Marathwada Shikshan Prasarak Mandal's
Deogiri Institute of Engineering and Management Studies, Aurangabad

Seminar Report

On

K-Means Clustering

Submitted By

Shubhangi Dilip Rokade (36066)

Dr. Babasaheb Ambedkar Technological University
Lonere (M.S.)

Department of Computer Science and Engineering
Deogiri Institute of Engineering and Management Studies, Aurangabad
(2019- 2020)

Seminar Report
On

K-Means Clustering

Submitted By

Shubhangi Dilip Rokade (36066)

In partial fulfillment of
Bachelor of Technology
(Computer Science & Engineering)

Guided By
Mrs .Amruta Joshi

Department of Computer Science & Engineering
Deogiri Institute of Engineering and Management Studies, Aurangabad
(2019- 2020)

CERTIFICATE

This is to certify that, the Seminar entitled "**K-means clustering**" submitted by **Shubhangi Dilip Rokade** is a bonafide work completed under my supervision and guidance in partial fulfillment for award of Bachelor of Technology (Computer Science and Engineering) Degree of Dr. Babasaheb Ambedkar Technological University, Lonere.

Place: Aurangabad
Date:

Mrs.Amruta Joshi
Guide

Mr. S.B. Kalyankar
Head

Dr. Ulhas D. Shiurkar
Director,
Deogiri Institute of Engineering and Management Studies,
Aurangabad

Abstract

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Since K-means is widely used for general clustering, its performance is a critical point. This performance depends highly on initial cluster centers since it may converge to numerous local minima .There are many variation of K-means algorithm such as Lloyd's K-means clustering algorithm, Grid based K-means algorithm, Hierarchical K-means algorithm etc.

Keywords: Clusters, K-means clustering algorithm, Euclidian Distance.

Contents

List of Figures

List of Tables

List of Screens

1 .INTRODUCTION

In recent years there has been a tremendous growth in the volume of data. To draw meaningful insights from this mountain of data we need algorithms which can perform analysis on this data. Clustering is the process of grouping of data or dividing large data set into smaller data sets of some similarity so that the objects in the same cluster are more similar to each other and more different from the objects in the other group . Clustering is important analysis techniques that is employed to large datasets and finds its application in the fields like search engines, recommendation systems, data mining, knowledge discovery, bioinformatics and documentation. Nowadays, the data being generated is not only huge in volume, but is also stored across various machines all around the world. We need to process this data in parallel to reduce the cost of processing. But clustering is not free from weakness such that user has to specify the number of clusters to be generated before the start of algorithm, which can be very difficult especially when we are dealing with big data.

Another problem which may arise in applying K-Means clustering algorithm is the resolution problem, which are over-resolution or under-resolution. Over-resolution is a state when the final number of clusters generated is more than the actual number of clusters and underresolution is the converse of over-resolution.

1.1Criterion to benchmark clustering

When evaluating clustering methods for big data, specific criteria need to be used to evaluate the relative strengths and weaknesses of every algorithm with respect to the three-dimensional properties of big data, including Volume, Velocity, and Variety. Volume refers to the ability of a clustering algorithm to deal with a large amount of data.

To guide the selection of a suitable clustering algorithm with respect to the Volume property, the following criteria are considered:

(i)size of the dataset,

(ii) handling high dimensionality and

(iii) handling noisy data. Variety refers to the ability of a clustering algorithm to handle different types of data.

To guide the selection of a suitable clustering algorithm with respect to the Variety property, the following criteria are considered:
(i)type of dataset and
(ii) clusters shape. Velocity refers to the speed of a clustering algorithm on big data.
To guide the selection of a suitable clustering algorithm with respect to the Velocity property, the following criteria are considered:
(i)complexity of algorithm and
(ii) the run time performance. International Journal of Modern Trends in Engineering and Research

