General Applicability of K-means Algorithm with Enhanced Centroids

Gopal Behera[1], Ashok Kumar Bhoi[2]
[1, 2]*Department of Computer Science & Engineering,*
Government College of Engineering Kalahandi, Bhawanipatna, Odisha, India

***Abstract*— K -means clustering is the simplest unsupervised learning algorithm in the data mining that can solve the large amount data set or objects by grouping them into k number groups or cluster. The grouping of object done in such way that objects in the similar groups or clusters are more similar to each other than those in the other groups/clusters. The data are randomly chosen to the clusters resulting the clusters that have same number of data set or objects, in this paper a new technique is adopted which results for selecting the better cluster from both uniform and non-uniform data set equally.**

Keywords: **Clustering, Initial Centroids, k-means Algorithm.**

I. INTRODUCTION

Huge amount of data was encountered in various activities of our day-to-day life. The most common activity is to classify or groups the data using various data exploration technique, clustering is one of the technique which groups objects with similar properties together to further facilitate processing of these groups, on other hand classification of data can be either supervised or unsupervised machine learning technique whereas clustering is a unsupervised classification technique. In clustering techniques the objects are grouped together on the basis of subjectively chosen measure of similarity. Similarity between the objects within a group is higher than the similarity between the objects belonging to different groups. Data mining technique has wide range applications in various fields, such as engineering, medical sciences, computer science, life science, earth sciences, social sciences, psychology, and economic etc. Clustering [1] is one of the tasks of data mining & is useful technique for the discovery of data distribution and patterns in the underlying data. The aim of clustering is to discover both dense and sparse regions in data set. The main two approaches to clustering - exclusive clustering and overlapping clustering. In case of exclusive clustering data is grouped in an exclusive way, so that if certain datum belongs to a definite cluster then it could not be included in another cluster. K-means algorithm is a main category of exclusive clustering algorithm. Whereas, overlapping clustering uses fuzzy sets to cluster data so that each point may belong to two or more clusters with different degrees of membership.

II. RELATED WORKS

In [5] the researchers introduce various application of k-means clustering algorithm. Due to its random selection of 'k' initial centers the algorithm has less accuracy. Therefore, in this paper surveyed different approaches for selecting initial centers k-means algorithm. In [5] the research paper study of different approaches to k- means clustering and analysis of different datasets using Original k-means and other modified algorithms. In [12, 17] researchers aim to minimize the initial centroid for k-means algorithm and it uses all the clustering algorithm results of k-means and reaches its local optimal. Further the algorithm is used for the complex clustering cases with large numbers of data set and many dimensional attributes. In [13] researchers introduced k-means clustering algorithm. This paper proposes method for the making k-means clustering algorithm more efficient and effective.

2.1 Basic k-meanss clustering

k-means clustering is a partitioning [17] clustering technique in which clusters are formed with the help of centroids. On the basis of these centroids [5], clusters can vary from one another in different iterations. Moreover, data elements can vary from one cluster[5] to another, as clusters are based on the random numbers known as initial centroids. k-means is one of the most widely used partition based clustering algorithm. But the initial centroids [16] generated randomly by k-means algorithm cause the algorithm to converge at local optimum. So to make k-means algorithm globally optimum [15], the initial centroids [16] must be selected carefully rather than randomly. In this paper formulate a new method to formulate the initial centroids which results in better clusters equally for uniform and non-uniform data set.

According to this algorithm, firstly select k data value as initial cluster centre[17], then calculate the distance between each data value and each cluster centre and assign it to the closest cluster, update the averages of all clusters, repeat this process until the criterion is not match. K-means clustering [14] aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean. The k-means algorithm takes two input parameters; the database of n objects, and 'k' the number of clusters to be created. The algorithm partitions the dataset of n objects into k clusters. Cluster similarity is measured by taking Euclidean or Manhattan distance between objects. In this way k-means find spherical or ball shaped cluster, can be viewed as the cluster's centre of

gravity.The algorithm works in two phases: in the first phase k initial centroids are selected randomly, one for each cluster. In the second phase each objet of the given input dataset is associated with the cluster having nearest centroid. Euclidean distance is commonly used as a measure to determine the distance between the objects and the centroids. However, other measures like Manhattan, etc can also be used. When all the objects from input dataset are assigned to some clusters, the first iteration is completed and an early grouping is done. At this point, the algorithm starts new iteration and recalculates the new centroids. 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 or the data objects do not change their clusters, this signifies the convergence criterion for clustering.

