A Mid – Point based k-mean Clustering Algorithm for Data mining

Neha Aggarwal
Department Of Computer Science
MRCE
Faridabad, India
aggarwal.neha83@gmail.com

Kirti Aggarwal
Department Of Computer Science
MRCE
Faridabad, India
kirtibansal06@gmail.com

***Abstract*—In k-means clustering algorithm, the number of centroids is equal to the number of the clusters in which data has to be partitioned which in turn is taken as an input parameter. The initial centroids in original k-means are chosen randomly from the given dataset and for the same dataset different clustering results are produced with different randomly chosen initial centroids. This paper presents a solution to this limitation of the original K-means Algorithm.**

Keywords-K-means,centroids,mid-pont,clustering, computationally expensive.

I. INTRODUCTION

In cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e., chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups., Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering. The k-means algorithm [6, 7, 8, 10, 11] is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high, especially for large data sets. Moreover, this algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers for improving the performance of the K-Means clustering algorithm. This paper deals with a method for improving the accuracy and efficiency of the k-means algorithm.

II. ORIGINAL K-MEANS ALGORITHM

This section describes the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering. The k-means algorithm is a popular method for automatically classifying vector-based data. It's used in various applications such as vector quantization, density estimation, workload behaviour characterization, image compression, and automatic topic identification, among many others. It's been identified as one of the top-10algorithms in data mining [13].

Pseudo code for the k-means clustering algorithm is listed as

Algorithm 1 [14].

Basic K-Means Algorithm:[16]

1: Select K points as initial centroids

2: repeat

3: Assign each point to its closest centroid to form K clusters

4: Recompute the centroid of each cluster

5: until Termination condition is met

To determine a distance from a centroid c to the given point x an Euclidean metric is used, although other metrics are also possible:

$d(x; c) = \sqrt{[(xi-ci)2]}$

For algorithm termination few termination conditions that

may be used

• Maximum number of iterations is reached,

• Centroids converged to (local or global) minima, i.e. there is no further change in centroids positions.

LIMITATIONS OF K-MEANS ALGORITHM

Even after being so popular algorithm for clustering, K-Means Algorithm [16]suffers from two major limitations

1. It is computationally very expensive as it involves several distance calculations of each data point from all the centroids in each iteration.
2. The final cluster results heavily depend on the selection of initial centroids which causes it to converge at local optimum.

The first limitation has been solved using the enhanced K-means algorithm in our paper at an international conference on emerging trends in engineering and management.[16],[17]

The second limitation states that the results of k-means clustering vary along the initial centres chosen . So, the solution of this problem is to find a way to chose the initial centres. This paper proposes the algorithm to chose initial centres. To apply k-means algorithm to a certain clustering problem, first of all we must determine the initial centroids. The number of initial centroids is equal to the number of desired clusters (k) in which the given input data is to be partitioned. This number k is determined by the user i.e. this is an input parameter.

In the original k-means algorithm the initial centroids are taken just randomly out of the input data set. But this random selection of initial centroids leads the computation of the algorithm into local optima. Say, k is determined to be 3. If from a given data set we select first 3 points as initial centroids and compute the k-means algorithm. Next time suppose we select the last 3 points as the initial centroids and further third time let we select any 3 arbitrary data points as initial centroids and compute the k-means algorithm. Each time the end clustering results will come out to be different. Then we have to analyze which one is the most appropriate result. Thus with the random selection of initial centroids there is no guarantee that the k-means algorithm will converge into best results. This is the limitation which needs to be dealt with in order to make the k-means algorithm more efficient.

This can be shown with an example dataset of 8 employees giving different clustering results with different initial centres.