All rights Reserved 219 The corresponding criterion of each property of big data: Type of Dataset: The collected data in the real world often contain both numeric and categorical attributes. Clustering algorithms work effectively either on purely numeric data or on purely categorical data; most of them perform poorly on mixed categorical and numerical data types. Size of Dataset: The size of the dataset has a major effect on the clustering quality. Some clustering methods are more efficient clustering methods than others when the data size is small, and vice versa. Input Parameter: A desirable feature for practical clustering is the one that has fewer parameters, since a large number of parameters may affect cluster quality because they will depend on the values of the parameters. Handling Noisy Data: A successful algorithm will often be able to handle noisy data. Also, noise makes it difficult for an algorithm to cluster an object into a suitable cluster. Time Complexity: Most of the clustering methods must be used several times to improve the clustering quality. Therefore if the process takes too long, then it can become impractical for applications that handle big data. Stability: One of the important features for any clustering algorithm is the ability to generate the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

1.2Types of clustering

Clustering algorithms have many categories like hierarchical-based algorithms, partition based algorithms, density-based algorithms and grid based algorithms. Partition-based clustering is centroid based which splits data points into k partition and each partition represents a cluster . K-means is a clustering algorithm which is used widely. This technique will be useful in extraction

of useful information using cluster from huge Database. The overall purpose of the process of data mining is to extract useful information from a huge set of data and converting it into a form which is understandable for further use. For example Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing.

1.3 Types of learning method in data mining

In data mining, two learning methods used to mine data i.e. supervised learning and unsupervised learning.

Supervised learning: In this learning, data includes together the input and the desired result. It is the fast and perfect learning method. The accurate results are known and are given in inputs to the model during learning procedure. Neural network, Multilayer perception, Decision tree are supervised models.

Unsupervised learning: The desired result is not provided to the unsupervised model during learning procedure. This method can be used to cluster the input data in classes on the basis of their statistical properties only. These models are for various types of clustering, k-means, distances and normalization, self-organizing maps.

1.4 K-Means clustering

Clustering is important and essential concept of data mining field used in various applications. In Clustering, data are divided onto various classes. These classes represents some important features. Means, classes are the container of similar behavior of objects. The objects which behave or are closer to each other are grouped in one class and who are far or non-similar are grouped in different class. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity.

K-means clustering technique is a technique of clustering which is widely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means clustering: K-Means

clustering is unsupervised clustering technique in which data points are given as input and

without and predefined result it generate clustering results. It is heavily used in scientific and industrial applications. E.g. clustering of similar gene expression, weather data, text classification etc.

- Select K points as initial centroids.
- Repeat
- Form K cluster by assigning each point to its closest centroid.
- Recomputed the centroid of each cluster until centroid does not change

K-means clustering algorithm is famous clustering technique. It is used in many areas such as information retrieval, computer vision and pattern recognition. K-means clustering assigns n data points into k clusters so that similar data points can be grouped together. It is an iterative method which assigns each point to the cluster whose centroid is the nearest. Then it again calculates the centroid of these groups by taking its average. Properties of k-means algorithm :

1. Large data set are efficiently processed.
2. It often terminates at a local optimum.
3. It supports numeric values. International Journal of Modern Trends in Engineering and Research
4. The shape of clusters is convex

Clustering is an essential task in data mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering has been a widely studied problem in a variety of application domains including neural networks. Cluster analysis divides data into meaningful or useful groups or clusters. If meaningful clusters are the goal, then the resulting clusters should capture the natural structure of the data. Clustering is an important area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics etc. In this paper the cluster analysis is doing with method K-means clustering algorithm. K-means is a clustering algorithm that deals

with numerical attribute values (NAs) primarily, although it can also be applied to categorical datasets with binary values, by viewing the binary values as numerical. The K-means clustering algorithm for numerical datasets requires the user to specify the number of clusters to be produced and the algorithm builds and refines the specified number of clusters. But due to number of iterations in the loop, the basic K-means is computationally more time consuming and also it produces different results with different dataset .

