LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

Enhancement in Asymmetric Clustering Algorithm to Remove Noise from Dataset to Reduce Escape Time

A Dissertation

Submitted

By

Jatinder Singh

(11306702)

To

Department of Computer Science

In partial fulfillment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science

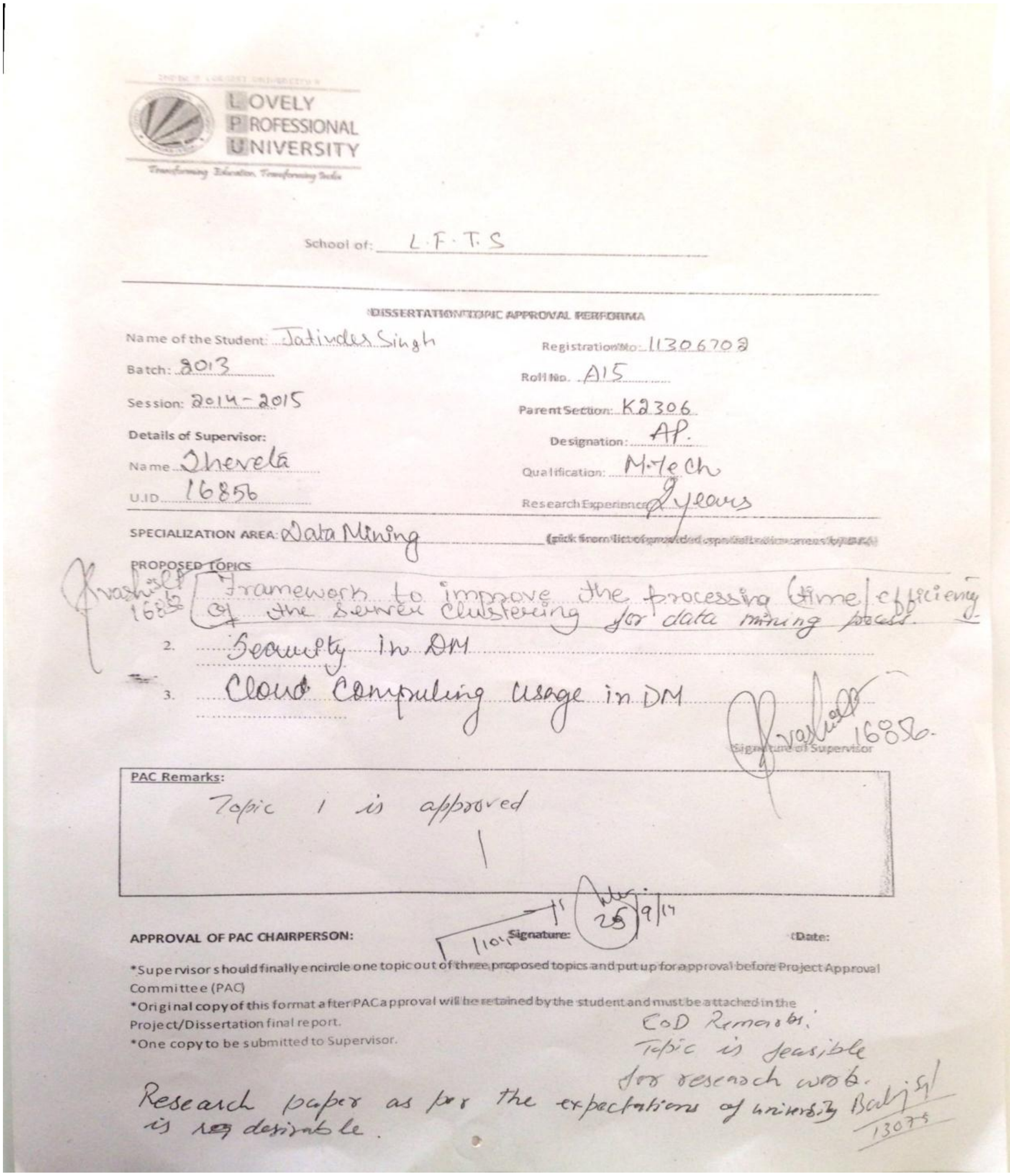
Under the guidance of

Sheveta Vashisht

(16856)

(May 2015)

Approval of PAC



LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: L.F.T.S

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Jatinder Singh

Registration No: 11306702

Batch: 2013

Roll No. A15

Session: 2014-2015

Parent Section: K2306

Details of Supervisor:

Designation: AP.

Name Shweta

Qualification: M.Tech

U.ID 16856

Research Experience 2 years

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

PROPOSED TOPICS

Vashist 16856 1. Framework to improve the processing time/efficiency of the server clustering for data mining process.

2. Security in DM

3. Cloud Computing usage in DM

Vashist 16856.

Signature of Supervisor

PAC Remarks:

Topic 1 is approved

APPROVAL OF PAC CHAIRPERSON: Signature: 25/9/14 Date:

*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.

CoD Remarks:
Topic is feasible
for research work.

Research paper as per the expectations of university is ~~req~~ desirable.

Baljit 13075

ABSTRACT

There is large amount of data is present in the world. This data is coming from various sources like companies, organizations, social networking sites, image processing, world wide web, scientific and medical data etc. Peoples do not have time to look all this data. They attended towards the precious and interested information. Data mining is technique which is used to extract meaning full information from huge databases. Extracted information is visualized in the form of statics, graphs, and tables and vides etc. There are number of data mining techniques and asymmetric clustering is one of them. Asymmetric technique is type of unsupervised learning. In this, data sets which have similarity are placed in one cluster and others are in other clusters. From, number of years various asymmetric clustering technique are introduced which work well with datasets. These techniques do not work well with the complex and strongly coupled data sets. To reduce processing time and improve accuracy neural networks are combined with asymmetric clustering algorithms.

ACKNOWLEDGEMENT

First I offer my sincerest gratitude to my supervisor, Sheveta Vashisht, who has supported me throughout my thesis .Without her this thesis would not have been completed or written. I am thankful for her aspiring guidance, invaluably constructive criticism and friendly advice during the work am sincerely grateful to her for sharing their truthful and illuminating views on a number of issues related to the research. Finally, I thank my parents for supporting me throughout all my studies at University.

I would also like to thanks my family and friends who have been a source of encouragement and inspiration throughout the duration of the research.

DECLARATION

I hereby declare that the dissertation entitled, "**Enhancement in Asymmetric Clustering Algorithm to Remove Noise from Dataset to Reduce Escape Time**." submitted for M-Tech degree is entirely my original work and all ideas and references have been duly acknowledged .It does not contain any work for award of any other degree or diploma.

Date **Jatinder Singh**

11306702

CERTIFICATE

This is to certify that Jatinder Singh has completed M.TECH dissertation titled **“Enhancement in Asymmetric Clustering Algorithm to Remove Noise from Dataset to Reduce Escape Time**.**”** 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 is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date: **Sheveta Vashisht**

(16856)

TABLE OF CONTENTS

LIST OF FIGURES

Chapter -1

INTRODUCTION

The sheer amount of data is stored in world today called big data. In 2001, it is assumed that about 8, 50,000 petabytes of data is stored in the world and it is expected that it will be about 35 zettabyte in 2022[1]. Mostly, data is generated by the social websites, market analysis medical field, web mining and image processing etc. This data is stored in large databases in the forms of tables, images and videos etc. called data warehouses. The process of extracting useful patterns or knowledge from data base is called data mining. The extracted information is visualized in the form of charts, graph and tables etc. Data mining is also known by another name called KDD (knowledge discovery from the database). In data mining, frequent item set is used to find relations between numerous numbers of fields in data mining. Association rules are used to discover the frequent data item sets. The concept of association rules is used in various fields like retail stores, market strategy and stock market etc.

