, s_k\}$ data groups which are divisible by *k* and c_i clusters with $i \in Z$ and

$C_i = \{c_1, c_2, \dots, c_k\}$. The K-Means algorithm must comply with the following conditions [6]:

1. $S_i \neq$ void lot
2. $\bigcup_{i'1}^{k} S_i = X$
3. $S_i \cap S_j = mulțulțime vidă, i \neq j; i, j = \overline{1,k}$

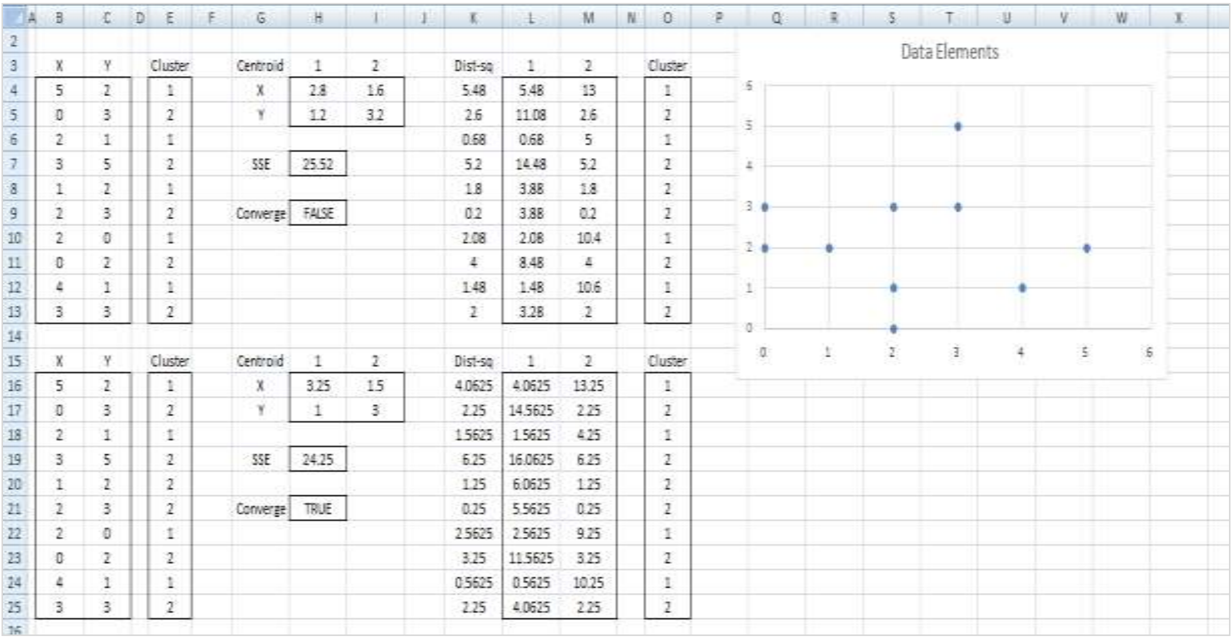
And the Euclidian distance is given by:

$$d(X_i, X_j) = \sqrt{\sum_{i=1}^{d} (x_i - x_j)^2}$$

$$d(X_i, X_j) = \|x_i - x_j\|$$

2. ALGORITHM IMPLEMENTATION

To begin with, I would like to implement the K-Means algorithm on a S data set with k clusters, noted with *c* and *n* tuples.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
2															
3		X	Y		Cluster		Centroid	1	2		Dist-sq	1	2		Cluster
4		5	2		1		X	2.8	1.6		5.48	5.48	13		1
5		0	3		2		Y	1.2	3.2		2.6	11.08	2.6		2
6		2	1		1						0.68	0.68	5		1
7		3	5		2		SSE	25.52			5.2	14.48	5.2		2
8		1	2		1						1.8	3.88	1.8		2
9		2	3		2		Converge	FALSE			0.2	3.88	0.2		2
10		2	0		1						2.08	2.08	10.4		1
11		0	2		2						4	8.48	4		2
12		4	1		1						1.48	1.48	10.6		1
13		3	3		2						2	3.28	2		2
14															
15		X	Y		Cluster		Centroid	1	2		Dist-sq	1	2		Cluster
16		5	2		1		X	3.25	1.5		4.0625	4.0625	13.25		1
17		0	3		2		Y	1	3		2.25	14.5625	2.25		2
18		2	1		1						1.5625	1.5625	4.25		1
19		3	5		2		SSE	24.25			6.25	16.0625	6.25		2
20		1	2		2						1.25	6.0625	1.25		2
21		2	3		2		Converge	TRUE			0.25	5.5625	0.25		2
22		2	0		1						2.5625	2.5625	9.25		1
23		0	2		2						3.25	11.5625	3.25		2
24		4	1		1						0.5625	0.5625	10.25		1
25		3	3		2						2.25	4.0625	2.25		2
26															