Recent years, there are tremendous increase in the usage of internet. The usage of internet generates lots of data. These data are gaining its size as the year passes. The data are generated at record rate every day. To analyze those data and group into cluster is tedious task. The problem also lies in storing and retrieving of data. The analysis of these data points into different cluster is also a challenging task. Researchers have estimated that amount of information in the world doubles for every 20 months. However raw data cannot be used directly. Its real value is predicted by extracting information useful for decision support. In most areas, data analysis was traditionally a manual process. When the size of data manipulation and exploration goes beyond human capabilities, people look for computing technologies to automate the process

Data mining is process of extraction, transformation and loading of information to/from database or warehouse system. Storing and managing data, provide access to data analyst and data scientist to analyses the data for benefit of their business. There are two learning method presents to mine useful data from raw data. 1. Supervised Learning: In this type of learning, dataset is given as input and get output as desired, but in presence of trainer. Trainer generally trains the input dataset and classify it. Example of supervised learning techniques are: Neural network, Multilayer perception, Decision tree. 2. Unsupervised Learning: The desired result is not

provided to the unsupervised model during learning procedure. This method can be used
cluster the input data in classes on the basis of their statistical properties only. These models are for various type of clustering, k-means, distances and normalization, self-organizing maps. This paper reviews various methods and techniques used in literature and its advantages and limitations, to analyze the further need of improvement of k-means

2. LITERATURE SURVEY

Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. The most popular clustering algorithm used in several field is K-Means since it is very simple and fast and efficient. K-means is developed by Mac Queen [36]. The K-Means algorithm is effective in producing cluster for many practical applications. But the computational complexity of the original K-Means algorithm is very high, especially for large datasets. The K-Means algorithm is a partition clustering method that separates data into K groups. Main drawback of this algorithm is that of a priori fixation of number of clusters and seeds. To rectify the drawbacks of traditional K-Means clustering algorithm various measures were carried out in recent years by the researcher from multiple fields. Because of the modified approaches, cluster analyses were upgraded in multiple dimensions such as increasing the performance, decreasing the computational complexity, reducing the cluster error 14 15 and to increase cluster uniqueness.

In the recent years, many clustering algorithms for big data have been proposed which are based on distributed and parallel computation. Mac Queen in 1967 firstly proposed this technique. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulsecode modulation. Sometimes it is referred as Lloyd-Forgy because in 1965, E.W.Forgy published essentially the same method . According to K. A. Abdul Nazeer et al the major drawback of the k-means algorithm is about selecting of initial centroids which produces different clusters. But final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean. But this enhanced clustering method uses both the phases of the original k-means algorithm. This algorithm combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. But still there is a limitation in this enhanced algorithm that is the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points. According to Y. S. Thakare et al. , the performance of k-means algorithm which is evaluated with various databases such as Iris, Wine, Vowel,

Ionosphere and Crude oil data Set and various distance metrics. It is concluded that performance of k-means clustering is depend on the data base used as well as distance metrics. Soumi Ghosh et al. proposed a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based Fuzzy C-Means clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms. The result of this comparative study is that FCM produces closer result to the K-means but still computation time is more than k-means due to involvement of the fuzzy measure calculations. Sakthi et al. proposed that due to the increment in the amount of data across the world, analysis of the data turns out to be very difficult task. To understand and learn the data, classify those data into remarkable collection. So, there is a need of data mining techniques. R. Amutha et al. proposed that when two or more algorithms of same category of clustering technique is used then best results will be acquired. Two k-means algorithms: Parallel k/h-Means Clustering for Large Data Sets and A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets. Parallel k/h-Means algorithm is designed to deal with very large data sets. Novel K-Means Based Clustering provides the advantages of using both HC and K-Means. Using these two algorithms, space and similarity between the data sets present each nodes is extended. Nidhi Singh et al. proposed the comparative analysis of one partition clustering algorithm (k means) and one hierarchical clustering algorithm (agglomerative).On the basis of accuracy and running time the performance of k-means and hierarchical clustering algorithm is calculated using WEKA tools. This work results that accuracy of k-means is higher than the hierarchical clustering for iris dataset which have real attributes and accuracy of hierarchal clustering is higher than the k-means for diabetes dataset which have integer, real attribute. So for large datasets k means algorithm is good. Shi Na et al. Proposed the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process affects the efficiency of clustering algorithm. Cui, Xiaoli, et al. proposed optimized big data K-Means using Map-Reduce in which they claimed to counter the iteration dependence of Map-Reduce jobs. They used a sequence of three Map-Reduce (MR) jobs. However, in their approach sampling technique is used in the first M-R job and in the final M-R Job the data set is mapped to centroids using the Voronoi diagram. Variety is an important feature in big data so using sampling techniques is questionable when applied to huge data sets in maintaining the quality of clustering. According to Weizhong Yan et al. an MPI based spectral clustering algorithm.