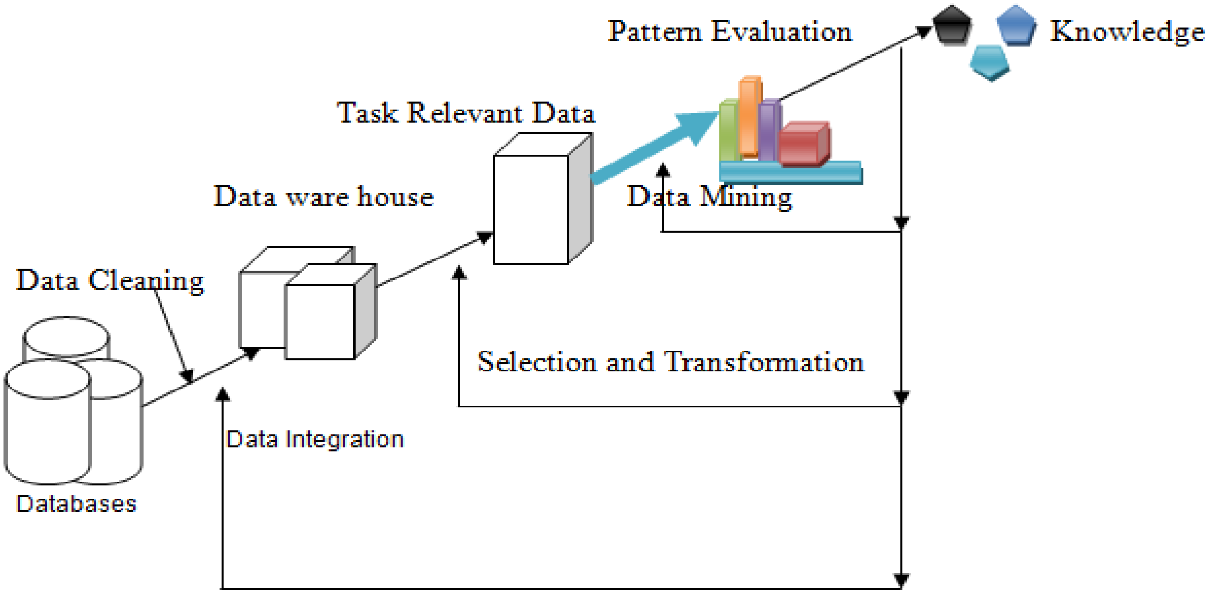


Figure 1.1: Data Mining Process

We know that these days Informational technology is mounting and databases created by organizations and companies like telecommunications, banking, marketing, transportation, manufacturing, and social networking sites etc. are becoming huge day by day. Knowledge discovery process is used to store this data in databases and efficiently access the interested or useful data from databases. Knowledge discovery consist of following steps [10]:

i. **Data Cleaning:** It is the step in which the process of detecting and removing of data which is not correct, irrelevant, containing missing values, duplicate values and noise that is dirty data from the database.

i. **Data Integration:** It is the step in which data from different sources is collected in one source to provide unified view of data.

ii. **Data Selection:** It is the step in which data analysis is done in way that the selection of relevant data from databases.

iii. **Data Transformation:** It is the step in which the data which is selected is reformed to correct form performing various operations like summary, aggregations, generalizations and normalized operations.

iv. **Data Mining:** This is important technique in which intelligent operations are used to extract the useful pattern from the database.

v. **Pattern Evaluation:** It is the step in which the required pattern are evaluated from the given database.

vi. **Knowledge Representation:** It is the step in which output of whole process is visualized to user. There are many techniques to represent the data like graphs, tables and graphs etc.

1.1 Advantages of Data Mining

Marketing / Retail

Data mining plays important role in marketing companies. Market companies develop new models which are based on historical data. It is helpful for predicting to response of new marketing campaigns. These marketing campaigns may be direct mail, online marketing campaign and others. Data mining helps to retailer to select best products for customers. Profits of products also increase by using data mining.

Different benefits are provided to retailer through data mining. Market basket analysis is very beneficial for companies because companies can arrange the products according to frequent purchased products. In this way, Customer can buy products with less time. It is helpful for both retailer and customer. By good arrangement of products, customer takes less time while selection of products. By using market basket analysis, retailer can provide discounts on products by which sale of products will increase.

Finance / Banking

All financial information can be provided by data mining. It provides information about loans. Banks can get idea about loans from historical data. It can also help for determine fraud people**.**

Manufacturing

Data gives information about the fault equipment's. Through this information, manufacturers can take decision to remove the fault from equipment's.

Governments

Data mining plays important role in government agency. It analyzes records of financial transaction which are used to create patterns. These patterns detect information related to criminal.

1.2 Disadvantages of Data Mining

Privacy Issues

There is fear in the mind of people because someone can hack the information of people. Information related to customers can be getting by businesses for knowing buying habits of people. At this time, personal information can be leak.

Security issues

Major issue related to data mining is security. Businesses want all personal information about customers which may be hacked by hackers.

Misuse of information/inaccurate information

Data mining gives information related to customer that can be misuse by unauthorized people.

1.3 Classification of Data Mining System: Data mining system is classified according to following categories:

i. **According to Data source to be mined:** Data mine system can be classified according to kinds of mined techniques used like spatial data, multimedia data etc
ii. **According to Data models:** Data mine systems may use many models like relational model, object oriented model and transactional models**.**
iii. **According to kind of Knowledge mined:** Data mine system can be classified according to the type of knowledge is used like classification, prediction, cluster analysis and outlier analysis.
iv. **According to utilized Mining technique:** Data mine system can be classified according to techniques used for data mining techniques like decision tree, neural network etc.
v. **According to adapted applications:** Data mine systems can be classified according to applications adapted like in finance, data mining system related to finance is used.

1.4 Major issues in Data Mining

There are various data mining algorithms and techniques but now there is large volume of data in world and this data is increasing day by day, issues that can be raised in data mining systems can be scalability and reliability of performance of data mining system. Various performance issues are:

i. **Effective, Efficient and Scalable data mining**: in order to efficiently extract the useful knowledge from the large amount of databases, the technique of data mining which we are using should be effective, efficient and scalable, gives desired outputs in the desired time.
ii. **Parallel, Distributed and Incremental mining algorithms**: The volume of data present in the databases is very huge and to maintain the complexity of data, data

mining techniques prompt to develop the parallel and distributed data mining algorithms. Data in these algorithms is stored in different partitions and processed parallel. The output which comes from these partitions is combined to provide desired results and this is quite tough job to mine data without any scratch.

1.5 Data Mining Techniques: There are several major *data mining techniques* have been developing and using in data mining. These techniques are as follow:

1. Association: Association data mining is one of the techniques which are most openly. Association data mining technique is also called as relation technique because on the basis of relationship between items patterns are discovered. In market basket analysis customer's behavior is judged by observing the items which customer's buys frequently. Association techniques are used to research customer's habit in buying items. On the basis of past records salesman can attract customers by giving them combo offers on items which they mostly buys. Like, customer come to buy computer and for salesman it will be good to show them peripheral devices along with computer. So, customer gets attracted toward these peripheral devices like mouse, speakers, and pen drives etc.

Association rule mining (ARM) is used to discover exciting relationship between the items. After ARM many algorithm are produced one of them is apriori algorithm and then improvement is done in the apriori algorithm. Han and Fu change the minimum support threshold for association rule; the algorithm that is F-P algorithm there is no need of generating the candidate item in this algorithm. Some of these algorithms very slows to show the result in reasonable time.

There are two steps in association rule mining:

- **Creating item sets which are frequent:** the present item sets must be equal to or more than the min support count.
- **Generate strong rules:** the condition for having a rule is strong is that it must be satisfying the min support and min confidence. Also introducing the following concepts:

Item set defines the total items present in the set. K item set shows the existence of k items in the set. Example can be taken as, {laptop, Software, pen drive} which is a 3-Itemset.

Support count provides the occurrence of items in the given item set. Frequent item set contains the items which satisfy the min support count [6].

