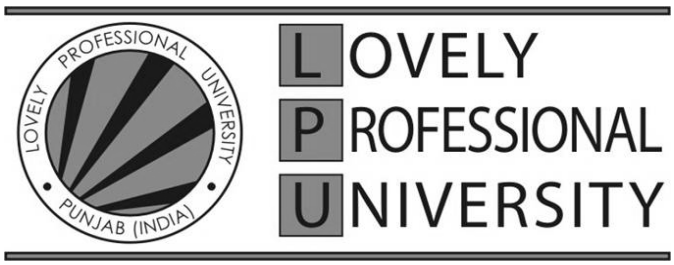

To Intensify Cluster Quality By Enhancing K-means Clustering Algorithm In Data Mining

A Dissertation Proposal submitted

By

Twinkle Garg

to

Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Mr. Arun Malik

(May 2015)

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of:Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	:Twinkle Garg	Registration No	:11002521
Batch	:2010-2015	Roll No	:RK2006B23
Session	:2014-2015	Parent Section	:K2006

Details of Supervisor:

Name	:Arun Malik	Designation	: Assistant Professor
UID	:17442	Qualification	: M.Tech
		Research Exp.	:3 years

Specialization Area:Databases (pick from list of provided specialization areas by DAA)

Proposed Topics:-

1. K-Means Algorithm in clustering based Data Mining.
2. Association rule based data mining.
3. Classification techniques in Data mining.

Signature of supervisor

PAC Remarks: Topic 1 approved. Paper expected.

APPROVAL OF PAC CHAIRMAN Signature: Date:

*Supervision should finally encircle one topic out of three proposed topics and put up for an 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.

CERTIFICATE

This is to certify that Twinkle Garg has completed M.Tech dissertation proposal titled "**To Intensify Cluster Quality By Enhancing K-means Clustering Algorithm In Data Mining**" under my guidance and supervision. To the best of my knowledge, the present work is the result of her 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 fulfilment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name: Mr. Arun Malik

UID:

ABSTRACT

Data Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term data mining is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Clustering is an unsupervised classification method aims at creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. K-means clustering algorithm is a partitional based algorithm. To overcome the deficiencies of existing k-means algorithm our motive is to propose a new improved k-means algorithm.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my mentor, **Mr. Arun Malik**, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my literature survey especially in writing this report. My thanks and appreciations also go to my colleague and people who have willingly helped me out with their abilities.

DECLARATION

I Twinkle Garg declare that the dissertation proposal entitled, **"To Intensify Cluster Quality By Enhancing K-means Clustering Algorithm 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:

Investigator

Regn.No:11002521

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION

1.1 Data Mining

Data Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term data mining is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Based on the patterns we look for the Data Mining models and tasks are divided into two main categories Predictive models and Descriptive Models. Whereas the Predictive Model is used to predict the feasibility of outcome, the other Descriptive model is used to describe the important features of dataset. The types of Predictive model are classification, regression, prediction and time series analysis. The various models included in descriptive model are clustering, summarization, Association rules and sequence discovery. The data mining technique collects the data from different sources such as database, World Wide Web, data marts and then clean the data to remove outliers and then integrate the data. After integration data is sent to data warehouse and after that data selection and transformation operations are applied. Next data is sent for data mining and then pattern evaluation is performed to find out hidden patterns. Lastly visualization of patterns is done in knowledge presentation.

There are needs to define a technique to compare the data for each function. Its functionalities are specified below:

Characterization and Discrimination: - Summarization of data of the class under study is called data characterization and comparison of the target class with one or a set of comparative classes is known as data discrimination. Descriptions of Class/concept are derived using these two functionalities.

The Mining of Frequent Patterns, Associations and Correlations: - The patterns that occur frequently in data are known as frequent patterns. Association rule mining is the process of determining frequent item sets and generate strong association rules using frequent item sets in a large transactional database.

Classification and Regression:-Classification is a supervised learning data mining technique used to predict group membership for data instances. Regression analysis is a methodology that is mainly used for numeric prediction.

Cluster Analysis: - clustering is defined as a process used for organizing/grouping a large amount of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters

Outlier Analysis:-Some objects in a data set do not have similar characteristics with other data within a cluster. These data objects are outliers and analysis of outlier data is known as outlier analysis.

KDD is an iterative process which contains following steps.

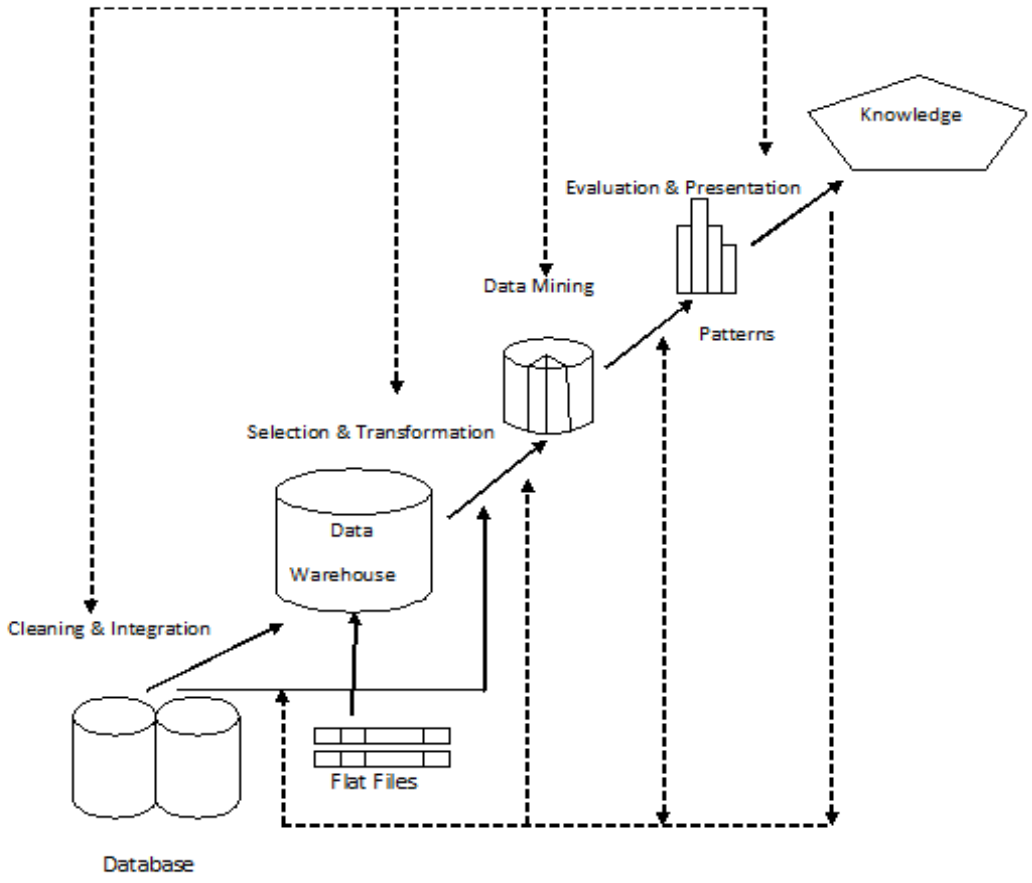

Fig 1.1 KDD process

Data cleaning: - this step is used to remove outliers and unreliable data from huge set of data by applying processes like find missing values, and noisy data is removed by applying binning method.

Data integration: - in this process data from various data sources is integerated. It is process of integrating data from different sources of data.

Data selection: - in this step meaningful and useful data for analysis is selected from different databases.

Data transformation: - in this step, aggregation and summary operations are applied to transform and consolidate data into forms that is required for suitable for mining

Data mining: -this is an important step which is used to extract meaningful patterns from bulk amount of data.

Pattern evaluation: - in this step, interestingness measures are used to symbolize knowledge.

Knowledge presentation: - in this step, visualization or other knowledge representation techniques are used to present extracted knowledge to user.

1.2 Data Mining Applications:

Market analysis –Finds out the general features of purchaser who buy the similar item.

Customer churn - Identify which purchaser are most frequently to leave your company and go to a other company that is your opposition

Direct marketing – Find out which proposal can be added in a mailing list to obtain the highest response rate.

Interactive marketing – Search if the Web site is interesting in viewing that is accessed by most of the individual.

Market basket analysis - Identify what items are frequently buy together; e.g., bread and jam.

Fraud detection - Predict which activities are mostly to be fake.

Trend analysis – Tell about the difference between a typical customer for present and the previous month

1.3 Advantages

Data Mining Predict future trends and purchase habits of customer

It also helps in decision making

It improves company revenue and lower costs

Market basket analysis is another advantage

It also helps in Fraud detection

1.4 Disadvantages

A great threat to security of user.

Unlimited data

Data Mining can also lead to misuse of information

Inaccurate data

1.5 Clustering in Data Mining

Clustering an unsupervised learning technique established in the area of data mining .Clustering or cluster analysis can be defined as a data reduction tool used to create subgroups that are more manageable than individual datum. Generally, clustering is defined as a process used for organizing/grouping a large amount of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters. The applications areas where clustering plays an important role are machine learning, image processing, data mining, marketing, text mining. The terms clustering and classifications are always confused with each other, since they are two separate terms. Whereas Clustering is unsupervised learning process because the resulting clusters are not known before the execution which implies the absence of predefined classes in clustering. On the other hand classification is a supervised learning process due to presence of predefined classes. The high quality clustering is to obtain high intra cluster similarity and low inter-cluster similarity. There exists number of algorithms that are used for clustering the data.

The major application areas of cluster analysis include pattern recognition, image processing, market research, data analysis.

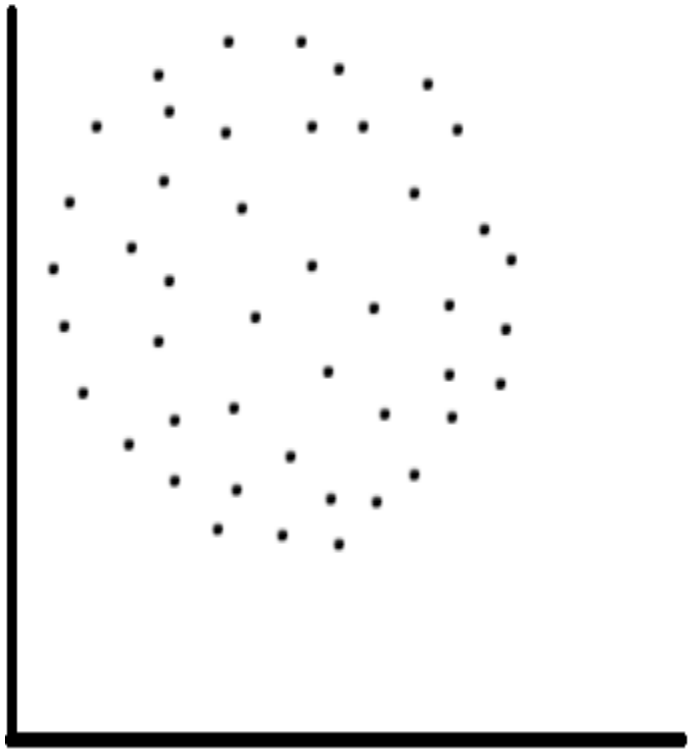

Fig. 1.2 Data without clustering

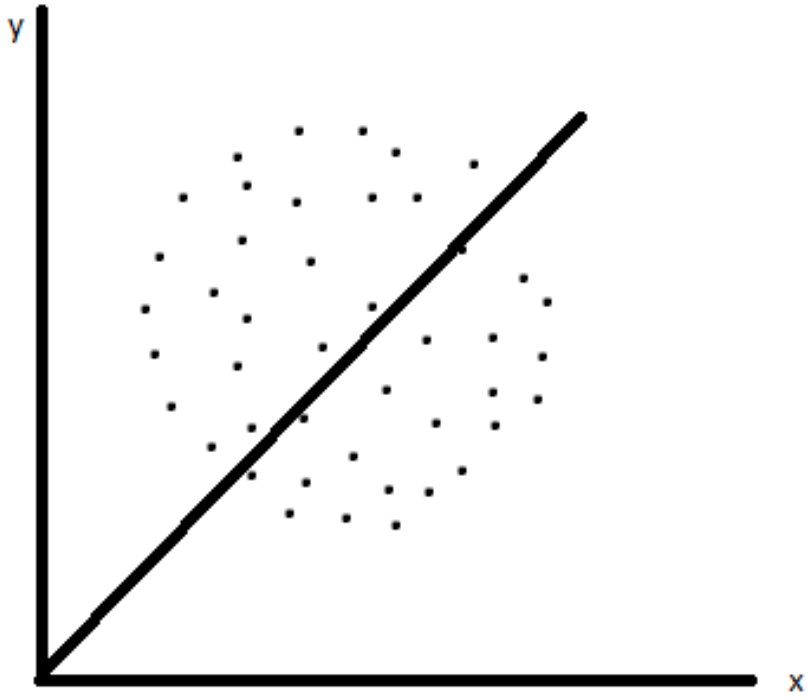

Fig.1.3 Data with Partitional clustering

The basic clustering methods used in data mining can be categorized into following types:

Partitioning Clustering

Grid Based Clustering

Density Based Clustering

Hierarchical Clustering

1. Partitioning Based Clustering: The partitioning algorithms divide data into clusters**.**In a good partitioning the data with common characteristics are grouped in the same cluster whereas data with dissimilar features are grouped in different clusters. Many applications have adopted popular heuristic methods like greedy approaches like the k-means and k-medoids algorithms which effectively intensify the clustering quality and reach a local optimum. This clustering algorithm produces spherical shaped clusters in small to medium size databases. The basic functionality of partitioning method is to increase similarity of the data inside the clusters and minimize dissimilarity between different clusters. Most partitioning methods are distance-based. K-means algorithm is the popularly used partitional based clustering algorithm that is used to group data based on Euclidean distance.

2. Grid Based Methods: Grid based clustering method forms a grid structure by quantizing the object space into a restricted number of cells. A set of grid cell is initially defined and objects are assigned to these grid cells and then density is computed. Next cell are eliminated whose density is less than desired threshold. It is a fast method and is not dependent on the number of data items and is dependent only on the number of cells in each dimension in the quantized space. In the grid-based method, the objects together form a grid. Fast processing time is the basic advantage of this clustering.

3. Density based Methods: In this clustering, the density of the neighbourhood instances is checked to form the clusters. The areas of higher density are defined as clusters. DBSCAN and OPTICS are the popularly known density based clustering method. It is based on connecting points with certain threshold distance. So for arbitrary shapes new methods are

used known as density-based methods which are based on the parameter of density. In these methods the cluster continues to extend until the density in the neighbourhood crosses some threshold.

