Double Centers based Efficient Initial Clusters for K-Means

[1]Aarti Chaudhary, [2]Dharmveer Singh Rajpoot

[1, 2]Department of Computer Science Engineering / Information Technology
[1, 2]Jaypee Institute of Information Technology, A – 10, Sector – 62, Noida, India
[1]chaudhary.aarti111@gmail.com,
[2]dharmveer.rajpoot@jiit.ac.in

Abstract

Now a days, K-Means is one of the widely used algorithm for partitioning the data into clusters. The main advantage is easy to use and understand. By seeing the drawback side, the random selection of clustering initial centers will not always the optimal clustering structure. To acquire optimal clustering the initial centers selection is also efficient one. In the proposed method by using twice operations of K-means would give us good start of clustering. From the experiments we can show that intra-cluster i.e. similarity within the clusters is maximized and inter-cluster i.e. dissimilarity between the clusters is minimized which proves our propose algorithm choosing the initial centers effectively.

Keywords:-Cluster; initial center; dataset; validity measure; k-means.

I. INTRODUCTION

In today's era lots of raw data like web log data, E-commerce, purchases at stores is being accessed, updated and stored in the warehouse. To find the useful information from the raw data by using the traditional techniques like data collection, data access, data warehousing and decision support is quiet infeasible [2]. As data mining gives the expected results of future with anxious data transmission. Mining the data to get useful information, which cannot be finding by human analysts also? The analysts cannot analyze all of the data present in the warehouse. So, an autonomous or semi-autonomous method is employed to get useful and quantitative information from raw data warehouse. The data mining consists of prediction and description methods [2]. The prediction methods used to predict the future values of other values by using some variables. It includes classification, regression and deviation detection. The description methods are used to find patterns that best describe the dataset. It includes clustering, association rule and sequential pattern discovery.

Clustering is one of the popular techniques that classify the dataset into small subsets of data called as clusters or groups. The main aim of clustering is to maximize the intra-cluster distance and minimize the inter-cluster [1]. There are different ways to cluster the dataset i.e. partitioning and hierarchical. In hierarchical the approach is to start clustering in divisive i.e. top to bottom approach, first all the dataset considers as one cluster and splitting one by one by seeing the similarity measure between the data points. On the other hand, the agglomerative is bottom to top approach, all data points are considered as singleton cluster and then tries to merge the clusters according to proximity measure. The main drawback of hierarchical clustering is that there is no movement between the clusters. On the other hand, partitioning clustering start with the selection of initial centers by using them we can cluster the dataset. So, the starting point selection is quiet complex process. In past many methods were discovered the selection of initial centers process. K-Means, K-Medoids is one of well-known technique to find initial centres on random basics. The difference comes with the selection of intermediate centers in case of K-Means the centers come from the mean of the data point present in the clusters but, K-Medoids only selects the centers from the dataset only. The drawback with the traditional technique is that error minimization changes with every run of method. The wrong initial centers are selected leads to local cluster results.

The paper is formulated as follows: In section II indicates the related study of existed initial centers problem. Section III, shows our proposed algorithm and its pseudo code. Section IV consists of experimental setup used for the implementation and datasets on which clustering takes place and proves the algorithm by comparing via using the cluster validity measure. Section V includes conclusion and scope of future work.

II. RELATED STUDY

There have been a number of methods to handle the initial problem. Juifang Chang [3] uses the re-centroid separation technique to remove the unstable results that comes from firstly choosing a set of incorrect initial centroids and lastly the data points contains non-circular patterns of different size and pattern. Yunming Ye [4] proposed the NK-Means which is integration of NBC (Neighborhood Based clustering) and K-Means. In the starting initial centres are selected in high dense neighborhood using NBC. But it is time consuming operation because for every point indicates that it is located in high dense or less dense or evenly shattered space. The dataset plotted in space is split into hyper cubes of same sides in every dimension. The searching for high density is operated in cells rather in whole space. It is also very time-consuming process in case of high-dimensional