2. Classification: Classification is data processing technique that is predicated on machine learning. Basically, classification is of two sorts supervised classification and unsupervised classification. In international organization supervised classification result's unknown and it referred to as clustering technique. Supervised classification is employed to classify the set of information in each item into one in all predefined set of categories or teams. Mathematical techniques like applied math, linear programming and statistics are utilized by classification. In classification, computer code is developing which may find out how to classify the info things into teams. Classification is apply in application that carry all records of staff UN agency left the corporate, and predict UN agency are left in future" during this case, divide the records of staff into two teams tagged as "leave" and "stay". At the top data processing software can classified them. It's primarily prediction of outcome of output primarily based upon given input.

3. Prediction: The prediction is one in every of a knowledge mining techniques that confirm association between independent variables and relationship between dependent and independent variables. It's used to predict the profit of the corporate by analysing its history record. It's based mostly upon dependent and independent variables each. Then supported the historical sale and profit information, a fitted regression line that's used for profit prediction is drawn.

4. Sequential Pattern: Sequential patterns analysis is one in every of data processing technique that identifies similar patterns, regular events or trends in group action information over a business amount. In sales, businesses will determine a group of things that customers obtain along totally different times during a year with historical group action information, then businesses will simply use this data to advocate customers buy it with higher deals supported their buying frequency within the past.

5. Decision Tree: Decision tree is one in every of the foremost used data processing techniques. It's terribly model is straightforward to understand for users. In decision tree technique, the basis of the choice tree could be a straightforward question or condition that has multiple answers then results in a group of queries or conditions that facilitate us confirm the information so we will build the ultimate decision supported it. it's like hierarchical technique .

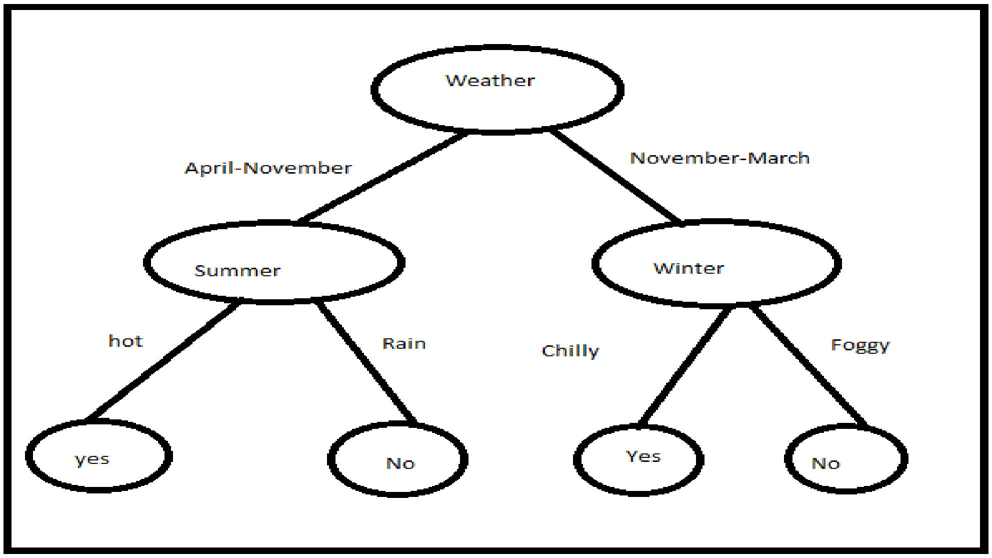


Figure 1.2 Decision Tree

1.6 Clustering in Data Mining

Clustering means putting objects having similar properties into one group and objects having dissimilar properties into another. For example, object having values above threshold values can be placed in one cluster and values below into another cluster. Clustering divides the large data set into groups or clusters according to similarity in properties. Clustering is an unsupervised learning technique as there are no classifiers and their labels .It is form of learning by observation. Cluster analysis can be used in the areas such as image processing, analysis of data, market research (buying patterns) etc. Using clustering we can do outlier detection where outliers are values lying outside the cluster.

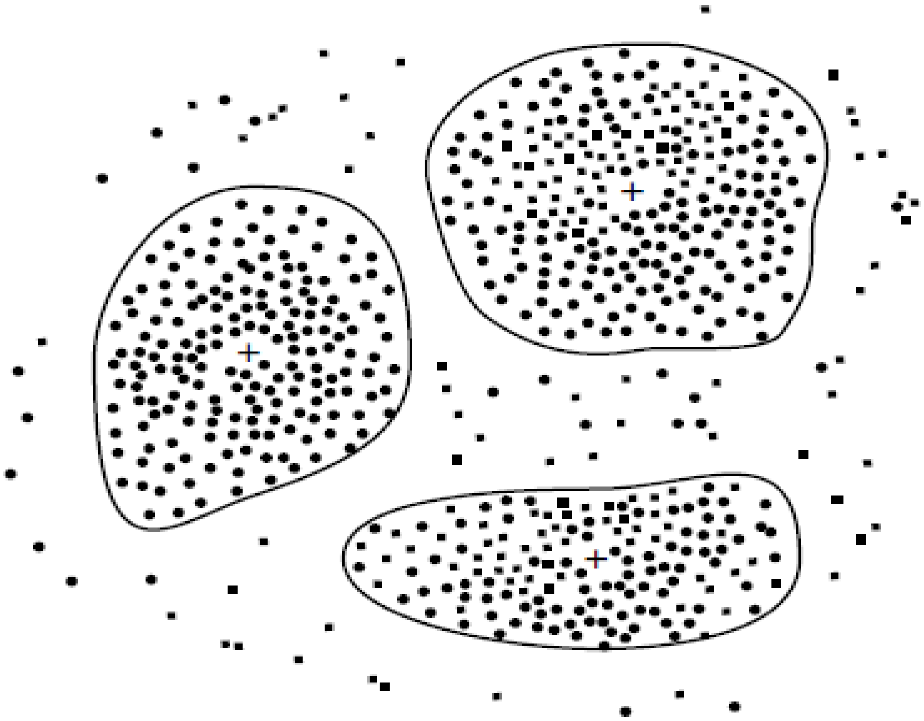


Figure 1.3 Clusters and Outliers [29]

In above figure 1.3 the dots which are outside the clusters represent outliers and there are clusters of object with similar properties.

The main difference between clustering and the nearest neighbor technique is that clustering is an unsupervised learning technique on the other hand nearest neighbor is generally used for prediction or a supervised learning technique. Unsupervised learning techniques are not under some supervision whereas supervised technique is under supervision like learning under teacher. In prediction, the patterns that are found in the database and presented in the model are always the most important patterns in the database for performing some particular prediction. In clustering there is no particular sense of why certain records are near to each other or why they all fall into the same cluster.

