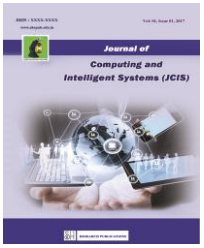
Original Research Article

J. Computing & Int. Systems (2018) 45–49

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in



ISSN: 2456-9496

AN OPTIMIZED CLUSTER CENTER INITIALIZATION USING K-MEANS AND CLUSTERING WITH LARGE APPLICATIONS

N. Nivetha[#1], R. Pugazendi [#2]

Received on 8th Aug 2018, Accepted on 20th Nov 2018

Abstract *— Cluster analysis is an essential and unsupervised data mining technique that is used for grouping objects automatically. Partitional clustering algorithms gather an individual partition of the data rather than a clustering structure. K-means is the most popular, simple and efficient method. The computational complexity of k-means cannot have problems with the size of the data set. One of the major drawbacks of Traditional k-means clustering algorithm's is the selection of initial centroids and do not work with huge data sets. To solve this problem a new hybrid technique has been proposed for k-means clustering and Clara. This algorithm is more efficient for working with large data sets. We have estimated the performance using three real data sets. The proposed technique is much more effective for improving accuracy and reduces computation time as well as iterations.*

Keywords: Data mining, Clustering, K-Means, Clara, Euclidean distance, Complexity.

I INTRODUCTION

Data mining is recognized as knowledge mining. Data mining is the greatest approach for extracting meaningful patterns from the vast quantity of data. The motivation of knowledge mining is for finding information that is unknown and unexpected. The pattern discovery process that leads to data mining automatically, otherwise semi-automatically deprived of any human intervention [1].

In data mining, the data considered in two alternative ways which include Supervised Learning and Unsupervised Learning methods. The objective of the supervised learning method is to discover the training data to identify the unknown value. Data which is unable to need specifically selected attribute is known as unsupervised learning method.

Clustering is the method of dividing a collection of articles into modules of similar objects. It is similarly called unsupervised classification. In further words, the data is organized into high intra period similarity and small inter period similarity. In clustering "related" objects are gathered in one group and "unrelated" objects are gathered in another group. Furthermost clustering techniques split data separation hooked on non-overlapping groups. Clustering is a vital task that discovers many fields, including image analysis, design recognition, information retrieval, and PC graphics [2].

Nowadays clustering numerous forms of algorithms are used in data mining, but only some popular algorithms are commonly used. The standard clustering algorithm generally consists of five different categories, they are partitioning clustering, hierarchical clustering, density-based clustering, grid-based clustering and model-based clustering. Essentially, all of the clustering algorithms use the distance measurement method. Each algorithm follows a distinctive method to find similar characteristics among the information factors. The simplest and essential model of cluster analysis is partitioning method, which classifies the items of a set into several amazing clusters.

Partition-based clustering strategies generate the clusters by making different partitions of the dataset. This strategy is used for analyzing the quality based on the distance between clusters. These kinds of algorithms iteratively refine the excellence of clusters to maximize intra-cluster similarity and consequently minimize inter cluster similarity. The excellence of the clustering result is dependent upon either the similarity measure utilized by the method and its implementation. The excellence of a clustering approach can be measured by its capability to find out the entire hidden pattern.

In partition based clustering, k-means algorithm is an unsupervised and an iterative method. It is easy and very fast, so in lots of sensible applications, the technique has proven to be a completely powerful manner that can produce top clustering effects. The classic k-means process is efficient for generating clusters for lots of realistic applications [12].

The original k-means algorithm consists of certain problems: 1) Determine the k value is difficult 2) algorithm is susceptible to the initial cluster, the selection of the initial centers can affect the clustering results and the efficiency of the algorithm. 3) The algorithm is vulnerable to irregular data, can produce the results in the local optimum solution. To overcome the above problems, lots of research scholars have been proposed, some development strategies to enhance the k-means algorithm.

.

* Corresponding author: E-mail: nivetha561992@gmail.com, pugazendi_r@gmail.com.

.

[1]Research Scholar, Dept. of Computer Science, Government Arts College, Periyar University, Salem – 7, Tamilnadu, India.

[2]Assistant Professor, Dept. of Computer Science, Government Arts College, Periyar University, Salem – 7, Tamilnadu, India.

II RELATED WORK

A novel clustering process built on density and k-means algorithm, which has proved the clustering accuracy, is proposed in [3]. This algorithm is used to overcome the group center initialization problem. The proposed model is implemented using the Matlab tool. Jain dataset and path base synthetic datasets are used for evaluating the accurateness of the clustering algorithm. An enhanced method that avoids for calculating the distance for each information object towards the group centers repeatedly.