Pseudo code for the k-means clustering algorithm [15,16] is listed as:

Algorithm: The k-means clustering algorithm

Input: A dataset D of n objects

- D={ $d_1, d_2, \ldots\ldots\ldots, d_n$}
- K= The number of desired clusters

Output:

- A set of k clusters containing data from dataset D.

Method:

1. Randomly select k objects from the dataset D as initial centroids;
2. Repeat
   a. Assign each object d_i from dataset D to the cluster to which the object is most similar i.e., has the closest centroid;
   b. Calculate new mean for each cluster;
   c. Until a convergence criterion is met (there is no change in cluster centres).

The k-means is simple and easy to implement and can be used for processing large datasets. But the algorithm has several limitations:

- The k-means algorithm is applicable to numeric data only. It cannot be applied where categorical attributes are involved.
- It is computationally expensive, time complexity being O (nkl), where n is the total number of objects in the dataset, k is the required number of clusters and l is the number of iterations.

Moreover, the quality of final clusters heavily depends on the selection of initial centroids. It means the results may be different for multiple run of algorithm for the same input data.

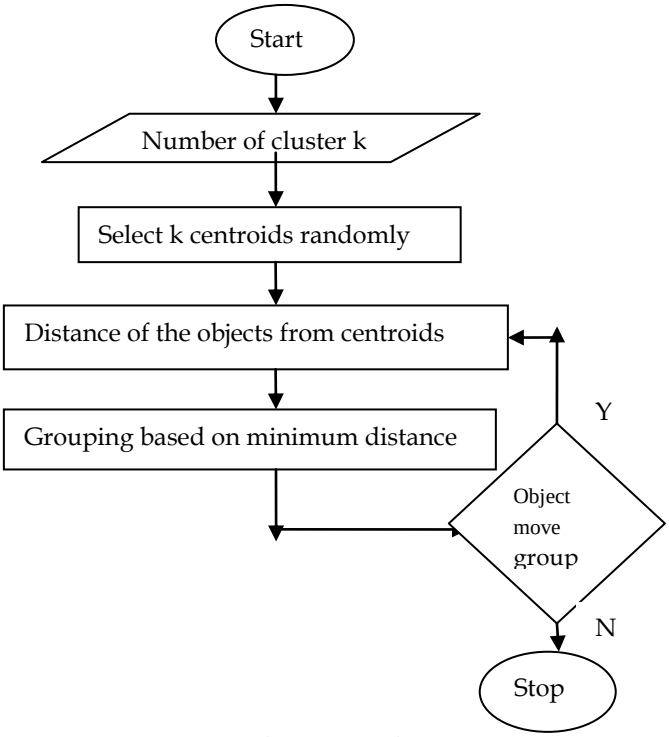


Fig.1: Process flow diagram of basic k-meanss

III. PROPOSED ALGORITHM FOR IMPROVEMENT OF INITIALCENTROIDS

An approach to systematically selecting the initial centroid has been proposed here. The centroids are determined following a systematic approach so that different runs of the algorithm on same dataset produce the same and good quality results. First of all, the given data points are plotted in a two dimensional space. All the data points should have positive valued attributes. If not then the negative value attributes should be first transformed into positive value by subtracting each data point attribute with minimum attribute value in the given dataset. This transformation is required because in the proposed algorithm the distance of each data point from the origin has to be calculated. If data points are not transformed there is a chance that for different data points, the same Euclidean distance from the origin is obtained, which will result in incorrect selection of initial centroids.

The algorithm is as under:

Input: A dataset X containing n data points.

- X={ $x_1, x_2, \ldots\ldots\ldots, x_n$}
- K= The number of desired clusters

Output: k number of initial centroids.

Steps:

1. For each data point d calculate the distance from the origin.

2. Sort the distances obtained in the previous step. In accordance with these distances sort the original data points.
3. Divide the sorted data points into k number of equal partitions.
4. In each partition, calculate the mean of the data points. These mean values will be taken as initial centroids to be used in the k-meanss algorithm.

For each data point the distance from the origin is calculated using the Euclidean distance measure as given below:

Origin: O(0,0)

Data point: P(x,y)

Euclidean distance between O-P will be: $\sqrt{((x-0)^2+(y-0)^2)}$

Then these distances are sorted in ascending or descending order. According to these sorted distances the corresponding original data points are also sorted, this sorted list of data is divided into k equal partitions. Then for each partition mean of data points is calculated.The mean values for each partition are taken as initial centroids. The centroids thus taken will suit each type of dataset. The data points with uniformly and evenly distributed values and also for those in which values are not uniformly distributed over the partition but most of them are concentrated towards any of the boundaries of the partition.