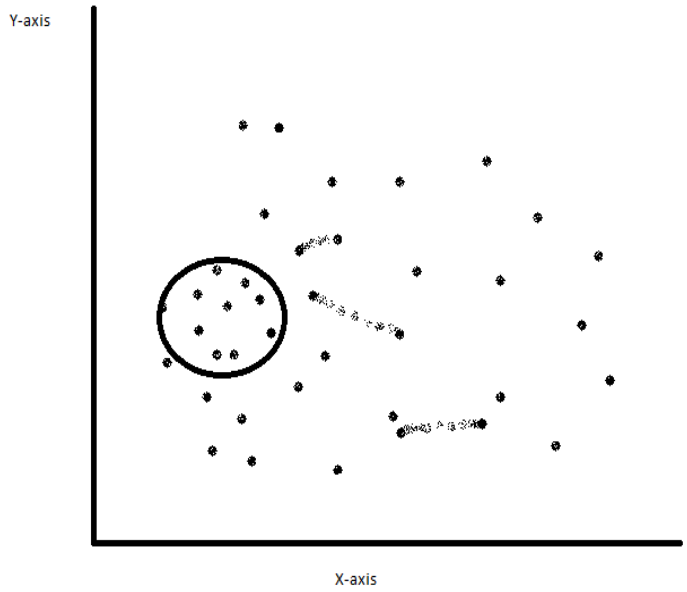

Fig1.4 DBSCAN

4. Hierarchical Methods:-This method is used to create a tree like structure by combining data objects. Based on hierarchical decomposition, it is subdivided into agglomerative hierarchical clustering or divisive hierarchical clustering. Agglomerative hierarchical clustering is the bottom up approach starts with each object that represents its own cluster. It then merges all objects till required cluster is achieved. Divisive approach is top down approach that starts with a large cluster and then subdivides the cluster in tiny clusters till required cluster is reached.

1.6 Procedure of Cluster Analysis

Cluster analysis is mainly divided into four basic steps as shown in Figure:

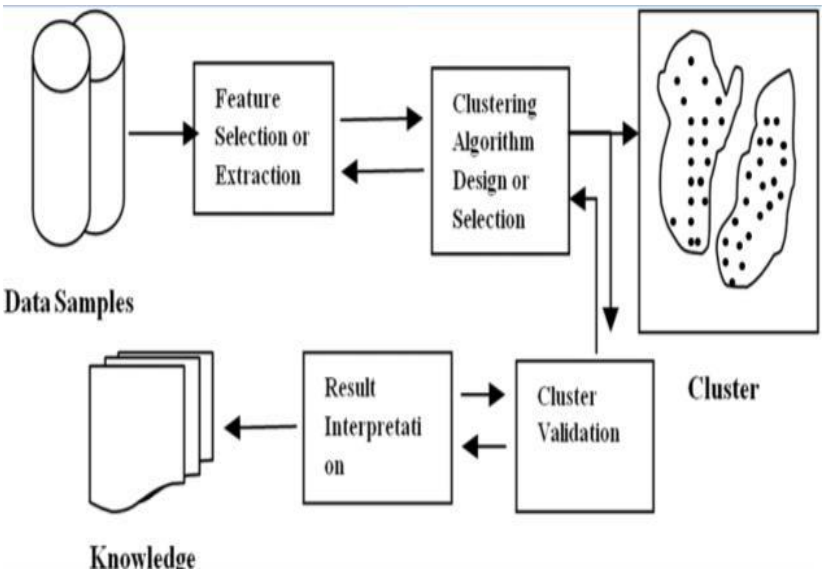

Fig 1.5: Clustering procedure steps

1 **Feature Selection or Extraction**

Feature selection is the process of discovering the most relevant attribute of a dataset to the data mining task. It is commonly used and powerful technique for reduction the dimensionality of a problem to more manageable task. Feature extraction uses some transformations to generate important and beneficial features from original data. It does not remove any of the original attribute.

2 Clustering Algorithm design and selection

In this step, the proximity (similarity or dissimilarity) measure and criterion function is selected. Proximity measure greatly affects the resulting clusters.

3 Cluster Validations

This step is used to validate that whether the clusters generated are of meaningful or just an artefact of an algorithm.

There are three methods of validating criteria:

i).External indices: based on prior knowledge and used as a basis to identify clustering solutions.

ii).Internal indices: independent or prior knowledge. The Clustering structure is evaluated directly from the original data.

iii).Relative criteria: compares different clustering structure to decide which one may best identify the features of the objects.

4 Result Interpretation

The motive of the clustering algorithm is to extract the important hidden information from the original dataset and to provide user with meaningful insights. The result should be easily interpretable and usable by the user.

1.7 K-Means Clustering Algorithm

The k-means clustering algorithm is one of the important and simpler partitioning based algorithms different from hierarchical algorithm such as divisive and agglomerative algorithm. K-means algorithm uses k as a parameter, divide x data items into k clusters so that the items in the one cluster are similar to each other but dissimilar to other items in other clusters. The algorithm was proposed by Mac Queen in the year 1967[5].It was introduced to solve various clustering problems. The algorithm aims to group data into k clusters based on randomly selected initial centroids. The grouping is done by minimizing the Euclidean distances between the data items and its related centroid. K-means algorithm itself is unsupervised and iterative in nature. The clusters generated by k-means are non-hierarchical in nature. The following example summarizes the k-means algorithms:

Consider arbitrary data set

Table 1.1

Student	P1	P2
1	1.0	1.5
2	2.0	2.0
3	3.5	4.0
4	4.0	4.5
5	4.5	5.0

1. Choose any random data as initial clusters

Table 1.2

No.	Student	Centroid
Cluster1	1	(1.0,1.5)
Cluster2	4	(4.0,4.5)

2. Assign objects to its nearest cluster by calculating mean

Table 1.3

	Cluster 1		Cluster 2	
Step	Student	Centroid	Student	Centroid
1	1	(1.0,1.5)	4	(4.0,4.5)
2	1,2	(1.5,1.7)	4	(4.0,4.5)
3	1,2,3	(2.2,2.5)	4	(4.0,4.5)
4	1,2,3	(2.2,2.5)	4,5	(4.25,4.75)

3. The two clusters now consist of following data

Table 1.4

No.	Student	Centroid
Cluster1	1,2,3	(2.2,2.5)
Cluster2	4,5	(4.25,4.75)

Assignments may not be correct, so now calculate Euclidean distance of each student from its own cluster and opposite cluster

Table 1.5

Student	Euclidean distance to centroid of cluster 1	Euclidean distance to centroid of cluster 2
1	1.6	4.6
2	.5	3.5
3	2.0	1.1
4	2.7	.3
5	3.4	.3

We see that the student 3 is nearer to cluster2 than its own cluster, so final assignment of data is

Table 1.6

No.	Student	Centroid
Cluster 1	1,2	(1.5,1.7)
Cluster2	3,4,5	(4,4.5)

Algorithm: The k-means algorithm is one of partitioning algorithm, in which each cluster's centre is represented by the mean value of the objects in the cluster.

Input:

K: number input clusters

D: represents data set containing n objects.

Output:

K clusters

Method:

(1) Choose randomly m objects from D as the initial cluster centres;

(2) **Repeat**

(3) Each object is re-assigned to the cluster based on the mean value to which the object is the most similar.

(4) Updation of cluster mean is done, i.e., the mean value of the objects for each cluster is calculated;

(5) **Until** no change;

The random selection of initial centroids has a large effect on the final result.

1.8 Example of k-means clustering in MATLAB

The fig 1.4 shows the formation of two clusters one in blue color and the other in red color each with its centroids

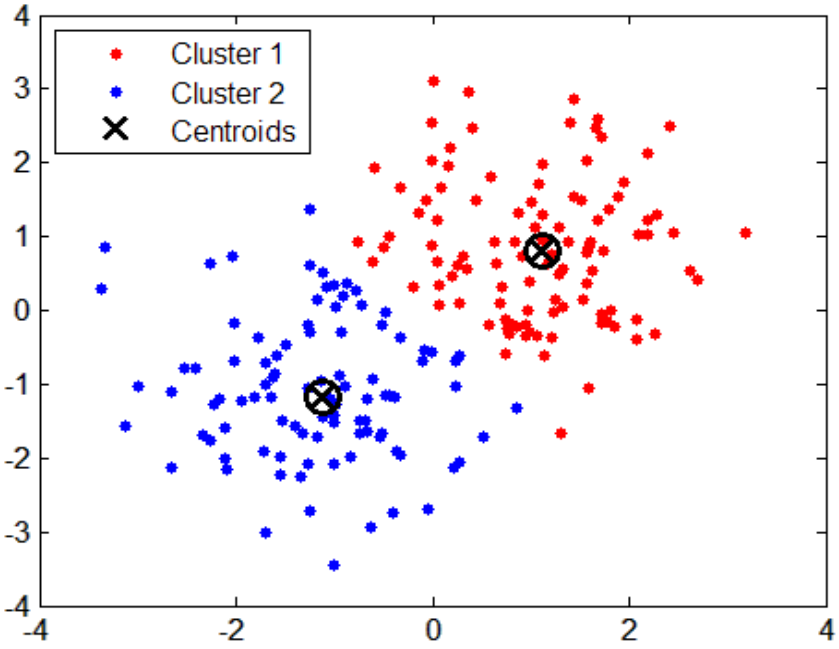

Fig 1.6 K-means clustering in MATLAB

1.9 Advantages of k-means

It is simple and robust

Comparatively K-Means is computationally faster than hierarchical clustering, if we keep k smalls and large number of data exists.

K-Means produce more compact clusters than hierarchical clustering, if the clusters are globular in shape.

More efficient than k-mediod

1.10 Disadvantages of k-means

The desired number of clusters as output is required to be initialized

Random selection of centroid

The k-means algorithm is sensitive to noise or outlier

Chapter 2
REVIEW OF LITERATURE

(Amar Singh, 2013)**-** Paper provides new method of Weighted Page Ranking algorithm to modify K-means algorithm to improve its accuracy and performance. The modified algorithm takes website link as input whose code is saved in text file. Then all the weights of in links and out links are computed and lastly weighted page algorithm is applied.

(Amanpreet Kaur Bhogal, 2010)This paper purposes idea of segmenting color image using k-means clustering algorithm. The procedure starts with randomly choosing k cluster centroid. Then each pixel in the image is assigned to the nearest cluster. Again the new centroids are computed and pixels are reassigned to the closest cluster till convergence criteria are met. This paper deals with providing explanation for image segmentation and produce good results.

(Anwit Jain, 2010) This paper presents modified k-means algorithm to improve efficiency and scalability for large data set. The tool used for implementation is Visual Studio .net. The purposed algorithm reduces the acceptance of cluster error criterion. Comparison is done between k-means algorithm, k-medoid and modified k-means algorithm. The factors used for comparison are various no. of records and execution time.

(Astha Joshi, 2013) This paper presents review of six different types of clustering techniques like partitioning method, hierarchical method, DBSCAN, grid method, STING, OPTICS. It was concluded that performance of K-means Clustering algorithm is much better than Hierarchical Clustering algorithm. On the other hand K-means algorithm is just restricted to numeric value. DBSCAN is used to find non-spherical shaped clusters whereas K-means and Hierarchical are used for spherical shaped clusters only.

(Chien-Hsing Chou, 2013) This paper presents an idea of line symmetry for discovering the line symmetry objects from image. The steps used are: firstly for the extaxtion of object

pixels from original image thresholding technique is applied. Followed by Fuzzy clustering for the labelling of object pixels is applied. Simulation results are used to define the performance.

(Dharmendra K Roy, 2010) In this paper, they introduced that clustering is defined as a process used for organizing/grouping a large amount of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters. The applications areas where clustering plays an important role are machine learning, image processing, data mining, marketing, text mining. This paper presents a clustering algorithm based on Genetic k-means algorithm is used for data that has features of both mixed numeric and categorical dataset. The performance of this algorithm has been studied on benchmark data sets.

(FAHIM A.M, 2006) proposes an efficient enhanced k-means algorithm to overcome problems in existing k-means. Original k-means is famous due to its ease, simplicity, speed of convergence and adaptability to sparse data. In spite of its large number of advantages, it suffers from certain disadvantages. These problem are the initialization of centroids, problem to converge to local minimum i.e updation of centroids till local minimum is not fount & execution of repeated while loops .All these problems are handled by the proposed k-means clustering algorithm. The enhanced algorithm firstly assigns datasets to its closest centroid and then computes distance with other centroids. In next step the two distances are compared and if the new distance is smaller than the previous distance then the data point is moved to new cluster otherwise if is small then it is assigned to same cluster. This process will save a lot of time and improve the efficiency. This algorithm uses two new functions .The first one is distance() function that is used to compute distance between each data point and its nearest cluster head. The second one is distance_new() function used to compute distance between data points and other remaining clusters. The experimental results show that the enhanced k-means algorithm is much fast and efficient than the original k-means.

(Huang) presented an extension to k-means algorithm named as k-modes algorithm. In k-modes algorithm mean is replaced by mode to improve clustering cost function. K-means algorithm was used to cluster only numeric data whereas k-modes is presented to cluster categorical data. Similar to k-means, k-mode algorithm was used to give local optimal solution .In k-mode algorithm the first step is used to select initial k-mode for each cluster. Next step is to allocate an object to the cluster whose mode is the nearer. Then updation operation is performed on all mode of the cluster. Third step is to reset the dissimilarity of objects against the current mode. Now if the object chosen is closer to some other cluster instead of its own, then reallocation of object is done. The results were evaluated using dataset on soyabean disease.

(Juntao Wang, 2011) Discuss an improved k-means clustering algorithm to deal with the problem of outlier detection of existing k-means algorithm. The proposed algorithm uses noise data filter to deal with this problem. Density based outlier detection method is applied on the data to be clustered so as to remove the outliers. The motive of this method is that the outliers may not be engaged in computation of initial cluster centers. In the next step fast global k-means algorithm proposed by Aristidis Likas is applied to the output generated previously. The results between k-means and improved k-means are compared using Iris, Wine, and Abalone datasets. The Factors used to test are clustering accuracy and clustering time. The disadvantage of the improved k-means is that while dealing with large data sets, it will cost more time.

(Liu Guoli, 2013) Presented algorithm named as K-means clustering based on iterative density known as **IDKM.** The algorithm was introduced to improve the dependence on initial value in existing K-means. **IDKM** applies continuous modifications to density threshold to get more number of clustering centers and thereafter combines them till the required number of clustering center is obtained. As the result **IDKM** reduces the time and space complexity when applied to IRIS data set for testing.

(Manpreet Kaur, 2013) In this paper introduces Query redirection (QR).Each time when a request is satisfied by more than one LTS, QR offers a way for BI Server to decide the set of logical table sources (LTS) appropriate to a logical request. Using BI applications

metadata content is contained in Oracle Fusion applications for real-time reporting analysis. QR technique is used to improve accuracy of K-means clustering algorithm. In this paper analysis is done by applying validation measures like entropy-measures, time etc. on k-mean algorithm and hierarchical algorithm. The comparison results show that k-mean algorithm gives better results as compared to hierarchical algorithm and also less execution time.

(Md. Sohrab Mahmud, 2012) Gave an algorithm to compute better initial centroids based on heuristic method. The newly presented algorithm results in highly accurate clusters with decrease in computational time. In this algorithm author firstly compute the average score of each data points that consists of multiple attributes and weight factor. Merge sort is applied to sort the output that was previously generated. The data points are then divided into k cluster i.e. number of desired cluster. Finally the nearest possible data point of the mean is taken as initial centroid. Experimental results show that the algorithm reduces the number of iterations to assign data into a cluster. But the algorithm still deals with the problem of assigning number of desired cluster as input.