dataset .The solution is density-aware method is done on the space ,if the space is dense then division or else merging of cells takes place. Partha Sarathi Bishnu [5], use 4-way branching tree i.e. Quad-Tree data structure for selection of the initial clusters centers. The main drawback is Quad-tree unbalanced nature. So, at the run time it may be going into the worst scenario where all the data points on one side of tree. Kai Lei [6] tries to solve the problem of empty clusters and too more or too few data points in the clusters by using group size of every cluster within a fixed range. Samuel Sangkon Lee [7] improves the execution of K-means by choosing the optimal initial cluster centre by maximized the distance between them. After choosing initial cluster which results in cluster centres are evenly scattered in the data space which produce more accurate results as compared to random selection method. Xin Chang [8] proposed the algorithm that is combination of supervised learning and unsupervised learning. In case of supervised learning the selection of initial centres is from labeled data which tries to classify the unsupervised dataset. When all the initial seeds come for labeled data known as CSK(Complete Seed K-Means) which results in efficient clustering configuration. Secondly, when only "j" i.e. j<k initial centres are selected from labeled data and other "k-j" initial centres are selected as randomly or max-distance apart from remaining "n-j" dataset known as ISK(Incomplete Seed K-Means). Young Jun Zhang [9] proposed the algorithm which chooses initial centres according to similarity density calculated with help of similarity degree between data points. The data points which are having minimum density are chosen as potential initial centres which are used to generate feasible clustering results.

All surveyed papers have the efficient selection of initial centres problem. Many of papers concentrate on high-density area[4] to select the initial centres. The use of Quad-tree data structure to identify the dense area, supervised knowledge [5] to identify the initial centers from supervised labeled data [8] and size constrained [6] in the clusters with nearest neighbor approach [9]. All papers aim for start the clustering with optimal initial centres only.

III. PROPOSED ALGORITHM

This section enclosed with the process in proposed algorithm. The algorithm contains of four parts. In first step, K-Means is executed using the maximum separation method. The resultant "k" final centers arrive from the step first feeded to K-Means and run until the centers does not convergence. In second step, we calculate the double centers of all the "k" resultant centers from the previous step. In next step, the execution of K-Means by using the "2k" doubles centers. In the last step, "2k" centers are shrinking using similarity measure up to "k" centers. The steps combined into proposed algorithm are enhanced into four sections as follows:

Execution K-Means by Maximum Separation Method: In this we are choosing the first "k" continuously objects from the dataset. So, after that we apply the maximum separation method on "k" objects. At the last, we get "k" resultant initial centers which is feeded to K-Means algorithm and run until centers convergence.

Double Centres: In this we are calculating the distance between particular centre and other centres, then by taking one - fourth closest distance from particular centre is taken as radius for the double centres. The radius is added and subtracted from particular centre except the last dimension which kept same. The resultant "2k" centres come from it.

Execution K-Means by Double Centers: The K-Means is executed using the double centres until final centres do not convergence.

Merge Stage: The "2k" centres will shrink up to "k" centers. The similarity measure is employed for the shrink centres.

A. Execution of K-Means by using Maximum Separation Method

Let, assume D is the dataset and m is the number of objects in dataset $D = \left(X_i\right)_{i=1}^{m}$

Now consider,

L is likely initial centers and expressed as $L = \left(X_1, X_2, \ldots\ldots X_k\right)$ and k indicates as number of clusters.

For every X_j

Loop j to (k+1 to m)

Determine Euclidean distance between X_j and L as $[dis\left(X_{k+1}, X_1\right), dis\left(X_{k+1}, X_2\right), \ldots\ldots, dis\left(X_{k+1}, X_k\right)]$

Then pick the maximum separation data point from X_j and replace X_j with it in L set.

Run K-Means(k,L)

Repeat the process until the final centers does not change and update the final centers in L.

B. Double Centers

From updated L, calculate the distance between the final centers.

Radius for first centre = $\min\left(dis\left(X_1, X_2\right), dis\left(X_1, X_3\right), \ldots, dis\left(X_1, X_k\right)\right) / 4$

L` = (first centre + radius for first centre) and (first centre - radius for first centre) except the last dimension.

Repeat the process for all "k" centers in L.

C. Execution 2K-Means by using Double Centers

Run K-Means (2k,L`)

Repeat the process until the final centers do not change.

D. Merge Stage

The resultant final centers are merge up to "k" clusters will not come. The merging process was done by maximum radius R_i i.e. farthest data point present in particular clusters. The radius for every clusters is calculated as follows:

$$R_i = \max_{j=1 to n_i} \left| y_j - z_i \right| \; y_j \in c_i, c_i \in C$$

where

z_i—centroid of cluster c_i

C—"k" cluster centers

n_i—number of data points present in cluster c_i

y_i---Object present in cluster c_i

Then, the similarity measure $d\left(C_i, C_j\right)$ between two clusters is calculated and it combines the centers which are having minimum similarity and parellely updated the L` set until k centers are not remaining.

$$d\left(C_i, C_j\right) = \frac{\left| z_i - z_j \right|}{R_i + R_j} \qquad (1)$$

E. Proposed Algorithm

In this approach the initial points are selected by using max separation method. Then, the "k" centers are feeded to K-Means until centers do not convergence. The final centers are doubled using the search radius except the last dimension which is kept same. After, double K-Means run on the "2k" centres until the centres does not change. At last, the centres are merged according to similarity measure between the clusters up to "k" clusters will not come.

