Overcoming the Defects of K-Means Clustering by using Canopy Clustering Algorithm

Ambika S[1] Kavitha G[2]
[1]M.Tech. Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]SCE College of Engineering, Bangalore, India

Abstract— High dimension data clustering is the study of data that contains hundreds of dimensions. To improve the processing time of K-means clustering algorithm on high dimensional dataset by making use of canopy clustering algorithm. A canopy clustering algorithm uses the synthetic sampling method as the preprocessing step, as well as it uses the created T1 & T2 parameter values to create canopies and also provides initial cluster centers. Existing clustering algorithm normally works with the small dataset and it doesn't works with the high dimensional dataset because the algorithm may yields the inaccurate clusters by selecting the random cluster centers, and another problem is the number of required cluster or k-values are predefined by the user. The proposed algorithm works well with the high dimensional dataset and it over comes the limitations of the K-means clustering algorithm and minimizes the execution time of the existing algorithm.

Key words: High Dimensional Dataset, Data Mining, Synthetic Sampling, Parameter Estimator, K-Means Clustering Algorithm, Canopy clustering Algorithm

I. INTRODUCTION

Grouping is the procedure of sorting out data objects into a set of disjoint classes called groups. Grouping is a case of unsupervised order. Classification allows assigning objects in to a set of classes. Unsupervised implies that grouping does not relies on upon predefined classes and training while ordering the data objects. Group analysis tries to segment given information set into groups in light of determined features so that the data points inside a cluster are more like each other than the points in various clusters. Subsequently, a group is a gathering of objects that are comparable among themselves and unlike the objects having a place with different groups. Grouping is an essential range of exploration, which discovers applications in numerous fields including bioinformatics, data mining, and data mining pattern recognition, image processing, marketing, economics, and so on. Group investigation is a one of the essential information tool in the data mining.

II. LITERATURE SURVEY

A. "A Survey on K-mean Clustering and Particle Swarm Optimization" [1]

Partition-based cluster analysis is called k-means clustering. Concurring to the algorithm we firstly select k-values as starting group centroids, and then compute the separation between every cluster point and every item and assign it to the closest cluster, Modify the midpoints of all groups, repeat this procedure until the algorithm get terminates.

B. "A Novel Density based improved k-means clustering Algorithm – Dbkmeans" [2]

Proposed in our impersonation of the calculation, we just accepted covered groups of round or spheroid in nature. So the criteria for part or joining a group can be resolved in view of the quantity of assessed points in a group or the evaluated density of the group.

1) Limitations:

Joining of clusters is time consuming for the large dataset.

C. "Standard K-Means Clustering Algorithm"[3]

Yields means which can be known as the last constant around which every single other point in the dataset get grouped. This is so in light of the fact that the K-Means ends when either the groups repeat in the following cycle or when the methods repeat in the following emphasis.

D. "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points," [4]

Proposed a technique for making the K-Means calculation more important and expert; to show signs of improvement grouping with condensed many sided quality. The best algorithm was discovered in view of their presentation utilizing uniform appropriation data points. The exactness of the algorithm was examined amid various execution of the project on the data points. From the trial results, the uniform distribution data points are utilized effortlessly understand the values and get a predominant result.

E. "Research on K-Means Clustering Algorithm

An Improved k-means Clustering Algorithm" [5] Proposed an enhanced k-means calculation to comprehend inadequacies of standard k-means grouping algorithm, requiring a straightforward data structure to store some data in each iteration, which is to be utilized as a part of the following emphasis. The enhanced strategy abstains from registering the separation of every data item to the group canters repeatedly, sparing the running time.

