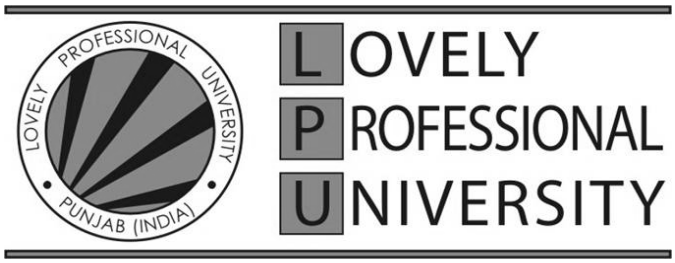

Multi - Assignment Clustering of Boolean data

A Dissertation proposal submitted

By

Suraj Nahar

Department of Computer Science and Engineering

In Partial fulfillment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Assistant Professor Ms. Karanvir Kaur

(May 2015)

CERTIFICATE

This is to certify that **Suraj Nahar** is proposing M. Tech. dissertation titled **Multi - Assignment Clustering of Boolean data** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation 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 and Engineering.

Date: 03rd June, 2015

Signature of Advisor

Name : Karanvir Kaur

UID : 14856

DECLARATION

I hereby declare that the dissertation proposal entitled Clustering of Boolean data and problem of role 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 03/06/2015

Suraj Nahar

41200367

LOVELY PROFESSIONAL UNIVERSITY
Transforming Education, Transforming India

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	: Suraj Nahar	Registration No	: 41200367
Batch	: 2012	Roll No	: RK2213A11
Session	: 2014-2015	Parent Section	: K2213

Details of Supervisor:

Name	: Karanvir Kaur	Designation	: Assistant Professor
UID	: 14856	Qualification	: M.E
		Research Exp.	: 4 years

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

Proposed Topics:-

1. Multi-Assignment Clustering of Boolean Data.
2. Parallel Spectral Clustering in distributed systems.
3. Data Spectroscopic Clustering.

Signature of supervisor 14856

PAC Remarks: Topic 1 is approved. Paper Expected. 7/10/17

APPROVAL OF PAC CHAIRMAN Signature: 24/11/17 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.

ABSTRACT

Data mining has been evolved as a very useful technology for many of the challenging problems in real world like predication of the sales, decision making and planning of the marketing strategies for big business houses. One of the mostly widely used techniques in data mining is cluster analysis. Cluster analysis refers to grouping of similar type of data in one set and other data in other set. Mining of Boolean data has been a great area of research these days. We will work on the mining of the Boolean data. We try to increase the efficiency of the existing algorithms and we also try to reduce to make the algorithm more prone to noise levels.

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to **Ms. Karanvir Kaur Assistant Professor, Department of computer science and engineering Lovely professional University** for her patient guidance and support throughout the Dissertation. I am truly very fortunate to have the opportunity to work with her found her guidance to be extremely valuable. I would like to thank entire faculty and staff of computer science and engineering department and then my friends who devoted their valuable time and help me in all possible ways towards successful completion of this work. I thank all those who have contributed directly or indirectly to this work.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Data mining is playing a vital role in many of the field such as market-basket analysis, classification, etc. In data mining, frequent item sets have significant role which is used to find out the correlations between the fields of database. Data mining is known as Knowledge Discovery in Database (KDD).Association rule is based on discovering frequent item sets. Association rules are frequently used by retail stores to manage in inventory control, predicting, marketing, advertising, faults in telecommunication network.

Data Mining is defined as the process in which a set of data is feed as the input and that input is processed and provide certain valuable knowledge which is useful to us in certain way. It can also define as data mining is the process of extracting the valuable information from input data. In the field of Information technology, it has enormous amount of data available that require being bitter information that is most appropriate for our use. This information further can be used for various applications like analysis of markets, Maintaining good customer relations and holding customers, controlling the production variables, detection of the various types of fraud, exploration that are valuable in scientific projects etc (Ashish Jain, 2009).

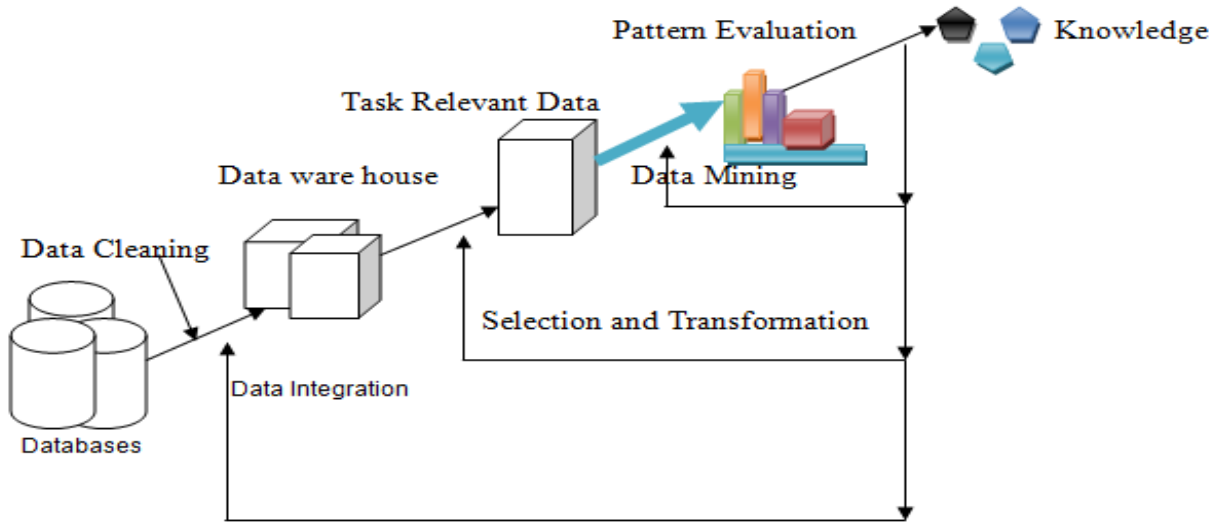

Figure 1.1: Data mining

It can be used with any type of data store. Various types of techniques are available for various types of data. Data mining has been applied to diverse domains like database containing multimedia data , data marts and data ware houses.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

1.1 Data Mining System: The data mining systems are classified into various classifications according to their criteria. These classifications are as follow (Manne, 2011):

1. **Classification according to the type of data source mined:**
   In this type of classification data mining system is categorized according to the data to be handled such as time series data, multimedia data, spatial data etc.

2. **Classification according to the data model:**
   In this type of classification data mining system is categorized according to the data to be handled such as relational database, object oriented databases, relational database etc.

3. **Classification according to the king of knowledge discovered:**
   In this type of classification data mining system is categorized according to the functionalities of the system and knowledge discovered like association, characterization, clustering, classification etc.

4. **Classification according to the mining techniques to be used:**

   In this type of classification data mining system is categorized according to the data analysis approach to be used in the system. These approaches are like machine learning approach, neural network, genetic algorithm and visualization and data ware house oriented. It also take consider of user coverage and degree of user interaction.

   Datamining has evolved over the time. There are various steps involved in it like extraction of data , transforming of data cleaning of data and also include storage and retrieval of data from the repository .one of the most prominent repository architecture

that has been evolved over the time is the architecture that greatly relies on the data warehouse. The storage facility can hold variety of data from different sources and all this managed under single schemata a standalone site which helps the governing body to take good business decision. This technology comprises of the various phases few of them being the makeover of the data or data purification, data amalgamation and the olap analysis technique which provided wide range of functions like sum up, association of data, gathering of data as well as the capacity to see the information from the various prospects . The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task.

As we know that the amount of data we are dealing is very large. We need more competent or powerful tools to analyse this data. This situation has lead us to a point that we have become more wealthy in terms of data but in actual we have become low on the amount of knowledge it give to us. In today's modern era all the activities across the web are stored and monitored in large data centres this huge amount of data has gone far beyond the reach of the human tendency to manage without the help of the competent tools. As a result of this the collected data has just become a grave of the data which is visited very less

Database, data warehouse, information repository and their server which are dependable for extraction the data that is most relevant which depends on the type of request entered by the user. Knowledge base is also its parts which are used to conduct the search, or calculate how the patterns that are evolved are of user's interest or not. After that data mining engine is there lies a set or group of the functional modules which contains the responsibilities [2]. Pattern evaluation is that module which keeps in touch with the data mining modules so as to focus the look for interesting patterns and graphical user interface which acts as a communication link between users and the data mining system facilitates the interaction of the user with the system.

1.2 Advantages of Data Mining:

1. To Investment firms and financial companies

This technology facilitates the banks to get an idea about the customer loan information and spending habits. By using this legacy data the banks can have an idea that which loans are good ad which loans are bad. This technology also helps in detecting the credit cards frauds by analyzing the buying habits of the customers if there is any change in it may be a credit card fraud.

2. In decision making and marketing of the product

Data mining facilitates the marketing firms to make such a plans or model which is based on the previous data of the legacy data and with help of this data they make an assumption that who are the intended customers. Which customers will most likely purchase the product and which may not then according to analysis they make strategy about giving discounts on the item. This helps the companies to sell the product to the intended customers. This technology generates lots of revenue for the big retails companies. They make use of techniques like the market basket analysis to predict the buying habits of the customers. This allows companies to make a set of pack of items which are most likely to be sold together like bread and butter.

3. In manufacturing of Products

With the help of this technology and good engineering practices the manufacturers can have an idea that whether the equipment is faulty or the or not and they can determine certain optimal control parameters which helps in obtaining good quality of finished product. Let's consider the example of manufacturing of semi- conductor devices. The biggest challenge in the manufacturing of the semi- conductor devices is that even though the manufacturing environment is same but still the quality of device made is different. This technology has been used to get limit of the parameters to be controlled so that the finished product will same quality as previous one. One the control parameter are finalized we will get final product of required quality.

4 State and law enforcement agencies

As the amount of data that has been stored and retrieved has increased many folds in recent times. This has present a challenge to the law enforcement agencies to tackle the cyber thefts etc. but data mining technologies enables the law enforcement agencies to get details of the various financial dealings that are occurring online. They can use this data to detect the outliers and they can detect cases of the fraud that are happening online. As there are cases of illegal purchase of goods and money laundering on net. Data mining can detect it and helps the law enforcement agencies to keep a check on such activities.

