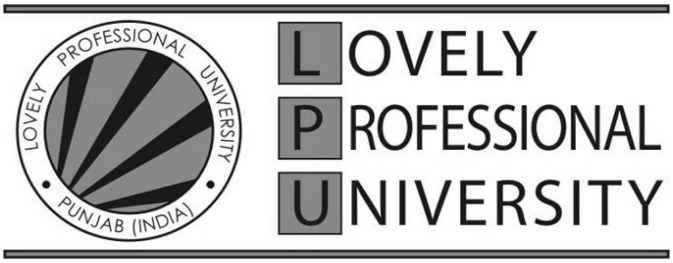


A New Approach to Construct the Affinity Graph of Spectral Clustering Using Euclidean Distance

A Dissertation Submitted

By

Gurpinder Kaur

11304804

To

Department of CSE

In fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science and Engineering

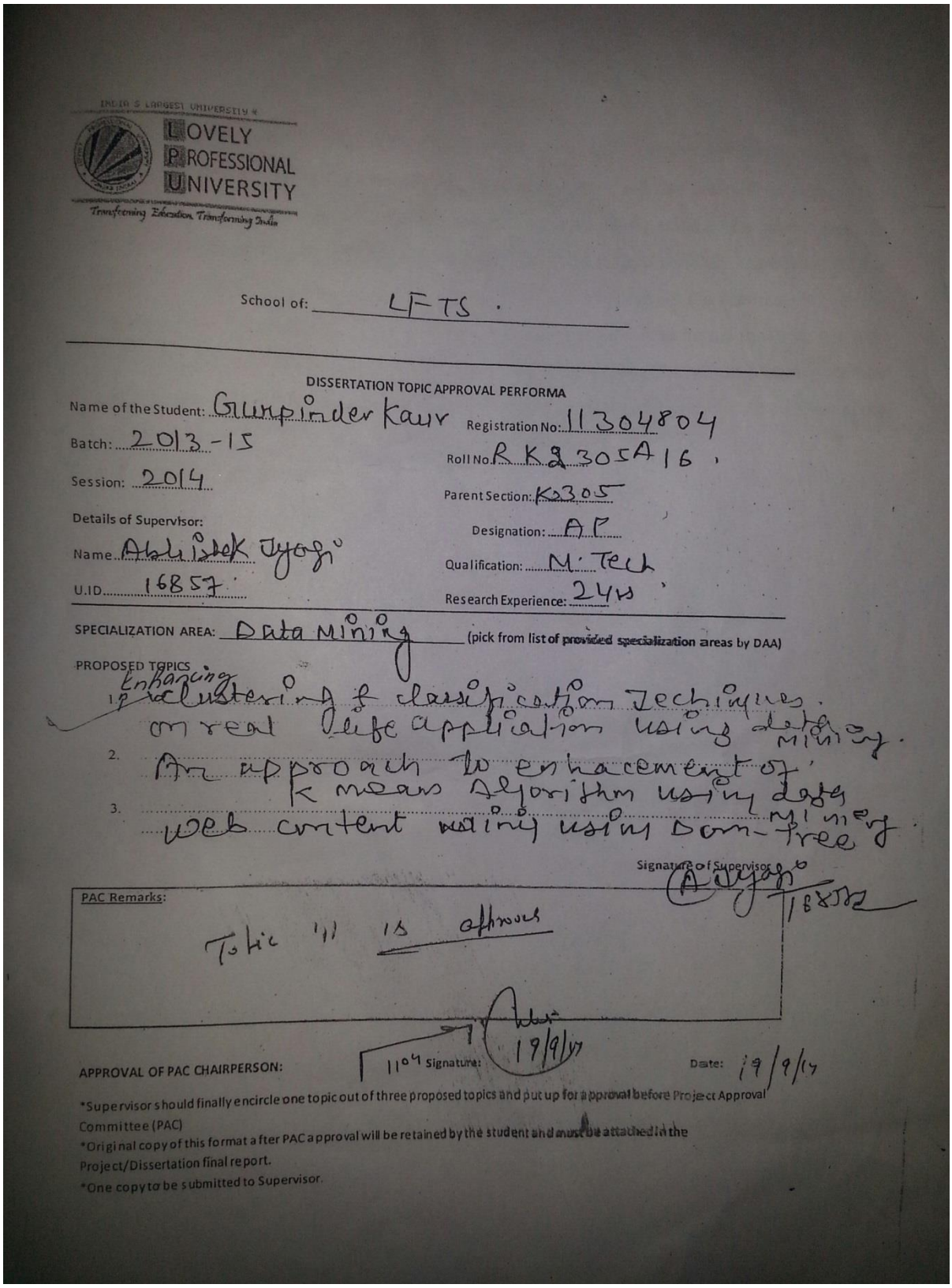
Under the guidance of

Mr. Abhishek Tyagi

(Assistant Professor, Lovely Professional University)

(May 2015)

PAC FORM



INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

School of: LFTS.

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Gurpinder Kaur Registration No: 11304804

Batch: 2013-15 Roll No: RK2305A16.

Session: 2014 Parent Section: K2305

Details of Supervisor: Designation: AP

Name: Abhishek Tyagi Qualification: M.Tech

U.ID: 16857. Research Experience: 2yrs.

SPECIALIZATION AREA: Data Mining (pick from list of ~~provided~~ specialization areas by DAA)

PROPOSED TOPICS

1. Enhancing clustering & classification Techniques on real life application using data Mining.
2. An approach to enhancement of K means Algorithm using data mining
3. web content mining using Dom-tree

Signature of Supervisor: A Tyagi 18/9/17

PAC Remarks:

Topic 1) is approved

Signature: 19/9/17 Date: 19/9/17

APPROVAL OF PAC CHAIRPERSON:

*Supervisor should finally encircle one topic out of three proposed topics and put up for 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

The core part of the spectral clustering is to construct the affinity matrix. There exist several ways to construct the affinity matrix and most of them are constructed directly using the weight matrix of the whole dataset and some functions like Gaussian functions. Although these approaches are good enough to perform the process of spectral clustering but they cannot handle the large datsets efficiently. In this work, three additional approaches are used for the construction of the affinity graphs named as: normal KNN affinity graph, mutual KNN affinity graph and epsilon affinity graph. In all these approaches, the Euclidean distance approach is also used along with the Gaussian functions. By using this approach, the most relevant data points are grouped together with which the large data sets can be clustered easily and the quality of the clusters is also improved. The experimental results prove that the proposed approach is helpful in achieving the improved quality of clusters.

CERTIFICATE

This is to certify that she has completed M. Tech dissertation proposal titled ‘A New Approach to Construct the Affinity Graph of Spectral Clustering Using Euclidean Distance’ 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 degree or diploma.

The Report is fit for the submission and the fulfilment of the conditions for the award of M. Tech Computer Science & Engineering.

Date: _____

Signature of Advisor

Name:

UID:

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without mentioning the name of those people who made it possible, because success results, not only from hard work, but also from steadfast determination, dedication and above all adepts advises. I would like to express my special gratitude to my mentor Mr. Abhishek Tyagi for his guidance and support. I would like to thank him for encouraging my research work and also for his suggestions.

I would also like to appreciate the guidance given by the Department of Computer Science thanks for their valuable advice.

Gurpinder Kaur

(11304804)

DECLARATION

I hereby declare that dissertation proposal entitled "A New Approach to Construct the Affinity Graph of Spectral Clustering Using Euclidean Distance" 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

Reg. number 11304804

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Data mining (also known as data or knowledge discovery process) is a process of evaluating and collecting useful data from a large amount of stored data by analysing it from different perspectives [14].It is a process of extracting useful, meaningful and relevant data from a huge amount of data.

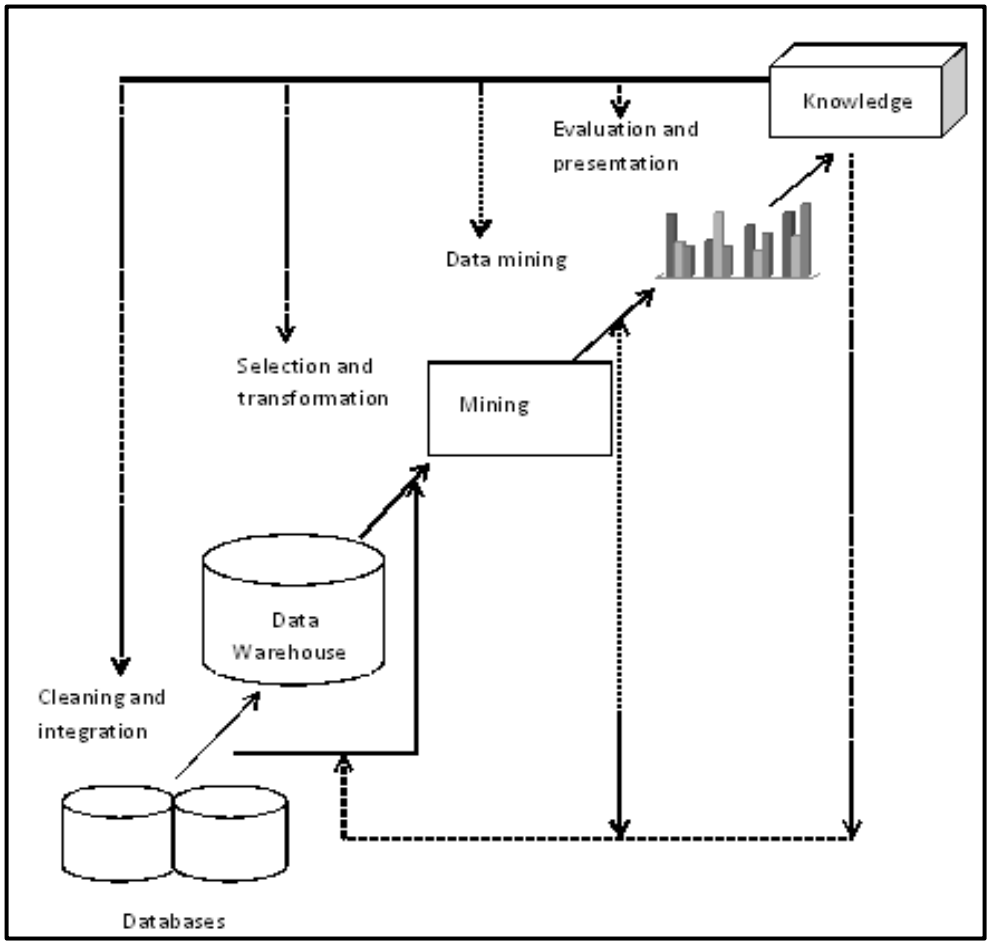


Fig.1.1 Data Mining as a step in the process of knowledge discovery

The main steps that are involved in the knowledge discovery process includes cleaning of the data (removal of noise and inconsistent data), after that integration of the data (data collected from the multiple sources is combined), selection of the relevant and useful data (the data relevant to the analysis task are retrieved from the database), transformation of the data (data

are transformed and consolidated into forms appropriate for mining by performing summary or aggregate operations), mining of the data (an essential process where intelligent methods are applied to extract data patterns), evaluation of the patterns (identification of the truly interesting patterns representing knowledge based on interestingness measures such as support and confidence) and presentation of the knowledge (visualization and knowledge representation techniques are used to present the mined knowledge) [4].

In the architecture of the data mining the basic parts includes information repository blocks, data warehouse server, data mining engine, pattern evaluator, knowledge base and user interface. The information repository blocks may include databases, data warehouse, world wide web or any other repository. The data obtained from these sources is cleaned, then integrated and then useful and relevant data is selected and stored in data warehouse server. After that data mining engine is there which contains a set of functional modules for performing various kinds of responsibilities. The main responsibilities of this engine include cluster analysis, outlier analysis, classification and association and correlation analysis. Knowledge base is a part which is used to conduct the search, or calculate the interestingness of resulting patterns. Pattern evaluation is that module which evaluates the patterns of the data that is mined with the help of data mining engine and user interface is that which attempts to communicate between users and the data mining system to allow the users to interact with system.

The architecture for the data mining is shown in fig.1.1

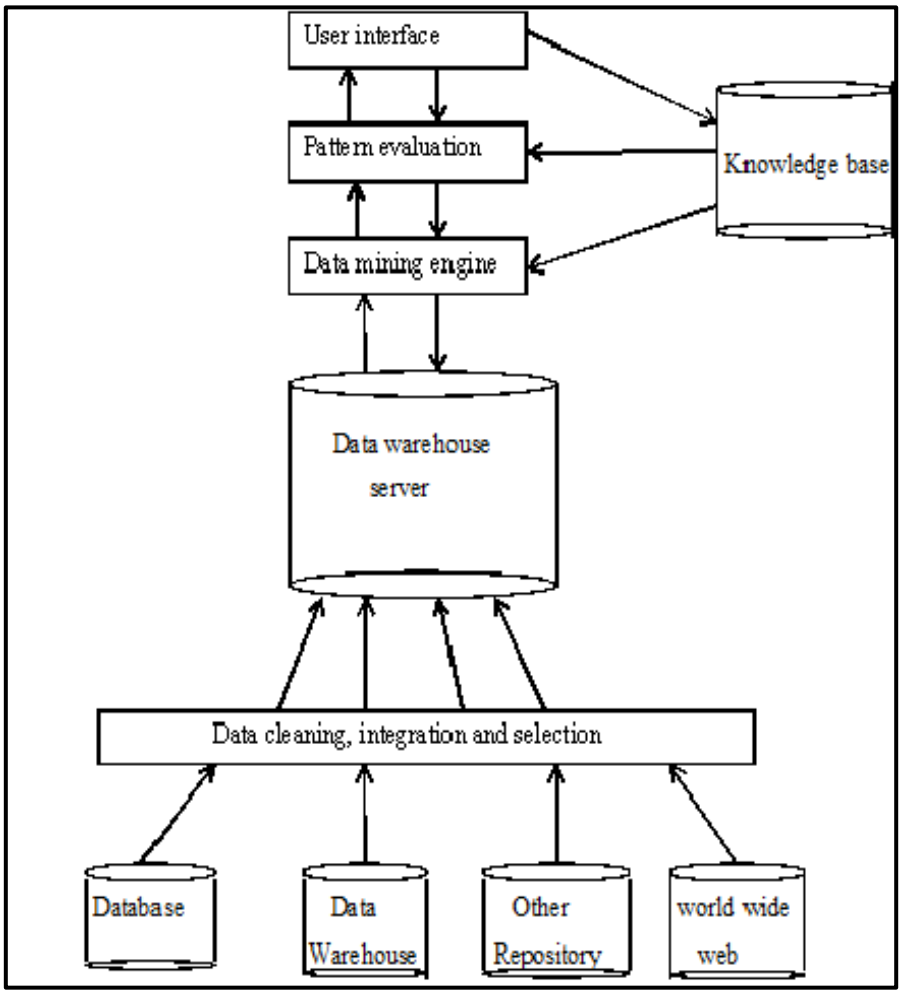


Fig.1.2 Architecture for Data Mining

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. Data clustering is a method in which the whole data under consideration is divided into clusters and this division process depends upon the characteristics of the data. The data with similar characteristics is kept in one cluster and those with different are kept in different clusters. Hence the main motive of the clustering is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. The major application areas of cluster analysis include pattern recognition, data analysis, image processing and outlier detection applications such as detection of credit card fraud. There are some requirements for the clustering that must be fulfilled. These requirements include:

- The clustering algorithms must be scalable so that they can easily deal with large datasets.
- The algorithms must be capable enough to detect the noisy data and can differentiate it from the other data.
- They must be capable of finding out the clusters of different shapes but not just spherical shapes.

1.2 TYPES OF CLUSTERING

There exist many different types of clustering methods including:

1.2.1 Hierarchical method

1.2.2 Partitioning method

1.2.3 Density based method

1.2.4 Model-based method

1.2.5 Grid-based method.

1.2.1 Hierarchical method: This method is also known as connectivity-based methods. It uses the approach of recursively subdividing the instances or data points in either a top-down fashion or bottom-up fashion in order to form the clusters. Depending on this top-down and bottom-up fashion, this method can be further divided into two types- agglomerative hierarchical and divisive hierarchical. Agglomerative is a bottom-up approach in which each object initially represents a cluster of its own and then they are repeatedly merged until the desired cluster structure is obtained. Divisive-hierarchical clustering uses top-down approach where initially all the instances are in one single cluster which is then divided into sub-clusters and then these sub-clusters are in turn divided into sub-clusters until the desired cluster structures are obtained [13]. The major disadvantage of hierarchical methods is that any action (split or merging) once performed can never be undone [16].

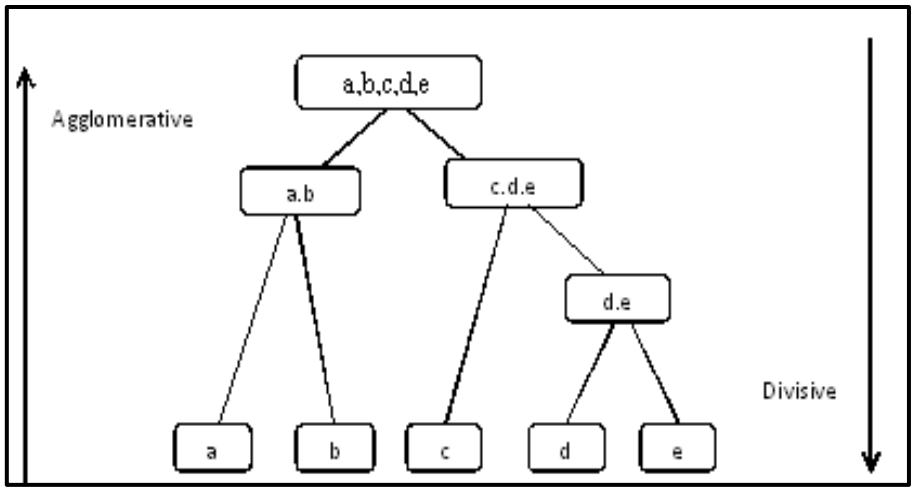


Fig.1.3 Hierarchical Clustering

1.2.2 Partitioning methods: Partitioning methods (also known as centroid-based methods) rearrange instances by repeatedly moving them from one cluster to another. Such methods typically demand the number of required clusters to be predicted in advance [13]. The most common example of this clustering method is k-means clustering which requires the number of clusters to be determined in advance and the same number of data points are selected to act as centroids.

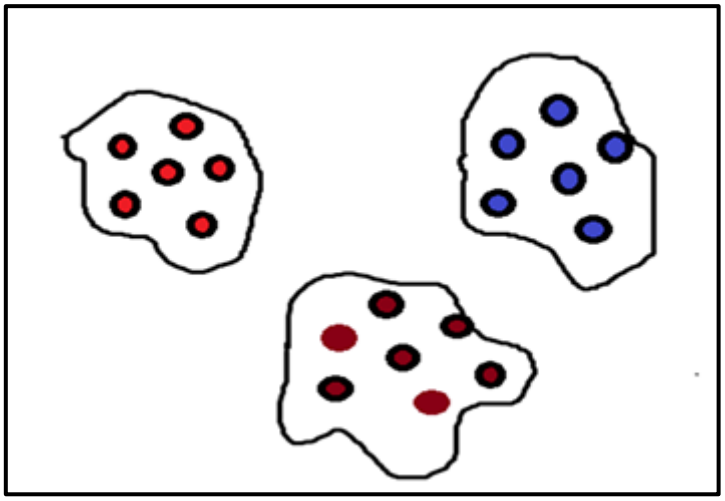


Fig.1.4 Partitioning Clustering

1.2.3 Density-Based Clustering: In Density-based clustering, the density of the neighbourhood instances is checked to form the clusters. The areas of higher density are defined as clusters. The most popular density based clustering method is DBSCAN. It is based on connecting points with certain threshold distance [15]. DBSCAN searches for clusters by checking the ε-neighbourhood of each point in the dataset. If ε-neighbourhood of

a point P contains more value than the minimum point then a new cluster with P as core object is created. DBSCAN then iteratively collects density-reachable objects from these core objects which may involve the merge of few density-connected objects. The process terminates when no new point can be added to any new cluster.

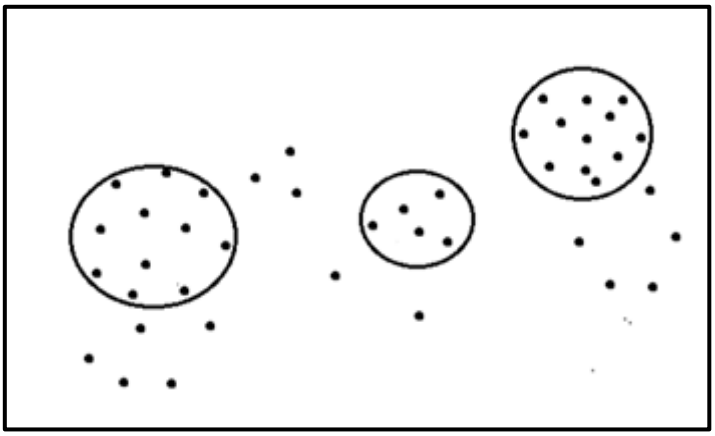


Fig.1.5 Density-based Clustering

1.2.4 Model-based clustering: This type of clustering is also known as distribution based clustering. This method assumes that the data were generated by model and tries to recover the original model from the data. The model that is recovered from the data then defines clusters and assignment of documents to clusters [17].

1.2.5 Grid-based clustering: In the grid-based method, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure. The major advantage of this method is its capability for the fast processing time [16], that is, it takes very little time for the clustering process.

1.3 SPECTRAL CLUSTERING

Spectral clustering is a form of clustering that originates from graph theory in which all the data points of the data set are represented in the form of nodes and the relationships between them are represented with the help of edges which carry some weight. High weight indicates high similarity between the two nodes and the low weight is an indication of the dissimilarity between the two nodes. These weights between the nodes are recorded in the form of a matrix and the matrix so produced is known as distance or weight matrix. This distance matrix, just created, is then converted into similarity matrix using some conversion method. After this, laplacian matrix is constructed which is usually the difference between the distance matrix and the similarity matrix. The next step is to calculate the Eigen values and

construct a matrix that consists of these Eigen values as the columns. Then cluster the data points using some clustering algorithm.