1.6.1 Types of Clustering

There are different types of clustering in data mining. These are as follow**:**

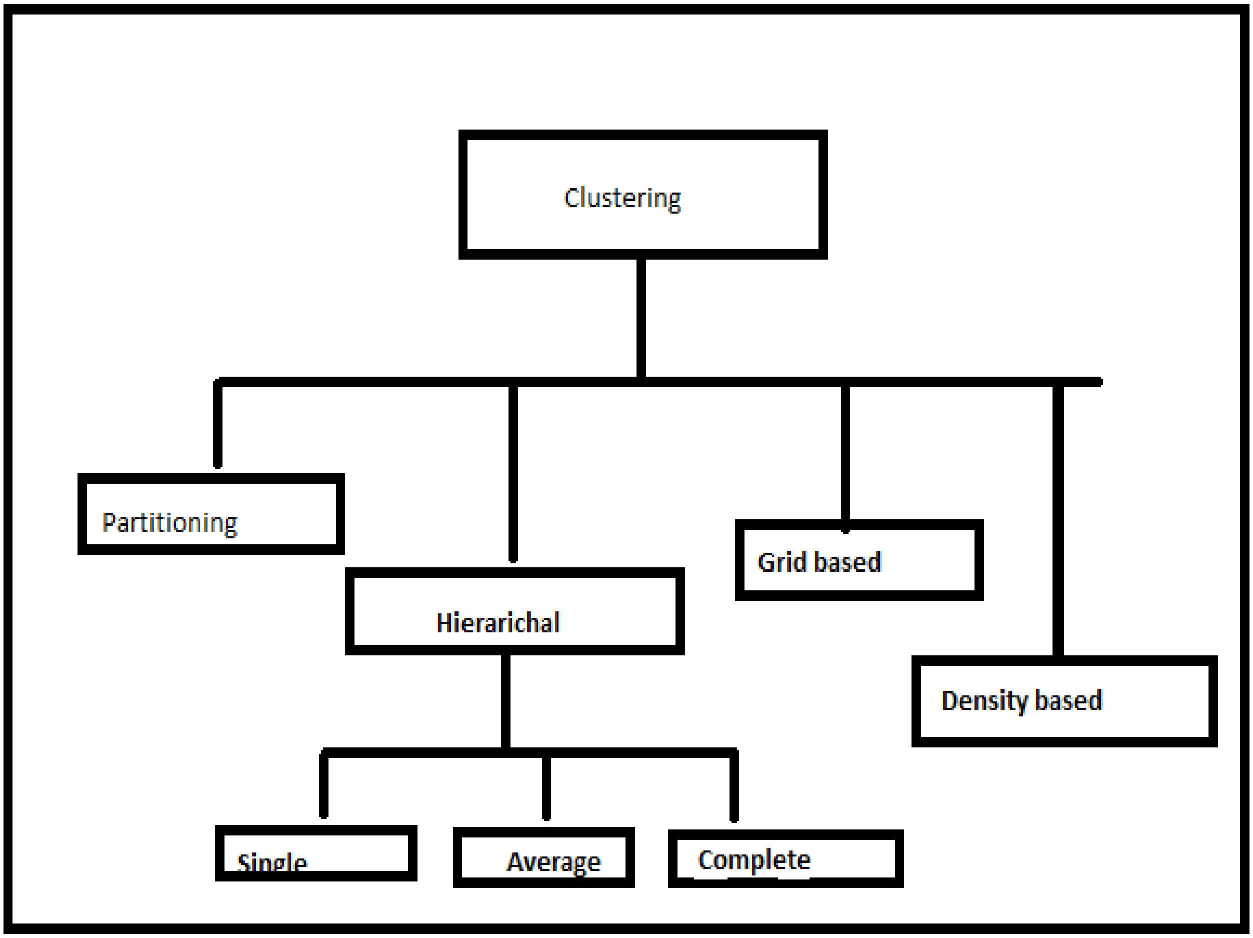


Figure1.4: Clustering in Data Mining

1.6.1.1 Partitioning Clustering: In partitioning clustering the partition of objects is done and is called clusters. The clusters formed to optimize an objective partitioning criterion in way that on the basis of distance the dissimilarity function, the similar objects are in cluster and the objects of different cluster are not similar with respect to attributes of data set. There are two partitioning methods one is k-means and other is k-medoids. In k-means algorithm from the input of n objects k clusters are made, the intra cluster similarity is high as compared to inter cluster similarity in k means algorithm. For, object with large values k means algorithm is sensitive to outliers and which may distort the data distribution. In k-medoids, the representative object is medoid called centrally located object, this is the basis of k-medoid method. The initial object in k-medoid is arbitrarily choose. The iterative process of swapping representing objects by other objects is running till the quality of clustering is enhanced.

1.6.1.2 Density Based Clustering: Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. In these methods the cluster is continue to grow as long as the density in the neighborhood exceeds some threshold [3].

1.6.1.3 Hierarchical Clustering: In hierarchical clustering data objects are grouped into tree of clusters. It is further classified into agglomerative and divisive clustering which depends on hierarchical breakdown formed by merging called bottom-up and splitting called top-down approach. The output of hierarchical clustering depends from its ability to do placements once a merge or split decision been executed. Agglomerative clustering is bottom-up strategy in which objects are placed in its own cluster and atomic clusters are merged in larger and larger cluster. In divisive hierarchical clustering the top-down strategy does the reverse of agglomerative clustering by starting the placing all objects in one cluster. It decomposes clusters into smaller pieces until each object forms a cluster and satisfies some termination conditions.

1.6.1.4 Grid Based Clustering: Data structure used in grid based clustering is multitier. It utilize the space of object in the finite number of cells which makes gris structure on which all clustering operations are performed. This approach has fast processing time and independent of number of data objects but it is dependent on the number of cells in each dimension in given space. STINGS, CLIQUE are the examples of grid based methods.

1.7 Asymmetric Clustering

In an asymmetric clustering, the partitions can be affirmed vigorously and usually run on a single cluster at a time. The task which is specific to an appropriate cluster and it can be routed to that cluster. The Asymmetric Clustering applications are used in electronic trading systems which are utilized by banks. Asymmetric clustering architecture is contrary to the typical stateless server farm where the whole application is simulated across machines. From time to time using distributed caching products for performance increasing. In an asymmetric cluster, business logic is dividing into partitions, where every partition can be the singular accessory of a set of underlying data. As a resultant, in each node in the cluster implementing its own local cache resulting in high performance reading and writing without the need to maintain a distributed cache between cluster nodes. An asymmetric clusters is practically the opposite of a symmetric cluster:

i. Applications can declare named partitions at any point while it’s running

ii. Partitions are highly available uniquely named singletons and run on a single cluster member at

iii. Incoming work for a partition is routed to the cluster member hosting the partition

The application is a modal. Partitions have a lifecycle of their own and can start background threads/alarms as well as respond to incoming events whether they are IIOP/HTTP or JMS/foreign messages. The other situation where asymmetric clustering beans plus partitioning improves performance is that both these features provide low level primitives that can be used to practically build a custom application container that fully leverages the features of the application server.

The J2EE specification provides a very high level set of services to an application developer. If this set of services isn’t what the application developer requires then the standard spec does not allow the developer to provide an alternative. The normal option is write a standalone J2SE application that does it or try to have features added to the next release of the commercial application server. The partitioning facility and asymmetric beans provide low level primitives that allow this code to run on top of the application server in a fully supported manner. This approach allows advanced customers to have the benefit of using a commercial application server without the normal limitations of the J2EE one size fits all philosophy.

Chapter- 2

REVIEW OF LITERATURE

Hao Huang *et al* (2014) mentioned that [2] mining arbitrary shaped clusters in large data sets is an open challenge in data mining. There are number of solutions to this problem have been proposed with high time complexity. Computational cost can be saved by using some algorithms that try to shrink a data set size to a smaller amount of representative data examples. But, the user-defined shrinking ratios can affect the clustering performance. CLASP an effective and efficient algorithm for mining arbitrary shape clusters presented in this paper. It automatically shrinks the size of a data set while effectively preserving the shape information of clusters in the data set with representative data examples. At that time, it alters the positions of these representative data examples to enhance their intrinsic relationship and make the cluster structures more clear and distinct for clustering. At last, it makes agglomerative clustering to identify the cluster structures with the help of Pk metric which is mutual k-nearest neighbours based metric. Extensive experiments on both synthetic and real data sets are performed, and the results prove the effectiveness and efficiency of this approach.

Gunnar Carlsson *et al* (2014) introduce [3] hierarchical quasi clustering methods; a generalization of hierarchical clustering the output structure of asymmetric networks in asymmetric networks preserves the asymmetry of the input data. Asymmetric network's output structure is equivalent to a finite quasi ultra-metric space and study admissibility with respect to two desirable properties. They prove that a modified version of single linkage is the only admissible quasi-clustering method. Moreover, show stability of the proposed method and they fulfil invariance properties established by it. New algorithms are developed and the value of quasi-clustering analysis is illustrated with a study of internal migration within United States.

R.Jensi *et al* (2013) proposed [4] that text Document Clustering is one of the fastest growing research areas because of huge amount of information is available in an electronic form. The number of techniques designed for clustering documents in such a way that documents with high intra-similarity are in same cluster and low inter-similarity documents are in same cluster. Mostly clustering algorithms in documents provide localized search in effectively navigating, summarizing, and organizing the information. The solution of this can be obtained by applying high-speed and high-quality optimization algorithms. This optimization algorithm globalized search the entire data. A brief survey on optimization approaches to text document clustering is tried to find out in this paper. This survey on text document clustering starts with a introduction about clustering in data mining then soft computing after this explore various research papers.