Existing System	Definition	Related Works
Survey on k-means clustering algorithm	k-means clustering algorithm doesn't applicable for high dimensional dataset	James MacQueen Stuart Lloyd
Density based improved k-means	Joining of clusters is time consuming for the large dataset.	K. Mumtaz et al.
Standard K-means clustering algo	Algorithm chooses the random cluster center	H. kayacik a. Zincir-heywood

		m. Heywood
Efficient k-means algo for reducing time complexity	Reducing time complexity for the high dimensional dataset is expensive	D. Napoleon P. Ganga lakshmi

Table 1: Comparisons of Existing Systems

III. EXISTING SYSTEM

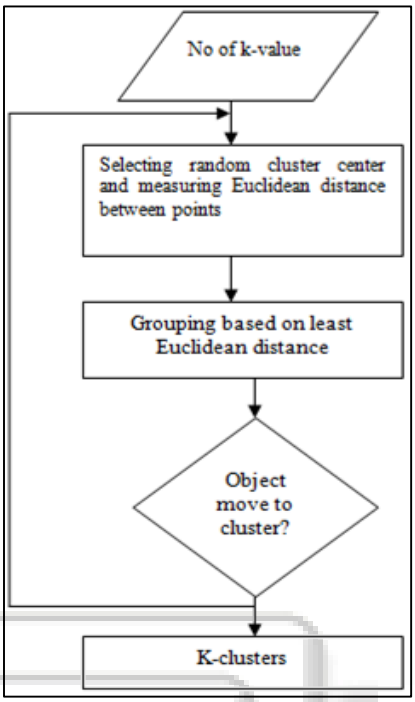

Fig. 1: Flow structure of K-means clustering algorithm

IV. PROPOSED SYSTEM

The goal of the proposed system for finding the initial seed selection and to minimize the execution time of the given large dataset.

The Fig 1depicts the flow structure of the proposed system.

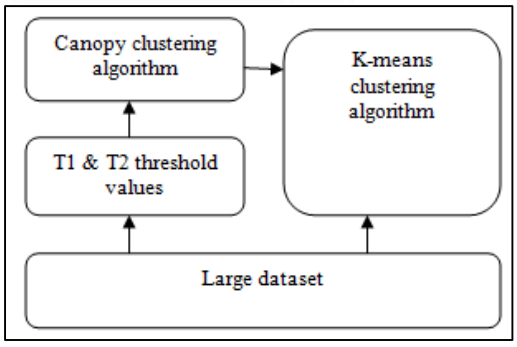

Fig. 2: Flow structure of proposed system

There are some modules used in the proposed system that are shown below

- Large dataset
- T1 & T2 threshold values
- Canopy clustering algorithm
- K-means clustering algorithm

A. Large Dataset

Large dataset consists of more than 500 attributes is called high dimensional dataset.

B. T1 and T2 Threshold Value

Using sample records the parameter estimator will create the T1 and T2 values for a given large dataset.

C. Canopy clustering algorithm

This algorithm results the no of canopies created and gives the cluster centre for the given dataset.

D. K-means clustering algorithm

K-means algorithm doesn't applicable for the large dataset to overcome this problem extracting the output from the canopy clustering algorithm and feeding it to the K-means algorithm as a input to reduce the execution time.

The proposed Algorithm as follows:

Algorithm: Canopy clustering algorithm

- Input: Dataset
- Output: K-value, initial seed selection
- Step 1: Select any point at random from the list to form a canopy center.
- Step 2: Approximate its distance to all other points in the list.
- Step 3: Put all the points which fall within the distance threshold of T1 into a canopy.
- Step 4: Remove from the list all the points which fall within the threshold T2
- Step 5: Repeat from step1 to 4 until original list becomes empty.

V. RESULT AND ANALYSIS

The processing time and the comparison result of the clustering algorithms are show below:

