Density base k-Mean's Cluster Centroid Initialization Algorithm

Kabiru Dalhatu
Department of Computer Science,
KUST, 2016
Kano University of Science
and Technology, Wudil

Alex Tie Hiang Sim
Department of Computing,
UTM, 2016.
Universiti Of Technologi
Malysia

ABSTRACT
A spatial data mining is a process of extracting valid and useful information out of generated data, which recently becomes a highly demanding field due to the huge amount of data collected everyday across various applications domains which by far exceeded human's ability to analyses, this brought about the development of many data mining tools among which clustering is recognized to be the efficient data mining method that categorized data based on similarity measures, where k-Means is a well-known clustering algorithm used across different application domains. Similarly, k-Means suffer from multiple limitations with its clustering accuracy fully depend on cluster center positioning. In this paper, a density base k-Means cluster centroid initialization algorithm has been proposed to overcome k-Mean's cluster center initialization problem. To prove the accuracy of the proposed algorithm the evaluation test was conducted using two synthetic datasets called Jain and Path base dataset. The clustering accuracy result of the proposed algorithm is compared with that of traditional k-Means algorithm where it proved that the clustering accuracy of the proposed algorithm is better than that of traditional k-Means algorithm.

General Terms
Clustering, k-means, centroids allocation, clustering accuracy.

Keywords
Temporary Matrix(TMAT), cluster center C, Dataset D.

1. INTRODUCTION
Clustering is a process of classifying a group of patterns (data items, feature vectors or observation) into clusters based on similarity measures, where patterns within the same cluster are highly related and different from those in the remaining clusters. The quality of clustering accuracy is highly dependent on great intra-cluster similarity and small inter-cluster similarity, clustering is one of the difficult problems to be solve based on assumption due to its unsupervised nature and it is of two types that is Fuzzy and hard clustering, where in fuzzy clustering a data point can belong to more than one clusters while in hard clustering a data point can belong to one and only one cluster and its further classified into partitioning and hierarchical clustering respectively[1]. A partitioning method is the iterative process of dividing an n data objects into non overlapping k clusters, such that each data object belong to one and only one cluster, this method consist of different renown algorithms among them is k-means clustering algorithm, this algorithm is a well-known clustering algorithm used across diverse fields due to its simplicity in nature [2] below are the steps of k-means algorithm
Input: An n dataset and k number of clusters

1. Select k centroid out of n dataset objects.
2. Calculate the distance between each k centroid and the remaining data objects
3. Assign each data object to its closest centroid based on the calculated distances
4. Find the mean of each group (cluster).
5. Repeat steps 3-5 until convergence.

Similarly, some of the advantages of this algorithm which made it to be popularly known among its counterpart are:

1. Easy to understand and implement
2. Relatively efficient with time complexity of O(kni)`
3. It attempts to reduce an objective function (square error function) [3],

However, even though k-Means algorithm is well known due to its subsequent strengths among others yet, it suffers a lot of limitations among which are:

1. The number of clusters in a given dataset has to be defined in advance.
2. Its sensitive to outliers.
3. Its result solely depend on cluster center initialization.
4. It often generate an empty clusters
5. It can only detect a globular shape cluster.
6. It can't works with categorical data

To handle the above mentioned problems of k-means clustering algorithm, a series of k-means enhancement algorithms were proposed by different researchers across the world in which each is trying to handle either of the above mentioned limitations such as EDCA, k-means++ k-means-- and ODBD-k-means among others. This research work was conducted with the intention of handling k-Means limitation III mentioned above, through which limitation IV can also be resolved, this is because the clustering accuracy of k-Means algorithm is highly affected when its cluster center is mistakenly selected. This is because k-Means algorithm result change with different selection of initial centroids[4]. The performance and convergence criteria of k-means algorithm are negatively affected if the cluster's initial centroids were wrongly selected [5]. In this paper, a density base k-Means cluster centroid initialization algorithm is present, by assigning a points with the highest density value to be a cluster center, this researched was done with a view of improving the efficiency of k-Means algorithm through handling cluster centroid allocation problems. The rest of the paper is organized as follows. The literature review of clustering algorithms will be discussed in section 2, a proposed algorithm methodology will be presented in section 3, an experimental evaluation test in order to prove the efficiency of the proposed algorithm using the subsequent mentioned synthetic datasets will be presented in section 4,

while section 5 present the conclusion of the study with some directions for future research.