The diagrammatic view of the rough idea of spectral clustering is shown in fig. 1.2

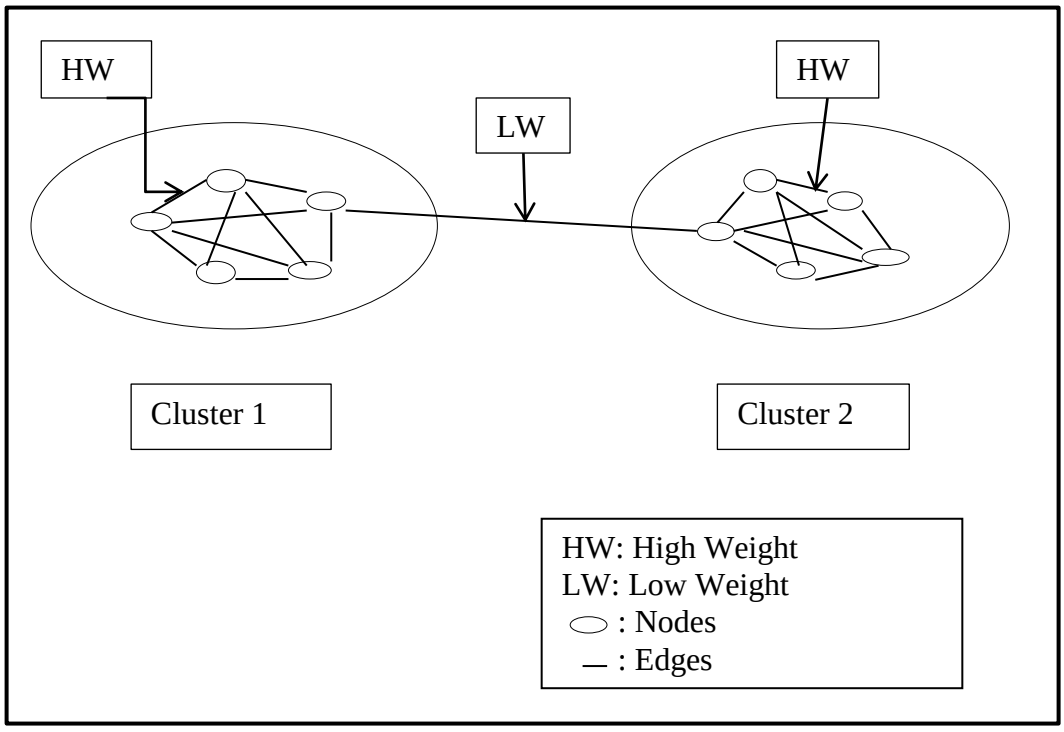


Fig. 1.6 Basic Overview of Spectral Clustering

Spectral Clustering Stages: There are few stages of spectral clustering. These stages are:

- **Pre-Processing:** In this stage the distance matrix and the affinity graphs are constructed that are required for further processing.
- **Spectral Representation:** The Laplacian matrix is constructed in this stage. For constructing the Laplacian matrix, distance matrix and the affinity matrices are required which were calculated in the previous stage. The eigen values and the eigen vectors are also calculated in this stage.
- **Clustering:** Clustering is usually performed using k-means clustering method. The matrix is generated from the eigen vectors calculated in the previous stage and on this matrix clustering operation is performed.

1.4 ADVANTAGES OF DATA MINING

1. Finance and Banking: Data mining is very helpful for financial institutions to collect information about loans and credit reporting. Models can be built using historical customer's data and accordingly the bank and financial institution can have an idea about the good loans and bad loans. Data mining technique also helps banks to detect if there is any fraudulent credit card transaction to protect credit card's owner from any kind of fraud [18].

2. Retail and Marketing: Data mining helps marketing companies also to construct models based on historical data to guess who will respond to the new marketing campaigns. It helps them to have an estimate about the popularity of the new decisions by considering the older data. By analysing the results, marketers will have correct move toward to sell gainful products to directed customers [18].

Data mining helps a great deal to bring a lot of profit to retail companies in the same way as in the marketing. For instance, by considering the technique of market basket analysis, a store can have an efficient and suitable production arrangement in a way that customers can buy normal buying products together with pleasant. This technique is beneficial for both: customers as well as for the owners of the companies. In addition to this, it also helps the marketing companies to make decisions to offer certain discount for particular products that will attract more customers [18].

3. Manufacturing Industry: By applying data mining in equipped engineering data, manufacturers can detect equipment having any sort of fault and can also determine optimal control parameters [18].

4. Governments: Data mining also helps government agencies by mining and analysing records of financial transactions to construct patterns that can detect money laundering or criminal activities [18].

1.5 DISADVANTAGES OF DATA MINING

1. Security issues: Companies store a vast amount of information about their employees and customers including social security number, birthday, payroll, address, account number and etc. However how properly they take care of this information is still in questions. There have been a lot of incidents that hackers accessed and stole the data of customers from big and reputed corporations. Because of the huge amount of personal and financial information

available, the threat of the stealing of the credit cards and identity theft becomes a big problem [18].

2. Misuse of information/inaccurate information: Information collected through data mining, intended for the ethical purposes can be misused. This information may be used by some corrupt people or businesses to take benefits of other people [18]. Moreover, if any sort of mistake is committed while obtaining information then that inaccurate information may have a serious consequence [19].

3. Privacy Issues: With the widespread use of the internet, the concerns about privacy have increased tremendously. Some people are afraid of giving their information to sites because someone may have access to their information and then may misuse it. Due to this reason they are afraid of doing online shopping [19].

CHAPTER 2
REVIEW OF LITERATURE

Hao Huang et.al (2014) they mentioned that it is an open challenge to mine the arbitrary shaped clusters in large data sets. Various approaches to this problem have been proposed but they all have very high time complexity. In order to save the computational cost, some algorithms have made attempts to shrink/collapse the size of the data set to a smaller amount of representative data examples. However, this kind of user-defined shrinking ratios may significantly affect the performance of the clustering. In this paper, they present CLASP, by adopting three phase strategy that has been proved to be an effective and efficient algorithm for mining the clusters having an arbitrary shape. The first phase of their approach attempts to shrink the size of a data set automatically while effectively retaining the information about the shape of clusters in the data set with representative data examples. Then in the second phase, it attempts to adjust the positions of these representative data examples to increase their intrinsic relationship and make the structures of the clusters more clear and distinct for clustering. Finally in the third and last phase, it performs agglomerative clustering to identify the structure of the cluster with the help of a mutual k-nearest neighbors-based similarity metric called Pk for completing the mining of arbitrary shaped clusters [3].

Xiatin Zhu et.al (2014) proposed a new method for constructing the affinity matrix. They did not take into consideration all the available features for constructing similarity matrix. Instead, they avoided les important features by measuring between-sample proximity and focused only on the more important features. They derived pairwise similarities of arbitrary sample pairs from a set of comparative tests using different available features. Such subtle similarities distributed over discriminative feature subspaces are combined automatically and effectively for producing robust affinity matrices. The affinity matrix constructed by this method automatically possess the local neighborhood [11].

Sumuya Borjigin and Chonghui Guo (2013) proposed non-unique cluster number determination methods based on stability. They used Gaussian kernel parameters (global scale and local scale) to convert the distance matrix into the similarity matrix and then used the multi-way normalized cut algorithm to cluster the data points. In their work they also

focused on determining whether the chosen cluster numbers are stable and reasonable which is helpful in improving the performance of the clustering procedure. Also the coherence is measured based on the gap to measure clustering quality [9].

Hongjie Jia et.al (2013) introduces basic concepts of spectral clustering and the latest research developments in this field. In the basic concepts, they discussed about the adjacency matrix, laplacian matrix (normalized and un-normalized), probability transition matrix and modularity matrix. In the latest research development, they discussed about the construction the similarity matrix, construction of the laplacian matrix, selection of the eigenvectors, selection of the number of clusters and the various applications of spectral clustering [4].

Hsin Chien Huang et.al (2012) discovered that it is useful to construct multiple affinity matrices on the basis of different useful features. They found that in many applications there could be different useful features and hence leading to the construction of different affinity matrices. They proposed an affinity aggregation spectral clustering (AASC) algorithm that attempts to extend the spectral clustering to a setting with multiple affinities available. AASC shares similar ideas with multiple kernel learning (MKL) method. However, this method was different from MKL in these two points: this method is supervised i.e.no labels are available for data and they assume that affinity matrices are symmetric [5].

Xianchao Zhang and Quanzeng You (2010) proposed a random walk based approach to process the Gaussian kernel similarity matrix. In order to make the similarity or affinity matrix close to the ideal matrix, the pairwise similarity between two data points they consider was not only related to two data points, but also related to their neighbors. To deal with the noisy items, initially the noisy items were ignored and only the other items were taken into account to perform the clustering. After clustering the non-noisy items, the correct cluster for each noisy item is determined [10].

Cuimei Guo et.al (2010) Discussed the basic framework of spectral clustering. They introduced the basic theories relative to spectral clustering and some algorithms also. Some of the partition criterions they included in their paper includes minimum cut, normalized cut and multiway normalized cut. Besides this, they also focussed on the problems related to the spectral clustering which includes (a) constructing the efficient similarity matrix and the

graph laplacian, (b) to decide the parameters of spectral clustering such as number of clusters and sigma and (c) extending spectral clustering to large data sets [1].

Gamila Obadi (2010) et.al focussed on the problems related to the high dimensionality data generated by social networks. They used spectral clustering to find the student's behavioural patterns performed in e-learning system. They used spectral clustering to find out the co-relations that exist between the similarity in the student's behaviour and their respective grades they secured. In order to conduct this research they developed a software that permits the users to define the input data values. In their results they found the students clustered together with similar behaviour but different grades. The main drawback of their approach was that it was applicable only to a small and homogeneous data set [2].

Xu-Degang et.al (2009) focused on building the affinity matrix, which is the most important part of spectral clustering and affects the process and quality of spectral clustering to a great extent. They proposed four different methods to build the affinity matrix. These matrices include Gaussian kernel function, the minkowski function, the nearest co-relation function and the local scale function. Then, they developed four new algorithms to contrast the clustering results and concluded that building appropriate local scale function is the most available method to make the affinity matrix [12].

Peter Kontschieder et.al (2008) discussed different possibilities for normalizing the similarity matrix and what could be their effect on the retrieval steps. They proposed to use a modified version of the mutual KNN graph to analyse the underlying structure of the datasets. They proposed a two-way normalization and analysis scheme which was having two main goals: the first goal aims on modelling object interdependence by neighbourhood incorporation and the second goal aims on improving the information retrieval performance [8].

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

Sumuya Borjigin and Chonghui Guo utilized three stages to determine non-unique cluster numbers of a data set. First of all, they utilized the multiway normalized cut spectral clustering algorithm to make the clusters of the data points of the data set for some cluster number k. Then they used the ratio value of the multiway normalized cut criterion of the obtained clusters and the sum of the leading eigenvalues of stochastic transition matrix as a standard to decide whether the k is a reasonable cluster number. In the third stage, they varied the scaling parameter in the Gaussian function to judge whether the cluster number k is also stable or not.

The algorithms they used in their work are described below:

Algorithm 1: Meil˘*a*-Shi multiway normalized cut spectral clustering algorithm:

– **Input:** Data set $P = \{p1, p2, \ldots, pn\}$, cluster number k.

– **Step 1** Compute the distance matrix W, construct similarity matrix S according to W, where $W(i, j)$ is the distance between pi and pj , $i = 1, 2, \ldots, n$;

– **Step 2** Calculate the Laplacian matrix, $L = D - S$;

– **Step 3** Compute the first k eigen vectors $\{v1, \ldots, vk\}$ of the generalized eigen problem $Lv = \lambda Dv$;

– **Step 4** Let $V \in Rn \times k$ be a matrix composed of the vectors $\{v1, \ldots, vk\}$ as columns;

– **Step 5** For $i = 1, \ldots, n$, let $yi \in R1 \times k$ be the vector corresponding to the ith row of V;

– **Step 6** Cluster the points $\{yi \in R1 \times k \mid i = 1, 2, \ldots, n\}$ with the k-means algorithm into clusters $C1, \ldots, Ck$, if $yi \in Cj$ then $pi \in Pj$, $1 \leq i \leq n$, $1 \leq j \leq k$.

–Output: k clusters $P1, \ldots, Pk$.

The similarity matrix used in this algorithm is calculated using gaussian kernel functions. The gaussian kernel functions can be divided into global scale gaussian kernel parameter and the local scale gaussian kernel parameter.

Global scale gaussian kernel parameter method is defined as

$$S(i,j)=\exp\left(-\frac{d^2(p_i,p_j)}{\sigma^2}\right)$$

where d(pi,pj) is the distance between point pi and pj and sigma is gaussian kernel parameter. Local scale gaussian kernel parameter is defined as

$$S(i,j)=\exp\left(-\frac{d^2(p_i,p_j)}{\sigma_i\sigma_j}\right)$$

Algorithm 2: Non-Unique Cluster Number determination method based on stability under global scale Gaussian kernel parameter:

– **Input**: Data set $P = \{p1, p2, \ldots, pn\}$, $\epsilon > 0$, $\delta > 0$, user-specified upper threshold $C\max \geq 2$ for cluster number to be testified, number of Gaussian kernel parameter $|\Sigma|$.

– **Step 1** Calculate the distance matrix W;

– **Step 2** Let $\sigma\min \triangleq \min\{Wi\,j \,|Wi\,j \neq 0, i, j = 1, 2, \ldots, n\}$, $\sigma\max \triangleq \max\{Wi\,j \,|i, j = 1, 2, \ldots, n\}$, $\sigma t \triangleq \sigma\min + \frac{\sigma max - \sigma min}{\Sigma - 1} * (t - 1)$, for every σt run step 3∼4;

– **Step 3** Calculate the similarity matrix S, where $S(i, j) = \exp\left(-\frac{W_{i,j}^2}{\sigma_t^2}\right)$;

– **Step 4** For every $k = 2, \ldots, C\max$, make use of the Meilˇa–Shi spectral clustering algorithm to cluster the data set P into k clusters and calculate the value of index *Ratio(k)* for obtained clusters;

– **Step 5** To determine whether the candidate cluster number $2 \leq k \leq C\max$ is an *ϵ-reasonable* and *δ-stable* cluster number according to the results of step 3 and step 4;

– **Output**: The set of *ϵ-reasonable* and *δ-stable* cluster numbers.

Algorithm 3: Non-Unique Cluster Number determination method based on stability under local scale Gaussian kernel parameter:

– **Input**: Data set $P = \{p1, p2, \ldots, pn\}$, $\epsilon > 0$, $\delta > 0$, user-specified upper threshold $C\max \geq 2$ for cluster number to be testified, user-specified maximum number of neighbors $K\max \geq 2$.

– **Step 1** Calculate the distance matrix W;

– **Step 2** For $i = 1, 2, \ldots, n$, sort the ith row of W, then calculate $pi\,K$, which is the Kth neighbor of pi , $K = 2, \ldots, K\max$;

– **Step 3** For $K = 2, \ldots, K\max$ run step 4∼5;

– **Step 4** Calculate the similarity matrix *S*, where $S(i, j) = \exp\left(-\frac{W_{i,j}^2}{\sigma_{iK}\sigma_{jK}}\right)$;

– **Step 5** For every $k = 2, \ldots, C\max$, make use of the Meil˘*a*–Shi spectral clustering algorithm

to cluster the data set *P* into *k* clusters and calculate the value of index *Ratio(k)*

for obtained clusters;

– **Step 6** To determine whether the candidate cluster number $2 \leq k \leq C\max$ is an ϵ-*reasonable* and δ-*stable* cluster number according to the results of step 4 and step 5;

– **Output**: The set of ϵ-*reasonable* and δ-*stable* cluster numbers.

The problem exists in the defined algorithm is of cluster quality and computation cost. In this work we will enhance spectral clustering technique to improve accuracy and clustering quality of the algorithms and to reduce the computation cost.

3.2 OBJECTIVES OF THE STUDY

1. To study and analyse various kinds of existing methods for performing spectral clustering techniques for clustering the data in data mining.
2. To identify the problem of data accuracy and computation cost in spectral clustering techniques.
3. To enhance the spectral clustering technique using geometric transformation technique for data mining.
4. Implementation of the proposed technique and existing techniques and analyse the performance in terms of accuracy, computation cost and cluster quality.

3.3 SCOPE OF THE STUDY

Spectral clustering is a form of density-based based clustering method in which all the data points of the data set are represented in the form of nodes and the relationships between them are represented with the help of edges which carry some weight. High weight indicates high similarity between the two nodes and the low weight is an indication of the dissimilarity between the two nodes. These weights between the nodes are recorded in the form of a matrix and the matrix so produced is known as distance or weight matrix. This distance matrix is then converted into similarity matrix using some conversion method. After this, laplacian matrix is constructed which is usually the difference between the distance matrix

and the similarity matrix, the next step is to calculate the eigenvalues and construct a matrix that consists of these eigen values as the columns. Then cluster the data points using some clustering algorithm. The special clustering will be enhanced in this work to improve the cluster quality and to increase the accuracy of the algorithm. The performance of the spectral clustering is very much dependent on the similarity matrix. More the similarity matrix is close to the ideal matrix more improved is the performance. The ideal matrix is the matrix in which the intra-cluster data points are assigned value 1 whereas the values assigned to the inter-cluster data points is 0.

3.4 RESEARCH METHODOLOGY

In this proposed work, when the dataset is loaded then it is recommended to normalize the data so that the relations between the data can be detected. After normalizing the data, affinity graphs are created. There are four ways with which affinity graphs can be constructed in this work: full, normal-k nearest neighbour, mutual-k nearest neighbour and epsilon. In the full approach, the traditional approach of constructing the affinity graph is used; that is, using simple distance matrix and then Gaussian functions the affinity matrix is created. In the other two methods, the whole dataset is first divided into partitions on the basis of the Euclidean distance. The random data points are chosen to initiate the process. The Euclidean distance of all the data points from these chosen points is calculated and the points having the almost similar values or the closer values are collected in same cluster. This way the most relevant data points are placed close to each other.

After creating the affinity graph, the next step is to perform the clustering. The three types of clustering methods are compared: unnormalized spectral clustering, normalized Spectral Clustering Using Euclidean Method and normalized JW method.

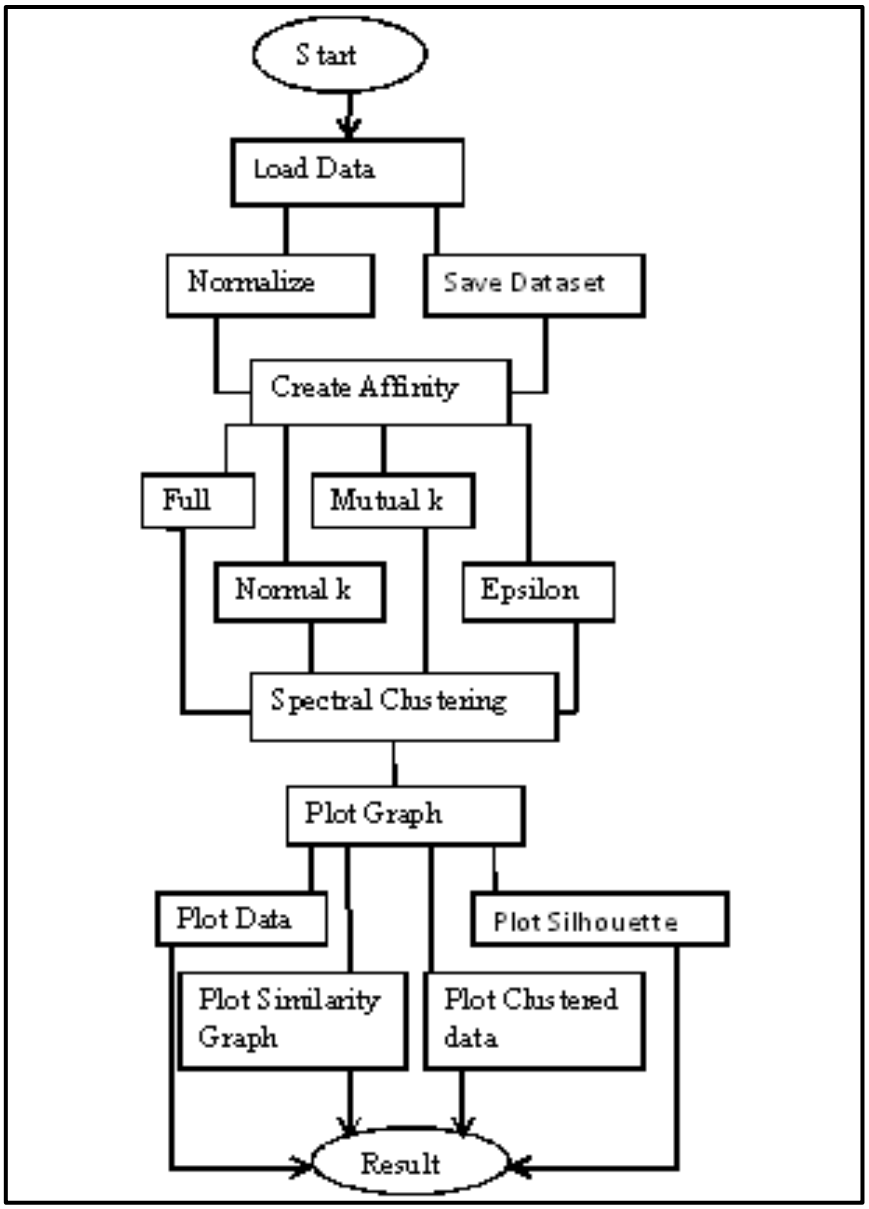


Fig.3.1 Research Methodology of the Proposed Work

CHAPTER 4
RESULTS AND DISCUSSION

4.1 INTRODUCTION TO THE TOOL

The tool that is used for the accomplishment of this task is MATLAB. The name MATLAB stands for MATrix LABoratory. MATLAB is a high-performance language for technical computing. It integrates computation, visualization and programming environment. It has powerful built-in routines that enable a very wide variety of computations. Specific applications are collected in packages known as toolbox. There are toolboxes for database, signal processing, image processing, statistics and many other[16]. It can produce nice pictures in both 2D and 3D.

