



**L** LOVELY  
**P** ROFESSIONAL  
**U** NIVERSITY

---

**ENHANCEMENT IN K-MEAN CLUSTERING BY REDUCING ITS  
PROCESSING TIME AND NUMBER OF ITERATIONS IN DATA MINING**

A DISSERTATION

SUBMITTED BY

MANDEEP KAUR (11305313)

**Department of Computer Science and Engineering**

In partial fulfillment of the requirements for the

Award of the degree

Of


**Master of Technology in computer Science & Engineering**

**Under the guidance of**

Mrs. RAJDEEP KAUR

**(May 2015)**

### PAC APPROVAL

 **LOVELY PROFESSIONAL UNIVERSITY**  
*Transforming Education. Transforming India*

School of: Science & Technology

---

**DISSERTATION TOPIC APPROVAL PERFORMANCE**

Name of the Student: <u>Hardeep Kaur</u>	Registration No.: <u>11305812</u>
Batch: <u>2K13-2K15</u>	Roll No.: <u>009</u>
Session: <u>14-15</u>	Parent Section: <u>RK2305</u>
Details of Supervisor:	Designation: <u>A.P</u>
Name: <u>Rajdeep Kaur</u>	Qualification: <u>M. Tech C.S.E</u>
UID: <u>10973</u>	Research Experience: <u>2.5 years</u>

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Data mining (Clustering) (Image Homogeneity)
2. Pre processing
3. Association rules

Signature of Supervisor: Rajdeep Kaur  
10973

PAC Remarks:

First topic approved  
11/11/14

APPROVAL OF PAC CHAIRPERSON: 11/11/14 Signature: [Signature] Date: 30/11/14

\*Supervisor should finally encircle one topic out of three proposed topics and put up for a approval before Project Approval Committee (PAC)  
\*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.  
\*One copy to be submitted to Supervisor.

**ABSTRACT**

This work presents an overview of the K-means clustering algorithm & various enhanced variations done on K-means clustering algorithm. K-means is the basic algorithm used for discovering clusters within a dataset. The initial point selection effects on the results of algorithm, both in the number of clusters found and their centroids. Methods to enhance the k-means clustering algorithm are discussed. By using these methods efficiency, accuracy, performance and computational time are improved. Some enhanced variations improve the efficiency and accuracy of algorithm. Basically in all the methods the main aim is to reduce the number of iterations which will decrease the computational time. Studies shows that K-means algorithm in clustering is widely used technique. Various enhancements done on K-mean are collected, so by using these enhancements one can build a new hybrid algorithm which will be more efficient, accurate and less time consuming than the previous work.

**CERTIFICATE**

This is to certify that **Mandeep Kaur** has completed M.Tech dissertation proposal titled **“Enhancement in k-mean clustering by reducing its processing time and number of iterations in data mining”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name: Rajdeep Kaur

UID: 16973

### ACKNOWLEDGEMENT

First and foremost, I want to thank the Department of CSE of Lovely Professional University for giving me permission to begin Thesis in first instance, to do necessary research work and to use required data.

I would like to acknowledge the assistance provided to me by the library staff of L.P.U. Inspiration to action is the most important ingredient required throughout the task. I am deeply indebted to my mentor Mrs. Rajdeep Kaur (Asst Prof) whose help, stimulating suggestions and encouragement helped me in all the time of research.

I express my gratitude to my parents for being a continuous source of encouragement and for their financial aids given to me. Finally, I would like to express my gratitude to all those who helped and supported me.

## DECLARATION

I hereby declare that the dissertation proposal entitled, “**Enhancement in k-mean clustering by reducing its processing time and number of iterations in data mining**” submitted for the M. Tech. Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 29 April 2015

Investigator:

Reg. No 11305313

## TABLE OF CONTENTS

<b>Chapter No: Chapter Name</b>	<b>Page no.</b>
Chapter 1: Introduction.....	1
1.1 Introduction data mining.....	1
1.1.1 Data mining KDD.....	2
1.1.2 Data mining functions.....	3
1.1.3 Need of Data Mining.....	4
1.1.4 Data Mining process.....	4
1.2 Clustering in Data Mining.....	5
1.2.1 Clustering requirements.....	9
1.2.2 Applications of clustering.....	9
1.3 K-mean clustering algorithm.....	10
1.3.1 Drawbacks of KMCA.....	13
1.3.2 Various enhancement in k-mean.....	13
Chapter 2: Review of literature.....	16
Chapter 3: Present work.....	30
3.1 Problem formulation.....	30
3.2 Objectives and Methodology.....	30
3.2.1 Steps for performing proposed work .....	30
3.3 Basic Design of Proposed Work.....	31
Chapter 4: Result and Discussion.....	33
Chapter 5: Conclusion and Future scope.....	42
5.1 Conclusion.....	42
5.2 Future Scope.....	42
Chapter 6: References.....	47
Chapter 7: Appendix.....	49

**LIST OF FIGURES**

<b>List of figures</b>	<b>Page No</b>
Figure 1.1 Data mining concept.....	1
Figure 1.1.1 DM as a step in KDD.....	2
Figure 1.1.4 DM process.....	5
Figure 1.2.1 Clustering in DM.....	6
Figure 1.2.2 Types of clustering in DM.....	6
Figure 1.2.3 Partitioning clustering.....	7
Figure 1.2.4 Hierarchical clustering.....	7
Figure 1.2.5 Density based clustering.....	8
Figure 1.3.1 Initial clustering.....	11
Figure 1.3.2 Iterate step.....	12
Figure 1.3.3 Final clustering.....	12
Figure 3.3 Design of proposed work.....	31
Figure 4.1.1 Dataset clustered.....	33
Figure 4.1.2 3D plane clustering.....	34
Figure 4.1.3 First iteration of clustering.....	35
Figure 4.1.4 Coloring of clusters.....	36
Figure 4.1.5 Data points represent by color for clustering.....	37
Figure 4.1.6 Clustering of data.....	38



<b>List of figures</b>	<b>Page No</b>
Figure 4.1.7 3D representation.....	39
Figure 4.1.8 Iteration graph.....	40
Figure 4.1.9 Time graph.....	41

# CHAPTER-1 INTRODUCTION

---

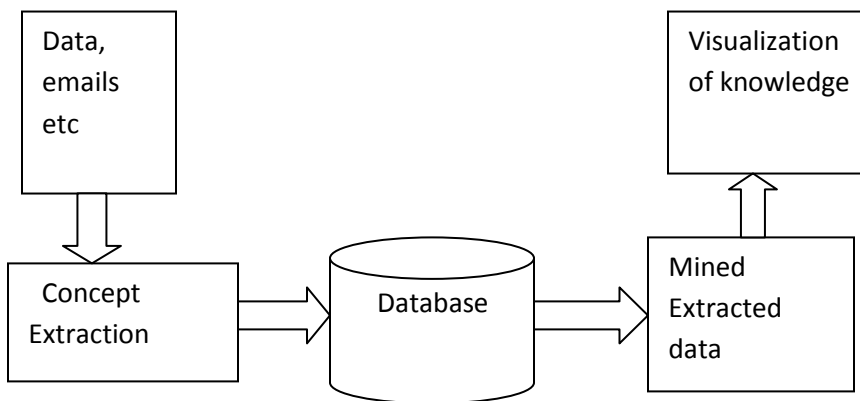
This chapter gives introduction about data mining, need of DM, various concepts in DM and methods of DM.

## 1.1 Introduction of Data Mining:

DM is defined as course of taking out of the inherent and formerly unfamiliar and really possibly useful information from huge amount of data. DM is also called as withdrawal of hidden patterns.

It is used to determine pattern in data, process should be fully programmed or semiautomatic pattern discovery. And discovery must be expressive and meaningful. It is also a procedure of discovery hidden information in the database.

This process suggests either one or greater than one computer method to spontaneously examine and abstract information from raw facts contain inside the database, it is part of information discovery process.



**Fig.1.1 Data mining concept**

DM relates many techniques to huge facts to generate models or attractive patterns for user and abstract hidden patterns. DM is also called knowledge discovery process as shown diagrammatically.

DM is method of finding relationship between large amount data. It may fully automate or semi automated process to discover knowledge that is useful for user.

DM applies technique to huge amount of facts to generate models or interesting pattern for client and will extract hidden patterns. [17]

### 1.1.1 Data mining KDD:-

KDD is an iterative process which contains following steps.

1. Data cleaning: - Noisy and conflicting data is removed from information by using various techniques like find missing values; remove noisy data by applying binning method, regression, outlier detection.
2. Data integration: after data cleaning process means after removing noise from data integration process is performed in data preprocessing process. Data integration means collecting or combining all cleaned data.

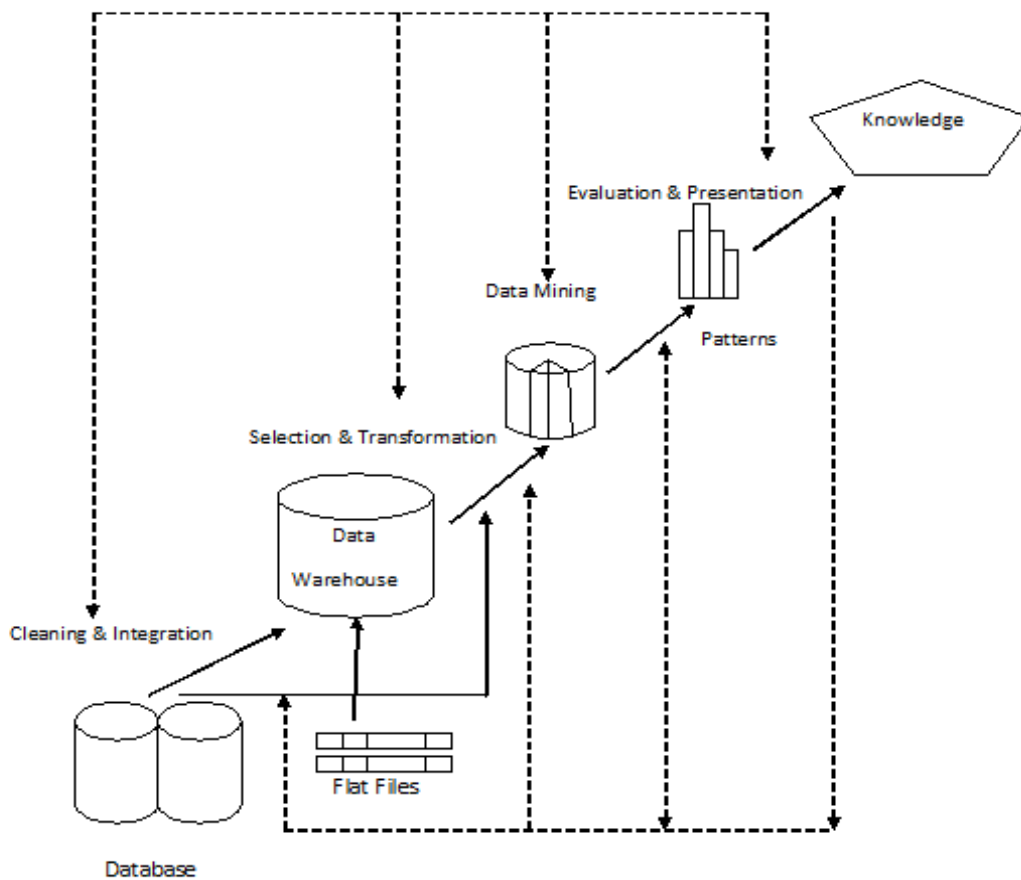


Fig 1.1.1 DM as a step in KDD

3. Data reduction: Data reduction is defined as data is reduced to smaller volumes but yet produces same results is called as data reduction data reduction contains histograms, sampling and clustering in it
4. Data transformation: Data is changed into forms that are suitable for taking out of facts is called as data transformation. Data is consolidated into forms that are usable for data mining.
5. DM: - it is very necessary procedure in which various bright methods are used to haul out unseen patterns.
6. Pattern evaluation: - Knowledge is represented by using interesting pattern depend on likeness calculation.
7. Knowledge appearance: - Present mined knowledge to consumer using mental picture system. [17]

**1.1.2 Data mining functions:-**The course of discovery a form to fit the facts is called data mining functions. Each function required to set criteria to create one model over another. There is necessitating defining a procedure to evaluate facts for each function. Its functionalities are specified below:

- 1. Characterization and Discrimination:** - Summarization of data of class underneath learning is pronounced data characterization and assessment of the aim class with one or a group of relative programme is known as data discrimination. Descriptions of Class/concept are derived using these two functionalities.
- 2. The withdrawal of Frequent Patterns, Associations and Correlations:** - The patterns that happen frequently in records are known as frequent patterns. Association rule mining is route of pronouncement appealing correlations, repeated design or relations stuck between sets of substance in operation databases, relational databases or else other information directory.
- 3. Classification & Regression:-**Classification is a DM (device learning) skill used to forecast group partisanship intended for data subsets. Regression analysis is style with the intention of is mainly second-hand for numeric calculation.

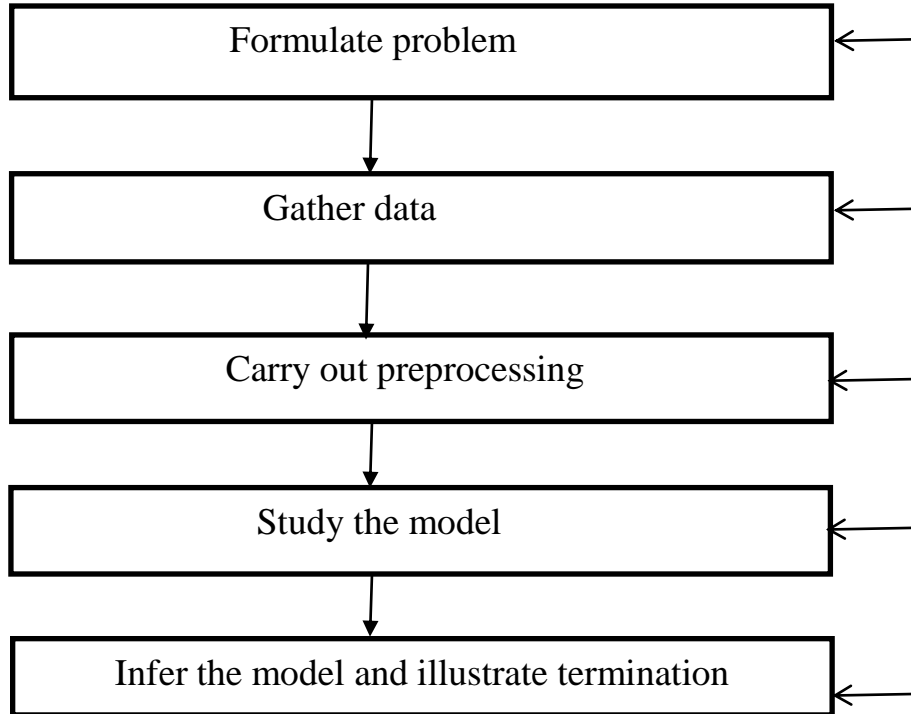
4. **Cluster Analysis:**-Cluster analysis is job of combining a set of things in such a way those things in the identical group are more related to each other than to those substances in other cluster.
5. **Outlier Analysis:**-Some objects in a data set do not fulfil with wide-ranging behaviour or replica of the data. This information stuff is outliers and examination of out of range data is pronounced outlier analysis.
6. **Evolution analysis:** It defines objects whose actions vary over times and also state model for them. It usually includes time-series data analysis, succession or periodicity pattern matching, and similarity-based data analysis.