2. RELATED WORK

There are several literatures written for classical k-Means enhancement through handling k-Means cluster centroid allocation problem by many researchers using different techniques and datasets across the globe, some of this work was conducted by[6] where they proposed a new method for improving the efficiency of k-Means clustering algorithm by optimizing the process of selecting initial cluster center and its calculation by enhancing early focal point while determine k-value, based on the experiment conducted, the result proved that the proposed algorithm not only outperform k-Means algorithm in term of stability, it also reduced the effect of noise, ensure accuracy and effectiveness of the final result. [7]proposed a new algorithm through the use of minimum spanning tree for k-Means initial centroid selection, with the aim of improving the accuracy rate of traditional k-Means algorithm. This algorithm consist of two phases upon which the first phase use the prim technique to find the minimum spanning tree of the randomly generated points and the second phase group the points with maximum amount of weights where all weights with small dissimilarity are selected as the initial centers. The evaluation test result shows that IKACP is more stable and improved the accuracy rate of traditional k-Means clustering algorithm. A new algorithm was proposed by[8] for enhancing the performance of standard k-Means algorithm, by improving the process of initial centroid selection through random selection of initial centroids. Even though this proposed algorithm improved the performance of k-Means algorithm, yet its sensitive to initial starting points. Hence it doesn't promise to provide equal clustering result.[9] proposed a new algorithm for initial center selection in k-Means clustering algorithm, by the use of space partitioning data structure (K-d tree) concept in relation to density based method. The performance evaluation experiment shows that the proposed algorithm is more effective for finding cluster than the well-known existing algorithms. Even though density based approach is very effective in outlier detection. [10] proposed a new algorithm for improving the efficiency of k-means algorithm, through best initial centroids selection, which will serve as the starting points of the proposed algorithm. The result proved that the proposed algorithm improved the accuracy of traditional k-Means algorithm.

3. METHODOLOGY

Given the subsequent effect of wrongly assigned cluster center to the traditional k-Means algorithm results, However, going by the above mentioned algorithm under section 2, it can be concluded that majority of the subsequent enhanced version of k-Means algorithm have their own drawback. This serve as a motivation factor that motivate us in formulating a new k-Means enhancement algorithm that could handle limitation III and IV above while maintaining consistent result given the same dataset and threshold value. In this section, a brief explanation of the terms, concept and analysis of the proposed algorithm for overcoming k-means algorithm limitations will be discussed.

3.1 Related Concept

Definition 1: Euclidean distance is one of the simplest formulas used for calculating distance between points, this formulas worked in the same way the distance between two points is measure using ruler. An Euclidean distance formula for similarity measures is shown below.

$$d(c,p) = \sqrt{\sum_{i=1}^{n}(\mathrm{Ci} - \mathrm{Pi})^2} \quad (1)$$

Where d(C,P) represent the distance between a cluster center C and any data point P in dataset D, n represent total number of dataset points and i is a counter.

Definition 2: The density value based on proposed algorithm can be computed using an indegree formula below, by using the above calculated(d(C,P)) as an input.

$$\rho_i = \frac{1}{d(C,P)+1} \quad (2)$$

Where ρi and d(C,P) stand for density and distances respectively.

Definition 3: Threshold (Eps) is a term main for representing the neighborhood radius of a data points say k denoted by NEps (k). The formula used for calculating the threshold value in this research was not standard; rather it was formulated by the researcher based on the dataset used, this threshold is mathematically defined below.

$$T = [\mathrm{sum}(\mathrm{Max}(dist)-\mathrm{Min}(dist)*\Gamma(0.5+1]/(n\sqrt{\pi}) \quad (3)$$

Where n stand for total number of dataset points, Max(dist) and Min(dist) stand for maximum and minimum distance respectively.

Definition 4: The process of defining a cluster center C of the proposed algorithm, is quite different with that of the traditional clustering method that used to assigned cluster center through random approach, this usually affect the clustering accuracy of those algorithm formulated based on the subsequent technique. Similarly, in the case of the proposed algorithm, a cluster center is assign to a point with maximum density value calculated by equ. 2 above. This cluster center C will be shifted from one point to another within the cluster in trying to absorb all points that are ought to be part of the cluster, but they out of it at initial stage due to the restriction put upon by the threshold value, until convergence. Below is a figure depicting this multi-center scenario

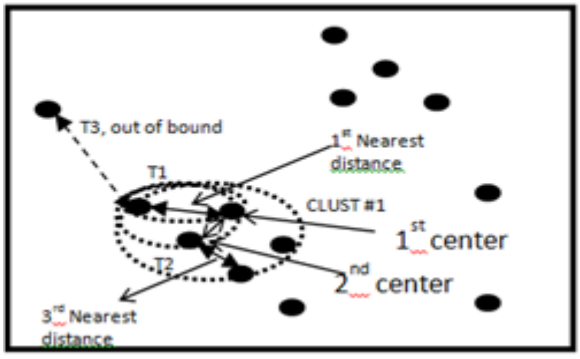

Figure 1: Multi-cluster center scenario

3.2 Proposed Algorithm Pseudo code

The ability to enhance k-means algorithm by handling its cluster center allocation problem was performed through the use of different techniques, explained in section 2 above. Similarly, this proposed algorithm provide an efficient cluster