Table 1. Employee dataset

EMPLOYEE	ATTRIBUTE1:X (Experience in No. of yrs)	ATTRIBUTE2:Y (Salary in Lacs/annum)
Emp1	0.5	2
Emp2	1	3
Emp3	2	3.5
Emp4	3	4
Emp5	3.5	5
Emp6	4	6
Emp7	4.5	7
Emp8	5	7.5

Now, after applying K-means clustering algorithm on this dataset for k=3 and initial centres as c1 (0.5, 2) ,c2 (1,3) , c3(2, 3.5) the result is

Table 2 clustering results with first set of initial centres.

EMPLOYEE	ATTRIBUTE1:X (Experience in yrs)	ATTRIBUTE2:Y (SalaryLacs/annual)	GROUP
Emp1	0.5	2	1
Emp2	1	3	2
Emp3	2	3.5	2
Emp4	3	4	2
Emp5	3.5	5	3
Emp6	4	6	3
Emp7	4.5	7	3
Emp8	5	7.5	3

Now, when we take different initial centres with same k=3 and initial centres as c1(0.5,2),c2(3,4) and c3(5,7.5) the results are

Table 3. clustering results with second set of initial centres.

EMPLOYEE	ATTRIBUTE1:X (Experience in yrs)	ATTRIBUTE2: Y (Annul Salary in Lacs)	GROUP
Emp1	0.5	2	1
Emp2	1	3	1
Emp3	2	3.5	2
Emp4	3	4	2
Emp5	3.5	5	2
Emp6	4	6	3
Emp7	4.5	7	3
Emp8	5	7.5	3

Thus different initial centres give different results. So solve this problem the algorithm to find initial centers is proposed.

III. ENHANCING K-MEANS WITH IMPROVED INITIAL CENTER USING MID-POINT METHOD

In this algorithm a systematic method to determine the initial centroids is explained. This method is quite efficient to produce good clusters using k-mean method, as compared to taking the initial centroids randomly.

The algorithm is as under:

Algorithm1: The enhanced method

Input: D= Set of n data points.

K=desired number of clusters

Output: k number of initial centroids

Steps:

1. In the given data set D, if the data points contain both the positive and negative attribute values then goto step 2, else goto step 4.
2. Find the minimum attribute value in the given dataset D.
3. For each data point attribute, subtract with the minimum attribute value.
4. For each data point calculate the distance from origin.
5. Sort the distances obtained in step 4. Sort the data points in accordance with the distances.
6. Partition the sorted data points into k equal sets.
7. In each set , take the middle point as the initial centroid.

In the above algorithm, if the input data set contains the negative value attributes, then all the attributes are transformed to positive space by subtracting each data point attribute with the minimum attribute value in the data set. This transformation is required because in the algorithm the distance from origin to each data point is calculated. So if there are both positive and negative values, then for different data points same Euclidean distance will be obtained which will result in incorrect selection of initial centroids. After transforming all attribute values to positive, next step is to calculate the distance of each point from the origin. Then the

original data points are sorted into k equal sets. In each set, the mid-point is calculated. All the mid-points are taken as the initial centroids.

IV. SCENARIO OF THE MID-POINT METHOD OF FINDING INITIAL CENTROIDS

Following is the example of k-mean algorithm using the enhanced method. The input data set contains 16 entities (condensing machines) which are described by two attributes: - Condensing temperature and corresponding pressure. The input parameter k is taken as 4. i.e. all the 16 entities have to be categorized into 4 clusters based on their efficiency.

Table 4. dataset of condensing machines

MACHINES	X ATTRIBUTE	Y ATTRIBUTE
Machine 1	15	58
Machine 2	50	93
Machine 3	25	130
Machine 4	40	130
Machine 5	25	165
Machine 6	50	170
Machine 7	25	225
Machine 8	60	220
Machine 9	40	250
Machine 10	43	270
Machine 11	50	280
Machine 12	50	320
Machine 13	43	360
Machine 14	60	360
Machine 15	60	405
Machine 16	60	540

Step 1: There are no negative values in the given data set, so goto step 4.