### 1.1.3 Need of DM:-

1. Very large data amount of data are generated every day. So that DM is used dig out constructive information and knowledge as of vast quantity of facts.
2. Data that are generated have different dimensionality. To deal with these types of dimensionality DM is used.
3. Variety of data is generated every day. So to deal with heterogeneity of data DM is used.
4. To work with new technology DM is used.
5. Conventional techniques infeasible.

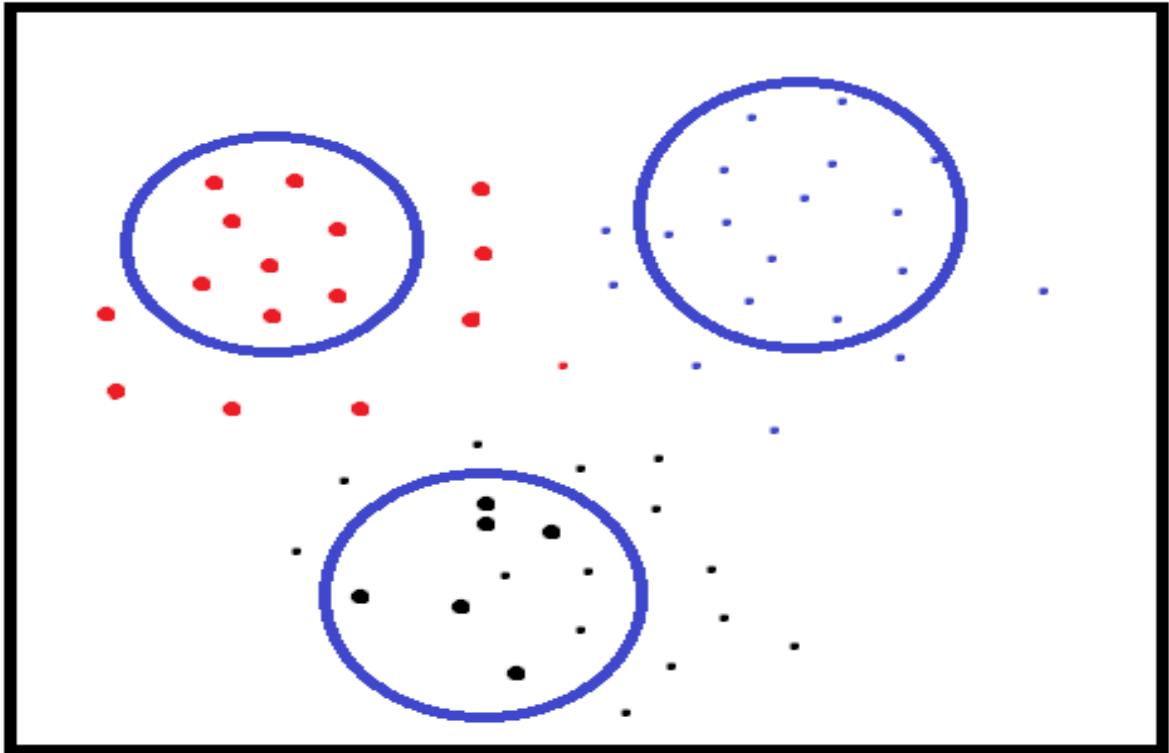
### 1.1.4 DM process: - A reparative process which include the subsequent stepladder

1. Originate or formulate problem for example:-classification or numeric calculation
2. Assemble the data relevant that relate with stated problem.
3. Perform preprocessing on data and characterize the records in the look of labels.
4. Study model or predictor.
5. Assess the model.
6. Well adjust the model as needed



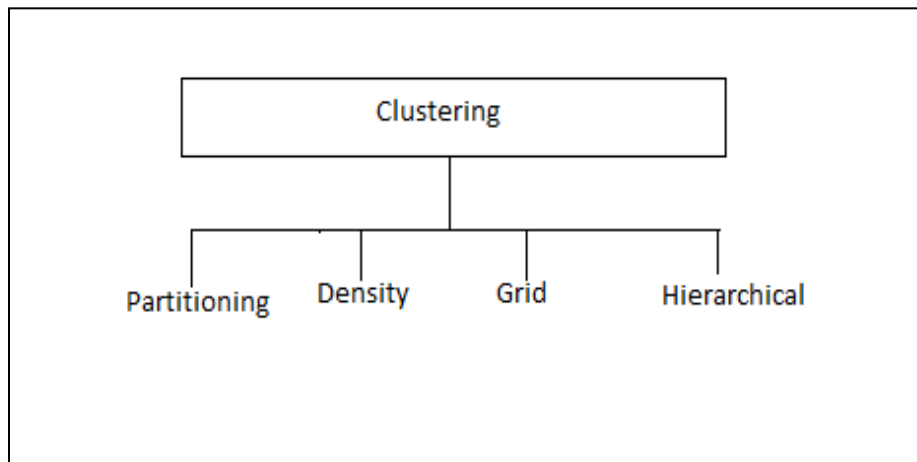
**Fig 1.1.4 DM process**

**1.2 Clustering in DM:** - Clustering is learning by observation process. It is procedure of dividing group of data into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. It is chief chore of exploratory DM and a general system for statistical data investigation. Several areas in which clustering is used [L.V.Bijuraj] like data/text mining, image processing, web mining, and voice mining. There are several methods of clustering. [15].



**Fig1.2.1 clustering in DM**

Many clustering algorithms used for clustering. Primary clustering methods can be organized into below given parts.



**Fig1.2.2 Types of clustering in DM**

1. **Partitioning Methods:**-The broad principle for partitioning samples inside a class have high resemblance between each one in the meantime sample in distinct classes are more contrasting. Most partitioning methods are distance-based. M numbers of

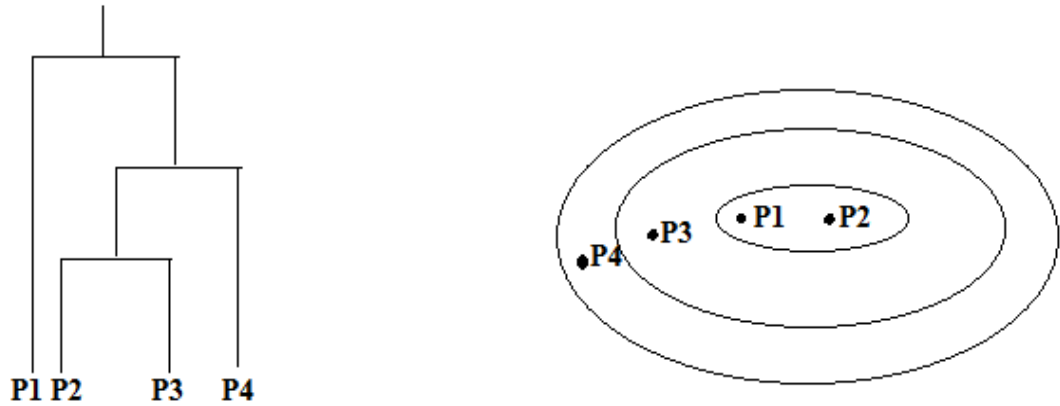
groups are created; this method create initial cluster center and then uses an iterative rearrangement skill adopt to get better partitioning by transferring items beginning single assembly to a different. In first-class partitioning stuff in equivalent huddle are nearest to each one while stuff in diverse huddle distant away from each other or different. The majority applications implement fashionable thumb method such as k-M and k-medoids algo which gradually progress bunch superiority. These clustering methods is generally used to develop spherical size clusters and used for small to medium size datasets.



**Fig. 1.2.3 Partitioning Clustering**

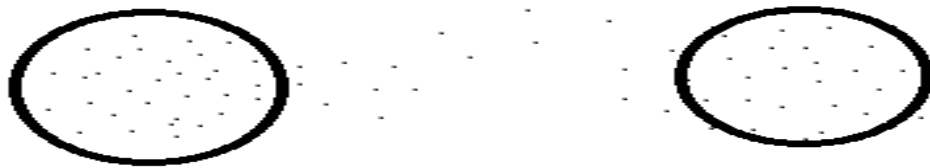
- 2. Hierarchical Methods:** - It produces a hierarchical breakdown of specified dataset of data objects are created. Hierarchical decay is characterized by a ranking arrangement like tree structures. It does not require clusters as inputs. In this type of clustering it is possible to view partitions at different level of granularities using different types of K.





**Fig. 1.2.4 Hierarchical Clustering**

3. **Density based:** - It is constructed on base of compactness. Fundamental thought is to carry on the rising known collect on condition that the density in the region cross a few porches that is for every data tip within a specified group, radius of certain group has to hold at least a smallest amount of tip. It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also.



**Fig. 1.2.5 Density based clustering**

4. **Grid Based Methods:** - Objects are collectively grouped together to form network. Object breathing space is determined into bound quantity of cells that build a grid configuration. It assigns to object grids cells and compute density of each cell. After that eliminate whose density is below threshold value. Now form cluster according to group of dense clusters. In this no distance computations so it is fast process. In this it is also easy to determine which cluster is neighboring. Here structure bound to the unification. Complexity of the clustering is depends on grouping of the cells.

### 1.2.1 Clustering have many requirements some of them are as listed below:

- **Scalability:** There is a need of greatly flexible clustering procedure to agreement with huge catalogs.
- **Compatible with diverse kind of attributes:** Every types of thoughtful raw facts like intermission based, definite, binary object algorithms should be accomplished to be applied.
- **Finding of clusters with attribute shape:** Here would not be constrained to merely detachment measures that lean near toward discovery sphere-shaped cluster that are of minor size. Clustering algorithm must be expert to notice cluster of random shape.
- **Great dimensionality:** Clustering algorithm must not just be capable of handling little- dimensional object and also great dimensional area.
- **Capacity to contract with outlier object–** Databases includes noisy, misplaced or inaccurate facts. Many techniques stay delicate towards such type of facts and might indication to reduced valuable groups.
- **Interpretability –** Its outcomes must be interpretable, understandable plus practical.

### 1.2.2 Application of clustering

1. It is commonly used in DM. It helps to analysis similar data from large amount of data. It puts similar data items into one group or cluster. It used by different company to analysis customer buying behavior.
2. It is used in text mining. It helps to analysis similar text. It finds out valuable information from huge amount of raw facts.
3. Search engine is also used clustering to classify data. When user query on search engine. Then search engine retrieve relevant information to user with the help of clustering.
4. It used in pattern recognition.
5. It used in image analysis.
6. Bioinformatics also used clustering.
7. It used in image processing

8. It used in whether report analysis

**1.3 K-mean clustering algorithm:** - KMCA is most familiar and easiest algorithms. It is unverified learning algorithm that is used to tackle with sound recognized clustering troubles. Course of action followed by it very straightforward and trouble-free manner to sort agreed data points. It is most commonly used algorithm. It is very older algorithm. Many research works going on it. Various enhancements are performed on it to remove its shortcomings. KMCA can apply on large datasets. It provides very easy and simple way to cluster data items. But it takes large amount time to cluster. Its processing speed is very low. It takes number of iteration to cluster data. [22]

- i. K-mean clustering algorithm has some properties that are specified below:
  1. There should be always k cluster.
  2. Each cluster always contains at least one item.
  3. Non-hierarchical clusters are formed and they do not overlie.

**Algorithm:** KMA is one of partition algorithm. Mean values of data is used to represent cluster centre.

**I/P:**

K: represent no of clusters,

D: specify a data set contain n items.

**O/P:**

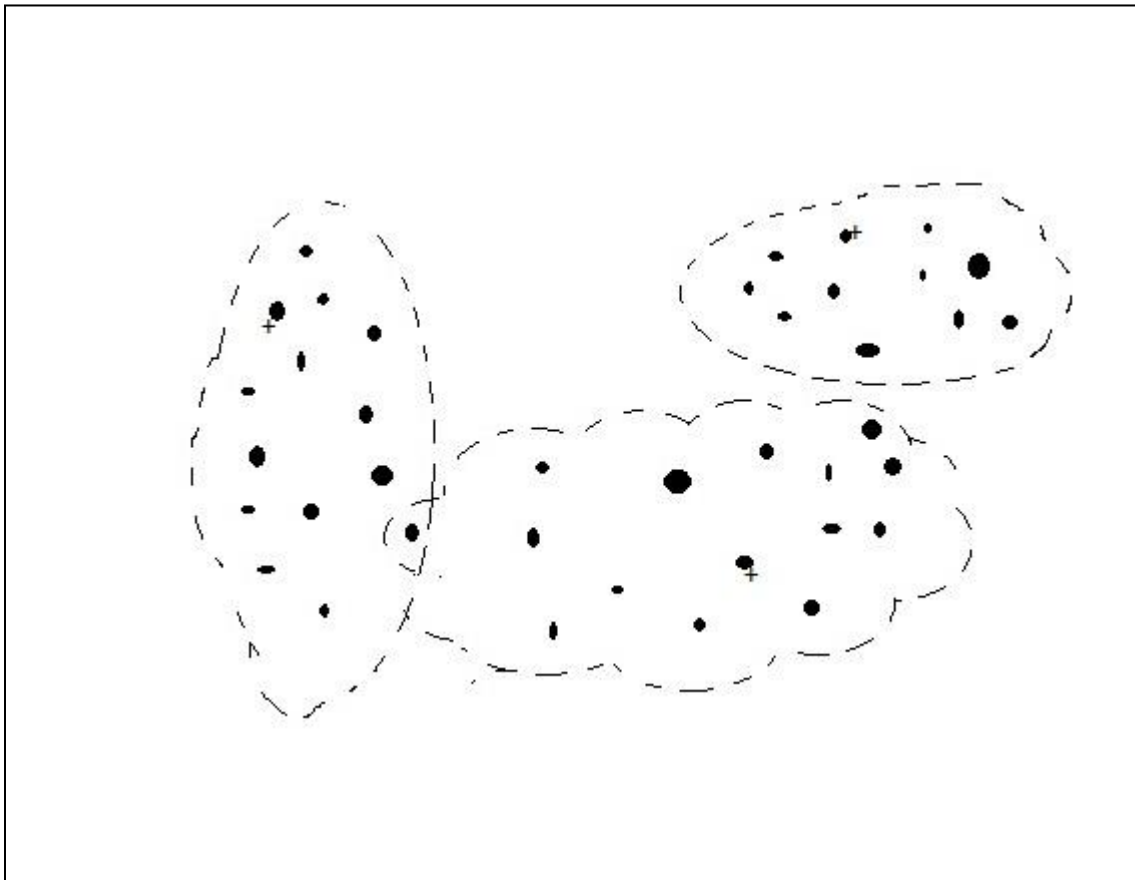
K clusters are generated.