center allocation by the used of point with maximum density value as a cluster center C, which will be used as a reference point for data clustering. The proposed algorithm divide a dataset points into cluster by the used of threshold value and cluster center C. Where the distance between the cluster center C and the remaining data points P computed by equ. 2, will be compare with the threshold value and assign all point whose distance is within the threshold bound to the current cluster, the cluster center will then be shifted to any other point within the cluster and calculate the distance between the new center and the remaining data points excluding those who were already inside the cluster as depicted in fig.1 above. This process will continue until all points belong to the first cluster are appropriately clustered and removed out of the dataset. The second cluster will appropriately be clustered out of the remaining dataset based on subsequent processes. The proceeding process will continue until all remaining dataset points are appropriately clustered. Below are steps of the proposed algorithm.

1. Input Datasets D, Temporary Matrix (TMAT) and Number of Cluster K.
2. Calculate distance between all dataset points di and dj using Equ.1
3. Calculate density of each data points di,jusing Equ. 2.
4. Calculate threshold value T using Equ. 3.
5. Find a point with maximum density value C (imaginary cluster center) and assign its copy into TMAT.
6. Find the distance between C and the remaining dataset points di,jϵD
7. Compare if the distance between C and di,j <= T, assign the copy of all data points that satisfy this condition into TMAT.
8. Check if the points inside TMAT are greater than one.
9. Calculate the distance between all points inside TMAT excluding original C with all di,jϵD who were initially compared with C and the condition didn't hold.
10. Assign the copy of all di,jϵD into TMAT, whose distance from any other point in TMAT <=T.
11. Repeated steps 9 and 10 until no di,jϵD whose distance from that of any other points in TMAT is less than or equal to T.
12. Subtract all points that are inside TMAT out of D.
13. Increment number of TMAT until number of TMAT<=k
14. Repeat from step 5-12 until all points in D are assigned to the appropriate TMAT.

3.3 Proposed Algorithm Handling Method of Limitations III and IV Above

The k-Means clustering algorithm result is highly dependent on initial centroid allocation as stated in limitation III above, by taking a look at Section 3.2 above we can easily get a clear picture upon what cause this limitation and how it was fully handled by our proposed algorithm. k-Means algorithm can often generate an empty cluster as stated in limitation IV, this is due to the random selection of cluster center nature of the algorithm, which may possibly lead to the selection of an entirely separated point within a particular dataset to be used as a cluster center, where if that separated point is being used as a cluster center, it may possibly lead to the formation of an empty cluster as a results of the fact that it doesn't have the most nearest neighborhood points that falls within the threshold value. Similarly, having carefully study the content of Section 3.2 above, it can easily be clarify that, the proposed algorithm have handled this limitation, since it doesn't a dopt the traditional way of random selection of cluster center, instead its cluster center always stand to be the points with maximum density value, which by the definition of the word density this points must fall within the mid of the dataset points which must have a number of nearest neighborhood points that falls within the threshold value.

3.4 Experimental Design

To further prove the efficiency of proposed algorithm over traditional k-Means algorithm, in this research paper a jain and path base synthetic datasets were used in conducting the experiment due to their complexity and irregularity shape in nature. Below is a table depicting the names, attributes and instances of these datasets.

.

Table 1: Joensuu Synthetic dataset for the experiment

Name	Attribute	Instance
Jain	373	2
Path base	300	2

Joensuu repository is a clustering datasets repository that group some complex synthetic datasets main for clustering algorithm testing. Before executing the proposed algorithm in section 3.2 using the above datasets, these dataset were normalized and divide into training and testing part by the used of k-fold cross validation technique, a value of cluster (K) was then set and computed the threshold value (T) using Equ. 3 above as depicted in Table 2 below

Table 2: Parameters used for running the Proposed algorithm

Name	Number of cluster (K)	Threshold Value (T)
Jain	2	0.0240
Path base	2	0.0100

The above threshold values were not standard as they were found through the use of the equ.3 above, which was formulated through try and error approach until an appropriate T is discovered, this is because it is quite difficult to formulate a standard threshold formula that can satisfy both datasets considering the nature and complexity of these datasets.

4. RESULTS AND DISCUSSION

having successfully got the parameters depicted in Tables 1 and 2 above it turn to run the proposed algorithm, the summary of the experimental results on percentages clustering accuracy found after successful running of the proposed algorithm with the above parameters as an input, is depicted in Table 3 below.