(Navjot Kaur, 2012) Enhanced the traditional k-means by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness. The tool used to implement the improved algorithm is Visual Studio 2008 using C#. The advantages of k-means are also analysed in this paper. The author finds k-means as fast, robust and easy understandable algorithm. He also discuss that the clusters are non-hierarchical in nature and are not overlapping in nature. The process used in the algorithm takes student marks as data set and then initial centroid is selected. Euclidean distance is then calculated from centroid for each data object. Then the threshold value is set for each data set. Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid. The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

(Purohit, 2013) Proposed an improved approach for original K-means clustering algorithm due to its certain limitations. The main reason for poor performance of K-means algorithm is

selection of initial centroids randomly. The proposed algorithm deals with this problem and improves the performance and cluster quality of original k-means algorithm. The new algorithm selects the initial centroid in a systematic manner rather than randomly selecting. It first find out the closest data points by calculating Euclidian distance between each data point and then these points are deleted from population and forms a new set. This step is repeated on new set by finding data points that are closest to each other. Performance comparison is done using Matlab tool. The proposed algorithm gives more accurate results and also decreases the mean square distance. But the proposed algorithm works better for dense dataset rather than sparse.

(Raju G, 2008) Gave a comparative analysis between k-means clustering algorithm and fuzzy clustering algorithm. In this paper the researcher also discuss the advantages and limitations of fuzzy c-means algorithm. K-means is a partional based clustering algorithm whereas Fuzzy c-means is non partional based clustering algorithm. Fuzzy c-means mainly works in two processes. In the first process cluster centers are calculated and in second the data points are assigned to calculated cluster center with the help of Euclidean distance. This process is almost similar to conventional k-means with a little difference. In fuzzy c-means algorithm membership value ranging from 0 to 1 is assigned to data item in cluster.0 membership represents the degree that the data point is not a member of cluster whereas 1 indicates the degree to which data point represents a cluster. The problem faced by fuzzy c-means algorithm is that the sum of membership value of data points in each cluster is restricted to 1.Algorithm also face problem in dealing with outliers. On the other hand comparison with k-means shows that the fuzzy algorithm is efficient in obtaining hidden patterns and information from natural data with outlier points.

(Sanjay Garg, 2006) Compared different existing variants of K-means clustering algorithm. The algorithms used for comparison were K-means, K-mediod ,h-k-mean. Whereas k-mean and k-mediod are partitional clustering algorithms, h-k-mean is heuristic based hybrid model of the other two algorithms. For k-mediod, PAM is used for comparison. Comparison is done using different factors i.e. average running time, average distance of points in each cluster etc.On comparison it was found that h-k-mean is better that the other two algorithms.

(Shi Na, 2010) extends existing K-means algorithm by introducing two simple data structures to store the labels of cluster and distance of all the data objects to the nearest cluster during each iteration to use in next iteration and so on. Since existing k-means algorithm calculates distance from each data object to all the centers of k clusters when it executes the iteration each time, the execution time increases. This improvement was done to improve the execution time, speed of clustering and accuracy, reducing the computational complexity of the k-mean.

(Shunye, 2013) Motivated by the problem of random selection of initial centroid and similarity measures, the researcher presented a new K-means clustering algorithm based on dissimilarity. This improved k-means clustering algorithm basically consists of 3 steps. The first step discussed is the construction of the dissimilarity matrix i.e. dm. Secondly, Huffman tree based on the Huffman algorithm is created according to dissimilarity matrix. The output of Huffman tree gives the initial centroid. Lastly the k-means algorithm is applies to initial centroids to get k cluster as output. Iris, Wine and Balance Scale datasets are selected from UIC machine learning repository to test the proposed algorithm. Compared to traditional k-means the proposed algorithm gives better accuracy rates and results.

(Shuhua Ren, 2011) Elaborates k-means clustering algorithm based on coefficient of variation. The coefficient of variation is defined as ratio of standard deviation to the mean value. Existing k-means algorithm uses Euclidean distance as the similarity metric which gives inaccurate results due to the effect of useless data. To overcome with this problem, proposed algorithm uses coefficient of weight factor to elicit the effect of ouliers. Weight values are assigned to all the features in clustering to remove irrelevant, noisy data so as to increase cluster quality. The results are evaluated using popular data sets i.e. Iris, Wine and Balance scale. The results prove that the modified algorithm presents more clustering accuracy and the numbers of iterations required for clustering are less than original k-means. The problem faced by proposed algorithm is that the numbers of clusters required as output are needed to be initially defined.

(Sing, 2011)This paper purposed a modified K-means algorithm to deal with problem of initial selection of centroid. The algorithm works in two steps. In Step 1segmentation on data set is applied. In step 2 frequencies is calculated of data point in each segment. Finally centroid is selected from segment that has the highest frequency. Comparison shows that purposed algorithm reduces the complexity and data is assigned to the cluster more effectively.

Chapter 3
PRESENT WORK

3.1 Problem Formulation

Data Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term DM is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Clustering plays a crucial role in various application areas. DM applies algorithms to large data to produce models or patterns that are useful for user and also extract hidden pattern. Partitioning method is simple and most essential version of cluster analysis. Partitioning method results in a set of K clusters, each cluster contain at least one object. The generally used efficient clustering algorithm is k-means clustering. The k-means clustering algorithm is one of the important and simpler partitioning based algorithms different from hierarchical algorithm such as divisive and agglomerative algorithm. Hierarchical algorithm method is used to create a tree like structure by combining data objects. K-means algorithm uses k as a parameter, divide x data items into k clusters with the intention that the items in the one cluster are similar to each other but dissimilar to other items in other clusters The k-means algorithm suffer from large number of limitations such as : problem of cluster initialization, cluster quality and efficiency of algorithm, iterations etc. . The main drawback of k-means algorithm is dependency on the random selection of initial centroids. This selection has a direct impact on the efficiency and accuracy of clusters. With increase in cluster quality the results of pattern recognition may improve. The proposed hypothesis will enhance the K-means clustering with the help of HMAC, which will increase the cluster quality and increase efficiency.

3.2 Objectives of Problem

i**.** To study various improved clustering algorithms in data mining and to identify the problem of cluster quality.

ii. To propose enhancement in K-mean clustering algorithm to increase cluster quality and efficiency of the algorithm

ii. To implement proposed and enhanced algorithms and evaluate the results on basis of cluster quality and efficiency in Matlab.

3.3 Methodology

3.3.1Basic Design of Existing Work

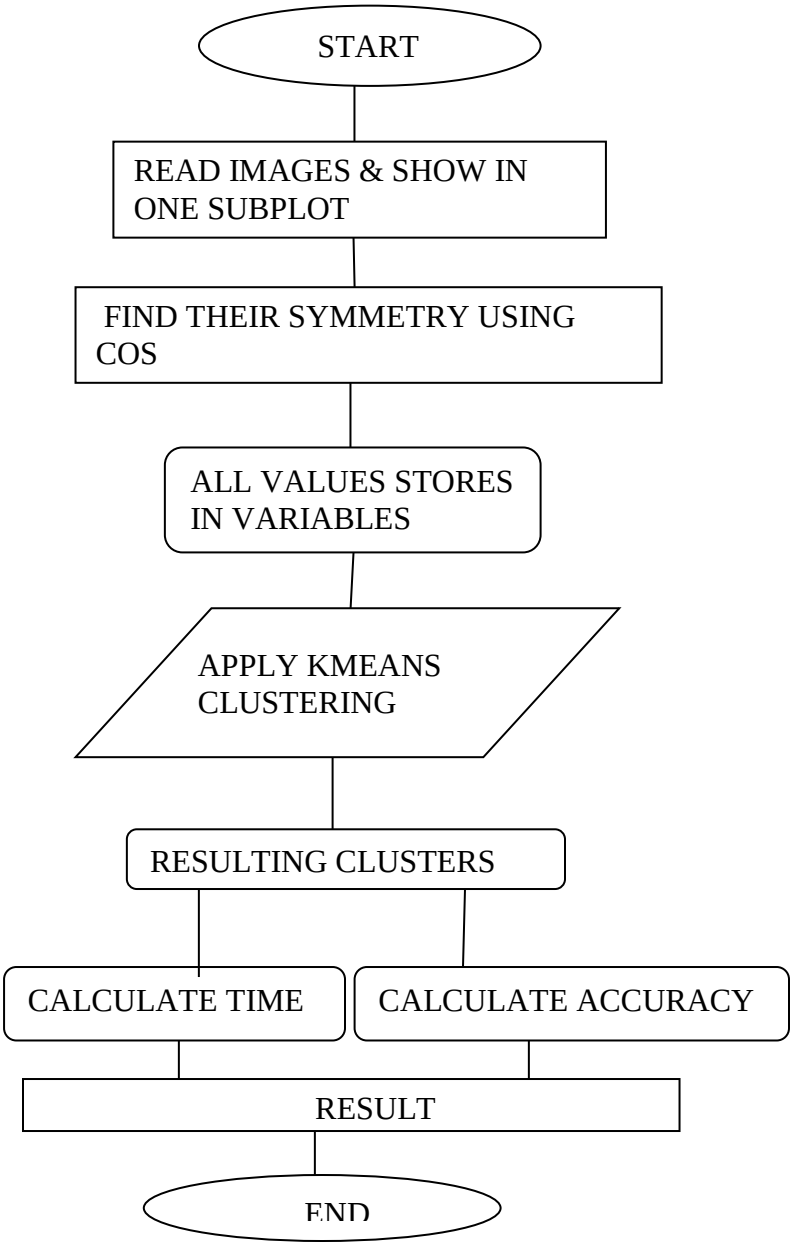

Fig 3.1: Flow Chart of Basic K-means algorithm

As illustrated in figure 3.1,

1. The first step is to read input images and show in one subplot rather than in different figures.
2. After input we find their symmetry using mathematical sign COS because if some time symmetry is zero and we know cos0=1, in the symmetry function all direction of images values will store in different variables.
3. After that we apply k-means function in which we cluster all the values of image and related variables.
4. Then we calculate time and accuracy of that process.

3.3.2 Basic Design of Proposed Work

The k-mean clustering algorithm is one of the simplest partitional clustering algorithms that is used to cluster huge amount of data in DM. The main problem exists in k-mean clustering algorithm is of efficiency and accuracy. The time required for data cluster is high and cluster quality is not so good. To improve cluster quality and to reduce time of the algorithm Hierarchal mode association clustering (HMAC) will be used in the enhancement.

1) The first step is to read input images and combine them in one figure.
2) After input we find their symmetry using mathematical sign COS because if some time symmetry is zero and we know cos0=1, in the symmetry function all direction of images values will store in different-2 variables.
3) Then segmentation of image into 3*3 parts is done.
4) After that k-means function is applied in which we cluster all the values of image and related variables.
5) At last HMAC is applied to further cluster the data. Then we calculate time and accuracy of that process.

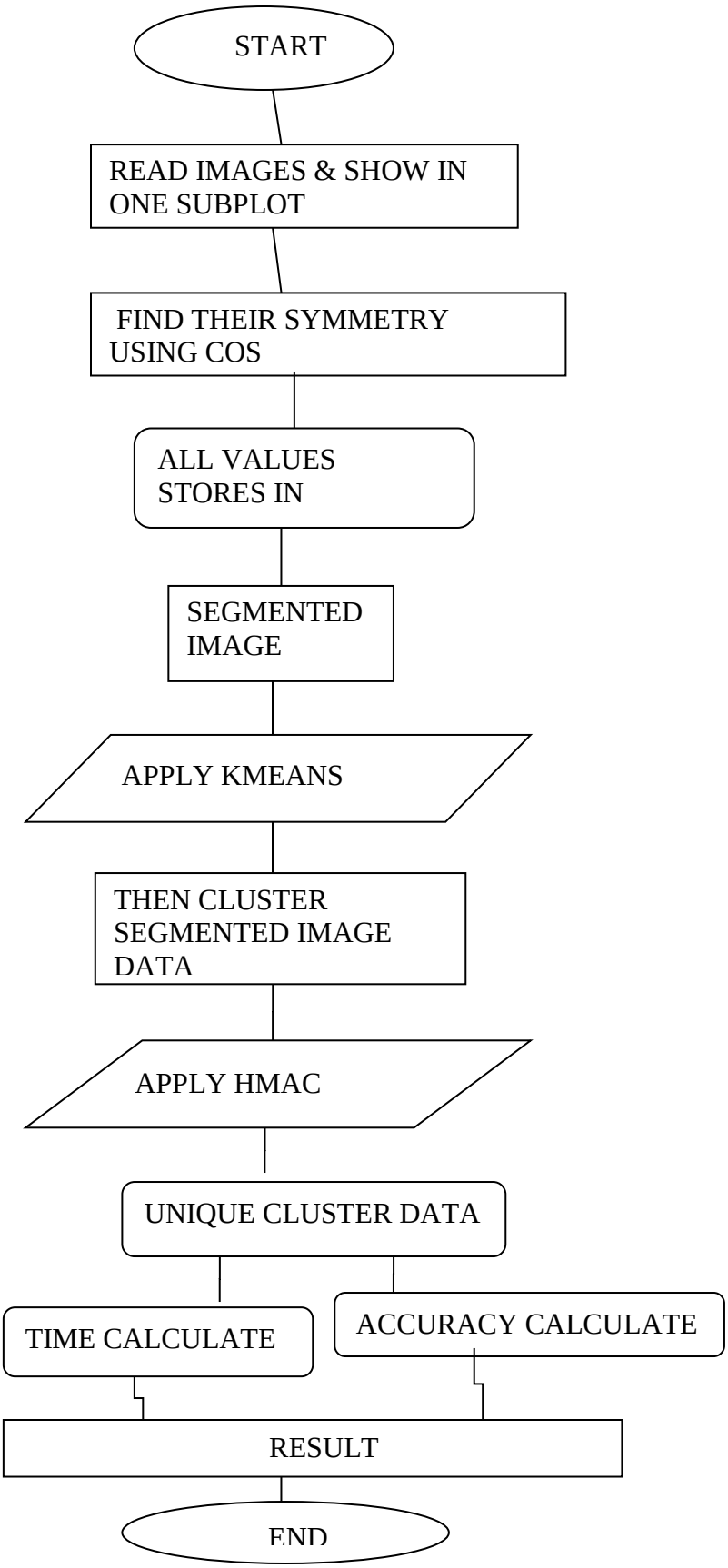

Fig 3.2: Flowchart of proposed technique

Algorithm of Proposed technique

Input: First image and second image

Output: n Cluster

i. Find symmetric of the two images and merge two input images and store data in matrix a
ii. Define n clusters, define centre point in each cluster which is defined as central=(1 to n)
iii. Apply segmentation on matrix a and store in matrix 1 to n1 (n1 is the number of segments) and define central point in each segment which is defined as central 1 to central n
iv. Apply Euclidian distance and find distance (r,n) from the matrix a
v. Find hierarchy of the similar elements in the distance (r,n) matrix and store in b matrix
vi. Find uniqueness in each segment and store similar values in 1 to n matrix n is the number of clusters
vii. Plot similar n number of clusters

Chapter 4

RESULTS AND DISCUSSION

4.1Experimental Results

4.1.1Existing algorithm results

MATLAB tool is used to implement the presented work and proposed work. MATLAB tool is a relevant tool for data mining because it consists of number of toolboxes. Large part of coding is shorten in MATLAB because it provides a lot of toolboxes for data mining. MATLAB is also a beneficial tool at the time of data manipulation and optimization.