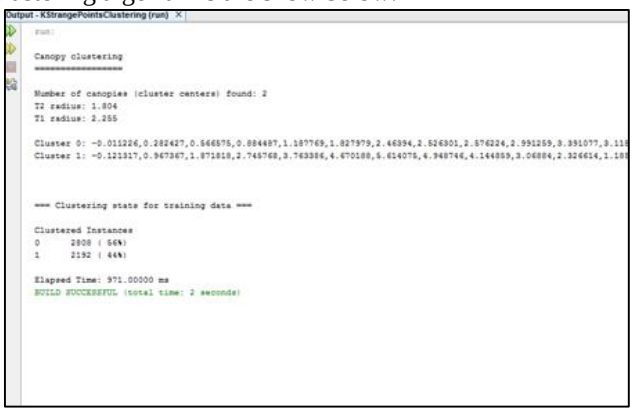

Fig. 3: Processing time of canopy clustering for waveform dataset

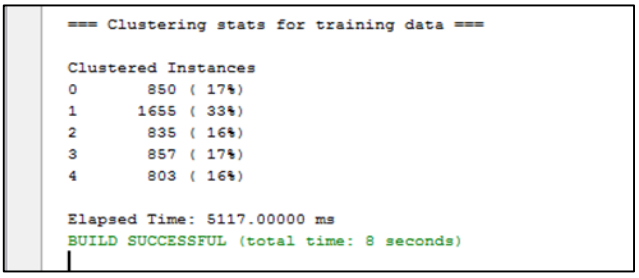

Fig. 4: Processing time of the K-means algorithm for waveform dataset

In this comparison result the K-means clustering algorithm will take 5000ms processing time for the waveform dataset because of the random initial seed selection likewise canopy clustering algorithm is taking 600ms for the given dataset then the proposed system canopy+ K-means is taking 400ms for a given waveform dataset.

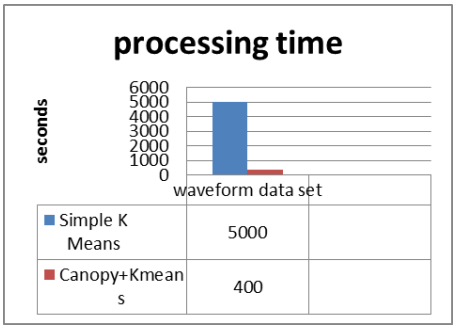

Fig. 5: comparison between k-means and Canopy clustering

VI. CONCLUSION

In this paper we are using a technique called canopy clustering module it automatically provides a number of k-values and the initial cluster centers for the given large dataset by taking these output values from the canopy clustering algorithm and then feeding it to the K-means algorithm as the input values to reduce the processing time for the given large dataset.

ACKNOWLEDGEMENT

I express deep thanks to Mrs. Kavitha G, Assistant Professor, Dept. of CSE for her valuable advice and support provided through out for completing this paper. I thank the anonymous referees for their reviews that significantly improved the presentation of this paper. Words cannot express our gratitude for all those people who helped us directly or indirectly in our endeavor. I take this opportunity to express my sincere thanks to all staff members of CS&E department of SCE for the valuable suggestion.

REFERENCES

[1] Dr. S. K. Singh and Terence Johnson, "K-strange points clustering algorithm, "Proceedings of International Conference on Computational Intelligence in Data Mining, 2014, Smart Innovation, Systems and Technologies, Springer publications, in press.
[2] Dr. S. K. Singh and Terence Johnson, "Improved collinear clustering algorithm in lower dimensions", proceedings of Second International Conference on Emerging Research in Computing, Information Communication and Applications, 2014, Elsevier publications, in press.
[3] Terence Johnson, "Bisecting collinear clustering algorithm", International Journal of Computer Science Engineering and Information Technology Research, TJPRC Pvt. Ltd, ISSN: 2249-6831, vol 3, Issue 5, Dec. 2013, pp. 43-46.
[4] FAHIM A.M., SALEM A.M., TORKEY F.A., RAMADAN M.A." An efficient enhanced k-means clustering algorithm" J Zhejiang Univ SCIENCE A 2006 7(10):1626-1633
[5] D. Napoleon & P. Ganga lakshmi "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points" IEEE,2010,pp,42-45
[6] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000
[7] M.P.S Bhatia & Deepika Khurana "Analysis of Initial Centers for k-Means Clustering Algorithm", IJCA, Volume 71 – No.5, pp 9 – 13, May 2013