S. No	Author(s)	Description(s)
1	Steven J. Phillips et al, 2002	Two simple modifications were carried out in KMeans Clustering Algorithms to improve the running time without Changing the Output. The two resulting algorithms were called Compare-Means & Sort Means.[
2	I.O. AyaquicaMartínez et al, 2005.	Described a new conceptual K-Means algorithm using similarity functions was proposed
3	Tong Zhaoet al, 2006.	A K-Means Clustering algorithm using Data Detection and Symbol-Timing Recovery for Burst-Mode Optical Receiver.
4	I.O. AyaquicaMartínez et al, 2006.	Described a new conceptual K-Means Algorithm based on Complex Features that did not use generalization lattices to built the concepts and allowed working with missing data was proposed.
5	Jan Carlo Barca et al, 2007.	Evaluated a modified K-Means algorithm that can be used for Removing Noise in multicolor motion Capture Image Sequences
6	Roy Varshavsky et al, 2007.	Evaluated Optimal K-Means (OKM) algorithm using a framework for Large Datasets.
7	IonutAlexandrescu et al, 2007.	Described a new K-Means clustering Algorithm using A Novel 3D Segmentation Method of the Lumenintima border on an "Intra Vascular Ultra Sound" (IVUS) Images.
8	JianwenXie et al, 2008.	A new K-means Clustering Algorithm using Meliorated Initial Centers and Its Application to Partition of Diet Structure showed its Feasibility and Validity

9	Grigorios F. Tzortzis et al, 2009.	The Global Kernel and K-Means Algorithm were used Clustering in Feature Space for identifying Nonlinearly separable Cluster.
10	Seraf´ın Alonso et al, 2011.	Describes a new K-Means clustering algorithm using 'Comparative Analysis of Power Consumption in University Buildings with EnvSOM modified by SelfOrganizing Map (SOM), was used to reduce the dimension of data and capture their electrical behaviors conditioned on the environment.
11	Mohammad SadeghRasooli et al, 2011.	Analyzed on K-Means clustering algorithm using 'Unsupervised Identification method'.
12	Ran Vijay Singh et al, 2011.	Evaluated a modified K-Means algorithm using data clustering approach for minimizing "Threshold distance" between data point and cluster's centroid
13	LihongZheng et al, 2011.	Described an improved K-means algorithm using Character Segmentation for License Plate Recognition (LPR) system
14	Roberto Rocci et al, 2011.	Analyzed the combination of Linear Discriminate Analysis and K-Means for Factor Discriminates KMeans (FDKM) clustering algorithm using an application on real-world data
15	Juan Carlos Rojas Thomas, 2011 et al,	A New Clustering Algorithm Based on K-Means Using a Line Segment as Prototype which captured the axis that presented the biggest variance of the cluster.
16	A.Pethalakshmi, et al,2011	Modification in K-Means clustering algorithm through affinity measure to increase the cluster uniqueness.
17	Mohammed Goryawala	Analyzed on 3-D Liver Segmentation using

	et al, 2012.	combined approach of K-means and segmentation algorithm.[
18	U. Siddiqui et al, 2012.	The Optimized K−Means (OKM) Clustering algorithm using "An Image Segmentation" with the capability of avoiding the Dead Centre and Trapped Centre at local minima.
19	MihaiVlase et al, 2012.	An Improvement of K-means Clustering algorithm Using various Patent Metadata
20	Xiaoping Li et al, 2012.	Evaluates Clustering in Image Indexing using KMeans clustering algorithm & Genetic Algorithm.
21	ManojKumar.V et al, Feb 2013.	This paper focused on enhancement of k-Means clustering by eliminating noise and unwanted region without any loss of important information in the image.