**Technique:**

1. Randomly decide k items or object from D as the preliminary cluster centres;
2. Repeat
3. (Re) assign every object to C to which object is closer, depend on the mean value of objects in C;
4. Update C means, i.e., estimate mean value of objects for every C
5. until no change;

KMCA is most familiar and easiest algorithms. It is unverified learning algorithm that is used to tackle with sound recognized clustering troubles. Course of action followed by it very

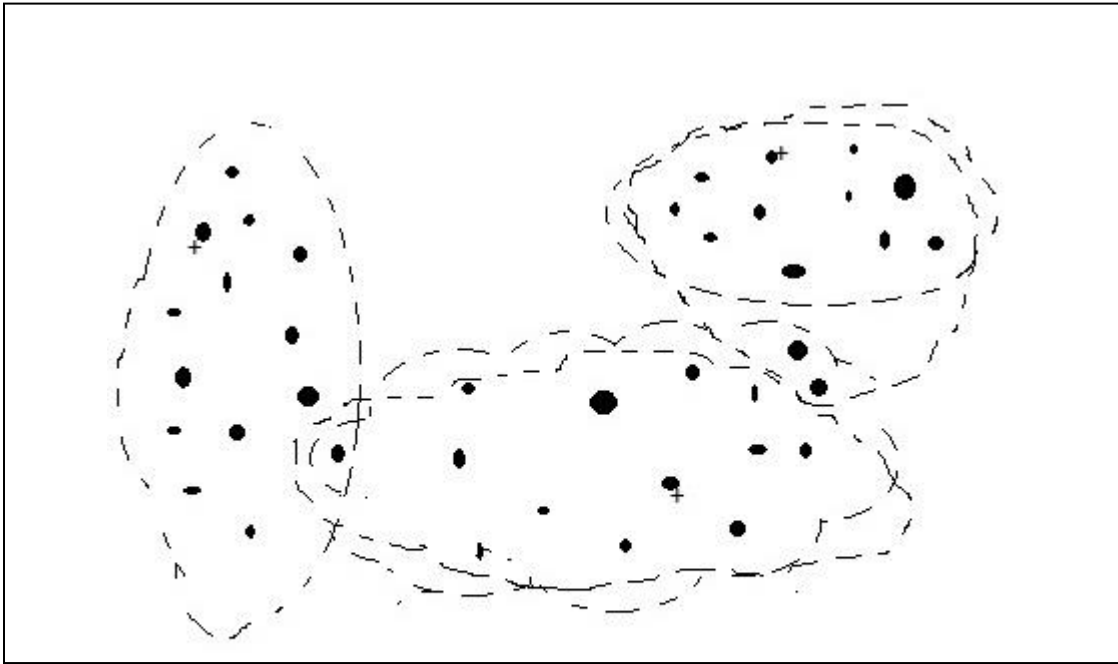
straightforward and trouble-free manner to sort agreed data points. It is most commonly used algorithm. It is very older algorithm. Many research works going on it. Various enhancements are performed on it to remove its shortcomings. KMCA can apply on large datasets. It provides very easy and simple way to cluster data items. But it takes large amount time to cluster. Its processing speed is very low. [17]



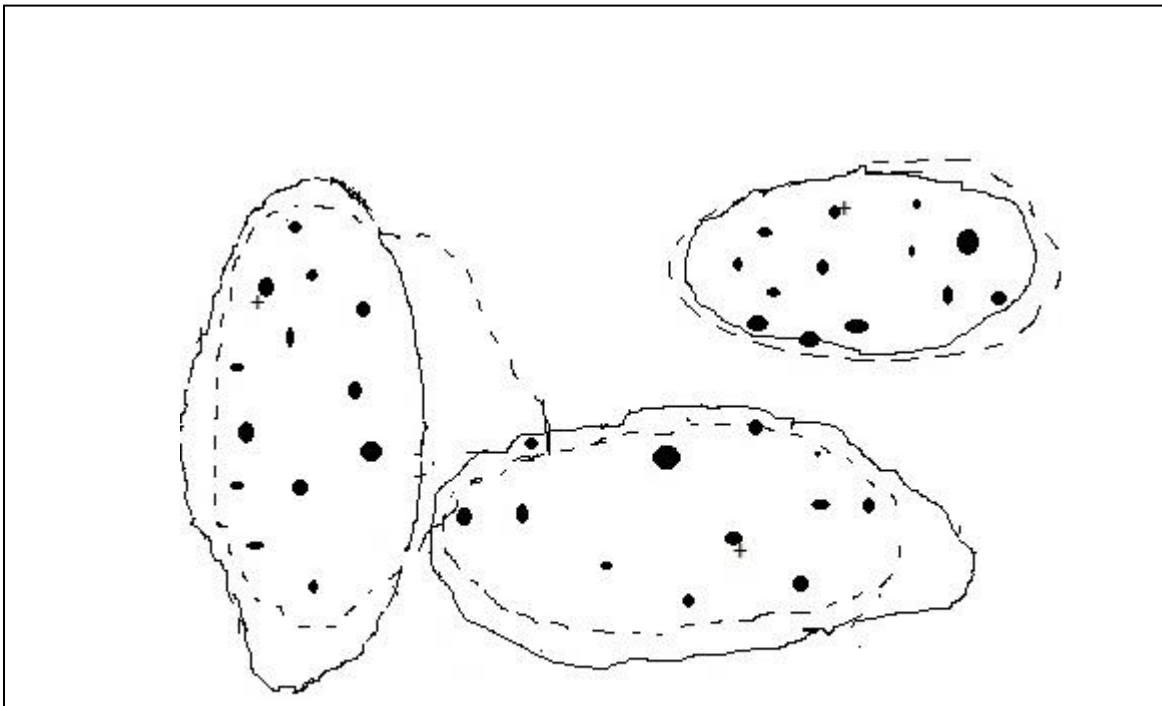
**Fig1.3.1 initial clustering**

Initially KMCA randomly choose number of clusters ( $k$ ) of data item that contain in data set  $D$ , these present group centriod. Then other data item puts in to that bunch to which those are closer. Distance between data items and group mean calculated by using Euclidean formula. KMCA is most familiar and easiest algorithms. It is unverified learning algorithm that is used to tackle with sound recognized clustering troubles. Course of action followed by it very straightforward and trouble-free manner to sort agreed data points. It is most commonly used algorithm. KMCA can apply on large datasets. It provides very easy and

simple way to cluster data items. But it takes large amount time to cluster. Its processing speed is very low.



**Fig 1.3.2 Iterate**



**Fig 1.3.3 Final clustering**

After first steps numbers of iterations are performance. Mean of cluster is restructured in each iteration. Then again distance between data items and cluster's mean is computed. The data item which have less and equal distance from cluster's mean stay in same cluster, other points that have more distance from existing cluster's mean move to cluster to which those are closer. All steps are repeatedly performance until no data point further move. At last we get final outcome of KMCA. [17]

Selection of initial division can significantly affect final k that formed.

It is widely used in many applications, but it has following drawbacks.

### 1.3.1 Drawbacks of KMCA

1. Number of clusters that are generated needed to specify in advance. But it is not true and possible in real-time applications.
2. It is reparative practice; the KMA is mainly responsive to origin centres choice.
3. The KMA may meet to difficulty of local minima.

There is efficiency problem in KMCA. Like it take large amount of time to perform computation and constructed clusters also does not have good

### 1.3.2 Various Enhancements in K-Mean algorithm: There are various types of enhancements which have done till now. These enhancements are as follow:

1. **First enhancement in terms of reduce number of iteration and Time complexity:**  
Paper new development in KMCA. This algorithm take input same as original KMCA. The entire procedure is separated in two steps. In step 1, group amount is fixed and starting clusters are produce with the help of outcome of first step. Sub array are formed by dividing elements of array and symbolize primary groups. In step two, taking dynamic size group and last (final) clusters are produce with the help of outcome of second step. Space from datasets help to produce initial cluster centre. Data which has less distance or equivalent distance from group can stay in same cluster but other element put in another cluster from which they have less distance. Algorithm gives final result when there no movement of element is possible.  
In Basic KMCA, clusters remain changing because initial cluster select based on input facts item. So every time when algorithm run number of iteration

and processing speed varied for similar facts. In proposed KMCA some computation are made to produce initial cluster centroid, so due to this reason number of iteration remain still and processing time also reduced. It proved better performance than basic KMCA. [6]

2. **Second enhancement in terms of improved initial centre:** K-mean one of the simplest algorithm to cluster large datasets. In this similar data items are placed into one cluster. But it has many problems like choosing initial cluster centroid. Its accuracy totally depend upon selection of initial centroid. Paper propose new approach of k-mean ,instead of randomly selecting initial cluster centroid some formulas are used to compute initial centroids. This approach help to improve processing speed of KMCA and also help to reduce number of iterations. It does not take any more input like threshold value. But this algorithm still contain problem because there is need to define value of k. Paper defines that KMCA is the most familiar algorithm. But it has many problems like choosing initial cluster centroid. Its accuracy totally depend upon selection of initial centroid. Basic KMCA traps in the problem of local minimum. Paper provides new method to find initial centroid and then defines accurate way to put data items into cluster to which it more closes. It helps to reduce time complexity and provide more accuracy. [38]
3. **Third enhancement in terms of improve accuracy and efficiency:** Paper defines that KMCA is most widely used algorithm. It used to cluster large datasets. It is very easy way to classify data items. But it has some limitation such as it take more time to cluster data items, cluster quality is not so good and its accuracy highly depend upon on selection of initial cluster centre. It provides different result with different initial centroid. Paper propose new algorithm to reduce limitation of KMCA. Enhanced algorithm defines systematic method to choose initial centroid and also defines accurate way to put data items into number of cluster. But enhanced algorithm still have problem such as this paper does not defines value of k. There is still need to define number of cluster in advance. [23]

In this work a new hybrid algorithm will be implemented by using its various enhancements. Remaining of work is planned in subsequence. In chapter two, previous effort done in the k-means clustering is overviewed and comparative study of various enhancements in k-means

clustering algorithm is done. In chapter 3, problem formulation, objectives, study of methodology used and basic design of the proposed work is done are discussed. In chapter 4 software implementation and results are defined. Then conclusion and future scope of proposed work is present. At last of this report various references are provided for detail study of the work.



## **CHAPTER-2**

### **REVIEW OF LITERATURE**

---

**Yugal Kumar and G. Sahoo, (2014) “A new initialization method to originate initial cluster center for k-means algorithm”:-** Paper define KMA is one of the fashionable partition algo. It is broadly second-hand to cluster items. KMA is trouble-free, straightforward and competent. But it also has a quantity of disadvantages such as origin cluster difficulty. They introduce new method to deal with primary cluster crisis in KMCA based on binary search technique. It is one of most well-liked penetrating way. This takes an entity in prearranged catalog of collection. Binary search property is used to define initial cluster center and then KMCA is practical to obtain most favorable representative of cluster in data sets. Planned scheme is practical on minkowaski subjective KMCA to verify its importance and usefulness. [38]

**Dibya Jyoti Bora and Anil Kumar Gupta, (2014) “A New Approach towards Clustering based Color Image Segmentation”:-** Paper proposes new approach in which watershed algorithm is merged with KMCA. But in this they used cosine distance instead of Euclidean distance. It applies on image segmentation. It clusters pixels of image on the bases of color. The final segment image is filtered by median filter, to remove noise from segmentation process and to provide clarity to final output image. [10]

**C.Deepika, R.Rangaraj, (2014) “An Efficient Uncertain Data Point Clustering based on Probability-Maximization Algorithm”:-** Handling uncertain data is much more difficult. Uncertain data means that data which has no certainty in them. Data are uncertain which has no proper location. Means moving objects like persons and living things like animals. Probability distributions are used to define uncertain data objects. Many circumstances are there that are used to define clustering of inexact data items. Basic examples of inexact data objects are as marketing research and one more example is weather position monitor weather circumstances. Accordingly some distributions we require to cluster the undefined items. Data uncertainty is represented with the help of probability theory. Probability density functions are there to represent an object. In this portioning method and density based

clustering is used to clustering the uncertain data. Portioning method contains k-mean clustering and density based clustering contain DBSCAN that are used in this paper to handle uncertain data or say clustering of uncertain data. In this purposed algorithm data sets are firstly preprocessed and in preprocessing step basic components are sizes, classes, attributes, standard deviations. After preprocessing step probability maximization algorithm is performed. In PM algorithm two partitions are there one is true partition and second is clustering results. [4]

**Pritesh Vora, Bhavesh Oza, (2013) “A Survey on K-mean Clustering and Particle Swarm Optimization”**:- Paper introduces that clustering is commonly used in many field like in DM, image segmentation, pattern recognition etc. KMCA is one of clustering algorithm, which commonly used to cluster large dataset and it is very easy and simple way to classify datasets. Paper defines particle swarm optimization which is also used in many applications to get optimized result. They give comparison between KMCA and swarm optimization. [31]

**Daljit Kaur and Kiran Jyot, (2013) “Enhancement in the Performance of K-means Algorithm”**:- Paper introduces KMCA is most commonly used algorithm to cluster large datasets. It used in many applications like image segmentation, data mining, data compression. It is very easy and simple way to classify the data items. But it has many problems like choosing initial cluster centroid. It accuracy totally depend upon selection of initial centriod. Paper defines new method to improve accuracy and efficiency of KMCA. It helps to reduce complexity of numerical calculation. It also reduce problem of dull units. [11]

**Chou Chien-Hsing and Chu Yung-Long “Extracting and Labeling the Objects from an Image by Using the Fuzzy Clustering Algorithm and a New Cluster Validity”, 2013**:- Paper proposed non metric distance measure based on live symmetry is used to measure cluster soundness. For this thresholding technique is applied first to extracts item from innovative representation, data patterns are designed by moving object pixels. Object pixels are labeled by applying the fuzzy clustering algorithm and number of objects are determining by applying proposed validity measure. To define performance of proposed measure simulation results are used. [7]

**Aastha joshi, Rajneet kaur, (2013) “A review of various clustering technique in data mining”:-** Paper describe comparison between various clustering techniques like PM, HM, DBSCAN, GM. CA are mainly provided to manage facts, categorized facts for facts reduction, model creation and also used for outlier discovery etc. Main motive each clustering technique is to find cluster center that represent each cluster. Then input data is compared with each cluster center, and then based on these cluster centers defined which cluster is nearest or similar one. Partitioning method like k-mean clustering algorithm is used for large datasets, as number clusters is increased its performance is also increased. But its use is limited to numeric values. Hierarchical algorithms are used for categorical data. DBSCAN is adopted to find cluster of arbitrary shapes. [1]

**Amar Singh and Navjot Kaur, (2013) “To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm”:-** Paper provides new method to improve accuracy and performance k-mean clustering that is a ranking based method. Analysis done on previous k-mean clustering method which fit in threshold value. Ranking method which is weighted page ranking applied on KMCA. In this in links and out links are used to compare performance in the form of execution time of clustering. Weighted page rank algorithm with k-mean provides better result than existing k-mean algorithm. It takes less computational time then existing k-mean algorithm. [2]

**Zhaohong, Kup-SZE, Choi Deng, (2013) “A survey on soft subspace clustering”:-** Subspace clustering is a very good clustering technology to identify clusters based on their associations with subspace in high dimensional space clustering can be classified into 2 different groups. Hard subspace clustering (HSC) and soft sub space clustering (SSC). HSC is been extensively studied by scientific community. SSC are relatively new but more attention on them due to better adaptability. In this paper comprehensive survey on existing SSC algorithm and recent developments use presented over here. SSC has mainly three type’s conventional subspace clustering, independent SSC, Extended SSC.As discussed three main categories of SSC are as listed below:

**Conventional SSC:-**Classical feature weighting clustering algorithms with all the clusters sharing the same subspace and a common weight.