Input: m indicates the number of objects in dataset $D = \left(X_i\right)_{i=1}^{m}$ and k indicates total number of clusters.

Output: Final Initial centers $FC = \left(X_i\right)_{i=1}^{k}$

Start

Step1: K-Means with Max separation method

Step1.1: Take a set $L = \left(X_1, X_2, \ldots\ldots X_k\right)$ where starting data points are taken in it.

Step1.2: Loop i from k+1 to m.

Step1.3: For every X_i calculate maximum separation between L and X_i and replace X_i with maximum separation data point in L.

Step 1.3: Run K-Means (k, L) where L taken as initial centers for K-Means and run until final centers do not change.

Step2: Double Centroid

Step2.1: Calculate the distance between every L potential centers to all other potential centers.

Step2.2: Select the one-fourth of closest distance from particular centre as search radius.

Step2.3: Subtract and add the search radius except the last dimension from particular centers resulting with 2 centers store them in L`.

Step2.4: Repeat step 2.3 for every L centers.

Step3: 2K-Means

Step3.1: Run K-means (2k, L`) until final centers do not convergence.

Step4: Merge among most similar potential centers

Step4.1: Calculate similarity between every L` centers based on formula (1)

Step4.2: Select minimum similarity and merge those centroids and update L` simultaneously.

Step4.2: Repeat step 4.1 and step 4.2 until "k" centers will not remain in L`.

Step5: Run k-means (k, L`) until final centers will not change.

Finish

IV. EXPERIMENT AND DISCUSSION

Proposed algorithm uses the Eclipse IDE Tool for Java programming language. The hardware configuration for running the algorithm is Intel Core i5 processor, 4GB RAM on Mac Book Pro having OS X Yosemite operating configuration.

A. Description of Datasets

The experiment consists of five real UCI datasets repository under the classification and clustering fields. The Iris datasets consists of 150 data points in R^5 having all objects are continuous, preprocessed, real and not null. The Wholesale customers contains 440 objects in 8 dimensions with all elements are continuous, preprocessed, real and not null in nature. The Wine Recognition contains 178 objects in 14 dimensions with all elements are continuous, preprocessed, real and not null in nature. The Seeds datasets contains of 210 data points in R^8 having all elements are continuous with not null values. The Blood Transfusion Service Centre contains 748 objects in 5 dimensions all elements is continuous, preprocessed, real and not null in nature. All the datasets are defined in Table I.

TABLE I. OBSERVATION OF DATASETS [11]

S. NO.	NAME OF DATASETS	NUMBER OF OBJECTS	DIMENSIONS
1	IRIS PLANT	150	5
2	WHOLESALES CUSTOMERS	440	8
3	WINE RECONGNITION	178	14
4	BLOOD TANSFUSION SERVICE CENTRE	540	21
5	SEEDS	210	8

B. Description of Traditional Methods

To test our experiments we are comparing the results of proposed algorithm with the existed popular techniques like K-Means, K-Medoids and K-Means++. The selection process of K-Means and K-Medoids is random selection of initial centers. The main difference between them is intermediate selection of centers from the mean of data points present in the clusters and with K-Medoids, no extra data point is generated as initial centers K-Means++ is the expansion of K-Means algorithm where the initial centers are selected by using D^2 weighting technique.

C. Validity measures

The cluster validity measures are used to compare the clustering done by existed popular algorithm and proposed algorithm. The measures what we are comparing in the experiments are as follows: RMSE(Root Mean Square Error),SSE(Sum of Square Error),SD Validity Index(SD index) and Davies - Bouldin Index(DB Index)[10].The convergence speed is also compared into it.