Step 4, 5, 6 ,7 : After calculating the distance of each data point from origin, the distances and the corresponding data points are sorted. Then they are divided into 4 equal groups. Then the mid-point of each group is taken:

Table 5. midpoint of each subset

Distance from origin	X attribute	Y attribute	Mid-point
59.91	15	58	(27.5,94)
105.59	50	93	
132.38	25	130	
136.01	40	130	
166.88	25	165	(42.5,192.5)
177.20	50	170	
226.38	25	225	
228.04	60	220	
253.18	40	250	(45,285)
273.40	43	270	
284.43	50	280	
323.88	50	320	
362.56	43	360	(51.5,450)
364.97	60	360	
409.42	60	405	
543.32	60	540	

Now using the calculated mid-points for each group as the initial 4 centroids, apply the k-mean algorithm on the input data. After three iterations of the k-mean algorithm, stability was achieved. The resulting clusters are as under:

Table 6.clustering results

MACHINES	X ATTRIBUTE	Y ATTRIBUTE	RESULTING CLUSTER
Machine 1	15	58	1
Machine 2	50	93	1
Machine 3	25	130	1
Machine 4	40	130	1
Machine 5	25	165	2
Machine 6	50	170	2
Machine 7	25	225	2
Machine 8	60	220	2
Machine 9	40	250	2
Machine 10	43	270	3
Machine 11	50	280	3
Machine 12	50	320	3
Machine 13	43	360	3
Machine 14	60	360	3
Machine 15	60	405	4
Machine 16	60	540	4

V. CONCLUSION

The K-Means Algorithm suffers from the two major limitations of being computationally very expensive as it involves several distance calculations of each data point from all the centroids in each iteration and secondly the final cluster results heavily depends on the selection of initial centroids which causes it to converge at local optimum. This paper presents the way to find the initial centres for the k-means so that every time the K-Means Algorithm produces same result for the same dataset and remove the second limitation of K-Means of producing different clustering results with different initial centroids.

VI. REFERENCES

[1] Amir Ben-Dor, Ron Shamir and Zohar Yakini, "Clustering Gene Expression Patterns," Journal of Computational Biology, 6(3/4): 281-297, 1999

[2] Chaturvedi J. C. A, Green P, "K-modes clustering," J. Classification, (18):35–55, 2001.

[3] Daxin Jiang, Chum Tong and Aidong Zhang, "Cluster Analysis for Gene Expression Data," IEEE Transactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

[4] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.

[5] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, (2):283–304, 1998.

[6] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

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

[8] McQueen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., (1):281–297, 1967.

[9] Merz C and Murphy P, UCI Repository of Machine Learning

[10] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[11] Stuart P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory, 28(2): 129-136.

[12] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conferenceon Machine Learning and Cybernetics, pages 26–29, August 2004.

[13] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. Knowl. Inf. Syst., 14(1):1–37, 2008.

[14] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and AngelaY. Wu. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell., 24(7):881–892, 2002

[15] Kirti Aggarwal, Neha Aggarwal, Priyanka Makkar" Analysis of K-Means Clustering Algorithm for Data Mining", national conference on emerging trends in electronics and infotmation technology,2012,"unpublished". R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[16] Kirti Aggarwal,Neha Aggarwal,Sunita Bhardwaj,Nikita Taneja: An Effective Enhanced k-mean Clustering Algorithm for Data Mining, international cofrence on emerging trends in engineering and management,"in press".

[17] Kirti Agarwal,Neha Aggarwal, Kanika Gupta: Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining,intrnational journal of compuer applications,"unpublished"

AUTHORS PROFILE.

Neha Aggarwal is pursuing her M.Tech,final year(computer Science and Engineering) degree from Manav Rachna College of Engineering,Maharishi Dayanand University . Her areas of interest are data mining, clustering

Kirti Aggarwal received M.Tech (Computer Science and Engineering) degrees with Hons. from Maharshi Dayanand University in 2009. Presently, she is working as an Assistant Professor in Computer Science and Engineering Department in Manav Rachna College of Engineering, Faridabad. Her areas of interest are Data Mining, Clustering