**Independent SSC:** - Multiple feature weighting clustering algorithms with all the clusters having their own weight vectors, i.e., each cluster has an independent subspace, and the weight vectors are controllable by different mechanisms

**Extended SSC:**-Algorithms extending the CSSC or ISSC algorithms with new clustering mechanisms for performance enhancement

Comprehensive survey of SSC is presented over here. These are systematically categorized into three categories, XSSC, ISSC, CSSC there are explained along with subcategories are explained in detailed .we see that transfer learning and multi-view learning will play an important role in development of SSC in future. A thorough understanding of SSC algorithm and insight into the advancement of SSC can be obtained through this survey. [41]

**ReshmaMR, Suchismitasahoo, (2013) “Management uncertainty and clustering in uncertain data based on KL divergence technique”:-**Many problems in data are comes due to uncertainty of data. From those problems clustering is one of them. Due to this uncertainty of data there are problems in clustering. Previously known methods i.e. portioning method and density based method adopted to handle uncertain data and cluster that uncertain data into a single cluster that is the data-items in one group are alike to each other. Portioning method plus density based methods are that which use or say based on geometric distance between objects. In purposed algorithm probability distributions are used which are the mandatory features of uncertain objects. In purposed algorithm Kullback-Liebler divergence is used. This algorithm is basically for measuring the resemblance among objects and integrates portioning plus thickness based method to group in exact objects. FCM method is used in this to cluster uncertain data objects and show effectiveness of the data objects. FCM means fuzzy c-mean clustering for data with tolerance. Data objects are represented by probability distributions. And probability distributions are described by probability mass function and objects that are in continuous distributions are described by probability density function. In this paper basically KL-divergence is integrated with portioning and density based clustering to handle uncertain data and clustering of uncertain data. [14]

**Raed T. Aldahdooh, Wesam Ashour, (2013) “Distance-based Initialization Method for K-means Clustering Algorithm”:-** Clustering is learning by observation process. It is procedure of dividing group of fact into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Clustering contains many methods to cluster datasets. Partitioning method is one of them. It clusters data into heterogeneous groups. It provide homogenous data within cluster and heterogeneous between clusters. KMCA is one of the partition algorithms. KMCA is the old algorithm. It provides very easy and simple way to classify data items. It outcome totally depend upon selection of initial cluster centriod. KMCA is suffering with many limitations. Paper proposes new method for choosing initial cluster centriod. It firstly selects cluster centre points which are far away from each other and which lie in different clusters. This step reduces number of computations. It gives optimum performance by decreasing KMCA functions. It provides better performance. It can apply on artificial and real data items. [32]

**Dr. M.P.S Bhatia and Deepika Khurana, (2013) “Experimental study of Data clustering using k-Means and modified algorithms”:-** Paper presents KMCA is one of the partition algorithms. It provides very easy and simple way to classify data items. Clustering is learning by observation process. It is procedure of dividing group of fact into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Clustering contains many methods to cluster datasets. KMCA is the older algorithm. Many research works going on it. Various enhancements of KMCA are collected. Then these algorithms are applied on different datasets and compare with basic KMCA. All work performs in MATLAB. Final outcome evaluated based on number of iterations, number of data items that are not classified properly. These algorithms are applied on iris datasets, wine datasets. These algorithms provide better performance. They are not sensitive initial cluster centriod selection and take less processing time to cluster large datasets. [26]

**Chunfei Zhang and Zhiyi Fang, (2013) “An Improved K-means Clustering Algorithm”:-** Paper presents K-mean one of the simplest algorithm to cluster large datasets. In this similar data items are placed into one cluster. But it has many problems like choosing initial cluster centroid. It accuracy totally depend upon selection of initial centriod. Other problem is that there is need to define number of clusters in advance. KMCA takes number of cluster mean value of k as input. It randomly selects initial cluster centriod. Paper

proposed enhanced KMCA which gives good performance than basic KMCA. It provides more stable clusters and it removes noise from datasets which help to provide more efficient and accurate result. [9]

**Ms.Chinki Chandhok et.al, (2012) “An Approach to Image Segmentation uses K-means Clustering Algorithm”:-** Paper proposes new method for image segmentation. It uses KMCA in image segmentation which provides color-based segmentation. It divides images into number of clusters. It defines same regions with respect to color and texture. Color and spatial features of image are used to group pixels into clusters. After completion of this process, then regions are formed by merging clusters. This approach helps to improve segmentation quality and processing speed. KMCA is most commonly used algorithm to cluster large datasets. It used in many applications like image segmentation, data mining, data compression. It is very easy and simple way to classify the data items. [8]

**Anwiti Jain, Anand Rajavat, (2012) “Design, analysis and implementation of modified k-mean algorithm for large data-sets to increase scalability and efficiency”:-** In this paper, they proposed MKMC algorithm to assemble huge datasets. Main proposes of MKMCA to get cluster representation to which other objects are much closet. It performs in every reparative step. MKMCA reduces problem error criterion of cluster and remove problem of trap in restricted minima. They compare MKMCA with KMCA, K-medoid on the basic of available reports on identical device and weigh against in alike organization surroundings. Results show that MKMCA take fewer instant to perform than existing KMCA and K medoid algorithm for small number of records as well as for large number of records. As compare to other algorithm MKMCA is stronger to missing values and out of range for the reason that it reduces a sum of common pair off differently without a sum of squared Euclidean distance. [3]

**Z.Volkovich, D.Toledano-Kitai, G.W.Weber, (2012) “Self-learning k-means clustering: a global optimization approach”:-** Paper proposed distance metric learning to represents data pattern much better in contrast with most usually used methods. Distance and KMA are used to add in object rescale and grouping, so that with sequential clustering data can iteratively produce in increased space. A global optimization problem is overcome with the purpose of reduce cluster distortions depending on cluster design. The true number of

clusters can be obtained based on weight metric which is used as a cluster validation feature. These algorithms are applied on different datasets like Gaussian synthetic dataset, iris dataset, picture dataset. [40]

**F.U Siddiqui and N.A Mat Isa, (2012) “Optimized k-means clustering algorithm for image segmentation”**:- Paper introduces OKMCA that segmented image homogeneously into regions of attention. It avoids the problem of dead center and rapt center at local minima phenomenon. Alteration is done on hard membership concepts. In this pixel is assigned to its nearest cluster, if cluster has same distance to two or closest cluster centers, the pixel assigned to cluster with void value or with lower strength value. The effectiveness and toughness of OKM clustering algorithm is determined by applying qualitative and quantitative analyses. In this approach pixel which have similar distance to two or more adjacent cluster is initially assigned to dead center. In additional iteration it is assigned to cluster to cluster which has lower cluster discrepancy, if no dead cluster is found. It provides outstanding consistency in its performance. So it can be used in different electronic product as an image segmentation tool. [13]

**Azhar Rauf et.al, (2012) “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”**:- K-mean one of the simplest algorithm to cluster large datasets. In this similar data items are placed into one cluster. But it has many problems like choosing initial cluster centroid. Its accuracy totally depend upon selection of initial centroid. Paper propose new approach of k-mean, instead of randomly selecting initial cluster centroid some formulas are used to compute initial centroids. This approach help to improve processing speed of KMCA and also help to reduce number of iterations. [6]

**Neha Aggarwal, Kirti Aggarwal, Kanika gupta, (2012) “Comparative Analysis of k-means and enhanced K-means clustering algorithm for data mining”**:- Paper introduces new method of k-mean which help to overcome problem of basic KMCA like more processing complexity and its accuracy highly depend upon selection of initial cluster centroid. It takes more time to cluster large dataset. In this paper both algorithms enhanced and basic KMCA applied on same datasets. Then performances of these algorithms check which show that enhanced algorithm more efficient and accurate than basic KMCA. [26]

**Shital A. Raut and S. R. Sathe, (2011) “A Modified Fast map K-Means Clustering Algorithm for Large Scale Gene Expression Datasets”:-** Paper introduce k-mean clustering algorithm that is used to extracts patterns from gene appearance datasets. Datasets are taken in terms of dimensions. So it increases CPU time and memory requirements in case proportions are increased, then try to increases process of KMA by providing extra step. This can be performed before the implementation of KMCA on the datasets. Fast map KMCA is two phase algo which is used to reduce CPU time and memory requirement. Gene is essential units of all the organisms. K-mean clustering algorithm is used to find out genetic information, to identify group of co related genes. Fast map algorithm is implemented in first phase to reduce dimensions and traditional k-mean clustering algorithm is introduced in second phase. It is implemented by using technique MATLAB7.0. [33]

**T. Velmurugan and T. Santhanam, (2010) “Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points”:-** In this research paper, KMA and k-medoid algorithm are analyzed and examined based upon distance between a choice of input data points. Groups are generated on basic of distance between data-item and cluster center is finding for each cluster. Bitterness of each algorithm is defined based on their performance. By using normal distribution and uniform distribution participation data are generated. Algorithms are performed in JAVA language and performance is compared for each execution. The effectiveness and accurateness of each algorithm is evaluated during implementation of program on contribution data point. In both cases usual time taken by KMCA is superior to k-medoid algorithm. Benefit of KMCA is its positive completing time. But it also has drawback like consumer has to be familiar with before how many clusters we required. It experiential that KMCA is capable f minor data sets and k-medoid algorithm is healthier for huge data sets. [35]

**Madhu Yedla et.al, (2010) “Enhancing K-means Clustering Algorithm with Improved Initial Center”:-** Paper defines that KMCA is the most familiar algorithm. But it has many problems like choosing initial cluster centroid. It accuracy totally depend upon selection of initial centriod. Basic KMCA traps in the problem of local minimum. Paper provides mew



method to find initial centroid and then defines accurate way to put data items into cluster to which it more closes. It helps to reduce time complexity and provide more accuracy. [23]

**Oyelade, O. J, (2010) “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance”**:-Paper introduce that it is critical issue to analysis growth of student in academic field. So they use clustering to analysis student performance and use different algorithm to give ranking to student based upon their performance. Clustering is learning by observation process. It is procedure of dividing group of data into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Paper use KMCA to observe student performance. This algorithm merged with deterministic model. Then this algorithm applies in Nigeria. It helps to make good decision by ranking student result. It accurately shows student progress in academic. [30]

**K. A. Abdul Nazeer, M. P. Sebastian, (2009) “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm”**:- Paper defines KMCA is most widely used algorithm. It used to cluster large datasets. It is very easy way to classify data items. But it has some limitation such as it take more time to cluster data items, cluster quality is not so good and its accuracy highly depend upon on selection of initial cluster centre. It provides different result with different initial centroid. Paper propose new algorithm to reduce limitation of KMCA. Enhanced algorithm defines systematic method to choose initial centroid and also defines accurate way to put data items into number of cluster. But enhanced algorithm still have problem such as this paper does not defines value of k. There is still need to define number of cluster in advance. [19]

**Malay K. Pakhira, (2009) “A Modified k-means Algorithm to Avoid Empty Clusters”**:- KMCA is the most commonly used algorithm to cluster large datasets. It used in many areas science and technology. KMCA suffers with many limitations one of them is that it generate dead clusters based on selection of initial centroid of cluster. This problem can be ignoring in case when KMCA perform statistical. Dead cluster problem can be reducing by running algorithm multiple times. But it create serious problem when KMCA used in other high level application. It produces empty clusters. It reduces performance of KMCA. Paper presents

enhanced KMCA that deal with this problem and removes dead cluster limitation. This algorithm provides better performance and it removes empty clusters. [25]

**Joaquín Pérez Ortega<sup>1</sup>, Ma. Del Rocío Boone Rojas, (2009) “Research issues on K-means Algorithm: An Experimental Trial Using Matlab”:-** Clustering is learning by observation process. It is procedure of dividing group of fact into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Clustering contains many methods to cluster datasets. Partitioning method is one of them. It clusters data into heterogeneous groups. It provide homogenous data within cluster and heterogeneous between clusters. KMCA is one of the partition algorithms. KMCA is the old algorithm. It provides very easy and simple way to classify data items. Clustering is used in many applications for e.g. information detection, data looseness; blue print recognition and blueprint categorization. This paper introduces various enhancements in KMCA, which are done to remove limitation of KMCA. All work performed in MATLAB. It takes datasets from UCI repository. It applies experiment on iris datasets. [16]

**Tina Eliassi-Rad Terence Critchlow, (2007) “Multivariate Clustering of Large-Scale Scientific Simulation Data”:-** In this paper simulation is explained composite technical procedure that is having the running of hugely parallel computer programs. These programs produce very huge amount of information on the daily bases. And these information or set of items are over the spatio temporal space. And produced this type of data has main step like modeling of such data it help the scientists to discover new knowledge from the computer simulation. In this algorithm cosine similarity is there. This would allow the reducing modeling time from  $O(n^2)$  to  $O(n * g(f(u)))$ . In this  $n$  is number of data points ( $u$ ) is the function of user defined clustering threshold. In this paper mesh format is used. In mesh format multivariate clustering is performed. Mesh data usually varies with time. And data is varies according to time and has multiple dimensions. And has many interconnected grid in them. In multivariate clustering algorithm inputs are list of Jones and outputs are list of clusters over here. And joneses used are of red and green .red that are already included in clusters and green that’s available for clustering. When a green jone is putted into a cluster then it would become red. This algorithm utilizes the cosine similarity measure to cluster

field variable in data sets. Cosine similarity measure has geometrical properties and we take advantage of these geometrical properties. This will reduce the modeling time. Spatial cluster play not much role in building a cluster. Linking algorithm is used in this paper for defining a proper location of cluster where it is actually located. Linking algorithm will connect the each cluster with appropriate node of data set topology tree. [18]