R.Jensi *et al* (2013) they represented [4] Text Document Clustering is one of the fastest growing research areas because of availability of huge amount of information in an electronic form. There are several number of techniques launched for clustering documents in such a way that documents within a cluster have high intra-similarity and low inter-similarity to other clusters. Many document clustering algorithms provide localized search in effectively navigating, summarizing, and organizing information. A global optimal solution can be obtained by applying high-speed and high-quality optimization algorithms. The optimization technique performs a globalized search in the entire solution space. In this paper, a brief survey on optimization approaches to text document clustering is turned out. This survey starts with a brief introduction about clustering in data mining, soft computing and explored various research papers related to text document clustering. More research works have to be carried out based on semantic to make the quality of text document clustering.

Mahendra Pratap Yadav *et al* (2012) explain [5] relationship between data mining and e-commerce with the continuously increasing growth of data in World Wide Web is discussed. The user wants to extract desirable information and resources. The main idea of this research is to find the behavior of customer that what they want or what are their requirements. For e-commerce conventional methods are no longer useful to find customers behavior. With the advanced technologies, large amount of data is stored in servers about thousands number of customers profiles and from they can search the data about customers' requirements. K-

Means algorithm in cluster customer is used for mining the input data coming from various e-commerce websites. To increase customer's behavior in online shopping strategy of attracting customers with good offers and combos is done by seeing their profiles. Age, gender and behavior are main attributes for analyzing the customers marketing in e-commerce.

Satoshi Takumi *et al* (2012) explains [6] algometric algorithms of hierarchal clustering using the asymmetric similarity measures. There are linkage methods proposed into this research are of two methods, first bottom up methods and other is top down methods. The bottom up method first searches similarity measure between objects and then searches similarity measure in the cluster. Whereas, in top-down approach is vice-versa of bottom up approach that is it first check similarity measures between cluster and after that it checks similarity measures between the objects. The tree diagram structure used to show result of hierarchical clustering called dendrogram result of the hierarchical clustering sometimes shows reversely. This paper gives emphasis to show no reversals in the dendogram. The first method of bottom-up approach does not show reversal in output of algometric hierarchal clustering and another method top-down approach use hypothesis. Example of this is based on real data which show these methods work.

Neelamadhab Padhy *et al* (2012) gave an overview of data mining and areas where it can be used .They told data mining can be used to extract information from very large amount of data .They mentioned the data mining techniques : Decision tree and rules ,classification methods and nonlinear regression etc. [11] .They told areas where data mining can be done to get information which can be used for making decisions .Areas are Healthcare ,Education Systems ,CRM ,Web Education ,Sports data mining ,E-Commerce etc. The various data mining techniques are used to extract the useful patterns.

S.R.Pande *et al* (2012) Provides [7] the data mining techniques of clustering. Cluster analysis divides data into the groups having similar properties. Clustering is unsupervised classification technique. Clustering is divided into two classes, first is hierarchical clustering techniques and other is partitioning technique. Partitioning clustering techniques include K-means, K-mediods, and CLARA etc. The hierarchical method forms tree like structure. It includes agglomerative and divisive technique. They also density based methods like

DBSCAN, DENCLUE. In this paper they process of clustering from the point of view of the data mining.

Ming-Yi Shih *et al* (2010) explained [8] about various clustering algorithms have been developed diverse domains in which data is stored in form of groups. The work in these clustering algorithms is either on pure numeric data or on pure categorical data, and on the mixed categorical and numeric data types. A new two-step clustering method is presented in this paper. In this method the items in categorical attributes are processed to construct the similarity or relationships among them based on the co-occurrence and after that categorical attributes are converted into numeric attributes on the basis of their relationships. All categorical data is converted into numeric, and the previously presented clustering algorithms can be applied to the dataset. However, the existing clustering algorithms has some disadvantages or weakness, the two-step method adds attribute to cluster with integrated hierarchical and partitioning clustering algorithm. This method explains the relationships between items and improves the weaknesses in single clustering algorithms. Experimental analysis shows that accurate and strong results can be obtained by applying this method to cluster mixed numeric and categorical method.

Wilhelmiina Hamalainen *et al* (2008) introduced [9] searching significant statically association rule is very important but it often neglected. It is consider it is not feasible to apply statistical significant rules to larger data sets. Author introduced pruning techniques, breadth first strategy and Stat Apriori algorithm to search all significant statistical association rules in reasonable time. Stat Aprior is used in two ways that it search k most association rule and passes the significant threshold and solve multiple testing problems. It prunes all the spare association rules. This experimental result shown by avoiding over fitting rule's quality can be improved. Mainly, main idea of experiment is to check speed ratio accuracy of Stat Apriori algorithm. Data in stat Apriori is selected on the basis of their minimum confidence. Data in Stat Apriori is selected on the basis of their minimum confidence.

Hui Xiong *et.al* (2010) presented [11] in data mining removing noise by removing objects which are the reason of noise in database is important goal of data cleaning. Most existing data cleaning methods mainly aims to remove noise that is show low level data errors and cause imperfect database but the objects which are not relevant can prevent data analysis.

The main objective is to update the data analysis at that stage on which these kinds of objects are considered as noise. Data set consist of huge amount of noise and these techniques need to remove large fraction of data. In this paper out of four, three are traditional outlier detection techniques which are distance based, clustering based and third one is the LOF (local outlier factor). The fourth technique is hyper clique based data noise remover. This paper shows by using hyper clique technique good clustering output and higher quality is achieved than other associated methods.

Yu Qian *et al* (2005) introduced [12] in last few years, the use of visualization is increased in all field specially in data mining.in this paper challenging issue of using virtualization in data mining is addressed by using some parameters for data cleaning methods. Through visualization performance of algorithm is improved and user can easily understand and provide the feedback by visualizing the data. There are many visualization models called waterfall model which is described for spatial data cleaning in four aspects: dimension-independent, data quality, algorithm parameter selection and measurement of noise removing on parameters.

Luis Daza *et a*l (2007*)* in this paper algorithm called Qclean Noise are introduced to analyze noisy instances. The effect of this algorithm is applied on three supervised classifiers: LDA, KNN and RPART, a decision tree classifier is explained. The comparison with other procedures is also described with well-known machine learning datasets. At last experiments results shows that this algorithm is better than other procedures used for detecting noisy instances using escape time.

Sumit Garg *et al* (2013) data mining is one of the fast growing fields in the computer science.in data mining new patterns are discovered in large datasets. There are many data mining algorithms are developed in last number of years. Single algorithm may not apply on all applications due to the reasons of the different data types of applications. That’s why right selection of algorithms does not depends upon the application used but it depends upon the type of data type is used in the applications.

Chapter -3

PRESENT WORK

3.1 SCOPE OF STUDY

In the previous times, various clustering algorithms had been developed to cluster data when diversion is seen in the given datasets. Now days, data in the world is increasing day by day like in social networking sites, market analysis, medical field, image processing and world wide web etc. volume of data is continuously increasing. To, store and efficiently access this data we need to make cluster of similar and dissimilar. There are many clustering techniques are used to perform various type of operation on data in databases. The clustering algorithms which are recommended can give good performance like provide good efficiency on the numeric and pure categorical type of data. This proposed algorithm perform good operation on simple and statistical database and but will not perform desired operation on data which are complex and of mixed category like plant dataset which we consider for this work.

In previous year an efficient clustering algorithm is proposed which works in two steps to find clusters for complex type of data. The two step algorithm works as every dataset had some of the attributes and then to cluster data the relationship between the attributes are maintained and similarity between attributes are derived, on the basis of similarity derived, the data will be clustered. This algorithm will also be applied on hierarchical and partitioning methods. This method shows relationships between the items or objects and tries to improve the weakness of using single clustering algorithms. In this research enhancement in asymmetric clustering is done on the basis on two-step algorithm. Clusters of complex data is made the conclusion taken from the output and enhancement will be main agenda of this work.