1) *RMSE:* The RMSE indicates the average of sum of square error. The minimum RMSE value shows the optimal cluster configuration.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{k}\left|Y_i - c_j\right|^2}{m}} \qquad Y_i \in Y, C_j \in C$$

where,

Y -Total element present in dataset

C - Set of all "k" cluster centers

k- Number of clusters

m-Number of dimensions in object

2) *SSE:* The SSE encloses the squared distance of all objects to the closest centers. The minimum value of SSE is very much favorable.

$$SSE = \sum_{i=1}^{k} \sum_{y_j \in c_i} \left| y_j - c_i \right|^2 \qquad y_j \in Y, c_i \in C$$

Where,

Y-Number of objects

C-"k" cluster centers

k-Number of clusters

3) *SD Validity Index:* SD index indicates both the compactness within the clusters and separation between clusters. The minimum SD value is more favorable.

$$SDIndex = \frac{1}{k} \sum_{a=1}^{k} \frac{\left|\sigma(C_a)\right|}{\left|\sigma((Y))\right|} + \frac{D_{\max}}{D_{\min}} \sum_{a=1}^{k} \left(\sum_{b=1, b \neq a}^{k} \left| C_a - C_b \right| \right)^{-1}$$

Where,

$\sigma(C_a)$-Variance of a[th] cluster

$\sigma(Y)$-Variance of whole dataset

$D_{\max}$ -Maximum distance between clusters

$D_{\min}$ -Minimum distance between clusters

4) DB Index: DB Index represents the average of homogeneity within the clusters and choose the most similar clusters. The minimum value of DB index indicates the optimal clustering configuration.

$$DB\ Index = \frac{1}{n_a} \sum_{k=1}^{n_a} R_k$$

Where,

$$R_k = \max(D_{kl}) k, l = 1,2,\ldots n_a \; k \neq l, \quad D_{kl} = \frac{dis_k + dis_l}{d_{kl}},$$

$$d_{kl} = d(u_k, u_l) \; and \; dis_{kl} = \frac{1}{|c_k|} \sum_{y \in c_k} d(y, u_k)$$

D_{kl} ---Similarity of clusters, dis_k ---Dispersion of the cluster, d_{kl} ---Dissimilarity of cluster

The iteration speed is time to acquire the final cluster centers. The less value of iteration is also favorable.

D. Results and Discussion

In this part we display the output of the proposed approach and other popular initial centers approaches like K-Means, K-Means++, K-Medoids to analyze clustering configuration against four quality measures and convergence speed on various datasets.

Note: The bold values indicate the improved clustering results in all the comparison tables.

TABLE II: COMPARISON OF RESULTS ON IRIS PLANT DATASET

QUALITY MEASURE	K-MEANS	K-MEDOIDS	K-MEANS++	PROPOSED
RMSE	122.27	113.34	97.34	**97.32**
SSE	3.191	2.180	1.419	**1.414**
SD INDEX	16.5	8.1	5.9	**3.1**
DB INDEX	0.09	0.08	**0.01**	0.08
ITERATION	7	29700	3	**3**

TABLE III: COMPARISON OF RESULTS ON BLOOD TRANSFUSION SERVICE CENTER DATASET

QUALITY MEASURE	K-MEANS	K-MEDOIDS	K-MEANS++	PROPOSED
RMSE	508040.6	516308.8	508640.6	**469637.8**
SSE	1607503.4	1253965.0	1607503.4	**631153.1**
SD INDEX	0.0059	**0.0014**	0.0059	0.0015
DB INDEX	3.8	1.9	3.8	**0.7**
ITERATION	10	80036	10	**3**

TABLE IV: COMPARISON OF RESULTS ON WHOLESALES CUSTOMER DATASET

QUALITY MEASURE	K-MEANS	K-MEDOIDS	K-MEANS++	PROPOSED
RMSE	**5428912**	5665174.5	5798929.0	5529996.0
SSE	2.61	7.79	2.80	**2.24**
SD INDEX	0.010	0.013	0.022	**0.006**
DB INDEX	107.5	393.2	172.2	**72.1**
ITERATION	13	31240	10	**10**

TABLE V: COMPARISON OF RESULTS ON SEEDS DATASET

QUALITY MEASURE	K-MEANS	K-MEDOIDS	K-MEANS++	PROPOSED
RMSE	313.73	381.90	313.21	**313.21**
SSE	**7.54**	12.79	8.97	8.97
SD INDEX	2.76	4.11	3.06	**2.56**
DB INDEX	0.021	0.022	0.012	**0.012**
ITERATION	6	33390	6	**3**

TABLE VI: COMPARISON OF RESULTS ON WINE RECOGNITION DATASET

QUALITY MEASURE	K-MEANS	K-MEDOIDS	K-MEANS++	PROPOSED
RMSE	16555.6	16593.7	**16555.6**	18123.0
SSE	68628.6	44081.8	68628.6	**4137.2**
SD INDEX	0.051	0.057	0.051	**0.027**
DB INDEX	1.84	1.22	2.18	**0.23**
ITERATION	9	61054	15	**5**

In the results we are comparing the proposed technique to the well-known techniques like K-Means, K-Medoids and K-Means++.