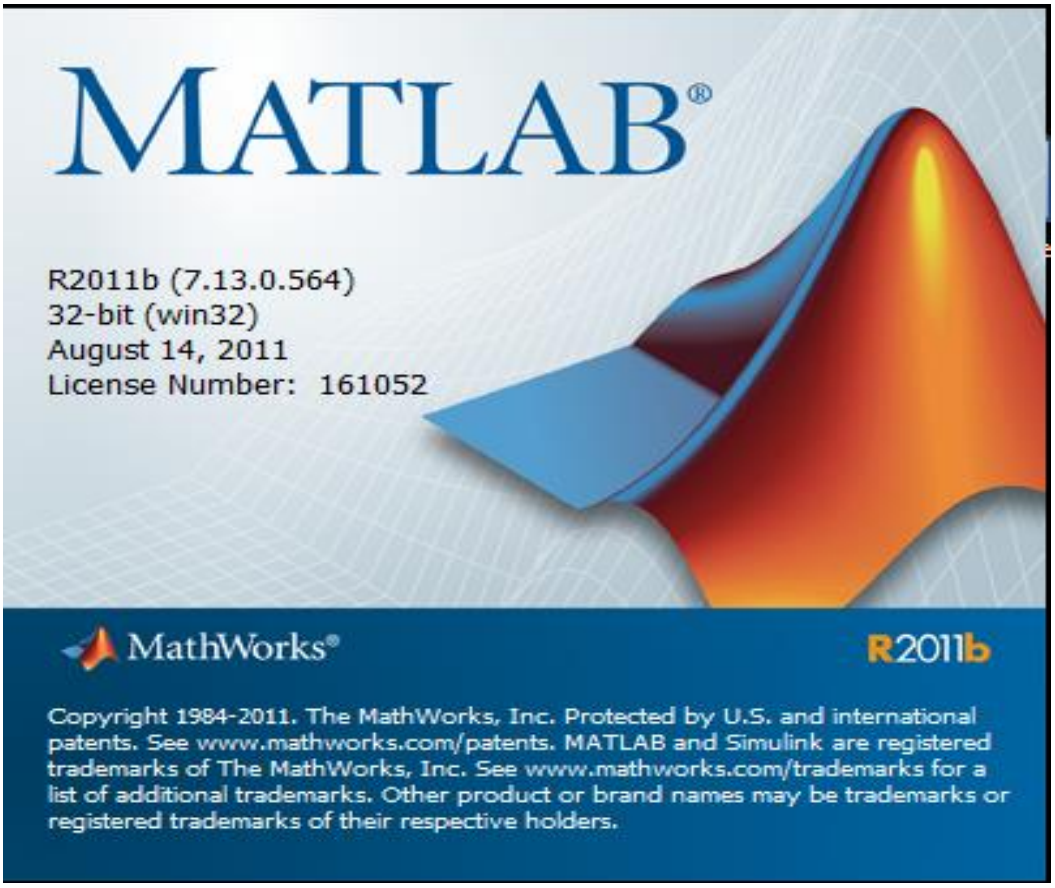

Fig 4.1: Matlab Tool

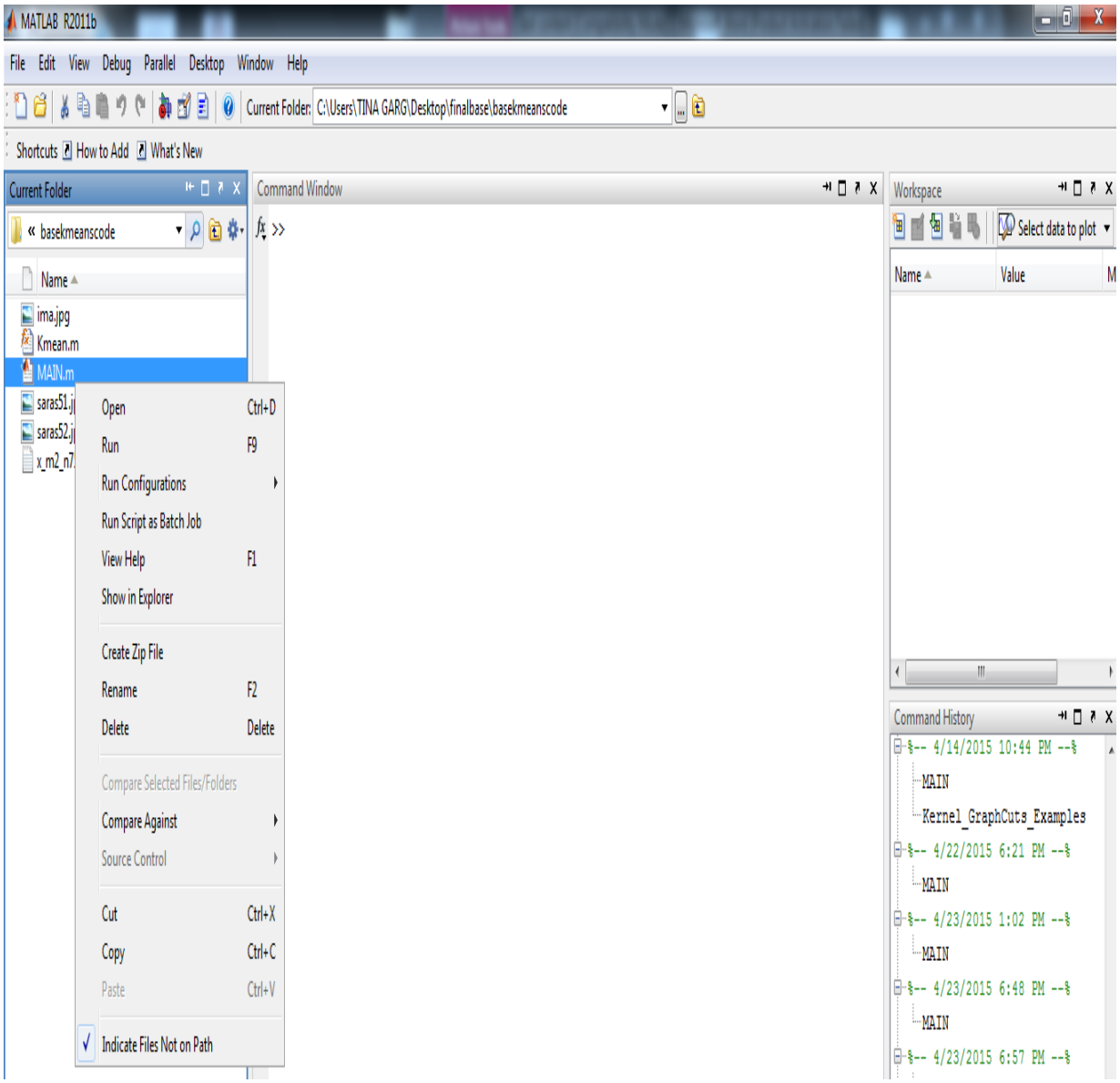

Fig 4.2: Implementing K-means Algorithm

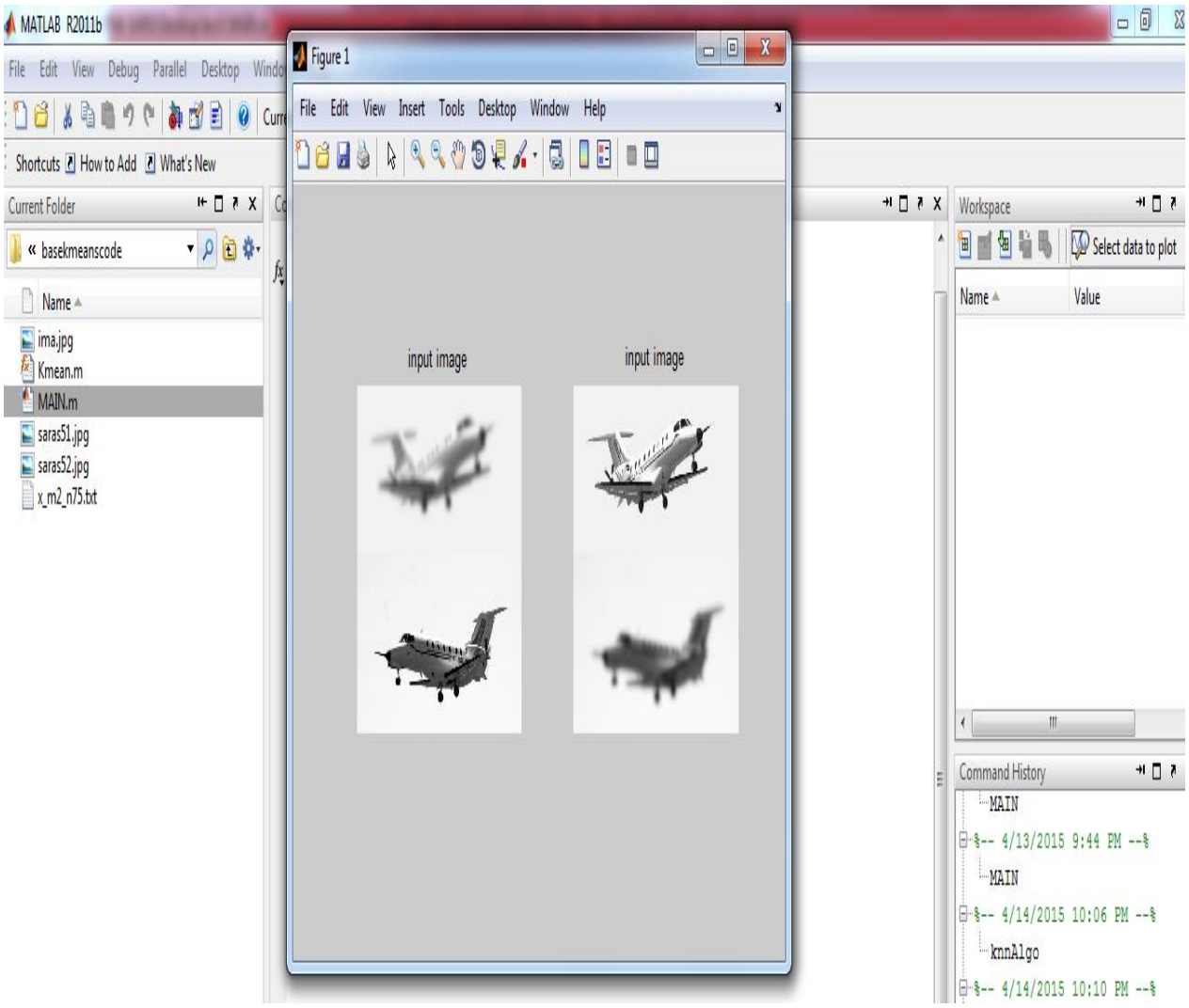

Fig 4.3: Two images are taken for clustering

As illustrated in figure 4.3, two different images are taken and these two images are adjusted according to sized and combined together

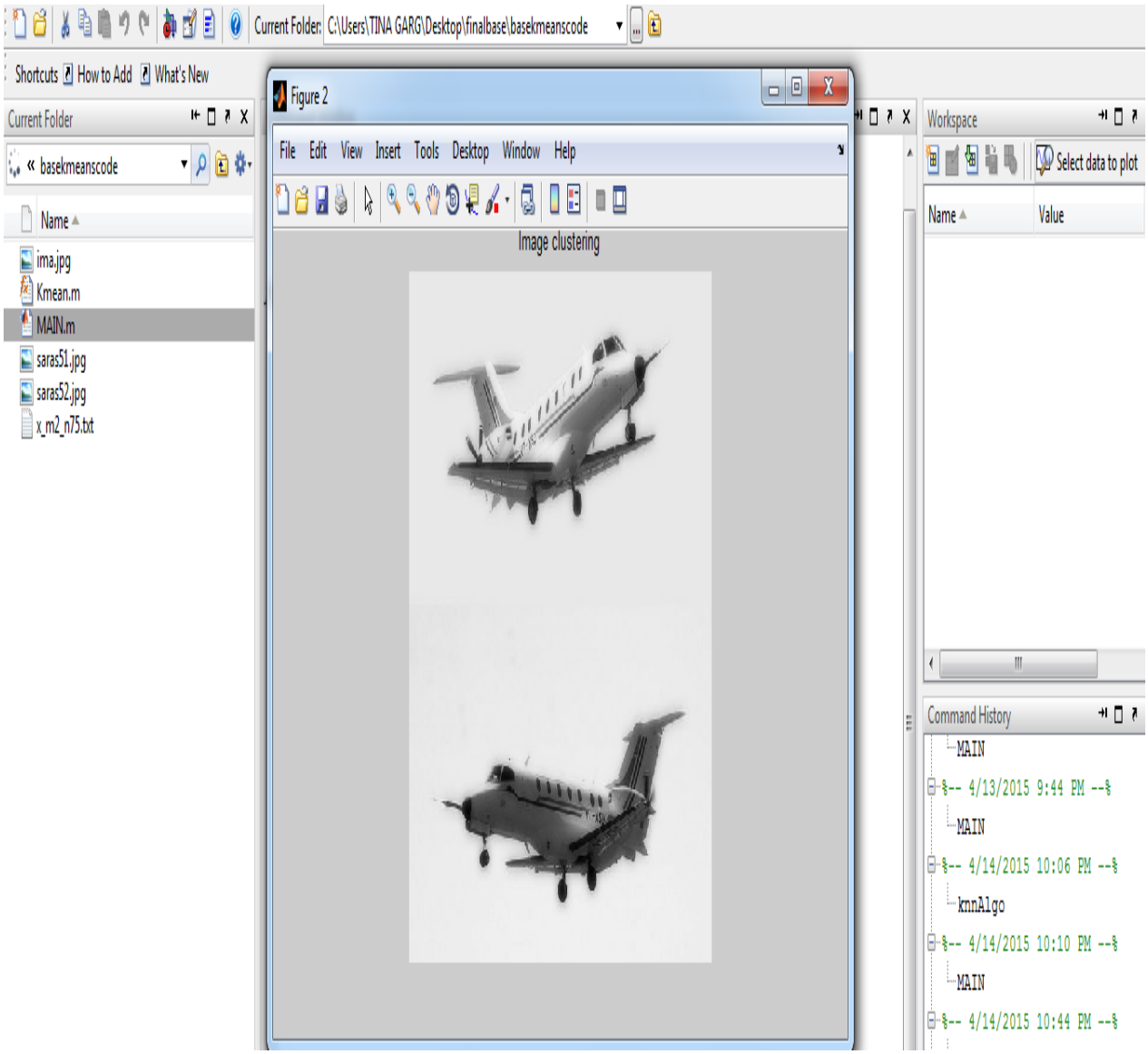

Fig 4.4: Combined Image

As illustrated in figure 4.4, the first two images are combined into single image is shown whose pixel values are to be clustered.