In the figure above it can be seen that we have chosen $X = \{x_1, x_2, \dots, x_{10}\}$, $Y = \{y_1, y_2, \dots, y_{10}\}$ and the number of clusters equal to 2. From the beginning we established each cluster centroids to be the average of all the elements in the group. In our case, the centroid of the first cluster is (2.8, 1.2), and the centroid for the second cluster is (4.6, 3.2), where: 2.8 is situated in the cell H4 („= AVERAGEIF (E4:E13,H3,B4:B13)"), 1.2 H5, 1.6 is situated in the cell I4 and 3.2 is situated in the cell I5. Further on we calculated the square distance of each of the elements for each centroid.

Namely, For L4 the formula is :

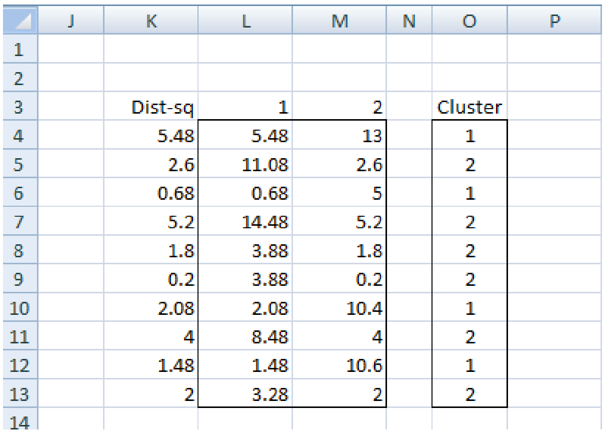
=(B4-H4)^2+(C4-H5)^2,

and for M4 the formula is:

=(B4-I4)^2+(C4-I5)^2

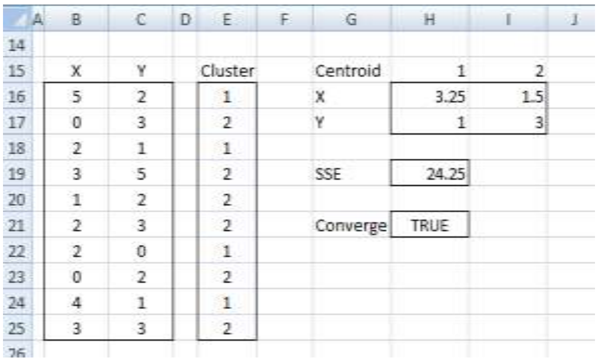
K4 =MIN(L4:M4)

O4 =IF(L4<=M4,1,2)



	J	K	L	M	N	O	P
1							
2							
3		Dist-sq	1	2		Cluster	
4		5.48	5.48	13		1	
5		2.6	11.08	2.6		2	
6		0.68	0.68	5		1	
7		5.2	14.48	5.2		2	
8		1.8	3.88	1.8		2	
9		0.2	3.88	0.2		2	
10		2.08	2.08	10.4		1	
11		4	8.48	4		2	
12		1.48	1.48	10.6		1	
13		2	3.28	2		2	
14							

We shall go on in a similar way in order to determine a reassigning of groups for the 10 items, and the value of this assignment is given by the SSE which is 25.52 and the algorithm is not convergent, which means that the process will continue. It can be noticed that in the next stage SSE is 24.25 and the algorithm is convergent, which indicates that it is the last stage in the implementation of the algorithm.