3.2 PROBLEM FORMULATION

Cluster analysis is being broadly used in several applications like basket analysis, e-commerce, image processing, scientific and medical field, data analysis, and word wide web etc. Today in business, stock market clustering can support marketers to determine interest's vendors and customers based on their record of purchasing patterns and distinguish groups of their customers who are interested in goods. In medical science, cluster analysis can be used to derive new plant like testing new hybrid species or estimating the conditions in which they grow well and observing soil and water quality. Animal taxonomies, classify their genetic factors with similar functionality. In geology, expert can use clustering technique to recognize areas of similar interests, lands, similar, houses and infrastructure in a city or in country etc. Data clustering technique is also useful in organizing data on the World Wide Web for interested knowledge or data.

Clustering is an unsupervised classification technique that aims at generating collections of items, or clusters in that way that object with similar properties are grouped together in same cluster and objects with different cluster are quite distant. Mining arbitrary shaped clusters in large data sets is an open challenge in data mining. The number of solutions of these problems has been proposed with high time complexity. Computational cost can be saved by using some algorithms by shrinking a data set size to a smaller amount data examples and user defined threshold ratios can affect the clustering performances. The CLASP (clustering algorithm for arbitrary shaped clusters) algorithm is an effective and efficient algorithm for mining arbitrary shaped clusters which automatically shrinks the size of a data set while effectively preserving the shape information of clusters in the data set with representative data examples. After this it changes the locations of these data examples to improve their intrinsic relationship and make the cluster structures more clear and distinct for clustering. At last, it does agglomerative clustering to find the cluster structures with the help of pk metric called mutual k-nearest neighbor-based similarity metric. In this work, the enhancement of the asymmetric clustering algorithms to increase the quality of cluster and improve the efficiency of algorithms.

3.3 OBJECTIVES

Objectives are the most vital part of our research which helps us to achieve our desired aims and expectations from the research. It is said that if objectives are clearly defined that half part of research is considered to be done. In this research, we are going to use neural networks along with two step clustering algorithm. The two step algorithm technique works in two steps and deals with tough and complex data types.

Here, we will use neural network along with two step clustering algorithm technique to enhance efficiency and accuracy of complex data sets. Here are the main objectives of our research:

1. To study and analyse various asymmetric clustering technique to cluster relevant and irrelevant data.
2. To propose enhancement in the asymmetric algorithm to improve accuracy of the algorithm.
3. To improve the clusters in the purposed work using neural networks.
4. To implement proposed algorithm and existing algorithms and analyse the results in terms of accuracy and cluster quality

3.4 RESEARCH METHODOLOGY

Asymmetric clustering 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, asymmetric clustering algorithm adopts that how many clusters k are present in database before which is not true in real time applications. Asymmetric is an iterative technique and these algorithms are sensitive towards initial centers selection.

Asymmetric clustering 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 asymmetric algorithms.in this research previous asymmetric clustering algorithm is enhanced to new asymmetric clustering algorithm to enhance the throughput in the term oh accuracy. Enhanced asymmetric algorithm uses k-means, normalization, mean shift, markov clustering algorithm and s-cluster to improve the accuracy of previous symmetric clustering algorithm. Flow chart (fig. 3.1) of previous research is defined as:

1. Define rows, column and load dataset :- This is the first step of the flowchart in which to define number of rows, columns and no of iteration for clustering on the dataset which is loaded for clustering. The number of rows and columns define size of the dataset and according to data set, number of iterations will be defined for best clustering results in terms of accuracy.

2. Scatter of data with k-mean clustering: - According to number of iterations, k-mean algorithm will be implemented which the base of asymmetric clustering. K-means is one of the easiest unsupervised learning algorithms that give the solution of the well-known clustering problem .The Lloyd’s algorithm is known as k-means algorithm. According to clustering central points are defined and loaded data will be defined under certain classes according to their similarity. The similarity between the data points will be calculated using Euclidian distance.

3. Apply Normalization on Scattered Data: - When the data points are plotted according to their similarity, normalization technique will be applied on the points. The normalization will assign classes to that points which are not assigned to any of the defined class. This technique will further improve cluster quality

4. Apply Strom Function: - The storm function is the function which is applied after k-mean clustering. This function will plot the x and y axis on the plotted data to recognize that which part of how much points are similar and which points are dissimilar

5. Random pick, pairing and exponential function: - These three functions are used for the asymmetric clustering after k-mean algorithm. In this functions random pick function will chose any random point from the dataset and apply mapping with pairing and exponential function to choose asymmetric points.

6. Apply N-cut algorithm and plot results: - In the previous step points are choose according to their symmetric and asymmetric relations, the n-cut algorithm will define its position in the plotting of final result.

The N-cut is the algorithm which is used for segmentation or to divide similar and dissimilar datasets. In this work, N-cut algorithm is applied on dataset which is clustered and to improve the cluster quality, dataset will be further clustered and in each cluster uniqueness will be calculated using the equation

$$\text{ncut}(A,B) = \frac{w(A,B)}{w(A,V)} + \frac{w(A,B)}{w(B,V)}$$

In the equation A and B are two clusters and w is the weight on each cluster. V is the variance of uniqueness between two clusters.