**Yi Ma, (2007) “Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression”:-** In this paper they introduced about grounded on thoughts from lossy data coding and solidity, segmentation of the data is produced here in this work. Much more amount of data is produced and data get up from many useful difficulties in many fields which are image processing, pattern recognition, computing vision are some of the multimodal multivariate distributions. Data segmentation is very significant phase in demonstrating data investigating data understanding and condensing such data. While doing this some questions arises in our mind that is data segmentation is effective or not? What would be the outcomes of data segmentation is there is gain or losses of data segmentation? Segmentation is performed in probabilistic distributions. Lossy data means when we perform compression on any data, there would be some changes in data or say data may be lost data would not remain same or lossless. In this paper data segmentation and estimation of model is described. It contains K-mean algorithm and expectation maximization algorithm. Whole algorithms are applied over probabilistic distributions. Model estimation and data segmentation decoupling is performed firstly. In very first model estimation is performed and then after model estimation data is divided into smaller parts or say individual components. Segmentation is performed through data compression, and coding length minimization is performed. Instead of using model based top-down approach to segment data in smaller segments work lead to new approach i.e. to obtain optimal segmentation data driven bottom-up approach is used. Greedy algorithm is used to lossy data compression. [40]

**Ming Chan hang et.al, (2005) “An Efficient k-Means Clustering Algorithm Using Simple partitioning”:-** This paper introduced that k-mean mainly used to cluster large datasets. It divides data items into number of groups. They provide new approach to cluster data using k-mean clustering algorithm. Firstly takes datasets, then divides datasets into blocks, which known as unit blocks. Each unit presents at least one pattern. By applying

small number of calculations, they able find the centroid of UB. CUB presents compressed datasets. Further this compressed datasets is used to calculate final centriod of original datasets. Each datasets calculate its distance from each UB. Then assign each dataset to cluster in which it more closes. This approach helps to improve performance of k-mean. [5]

**D T Pham, S Dimov, and C D Nguyen, (2004) “Selection of K in K-means clustering”:-** K-mean one of the simplest algorithm to cluster large datasets. In this similar data items are placed into one cluster. But it has many problems like choosing initial cluster centroid. It accuracy totally depend upon selection of initial centriod. Other problem is that there is need to define number of clusters in advance. KMCA takes number of cluster mean value of k as input. This paper firstly analysis previous algorithms for choosing k (number of clusters). Then defines factors that put impact on selection of value of k. further new technique is defined based on this analysis. It considers different value of k (number of clusters) in case different result gets with different requirement. It applies on large datasets not on complex datasets. Because it computationally expensive. [12]

**Tapas Kanungo and Nathan S. Netanyahu, (2002) “An Efficient k-Means Clustering Algorithm: Analysis and Implementation”:-** This paper provide new approach it use Lloyd’s algorithm with KMCA. It is heuristics approach for KMCA which is known as filtering algorithm. It is very simple algorithm and easy to use. It needs kd-tree. It defines two ways to check efficiency of filtering algorithm. In first way, they analysis aglo’s processing speed by using data-sensitive analysis process. It gives processing speed of algorithm is high then basic KMCA. Secondly they apply experiential studies on synthetically datasets and on real data items. It is mainly used in image segmentation, data compression and colour quantization. [36]

**Yiu-Ming Cheung, (2002) “k\_-Means: A new generalized k-means clustering algorithm”:-** Paper propose enhancement in KMCA. Clustering is learning by observation process. It is procedure to dividing group of facts into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Clustering contains many methods to cluster datasets. Partitioning method is one of them. It clusters data into heterogeneous groups. It provide homogenous data within cluster and heterogeneous between clusters. KMCA is one of the

partition algorithms. KMCA is the old algorithm. It provides very easy and simple way to classify data items. Paper propose enhancement in KMCA. It can apply on ellipse produced datasets. It does not contain empty cluster problem. It also provides good quality clusters. There no need to give value of k as input. This algorithm provides excellent performance, which observes through different experiments. It can also apply on ball produced cluster. [39]

**Kiri Wagsta and Claire Cardie, (2001) “Constrained K-means clustering with Background Knowledge”:-** Clustering is learning by observation process. It is procedure to dividing group of fact into a set of significant sub-classes, define as cluster. Data is organized into clusters such that there is more in-cluster similarity and little out-clusters similarity. Clustering contains many methods to cluster datasets. Partitioning method is one of them. It clusters data into heterogeneous groups. It provide homogenous data within cluster and heterogeneous between clusters. KMCA is one of the partition algorithms. KMCA is the old algorithm. It provides very easy and simple way to classify data items. It is supervised method. It helps to analysis data. Paper uses this feature of KMCA to analysis different datasets. Perform experiment on six datasets in synthetic environment. It shows improvement in cluster quality. It also applies on real datasets. It helps to find road laces by using GPS data. This technique improves performance it uses background knowledge. It combines background knowledge with KMCA. [20]

**Siddheswar Ray and Rose H.Turi, (1999) “Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation”:-** This paper overcome problem of k-mean cluster to determine number of cluster in advance. They present simple validity measure to define no of cluster automatically. VM performs based on in-cluster and out-cluster compute. Basic approach to segment image for two clusters equipped k-max cluster. Best clustering measure is defined based on validity measure. It is performed for artificial image for it numbers of clusters are identified. It can apply for natural images. Median cluster representation is used instead of mean cluster representation. [34]

**Martin Ester, Hans-peter Kriegel, (1996) “Density based algorithm for discovering clusters in large spatial database with noise”:-**Clustering algorithms are used for class identification purpose basically. Spatial clustering has many requirements like field

information to represent the involvement constraints and cluster of arbitrary shapes are discovered in spatial databases but these requirements are not fulfilled by well known clustering algorithms or say our well known clustering algorithm cannot fulfill the requirements of spatial databases. Spatial database means data related to space some spatial database systems are used to manage the spatial data. For identifying class clustering algorithms are used. In spatial data many requirements are there like minimum knowledge domain for determining input parameters, arbitrary shaped cluster discovery, mainly efficiency on the large databases. For clustering in spatial databases new algorithm is proposed i.e. density based algorithm DBSCAN algorithm for spatial databases is used. In partitioning method various partitions of data is done in many clusters like K-mean and K-mediod methods are there in partitioning method. But it is not useful for spatial databases due to not fulfill of the requirements of spatial databases. An algorithm CLARANS is used in clustering i.e. clustering large application based on randomized search. It is efficient and more effective it is a very improved algorithm of K-mediod algorithm but it is not much more efficient for spatial databases. DBSCAN algorithm is used in spatial databases. DBSCAN algorithm takes only one parameter as an input and support user to determine an appropriate value for it. [18]

## **CHAPTER-3**

### **PRESENT WORK**

---

#### **3.1 Problem Formulation**

DM is defined as process of taking out of the inherent and formerly unfamiliar and really possibly useful information from huge amount of data. DM is also called as withdrawal of hidden patterns. This process must be fully automated or semi automated. Patterns that are discovered must be meaningful. It is part of knowledge discovery process. In this one or more computer learning techniques are applied to finding the hidden information in database or to exclude and analyses knowledge from data include within database. DM applies technique to huge amount of data-item to generate models or patterns that are useful for user and also extract hidden pattern. Partitioning methods are simple and most essential version of cluster analysis. Partitioning method results in a set of K clusters, every cluster have at least one item. There are various partitioning clustering algorithm like KMCA, k-medoid clustering algorithm. KMCA is a very simple algorithm. It can also apply on large datasets, it provides good result. But main problem of in KMCA is efficiency. Time required to cluster data is very high but cluster quality is not so good.

#### **3.2 Objectives and Methodology**

- i. To study the concept of K-means clustering algorithms.
- ii. To study the concept of various enhanced K-means clustering algorithms.
- iii. To implement a new hybrid K-means clustering algorithm using its various enhancements.
- iv. To compare new hybrid technique with existing variants.

##### **3.2.1 Steps of Performing Proposed Method**

1. Study the basic concepts of data mining.
2. Study the knowledge discovery from data or KDD process.
3. Study the K-means clustering algorithm.
4. Study various enhanced K-means clustering algorithms.
5. Selecting best 2 enhanced algorithms based on performance.

6. Hybrid algorithm is developed by using the selected 2 enhanced algorithms.
7. Comparative Analysis of hybrid algorithm with other enhanced k-means clustering algorithms.

### 3.3 Basic Design of Proposed Work

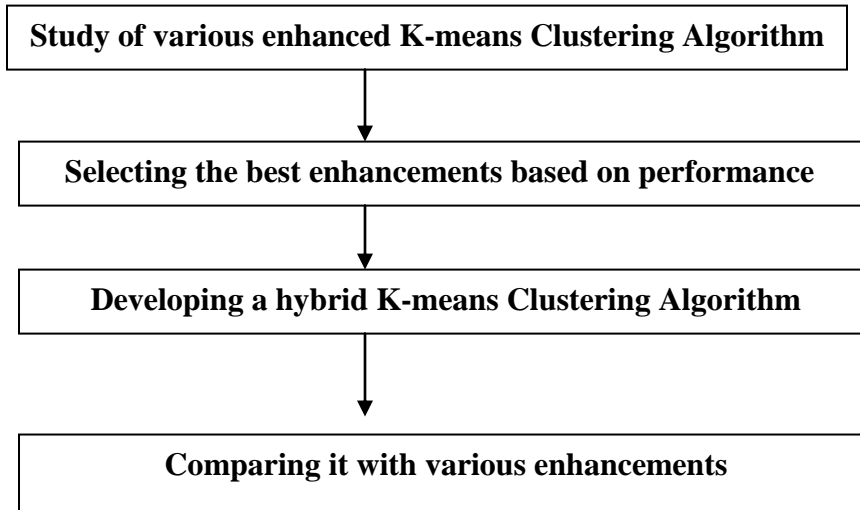


Fig 3.3 design of proposed work

#### HYBRID ALGORITHM

Input:

$D = \{d_1, d_2, \dots, d_n\}$  // set of n data items.

k// Number of clusters.

Output:

A set of k clusters.

Steps:

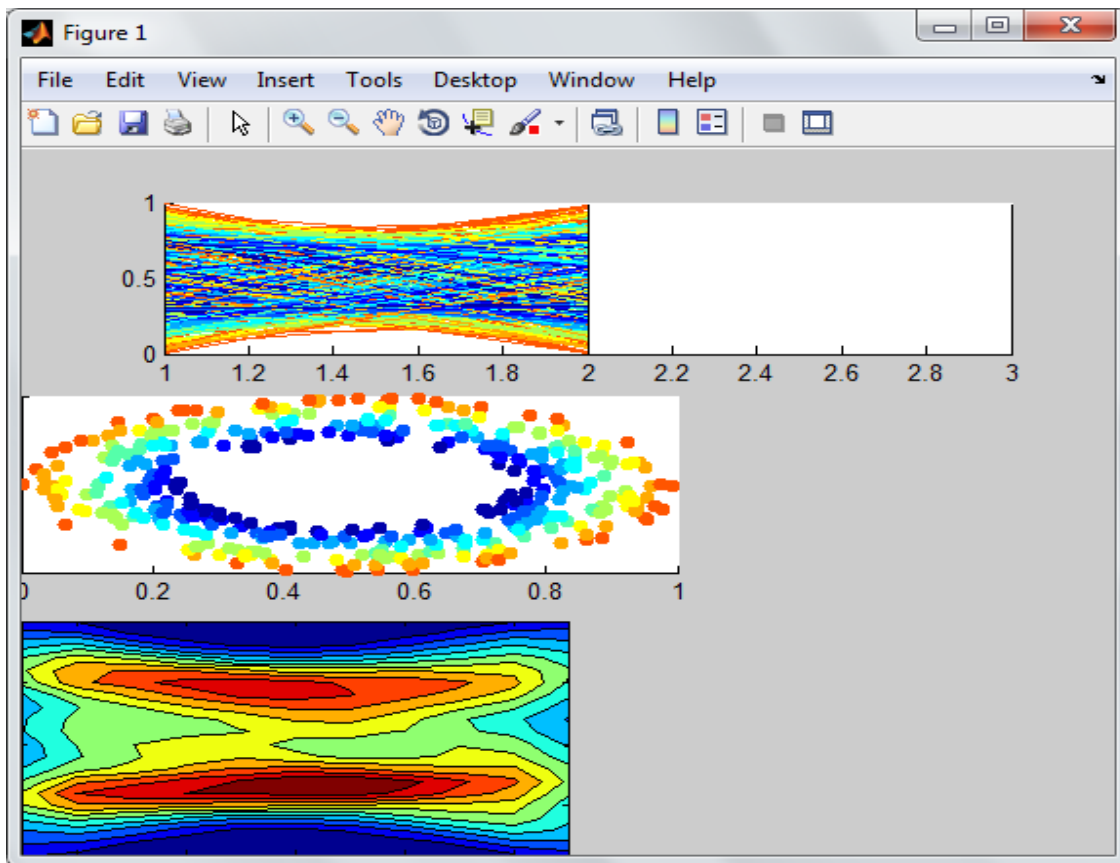
1. For each column of the data set, determine the range as the variation between the maximum and the minimum element;
2. Identify the column with maximum range;
3. Sort the entire data set increasing order based on the column having the maximum range;
4. The sorted data set are partitioned into k equal parts;
5. Determine the arithmetic mean of each part obtained in Step 4 as  $a_1, c_2, \dots, a_k$ ; Take these mean values as the initial centroids.



6. Compute the distance of each data-point  $d_i$  ( $1 \leq i \leq n$ ) to all the centroids  $c_j$  ( $1 \leq j \leq k+1$ ) as  $d(d_i, c_j)$
7. For each data-point  $d_i$ , find the closest centroid  $c_j$  and assign  $d_i$  to cluster  $j$
8. Set  $\text{ClusterId}[i] = j$ ; //  $j$ : Id of the closest cluster
9. Set  $\text{Nearest\_Dist}[i] = d(d_i, c_j)$
10. For each cluster  $j$  ( $1 \leq j \leq k+1$ ), recalculate the centroids
11. Repeat
12. for each data-point  $d_i$ 
  - 12.1 Compute its distance from the centroid of the present nearest cluster
  - 12.2 If this distance is less than or equal to the present nearest Distance, the data-point stays in the cluster,  
Else
    - 12.2.1 For every centroid  $c_j$  ( $1 \leq j \leq k+1$ ) Compute the distance ( $d_i, c_j$ ); End for
    - 12.2.2 Assign the data-point  $d_i$  to the cluster with the nearest Centroid  $C_j$
    - 12.2.3 Set  $\text{ClusterId}[i] = j$
    - 12.2.4 Set  $\text{Nearest\_Dist}[i] = d(d_i, c_j)$ ; End for
13. For each cluster  $j$  ( $1 \leq j \leq k+1$ ), recalculate the centroids; until the convergence Criteria is met.