Table 3: The algorithms results on datasets

DATASET	Proposed algorithm Clustering accuracy (%)	k-Means algorithm Clustering accuracy (%)
Jain	*93.41*	*83.78*
Path base	*88.95*	*88.78*

The above datasets results on traditional k-Means algorithm was found by running the above datasets in k-Means matlab built in function, for the successful comparison with the proposed algorithm result. Having carefully study the result presented in Table 3 above, it can easily conclude that the percentage clustering accuracy of the proposed algorithm is significantly improved in both datasets as compared with that of the traditional k-Means algorithm. This is because the traditional k-Means algorithm used the traditional way of random cluster center selection which usually affect its clustering accuracy, while the proposed algorithm used a point with maximum density value as an imaginary cluster center, this center may keep changing from one point to another within the same cluster, where each center will be used in connection with threshold value to retrieve all points within the same cluster, this made the proposed algorithm a multi imaginary cluster center algorithm. This is quiet efficient when compared with the traditional single center approach, where a cluster growth based on threshold length and any other point outside threshold coverage area will be regarded as a points which belong to different cluster or outlier, this may commonly happen when dealing with an irregular shape clusters. Below is the figure depicting the graphical representation of clustering accuracy results presented in Table 3 above.

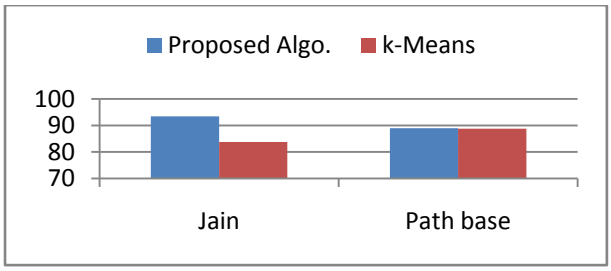

Figure 2: Clustering accuracy comparison between proposed and k-Means algorithms

Similarly, since the proposed algorithm is a k-Means improvement algorithm based on cluster's centre allocation it indicates that, the clustering accuracy of the proposed algorithm may solely depend on how efficient it can be able locate an appropriate cluster's center. Based on the above clustering result, it proved that the proposed algorithm outperformed traditional k-Means algorithm in term of clustering accuracy, based on this finding hope this proposed algorithm to be used as an enhancement version of traditional k-Means algorithm

5. CONCLUSION

In this paper, a new algorithm was proposed based on cluster centroid allocation, where a point with maximum density value is use as an imaginary cluster center which usually shift from on point to another within a cluster until all points belong to a particular cluster are well separated, then the number of cluster will be incremented and continue until all data points are appropriately clustered, this algorithm was proposed with the aim of handling traditional k-Means limitation III through which limitation IV was also handled. The experimental results found during the executions of the proposed algorithm remain fixed in different executions provided both datasets, number of clusters and threshold value remain the same, this proved to solved the subsequent k-Means limitations and similarly outperformed it, in term clustering accuracy. In future studies we hope to reformulate the threshold detection equation thereby standardizing it. 6.

6. REFERENCES

[1] Denning, D.E., *An intrusion detection model.* IEEE Transactions of Software Engineering, 1987. **13**(2): p. 222-232.

[2] Abubaker, M. and W. Ashour, *Efficient Data Clustering Algorithms: Improvements over Kmeans.* I.J. Intelligent Systems and Applications, 2013: p. 37-49.

[3] Rai, P. and S. Singh, *A Survey of Clustering Techniques.* International Journal of Computer Applications, 2010. **7**(12): p. 1-5.

[4] Abhilash, B.C., *A Comparative study on clustering of data using Improved K-means Algorithms.* International Journal of Computer Trends and Technology 2013. **4**(4): p. 771-778.

[5] Mumtaz, K. and D.K. Duraiswamy, *A Novel Density based improved k-means Clustering Algorithm.* International Journal on Computer Science and Engineering, 2013. **02**(02): p. 213-218.

[6] Zhang, C. and Z. Fang, *An Improved K-means Clustering Algorithm.* Journal of Information & Computational Science, 2013. **10**(1): p. 193–199.

[7] Wang, K., et al., *An Improved K-means Clustering Algorithm Based on Prim.* Journal of Information & Computational Science, 2013. **10**(13): p. 4303–4310.

[8] Yedla, M., S.R. Pathakota, and T.M. Srinivasa, *Enhancing K-means Clustering Algorithm with Improved Initial Center*International Journal of Computer Science and Information Technologies, 2010. **1**(2): p. 121-125.

[9] Zhang, X., et al., *A Density-Based Method for Initializing the K-means Clustering Algorithm.* International Conference on Network and Computational Intelligence, 2012. **46**: p. 46-53.

[10] Yuan, F., et al., *A New Algorithm to get the Intial Centroids* Proceedings of the Third International Conference on Machine Laming and Cybernetic, 2004: p. 1191-1193.

7. AUTHOR PROFILE

Mr. Kabiru Dalhatu received the B.Sc.(Hons) and M.Sc. degree in Computer Science from Kano University of Science and T echnology, Wudil, in the year 2007 and Universiti of Technologi Malaysia, in the year 2014 respectively.