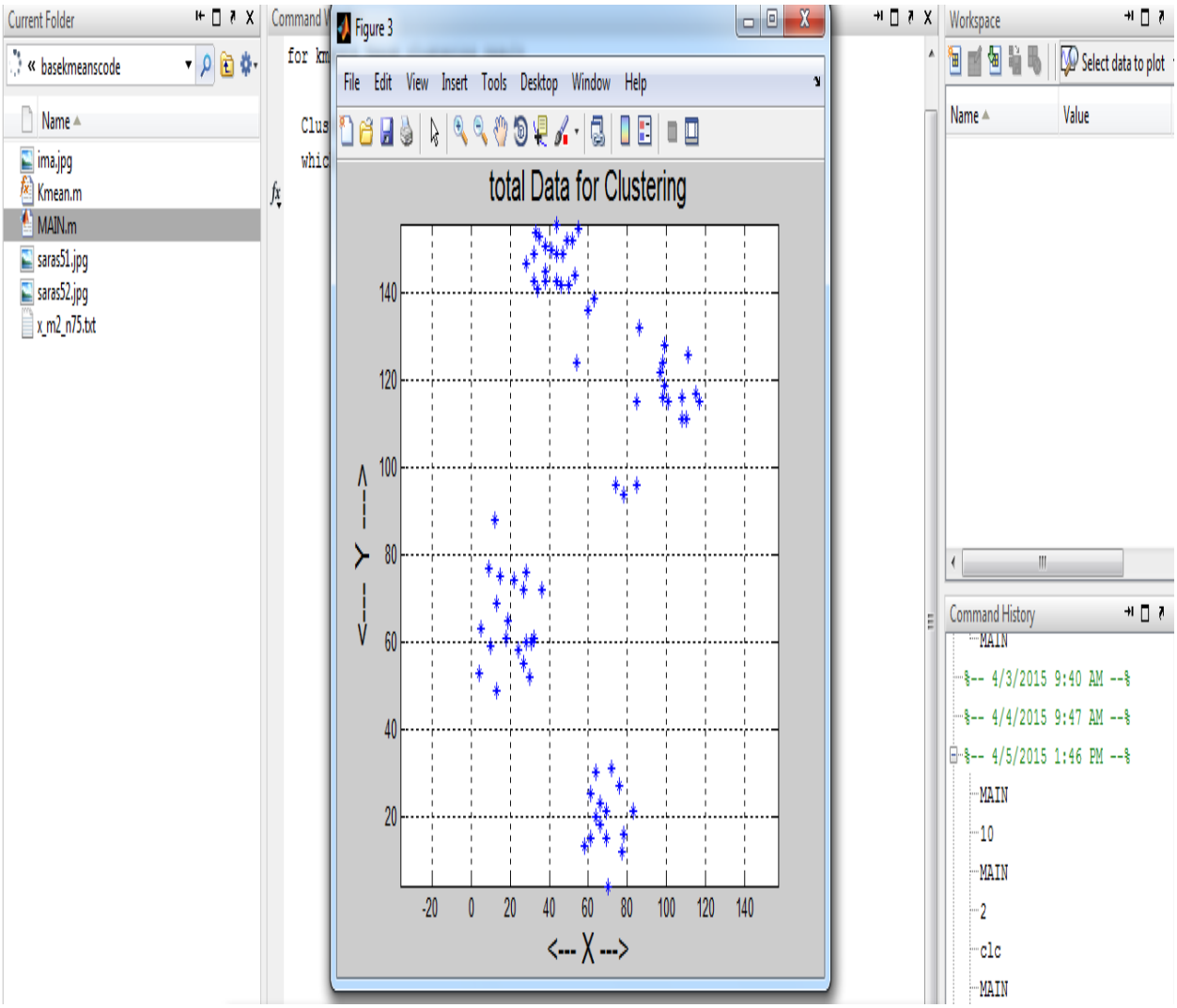

Fig 4.5: Data for Clustering

As shown in the figure 4.5, the pixel values of the combined image are stored in the text file. The stored information is clustered and shown on the 2d plane.

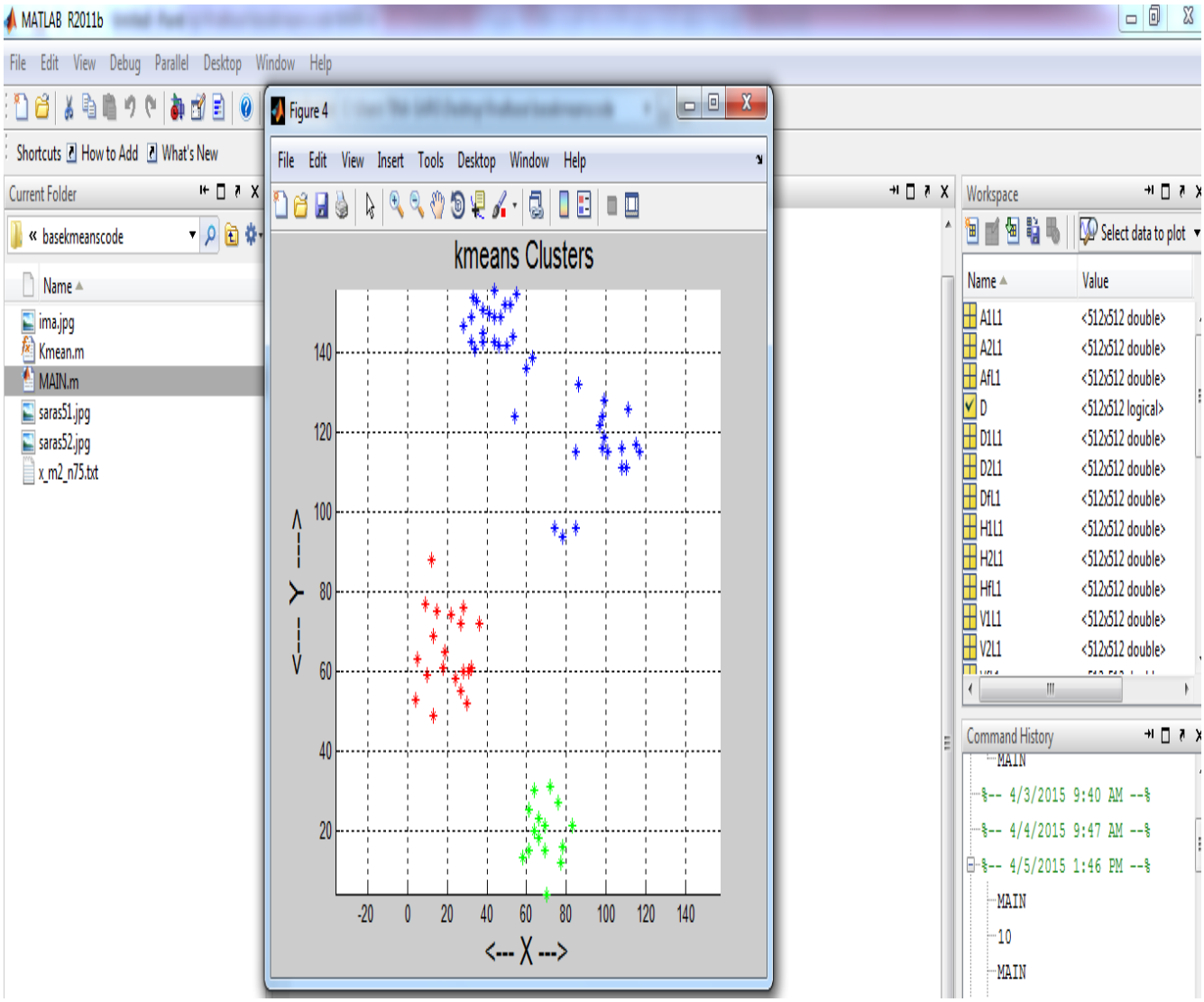

Fig 4.6: Clustered data are marked

As shown in the figure 4.6, k-mean clustering is applied on the pixel values of the combined image for clustering. The clustering data is shown on the 2 D plan and each cluster is marked with different colors.

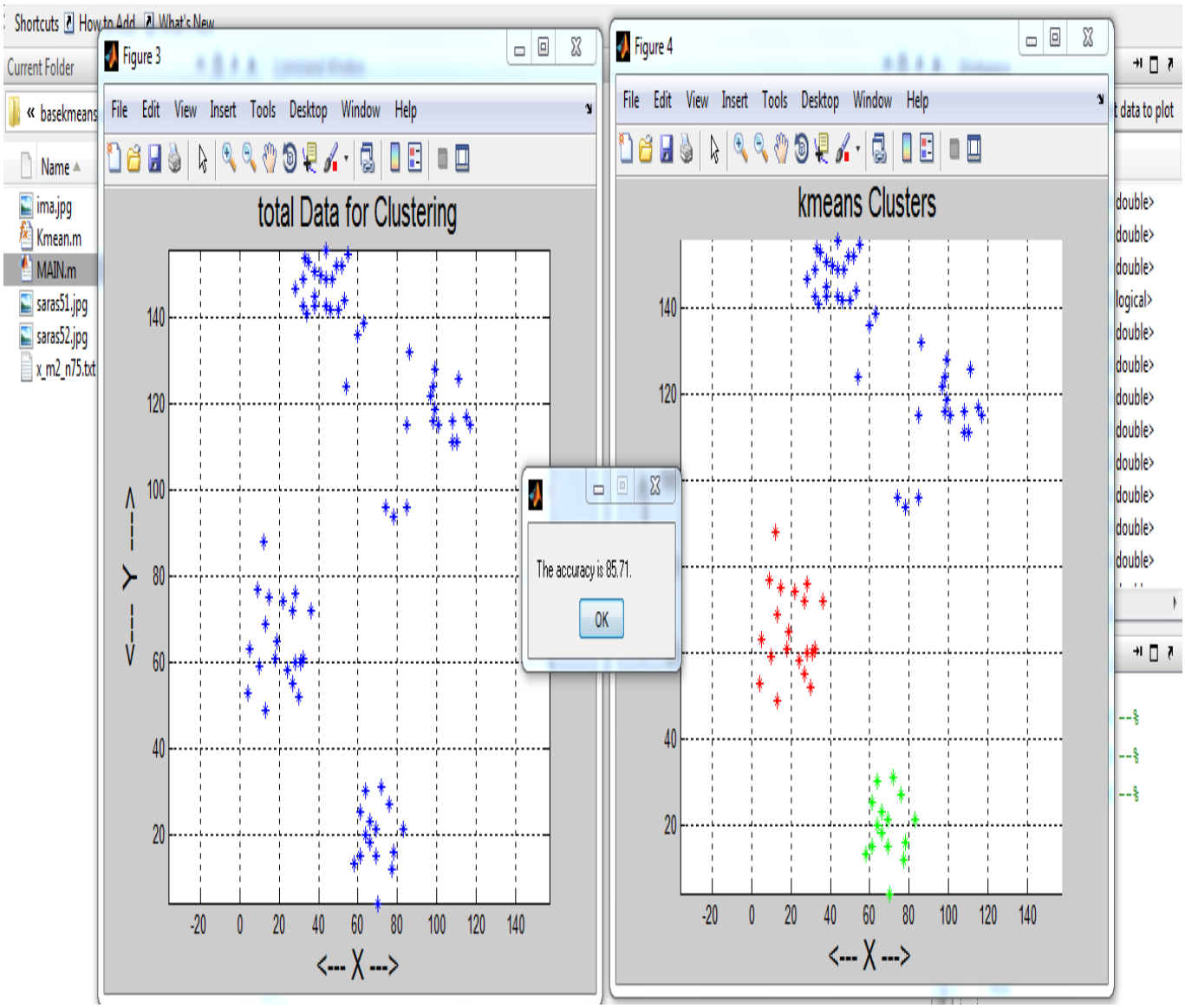

Fig: 4.7 Accuracy of K-means Algorithm

As shown in the figure 4.7, the accuracy is calculated and shown in the message box.

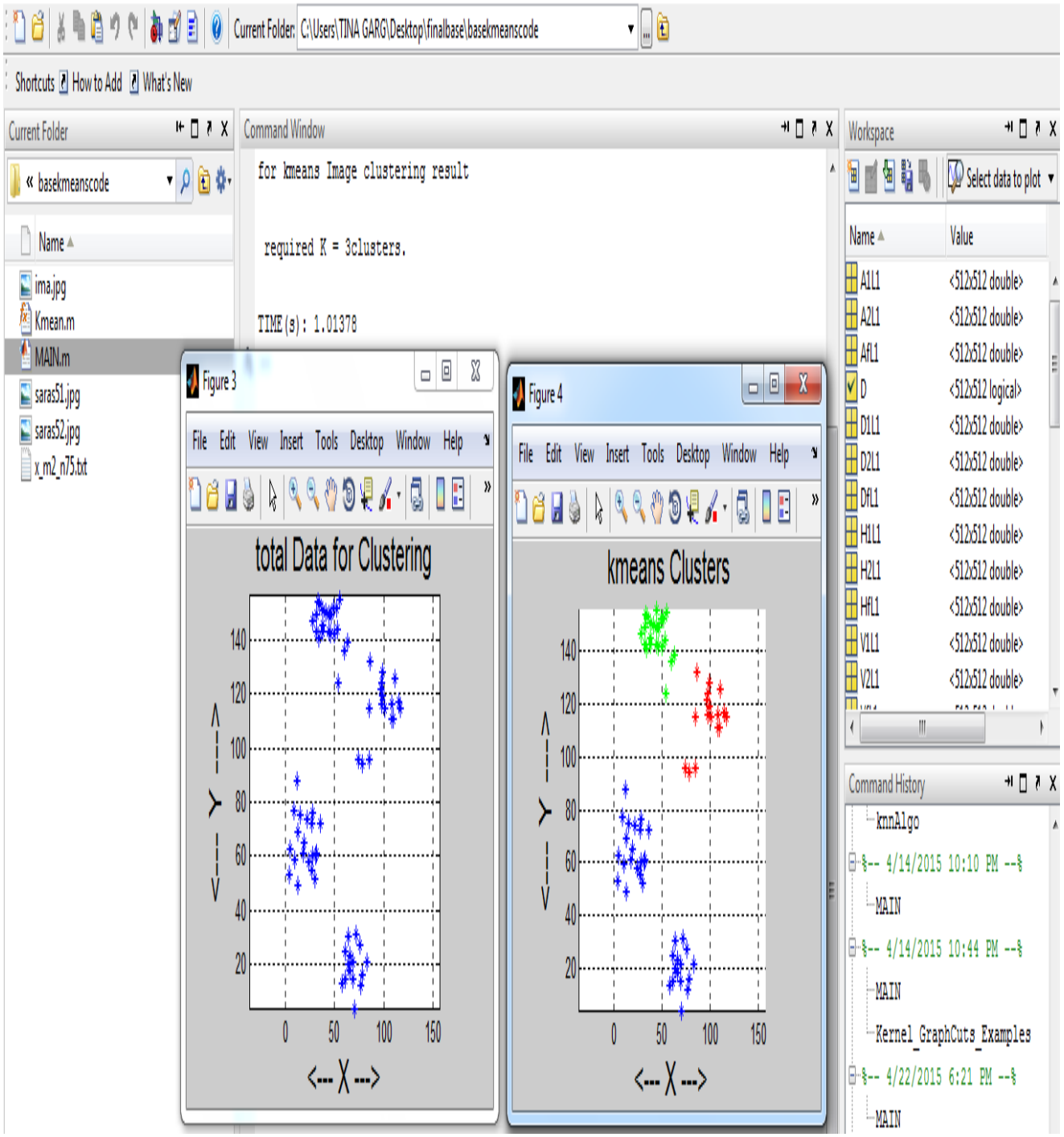

Fig 4.8: Time Taken by K-means Algorithm

As shown in the figure 4.8, the time for clustering is calculated and is shown on the command window.