	A	B	C	D	E	F	G	H	I	J
14										
15		X	Y		Cluster		Centroid	1	2	
16		5	2		1		X	3.25	1.5	
17		0	3		2		Y	1	3	
18		2	1		1					
19		3	5		2		SSE	24.25		
20		1	2		2					
21		2	3		2		Converge	TRUE		
22		2	0		1					
23		0	2		2					
24		4	1		1					
25		3	3		2					
26										

3 THE ALGORITHM SUGGESTED

Let be the vectors $X,Y\in R^d$ and $X=(x_1,x_2,...,x_k)$ and $Y=(y_1,y_2,...,y_k)$ with $i=\overline{1,k}$ and $d(x,y)$ the euclidian distance

$$d(x_i,x_j)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2}=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_k-y_k)^2}, \tag{3.1}$$

Namely,

$$\alpha(x,y)=d(x,y)=\sum_{i=1}^{d}(x_i-y_i)^2=(x-y)^T*(x-y)$$

Let $\alpha(X',z)$ be the euclidian distance between the subspace points from X ($X'\subset X$)and zwhich is the center.

Lemma 1: Let $X'\subset X$ be subset of date from X and $X\in R^d$ and let be $z'\subset X$ with z centroid. Then there is the relation:

$$\alpha(X',z'')=\alpha(x',z')+|X|*\alpha(z',z''), \tag{3.2}$$

Demonstration:

We will extend space $\alpha(X',z'')$, namely:

$$\alpha(X',z'')=\sum\alpha(x,z'')=\sum(x-z'')*(x-z''), \tag{3.3}$$

$$=\sum((x-z')+(z'-z''))*((x-z')+(z'-z'')), \tag{3.4}$$

$$=\sum[(x-z')*(x-z')+(x-z')*(z'-z'')+(x-z')*(z'-z'')+(z'-z'')*(z'-z'')]$$

$$=\sum((x-z')*(x-z')+2*((x-z')*(z'-z''))+((z'-z'')*(z'-z''))) \tag{3.5}$$

But $\alpha(X',z')=\sum(x-z')*(x-z')$ (3.5.1)

From (3.5) and (3.5.1)we will have:

$$\alpha(X',z'')=\alpha(X',z')+2*((x-z')*(z'-z''))+((z'-z'')*(z'-z''))$$

$$\alpha(X',z'')=\alpha(X',z')+2*(z'-z'')*\sum(x-z')+|X|*(z'-z'')*(z'-z'')$$

$$\alpha(X',z'')=\alpha(X',z')+2*(z'-z'')*\sum(x-z')+|X|*\alpha(z',z'') \tag{3.6}$$

But $\alpha(z',z'')=(z'-z'')*(z'-z'')$ (3.7)

and $\sum(x-z')=0$ (3.8)

From (3.6), (3.7) and (3.8) we have: $\alpha(X',z'')=\alpha(X',z)+|X'|*\alpha(z',z'')$

This relation is real if there is z center with $z\subset X'$, then $\sum(X-z)=0$

Lema 2: Let $X\in R^d$ be set of dataand $z\subset X$ with z center. For $\forall z'\in R^d$ there is the relation:

$$\alpha(x,z')=\alpha(X,z)+|X|*\alpha(z,z')$$

Demonstration

We shall extend $\alpha(X, z')$, namely:

$$\alpha(X, z') = \sum \alpha(x, z') = \sum (x - z')*(x - z') \quad (3.9)$$

$$= \sum [(x - z) + (z - z')]*[(x - z) + (z - z')] \quad (3.10)$$

$$= \sum [(x - z)*(x - z) + (x - z)*(z - z') + (x - z)*(z - z') + (z - z')*(z - z')]$$

$$= \sum [(x - z)*(x - z) + 2*(x - z)*(z - z') + (z - z')*(z - z')]$$

But $\alpha(X, z) = \sum (x - z)*(x - z)$

Therefore $\alpha(X, z') = \alpha(X, z) + 2*(z - z')*\sum(x - z) + |X|*(z - z')*(z - z')$ (3.11)

But $(z - z')*(z - z') = \alpha(z, z')$ (3.12)

$\sum (x - z) = 0$ (3.13)

from (3.11), (3.12) and (3.13)we will have:

$\alpha(X, z') = \alpha(X, z) + |X|*\alpha(z, z')$ (3.14)