4.2 INTERFACE

The abalone dataset is used to perform this clustering and to compare the results. Abalone are the marine snails. Abalone are single shelled snails with a large muscular foot to hold them to rocks. This dataset has 9 attributes: sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and rings.

The script-based GUI is prepared from which first of all dataset is loaded and then to perform the unnormalized clustering this dataset is not required to normalize and let it remain as it is. Whereas, to perform the Spectral Clustering Using Euclidean Distance and normalized JW this dataset is to be normalized.

After this affinity graphs are created. Out of the four ways, first one is full approach in which dataset is not partitioned. Results prove that the time required for performing the spectral clustering using full affinity graph approach is more comparative to the other three methods.

The interface that is made for performing all the process is as follows:

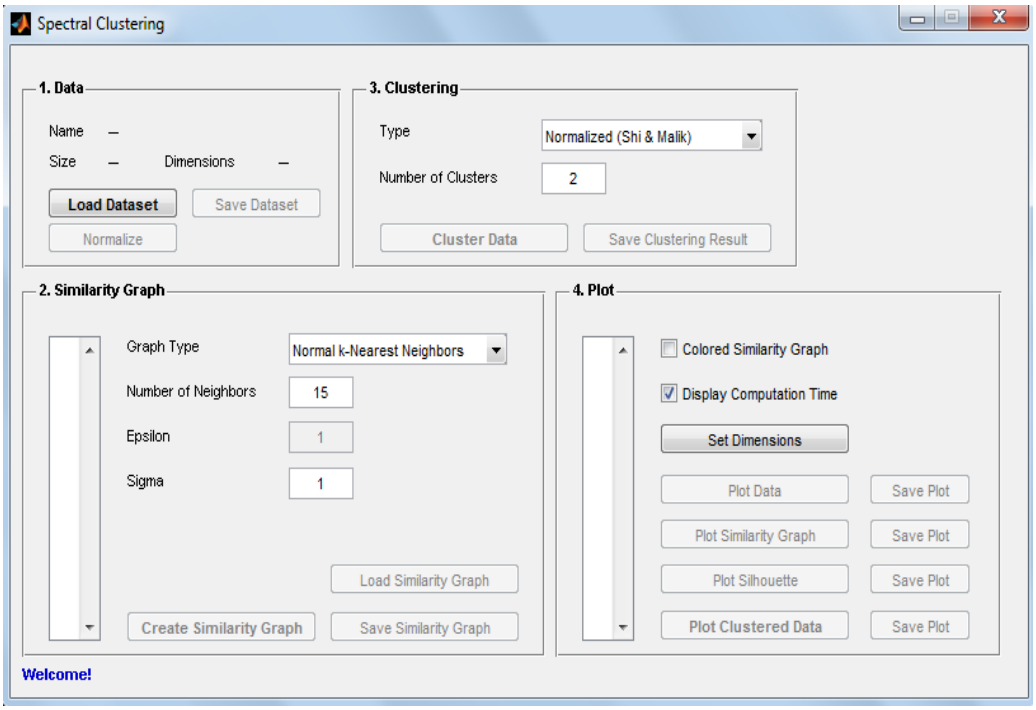


Fig.4.1 Interface for performing the clustering process

At the first step, dataset is loaded, from the first section of the interface, on which all the operations are to be performed. This dataset can be saved for future use also. The third button in the first section is to normalize the data, that is, the relation between the data are detected.

At the second step, affinity graphs are created. In this interface four kinds of affinity graphs can be created. The first type of the affinity graph that can be created is full affinity graph. In this approach, the traditional method of constructing the affinity graph is used. Only the sigma value is required for constructing this affinity graph as no partition of dataset will be there. The second kind of the affinity graph is normal k-nearest neighbour (KNN) affinity graph and the third kind is mutual KNN approach. In these approaches a random data point is selected to initiate the process and then the Euclidean distance of all the data points is calculated. The nearest neighbour graphs are supposed to model the local relation between each data point and its k-nearest neighbours. The normal KNN graph connects all vertices (v_i,v_j) if $v_i \in KNN(v_j)$ or $v_j \in KNN(v_i)$. The mutual KNN graph connects all vertices (v_i,v_j) if $v_i \in KNN(v_j)$ and $v_j \in KNN(v_i)$. For the epsilon affinity graphs only the number of epsilons are required as input.

At the third step clustering is performed. The three ways are introduced in this interface to perform this process. The first kind is unnormalized clustering, the second way is spectral clustering using Euclidean distance and the third way is normalized JW method.

The fourth step is to plot all the results. The results can be plotted in the matrix form and in star co-ordinates form. Out of the four buttons, the first button plots the original data points, the second button plots the similarity graphs that were created earlier, the third button is used to plot the silhouette value that is used to determine the quality of clusters and the fourth button plots the clustered data points.

The information about the status of the operations can be detected at the lower left corner of the interface.

6.3 RESULTS

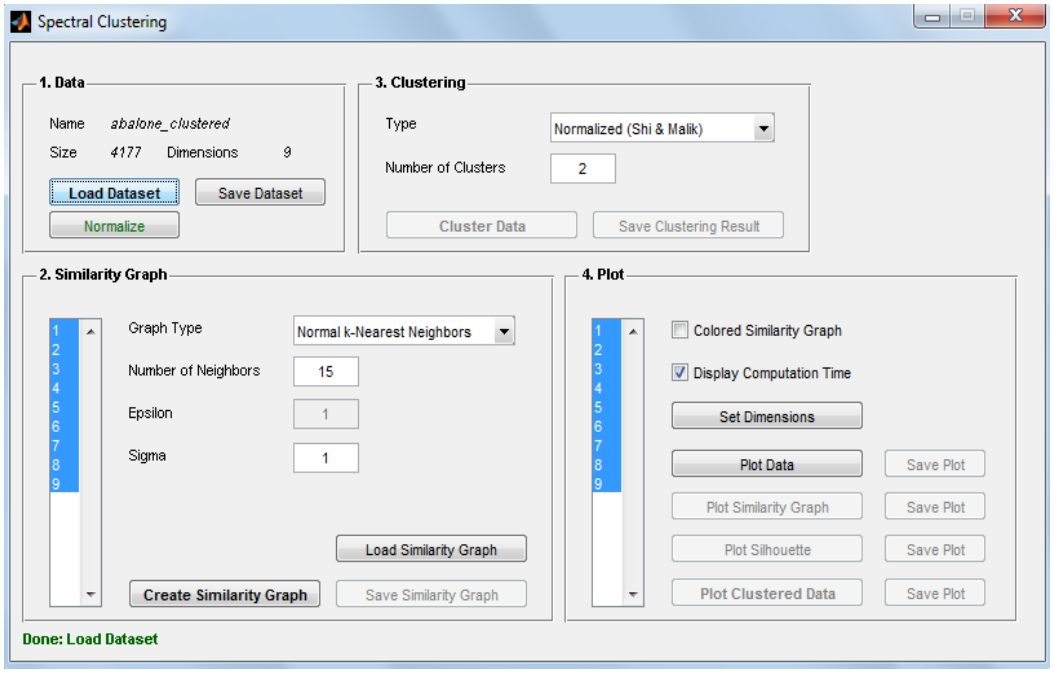


Fig.4.2 The loading of the dataset

As illustrated in above figure the status that the dataset has been loaded and its status can be seen at the lower left corner of the interface. The loaded dataset can be saved also for the further use using the "Save Dataset" button from the first part of the database. The next step is to normalize the dataset. After normalizing, the affinity graph is to be constructed.

As the dataset is now loaded so the plot data button will start working now and the original data points can be plotted now. The data points can be plotted in the matrix form and in the form of star co-ordinates. The following pop-up box will appear when one chose to plot the data.

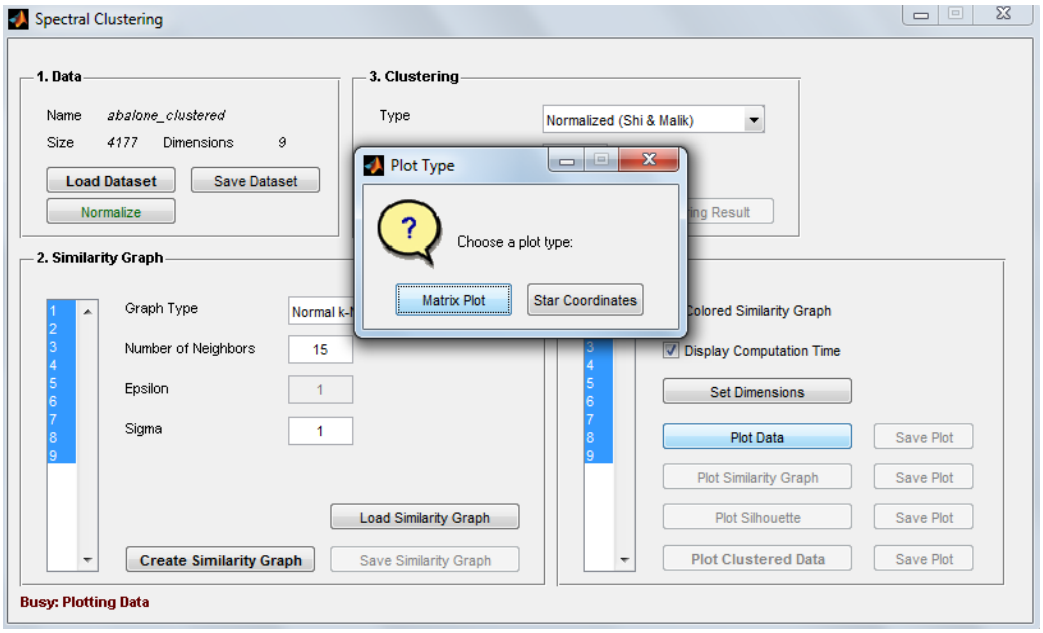


Fig.4.3 Plot type options

On choosing the Matrix Plot option the following screen will appear.

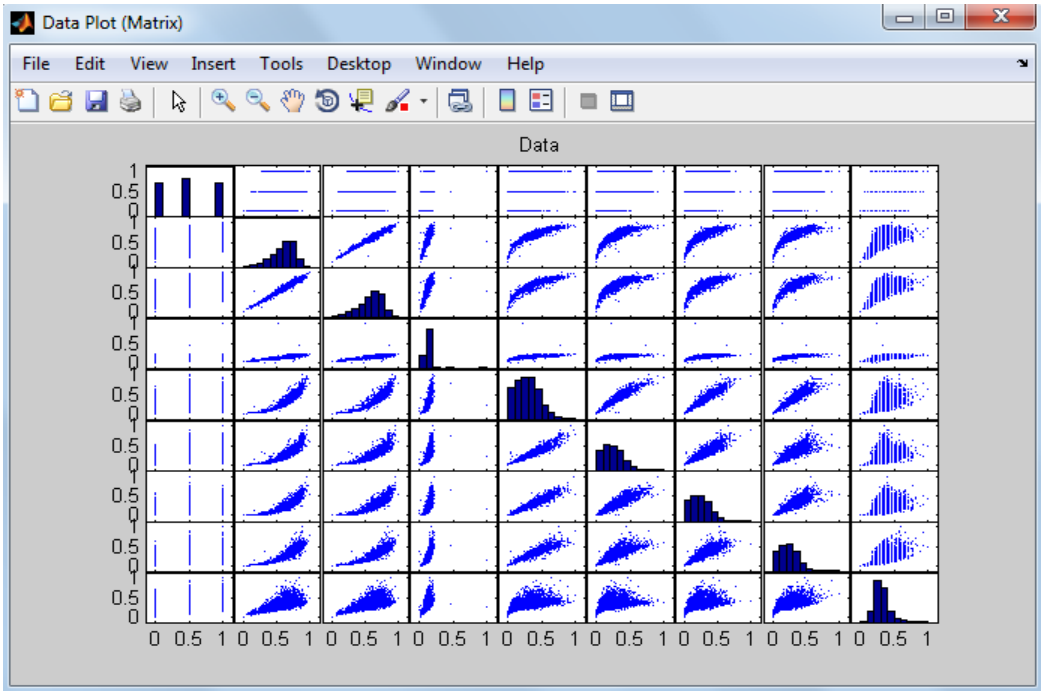


Fig.4.4 Matrix Plot of the Original data points

On choosing the Star Coordinates option the following screen will appear.

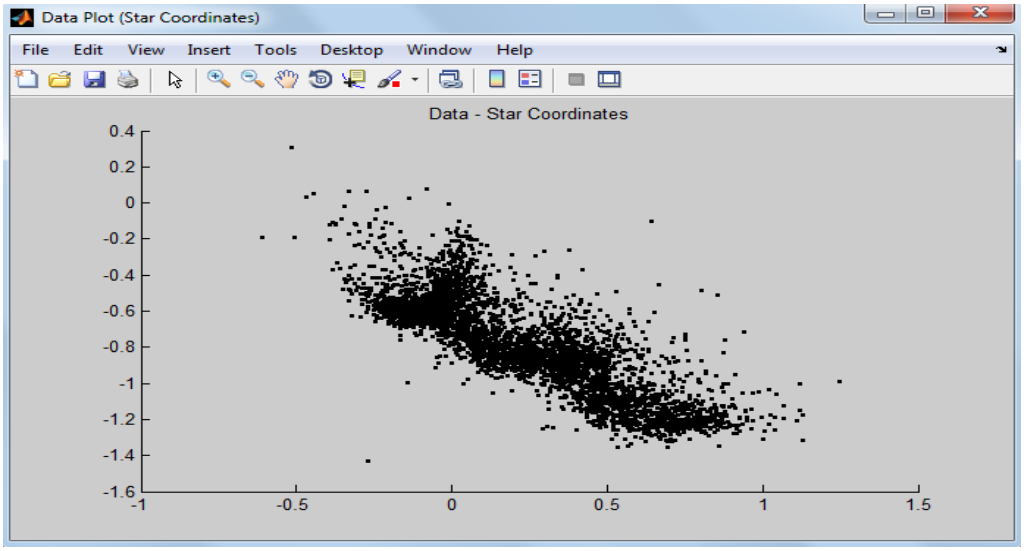


Fig.4.5 Original Data points in the form of Star Co-ordinates

After loading the dataset the next step is to calculate the affinity graph. First of all full affinity graph is calculated. Likewise the "Save Dataset button" the "Save Similarity Graph" button will save the created similarity graph for the further use and this similarity graph can be used again using "Load Similarity Graph" button. For calculating the full affinity graph the sigma value can be adjusted as per choice. The default value for the sigma is 1.

As soon as the similarity graph is created the "Plot Similarity Graph" button in the fourth section will also start working.

Full Affinity Graph

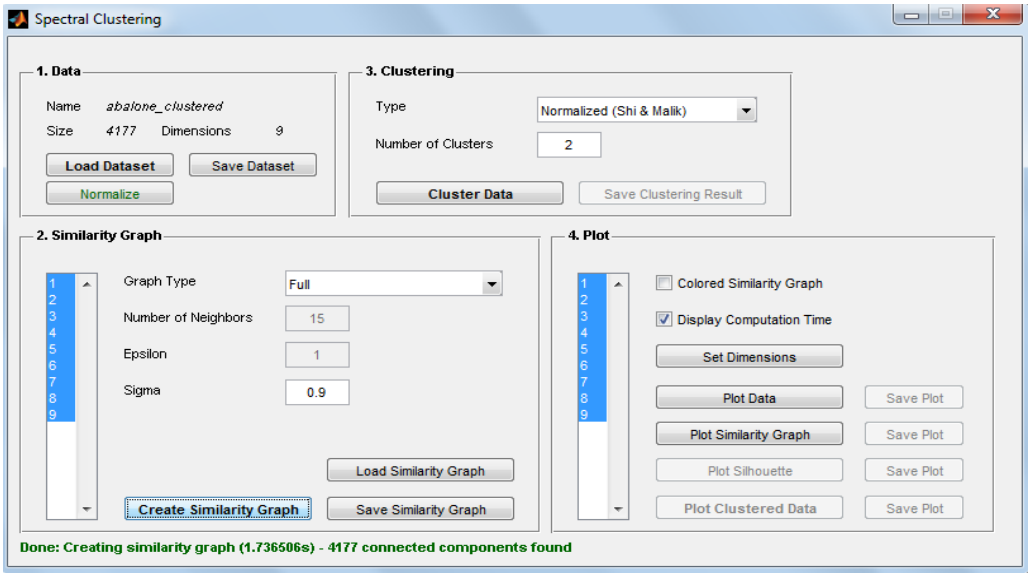


Fig.4.6 Time required to create the Full Affinity Graph

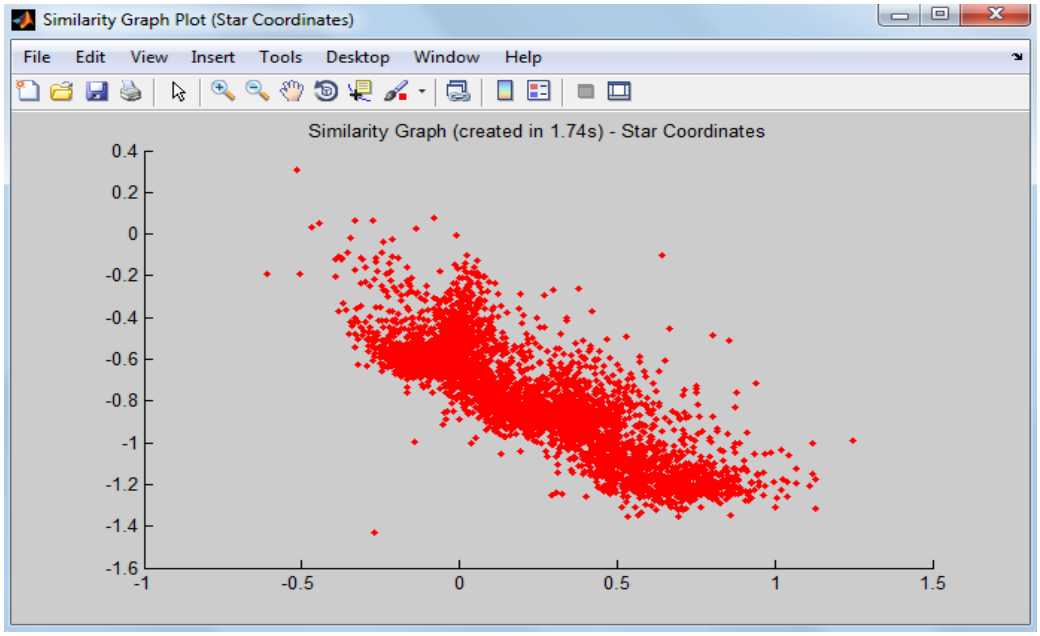


Fig.4.7 Full Affinity Graph

After the creation of the similarity graph the next step is to perform the clustering. This interface will allow to perform three kinds of clustering: unnormalized, SCUED and normalized JW clustering. The results of all the three kinds of clustering using full affinity graph are shown below:

Clustering Using Full Affinity Graph

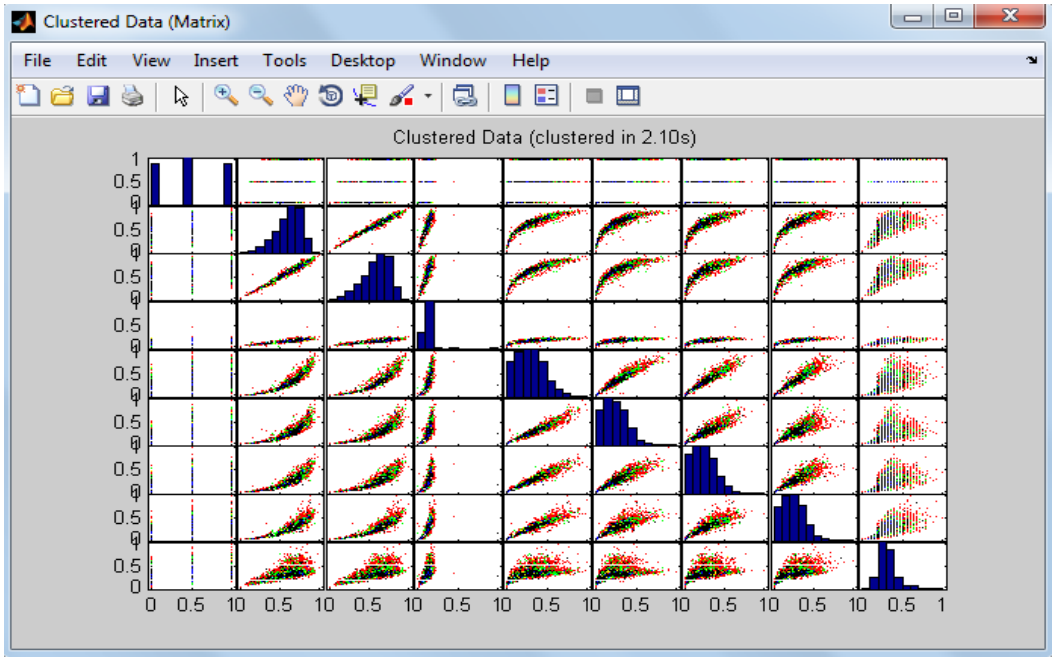


Fig.4.8 Matrix Plot of the Unnormalized Clustering using Full Affinity Graph

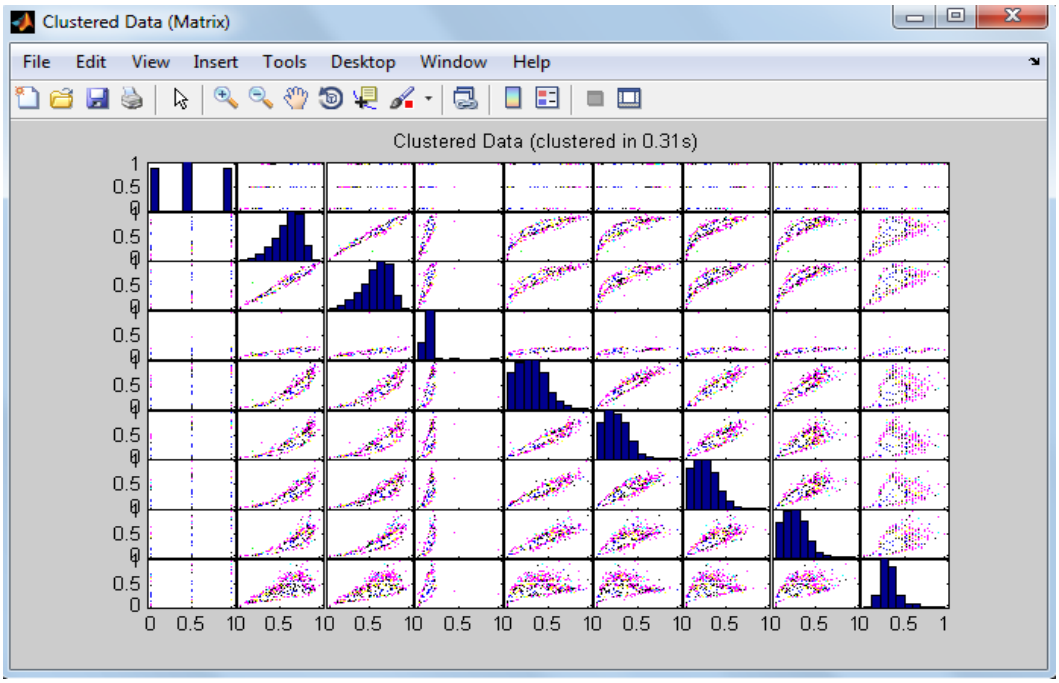


Fig.4.9 Matrix Plot of the SCUED using Full Affinity Graph

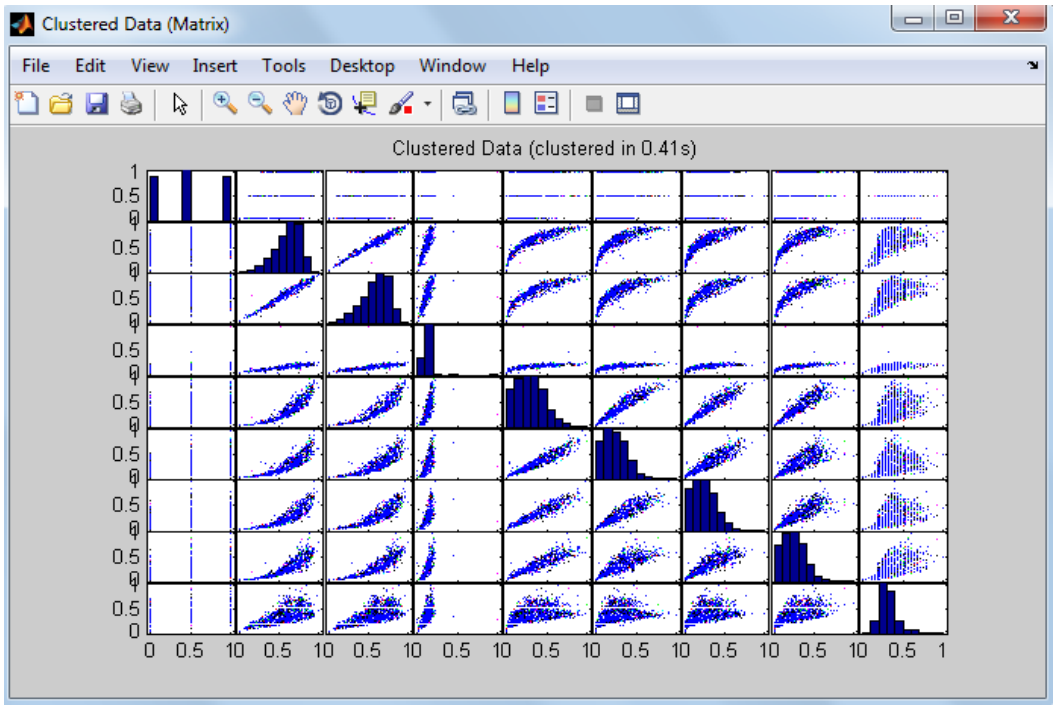


Fig.4.10 Matrix Plot of the Normalized JW clustering using Full Affinity Graph

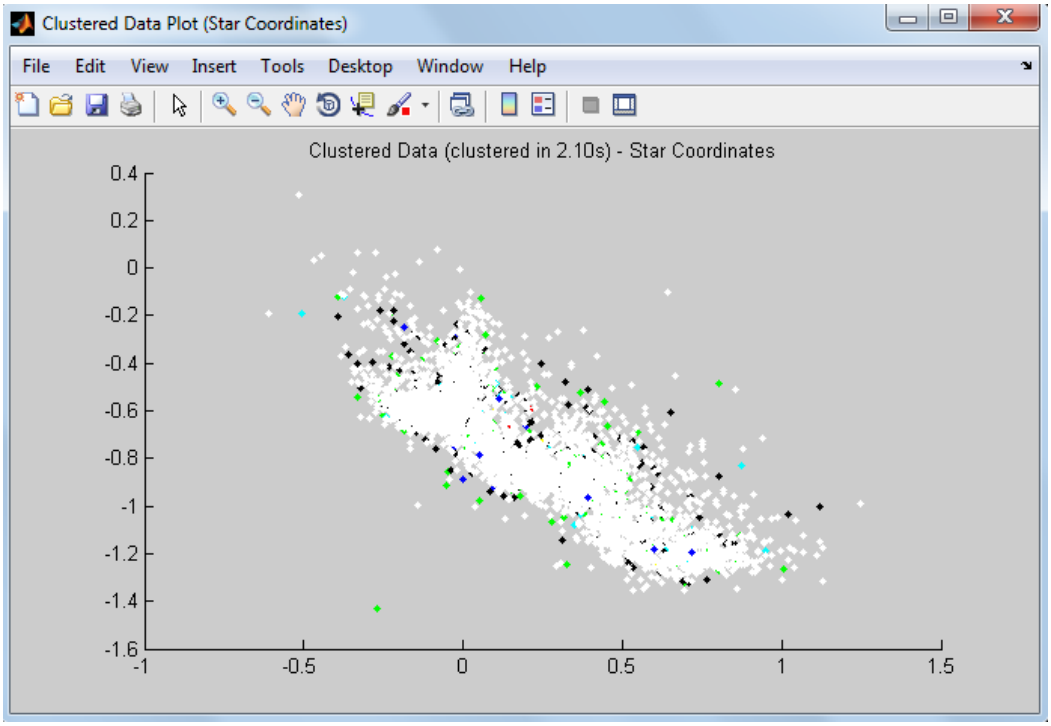


Fig.4.11 Unnormalized Clustered data in the form of Star Co-ordinates

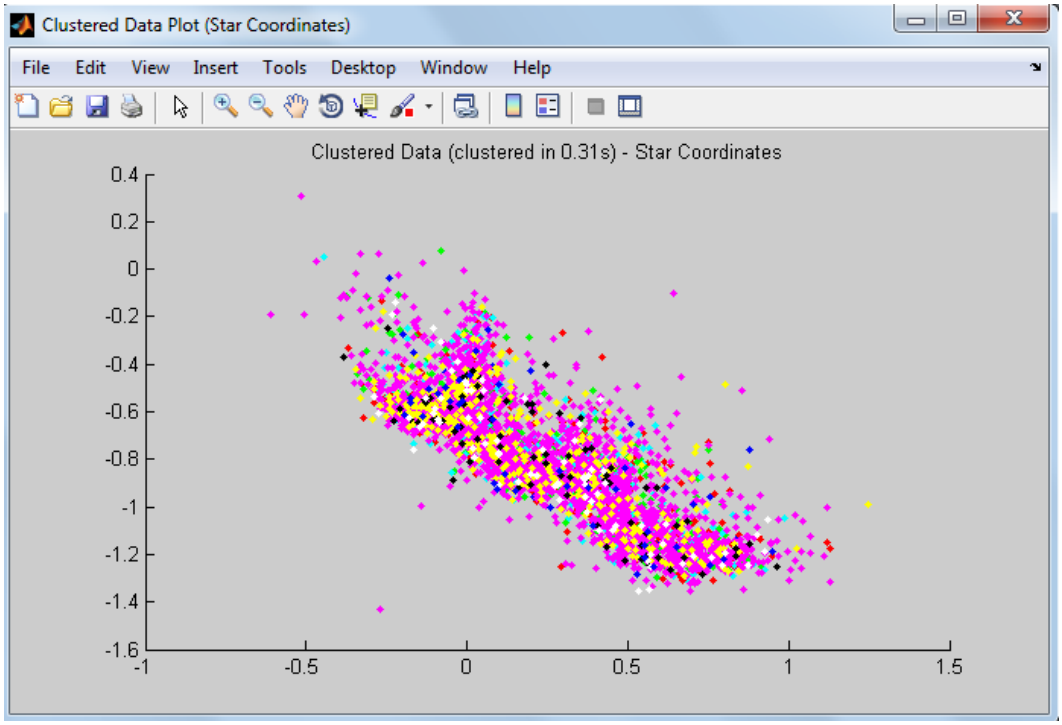


Fig.4.12 Data clustered with SCUED using full Affinity Graph in the form of Star Co-ordinates

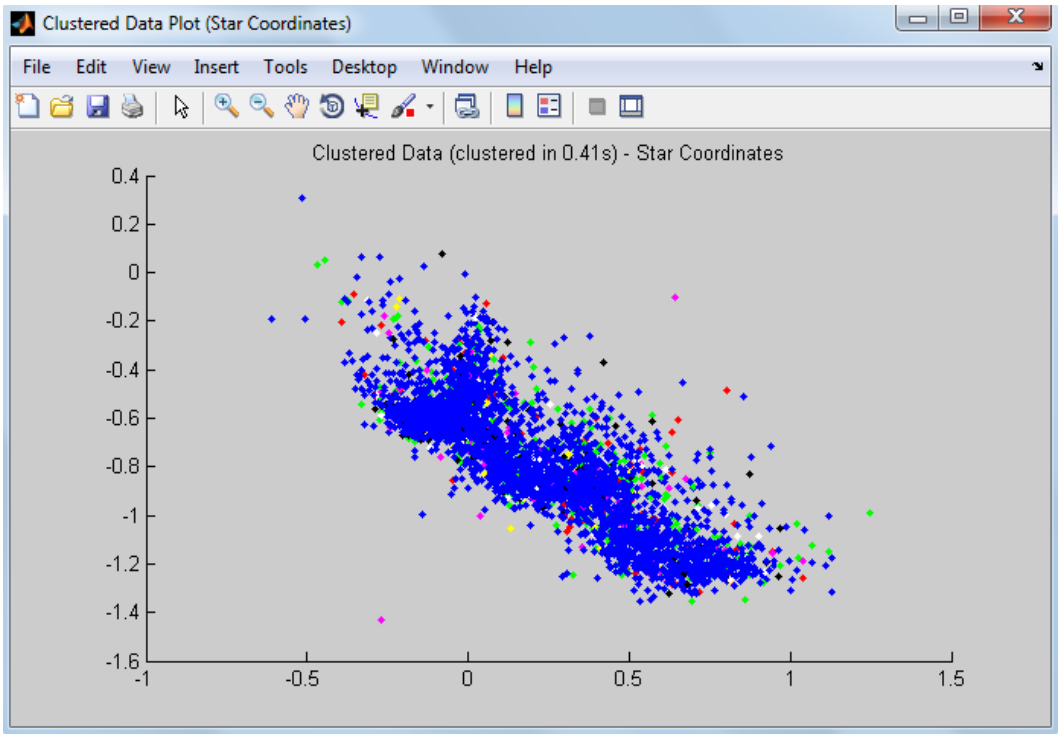


Fig.4.13 Data clustered with Normalized JW using Full Affinity Graph in the form of Star Co-ordinates

Silhouette values using Full Affinity Graph

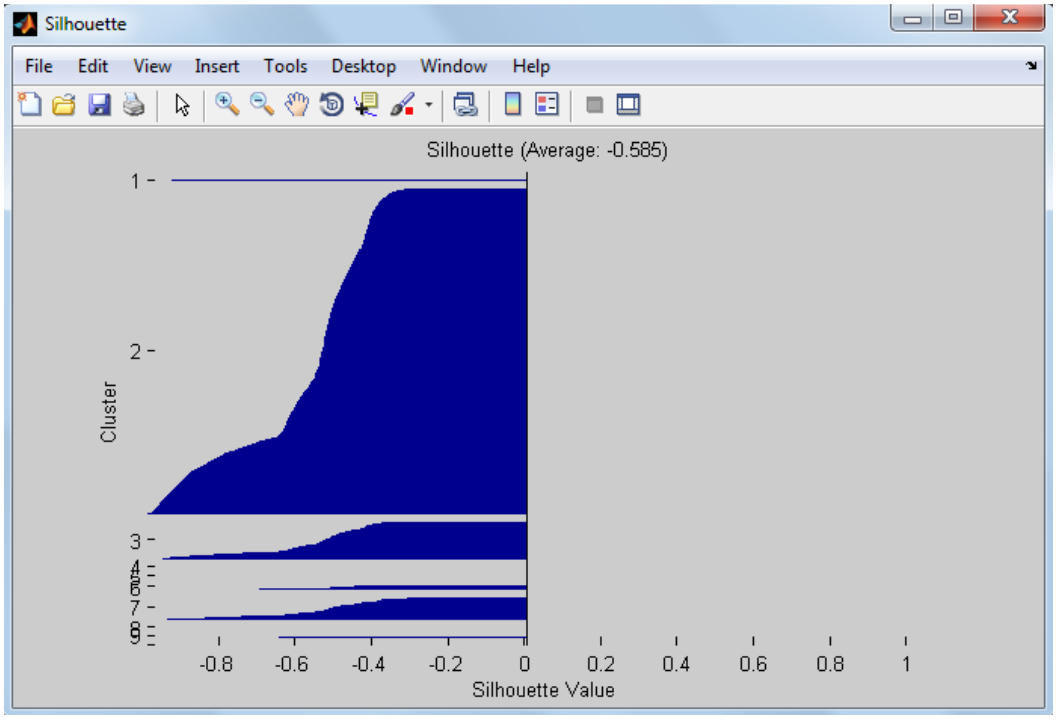


Fig.4.14 Silhouette value of Unnormalized Clustering using Full Affinity Graph

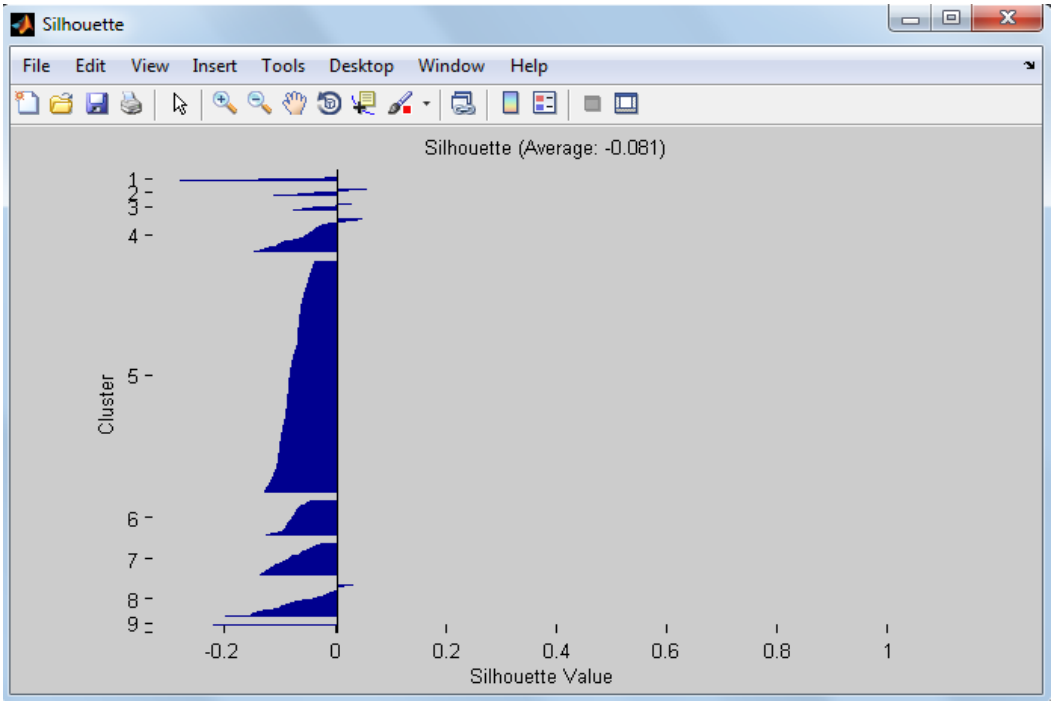


Fig.4.15 Silhouette Value of SCUED using Full Affinity Graph

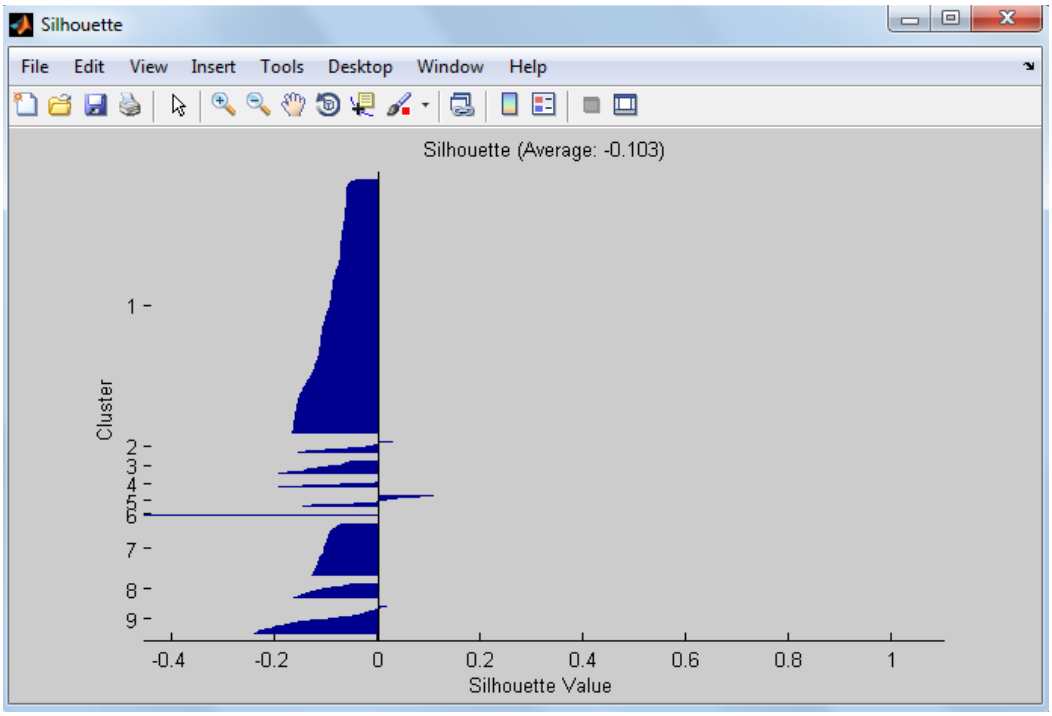


Fig.4.16 Silhouette Value of Normalized JW using Full Affinity Graph

This was the comparison of the three types of clustering using full affinity graph only. The higher silhouette value in the case of SCUED indicates that the data points clustered using

SCUED are more close to each other as compared to the unnormalized and normalized JW methods.

In the next section, clustering is done using Normal KNN Affinity graph.