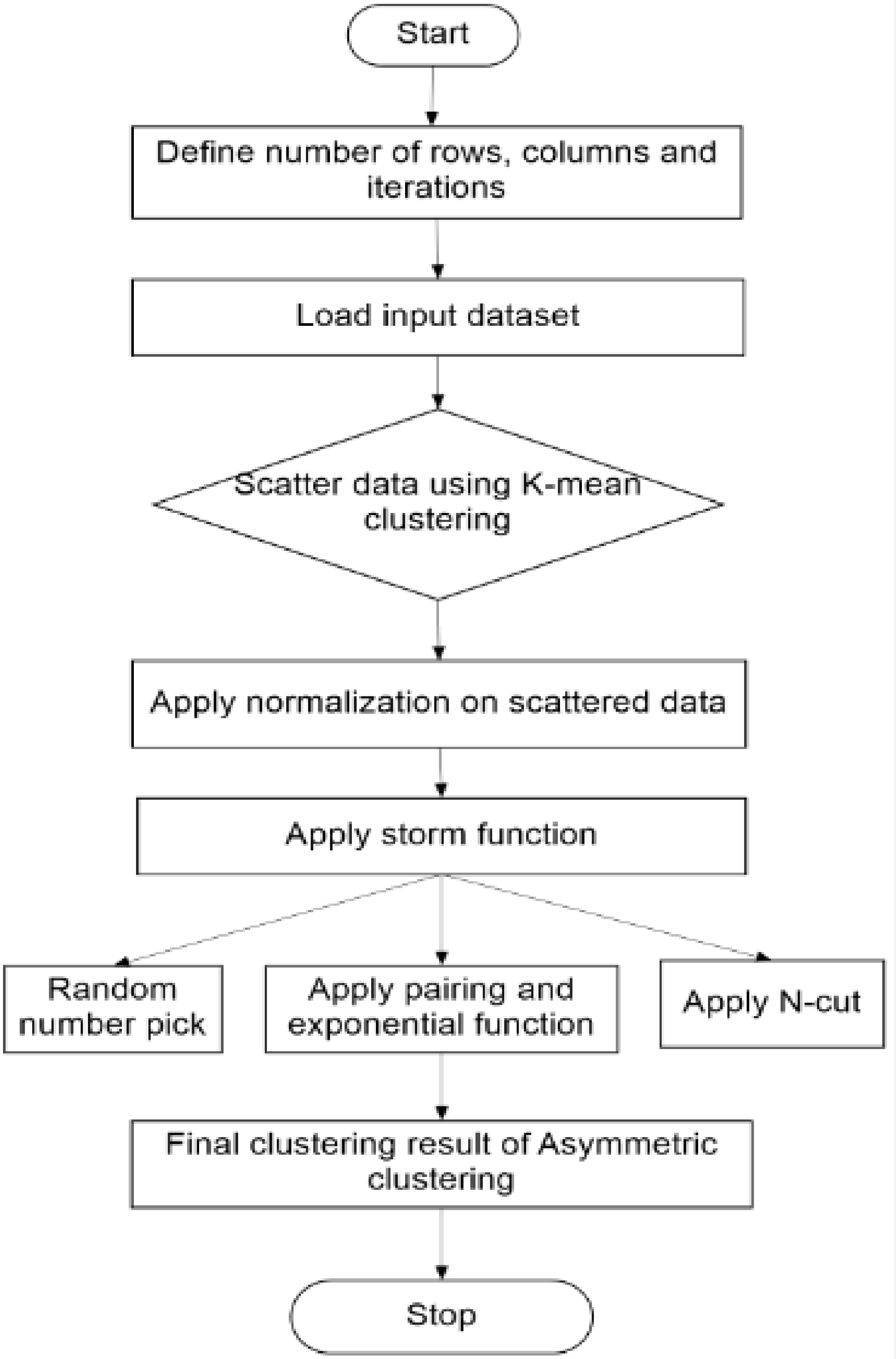

Start
Define number of rows, columns and iterations
Load input dataset
Scatter data using K-mean clustering
Apply normalization on scattered data
Apply storm function
Random number pick
Apply pairing and exponential function
Apply N-cut
Final clustering result of Asymmetric clustering
Stop

Figure 3.1 Flowchart of Previous work done

3.4.1 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.