Table 2.1: K-Means and its enhancement

S. No.	Author	Method Used	Data source	Review	Limitation
1	K. A. Abdul Nazeer et al.	K-Means Algorithm	Iris Dataset	An enhanced clustering method propose to find initial centroids efficiently assign data points to cluster. Improve the efficiency and accuracy of k means algorithm.	Limitation in this enhanced algorithm that is the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points.
2	Soumi Ghosh et al.	K-means algorithm, Fuzzy C-means algorithm,	Iris and plant Dataset	Comparative analysis of Fuzzy C-means and K-means on the basis of time complexity. K-means algorithm seems to be superior than Fuzzy C-means	computation time is more than k-means due to involvement of the fuzzy measure calculations
3	Shafeeq et al.	modified K-means algorithm	random numbers of 300,500 and 1000 data points	Number of clusters are find in the proposed method on the run based on the cluster quality output. It is work for both known no. of cluster in advance as well as unknown no. of cluster.	proposed approach takes more computational time than the K-means for larger data sets

4	Junatao Wang et al.	K-Means algorithm	Data set from UCI Repository of Machine Learning Databases	Modified algorithm decrease the impact of noise data on k-means algorithm and clustering results are more accurate	Impact of noise are more in forming cluster
5	Shi Na et al.	K-Means algorithm	Data set from UCI Repository of Machine Learning Databases	Improve the speed and accuracy of clustering, reducing the computational complexity of the k-means	Centroid selection algorithm is not effective

Table 2.2: K-Means and its Description

3. BRIEF ON SYSTEM

The k-means clustering algorithm and Euclidean distance to cluster the following eight examples into three clusters: A1= (2, 10), A2= (2, 5), A3= (8, 4), A4= (5, 8), A5= (7, 5), A6= (6, 4), A7= 1(1, 2), A8= (4, 9). Find the new centroid at every new point entry into the cluster group. Assume initial cluster centers as A1, A4 and A7.

Answer:

K-means clustering algorithm is one of the most well-known partitioning algorithms. In this algorithm we are taking the number of inputs, represented with the k, the k is called as clusters from the data set. The value of k will define by the user and the each cluster having some distance between them, we calculate the distance between the clusters using the Euclidean distance formula.

d(a,b) denotes the Eucledian distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b) = \sqrt{((xb-xa)^2 + (yb-ya)^2)}$

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

Iteration I:

A1:

d(A1, seed1)=0

d(A1, seed2)= $\sqrt{13} > 0$

d(A1, seed3)= $\sqrt{65} > 0$

A1 ∈ cluster1

A2:

d(A2,seed1)= √25 = 5

d(A2, seed2)= √18 = 4.24

d(A2, seed3)= √10 = 3.16

A2 ∈ cluster3

A3:

d(A3, seed1)= √36 = 6

d(A3, seed2)= √25 = 5

d(A3, seed3)= √53 = 7.28

A3 ∈ cluster2

A4:

d(A4, seed1)= 13

d(A4, seed2)=0

d(A4, seed3)= 52

A4 ∈ cluster2

A5:

d(A5, seed1)= √50 = 7.07

d(A5, seed2)= √13 = 3.60

d(A5, seed3)= $\sqrt{45}$ = 6.70

A5 ∈ cluster2

A6:

d(A6, seed1)= $\sqrt{52}$ = 7.21

d(A6, seed2)= $\sqrt{17}$ = 4.12

d(A6, seed3)= $\sqrt{29}$ = 5.38

A6 ∈ cluster2

A7:

d(A7, seed1)= $\sqrt{65}$

d(A7, seed2)= $\sqrt{52}$

d(A7, seed3)=0

A7 ∈ cluster3

A8:

d(A8, seed1)= $\sqrt{5}$

d(A8, seed2)=$\sqrt{2}$

d(A8, seed3)= $\sqrt{58}$

A8 ∈ cluster2

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

Cluster Formation Iteration I:

Cluster 1: (2, 10)

Cluster 2: (8, 4), (5, 8), (7, 5), (6, 4), (4, 9)

Cluster 3: (2, 5), (1, 2)

centers of the new clusters:

C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

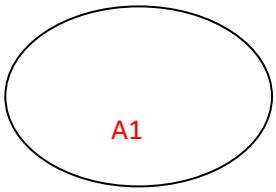


Fig:3.1:cluster1

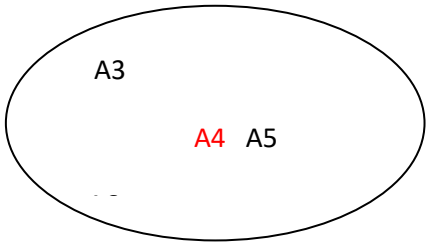


Fig:3.2 :cluster2

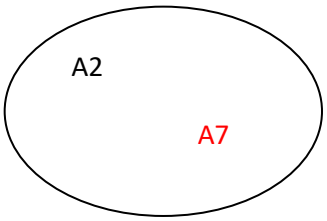
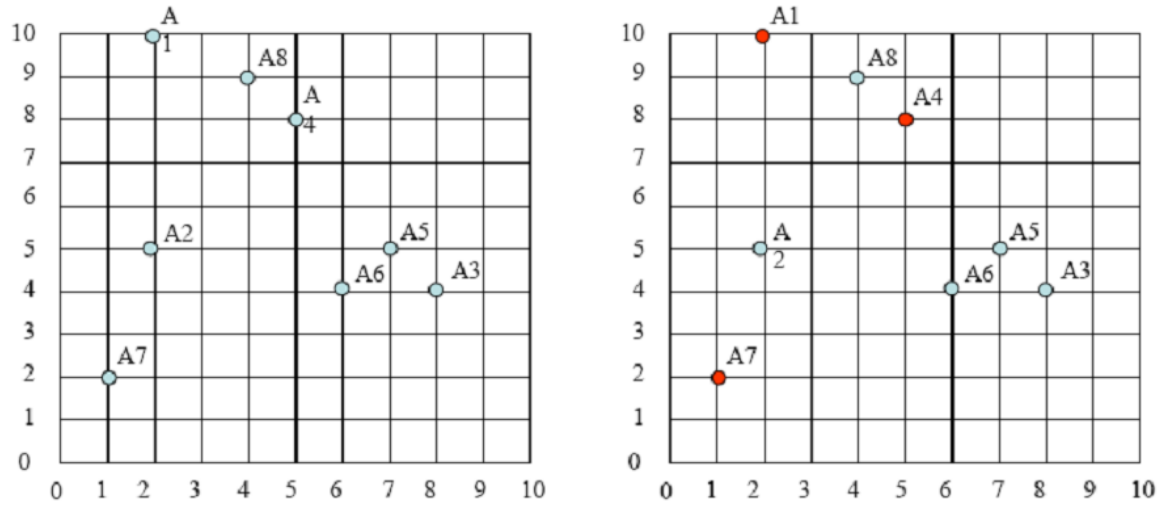


Fig:3.3:cluster3



Screenshot:3.1

Iteration II:

We would need two more epochs. After the 2nd epoch the results would be:

1: {A1, A8},2: {A3, A4, A5, A6}, 3: {A2, A7}

with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).