Normal KNN Affinity Graph

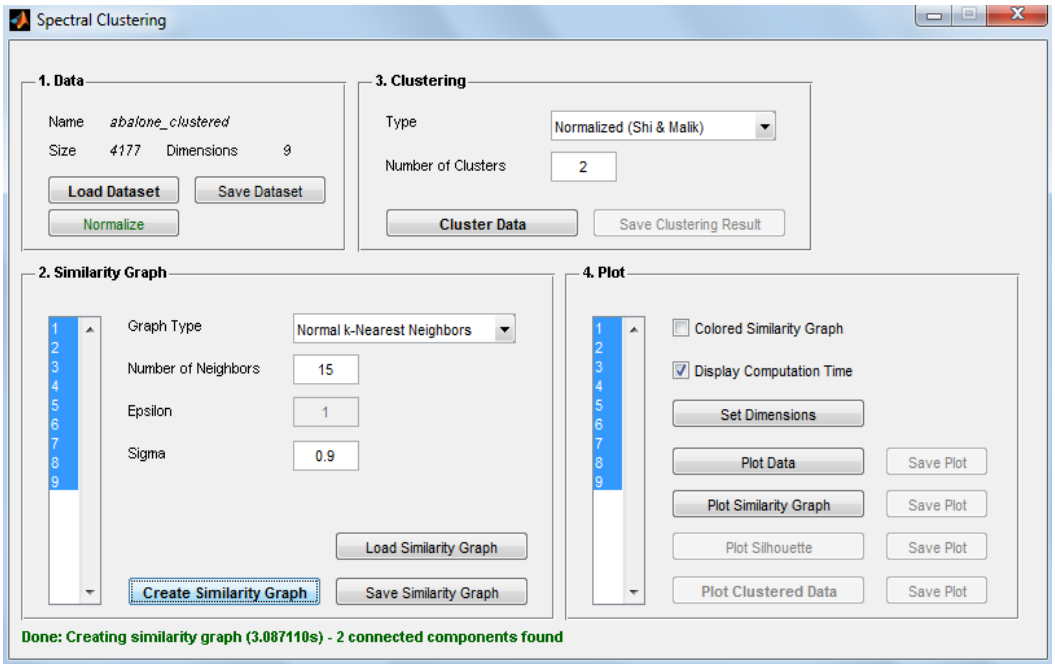


Fig.4.17 Time required to create the Normal KNN Affinity Graph

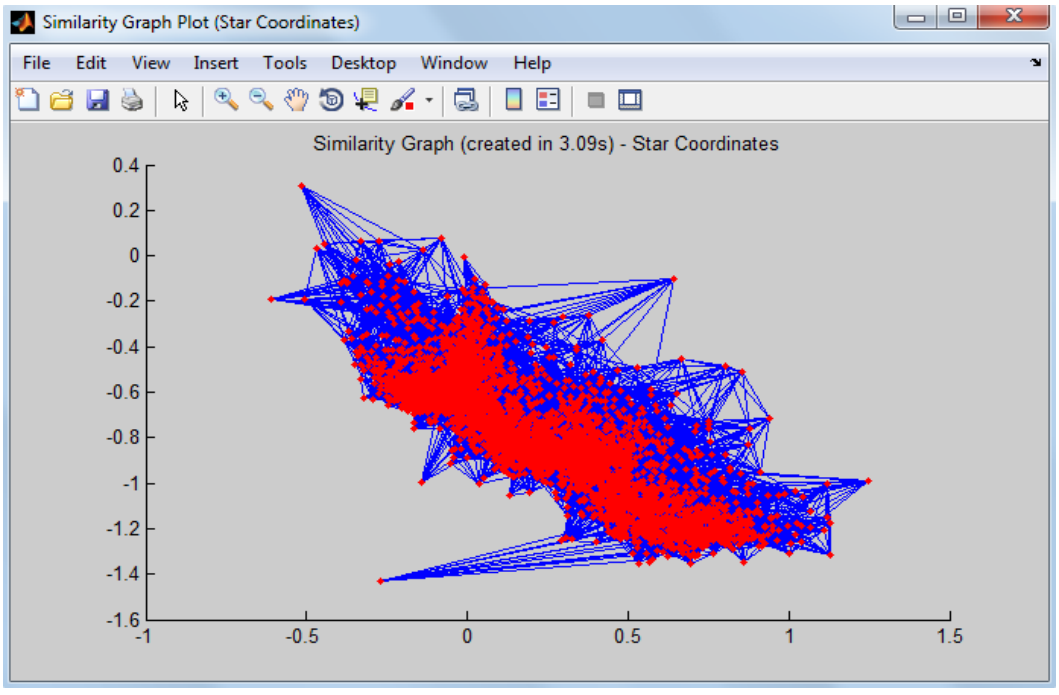


Fig.4.18 Normal KNN Affinity Graph

Clustering using Normal KNN Affinity Graph

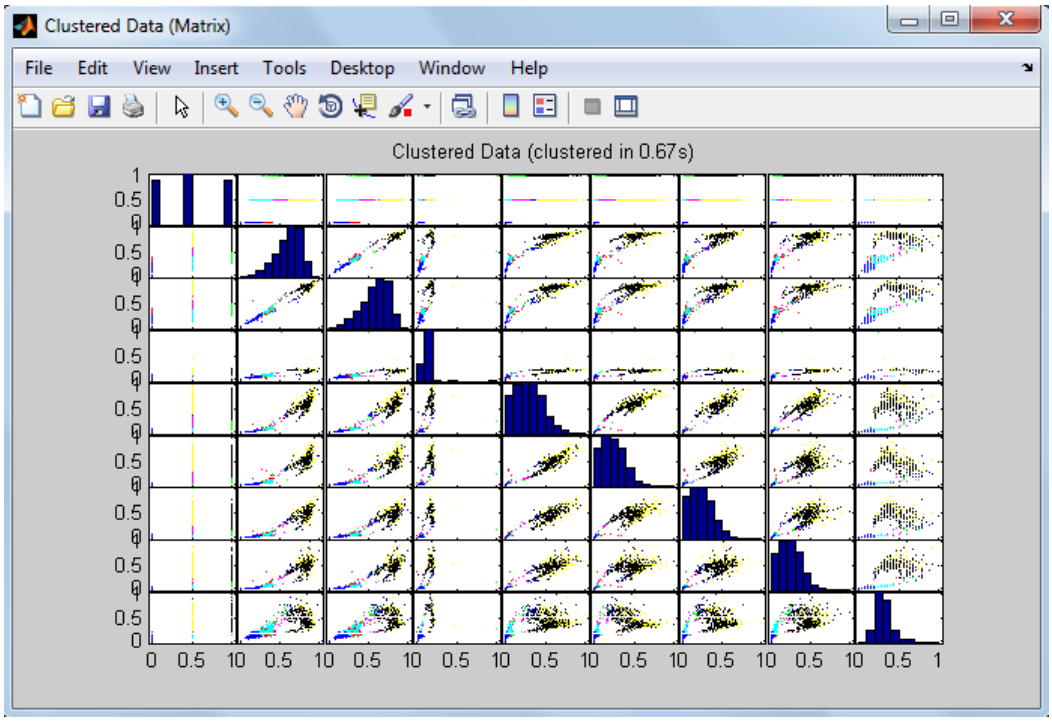


Fig.4.19 Matrix Plot of the Unnormalized Clustered Data Using Normal KNN Affinity Graph

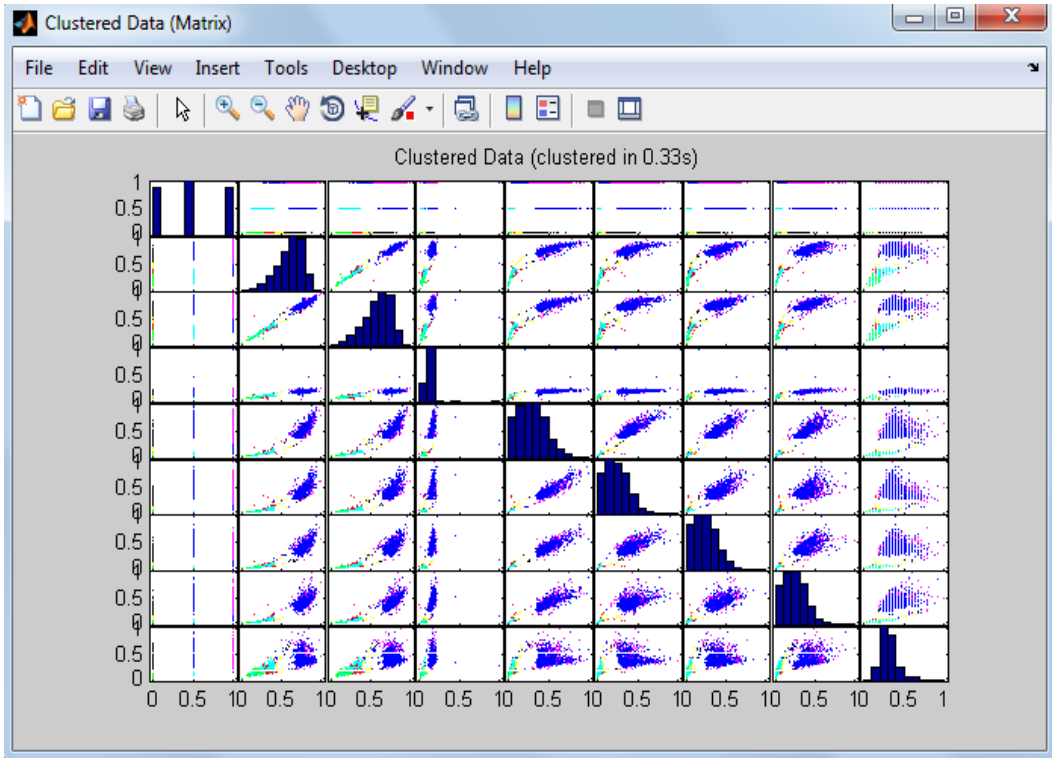


Fig.4.20 Matrix Plot of the SCUED Using Normal KNN Affinity Graph

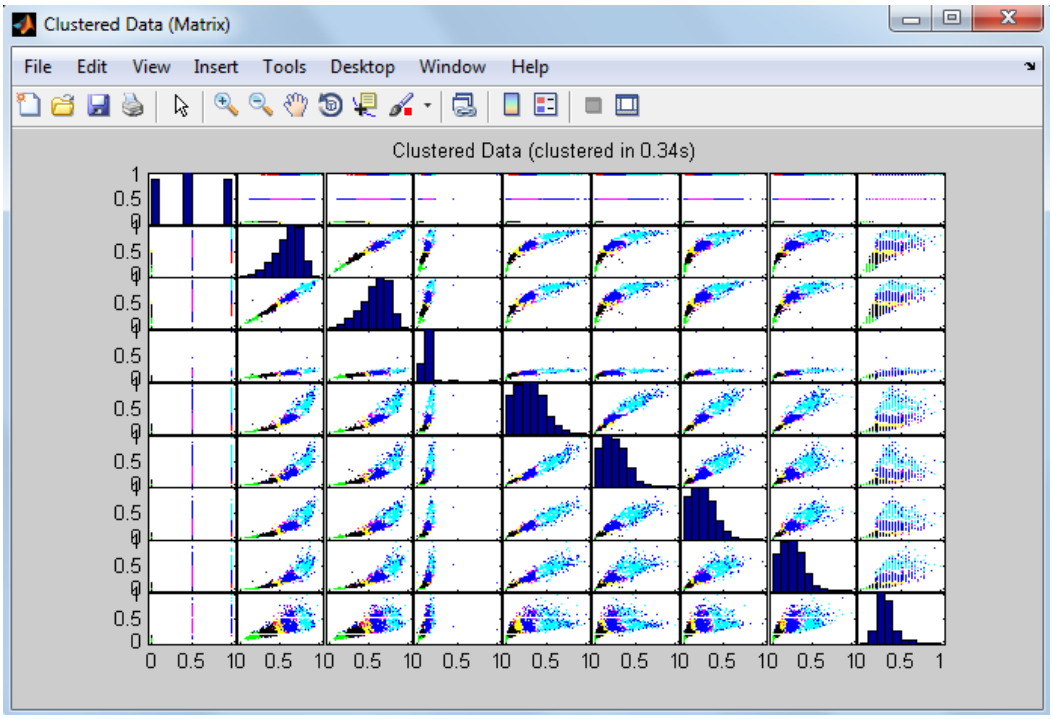


Fig.4.21 Matrix Plot of the Normalized JW Clustering using Normal KNN Affinity Graph

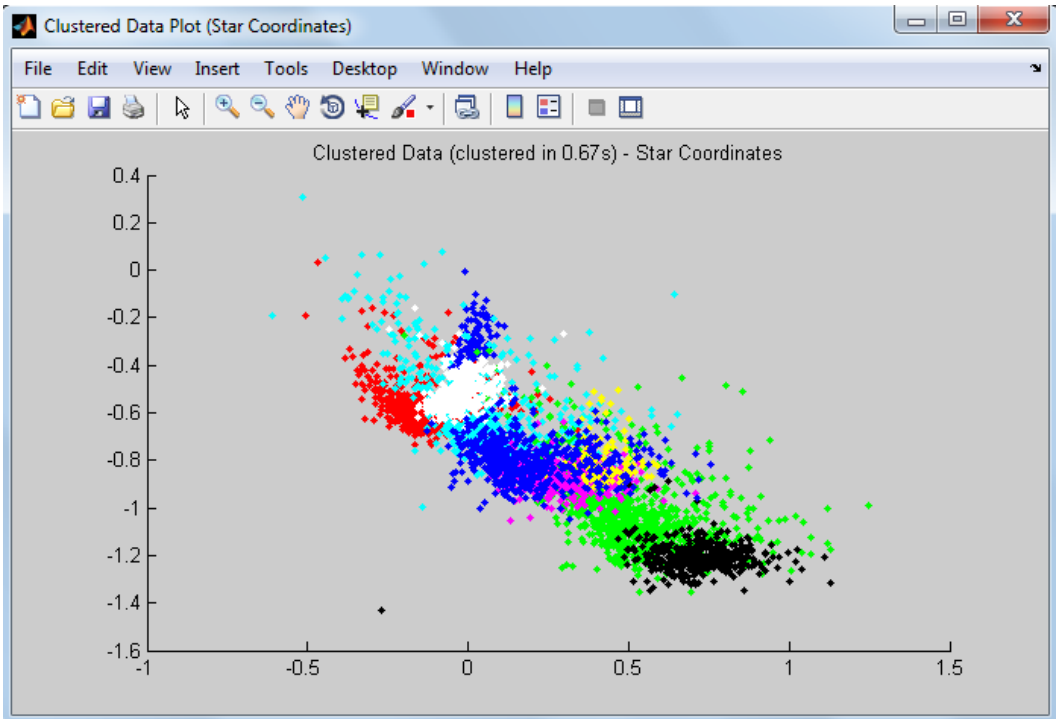


Fig.4.22 Unnormalized Clustering Using Normal KNN Affinity Graph in the form of Star Co-ordinates

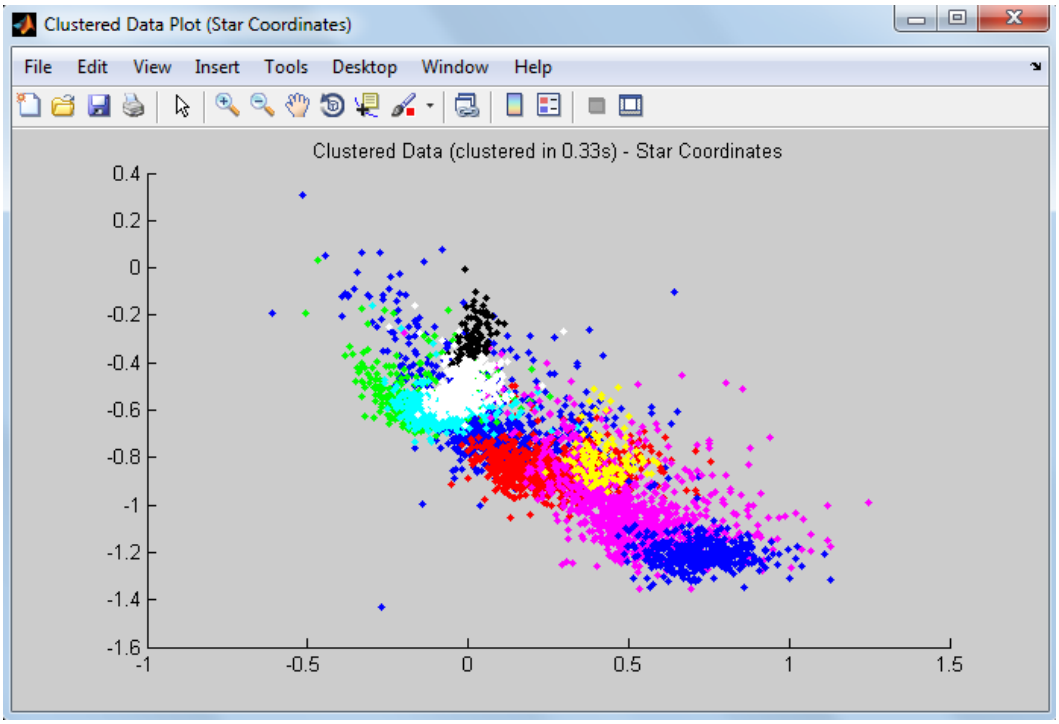


Fig.4.23 SCUED using Normal KNN Affinity Graph in the form of Star Co-ordinates

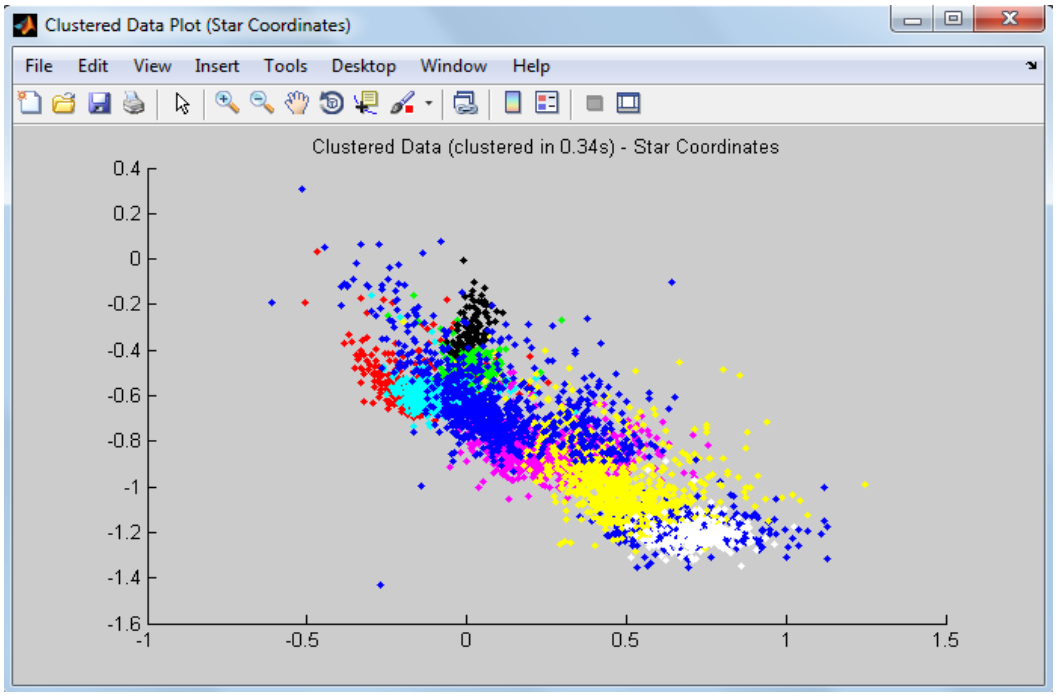


Fig.4.24 Normalized JW Clustering Using Normal KNN Affinity Graph in the form of Star Co-ordinates

Silhouette Values

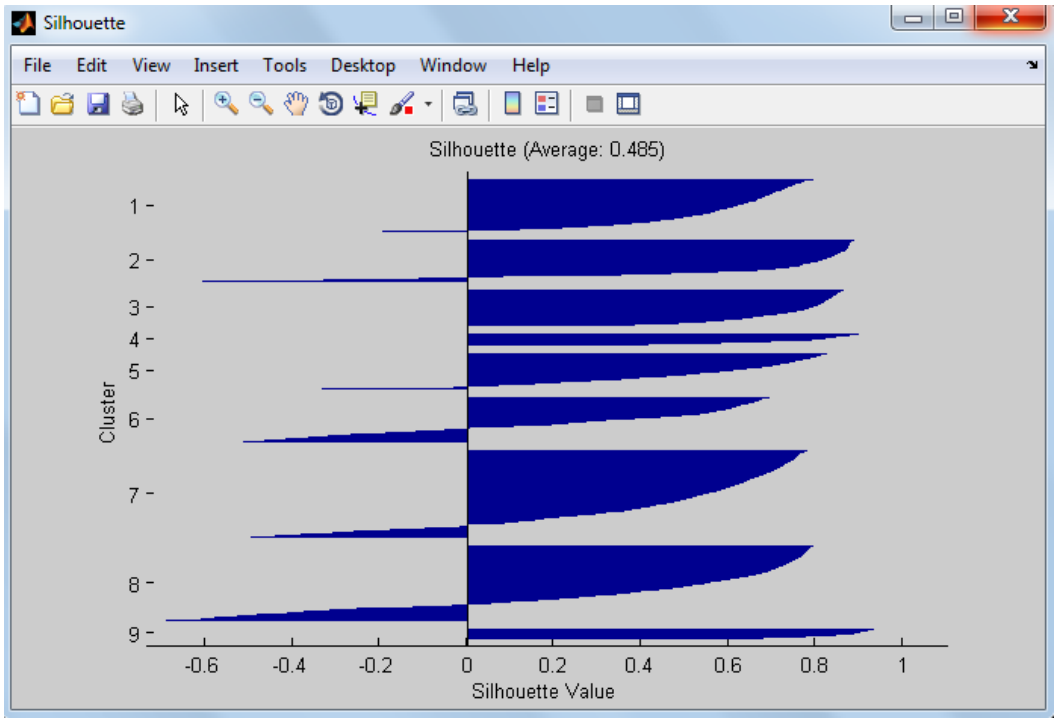


Fig.4.25 Silhouette Value of the Unnormalized Clustered data under Normal KNN Graph

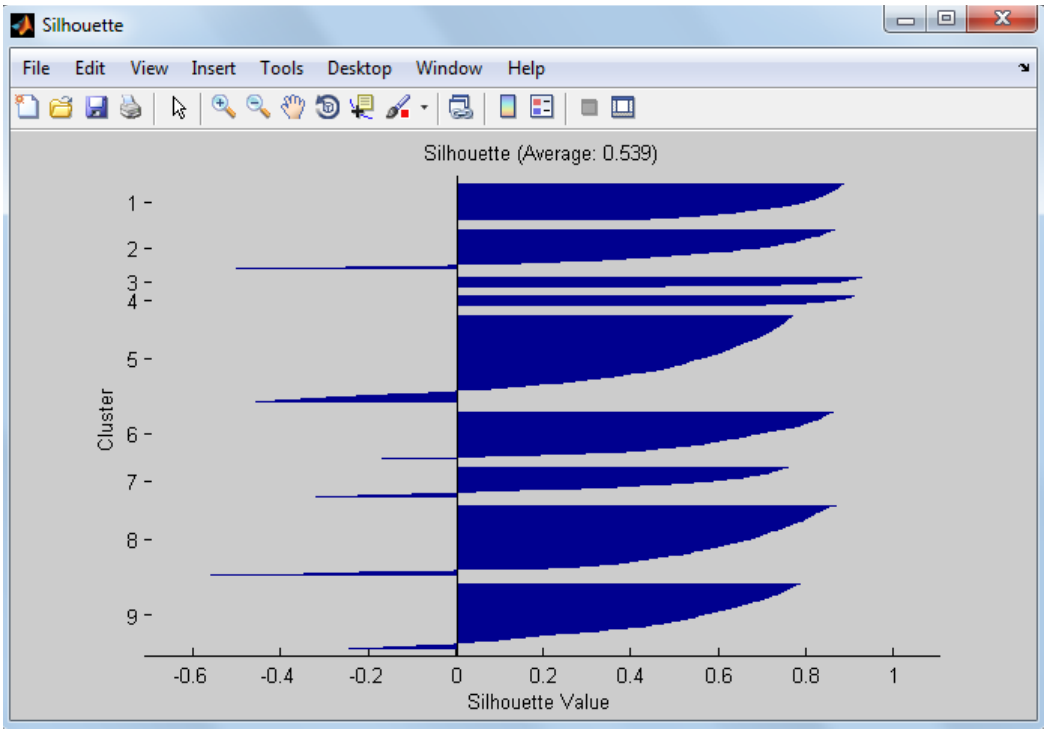


Fig.4.26 Silhouette Value of SCUED using Normal KNN Affinity Graph

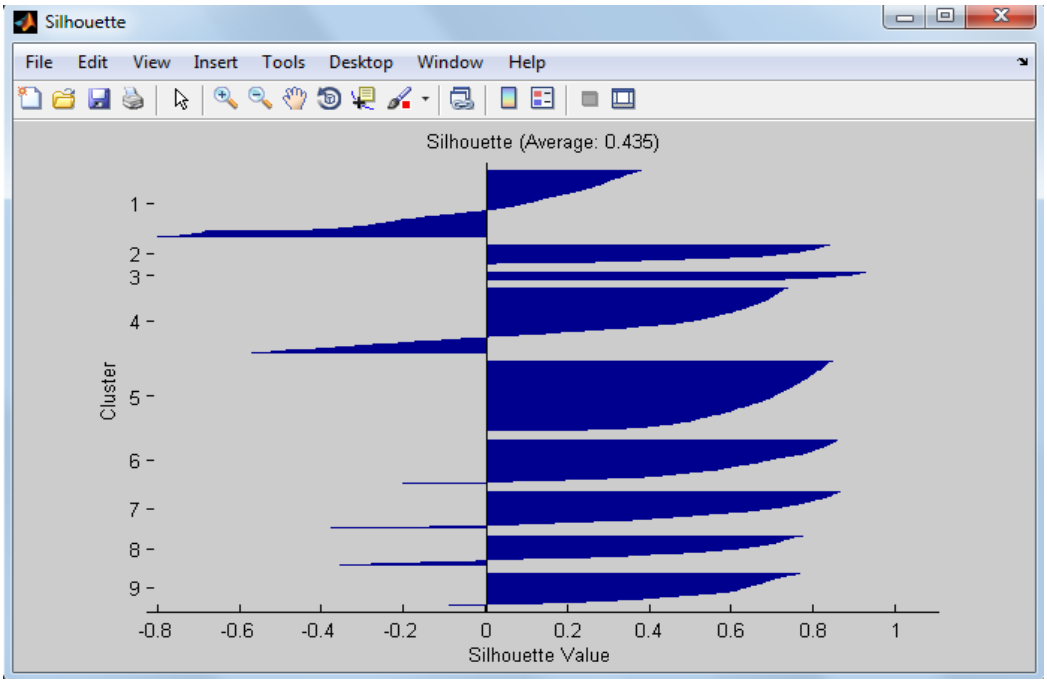


Fig.4.27 Silhouette Value of Normalized JW using Normal KNN Affinity Graph

Again, the silhouette value for the SCUED is greater than the other two methods and the time required for performing the clustering is less in case of SCUED; so, in this approach also the SCUED outperforms the other two methods.

