A Novel Benchmark K-Means Clustering on Continuous Data

K. Prasanna
Assistant Professor, Department of Computer Science & Engineering
Annamacharya Institute of Technology & Sciences::Rajampet
Kadapa, Andhra Pradesh, India
prasanna.k642@gmail.com

M. Sankara Prasanna Kumar
Assistant Professor, Department of Information Technology
Annamacharya Institute of Technology & Sciences::Rajampet
Kadapa, Andhra Pradesh, India
sankaraprasannakumar@gmail.com

G. Surya Narayana
Assistant Professor, Department of Computer Science & Engineering
Annamacharya Institute of Technology & Sciences::Rajampet
Kadapa, Andhra Pradesh, India
surya.aits@gmail.com

***Abstract* – Cluster analysis is one of the prominent techniques in the field of data mining and k-means is one of the most well known popular and partitioned based clustering algorithms. K-means clustering algorithm is widely used in clustering. The performance of k-means algorithm will affect when clustering the continuous data.**

In this paper, a novel approach for performing k-means clustering on continuous data is proposed. It organizes all the continuous data sets in a sorted structure such that one can find all the data sets which are closest to a given centroid efficiently. The key institution behind this approach is calculating the distance from origin to each data point in the data set. The data sets are portioned into k-equal number of cluster with initial centroids and these are updated all at a time with closest one according to newly calculated distances from the data set.

The experimental results demonstrate that proposed approach can improves the computational speed of the direct k-means algorithm in the total number of distance calculations and the overall time of computations particularly in handling continuous data.

Keywords: cluster analysis, data mining, k-means clustering algorithm and continuous data.

I. INTRODUCTION

Clustering is the process of partitioning or grouping a given set of patterns or data sets into disjoint clusters. Clustering is a process in which a group of unlabeled patterns are partitioned into a number of sets so that similar patterns are assigned to the same cluster, and dissimilar patterns are assigned to different clusters. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics.

Several algorithms have been proposed in the literature for clustering. Among the algorithms, k-means algorithm has gained to be more effective and prominent in producing good clustering results for much of the practical application. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for multi dimensional datasets and continuous data. Since the continuous data requires additional efforts to handle fractional or decimal data part in the given data sets.

We propose a novel approach for implementing the k-mans method. Our algorithm produces clustering results to the basic k-means algorithm. It has significant better performance with the basic k-means algorithm in most cases.

II. BASIC PRELIMINARIES

There are two goals for clustering algorithms:

- Determining good clusters and doing so efficiently.
- Clustering is used in application domains such as in data mining and knowledge discovery, statistical data analysis, data classification and compression, medical image processing and bioinformatics.

K-means clustering:

The data clustering [3, 5] is process of grouping raw data to find clusters or groups of similar behavior data. In each cluster, members have some similarity in type of data. The principles of data clustering are finding value of score in similarity, and assigning each member to be in the same group of other members that have similar or same score. The clustering method relies on the similarity measurement to automatically from groups of relevant or similar data members.

K-means clustering algorithm is a prominent technique to cluster data. K-means is a nonhierarchical clustering and use looping to group data into K groups [2]. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The summarization of the k-means algorithm [1] in following steps:

1. Specify group number and select initial centroid of each group randomly.
2. Calculate Euclidean distance for each data member and centroid to assign members to the closest centroid.
3. Calculate distance's mean of every data member and own centroid to define new centroid in each group.
4. Repeat steps 2 and 3 until each group has stable centroid or same centroid.

Therefore, the *k*-means clustering algorithm cannot satisfy the need for fast response time for some applications. How to reduce the computational time required to cluster a large dataset becomes an important operational objective. Especially with continuous data, the clustering becomes very tedious to cluster. To solve this and other related performance problems, we proposed a novel benchmarking approach based on the candidate centroid of a cluster. This method will reduce the number of distance calculations and the execution time.

III. RELATED WORK

The original k-means algorithm is very impressionable to the initial starting points. So, it is quite difficult for k-means to have refined initial cluster centers. Several methods have been proposed in the literature for finding the better initial centroids. And some methods were proposed to improve both the accuracy and efficiency of the k-means clustering algorithm.

In the original k-means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive. In [6] approach the required computational time is reduced when assigning the data elements to the appropriate clusters. But in this method the initial centroids are selected randomly**.**

In [7] it uses two algorithms; one is for finding the better initial centroids. And another one is an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time, and enhances the accuracy and efficiency. The standard k-means algorithm [8] the initial centroids are calculated systematically. It selects k-objects randomly from the given data set as the initial centroids. If different initial values are given for the centroids, the accuracy output by the standard k-means algorithm can be affected.

A novel clustering algorithm [9] is proposed for grouping of genes, called Divisive Correlation Clustering Algorithm. DCCA is able to produce clusters, without taking the initial centroids and the value of k, the number of desired clusters as an input. The time complexity of the algorithm is high and the cost for repairing from any misplacement is also high.

IV. PROPOSED WORK

In this section, we proposed an enhanced method for enhancing the performance of the k-means clustering on continuous data. In basic k-means algorithm, the problem is randomly initializing centroids. When dealing with continuous data it may leads to data loss and computational overhead on selecting centroids. Several works is carried under selection of centroids. In [6] the proposed enhancement will improve the efficiency of the k-means clustering algorithm. But this will suffers with initial centroid selection and it is very sensitive to handle continuous data and also to the initial starting points.