4.1.2Proposed algorithm results

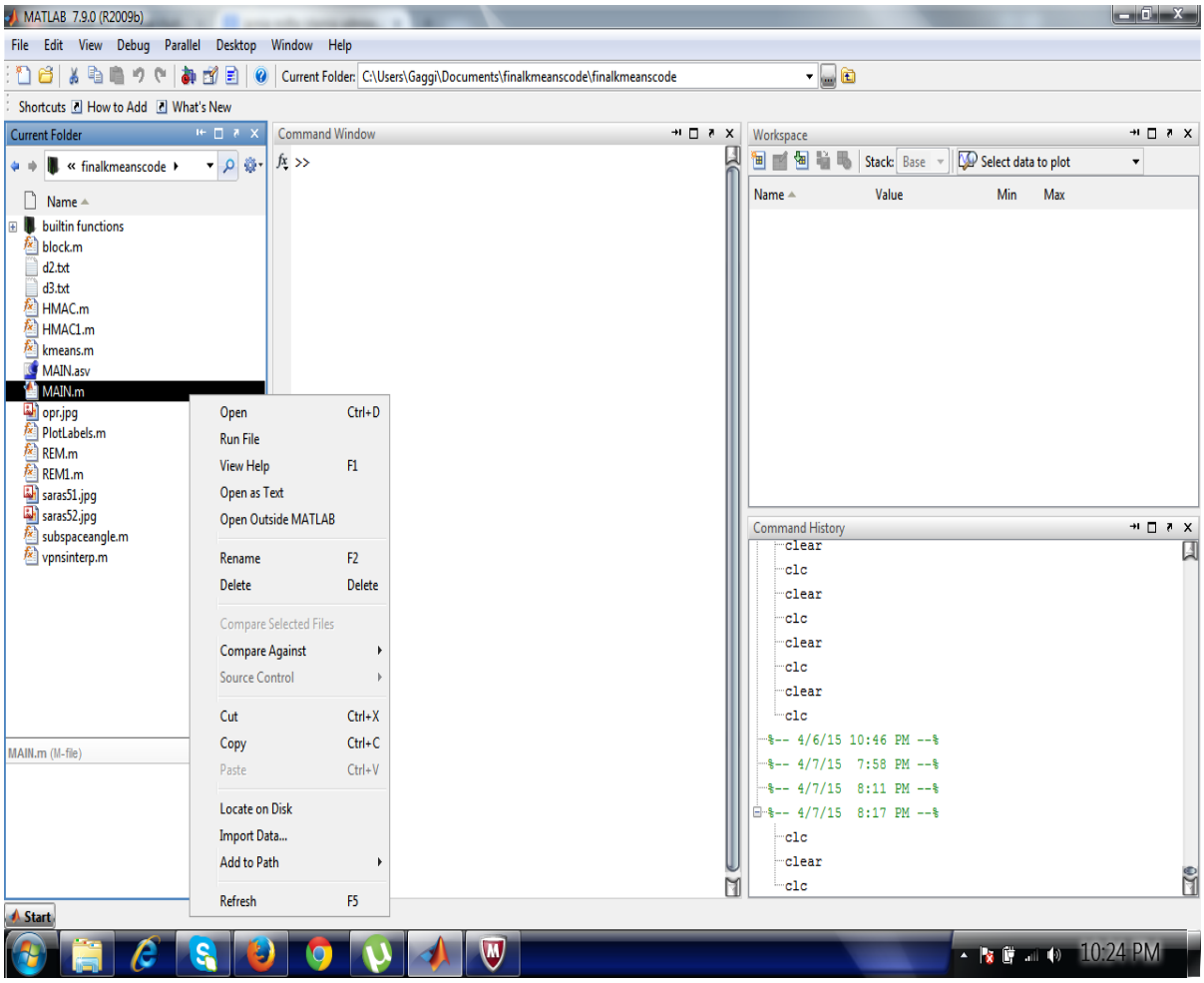

Fig 4.9 Implementing Purposed Algorithm

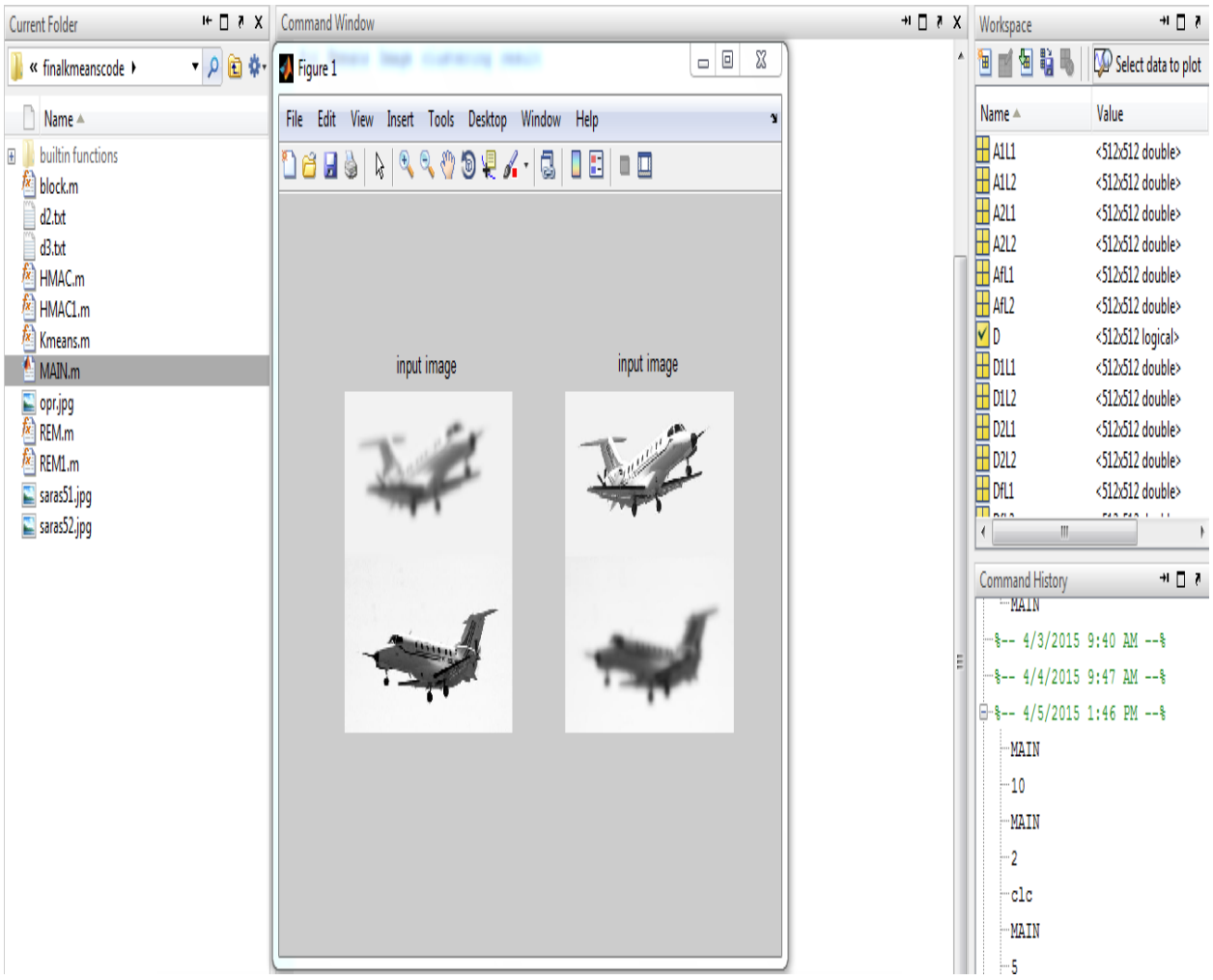

Fig 4.10: Two mages are taken for clustering

As illustrated in figure 4.10, two different images are taken and these two images are adjusted according to sized and combined together

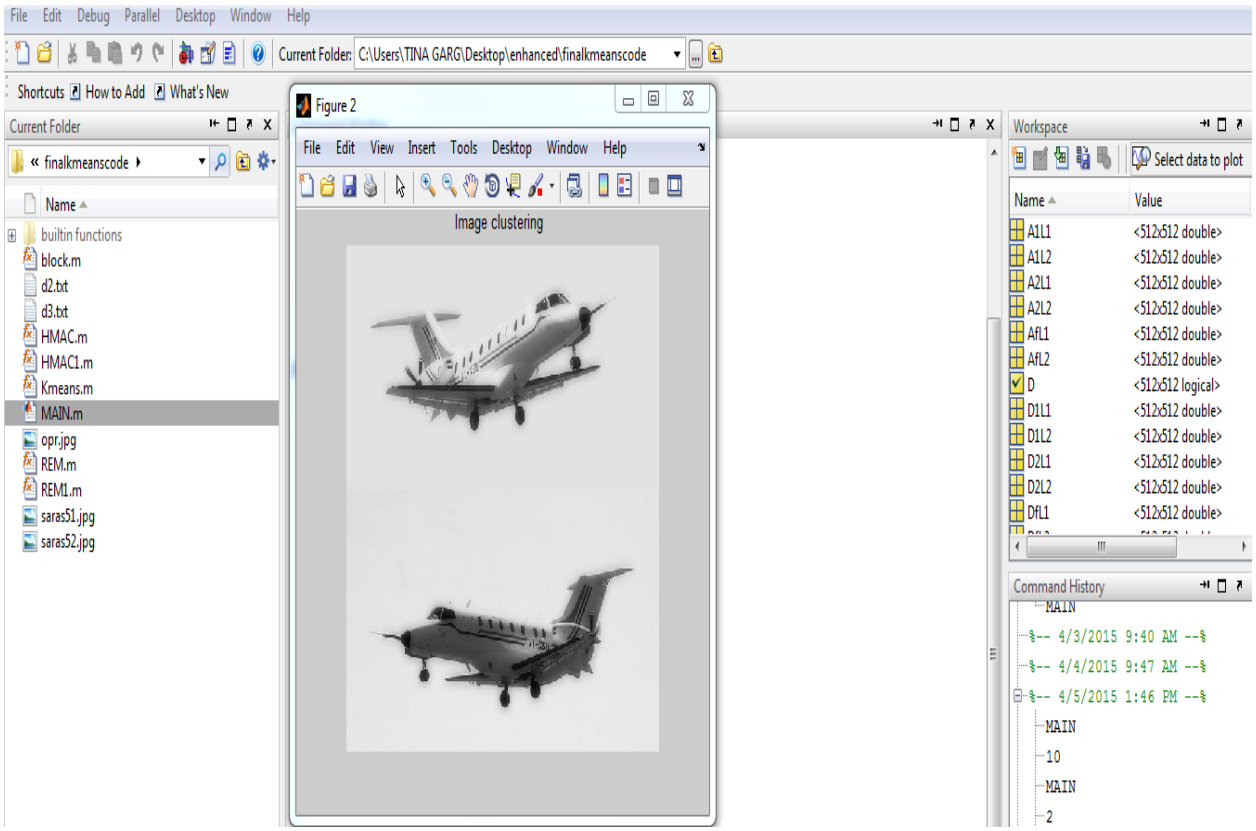

Fig 4.11: Combined image

As illustrated in figure 4.11, combined image is shown whose pixel values are to be clustered

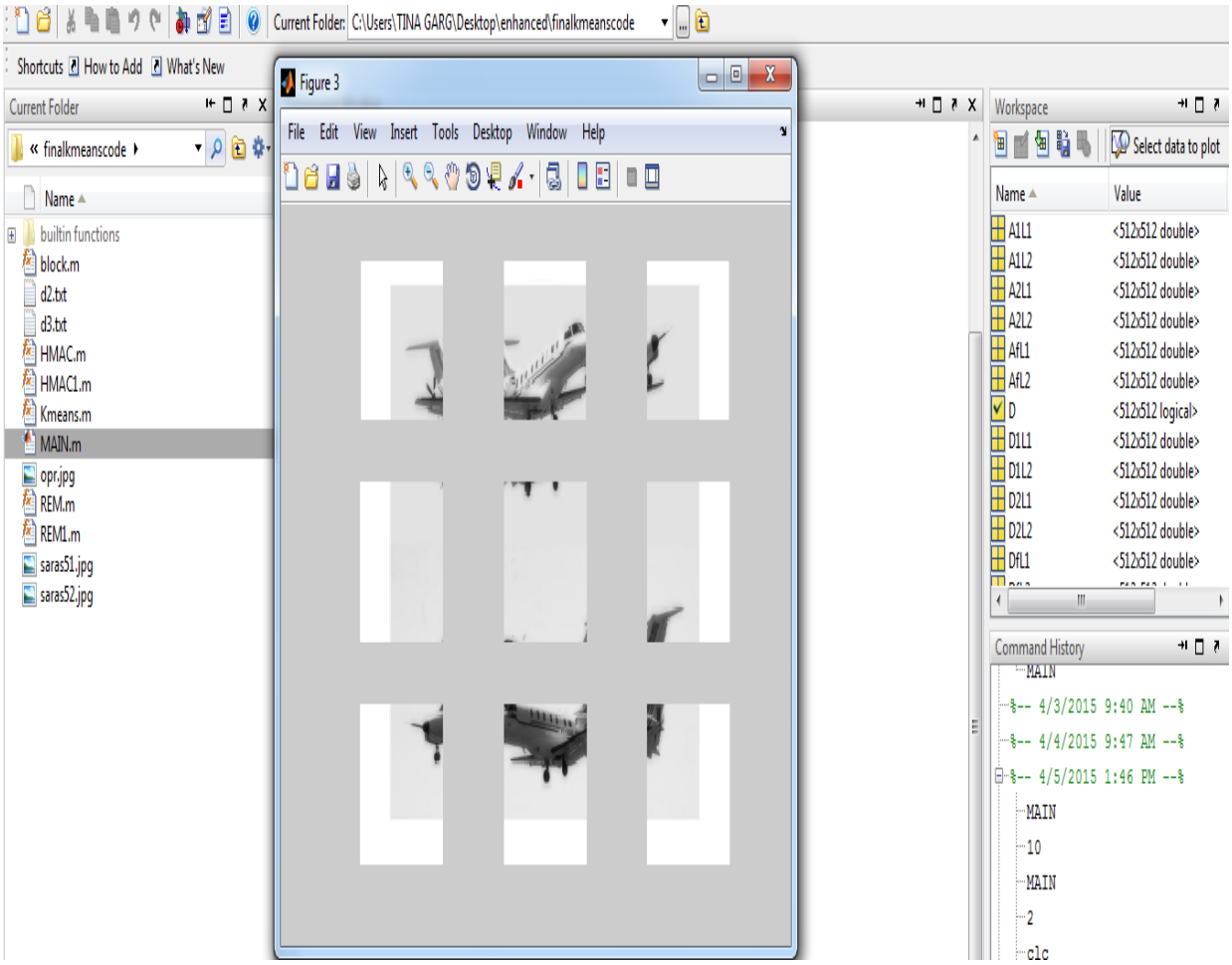

Fig 4.12: Segmentation of the image

As illustrated in the figure 4.12, the combined image is divided into nine segments. The nine numbers of clusters are formed and in each cluster, central point is calculated