1.3 Disadvantages of Data Mining:

1. Breach of security

Breach of security is the biggest concern of the data mining technology. Data mining include analysis of large amount of data which may include private and confidential information like birth a record of the person, provident fund details or driver license details etc. but how this data has been handled remains a matter of concern. In the past there has been number of cases when the intruders were able to break the security of the system and stole the data about the customer details and other financial data. There had been cases of the identity theft also in which a person pretend to be a person which he is not in actual.

2. Breach of privacy

With the ever growing amount of data collected on the networking sites and ecommerce websites. People are very much concerned about their privacy. Certain people lives in the fear that they are been traced online and information about them is stored online. If there is any leak of this information it may cause problem to the users.

3. Wrong use of the information

The information generated of the information collected from the data mining is intended to be used for the noble purposes like decision taking or analysis of data. This data may contain information that may be personal in nature and confidential too. But there has been incidents

where the big business houses has used this information for taking advantage of the situation and exploiting the vulnerable person

1.4 Issues in Data Mining: The various types of issues in data mining are

The Security and Moral Issues

The most discussed concern of the data mining is the security. Data is collected so that it can be analyzed and certain decisions can be taken based on that data. Data is related to customers and include the confidential information so there is always threat that this data can be used for the illegal purposes or wrong doings.

Issues related to performance

Effectively and scale up capability of the algorithms used in the data mining: As there is lot of data involved in data mining. We must find such algorithms that are more effective when it comes to computation power required and our algorithms must be smart enough so that they can be scaled up as per growing need.

.

Incremental, parallel and distributed mining algorithms: There is large volume of data involve in the process of the data mining. SO we need to develop the algorithms that are parallel in nature and must have distributed qualities as well. In parallel algorithms the task is divided into chunks of small sizes and then these chunks are processed in isolation and after that we amalgamate the result obtained from various chunks. The use of the algorithm that are incremental in nature saves our time as we need not to mine data from the grass root level again.

1.5 Applications of Data Mining:

Data mining has great approach and various fields of applications. These fields are:

1. **Business Applications:** In business application, database mining is one of the popular and important applications. During mining of historical data pattern and customer profile are built on the basis of the results. It is also use in retail database to find out customer

databases from the records. Using mining techniques models are used to build models for the stocks. It is also used for loan and credit applications.

2. **Science Application:** It is also used in various science activities. It is used in biology, molecular science, astrology and sky objects.
3. **Banking Sector:** It is also used in banking sector to retrieve information of old customer, their identity and many other things [13].

1.6 Clustering in Data Mining:

Clustering is one of the most widely used techniques for analyzing 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 research of market situation, recognition of patterns, analysis of the data, processing of the images and detection of the outlier applications such as detection of credit card frauds. There exist many different types of clustering methods including hierarchical, partitioning, density based, model-based and grid-based methods

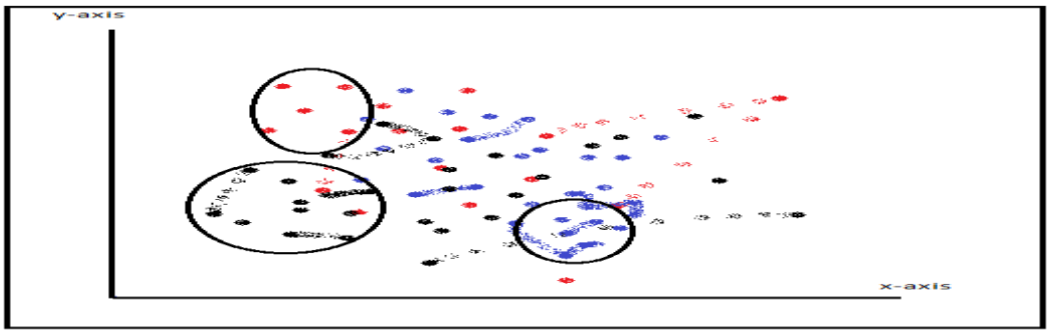

Figure 1.2: Clustering

1.7 Types of Clustering Algorithms: There are four types of clustering algorithms

1. Partitioning Clustering

2. Hierarchical Clustering
3. Grid based Clustering
4. Density based Clustering

1. **Partitioning Clustering:** In this type of clustering the data the entire data is divided into certain number of clusters say K and every cluster has certain number of object contained in it say nit is required the every cluster should have certain number of elements in it .No cluster should be left blank of empty. In this concept a divider of a data set accepting as k clusters are formed from n number of objects, so to minimize a benchmark. The goal is, given a k, find a partition of k clusters that enhances the preferred partitioning benchmark [20]. Here k is a input parameters. E.g. K-mean and K-centriod.

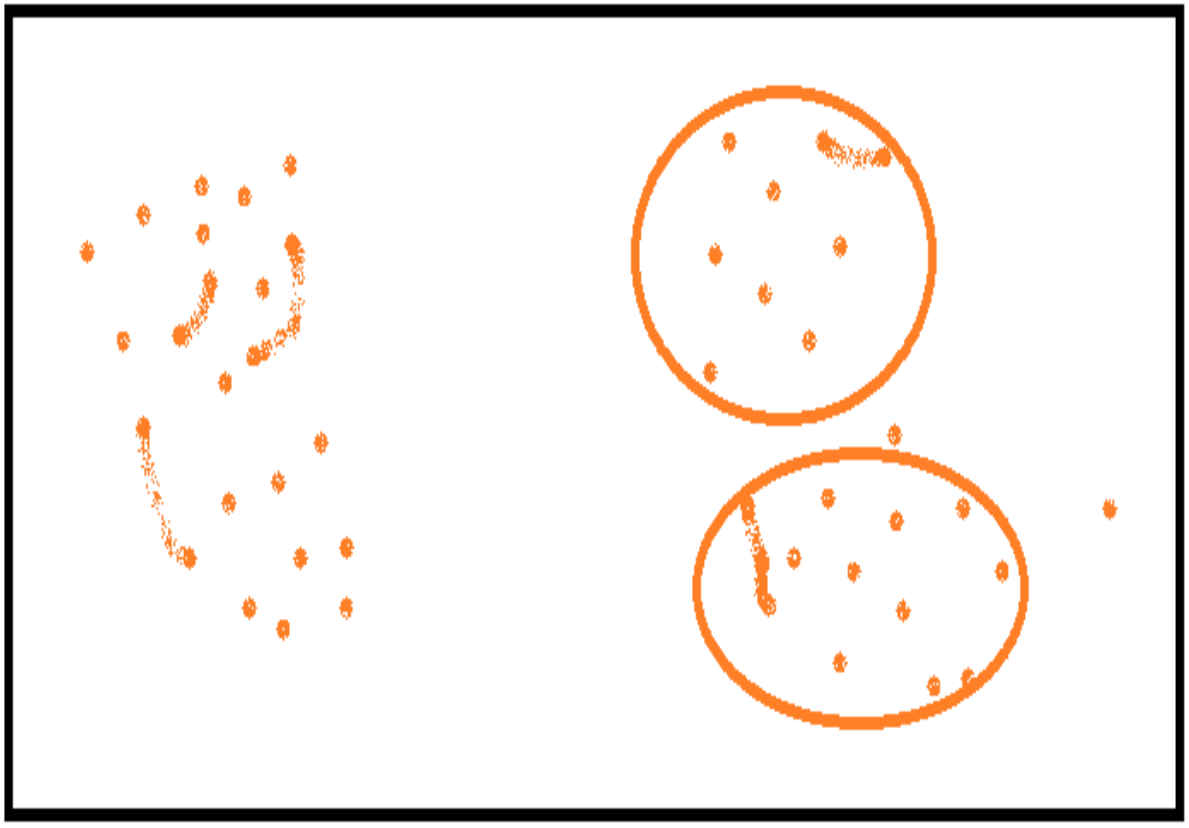

Figure 1.3: Partitioning Clustering

2. Hierarchal Clustering:

In this type of clustering the given dataset objects are produced in hierarchy. The clusters are shown in tree structure known as dendrogram [21]. No cluster input is required. We can view partions at various levels of granularity using different types of K. E.g. Flat Clustering