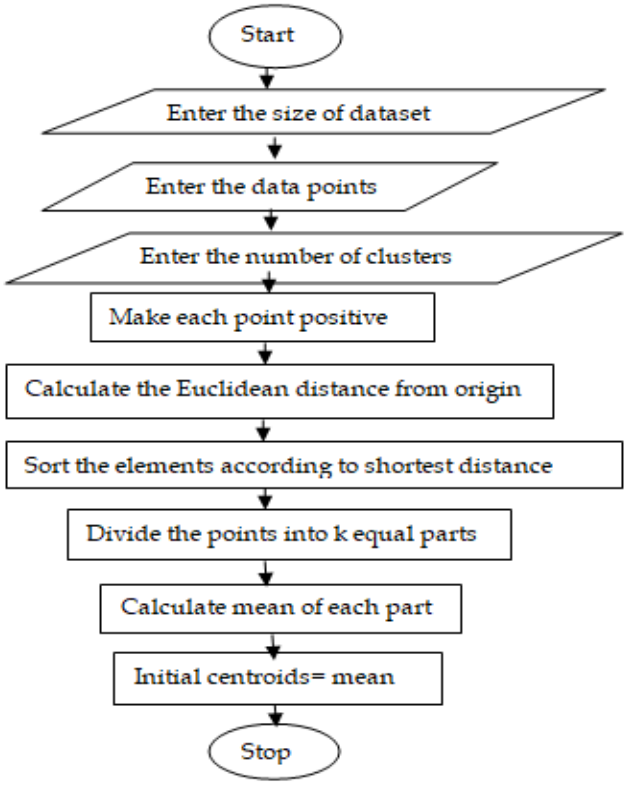


Fig 2: Process flow diagram of proposed k-meanss

3.1 Implementation and Result:

preg	plas	pres	skin	insu	mass	pedi	age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	88	31	0.248	26
10	115	0	0	0	35.3	0.134	29
2	197	70	45	543	30.5	0.158	53
8	125	96	0	0	0	0.232	54
4	110	92	0	0	37.6	0.191	30
10	168	74	0	0	38	0.537	34
10	139	80	0	0	27.1	1.441	57
1	189	60	23	846	30.1	0.398	59
5	166	72	19	175	25.8	0.587	51
7	100	0	0	0	30	0.484	32

Table 1: Diabetes data

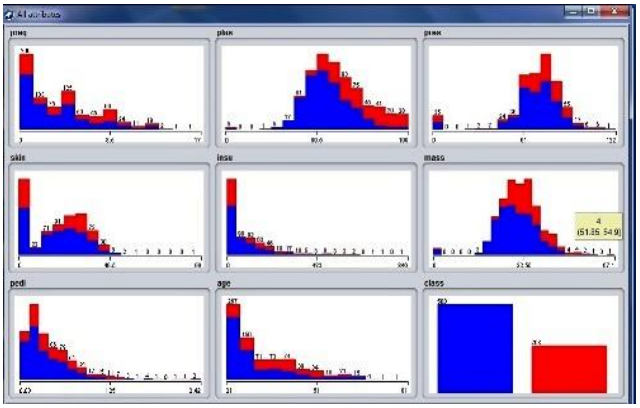


Fig 3: All attributes with test positive and negatives

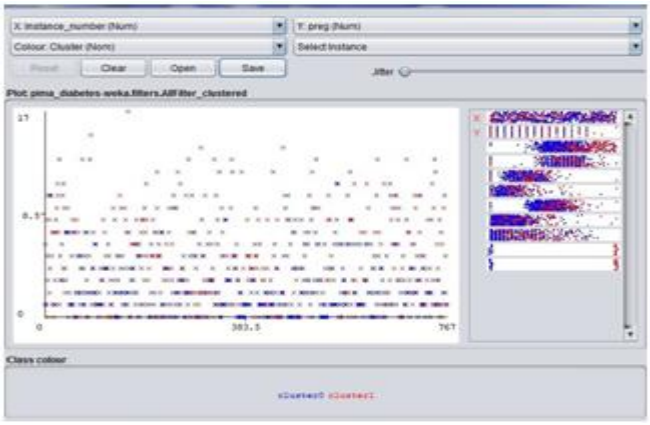
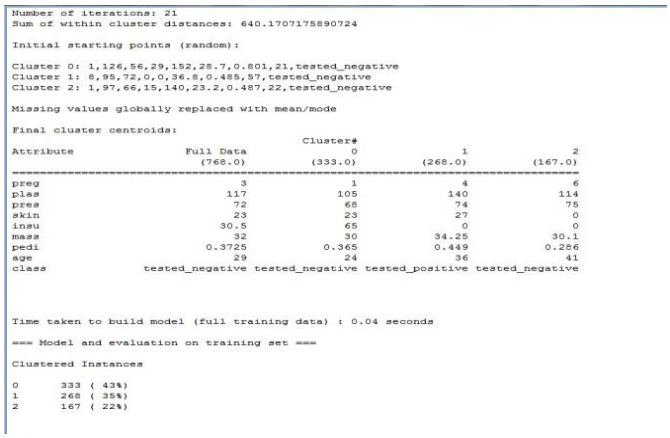