In the next section, the clustering is performed using Mutual KNN Affinity Graph.

Mutual KNN Affinity Graph

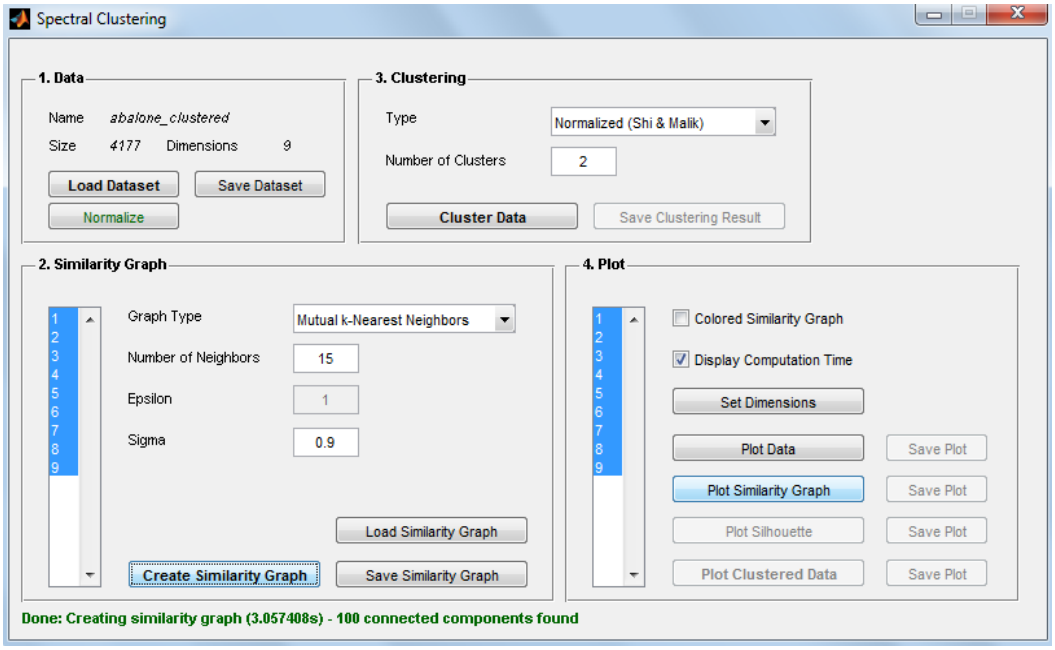


Fig.4.28 Time required to create the Mutual KNN Affinity Graph

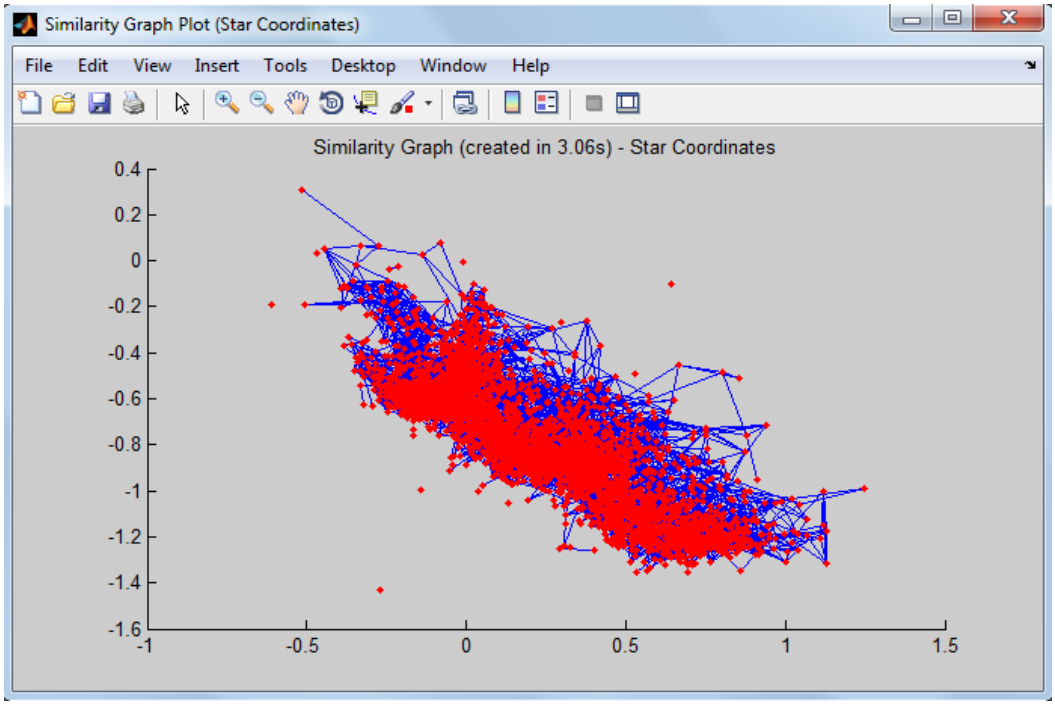


Fig.4.29 Mutual KNN Affinity Graph

Clustering Using Mutual KNN Affinity Graph

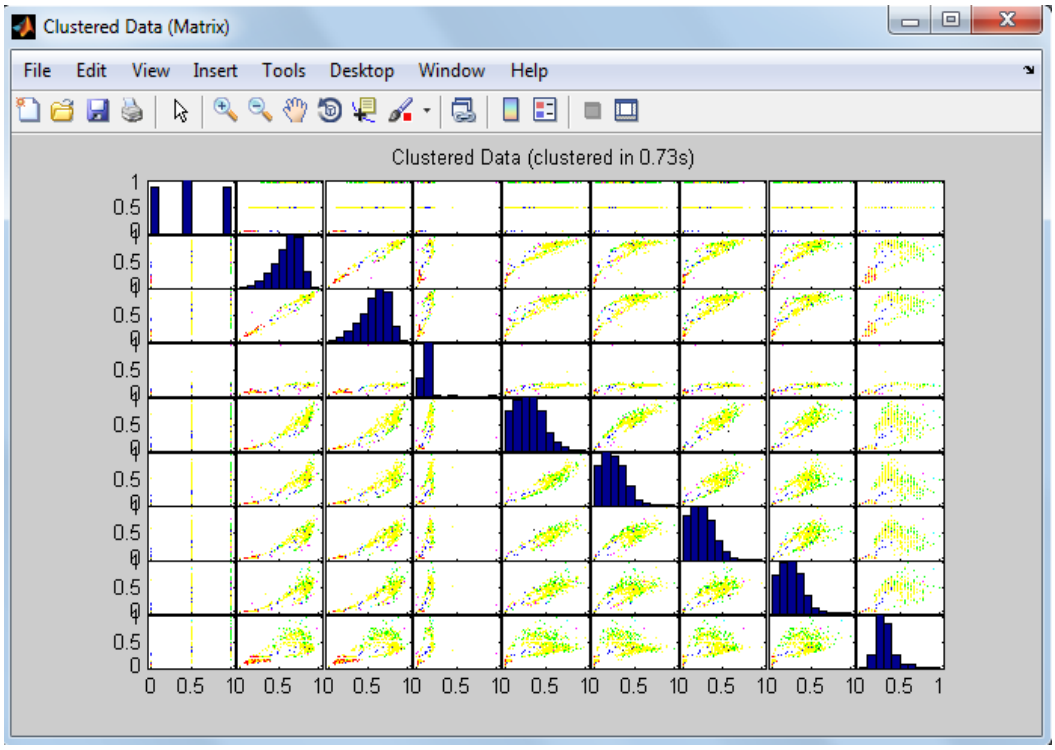


Fig.4.30 Matrix Plot of the Unnormalized Clustering using Mutual KNN Affinity Graph

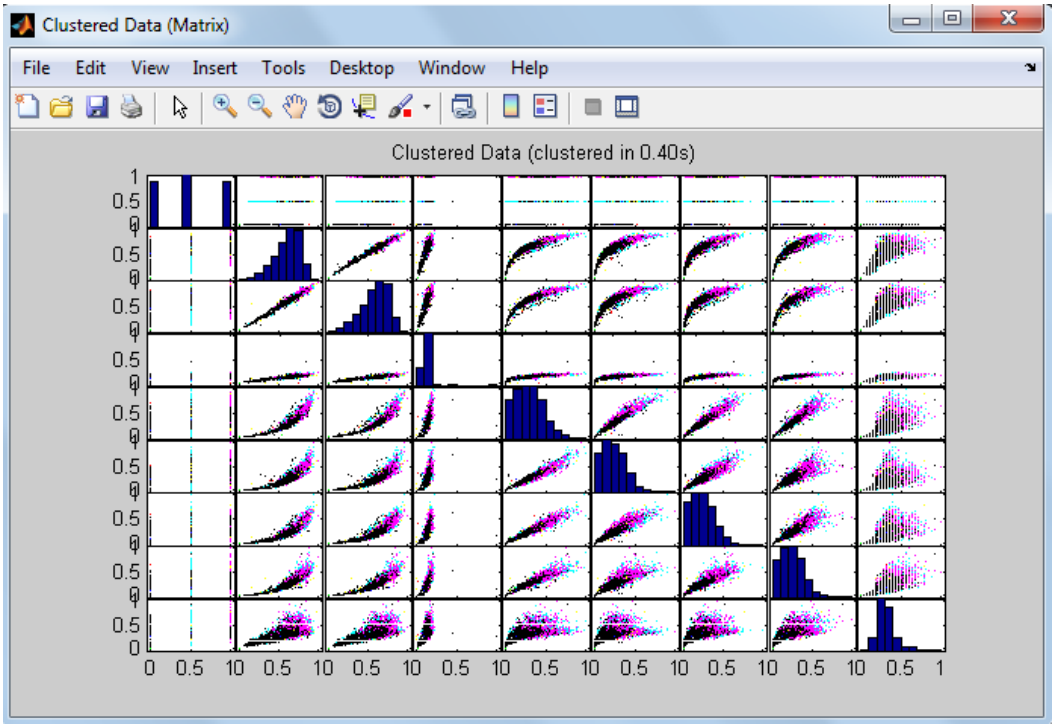


Fig.4.31 Matrix Plot of the SCUED using Mutual KNN Affinity Graph

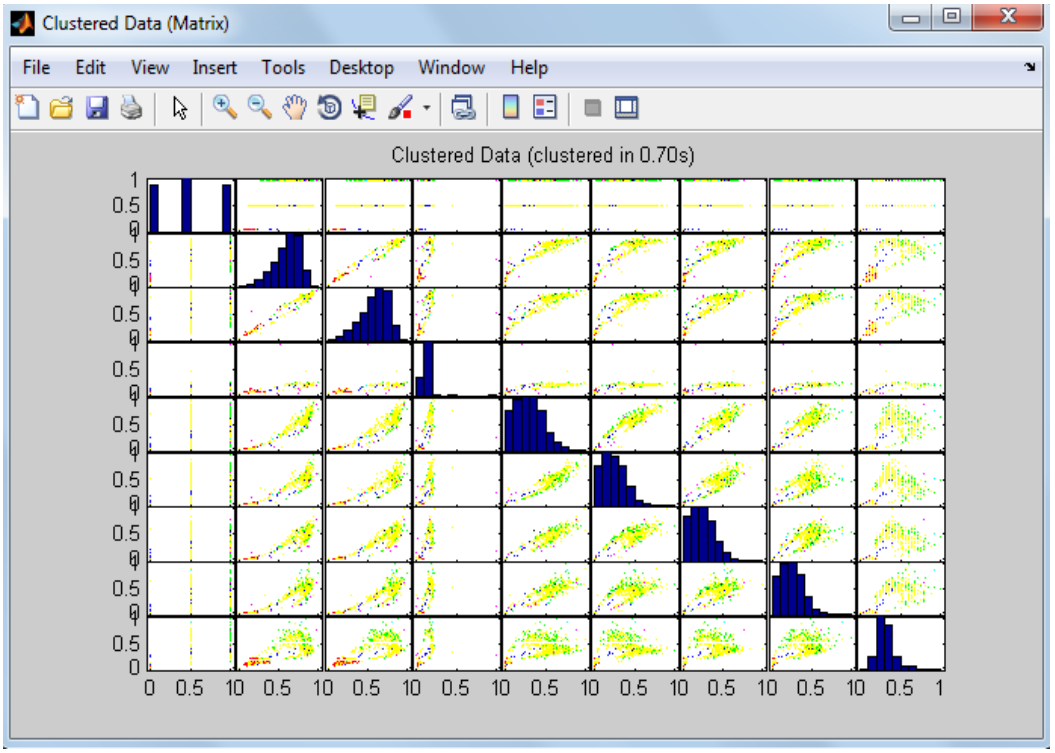


Fig.4.32 Matrix Plot of the Normalized JW Clustering using Mutual KNN Affinity Graph

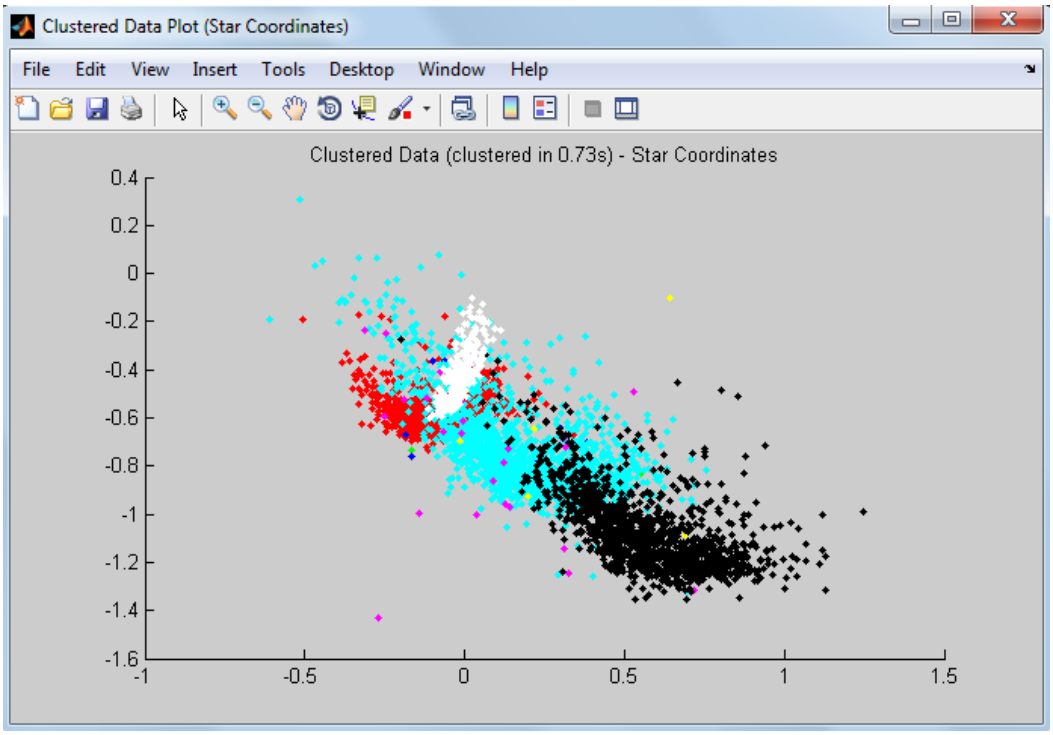


Fig.4.33 Unnormalized Clustering using Mutual KNN in the form of Star Co-ordinates

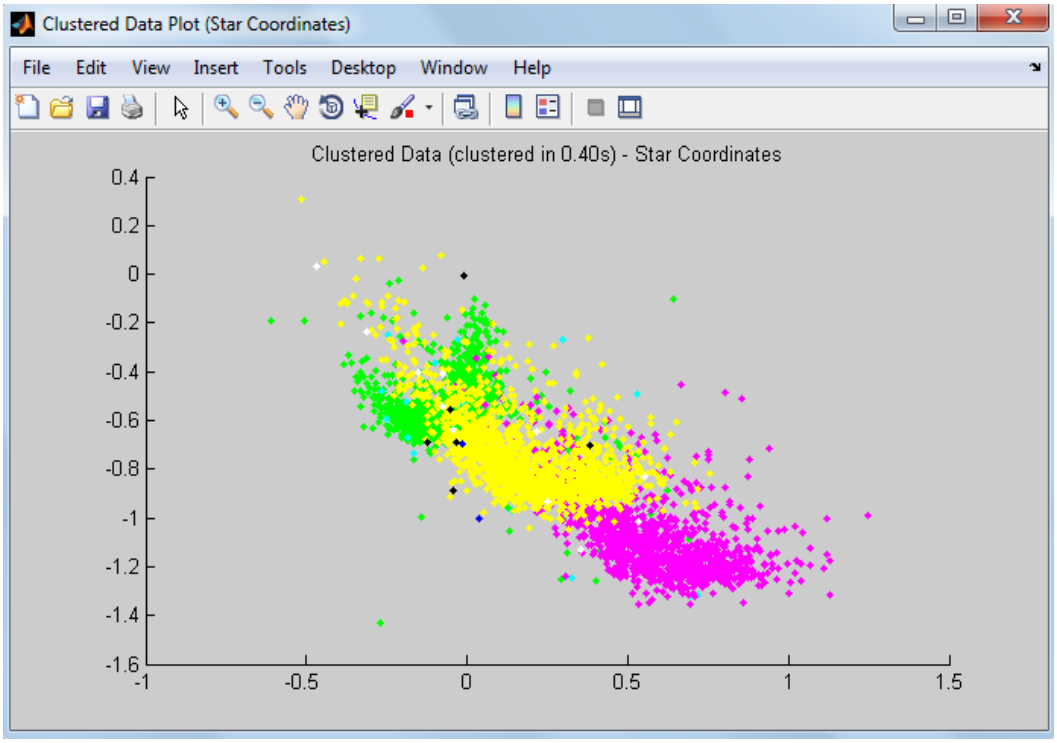


Fig.4.34 SCUED using Mutual KNN Affinity Graph in the form of Star Co-ordinates

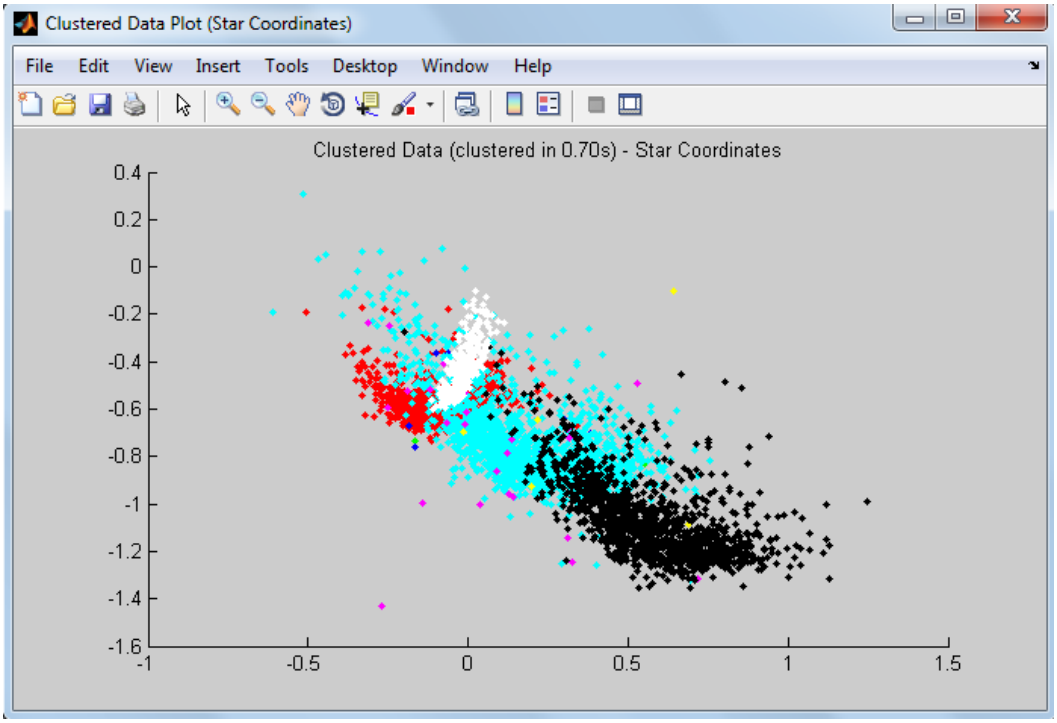


Fig.4.35 Normalized JW Clustering using Mutual KNN in the form of Star Co-ordinates