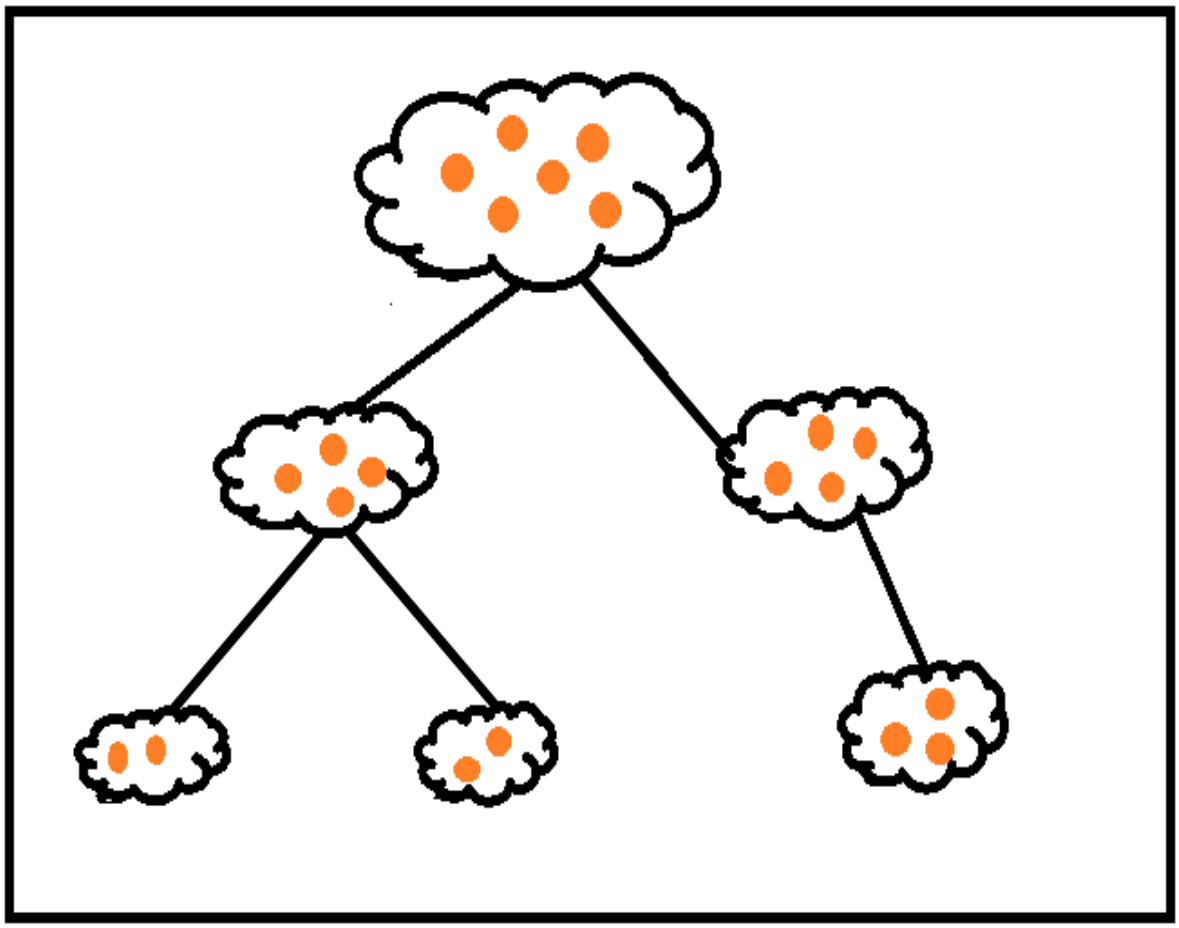

Figure 1.4: Hierarchical Clustering

3. Density based clustering:. This technique is based on the density. Density is used to get a cluster in this technique. In this technique we grow the cluster to an extent till the density of the nearest cluster beats some threshold. The Diameter of the provided cluster need to take over the prescribed number of points..It helps to discover arbitrary shape clusters [22]. This algorithm manages the noise in data and needs the density parameter and one time scan.

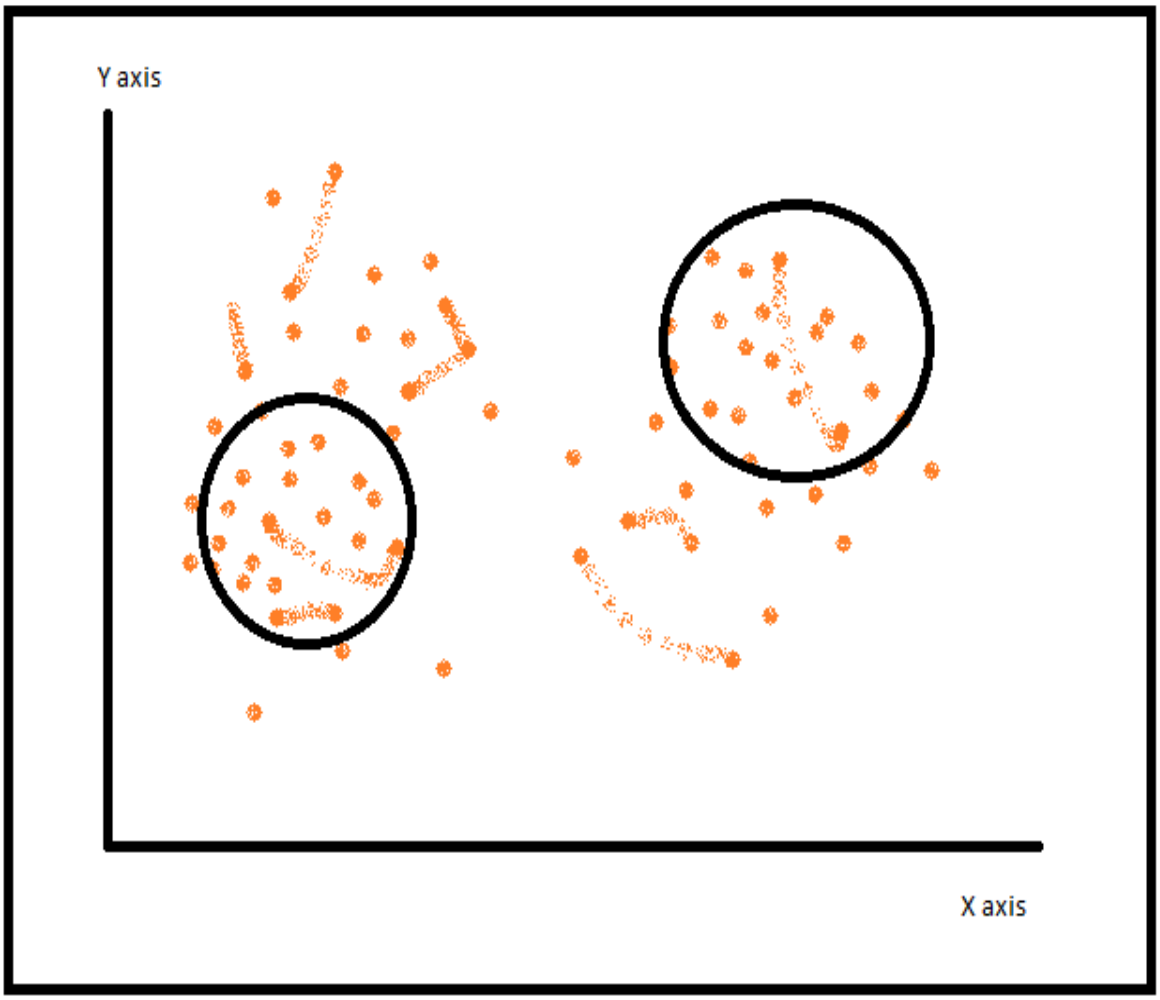

Figure 1.5: Density based clustering

4. Grid based Clustering:

In this technique the objects are so arranged that they form a network. The distance between the item adjusted to limited number of booths which gives an assembly of network. It assigns to object grids cells and compute density of each cell. After that eliminate whose density is below threshold value [23].by using the dense clusters we form the cluster. We don't need to compute distance in this case so it is fast. We can easily determine the closed neighbor cluster. Here in this figures are narrow to the union. The complexity is directly proportional

to the alignment of the cells . Grid clustering algorithms determine the space into a limited number of grids and accomplish all tasks on this determined space. This algorithm has fast processing time and self-controlled size of dataset and only relies on the count of parts in each direction in the assigned area.

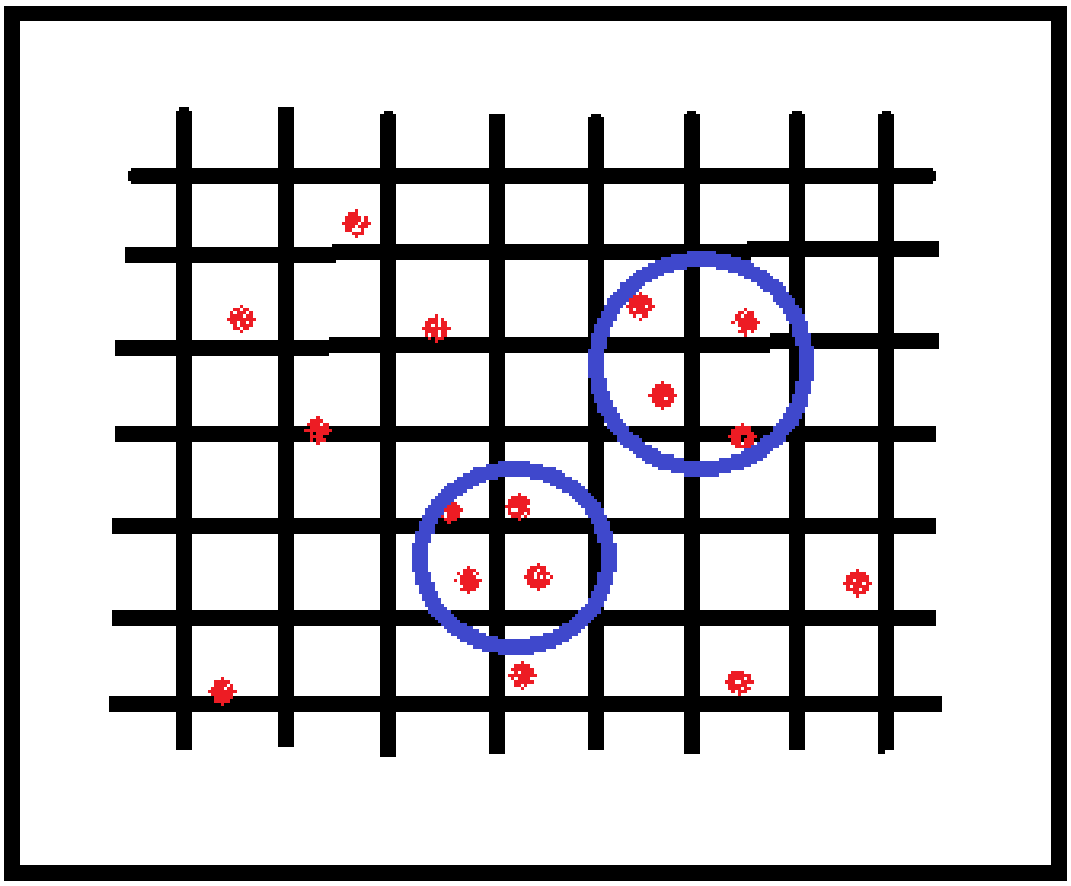

Figure 1.6: Grid based density

1.7 Multi Assignment Clustering: Multi-Assignment Clustering (MAC), for clustering Boolean vectorial data that can simultaneously belong to multiple clusters. In generative model, each constituent of each data vector is either drawn from a signal distribution that is given by the clusters in the data item belongs to [24]. It also forms an independent global noise distribution. It presents an expectation-maximization (EM-) algorithm where the source prototypes of the

clusters and the cluster memberships of each data item are simultaneously estimated as well as the mixture

Multi-Assignment Clustering for Boolean Data weight of the global noise source. It recovers the cluster prototypes with significantly higher accuracy than alternative methods, especially for datasets with high noise level. Furthermore, the assignment of data items to clusters has superior stability under resampling. The clustering of data in disjoint cluster is very simple but assignment of the cluster is very difficult due to its structured. Multiple clusters can be generating data item simultaneously using dependent link function [25].

MACM produce results precise and accurate as compare to the state-of-art clustering techniques. It uses a user role matrix and permission role assignment matrix from a Boolean relationship matrix and defines an access-control matrix. The generalization ability of our model in this domain outperforms other multi-assignment techniques.