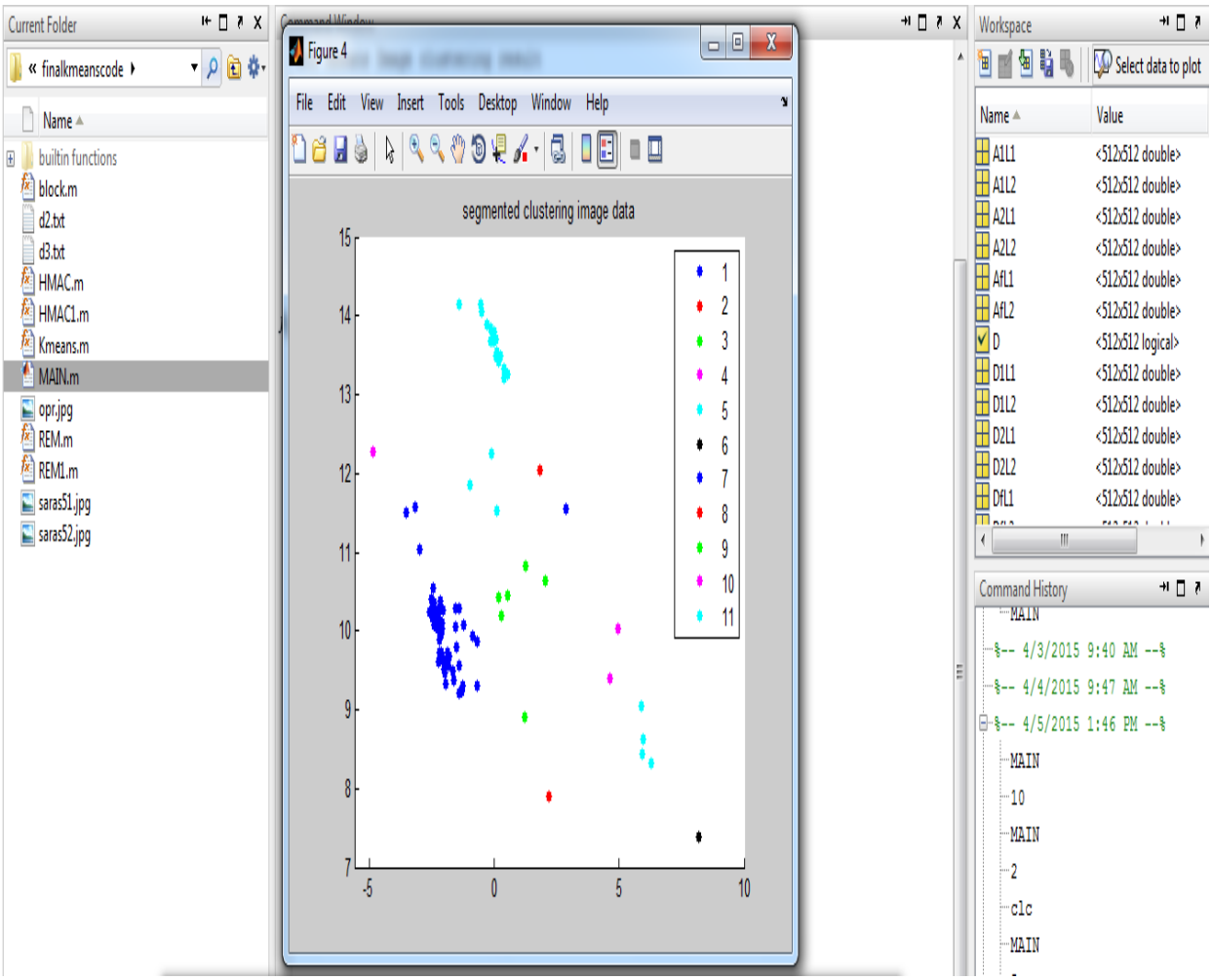

Fig 4.13: Formation of clusters

As shown in figure 4.13, the combined image is segmented into nine parts. At each segment central point is find out and each cluster is shown with the different color

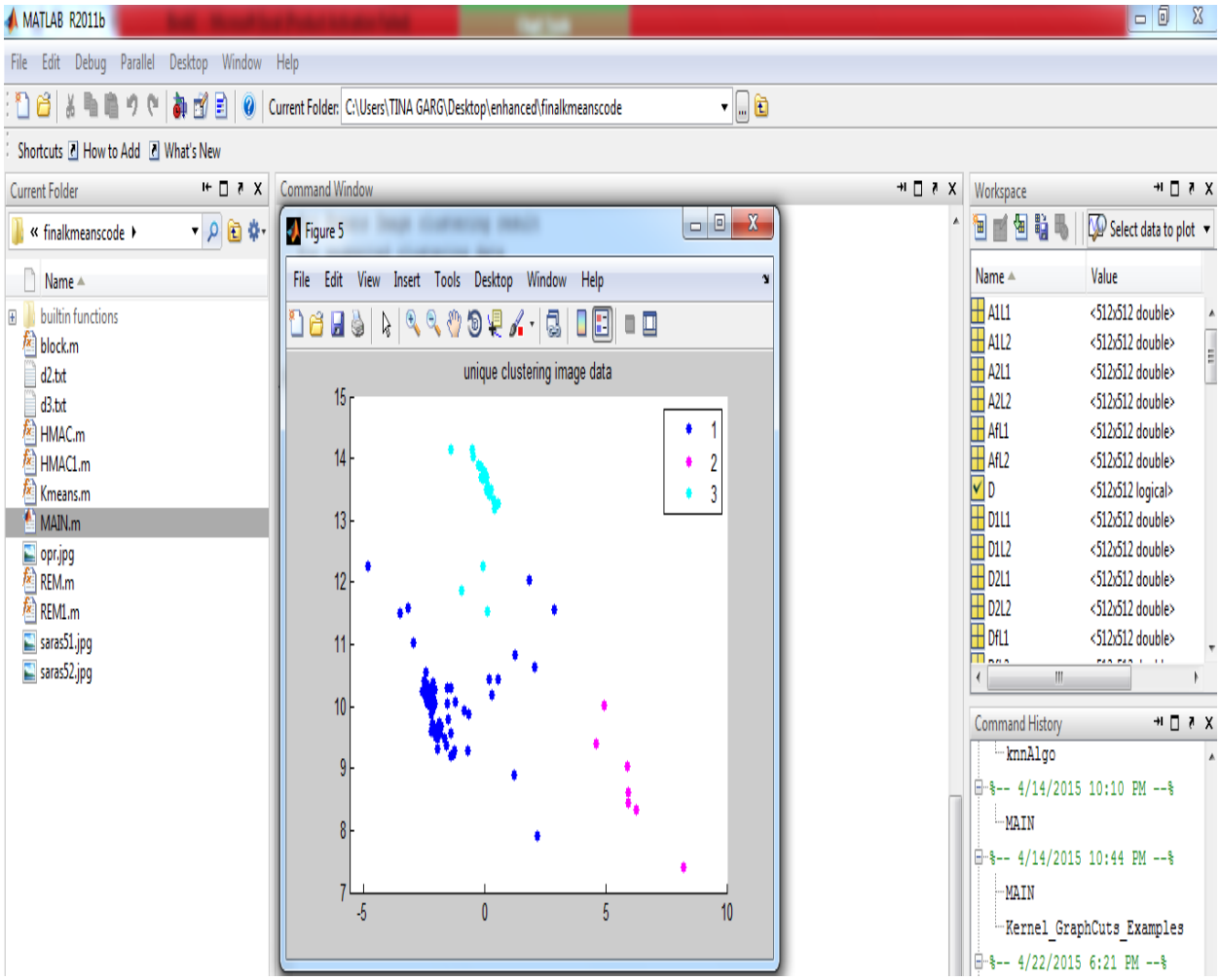

Fig 4.14: Final clustering result

As shown in figure 4.14, the nine numbers of segments are formed and in each segment central point is calculated. The uniqueness is find out in between central points are on the basis of uniqueness final clustering result is shown on 2d plot.

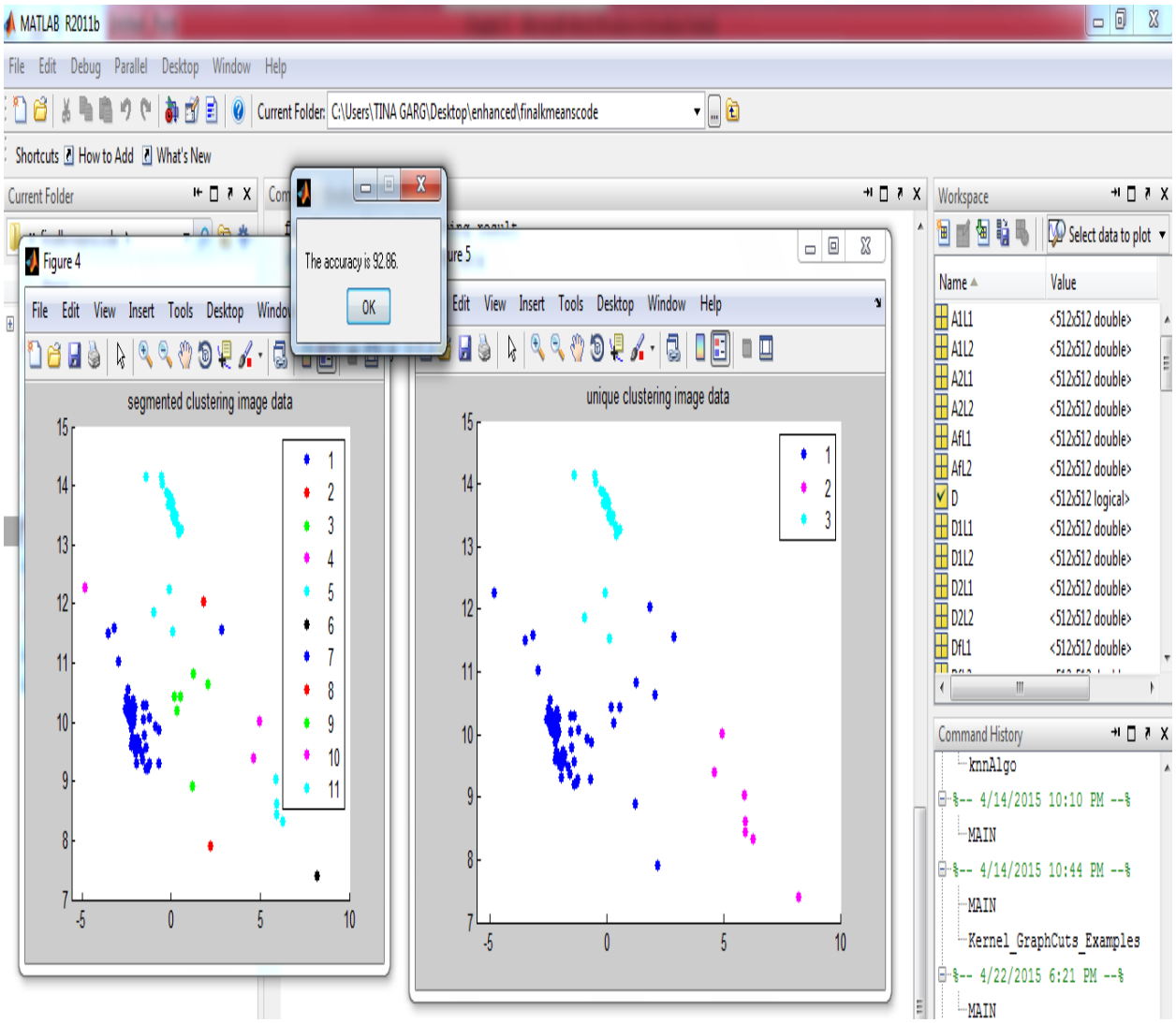

Fig 4.15: Accuracy of Purposed Algorithm

The accuracy is calculated and shown in the message box.

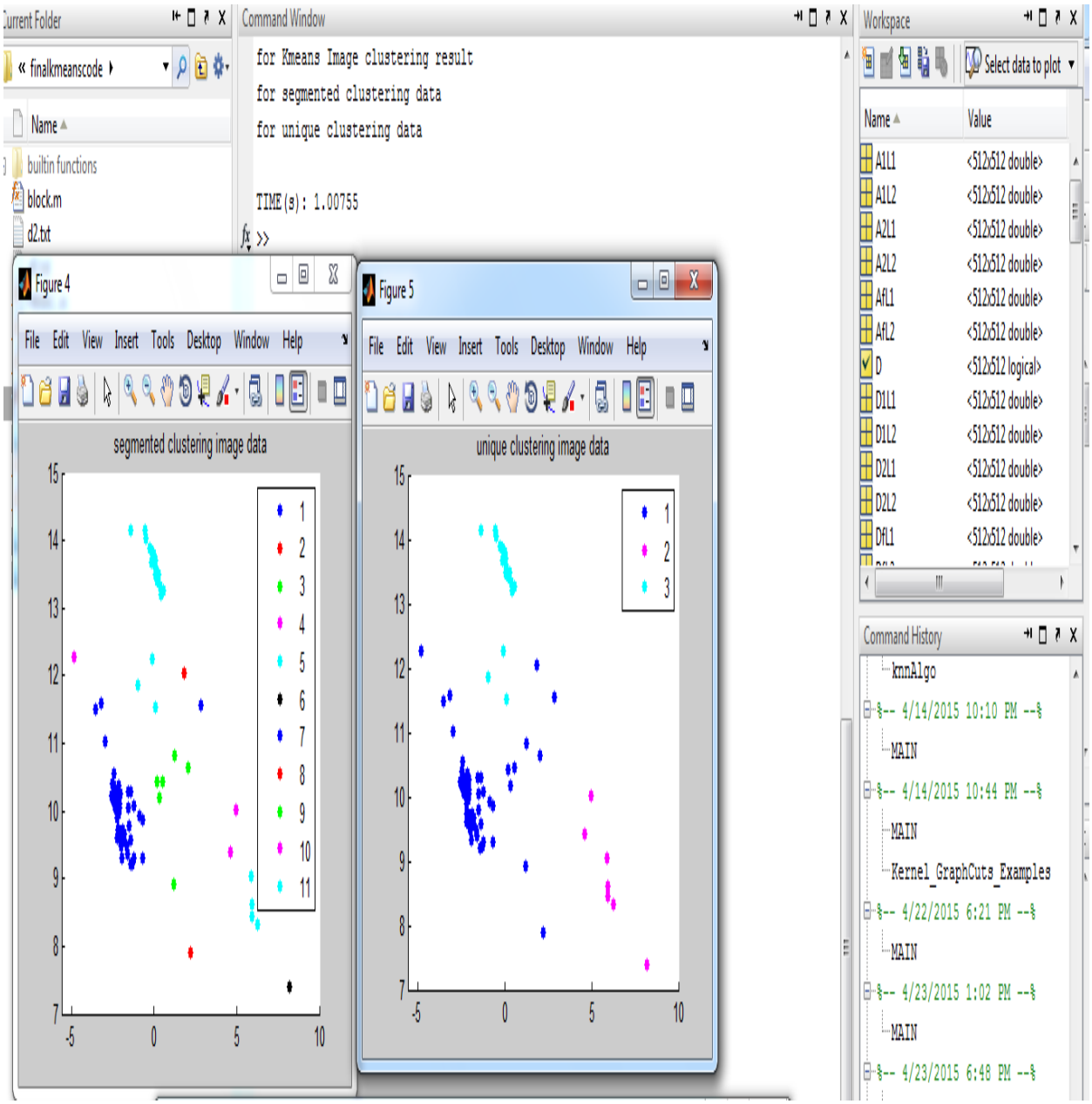

Fig 4.16: Time Taken by Purposed Algorithm

The time for clustering is shown on the command window.