Cluster Formation Iteration II:

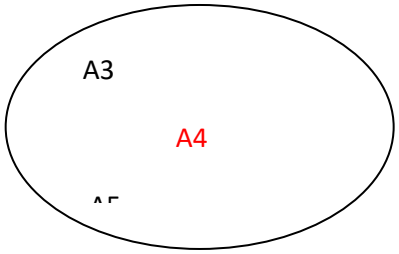


Fig:3.5:cluster5

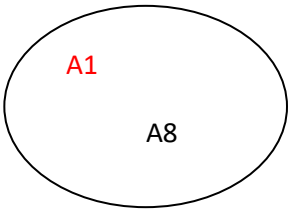


Fig:3.4:cluster4

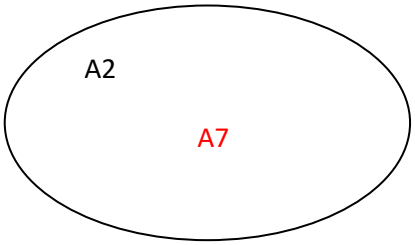
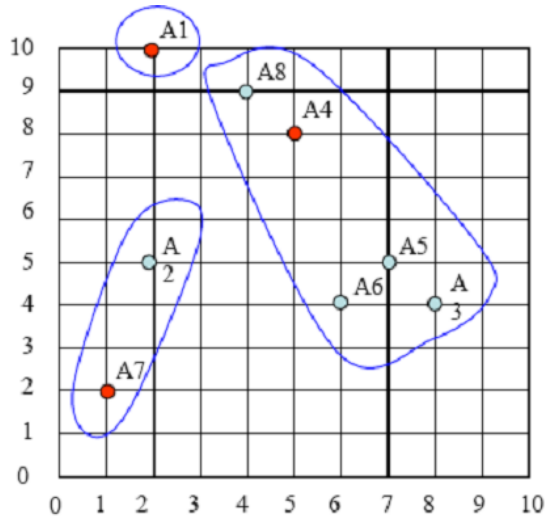
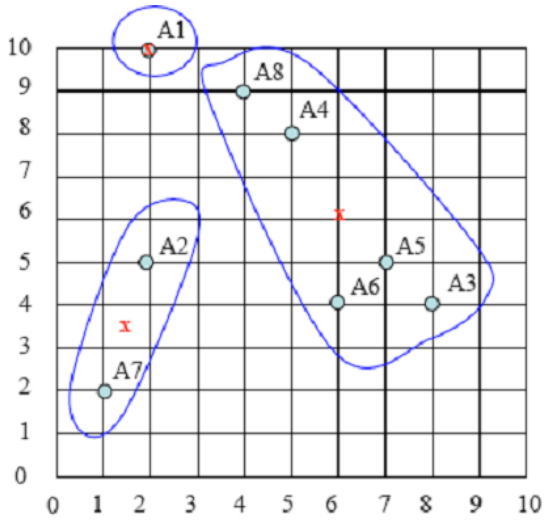


Fig:3.6:cluster6

Iteration III: After the 3rd epoch, the results would be:





Screenshot:3.2

1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}

with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).

Cluster Formation Iteration III:

Here we see the cluster values are no change between the Iteration 2 and the iteration 3, then we stop the iteration/ this data set with the 3 centroids generates the cluster groups in 3 iterations.