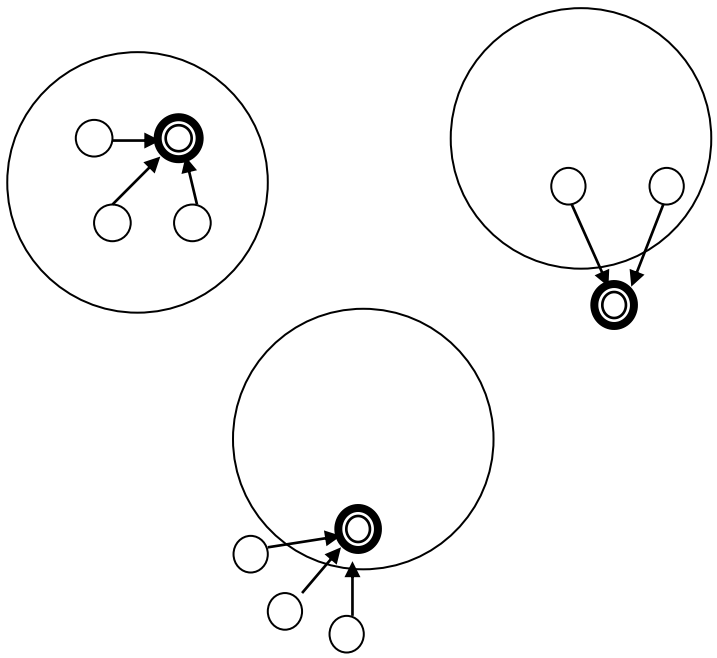

Figure 1.7: Multi-assignment clustering

1.8 Clustering has many requirements some of them are as listed below:

1. **Scalability**: The algorithm must be scalable as we need large extend of scalability.
2. **Should work on large types attributes**: The Algorithm must be flexible enough to work with large no of data types.
3. **Working with clusters of various sizes**: The Algorithm must not only find clusters of the spherical size but also of the other shape and sizes.
4. **Works with multi-dimensional data**: The clustering algorithm must be flexible enough to work with large multi-dimensional data.
5. **Handing the noisy data**: Clustering algorithm should be such that it should handle the noisy data to great extent.
6. **Clear results**: The results produced by the clustering algorithms should be easily understandable and accurate.

CHAPTER 2

REVIEW OF LITERATURE

Ming-Yi Shih et al. (2010) discussed about the different types of algorithm that are used for clustering data across various domain. But the limitations of these algorithms are that they work on the data which is numeric in nature or which are categorical data. But a majority of algorithm fails on mixed type of data so the author proposes a fresh two step clustering algorithm that can be used on, mixed kind of data. In this approach the attributes that are categorical in nature are evaluated to get the similarity or we can say the relationship between them based on the view of co-occurrence after that all attributes that are categorical in nature are being changed to numeric data. After that the state of the art clustering algorithms are applied to it without any difficulty. But there are certain disadvantages suggested by the author in the existing algorithms used in the clustering. But the algorithms suggested buy the author uses the two step method which joins the hierarchical and portioning clustering by addition of attributed to the cluster objects. This algorithm explains how one item is related to other item and eliminated the limitation of using only single cluster algorithms experimentally it has been show that by using the algorithm good result can be achieved in mixed kind of data.

R.Jensi and Dr.G.Wiselin Jiji (2013) large amount of data is being available in digital format which makes it easy to do text document clustering and it also evolved as one of the major research area. There are wide range of methods available which can cluster data in a way the electronic local to a cluster has large similarity among the cluster and very less similarity between the clusters large number of clustering algorithm provide with the facility of easily organizing and navigating. A widely accepted solution can be achieved by the using the algorithms that are high in quality and high in speed. This technique searching the global solution space various approaches to text document clustering has been discussed in it. This paper covers a brief detail about soft computing and data mining. Most of the research has been done by use if semantic to improve the quality of the clustering.

Dharmendra K Roy and Lokesh K Sharma (2010) the role of clustering in data mining and the target of organizing the data objects in certain classes that are meaningful inn nature in such a manner that the objects in one class have most of the similarities and each class must have

different properties from the other class. The author has proposed a genetic k means algorithm that can perform with mixed king of data. They had proposed certain modification to the center of the cluster so that limitation of numeric data can be overcome and centroid can be assigned to the cluster. They have analyzed the performance of this algorithm on various benchmark datasets.

Andreas P. Streich et al. (2009) the conventionally available clustering methods which assumes that every data object will be limited to only single cluster but in general this assumption is not true. To remove this limitation the author has proposed a new method for clustering or grouping of Boolean dat. Where there is a probability that one object can belong to multiple clusters. By the use of annealing scheme the algorithm breaks the observed data as participation of the each cluster and then guess there parameters. Experimental results on Boolean data have shown that this method gets higher accuracy in estimation of the source parameter and high stability of the cluster when compared against the available algorithms. This method has better performance as compared to other algorithm this method has been widely used in role mining. This method works good under high noise while other fails to do so.

Mario Frank et.al. (2012) a model based on the probability which is used in clustering of Boolean data in which an object can exists in two clusters at one point of time. This method provides good result even in situation where the sample quality is low. This model has been tested with different noise process and it performs better as compared to single assignment clustering. This method is used mostly in role mining. This method performs better as compared to traditional methods.

Sumuya Borjigin and Chonghui Guo (2013) a non-unique cluster determination method which depends on stability. They had made use Gaussian kernel parameter to change the matrix used for distance into matrix that represents the similarity and then after it uses multiple normalized algorithms which cluster the data point. The basic focus is to check whether the number of clusters formed is stable and which are reasonable which can be used to increase the cluster quality.

Ms Chinki Chandhole et al. (2012) focuses on getting a way for segmentation of image by using the classical K-Means algorithm. In segmentation of image we make frequent use of the

clustering algorithms because of their simplicity and user friendliness. In this paper we will be doing segmentation by using the K-Means algorithm. By applying the K-Means algorithm repeatedly we partion the images in K clusters. But the limitation of this algorithm is that we get exact segmentation only when images are defined by homogenous regions. In the first phase we first cluster the image pixels bases on their color and features after that we merge the clustered blocks to specific number of regions. Thus with use of this technique we can revive the image.

Ming Chan Hang et al. (2005) Focuses about the widely used clustering algorithm i.e K-means. But the main limitation of K-means algorithm is that we need to calculate the centroids which are expensive in nature. In this paper the author has proposed a new approach by which expensive calculations can be reduced. To do this we need to partition the dataset into blocks each block is known as a unit blocks (UB) which contains a minimum of one pattern. We by means of simple calculations find centroid of the unit block. Then these CUBs are used to find the final centroid. By using this technique we can greatly reduce the time in calculation the centroid. This algorithm has better results as compared to other algorithms with much improved performance.

Walea K Gad, Mohamd S.Kamal (2010) Focuses incremental clustering of documents is very essential for working on large datasets. There are various types of incremental algorithms present but the limitations of these algorithms are that they do not take into account linguistic and semantic properties of text. The algorithms proposed by the author works well with the World Wide Web. The algorithm proposed by the author binds the text semantics to clustering algorithm. This algorithm has better performance as compared to normal clustering methods.

Simon Jons, Ling Shao (2011) Focuses contextual information is related to an action. Consider an example of an object or scene can very much effect the capability if recognitions of human activities. But with the use of context to enhance unsupervised human actions clustering was never done before and it was not possible by using the traditional clustering methods. The author has proposed a new Dual K- Means algorithm. This is able to do co-occurring tasks simultaneously.

CHAPTER 3

PRESENT WORK

3.1 SCOPE OF STUDY

There are large numbers of clustering algorithm available that help to group similar data in a cluster. But these clustering algorithms always assume that an object will belong only to a single cluster but it may be not true. There may be a situation in which one object can belongs to

multiple clusters but the traditional algorithms are unable to do so. Our focus is to study multi – assignment clustering which is used for clustering of the Boolean data in which there is a probability that an object may belong to one or more clusters. We will use this technique to one of the, most challenging problems in data mining that is role mining.

We will also take a look at the different application specific process which is responsible for the deviation in the data and we will then take a look at the relationship between these objects. We try to show that with the use of mac we can get more accurate result as compared to the present very stable algorithms. we will focus on providing a solid solution to problem of role mining we will try to try to eliminate the problem that are faced in the technique of the decomposition of Boolean matrix and equivalent problems and the various approaches that are applied to solve it.

We will also focus on the problem like the noise level and the average hamming distance. Average hamming distance refers to the measure of how accurately the centroids of the clusters are being estimated. Our main focus will be to reduce the noise level that may be present due to irregularity of the data.

3.2 PRESENT WORK

Analysis of the clusters originated in data mining has been used in wide range of applications which include research in the field of the marketing a, in recognition of the pattern, analysis of the data and processing of the images. When it comes to business cluster analysis has been used to know the interest of the users' customers which is primarily based on the buying habits and features of groups of the customers. In the field of life science, cluster analysis can be used to describe out the plants and to drive out the taxonomies of animals and categorizing of genes which are similar in function and get insight knowledge about population structure. In the field of the geology a person who specializes in this technique can predict the area with similar land and area which have houses that are similar in nature. Clustering of data can very useful in categorizing text documents on the internet for discovery information. Clustering in generally referred as unsupervised method of classification in which we target to create certain group object or certain clusters in a way that object in one cluster had same features and objects in other cluster have other feature. The biggest challenge in data mining in mining of arbitrary

clusters. But the various methods that have been used to solve this problem depend on way by time computation. To minimize the cost of computation certain algorithms try to reduce the size of the data set. There is algorithm popularly known as CLASP algorithm which is very efficient for getting clusters that are arbitrary in size. This algorithm shrinks the dataset but it maintains the information about the shape after that it makes adjustment to the position of data so that the relationship can be enhanced and clusters can be made more clearly. Then it uses the similarity matrix called pk. In this work we will make an effort to improve the asymmetric clustering algorithm so that we can increase the quality of cluster and can improve the efficiency of the algorithm.

3.3 OBJECTIVES

1. To study and analyze various asymmetric clustering techniques to cluster relevant and irrelevant data.

2. To propose enhancement in the asymmetric algorithm to improve accuracy of the algorithm.

3. The proposed enhancement will be based on the neural network to improve cluster quality.

4. To implement proposed algorithm and existing algorithms and analyze the results in terms of accuracy and cluster quality.