Another enhancement proposed [7] to improve accuracy and efficiency of the k-means clustering algorithm. The method used for finding the initial centroids [7] is computationally expensive. In this paper we proposed a new approach for finding the better initial centroids with reduced time complexity. For assigning the data points we follow the paper [6], [7]. The pseudo code for the proposed algorithm is outlined in algorithm.

Algorithm:

Step1: Sort the continuous data in non decreasing order

Step2: For the sorted data set, calculate ‘S’ i.e. number of data units per each cluster.

Step3: Partition the sorted data set into k-equal sets and with each set contains ‘S’ data elements.

Step4: For each set

a) Initialize a group centroid by calculating a mean value among the data items.

b) Assign all centroids at a time in the group that has closest centroid.

c) Recalculate the new positions of the assigned centroid and its neighbors in the data space according to their distances from the data set.

Step5: Repeat step 4 until the centroid no longer move or the maximum iteration number reached.

In the proposed algorithm first we are checking, the given data set contains continuous data or not. Here, the transformation is required, because in the proposed algorithm we calculate the distance from origin to each data point in the data set. So, for the different data points as showed, we will get the same Euclidean distance from the origin. This will result in incorrect selection of the initial centroids. To overcome this problem all the data points are transformed to the continuous data space.

In the next step, for each data point we calculate the s i.e. the initial data elements per each cluster. Then, the original data points are sorted accordance with the sorted distances. After sorting partition the sorted data points into k equal sets. In each set take the mean value points as the initial centroids. These initial centroids lead to the better unique clustering results. Next, for each data point the distance calculated from all the initial centroids. The next stage is an iterative process which makes use of a heuristic approach to reduce the required computational time. The data points are assigned to all the clusters at the same time that are having the closest centroids in the next step.

V. EXPERIMENTAL RESULTS

The experiments are conducted on two different continuous data sets. The first data set contains N=250 data units and the second data set contains N=1000 data units. The results are shown in figure 1 and figure 2. For each data set we applied both Basic K-means algorithm and the proposed approach. Both the algorithms need number of clusters as an input. For the basic k-means algorithm, it requires initial centroids as additional input. The enhanced method finds initial centroids systematically. The enhanced method requires only the data values and number of clusters as inputs.

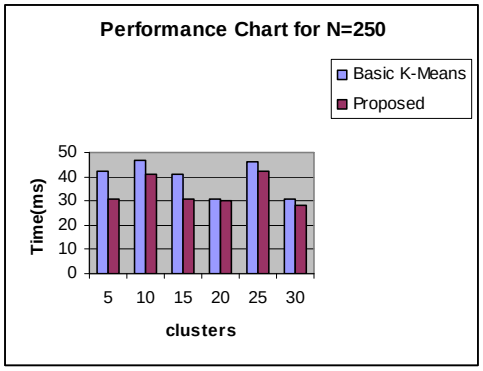

Figure 1: Performance comparison for N=250

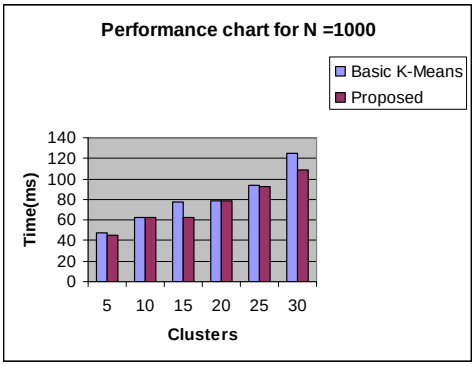

Figure 2: Performance comparison for N=1000

The basic k-means algorithm and the proposed approach are executed several times with different number of clusters for each N=250 and N=1000. At each execution the processing time is evaluated as shown in figure 1 and 2. The results obtained are also show that the proposed approach is producing better unique clustering results compared to the k-means algorithm in less amount of computational time

VI. CONCLUSIONS

The prominent clustering algorithm is the K-Means clustering algorithm. But it requires the number of clusters should relay on initial centers which are selected randomly. Moreover, the k-means algorithm is computationally very expensive and difficulty in clustering continuous data also. The proposed algorithm is found to be more accurate and efficient compared to the original k-means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the continuous data points to the suitable clusters. The proposed approach does not require any additional inputs and will produce accurate clustering results. As a future work, we can extend the approach for clustering discrete data which is having more scope for expansion.

REFERENCES

[1] Lawrence K.D., Kudyba S. and Klimberg R.K., "*Data mining methods and applications,"* USA: Auerbach Publications, 2008, pp. 83-104.

[2] Rahman H., "*Data mining applications for empowering knowledge societies,"* USA: Information Science Reference, 2009, pp. 43-54.

[3] Taniar D., *Data mining and knowledge discovery technologies*. USA: IGI Pub, 2008, pp. 118-142.

[4] Wang J., "*Data warehousing and mining: concepts, methodologies, tools, and applications,"* USA: Information Science Reference, 2009, pp. 303-335.

[5] Wu X. and Kumar V., "*The top ten algorithms in data mining,*" USA: CRC Press, 2009, p. 21, p. 93.

[6] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "*An Efficient enhanced k-means clustering algorithm*," journal of Zhejiang University, 10(7): 16261633, 2006.

[7] K. A. Abdul Nazeer and M. P. Sebastian, *"Improving the accuracy and efficiency of the k-means clustering algorithm,"* in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009),Vol 1, July 2009, London, UK.

[8] 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.

[9] A. Bhattacharya and R. K. De, *"Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles,"* bioinformatics, Vol. 24, pp. 1359-1366, 2008.