## CHAPTER-4 RESULTS AND DISCUSSIONS

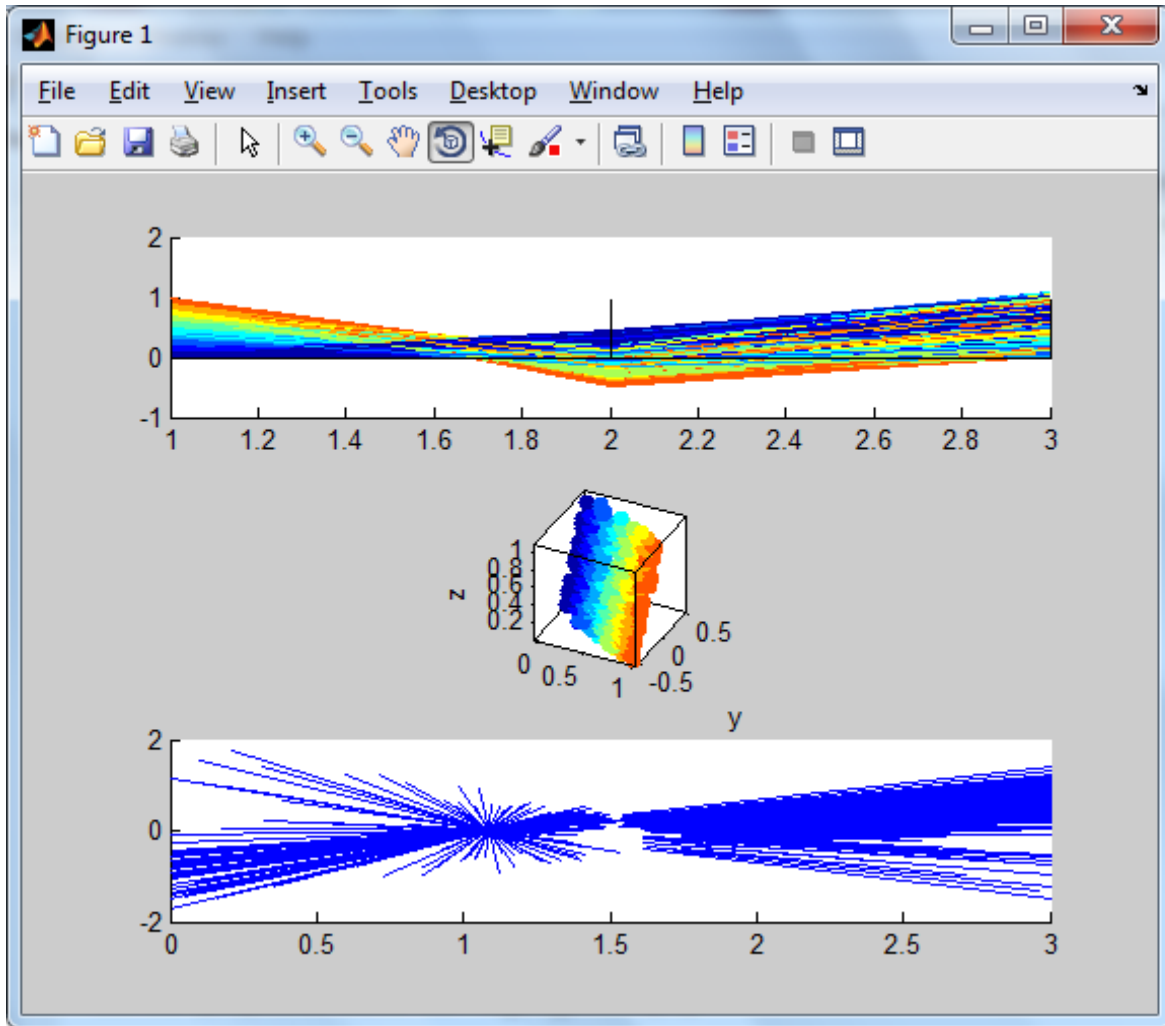
**4.1 Tool:-**The proposed idea will be implemented in MATLAB which is broadly used in all areas of research universities, and also in the industry. MATLAB is valuable for mathematics equations (linear algebra) and also beneficial for numerical integration equations. These equations are solved by MATLAB It is also a programming language, and is one of the simplest programming languages widely used for writing mathematical programs. It has a range of types of tool boxes that are very precious for optimization and many more.



**Fig 4.1.1 DATASET Clustered**

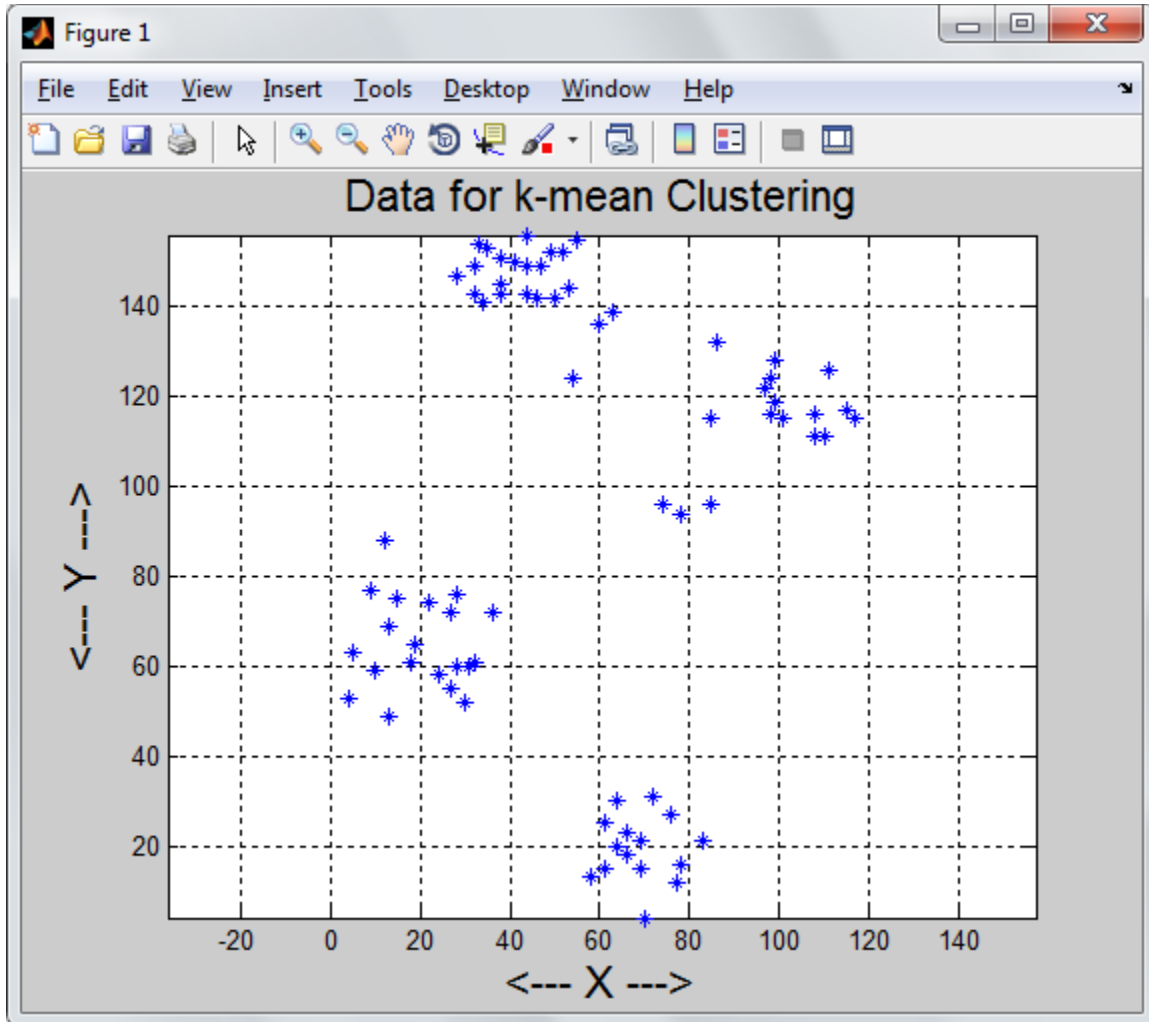
As shown in the figure 1, as explained in the previously the K-mean is the algorithm in which the data will be clustered according to Euclidian distance. The random center points had been selected from the data. The Euclidian distance will be calculated from the data centers to

other points and points will be clustered accordingly. The output of the clustered will be shown in the 2D plane. When the data will be shown in 2D plan, some points which are very close to each other cannot be shown which reduce the cluster quality.



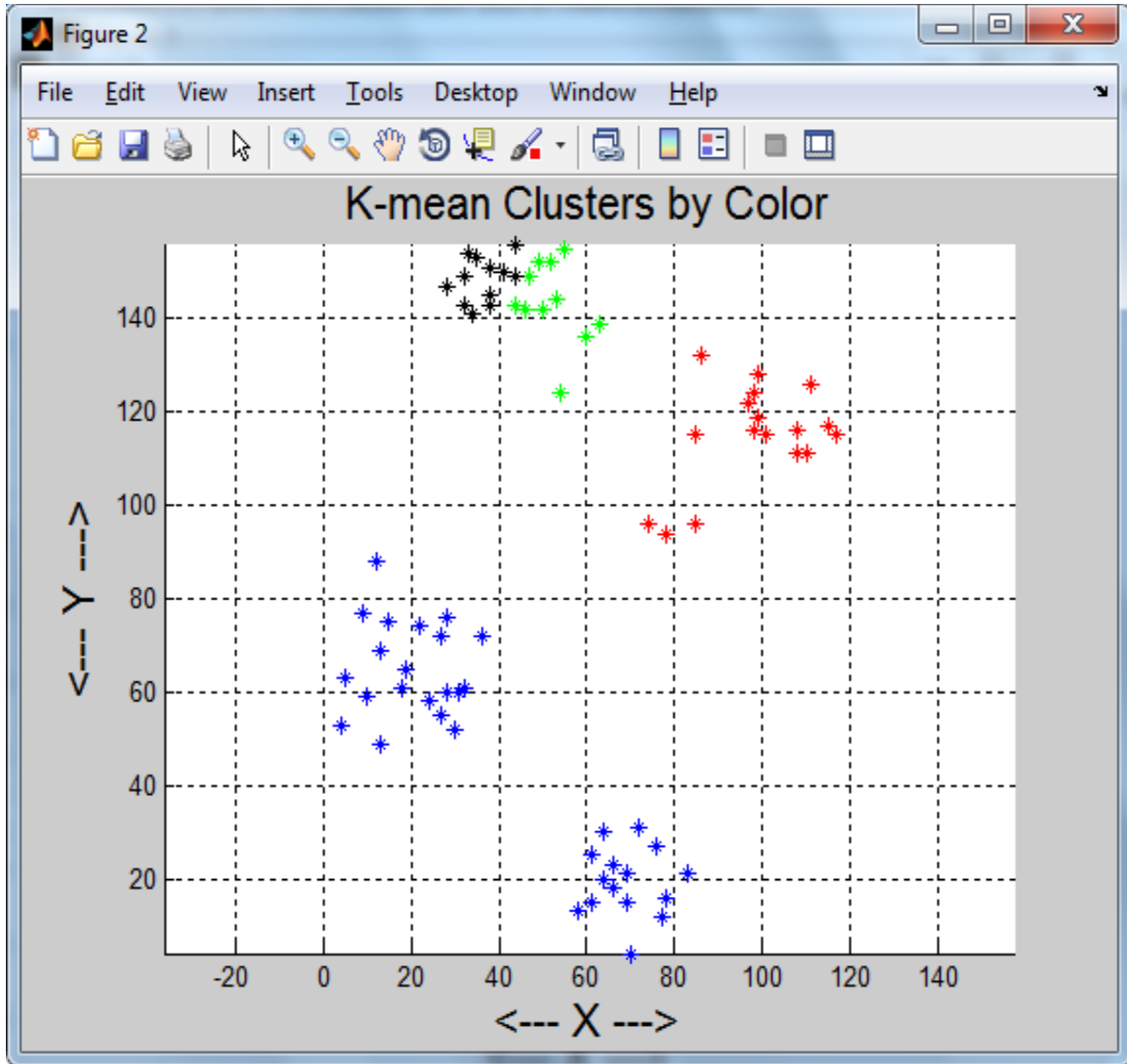
**Fig 4.1.2 3D plane clustering**

As shown in the figure2, in the last figure, the dataset will be shown on the 2D plane after clustering. In the figure2, the hybrid algorithm will be applied and clustering will be done on the basis of Euclidian distance. In new algorithm, centre point will be selected randomly and data will be plotted on the 3D plane. When the data will be plotted on the 3D plane, the cluster quality will be improved and points can be shown more accurately.



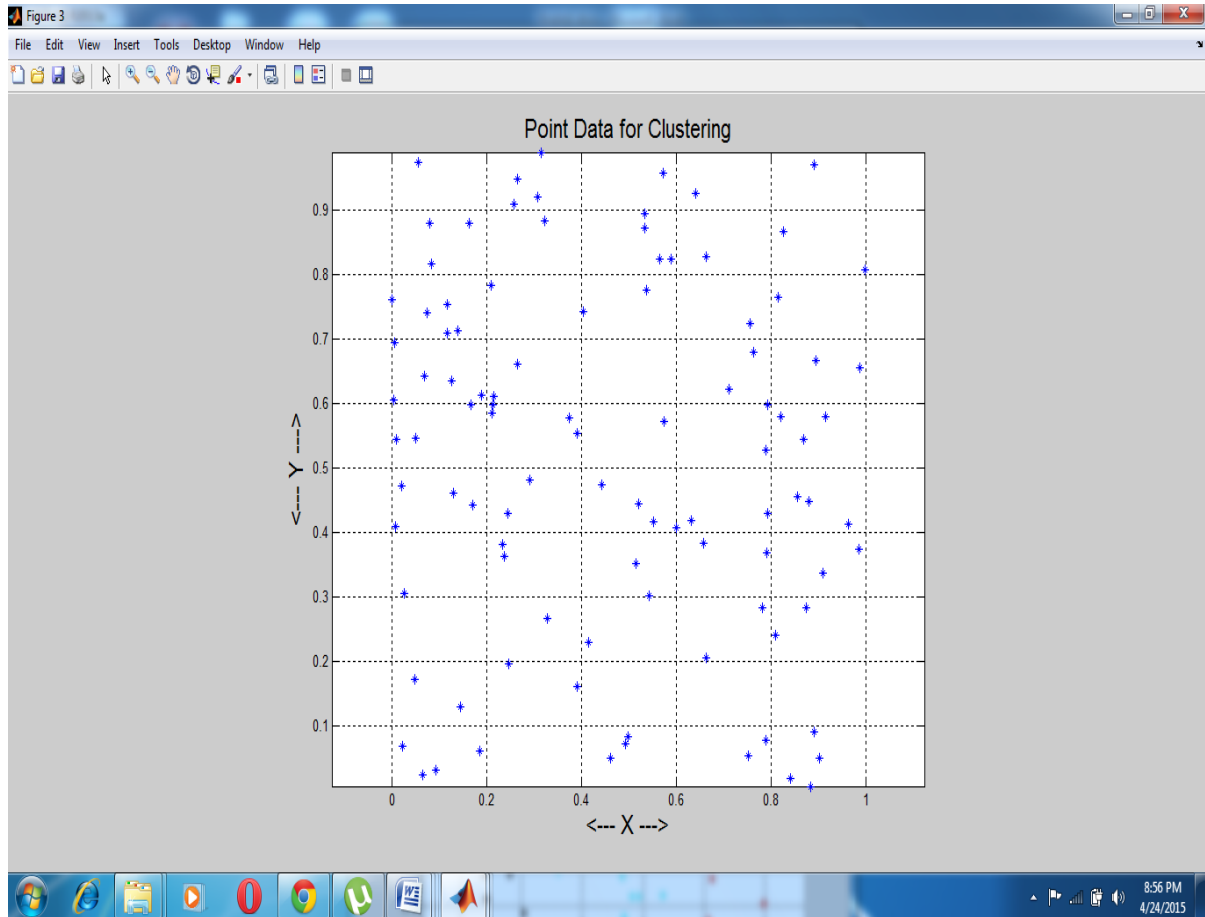
**Fig 4.1.3 First iteration of clustering**

As shown in the figure 3, to define performance of the algorithm. K-mean algorithm will be applied on another dataset. In this Dataset, various figures have been shown for data clustering. In the figure3, the data points have been plotted which we have to cluster and first centered point will be selected and according to first selected points data will be clustered. Initially KMCA randomly choose number of clusters (k) of data item that contain in data set D, these present group centriod. Then other data item puts in to that bunch to which those are closer. Distance between data items and group mean calculated by using Euclidean formula.