4.2 Performance Evaluation

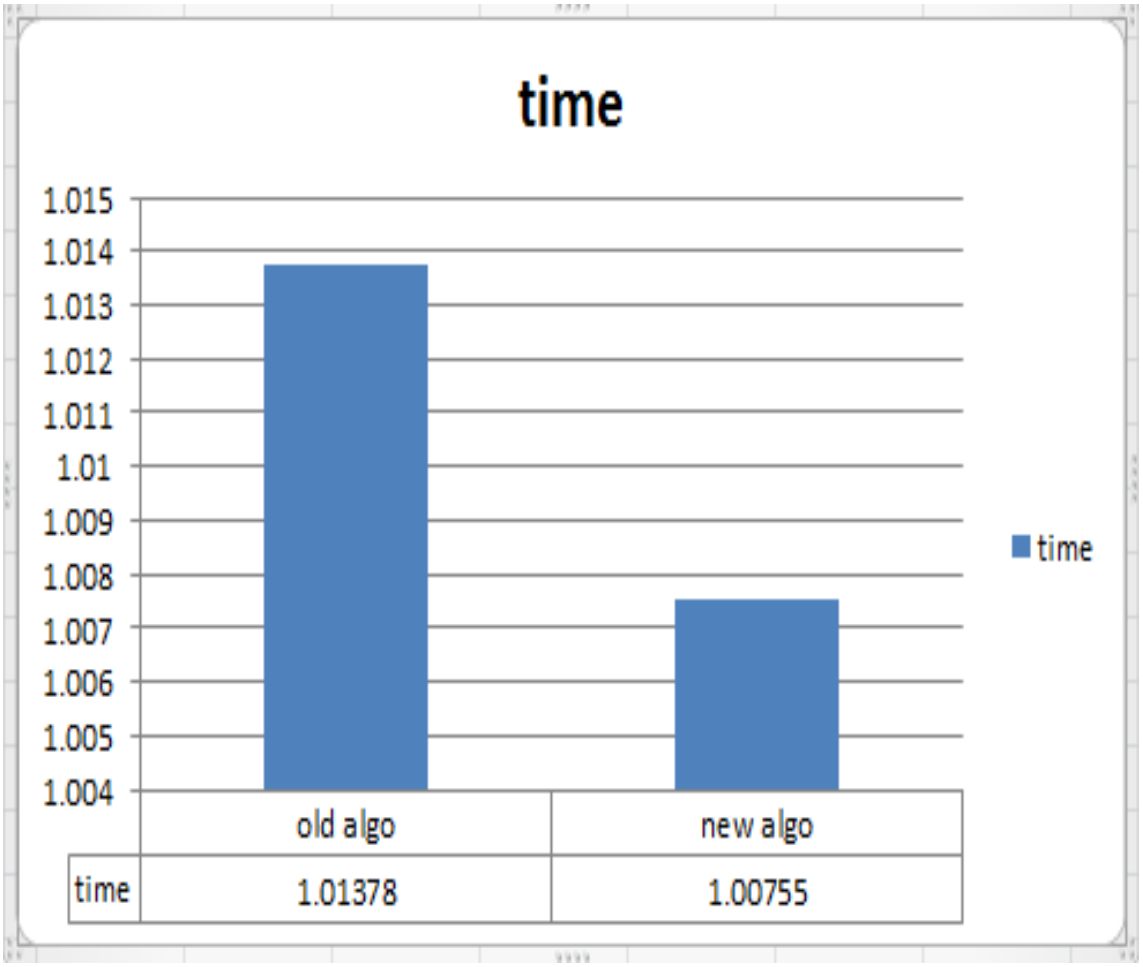

Fig 4.17: Time Comparison

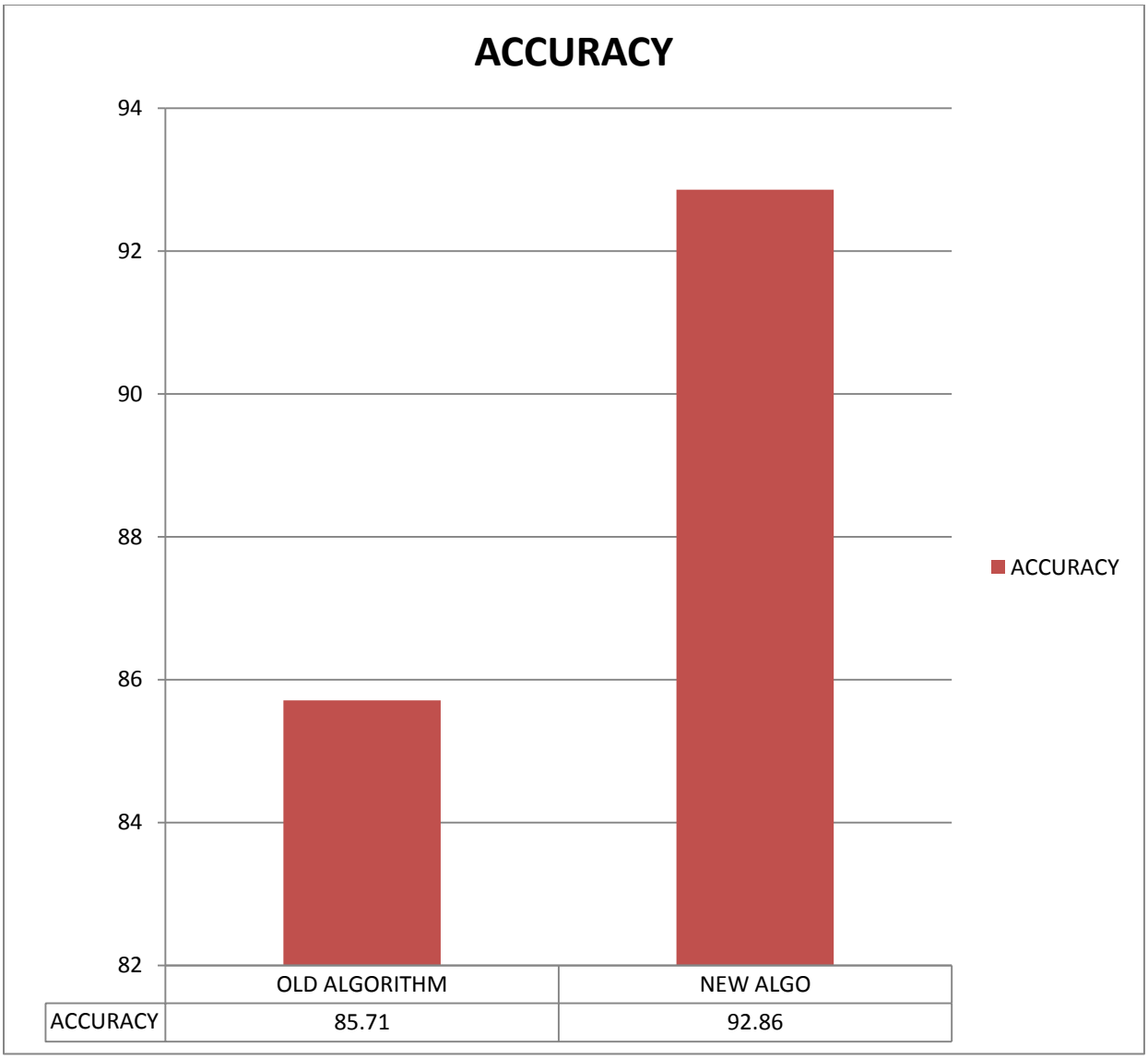

Fig 4.18: Accuracy Comparison

Chapter 5
CONCLUSION

Data Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term data mining is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Clustering is one of the most widely used techniques for analysing the data which attempts to keep similar kind of data together and dissimilar data apart from each other. Hence the main motive of the clustering is to increase the intra-cluster similarity and decrease the inter-cluster similarity. The major application areas of cluster analysis include market research, pattern recognition, data analysis, image processing and outlier detection applications such as detection of credit card fraud. Among the clustering techniques, the K-mean clustering algorithm is the most efficient algorithm which is used to cluster images, rough data etc. The k-means clustering algorithm is one of the important and simpler partitioning based algorithms different from hierarchical algorithm such as divisive and agglomerative algorithm. Hierarchical algorithm method is used to create a tree like structure by combining data objects. K-means algorithm uses k as a parameter, divide x data items into k clusters with the intention that the items in the one cluster are similar to each other but dissimilar to other items in other clusters The k-means algorithm suffer from large number of limitations such as : problem of cluster initialization, cluster quality and efficiency of algorithm, iterations etc. . The main drawback of k-means algorithm is dependency on the random selection of initial centroids. This selection has a direct impact on the efficiency and accuracy of clusters. To deal with this problem various enhancements has been purposed but there is a still scope to improve the efficiency and accuracy of K-mean clustering algorithm. So to increase cluster quality and efficiency of k-mean clustering we use HMAC. By comparing the results,it is shown that purposed k-means is better than original k-means algorithm as the time taken by purposed algorithm is less than existing and accuracy of new algorithm is much better than the original k-means algorithm.

Future Scope

The purposed algorithm based on HMAC still has some problem for further study. The problem of initialization of centroids is still present in the purposed algorithm. As compare to other algorithm Modified k-mean algorithm is tougher to noise and outliers because it minimizes a sum of general pair wise dissimilarities instead of a sum of squared Euclidean distance.

Chapter 6
REFERNCES

JOURNALS

Amanpreet Kaur Bhogal, N. S. (2010). Color image segmentation using k-means clustering algorithm. *International Journal on Emerging Technologies*.

Amar Singh, N. K. (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*.

Anwit Jain, A. R. (2010). Design, Analysis and Implementation of Modified KMean for Large Data-set to Increase. *2012 Fourth International Conference on Computational Intelligence and Communication Networks.* Indore: IEEE.

Astha Joshi, R. K. (2013). A Review: Comparative Study of Various Clustering Techniques in Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Chien-Hsing Chou, Y.-Z. H.-C.-L. (2013). Extracting and Labeling the Objects from an Image by Using the Fuzzy Clustering Algorithm and New Cluster Validity. *International Journal of Computer and Communication Engineering*.

Dharmendra K Roy, L. K. (2010). Genetic K-means Clustering Algorithm for Mixed Numeric and Categorical Data sets. *International Journal of Artificial Intelliginece and Applications (IJAIA)*.

FAHIM A.M, S. A. (2006). An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University SCIENCE* .

Huang, Z. (n.d.). A Fast Clustering Algorithm to Cluster Very Large Categorial Data Sets in Data Mining. *Cooperative Research Center for Advanced Computational Systems*.

Juntao Wang, X. S. (2011). An improved K-Means Clustering Algorithm. *IEEE*.

Liu Guoli, H. L. (2013). The Improved Research on K-Means Clustering Algorithm in Initial Values. *International Conference on Mechatronic Sciences,Electric Engineering and Computer.* Shenyang,China: IEEE.

Manpreet Kaur, U. K. (2013). Comparison Between K-means and Hierarchical Algorithm Using Query Redirection. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Md. Sohrab Mahmud, M. M. (2012). Improvement of K-means Clustering algorithm with better initial centroids based on weighted average. *7th International Conference on Electrical and Computer Engineering.* Dhaka: IEEE.

Navjot Kaur, J. K. (2012). Efficient K-means Clustering Algorithm using Ranking Method in Data Mining. *International Journal of Advanced Reasearch in Computer Engineering and Technology*.

Purohit, P. (2013). A New Efficient Approach towards K-means Clustering Algorithm. *International Journal of Computer Application*.

Raju G, B. T. (2008). Fuzzy Clustering Methods in Data Mining. *International Conference on Advanced Computer Theory and Engineering.* IEEE.

Sanjay Garg, R. C. (2006). Variations of k-mean Algorithm:A Study for High- Dimensional Large Data Sets. *Information Technology Journal* .

Shi Na, G. Y. (2010). Research on K-Means Clustering Algorithm. *Third International Symposium on Intelligent Informationa Technology and Security Informatics.* IEEE.

Shuhua Ren, A. F. (2011). K-means Clustering Algorithm Based On Coefficient Of Variation. *4th International Congress on Image and Signal Processing.* Dalian,China: IEEE.

Shunye, W. (2013). An Improved K-Means Clustering Algorithm Based on Dissimilarity. *International Conference of Mechatronic Sciences,Electrical Engineering and Computer.* China: IEEE.

Sing, R. (2011). Data clustering with modified K-means algorithm. *Recent Trends in Information Technology (ICRTIT), 2011 International Conference.* Chennai: IEEE.

BOOKS

Kamber, J. H. (2006). *Data Mining Concepts and Techniques.* ELSEVIER.

WEBSITES

http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

http://bus237datamining.blogspot.in/2012/11/advantages-disadvantages.html
http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/cluster/cluster.html

Chapter 7

APPENDIX

7.1 ABBREVIATIONS

DM-Data Mining

KDD-Knowledge Discovery in Databases

DBSCAN-Density Based Spatial Clustering of Applications with Noise

IDKM-K-means based on Iterative Density

LTS- Logical table Sources

HMAC-Hierarchal Mode Association Clustering

7.2 LIST OF PUBLICATIONS

1) A paper entitled “Survey on Various Enhanced K-Means Algorithms” has been published in International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) in November 2014, Volume 3, Issue 11.

2) A paper entitled “To Intensify Cluster Quality by Enhancing K-Means Clustering Algorithms in Data Mining ” has been accepted in International Journal of Applied Engineering Research (IJAER) in April 2015 which has SCOPUS indexing.(Paper code: 35206)

3) A paper entitled “To Intensify Cluster Quality by Enhancing K-Means Clustering Algorithms in Data Mining” has been accepted in International Conference on Advances in Applied Engineering and Technology (ICAAET)-2015. Conference Proceeding will be published in SCOPUS indexed journal International Journal of Applied Engineering Research (IJAER). (Paper Id: 1033)