3.4 RESEARCH METHODOLOGY

Multi assignment task is the type of unsupervised learning technique that is used to solve the clustering problems. The process goes through a way that it organizes the data set in the form of clusters fixed according to priori. Asymmetric clustering is used in number of applications. There are number of clustering methods, multi task assignment algorithm adopts that how many clusters k are present in database before which is not true in real time applications. It is an iterative technique and these algorithms are sensitive towards initial centers selection. Multi task assignment has many disadvantages that it works well with simple databases but it does gives desired outputs in mixed and tightly coupled data sets or items and by this accuracy and

efficiency of algorithms is reduced. So, a proper method needed that will balance both the accuracy and efficiency of the multi task assignment. Flow chart of research is defined as:

DATASET and Pre-processing: In the first step of flowchart, the dataset is extracted and then pre-processed to perform clustering on dataset.

Obtain various attributes: After the pre-processing phase, the dataset contains various attributes, in this phase relationship between various attributes are established.A1, A2, A3…. is the number of different clusters of particular attributes.

Feature Reconstruction: in this step, again pre-processing technique is performed on the clusters to remove noise from the attributes in the clusters

Drive relationship and training dataset:- to drive relationship between various attributes of the dataset, technique of neural network will be applied. To apply neural network we need a trained dataset. The trained dataset will establish relationship between various attributes

Threshold Analysis: in this step, threshold analysis done and values above threshold values are stored in one cluster and values below threshold values stored in another cluster called C1 andC2

Prediction: this step predicts the improved efficiency of the work in research.

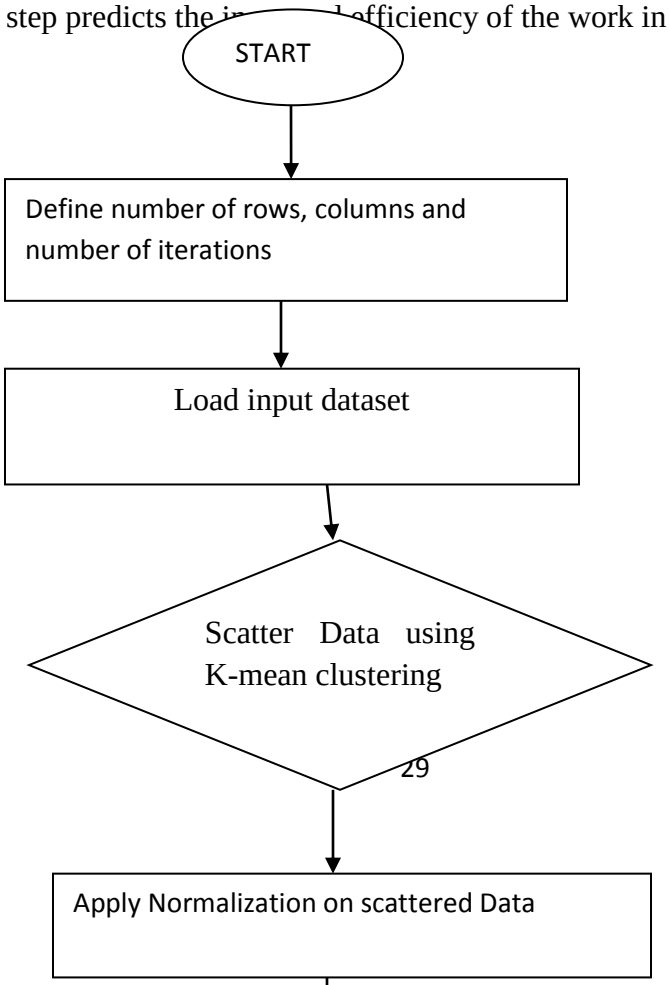

Figure 3.1: Previous work done

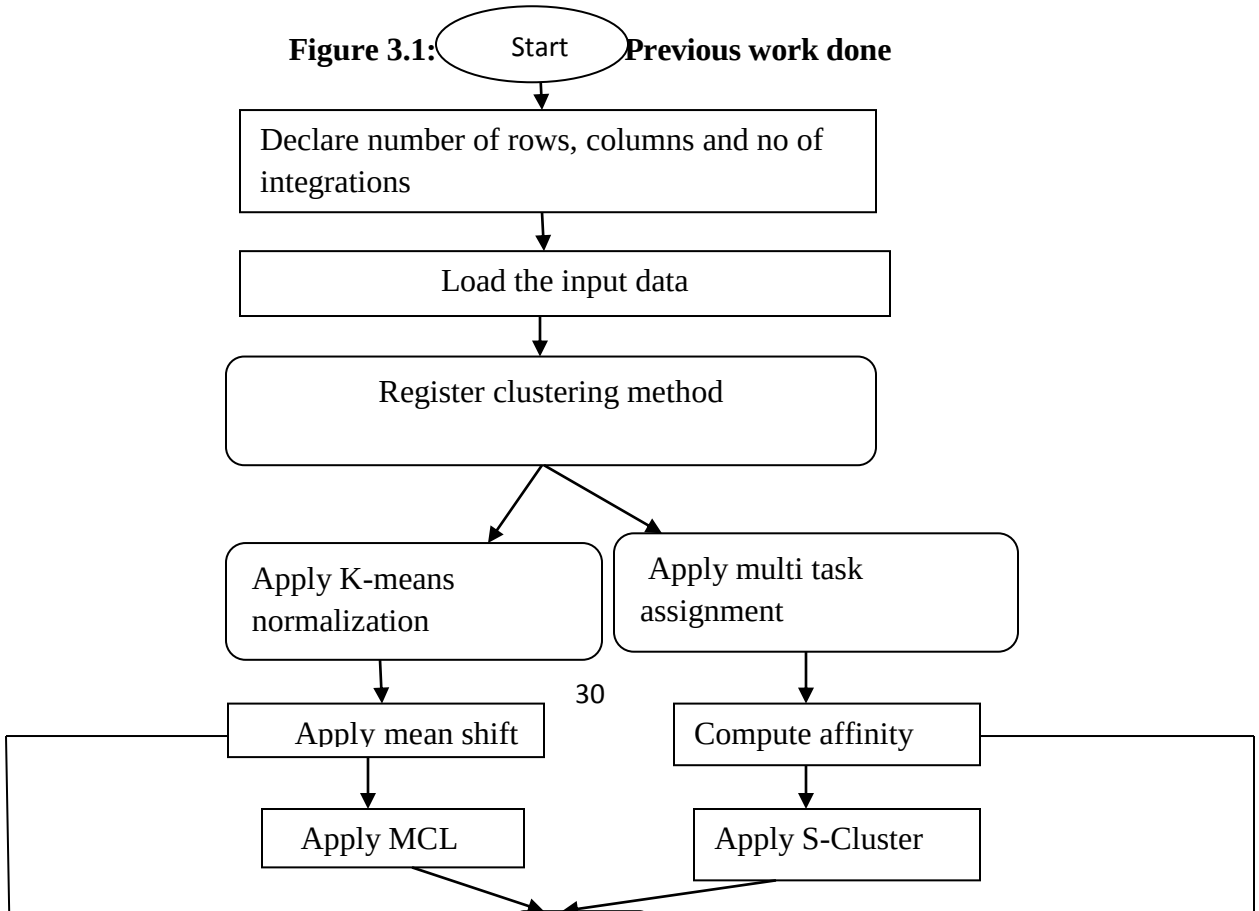

Figure 3.2: Flowchart Of proposed Algorithm

Explanation of Proposed Flowchart

1. **Declare rows, columns, integration and load dataset: -** This is the first set of algorithm in which the number of rows and columns are defined for the dataset. The second condition is defined to define number of iteration to define cluster quality. In the step of the flowchart the dataset will be loaded to perform clustering operation
2. **Register Clustering Method: -** To register clustering method is the second step of the flowchart in which we defined the two clustering method. The first method is K-mean clustering and second method is asymmetric clustering. According to the selected method the operation of clustering will be performed on the dataset.
3. **Apply K-mean Normalization and asymmetric Clustering: -** When the clustering

method is registered, it may be K-mean normalization method which is selected for clustering with the normalization equation. The normalization equation when implemented with k-mean the cluster quality can be improved. The second method is of asymmetric clustering which is implemented to cluster the asymmetric data from the loaded dataset.

4. **Apply mean shift and affinity metrics: -** In this step, two operations are performed. In the first step mean shift algorithm is applied on the loaded dataset. In the mean shift algorithm, the mean value is calculated on the dataset and left shift operation is performed to simplify the operation of clustering. The second method is of affinity metrics, it is equation which is applied to find relationship between various elements of the dataset.
5. **Apply MCL and S-clustering: -** The MCL is the markov clustering algorithm, which is the unsupervised clustering graph based algorithm. This algorithm is fast and reliable and has good cluster quality. The main concept behind this algorithm is mathematical theory behind it, its position in cluster analysis and graph clustering, issues concerning scalability, implementation, and benchmarking, and performance criteria for graph clustering in general. The second method is S-clustering which is applied to cluster the data on the basis of graph methods
6. **Plot and make clustering and normalize: -** In the previous step, two methods are applied which are MCL and S-cluster, to cluster the data. In this method clustered data will be plotted. When the data is plotted, the method of normalization will be applied on the plotted data to improve the cluster quality.
7. **Start of iteration, mean shift insertion and affinity insertion: -**In these steps of flowchart, the iterations which are defined in start of flowchart. The process of mean shift and affinity metrics is calculated and which are inserted on every iteration and with each iteration cluster quality had been improved.

Working of K-mean and Normalization Step

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 $P1, \ldots, Pk$.