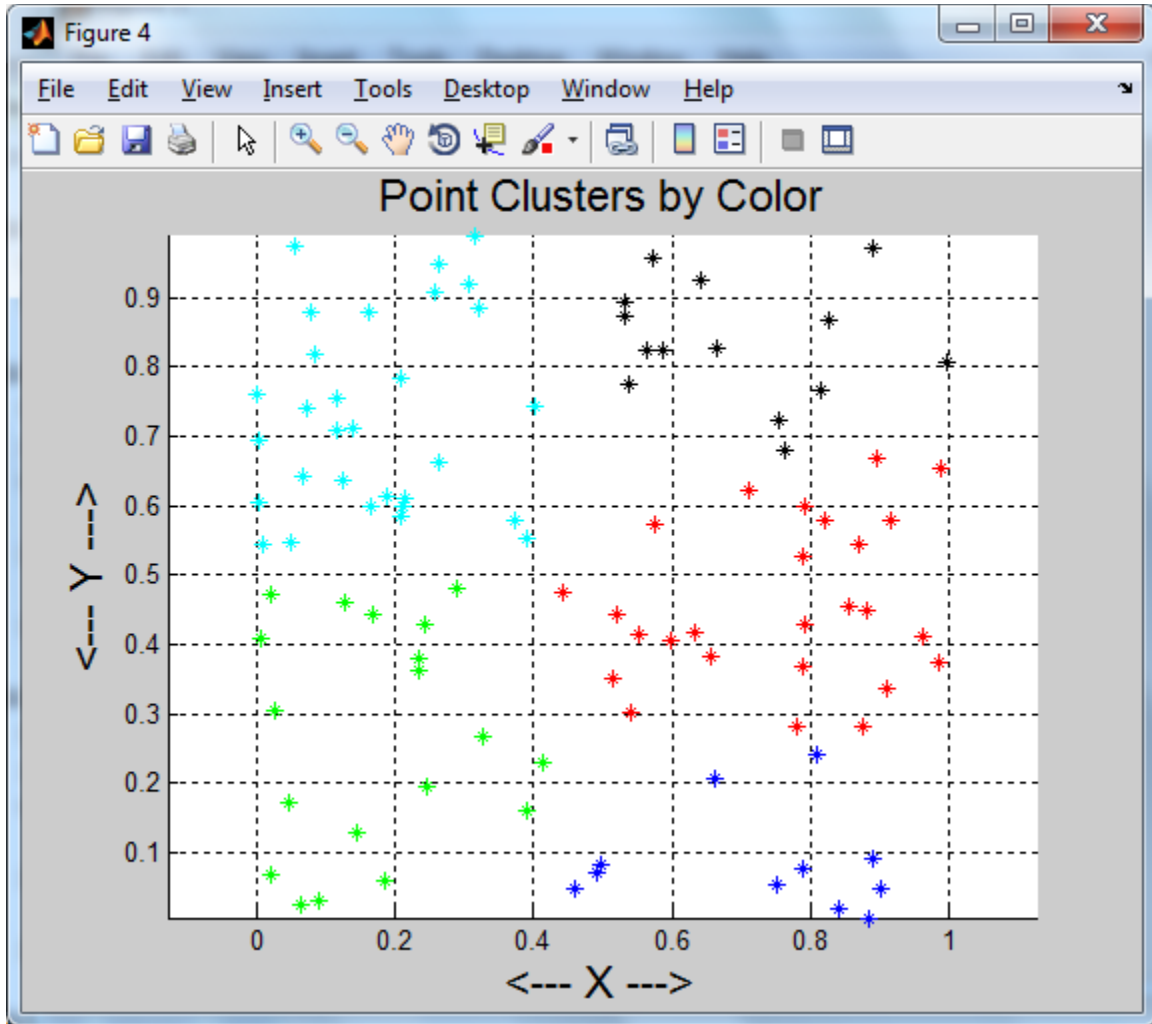
**Fig 4.1.4 Coloring of clusters**

As shown in figure4, in figure3, the first selected points are used for clustering of data. In this figure, the data will be clustered using Euclidian distance and this data will be shown with different colors for better analysis of data. Initially KMCA randomly choose number of clusters (k) of data item that contain in data set D, these present group centroid. Then other data item puts in to that bunch to which those are closer. Distance between data items and group mean calculated by using Euclidean formula.



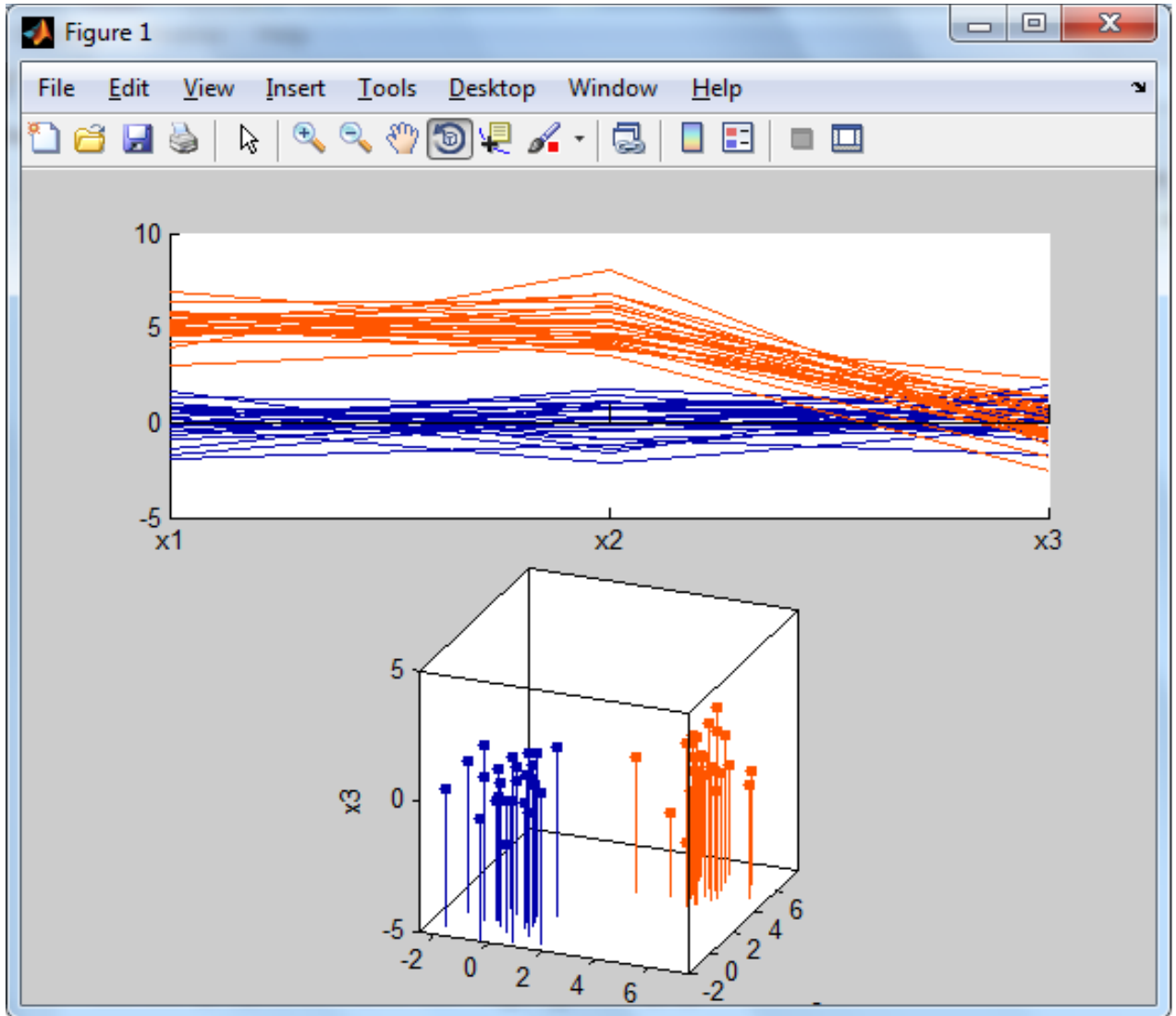
**Fig 4.1.5 Data points represent by color for clustering**

In this figure dataset that is used for clustering is pointed by using colors. Then apply KMCA on it. Find Euclidian distance of dataset from different centriods. Then cluster this data points. Initially KMCA randomly choose number of clusters (k) of data item that contain in data set D, these present group centriod. Then other data item puts in to that bunch to which those are closer. Distance between data items and group mean calculated by using Euclidean formula. Then these datasets are clustered based on their distance to centriod of each cluster. Data point assign to cluster to which they are closer. Data points within cluster have more similarity and data point which lies in another contain dissimilarity. It contains homogenous data within clusters and heterogeneous data points between clusters.



**Fig 4.1.6 Clustering of Data**

As shown in figure 5, data set which is used for clustering is been clustered and each cluster will be marked with different colors. In this figure, various iterations are runned, means at every iteration new centered point is selected and on the basis of that centered point, cluster assignment procedure will be done.



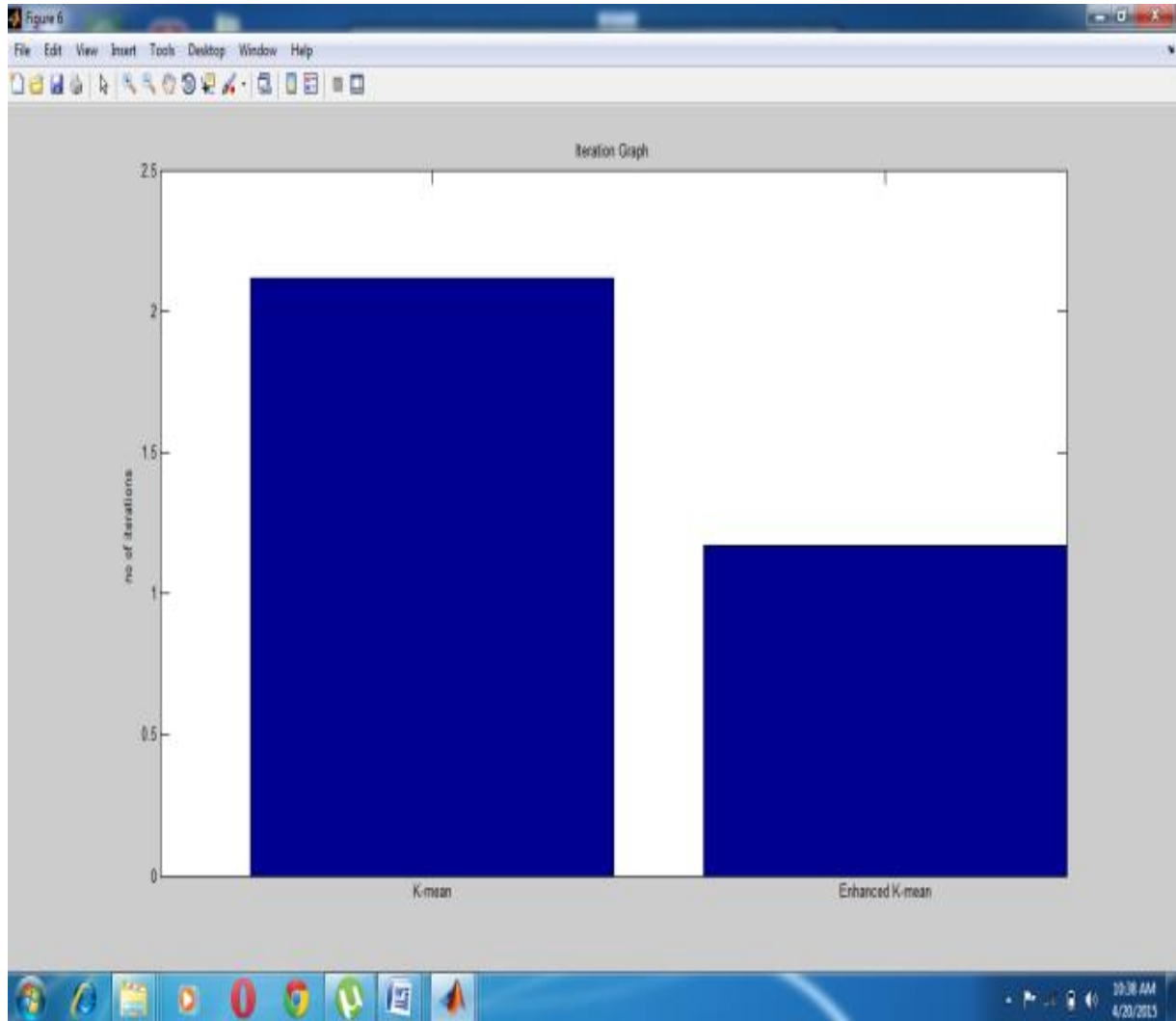
**Fig 4.1.7 3D Representation**

As shown in figure 5, the dataset which is used in the previous figure will be clustered using the hybrid type of k-mean clustering algorithm. When the dataset will be clustered using hybrid algorithm cluster quality will be improved and each point in the dataset will be shown on 3D plane for better analysis of dataset.