We explain the cluster validity measure on Wine Recognition, Iris, Wholesales Customer, Seeds, Blood Transfusion Service Centre datasets with various number of clusters. Table II shows the proposed method shows better clustering results with reference to RMSE,SSE,SD Validity Index as compare to K-Means, K-Medoids and K-Means++.The iteration speed of proposed is equal to k-Means++ and less than other two existed methods. In DB index proposed method shows optimal clustering results compare to random selection technique i.e. K-Means and K-Medoids, but in compare to K-Means++ showing less DB index. In table III, shows the best case of proposed approach. All the cluster validity metrics perform extremely well over the existed methods. The convergence speed is quiet fast than all of them. Table IV, indicates the efficient measurement in case of SSE,SD validity index and DB index. The speed of convergence is less than K-Means and K-Medoids but, in case of K-Means++ iteration speed is same. In Table V measuring the validity measure effective in case of RMSE,DB Index and SD Validity Index of proposed algorithm. The SSE measure is better than K-Medoids with equal measurement in case of K-Means++. But, K-Means shows the best result in all. The convergence speed of proposed algorithm is quiet faster and efficient as compare to existing one. In table VI the RMSE is not so good than the existing methods. Other measures like SSE, SD Validity Index and DB index shows optimal clustering results with all the others techniques. The speed to acquire the final clustering is faster in proposed algorithm as compare to other techniques.

V. CONCLUSION AND FUTURE WORK

This work improves the initial centers selection for many applications which are used for the analysis of raw data by using the maximum separation method, double centers and merging clusters using the similarity between them. As in the classic methods the random choose of initial centers leads to wrong clustering configuration or poor cluster validity measure. But, the proposed approach shows better clustering results and structure. In this paper, double centers around the maximum separation method centers results in maximized in homogeneity similarity within clusters and minimizing the heterogeneity between the clusters. The optimal results come in many datasets in terms of error minimization in case of RMSE,SSE, compactness and separation of the clusters. The convergence speed is faster or same to the well-known methods. All the evaluation measures experiments with five datasets have prove to get optimal clustering configuration.

In the future, we shall try to use more complex dataset. The time complexity due to twice operations of K-Means is quiet high in proposed approach. So, to reduce this we have to reduce the reluctant calculation and to make this efficient in term of time complexity.

REFERENCES

[1] A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", Journal ACM Computing Surveys (CSUR), Volume 31, Issue 3, pp. 264-323, 1999.

[2] K. P. Soman, S. Diwaker and V. Ajay, "Insight into Data mining: Theory and Practice", pp.17-18, 2006.

[3] J. Chang, "SDCC: A New Stable Double-Centroid Clustering Technique Based on K-Means for Non-spherical Patterns", Advances in Neural Networks, Springer Berlin Heidelberg, pp. 794-801, 2009.

[4] Ye, Y., Huang, J. Z., Chen, X., Zhou, S., Williams, G., and Xu, X., "Neighborhood density method for selecting initial cluster centers in K-means clustering", Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, pp. 189-198, 2006.

[5] Bishnu, P. S., and Bhattacherjee, V., "Software fault prediction using quad tree-based k-means clustering algorithm", IEEE Transactions on Knowledge and Data Engineering,volume 24, Issue 6, pp. 1146-1150, 2012.

[6] Lei, K., Wang, S., Song, W., and Li, Q., "Size-Constrained Clustering Using an Initial Points Selection Method", Knowledge Science, Engineering and Management, Springer Berlin Heidelberg, pp. 195-205, 2013.

[7] Lee, S. S., and Han, C. Y., "Finding Good Initial Cluster Center by Using Maximum Average Distance", Advances in Natural Language Processing, Springer Berlin Heidelberg, pp. 228-238, 2012.

[8] Wang, X., Wang, C., and Shen, J., "Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers", Web Information Systems and Mining, Springer Berlin Heidelberg, pp. 178-187, 2011.

[9] Zhang, Y., and Cheng, E., "An optimized method for selection of the initial centers of k-means clustering", Integrated Uncertainty in Knowledge Modeling and Decision Making, Springer Berlin Heidelberg, pp. 149-156, 2013.

[10] Kovács, F., Legány, C., and Babos, A., "Cluster validity measurement techniques", 6th International symposium of Hungarian researchers on computational intelligence, 2005.

[11] Dataset Collection: http://archive.ics.uci.edu/ml/datasets.html