Working of affinity and mean shift step

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

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 , $K = 2, \ldots, K\text{max}$;

Step 3 For $K = 2, \ldots, K\text{max}$ run step 4∼5;

Step 4 Calculate the similarity matrix S, where $S(i, j) = \exp(\ )$;

Step 5 For every $k = 2, \ldots, C\text{max}$, 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 $2 \leq k \leq C\text{max}$ is an - reasonable and δ-stable cluster number according to the results of step 4 and step 5;

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

3.4 PROPOSED WORK AND IMPLEMETATION

1. Clustering of data

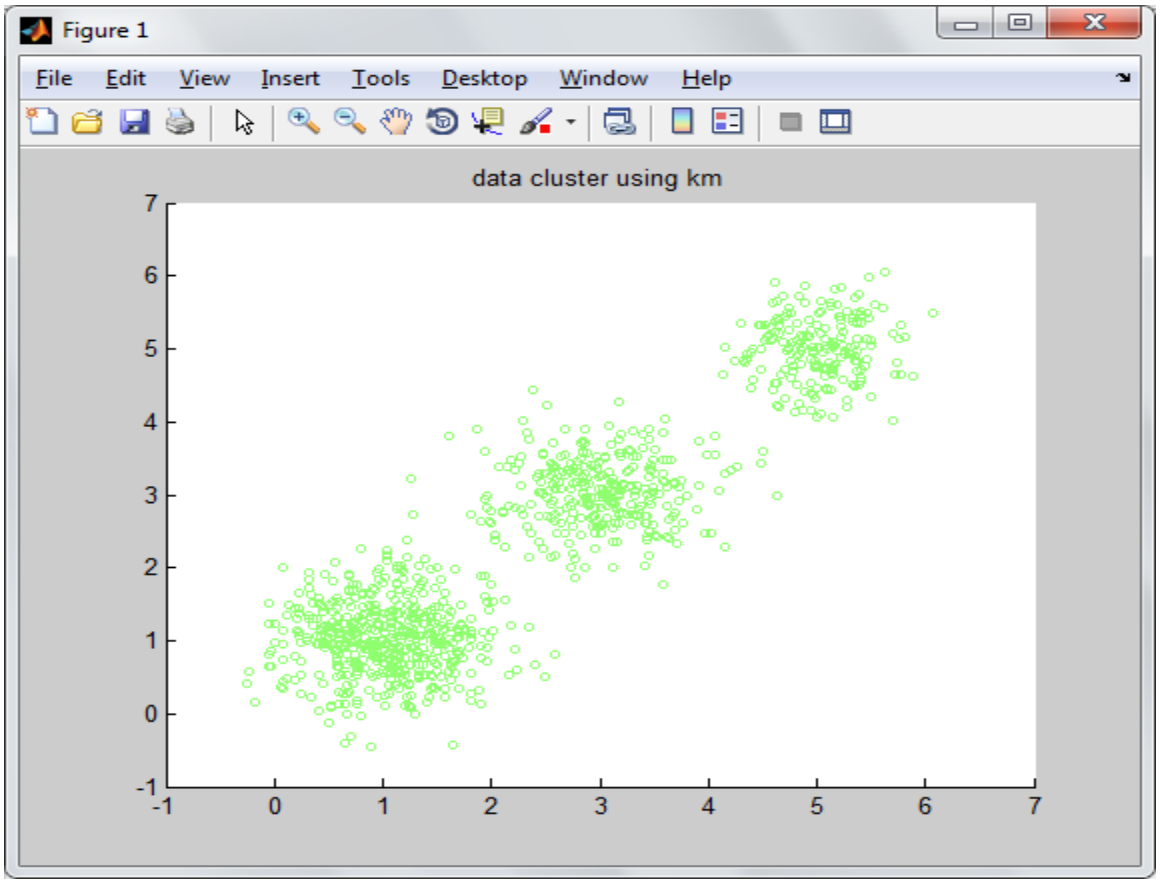

Figure 3.3: Scattering of data

As shown in figure 1.1, As in the previous chapter, the dataset which is loaded will be scattered and plotted on the 2D plane

2. Clustered Data

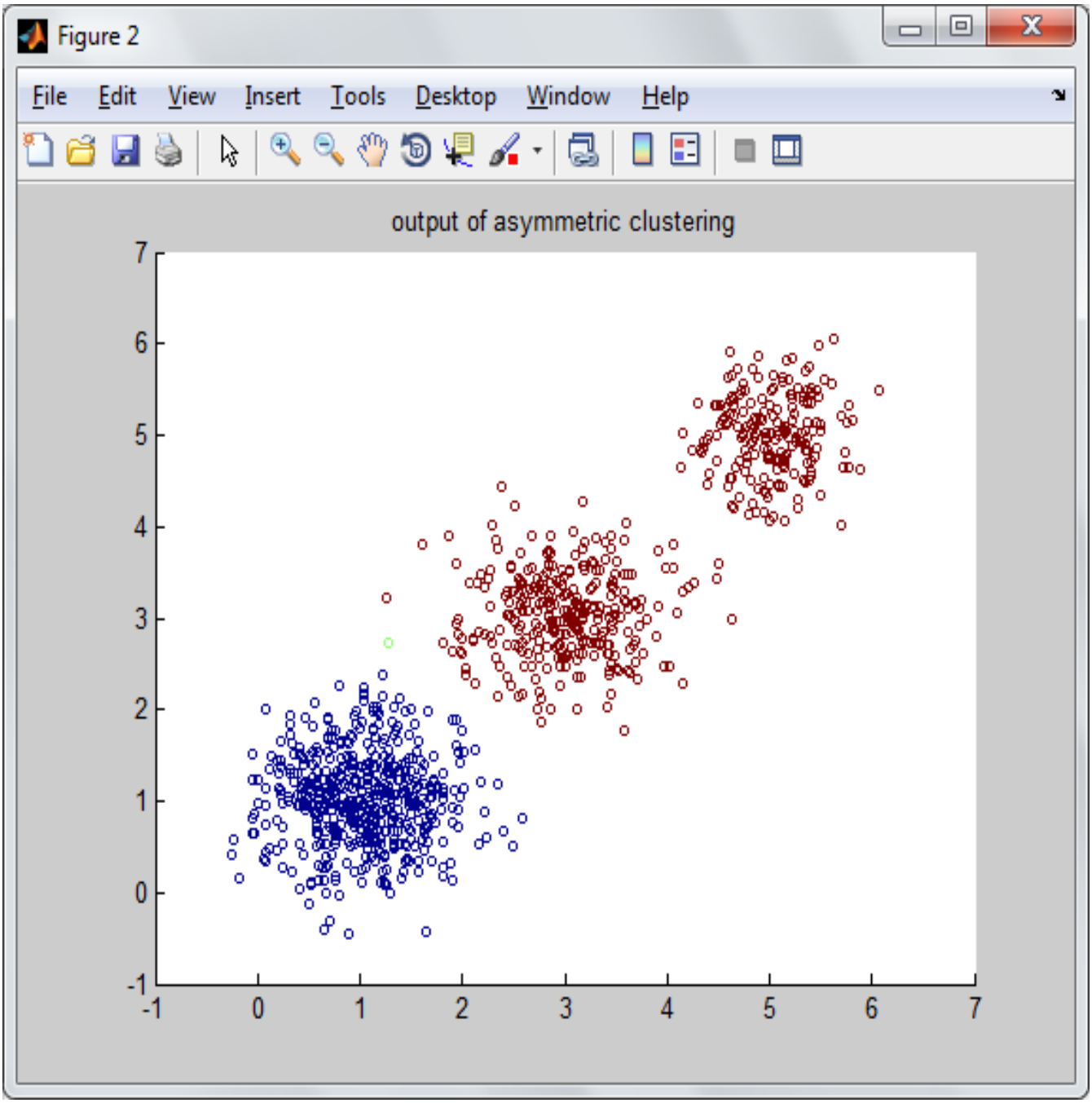

Figure 3.4: Clustering of data

As illustrated in the figure 1.2, the dataset which is loaded had been scattered and scattered data is clustered asymmetric according to asymmetric between the loaded data

3. Loading of data

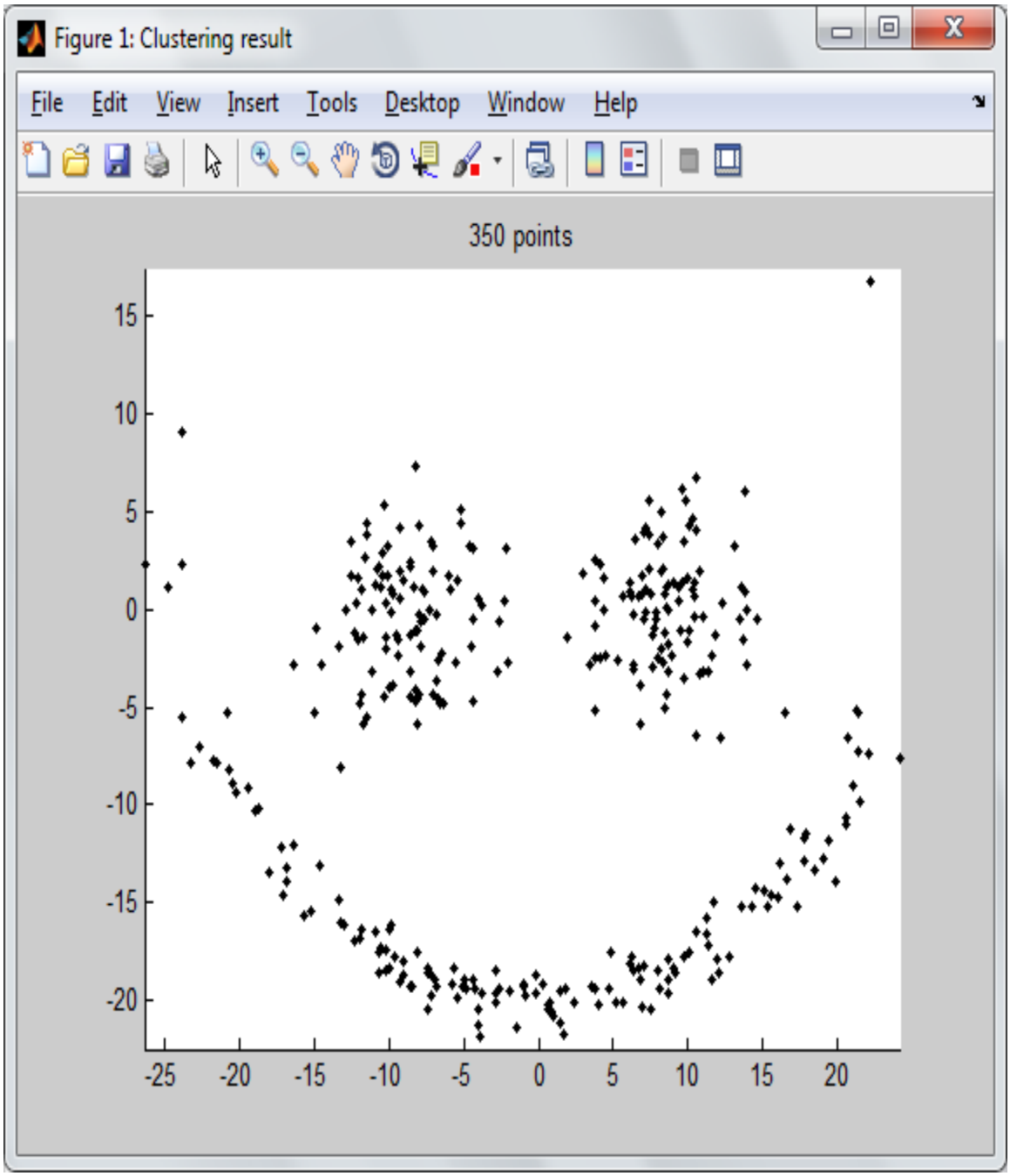

Figure 3.5: Loading of data

As illustrated in figure 1.3, the dataset is loaded and no of rows and columns are defied . The second step is to ask for iterations. According to no of iterations defined data is shown into the 2 D plane