Novel hybrid technique will improve processing speed. Because within few iteration it produce good quality clusters. KMCA takes more time to cluster datasets as compare to novel hybrid algorithm, because hybrid algorithm takes very small number of iteration to cluster iris datasets. Hybrid algorithm also improves cluster quality because it represents data in 3D plane. Following graph show comparison between basic KMCA and proposed hybrid

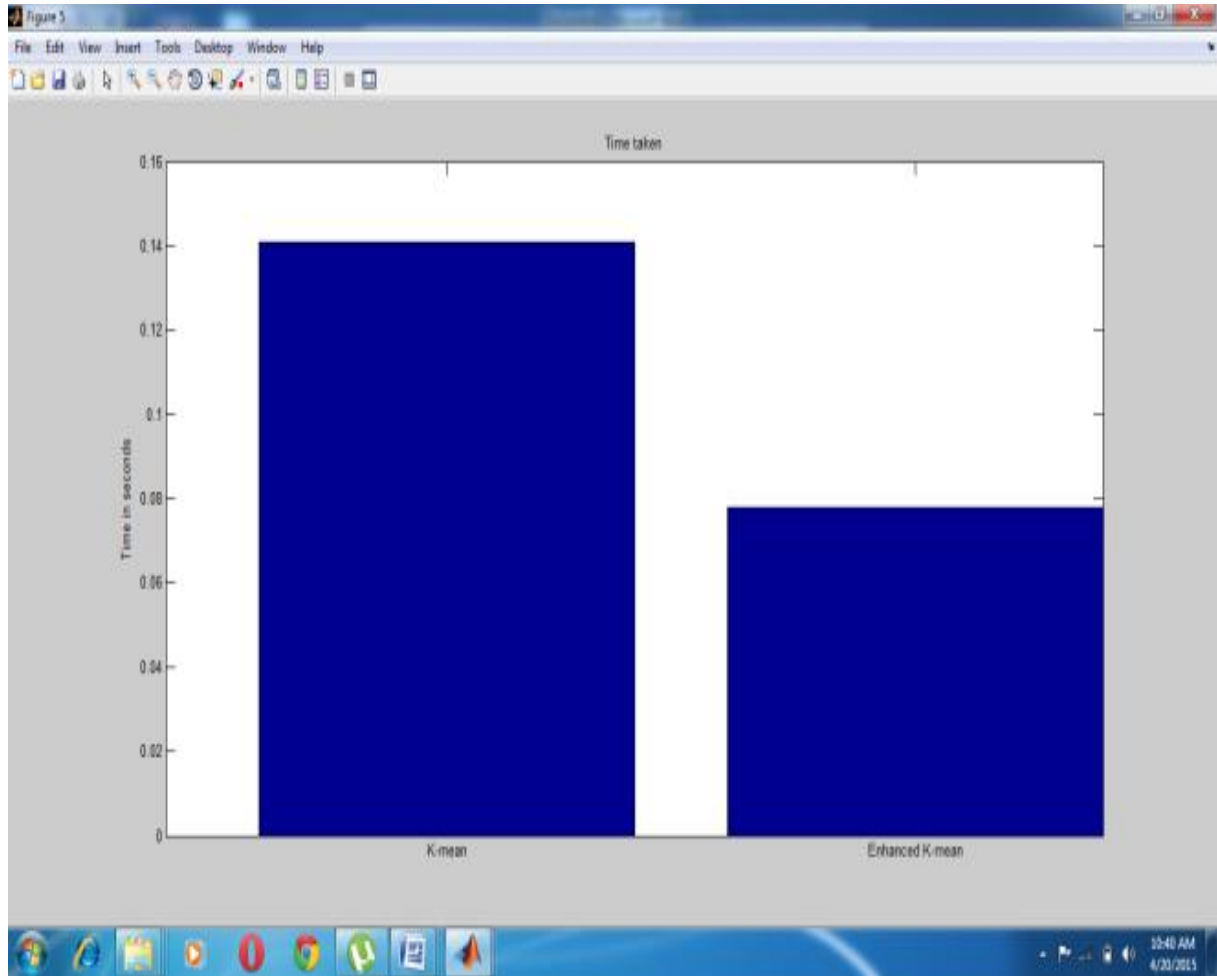


k-mean clustering in term of number iteration and time take by each algorithm to cluster datasets



**Fig 4.1.8 Iteration graph**

K-mean clustering algorithm processing speed is very low. So it takes number of iteration to form clusters. In compare to k-mean clustering algorithm, new hybrid takes very less amount of time to cluster data. So it takes very less amount of iteration to clusters larege amount of data.



**Fig 4.1.9 Time graph**

Enhanced KMCA take very less amount of time to cluster dataset as compare to basic KMCA. Because processing speed of enhance KMCA is more as compare to basic KMCA. A new hybrid technique takes less number of iterations to cluster dataset, so it cluster data in very less amount of time.

## **CHAPTER-5**

### **CONCLUSION AND FUTURE SCOPE**

---

#### **6.1 Conclusion**

Data mining is the process taking out hidden pattern from data. Various computer techniques are applied to get meaningful information from data in database. It is also part of knowledge discovery process. It provides knowledge to user that is required by them. Clustering is unsupervised learning technique. It is used to cluster data. Data is organized in clusters such a way that data within cluster is more similar. That there is high intra cluster similarity and high inter cluster dissimilarity. There is various clustering method such as PM, HM, DBSCAN, and GM. Partition method is very simple and most fundamental version of cluster analysis. It contains k clusters and each cluster contains at least one object. For Real world data arithmetic mean of data is select as centriod of cluster or representative of cluster. And many alternative methods are used to select the centriod of cluster. K-mean clustering algorithm is oddly used clustering technique. It comes under partitioning method. K-mean is very simple and easy way to classify data. It also applied on large datasets. But it has efficiency problem. It takes great amount of time to cluster data and cluster excellence is not so good. It also has another problem like there is need to define number of clusters in advance. So there is need to define good technique to define initial cluster center. I study various enhancement of KMCA, and then based on performance I select two algorithms. After that combines these two algorithms made a hybrid algorithm which compare with other enhanced algorithms. Hybrid novel technique has been providing following improvement i.) Improve the cluster quality. ii.) Increase processing speed. iii.) Decrease number of iterations.

#### **6.2 Future Scope**

In KMCA has a limitation like there is need to define number of cluster in advance. Mean before starting processing this algorithm we require to give value of k as input. This proposed hybrid algorithm also has this problem. We need give value of k as input. This algorithm can apply on large datasets. But need to develop more algorithms that can handle complex datasets. It many more weakness like it randomly select initial cluster center. Outcome of

KMCA totally depend upon selection of initial cluster centriod. Hybrid algorithm removes this problem of KMCA. It does not randomly select initial cluster centre. It also increase processing speed and reduce number of iterations.

## **CHAPTER-6**

### **REFERENCES**

---

[1] Aastha Joshi, Rajneet kaur, (2013) “A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3.

[2] Amar Singh, Navjot Kaur, (2013) “To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8.

[3] Anwiti Jain, Anand Rajavat, Rupali Bhartiya, (2012) “Design, analysis and implementation of modified K-mean algorithm for large data-set to increase scalability and efficiency”, Fourth international conference on computational intelligence and communication networks.

[4] Anand M. Baswade, Kalpana D. Joshi and Prakash S. Nalwade, (2012) “A Comparative Study Of K-Means and Weighted K-Means for Clustering,” International Journal of Engineering Research & Technology, Volume 1, Issue 10.

[5] Ahamed Shafeeq B M and Hareesha K S, (2012) “Dynamic Clustering of Data with Modified K-Means Algorithm,” International Conference on Information and Computer Networks, Volume 27.

[6] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, (2012) “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity,” Middle-East Journal of Scientific Research, pages 959-963.

[7] Chien-Hsing Chou, Yi-Zeng Hsieh, Mu-Chun Su, and Yung-Long Chu, (2013) “Extracting and Labeling the Objects from an Image by Using the Fuzzy Clustering Algorithm and a New Cluster Validity”, International Journal of Computer and Communication Engineering, Vol. 2, No. 3.

[8] Ms.Chinki Chandhok Mrs.Soni Chaturvedi, Dr.A.A Khurshid (2012)“An Approach to Image Segmentation using K-means Clustering Algorithm”, International Journal of Information Technology (IJIT), Volume -1,Issue.

[9] Chunfei Zhang, Zhiyi Fang, (2013) “An Improved K-means Clustering Algorithm”, Journal of Information & Computational Science 10: 1.

[10] Dibya Jyoti Bora, Anil Kumar Gupta, (2014) “A New Approach towards Clustering based Color Image Segmentation”, International Journal of Computer Applications (0975 – 8887) Volume 107 – No 12.

[11] Daljit Kaur and Kiran Jyot, (2013) “Enhancement in the Performance of K-means Algorithm”, International Journal of Computer Science and Communication Engineering Volume 2 Issue 1.

[12] D T Pham\_, S Dimov, and C D Nguyen, (2004) “Selection of K in K-means clustering”, Proc. Mache Vol. 219 Part C: J. Mechanical Engineering Science.

[13] F.U.Siddiqui, N.A.Mat Isa, (2012) “Optimized k-means clustering algorithm for image segmentation”, School of Electrical and electronic engineering, university Sains Malaysia, 14300, Nibong Tebel, Penang, Malaysia.

[14] Harpreet Kaur and Jaspreet Kaur Sahiwal, (2013) “Image Compression with Improved K-Means Algorithm for Performance Enhancement”, International Journal of Computer Science and Management Research, Volume 2, Issue 6.

[15]Henry Lin, “Clustering”, 15-381 Artificial Intelligence.

[16] Joaquín Pérez Ortega, Ma. Del Rocío Boone Rojas, María J. Somodevilla García, (2009) “Research issues on K-means Algorithm: An Experimental Trial Using Mat lab”.

- [17] Jiawei Han, Micheline Kamber, Jian Pei, “Data mining concepts and techniques”, Third edition.
- [18] Kajal C. Agrawal and Meghana Nagori, (2013) “Clusters of Ayurvedic Medicines Using Improved K-means Algorithm,” International Conf. on Advances in Computer Science and Electronics Engineering.
- [19] K. A. Abdul Nazeer, M. P. Sebastian, (2009) “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering, Vol IWCE.
- [20] Kiri Wagsta and Claire Cardie, (2001) “Constrained K-means Clustering with Background Knowledge”, Proceedings of the Eighteenth International Conference on Machine Learning.
- [21] L.V.Bijuraj, (2013) *clustering and its applications*, Proceedings of National Conference on New Horizons in IT – NCNHIT.
- [22] Macqueen, J.B: (1967) *some methods for classification and analysis of multivariate observations*. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, pp. 281–297.
- [23] Madhu Yedla, T M Srinivasa, (2010) “Enhancing K-means Clustering Algorithm with Improved Initial Center”, International Journal of Computer Science and Information Technologies, Vol. 1 (2).
- [24] Ming Chan hangs et.al, (2005) “An Efficient k-Means Clustering Algorithm Using Simple partitioning”, NSC.
- [25] Malay K. Pakhira, (2009) “A Modified *k*-means Algorithm to Avoid Empty Clusters”, International Journal of Recent Trends in Engineering, Vol 1, No. 1.

- [26] Dr. M.P.S Bhatia and Deepika Khurana, (2013) “Experimental study of Data clustering using k-Means and modified algorithms”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3.
- [27] M. N. Vrahatis, B. Boutsinas, (2002) “The New k-Windows Algorithm for Improving the k-Means Clustering Algorithm,” Journal of Complexity 18, pages 375-391.
- [28] Neha Aggarwal, Kirti Aggarwal and Kanika Gupta, (2012) “Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining,” International Journal of Scientific & Engineering Research, Volume 3, Issue 3.
- [29] Osamor VC, Adebisi EF, Oyelade JO and Doumbia S, (2012) “Reducing the Time Requirement of K-Means Algorithm” *PLoS ONE*, Volume 7, Issue 12.
- [30] Oyelade, O. J, Oladipupo, O. O, (2010) “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7.
- [31] Pritesh Vora, Bhavesh Oza, (2013) “A Survey on K-mean Clustering and Particle Swarm Optimization”, International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-3.
- [32] Raed T. Aldahdooh, Wesam Ashour, (2013) “Distance-based Initialization Method for K-means Clustering Algorithm”, I.J. Intelligent Systems and Applications, 02.
- [33] Shital A. Raut and S. R. Sathe, (2011) “A Modified Fast map K-Means Clustering Algorithm for Large Scale Gene Expression Datasets”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 1, No. 4.
- [34] Siddheswar Ray and Rose H. Turi, (1999) “Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation”, School of Computer



Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.

[35] T. Velmurugan and T. Santhanam, (2010) “Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points”, *Journal of Computer Science* 6 (3): 363-368, 2010 ISSN 1549-3636, © 2010 Science Publications.

[36] Tapas Kanungo, David M. Mount, (2002) “An Efficient k-Means Clustering Algorithm: Analysis and Implementation”, *IEEE Transaction on pattern analysis and machine intelligence*, VOL. 24, NO. 7.

[37] Tina Eliassi-Rad Terence Critchlow, (2007) *Multivariate Clustering of Large-Scale Scientific Simulation Data*, In *Proceedings of SIGMOD Record*, ACM Press 28(4):49-57.

[38] Vijay Jumb Mandar Sohani Avinash Shrivasa, (2014) “Color Image Segmentation Using K-Means Clustering and Otsu’s Adaptive Thresholding”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, Volume-3, Issue-9.

[38] Yugal Kumar and G.Sahoo, (2014) “A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm”, *International Journal of Advanced Science and Technology* Vol.62, (2014), pp.43-54.

[39] Yiu-Ming Cheung, “k-Means: A new generalized k-means clustering algorithm”, *Pattern Recognition Letters* 24 (2003) 2883–2893, 2002.

[40] Z.Volkovich, D. Toledano-Kitai · G.-W. Weber, (2012) “Self-learning K-means clustering: a global optimization Approach”, © Springer Science Business Media.

[41] Zhaohong, Kup-SZE, Choi Deng (2013) “A survey on soft subspace clustering”, *WAIM 2005, LNCS 3739*, pp. 475–491. © Springer-Verlag Berlin Heidelberg.

## CHAPTER-7 APPENDIX

---

1. CUB- Called unit block
2. C- Cluster
3. D- Set of data containing n objects.
4. DBSCAN- Density based clustering algorithm.
5. DM- Data mining
6. K- Number of clusters.
7. KDD- Knowledge Discovery
8. N- Number of objects in data sets.
9. OKM- Optimized k-means
10. KMCA- k-means clustering algorithm
11. MKCA- Modified K-mean clustering algorithm
12. PM- partitioning method
13. HM- hierarchical method
14. GM- grid method
15. CA- clustering algorithm
16. UB- Unit block
17. 3D- Three dimensional plane
18. 2D- Two dimensional plane
19. SSC- Soft subsurface clustering
20. XSSE-Extended soft subspace clustering
21. ISSC-Independent subspace clustering
22. CSSC-Conventional subspace clustering
23. HSSC-Hard subspace clustering
24. K-L-Kullback Lieber
25. FCM-Fuzzy C-mean