This paper presented by a hybrid clustering algorithm that combines optimized k-means algorithm (PKM) then the genetic algorithm (PKGM). The new method is used to predict distance with the high-density part such as initial cluster centers. This algorithm is used for automatically searching aimed at the best cluster. The new algorithm proves the compactness and cluster quality [4].

In this author, we describe the kernel density –based technique to calculate the k-mean the initial seed value of the k-means clustering algorithm. The initial seed value is used for avoiding outliers in the dataset. Internal then external validity measures are used to check the clustering quality. The proposed method for validating the Dunn index, MSE, Mirkin, ARI, and DBI improves reproducibility of a cluster than the traditional k-means algorithm [5].

The author discussed the hierarchical k-implies algorithm. The knowledge is to decrease the underlying centroid for means calculation. It utilizes a whole clustering algorithm, outcomes of means and achieves its neighborhood ideal. This procedure is applied to the large data set of complex clustering cases clustering results are more robust for efficiency and iterations [6].

An optimized k-means genetic optimization algorithm built on density for computing the clustering effect. This algorithm used for choosing the initial centers for evaluating the crossover then mutation operators with a genetic algorithm. An enhanced method that can be avoided for calculating the distance to each data articles for the cluster centers repetitively, and save for running time. The experimental results show that the enhanced performance of our future method [7].

An enhanced k-means clustering approach is proposed to discover better initial centroids and more accurate clusters with saving running time. The Experimental results establish that the improved algorithm produces clusters with improved accuracy consequently increase the overall performance of the k-means algorithm.

This paper gives a modified model of the clustering, k-means algorithm that effectively removes the empty cluster problem. We've become demonstrated that the proposed algorithm is semantically comparable with the exclusive k-means approach and having no efficiency degradation through incorporated change. Effects of simulation experiments the usages of several data units prove our claim [8].

III RESEARCH METHODOLOGY

A. Initialization for Clustering Approaches

The functionality of clustering algorithm alterations is to improve the effectiveness of the main algorithms by solving their weakness. In addition to given that randomness is one of the strategies used for initializing a lot of clustering techniques, and giving every point an equal possibility be an initial one, it will be considered the objective of their weakness that has to be solved. However, due to the sensitivity of KMeans to their initial points, considered very high, we need to make them as near to global minima as possible to be able to enhance the clustering performance.

K-MEANS

The K-means algorithm is among the unsupervised learning strategies requires the user to select the number of clusters (K), to be created. Recall that k-means each group is characterized by its center. Consequently, the closest centroid is described for using the Euclidean distance with object and cluster. Next the assignment phase, the algorithm analyzes the new mean value of every cluster. The term cluster centroid update is used to design this phase. All the objects are reallocated again using the efficient cluster means. Almost all of the objects are reassigned once again working with the updated cluster means. While clustering the similarity measures are applied for cluster centroid methods.

CLARA

CLARA stands for Clustering Large Application. It is a development of PAM to handle with large data sets. It operates PAM on multiple random samples rather than the whole data set. The emphasis of clustering a huge number of objects instead of a small amount of objects in large dimensions. It randomly chooses the data and selects the medoid making use of PAM. In this paper, the Clara algorithm developing cluster quality to increase accuracy and reduce time.

B. HYBRID TECHNIQUE OF K-MEANS AND CLARA

Hybrid technique included is the best technique for acquiring accuracy in rainfall prediction. K-Means clustering and Multilayer perceptron are called a hybrid method of K-Means clustering and Multilayer perceptron. In this paper, K-Means and Multilayer perceptron have been implemented using JAVA. The K-means Cluster and Multilayer perceptron algorithm used to find the best accuracy and reduce an execution time.

This paper suggests an alternative hybrid technique for initializing centroids of the k-means algorithm. This strategy is to find the initial cluster points and intra-class similarity as well as inter-class similarity depending on a certain methodology that can improve the good quality of the resultant cluster.

CLARA algorithm can easily handle with noisy data. This algorithm is to avoid local optimization. K-means algorithm requires an initial clustering to evaluate by using assurance inspiring principle. Although, it can converge faster. K-means clustering and CLARA algorithm are used to increase the cluster quality, compactness and execution time.

The working procedure of a hybrid clustering algorithm

The hybrid clustering algorithm described as follows:
Each and every data object is built to perform clustering with CLARA algorithm from a, k clusters together with cluster centers are selected
Input above information then grouping by k-means algorithm
Performance of the result.