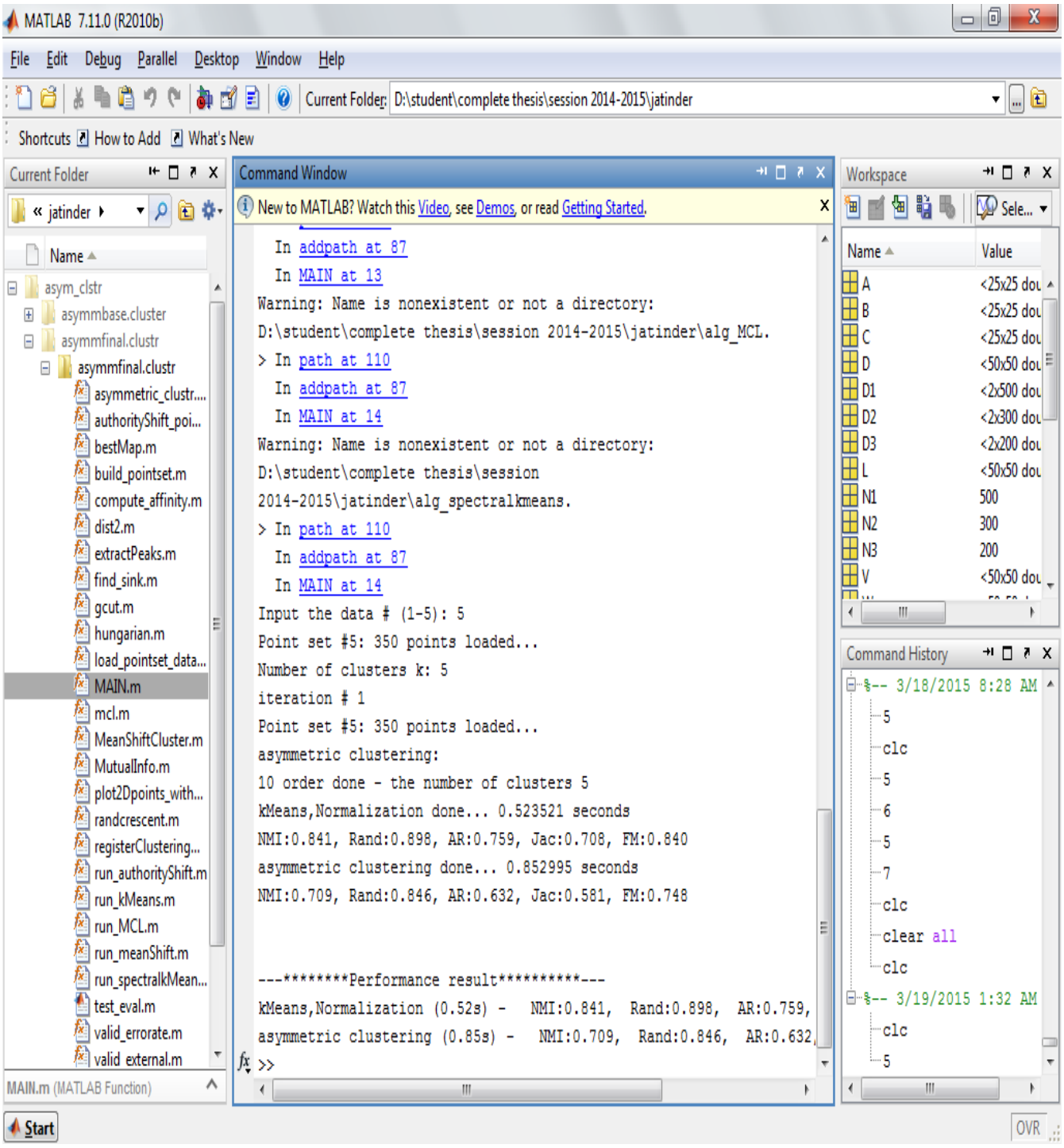

Figure 3.6: Selection of no of clusters

As shown in figure 1.4, The dataset which is loaded and on the loaded dataset, mean shift and affinity metrics is calculated, the MCL and S-clustering algorithm is applied. In this snapshot, the

normalization techniques will be applied and data will be shown in the graphically order.

1.5 Clustered data

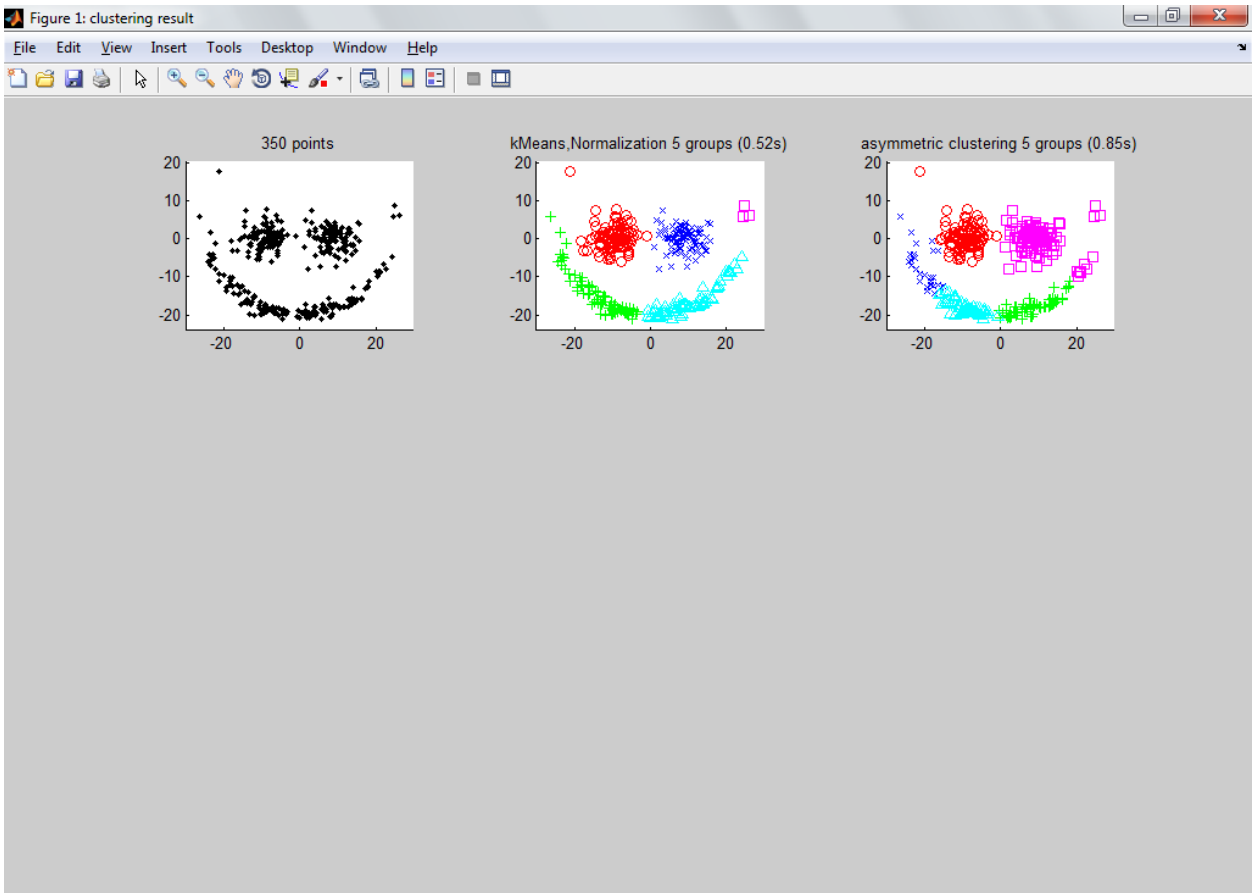

Figure 3.7: Clustering of data

As shown in figure, As illustrated in figure 1.5, the three different procedure as defined in this snapshot. In the first snapshot the different points are dataset have been scattered randomly. In the second figure, the MCL and S-Clustering is applied for graphically shown. In the third figure, the multi task assignment is applied for data clustering

CHAPTER 4
RESULTS AND DISCUSSION

The new technique implemented by us has better efficiency and result as compared to the previous algorithms. A comparison between the two techniques has been show below in form of graph.