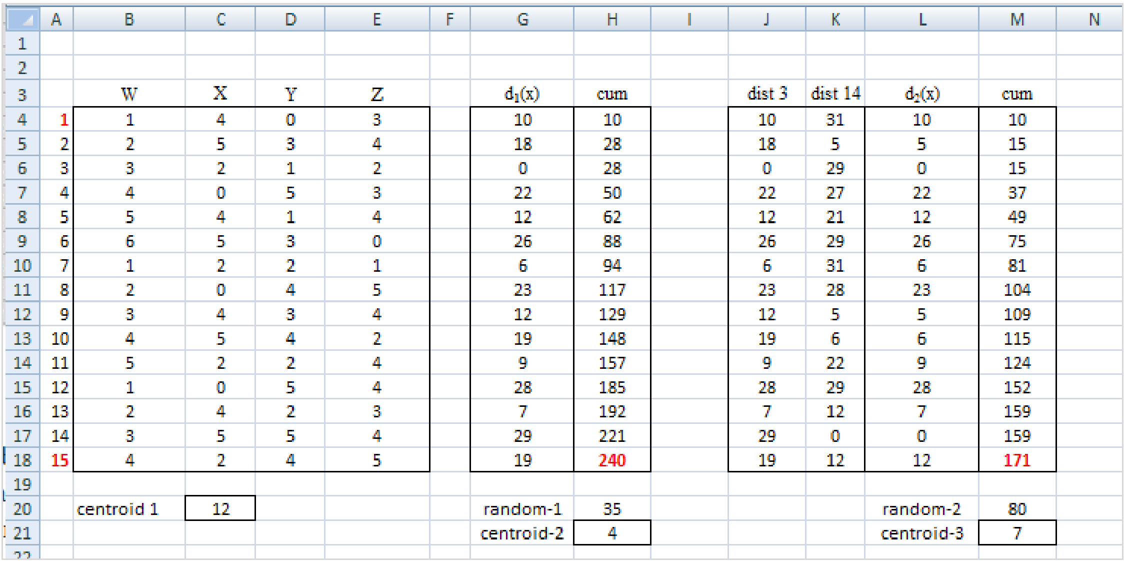
The expression is real if and only if $z \in X$ is the center and $\sum(x - z) = 0$

In order to achieve the algorithm the following steps need to be followed:

1. We must choose a centroid c randomly out of a set of data S
2. For $\forall x \in S$ we calculate the square root between the minimum distance x and the centroids (z) chosen. That is, $\forall (z_1, z_2, ..., z_k)$ centroids, then define $dist_k(x) = (dist^2(x, z_i))$, for $\forall i$
3. An item is chosen from the lot $S - (z_1, z_2, ..., z_k)$ and z_{k+1} is a centroid chosen randomly and $\forall x$ so that, X probability should be equal to $\frac{dist_k(x)}{\sum dist_k(y)}$, that is $P(X) = \frac{dist_k(x)}{\sum dist_k(y)}$
4. Step 2 is repeated until the centoid *k* is found.

Suggested application

According to the algorithm described we carry out an application where we have 3 clusters and 4 tuples(W,X,Y,Z).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2														
3		W	X	Y	Z		$d_1(x)$	cum		dist 3	dist 14	$d_2(x)$	cum	
4	1	1	4	0	3		10	10		10	31	10	10	
5	2	2	5	3	4		18	28		18	5	5	15	
6	3	3	2	1	2		0	28		0	29	0	15	
7	4	4	0	5	3		22	50		22	27	22	37	
8	5	5	4	1	4		12	62		12	21	12	49	
9	6	6	5	3	0		26	88		26	29	26	75	
10	7	1	2	2	1		6	94		6	31	6	81	
11	8	2	0	4	5		23	117		23	28	23	104	
12	9	3	4	3	4		12	129		12	5	5	109	
13	10	4	5	4	2		19	148		19	6	6	115	
14	11	5	2	2	4		9	157		9	22	9	124	
15	12	1	0	5	4		28	185		28	29	28	152	
16	13	2	4	2	3		7	192		7	12	7	159	
17	14	3	5	5	4		29	221		29	0	0	159	
18	15	4	2	4	5		19	240		19	12	12	171	
19														
20		centroid 1	12				random-1	35				random-2	80	
21							centroid-2	4				centroid-3	7	