Initialization of cluster centroids
Step 1: Select number of centroids k for the clustering process.
Step 2: Compute the average distance of every data point
di(avgdi)=(w1*x1+w2*x2+w3*x3+…..wm*xm)/m
Where x= attribute's value, m= number of attribute, w= weight for multiply the distribution of cluster.
Step 3: Arrange all data objects constructed on the entire distance of the objects.
Step 4: Evaluate the distance among the objects and centroids.
Step 5: Assign every point di towards the cluster that ensures the nearest centroid
Step 6: Compute the new centroids for all clusters
Step 7: Repeat step 3 then step 6 until convergence conditions are met.

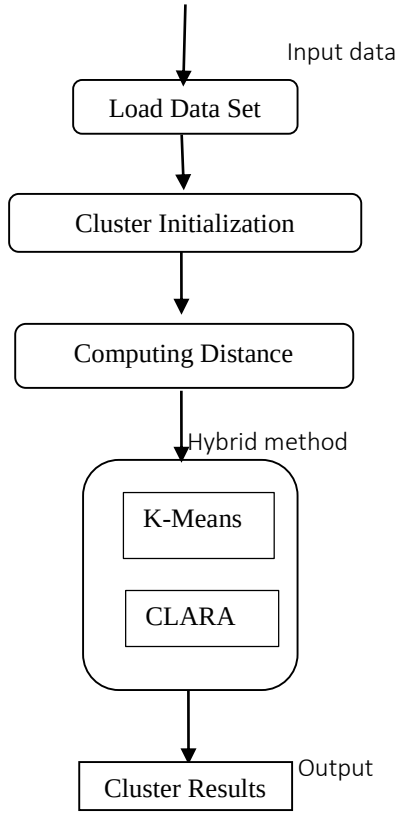


Figure I - System Architecture Diagram

IV EXPERIMENTAL RESULT

All the experimentations are conducted using a MATLAB programming language. The overall efficiency of our proposed algorithm is evaluated on the accuracy, computational time and a number of iterations preferred with various real-world data sets and compared to modified k-means clustering with hybrid method. K-means through random centroids initialization is unable to provide the unique clustering results, therefore, we made 5 times experiments and consider the average results. We estimated the proposed algorithm run on various datasets. The data sets are collected from the UCI repository [13]. Three different datasets are used for this work. They are Iris, Yeast and energy data sets. The iris, yeast and energy data sets are multivariate data sets. Each data set contains a different number of elements and instances. All of the three data sets are real data sets organized by the UCI repository. We operate them at the same experimental environment, evaluate the functionality effect. We looked at grouping comes about accomplished with the K-Means process utilizing arbitrary, starting focuses and initial centers determined by the proposed calculation. The effectiveness of cluster centroid is tested by applying k-means clustering with Clara. We observed that the results acquired by using a hybrid method are better than other algorithms.

Table 1- Data Set Description

Data Set	Size	Attribute
Iris	150	4
Yeast	900	8
Energy	1500	29

A. Accuracy

Accuracy implies most appropriate classification value for measuring the accurate instances through the algorithm for predicting cluster quality. Although evaluating the existing algorithm, a hybrid technique of K-means and CLARA algorithm accuracy and reliability value is higher for cluster quality.

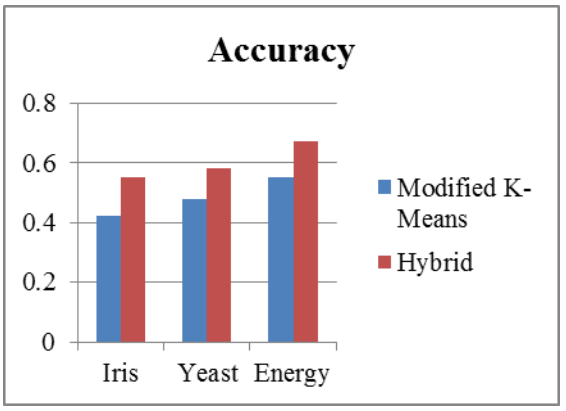


Figure II -1 Display Accuracy value evaluation with existing and proposed strategies graphically

B. Execution time

Execution time is also referred to as a development process in cluster centroid using three real data sets. A hybrid technique of K-means and CLARA algorithm reduce execution time for cluster quality for comparing the existing algorithms.