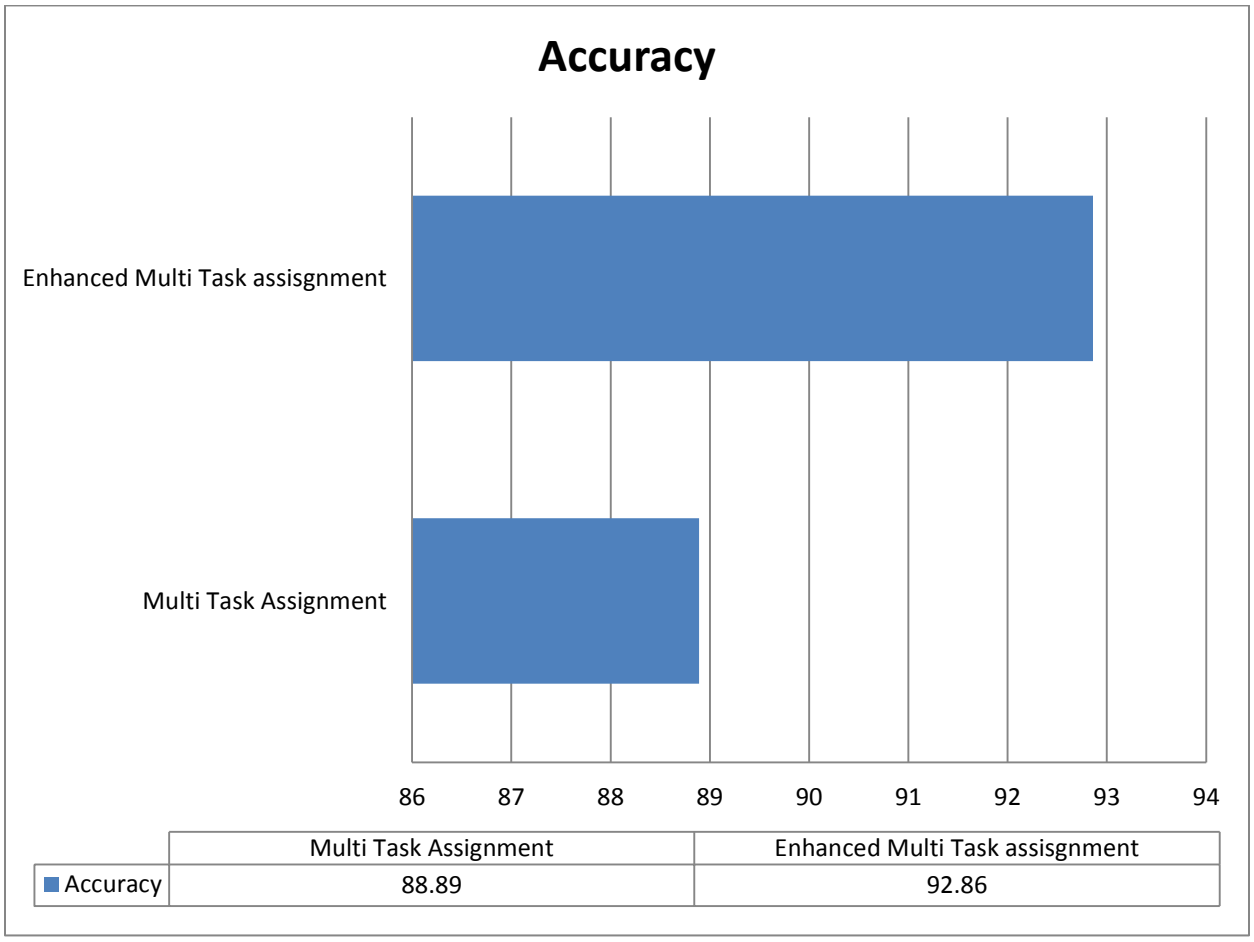

Fig.4.1 Illustrate that existing technique has accuracy 88.89 whereas proposed technique that is enhanced multi task assignment has 92.86. This shows that proposed technique is better than existing technique.

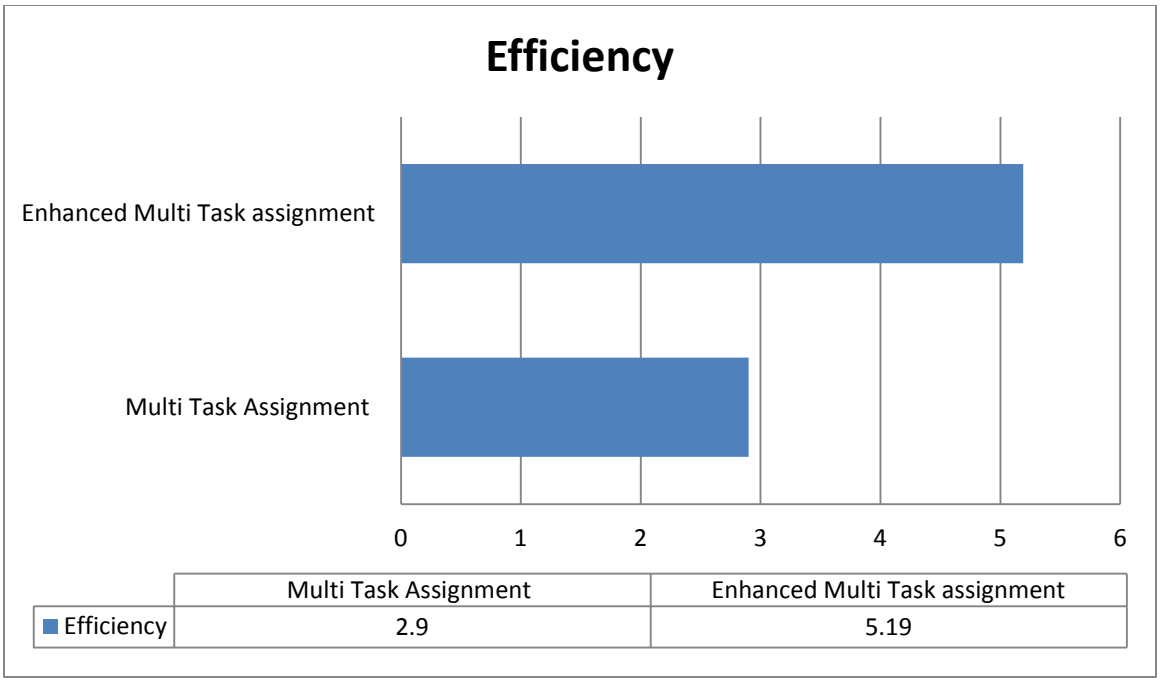

Fig.4.2 Illustrate that existing technique has efficiency 2.9 whereas proposed technique that is enhanced multi task assignment has 5.19. This shows that proposed technique is better than existing technique.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Data mining has been widely used in many application domains across various disciplines. Data mining has been applied to various data repository to get valuable information from the through the process of knowledge discovery. Data mining has been applied to solve some of the complex problems like the role mining. Mining of the Boolean data has leads to solving the problem of role mining were set of permissions is represented against the columns and user id against the rows. The main problem with the existing algorithms is that they are not much robust and are unable to handle the noise well. The algorithms proposed by us have better result in terms of accuracy, efficiency and more prone to noise levels in the dataset. By using our algorithm we can easily increase the efficiency and accuracy of the data mining system.

5.2 FUTURE SCOPE

We have implemented the new technique which has considerably increased the efficiency and accuracy of the existing algorithm. The accuracy of the new algorithm is 92.86 %. But still there is a scope of improvement in this accuracy percentage. More enhancements can be made to the algorithm to achieve the higher accuracy rate and to tackle the noisy data in better way.

REFRENCES

[1] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, "A Two-Step Method for Clustering Mixed Categroical and Numeric Data", *Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19 ,2010*

[2] R.Jensi and Dr.G.Wiselin Jiji, "A Survey On Optimization Approaches To Text Document Clustering", *International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013*

[3] Dharmendra K Roy and Lokesh K Sharma, "Genetic K-mean clustering algorithm for mixed numeric and categorical data sets", *International Journal of Artificial Intelliginece and Applications (IJAIA), Vol.1, No.2 April 2010*

[4] Andreas P. Streich , Mario Frank David Basin Joachim M. Buhmann, "Multi-Assignment Clustering for Boolean Data", Appearing in Proceedings of the 26th International Confer-ence on Machine Learning, Montreal, Canada, 2009.

[5] Mario Frank, Andreas P. Streich, David Basin, "Multi-Assignment Clustering for Boolean Data", Journal of Machine Learning Research 13 (2012) 459-489 2012

[6] Sumuya Borjigin and Chonghui Guo , "Non-Unique cluster number determination method based on stability in spectral clustering." Knowledge Info System(2013) 36:439-458.

[7] Deepika, C., and R. Rangaraj. "An Efficient Uncertain Data Point Clustering Based On Probability–Maximization Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, IJIRCEE, 2014

[8] ReshmaMR,Suchismitasahoo,"Management uncertainty and clustering in uncertain data based on KL divergence technique",IEEE, 2013

[9] Z. He, X. Xu, & S. Deng, " Scalable algorithms for clustering categorical data",*Journal of Computer Science and Intelligence Systems 20, 1077-1089, 2002*

[10] Jung, Jean Christoph, and Carsten Lutz. "Ontology-based access to probabilistic data with OWL QL." *The Semantic Web–ISWC 2012*. Springer Berlin Heidelberg, 2012. 182-197.

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

[12] Ms.Chinki Chandhok Mrs.Soni Chaturvedi, Dr.A.A Khurshid "An Approach to Image Segmentation using K-means Clustering Algorithm", *International Journal of Information Technology (IJIT),* Volume -1,Issue 1,page 110-113,August 2012

[13] Ming Chan hang et.al, "An Efficient k-Means Clustering Algorithm Using Simple Partitioning", *NSC,*2005

[14] Walaa K. Gad, Mohamed S. Kamel, "Incremental Clustering Algorithm Based on Phrase-Semantic Similarity Histogram", 2010

[15] Adrain Kuhn, Stephanne Ducasse, Tudor Girba, "Semantic Clustering: Identifying Topics in Source Code", *Elsevier,* 2006

[16] Simon Jones, Ling Shao, "Unsupervised Spectral Dual Assignment Clustering of HumanActions in Context", 2011

[17] Shaidah Jusoh and Hejab M. Alfawareh, "Techniques Applications and Challenging Issue in Text Mining uses, Applications", *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814*

[18] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, August, 2009*

[19] Shady Shehata , "Enhancing Text Clustering using Concept-based Mining Model", *Proceedings of the Sixth International Conference on Data Mining (ICDM'06) 0-7695-2701-9/06,2006*

[20] Akhil Khare, Amol N. Jadhav, "An Efficient Concept-Based Mining Model For Enhancing Text Clustering**",***IJAET/Vol.II/ Issue IV/October-December, 2011*

[21] Shady Shehata, "A WordNet-based Semantic Model for Enhancing Text Clustering", *IEEE International Conference on Data Mining Workshops, IEEE, 2009*

[22] Manne suneetha "Clustering of Web Search Results using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm*" PROCEEDINGS OF ICETECT, 2011*

[23] Ahamed Shafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified K-Means Algorithm," *International Conference on Information and Computer Networks,* Volume 27, pp-108-112, 2012.

[24] Manpreet Kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", I*nternational Journal of Advanced Research in Computer Science and Social ,* Volume 3, Issue 7, July 2013, pp-210-214 ISSN: 2277 128X

[25] Tapas Kanungo , David M. Mount , Nathan S. Netanyahu Christine, D. Piatko , Ruth Silverman and Angela Y. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation ," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 24, pp- 88-92, July 2002.

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