The first centroid was chosen randomly with the help of the formula=RANDBETWEEN(A4,A18), centroid situated in the cell C20 and its value is 12. In order to find the second centroid, we will calculate the square distance of all the elements towards the first centroid, namely, in the cell G4 we will calculate by means of the formula "= SUMXMY2 (B4: E4, B $ 6: E $ 6)". I know that these distances show the shares that occur in a random selection of the second centroid.To this end, in the area H14:H18 we calculated the cumulative values using the formula "=RANDOM (1, H18, TRUE)" named random-1 situated in the cell H20 which shows to be 35, and the centroid which is in the cell H21 will be calculated using the formula:

"=IFERROR(MATCH(H20,H4:H18,0), IFERROR(MATCH(H20,H4:H18,1)+1,1))". In order to calculate the third centroid we will first calculate the square distance between the the elements and the first centroid and the square distance between the elements and the second centroid. The share is located in cell M20 and we calculated it using the formula "=RANDOM (1,M18, TRUE)"., and the third centroid is situated in cell M2, has the value 7 and we calculated it using the formula

„=IFERROR(MATCH(M20,M4:M18,0),IFERROR(MATCH(M20,M4:M18,1)+1,1))"

By making use of the results. the initial clusters can be found.

4 DISCUSSIONS

In this article we have the studied the K-Means algorithm, that is we carried out an application using this algorithm where we applied the Euclidean distance. It can be noticed that we have chosen X and Y with i = 1 to 10, and the cluster number is 2. We have carried out two iterations: in the first iteration we found the first cluster centroid as (2.8, 1.2) and the center of the second cluster is (4.6, 3.2).

In the second iteration we acted similarly in order to determine the new allocation, in which case I noticed that the value is 25.52 by applying the SSE, what proved that the algorithm is convergent, that is the process continues.

Based on the above mentioned findings, we made an improvement thereof, that is we carried out an application where we increased the number of clusters and that of tuples, in order to find out, in the end, whether the algorithm is convergent or not, or if we can find the initial clusters, that is in this case the algorithm is non-convergent.

As described in the application, we designed an algorithm, which is based on the dataset and the number of centroids, having to find out, once the iterations designed, if the algorithm is convergent or if we can find the original clusters, in this case the algorithm is non-convergent.

CONCLUSIONS

We carried out the same application imposing the number of iterations. For the first application we took 10 as the number of iterations, while for the second one it is 50. It can be noticed that the greater the number of iterations is, the lower the SSE.

REFERENCES

[1] Juanying Xie, Shuai Jiang 2010, "A Simple and Fast Algorithm for Global K-means Clustering", 2010 Second International Workshop on Education Technology and Computer Science, 978-0-7695-3987-4/10 $26.00 © 2010 IEEE DOI 10.1109/ETCS.2010.347

[2] Pakhira, Malay K.: "A Modified k-means Algorithm to Avoid Empty Clusters". International Journal of Recent Trends in Engineering, Vol.1, No. 1, May (2009).

[3] Bradley, P. S. and Fayyad, U. M.: "Refining Initial Points for Kmeans Clustering", Technical Report of Microsoft Research Center, Redmond, California, USA, (1998)

[4] Wu, F. X.: "Genetic Weighted k-means Algorithm for Clustering Large-Scale Gene Expression Data". BMC Bioinformatics, vol. 9, (2008)

[5] Khan, S. S. and Ahmed, A.: "Cluster Center Initialization for Kmeans Algorithm. Pattern Recognition" Letters, vol. 25, no. 11, pp. 1293-1302, (2004)

[6] Ph.D., Brunel University (United Kingdom), 1989, M.Sc., University of Technology (Iraq), 1978, B.Sc., University of Technology (Iraq), 1976Dip.Tch., Auckland University (New Zealand), 1996, CONTRIBUTIONS TO K-MEANS CLUSTERING AND REGRESSION VIA CLASSIFICATION ALGORITHMS, Virginia Commonwealth University Richmond, Virginia, April, 2012