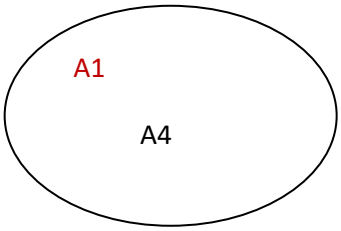


Fig:3.7:cluster7

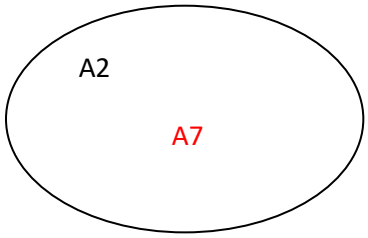


Fig:3.8:cluster8

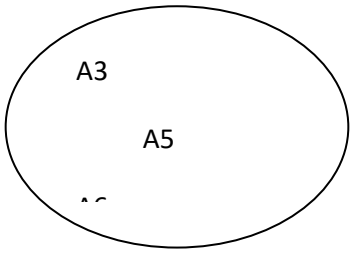
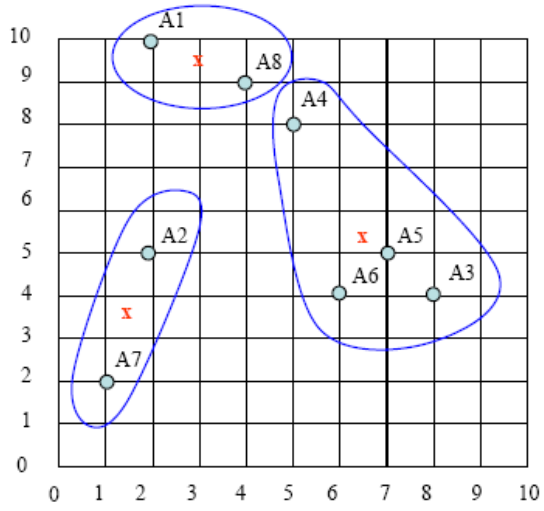
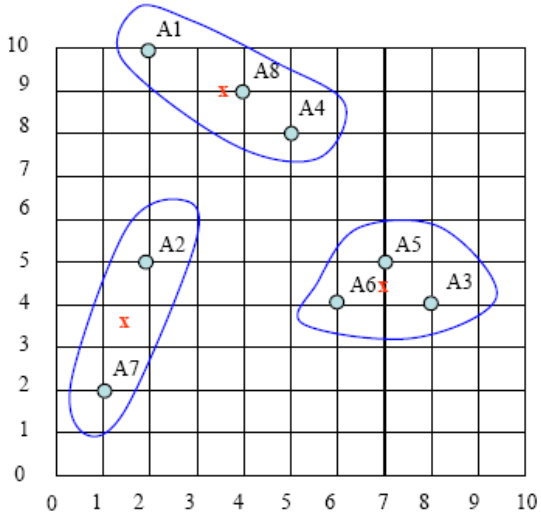


Fig:3.9:cluster9





Screenshot:3.3

4. CONCLUSIONS

4.1 Conclusion

In this paper k-means clustering techniques and method are reviewed. K-means being most famous among data scientist need further improvement in various section of algorithm. The outliers, empty clusters and selecting centroid for datasets are still a challenging task. Hence various further research needed to focus on these mentioned issues. Table I. presents various techniques and its limitation are present in proposed k-means algorithm. They need further enhancement due to increase of size of data as of now. This paper has make an attempt to review a significant number of papers to deal with the present algorithm of k-means. Present study illustrate that k-means algorithm can be enhanced by selecting centroid point appropriately.

4.2 Application

1. Document Classification

2. Delivery Store Optimization

3. Identifying Crime Localities

4. Customer Segmentation

5. Fantasy League Stat Analysis

6. Insurance Fraud Detection

7. Rideshare Data Analysis

8. Cyber-Profiling Criminals

9. Call Record Detail Analysis

10. Automatic Clustering of IT Alerts

REFERENCES

1] E. A. Khadem, E. F. Nezhad, M. Sharifi, “Data Mining: Methods & Utilities”, Researcher2013; 5(12):47-59. (ISSN: 1553-9865).

[2] Namrata S Gupta, Bijendra S Agrawal, Rajkumar M. Chauhan, ìSurvey On Clustering Technique of Data Mining, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN:2328-3491

[3] Malwindersingh, Meenakshibansal ,î A Survey on Various KMeans algorithms for Clustering, IJCSNS International Journal of Computer Science and Network Security, VOL.15 No.6, June 2015

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to **Prof. S.B.Kalyankar**, HOD-Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for his generous guidance, help and useful suggestions.

I express my sincere gratitude to **Prof. Amruta Joshi**, Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I am extremely thankful to **Dr.Ulhas Shiurkar**, Director, Deogiri Institute of Engineering and management Studies Aurangabad, for providing me infrastructural facilities to work in, without which this work would not have been possible.

Signature of Student

Shubhangi Dilip Rokade

(36066)