Fig 4: Cluster visualize for different clusters of k-means



```
Number of iterations: 21
Sum of within cluster distances: 640.1707175890724

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21,tested_negative
Cluster 1: 8,95,72,0,0,36.8,0.485,57,tested_negative
Cluster 2: 1,97,66,15,140,23.2,0.487,22,tested_negative

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                          (768.0)       (333.0)       (268.0)       (167.0)
=============================================================================
preg                             3             1             4             6
plas                           117           105           140           114
pres                            72            68            74            75
skin                            23            23            27             0
insu                          30.5            65             0             0
mass                            32            30         34.25          30.1
pedi                        0.3725         0.365         0.449         0.286
age                             29            24            36            41
class          tested_negative tested_negative tested_positive tested_negative



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       333 ( 43%)
1       268 ( 35%)
2       167 ( 22%)
```

Fig 5: Clusters output

IV. CONCLUSION

The proposed mean-based algorithm is easy to implement and it proves to be a better method to determine the initial centroids which can be used in the k-meanss clustering algorithm. Besides solving the problem of non-unique results, our proposed algorithm can also be widely applicable to different types of datasets. The problems associated with uniform as well as with non uniform distribution of data points are better addressed using the proposed algorithm. Another major advantage of this algorithm over the original k-means algorithm is that once the initial centroids are systematically determined, the number of iteration required reaching the convergence criteria are reduced to a great extent.The k-means algorithm can be applied to numerical data only. But in day to day life, scenarios with combination of both numerical and categorical data values are encountered. So this algorithm can be extended in the direction making the k-means algorithm applicable for mixed type of data.

REFERENCES

[1]. Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[2]. F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[3]. S. Deelers and S. Auwatanamongkol, "Enhancing K-means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4.

[4]. Huang, "Extensions to the k-meanss Algorithms for Clustering Large Data Sets with Categorical Values," Data Ming and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.

[5]. S.A. Rauf, S. mahfooz, S. Khusro, H. Javed," enchanted k-means clustering algorithm to reduce number of iterations and time complexity", Middle–East J. Sci. Res. 12(7), 959-963(2012)

[6]. C.S. Li, "Cluster center initialization method for K-means algorithm over dataset with two clusters," in Proceeding of international conference on Advances in Engineering, pp. 324-328, 2011

[7]. A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data. Englewood Cliffs," NJ: Prentice-Hall, 1988.

[8]. S. Z. Selim and M. A. Ismail, "K-means type algorithms: a generalized convergence theorem and characterization of local optimality," in IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 6, No. 1, pp. 81--87, 1984.

[9]. Jieming Zhou, J.G. and X. Chen, "An Enhancement of K-means Clustering Algorithm," in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing, 2009.

[10]. M.P.S Bhatia, Deepika Khurana, "Analysis of Initial Centers for k-meanss Clustering Algorithm ," International Journal of Computer Applications (0975 – 8887) Volume 71– No.5, May 2013

[11]. Dr. M.P.S Bhatia1 and Deepika Khurana, " Experimental study of Data clustering using k- Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013

[12]. Kohei Arai and Ali Ridho Barakbah "Hierarchical K-means: an algorithm for centroids initialization for K-means ,"Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering, Saga Univ. Saga University, Vol. 36, No.1, 2007 36-1 (2007),25-31

[13]. Napoleon, D. and P.G. Lakshmi "An efficient K-means clustering algorithm for reducing time complexity using uniform distribution data points ," 2010 in Trendz in Information Sciences and Computing (TISC), Chennai

[14]. S. Ray, and R. H. Turi, "Determination of number of clusters in kmeans clustering and application in colour image segmentation, "In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, pp.137-143.

[15]. Dong, J. and M. Qi, "K-means Optimization Algorithm for Solving Clustering Problem," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Moscow 2009.

[16]. Goyal, M. & Kumar, S. "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability" Journal of The Institution of Engineers (India): Series B, 2014, Volume 95, Number 4, Page 345.

[17]. Er. Nikhil Chaturvedi et al., "Improvement in K-means Clustering Algorithm Using Better Time and Accuracy", International Journal of Programming Languages and Applications (IJPLA) Vol.3, No.4, October 2013.