Silhouettes Values

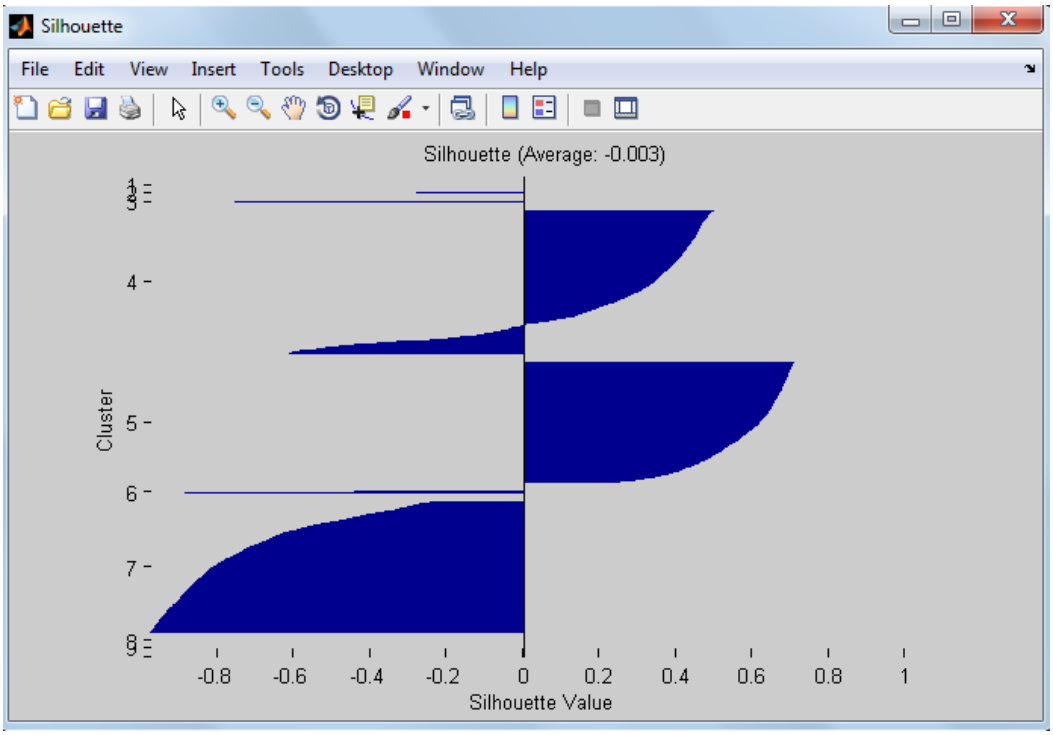


Fig.4.36 Silhouette Value for the Unnormalized Clustering using Mutual KNN Affinity Graph

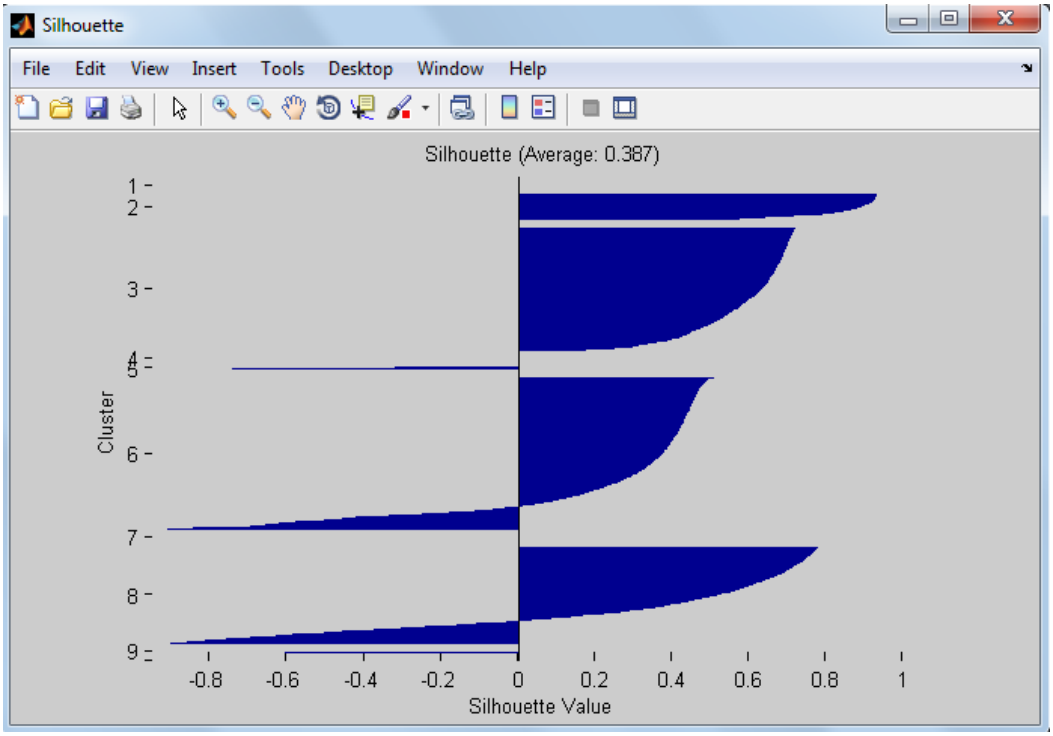


Fig.4.37 Silhouette Value for the SCUED using Mutual KNN Affinity Graph

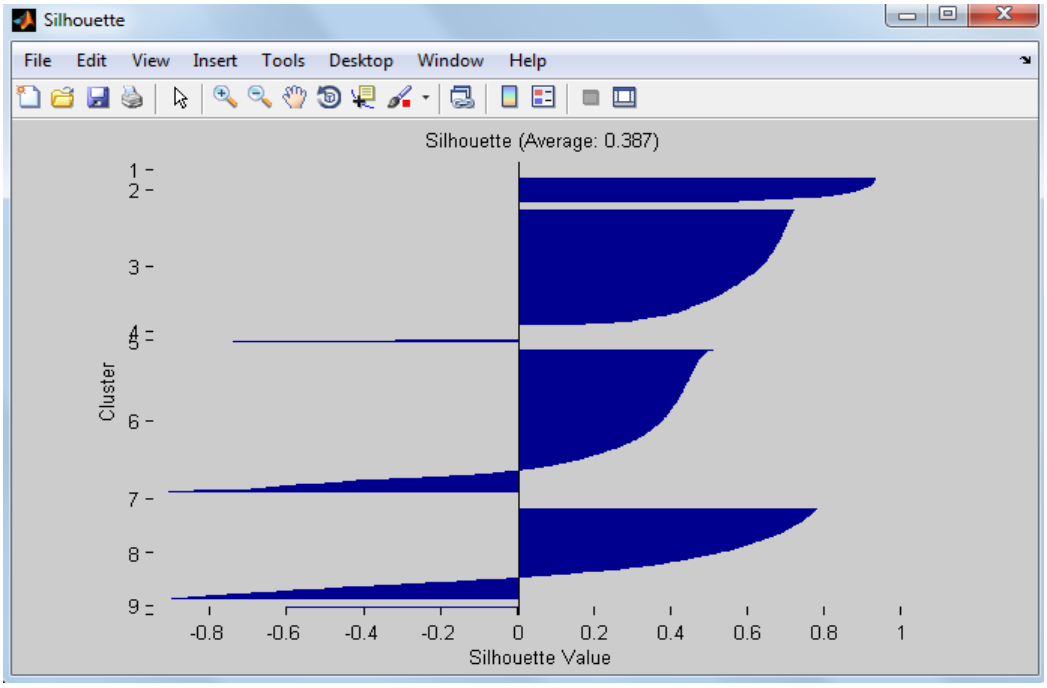


Fig.4.38 Silhouette Value For Normalized JW using Mutual KNN Affinity Graph

The above mentioned results again prove that the SCUED is helpful in reducing the time required for performing clustering and improving the quality of the clusters. Although the

silhouette values for the SCUED and Normalized JW are similar but the time taken by Normalized JW is more compared to SCUED.

Epsilon Affinity Graph

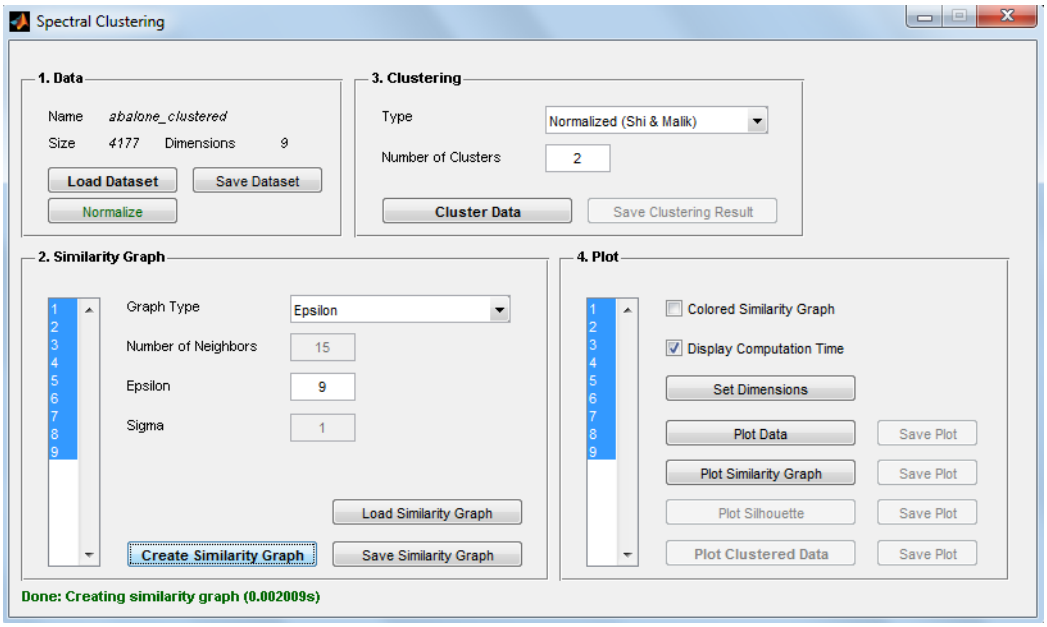


Fig.4.39 Time required to create the Epsilon Affinity Graph

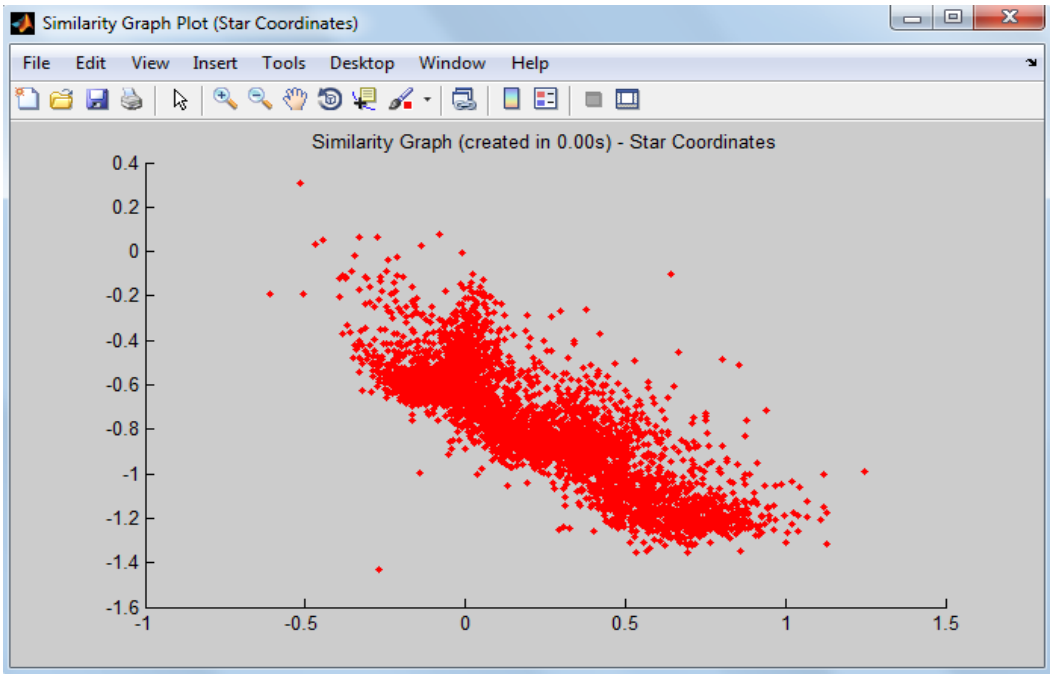


Fig.4.40 Epsilon Affinity Graph

Clustering Using Epsilon Affinity Graph

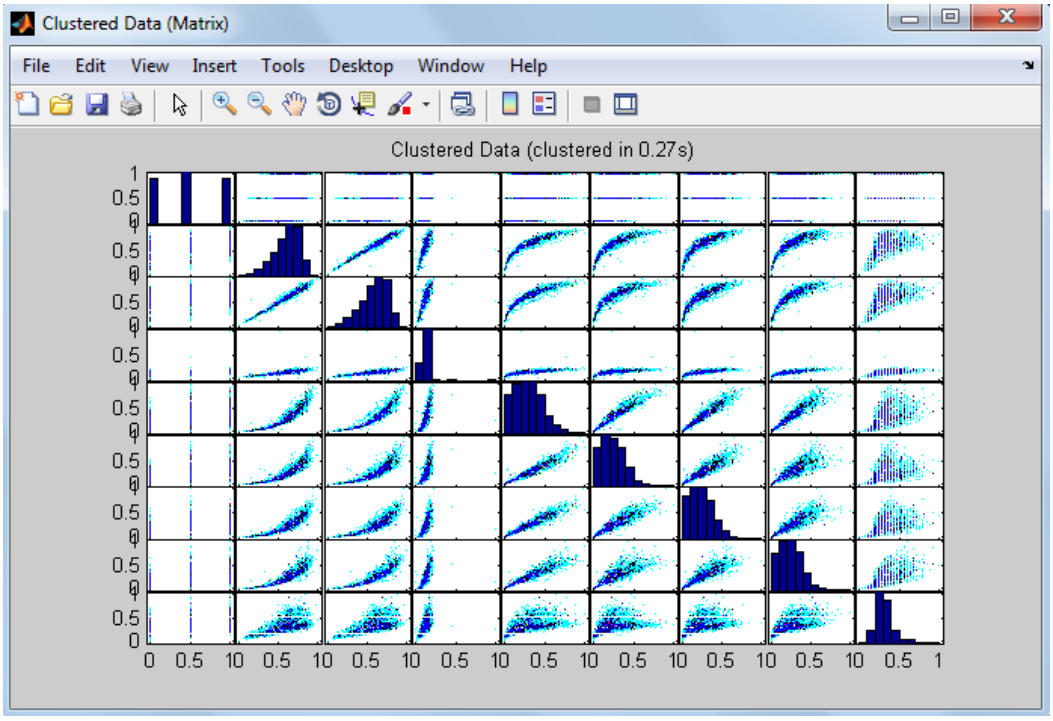


Fig.4.41 Matrix Plot of the Unnormalized Clustered Data using Epsilon Affinity Graph

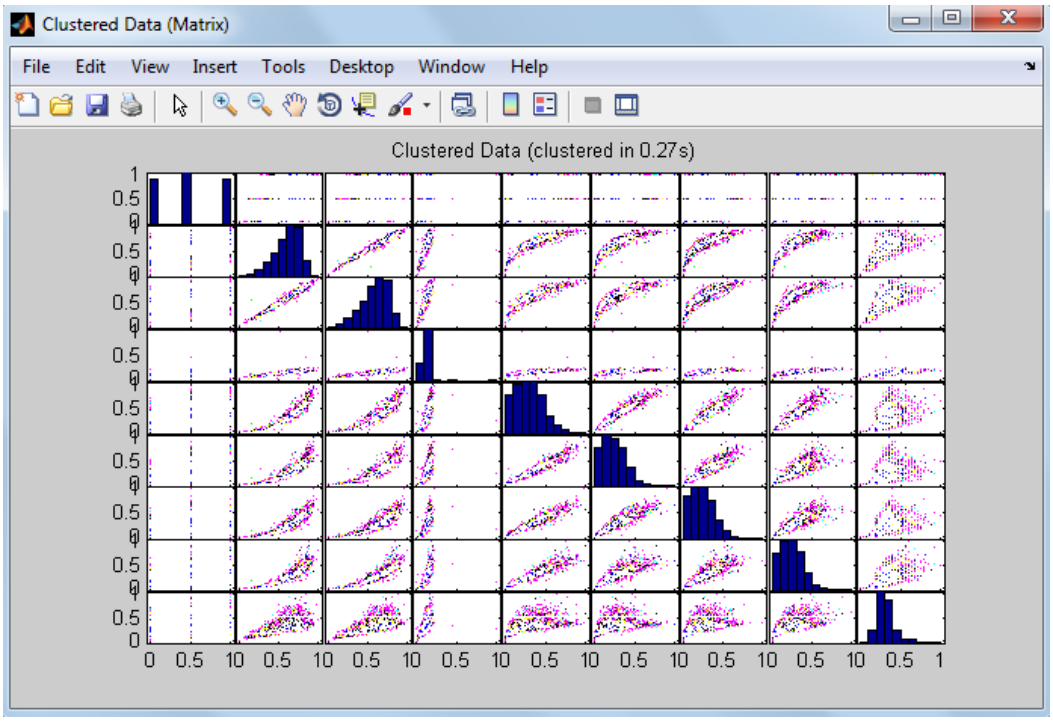


Fig.4.42 Matrix Plot of the SCUED using Epsilon Affinity Graph

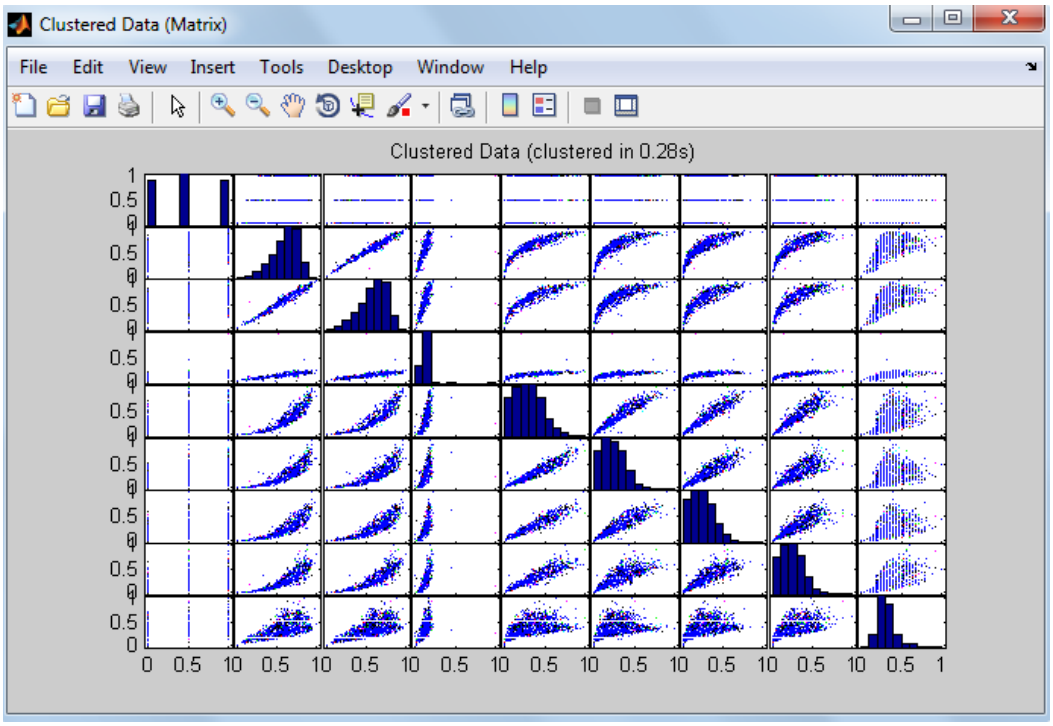


Fig.4.43 Matrix Plot of the Normalized JW clustering using Epsilon Affinity Graph