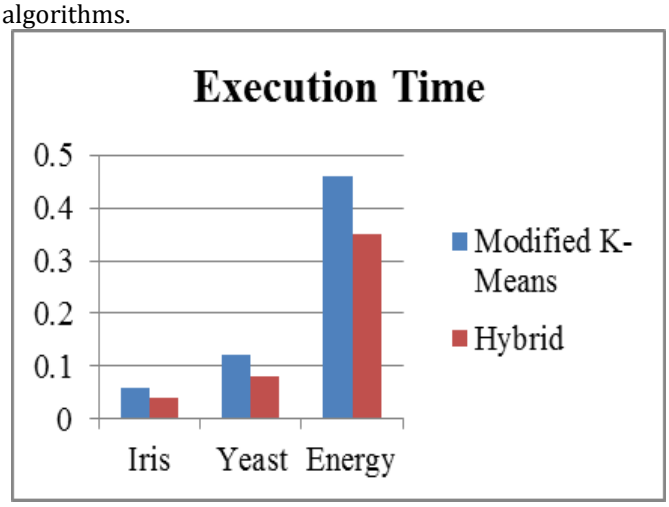


Figure II - 2 Display Execution Time comparing with existing and proposed strategies graphically

C. Iterations

Iteration is the strategy for dealing with a set of operations that manage with the clustering algorithm. Even though comparing the existing algorithm, a hybrid technique of K-means and CLARA algorithm reduce iterations for cluster centroid centroids.

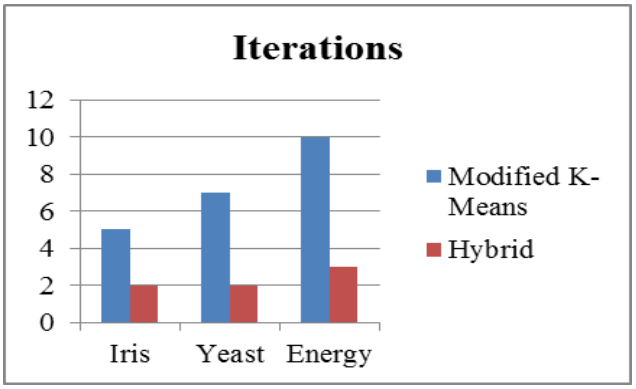


Figure II - 3 Display Iterations comparing with existing and proposed strategies graphically

V CONCLUSION

The selection of the initial centers of k-means affect the clustering results. The result of the traditional k-means clustering is usually unstable, that will be easy to reduce into a locally optimal solution. The proposed hybrid method of k-means and clara algorithms select initial points by high quality and outliers are screened out. The major improvement of the proposed approach is its simplicity because it creates the process of computation of initial centroids relatively simple. From the experimental performance, the algorithm can converge faster. The overall performance and accuracy of the algorithm are improved.

REFERENCES

[1] Han J and Kamber M. Data Mining: Concepts and Techniques, 2000 (Morgan Kaufmann, San Francisco, California).
[2] K.Jain and R.C.Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[3] Kabiru Dalhatu and Alex Tie Hiang Sim "Density base k-means Cluster Centroid Initialization Algorithm", International Journal of Computer Applications (0975-8887) Volume 137 – No.11, March 2016.
[4] Min Feng and Zhenyan Wang, "A Genetic K-Means Clustering Algorithm Based on the Optimized Initial Centers", Vol.4 No.3 ISSN 1913–8989 E-ISSN 1913-8997.
[5] Ajay Kumar and Shishir Kumar "Density Based Initialization Method for K-Means Clustering Algorithm" 08 October 2017.
[6] P.S. Bradley and U.Fayyad, "Refining Initial Points for K-means Clustering", Proc.15th International Conference on Machine Learning, pp. 91-99, 1988.
[7] "Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm", Hartono, EriantoOngko and Dahlan, International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 6, June 2015.
[8] Malay K.Pakhira, "A Modified k-means Algorithm to Avoid Empty Clusters", International Journal of Recent Trends in Engineering, Vol 1, No.1, May 2009.
[9] Tajunisha and Saravanan "Performance analysis of k-means with different initialization methods for high dimensional data" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010.
[10] https://en.wikipedia.org/wiki/Data mining
[11] https://www.datascience.com/blog/k-means-clustering
[12] http://www.sthda.com/english/articles/27-partitioning-clustering-essentials
[13] UCI Machine Learning Repository, https://archive.ics.uci.edu/m1/datasets.html
[14] K.Gaja Lakshmi and Dr.D.Prabha, "Clustering Big Data Using Normalization Based K-Means Algorithm", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3, March-2016.
[15] Sukhvir Kaur, "Survey of Different Data Clustering Algorithms", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May-2016.