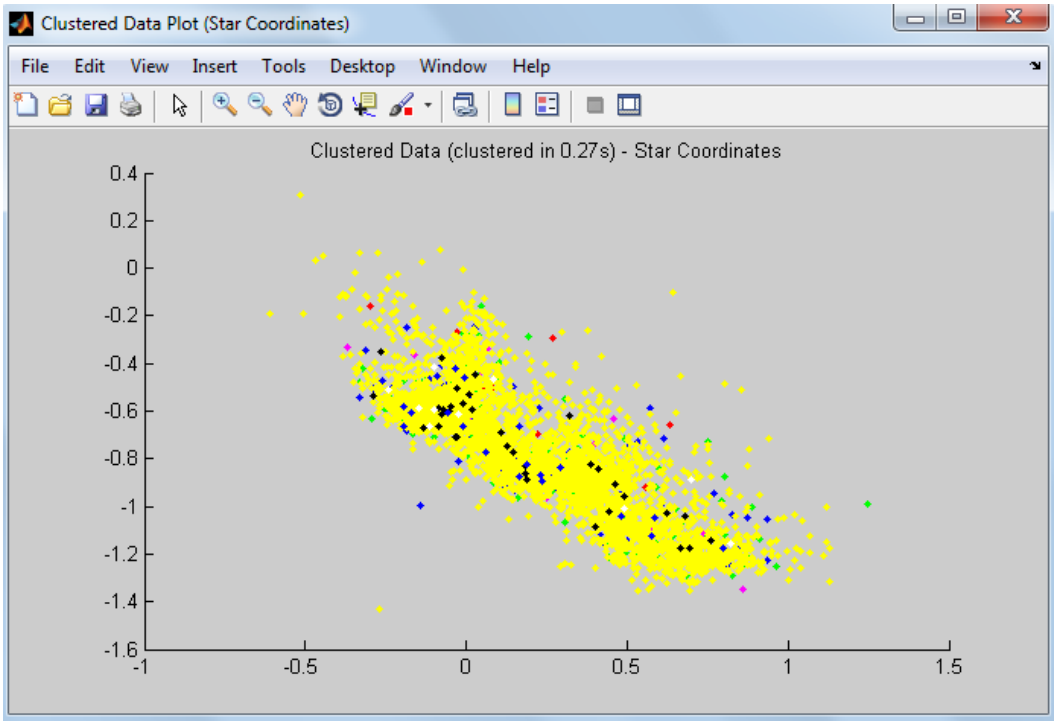


Fig.4.44 Unnormalized Clustered data using Epsilon Affinity Graph in the form of Star Coordinates

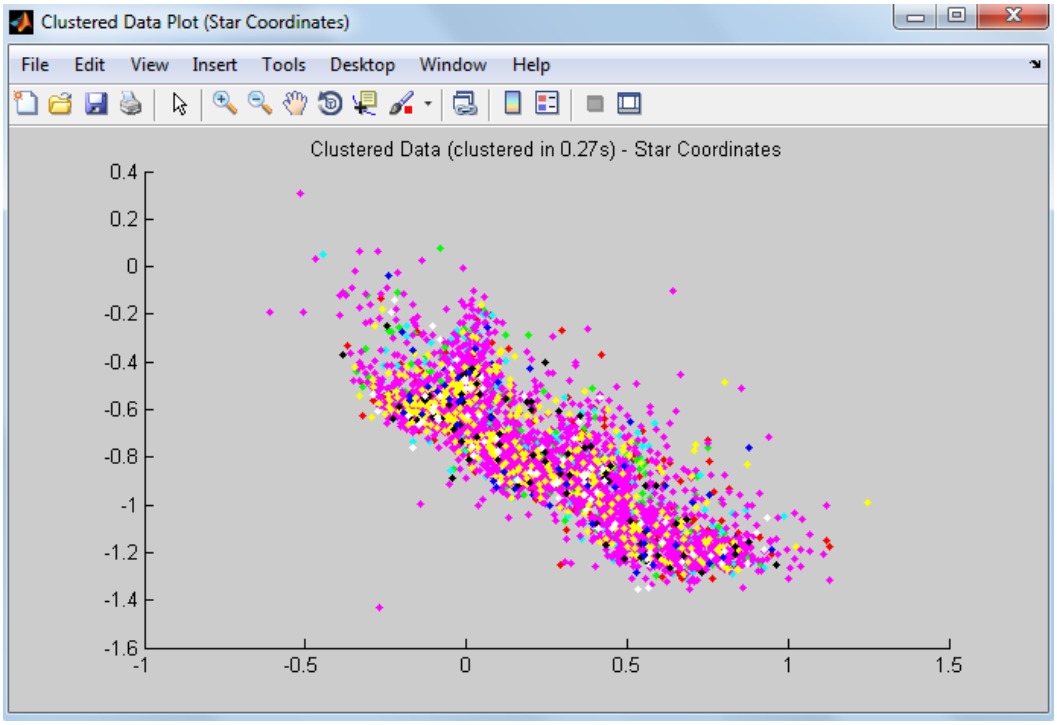


Fig.4.45 SCUED using Epsilon Affinity Graph in the form of Star Co-ordinates

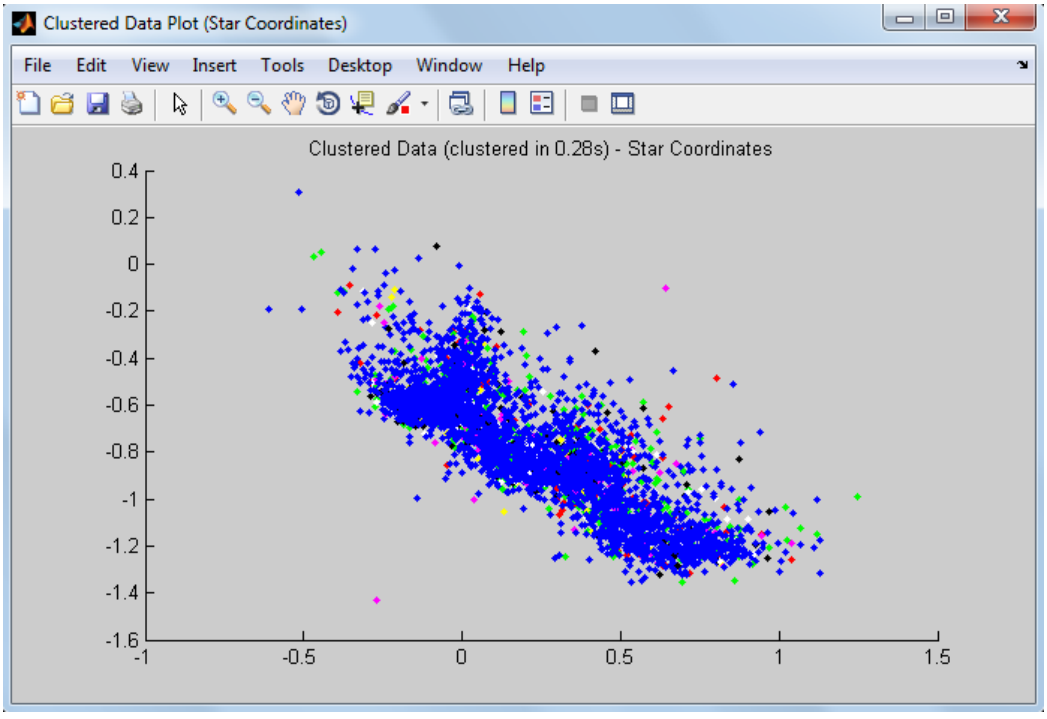


Fig.4.46 Normalized JW clustering using Epsilon Affinity Graph in the form of Star Co-ordinates

Silhouette Values

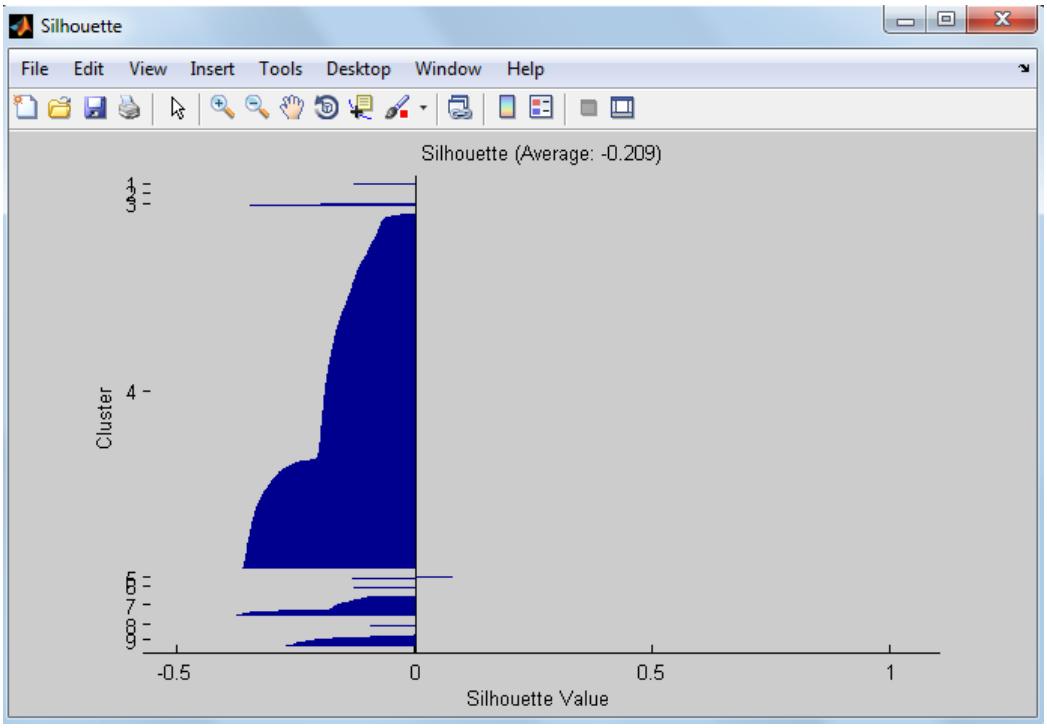


Fig.4.47 Silhouette Value for the Unnormalized Clustering using Epsilon Affinity Graph

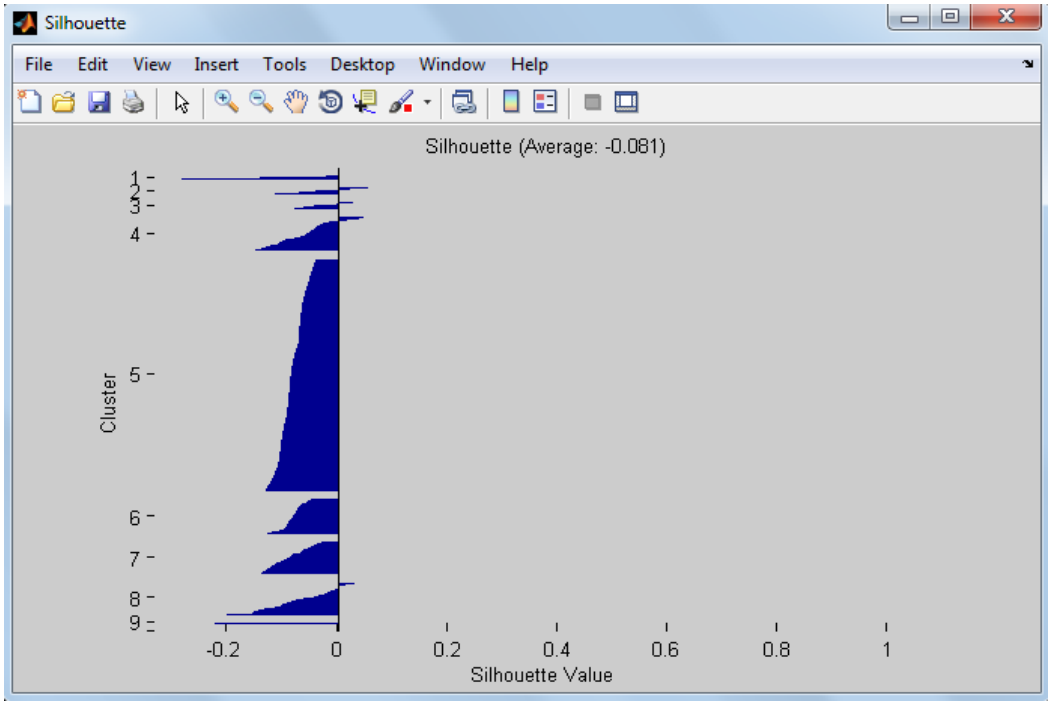


Fig.4.48 Silhouette Value for the SCUED using Epsilon Affinity Graph

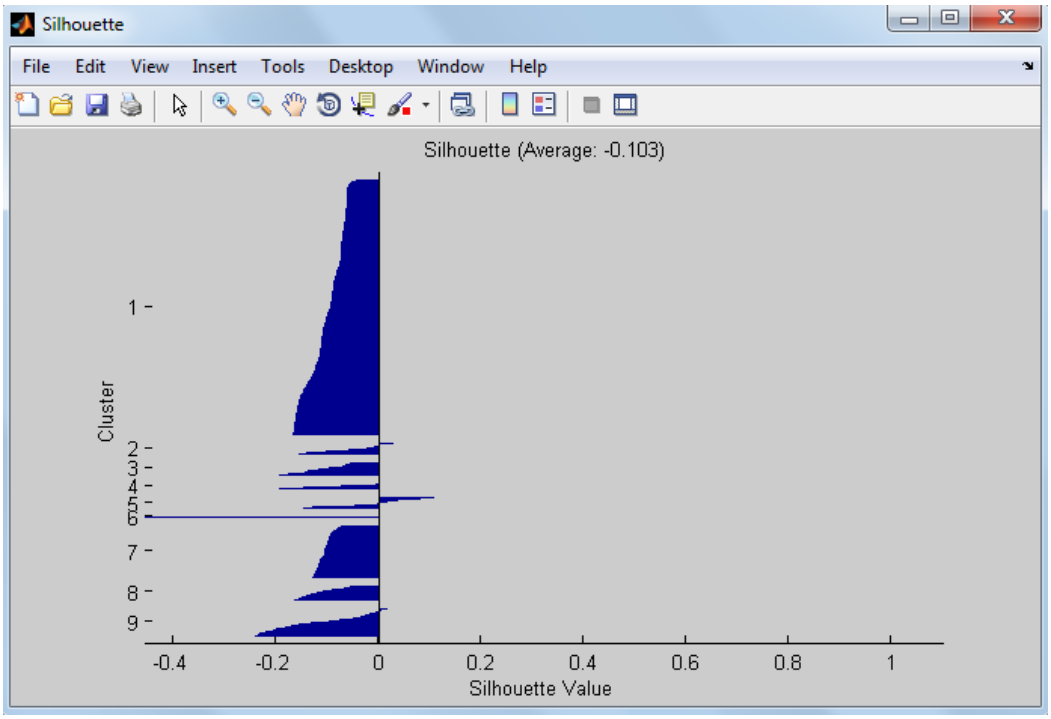


Fig.4.49 Silhouette Value for the Normalized JW clustering using Epsilon Affinity Graph

The above shown results again prove that the silhouette value for the SCUED is greater than the silhouette values of the unnormalized and normalized JW methods. Hence, the clustering done using SCUED is of more good quality than the other two methods.

The summary of the above results is shown in the following tables:

Type of Clustering → Type of Similarity Graph ↓	Unnormalized	SCUED	Normalized JW
	Time(sec)	Time(sec)	Time(sec)
1. Full	2.10	0.31	0.41
2. Normal KNN	0.67	0.33	0.34
3. Mutual KNN	0.73	0.40	0.70
4. Epsilon	0.27	0.27	0.28

Table.4.1. Comparison of Time among all types of clustering using different types of Similarity Graphs

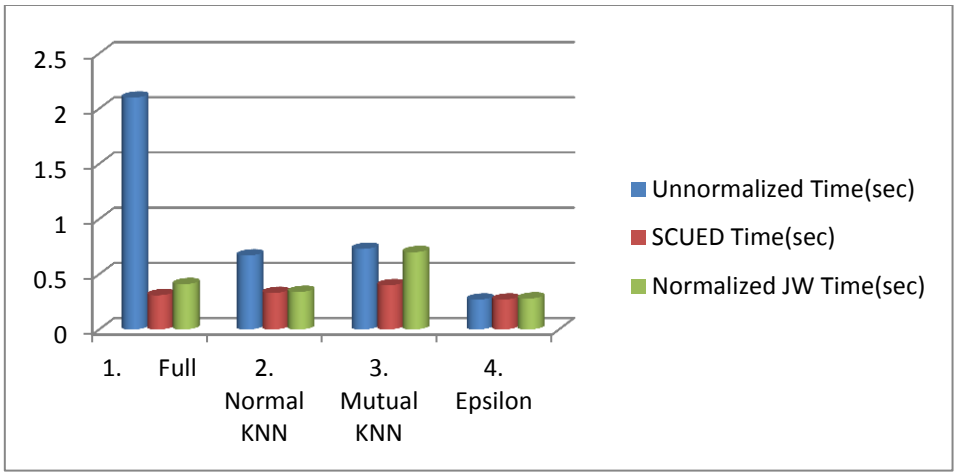


Fig.4.50 Comparison of time among all types of clustering using different types of Similarity Graphs

Type of Clustering → Type of Similarity Graph ↓	Unnormalized Silhouette Value	SCUED Silhouette Value	Normalized JW Silhouette Value
1. Full	-0.595	-0.081	-0.103
2. Normal KNN	0.485	0.539	0.435
3. Mutual KNN	-0.003	0.387	0.387
4. Epsilon	-0.209	-0.081	-0.103

Table 4.2: Comparison of Silhouette Values among all types of clustering using different types of Similarity Graphs

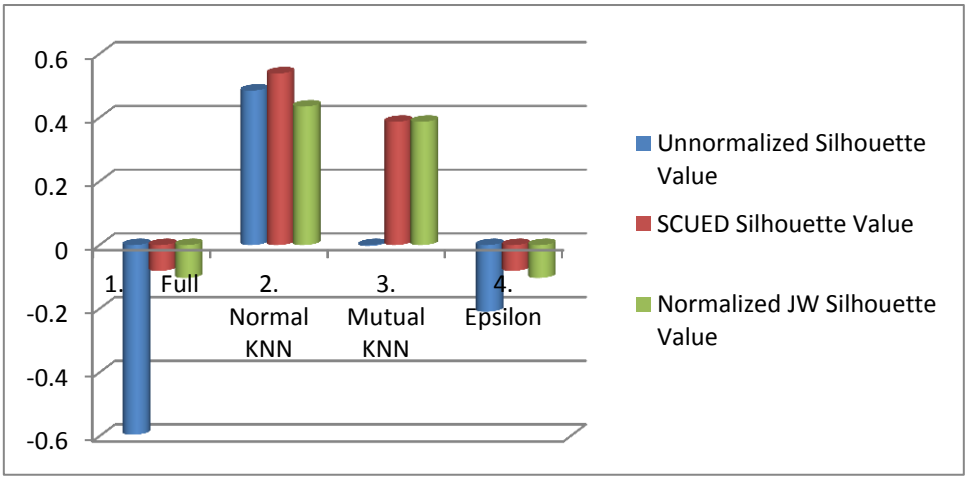


Fig.4.51 Comparison of silhouette values among all types of clustering using different types of Similarity Graphs

CHAPTER 5
CONCLUSIONS AND FUTURE SCOPE

CONCLUSION

In this thesis report, there was a brief description about the field data mining and then some introduction about the topic clustering and its types. After that, the spectral clustering was introduced and the existing technique used for the construction of the affinity matrix. In this work, four ways have been used to construct the affinity matrix. Out of them one is the traditional method in which instead of dividing the dataset into small parts the whole dataset is taken into consideration in a single attempt while in the other three methods the dataset is divided into small parts by calculating the Euclidean distance between the data points. The results prove that by adopting this technique for the construction of the affinity matrix the time for performing the clustering is reduced and the quality of the clusters is improved. The higher silhouette value in the SCUED in all the cases is a proof that the data points clustered using SCUED are more similar to each other as compared to the other methods.

FUTURE SCOPE

SCUED has reduced the time for performing the clustering and the quality of the clusters is also improved as compared to some other methods. But, in case of epsilon affinity graph, the quality of all kinds of clustering is not much good. In future, some new method needs to be introduced that can enhance the quality of clusters in case of epsilon affinity graphs.

CHAPTER 6
REFERENCES

References

1. Cuimei Guo, Sheng Zheng, Yaocheng Xei and Wei Hao, "A Survey on Spectral Clustering", IEEE, 2010

2. Gamila Obadi, Pavla, Jan, "Using Spectral Clustering for Finding Students'
Patterns of Behavior in Social Networks", Dateso 2010, pp. 118{130, ISBN 978-80-7378-116-3}

3. Hao Huang Yunjun Gao, Kevin Chiew Lei Chen Qinming He**,** "Towards Effective and Efficient Mining of Arbitrary Shaped Clusters", ICDE Conference, IEEE 2014

4. Hongjie Jia, Shifei Ding, Xinzheng Zu, "The latest research progress on spectral clustering", Neural Comput & Applic (2014) 24:1477-1486

5. Hsin-Chien Huang, Yung-Yu Chuang, Chu-Song Chen, "Affinity Aggregation for Spectral Clustering", IEEE, 2012

6. Jiawei Han, Micheline Kamber and Jian Pei (2012), *Data Mining- Concepts and Techniques*, Morgan Kaufmann Publishers, USA

7. Luxburg U, "A tutorial on spectral clustering", Stat Comput 17(4):395-416, 2007

8. Peter Kontschieder, Michael Donoser and Horst Bischof, "Improving Affinity Matrices by Modified Mutual kNN-Graphs"

9. Sumuya Borjigin and Chonghui Guo , "Non-Unique cluster number determination method based on stability in spectral clustering". Knowl Info System , 36:439-458, 2013

10. Xianchao Zhang, Quanzeng You , "An improved spectral clustering algorithm based on random walk". 5(3):268-278, 2010

11. Xiatin Zhu, Chen Change Loy, Shaogang Gong, "Constructing Robust Affinity Graphs for Spectral Clustering", IEEE Conference on Computer Vision and Pattern Recognition, 2014

12. Xu-Degang, Zhao Panlei, Gui Weihua, Yang Chunhua, Xie Yongfang, " Research on spectral clustering algorithm based on building different affinity matrix", IEEE, 2009

WEBSITES

13. http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf

14.http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.html

15. http://en.wikipedia.org/wiki/Cluster_analysis

16. http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

17. http://nlp.stanford.edu/IR-book/html/htmledition/model-based-clustering-1.html

18. http://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/

19. http://www.ustudy.in/node/6653

20. http://www.math.utah.edu/~wright/misc/matlab/matlabintro.html

21.https://www.mccormick.northwestern.edu/documents/students/undergraduate/introduction-to-matlab.pdf

CHAPTER 7
APPENDIX

1. KNN K Nearest Neighbours
2. SCUED Spectral Clustering Using Euclidean Distance