Working of affinity and mean shift step

Input: Data set $P = \{p1, p2, . . . , pn\}$, > 0, $\delta > 0$, user-specified upper threshold Cmax ≥ 2 for cluster number to be testified, user-specified maximum number of neighbors Kmax ≥ 2.

```
    **Step 1** Calculate the distance matrix $W$;
    **Step 2** For $i = 1, 2, . . . , n$, sort the $i$th row of $W$, then calculate $pi\ K$ , which is the $K$th
    **Step 3** For $K = 2, . . . , K$max run step 4∼5;
    **Step 4** Calculate the similarity matrix $S$, where $S(i, j\ )$ = exp( );
    **Step 5** For every $k = 2, . . . ,C$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 6 To determine whether the candidate cluster number $2 \leq k \leq C\max$ 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.

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.

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.

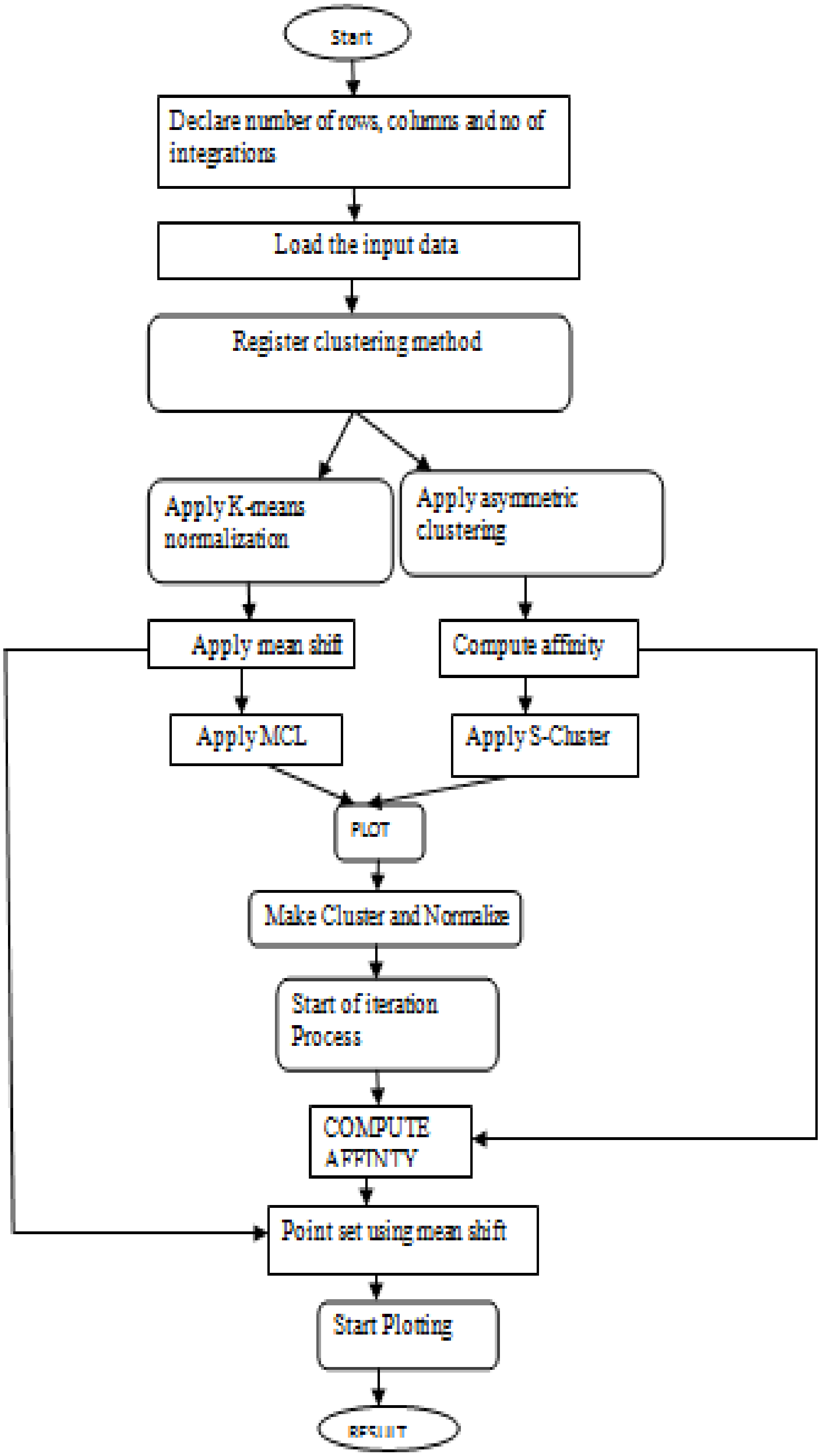

Start
Declare number of rows, columns and no of integrations
Load the input data
Register clustering method
Apply K-means normalization
Apply asymmetric clustering
Apply mean shift
Compute affinity
Apply MCL
Apply S-Cluster
PLOT
Make Cluster and Normalize
Start of iteration Process
COMPUTE AFFINTY
Point set using mean shift
Start Plotting
RESULT

Figure 3.2 Flowchart of Proposed work

3.4.2 k-means Clustering

Clustering means dividing large data sets into smaller data set of some similarity. In this we create clusters and elements of cluster will have same properties. K-means is one of the easiest unsupervised learning algorithms that give the solution of the well-known clustering problem .The Lloyd's algorithm is known as k-means algorithm.

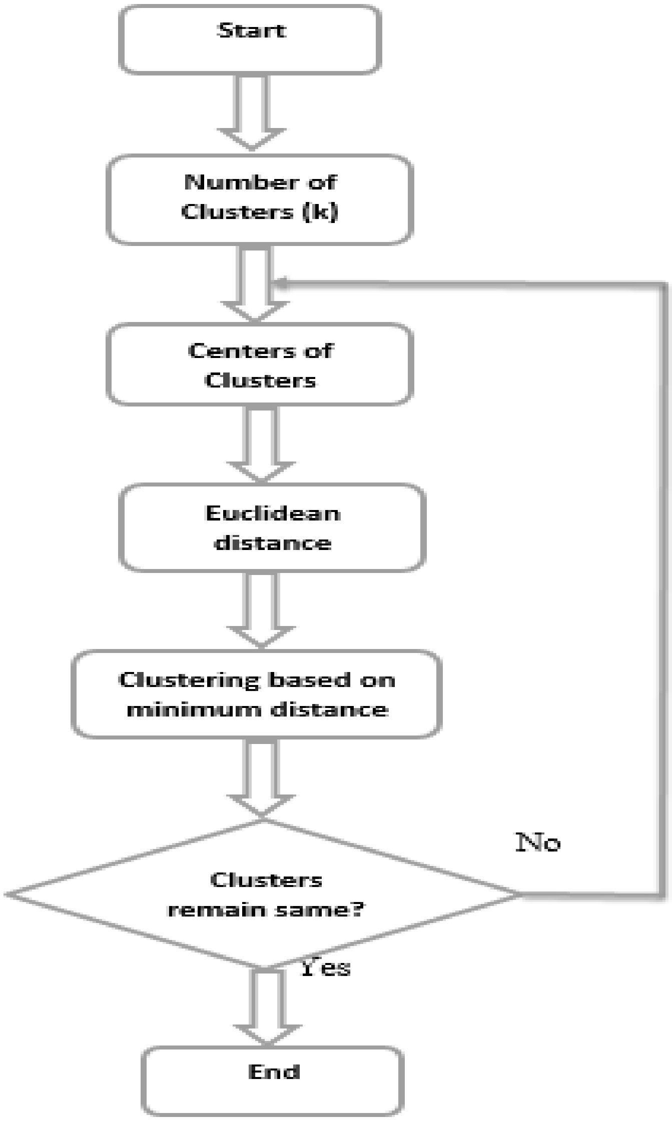


Figure 3.3: Flow chart of k-means

K-means algorithm works as follows:

Step 1: k represent number of clusters .First we have to decide number of clusters we want to make.

Step 2: Initialize the centers of clusters.

Step 3: Calculate the distance between each point and centers of clusters initialize above.

$$\| x_i - v_j \|^2$$

$\| x_i - v_j \|^2$ is the Euclidean distance between x_i and v_j.

Step 4: Assign the points to the clusters whose distance from the cluster center is minimum of all the cluster centers**.**

Step 5: Recalculate the cluster centers and repeat steps 3 and 4.

Step 6: If clusters data remains same then finish, otherwise repeat steps 3 and 4.

3.4.3 Overall work Flowchart

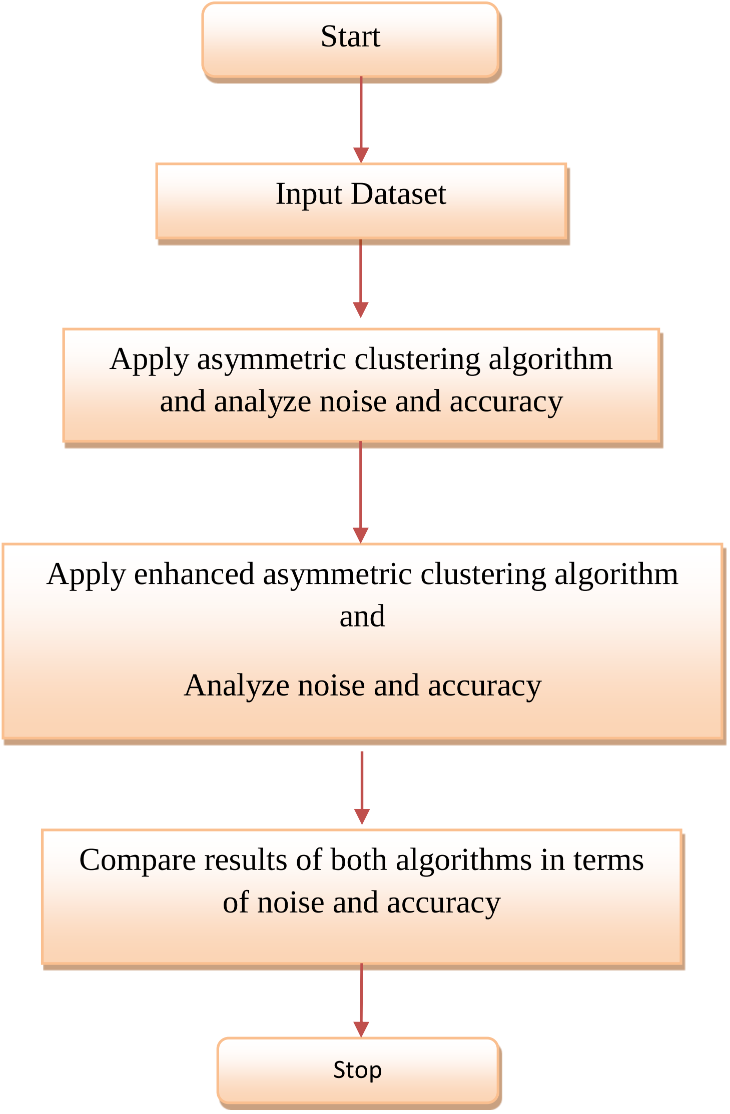

Start
Input Dataset
Apply asymmetric clustering algorithm and analyze noise and accuracy
Apply enhanced asymmetric clustering algorithm and
Analyze noise and accuracy
Compare results of both algorithms in terms of noise and accuracy
Stop

Figure 3.3 Flowchart of overall work

Explanation of Overall work Flowchart

1. **Input 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. **Apply asymmetric clustering algorithm and analyse noise and accuracy:** In this step, previous asymmetric clustering algorithm is applied on the data set and make clusters of input dataset. K-means and normalization techniques are used to make clusters. These three functions are used for the asymmetric clustering after k-mean algorithm. In this functions random pick function will chose any random point from the dataset and apply mapping with pairing and exponential function to choose asymmetric points. The n-cut algorithm will define its position in the plotting of final result.

3. **Apply enhanced asymmetric clustering algorithm and Analyse noise and accuracy:** The K-mean normalization method which is selected for clustering with the normalization equation. The second method is of asymmetric clustering which is implemented to cluster the asymmetric data from the loaded dataset. In this step, two operations are performed. In the first step mean shift algorithm is applied on the loaded dataset. The second method is of affinity metrics, it is equation which is applied to find relationship between various elements of the dataset. 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.

4. **Compare results of both algorithms in terms of noise and accuracy:** Both the algorithms are compared in the term of removing and improved accuracy.

.

3.5 Tool Used

Introduction to MATLAB

MATLAB is a high-level language and interactive environment by which we can perform computationally intensive tasks with more speed traditional programming languages such as C, C++, and FORTRAN [Mat lab Toolbox]. Or **MATLAB** (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language.
Mat lab has following functions:

1. Introduction and Key Features

2. Developing Algorithms and Applications

3. Analyzing and Accessing Data

4. Visualizing Data

5. Performing Numeric Computation

6. Publishing Results and Deploying Applications

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, numeric computation. It helps to solve more complex problems that are very difficult with traditional programming languages, such as C, C++, and FORTRAN. Add-on toolboxes (collections of special-purpose MATLAB functions, available separately) extend the MATLAB environment to solve particular classes of problems in these application areas. MATLAB gives various feature to simplify the task. MATLAB code can integrate with other languages and applications, and distribute MATLAB algorithms and applications.

Some features of MATLAB

1. MATLAB matrices and vectors

2. Dense matrices and vectors

3. Range operator

4. Size and shape operator

5. MATLAB arithmetic operators

6. Array operations

7. Back slash operator

8. Complex arithmetic

9. MATLAB software

10. Linear algebra

11. Nonlinear functions

12. Ordinary differential equations

13. Fourier Transformations

MATLAB Graphics

1. Plotting data(x, y)

2. Plotting data(x, y, z)

3. Movies

4. Saving Post script graph

MATLAB data handling

1. Importing ASCII data

2. Exporting ASCII data

Programming hints

1. Some basics

2. M files

3. Good Practices

Examples of problems solved with MATLAB

1. Partial differential equations

2. Matrix examples

Some more advanced examples

1. Using c++ with MATLAB

Advantages of MATLAB

1. It performs numerical calculations and visualized result programming

2. It provides graphics easily and produces code efficiently

Disadvantages of MATLAB

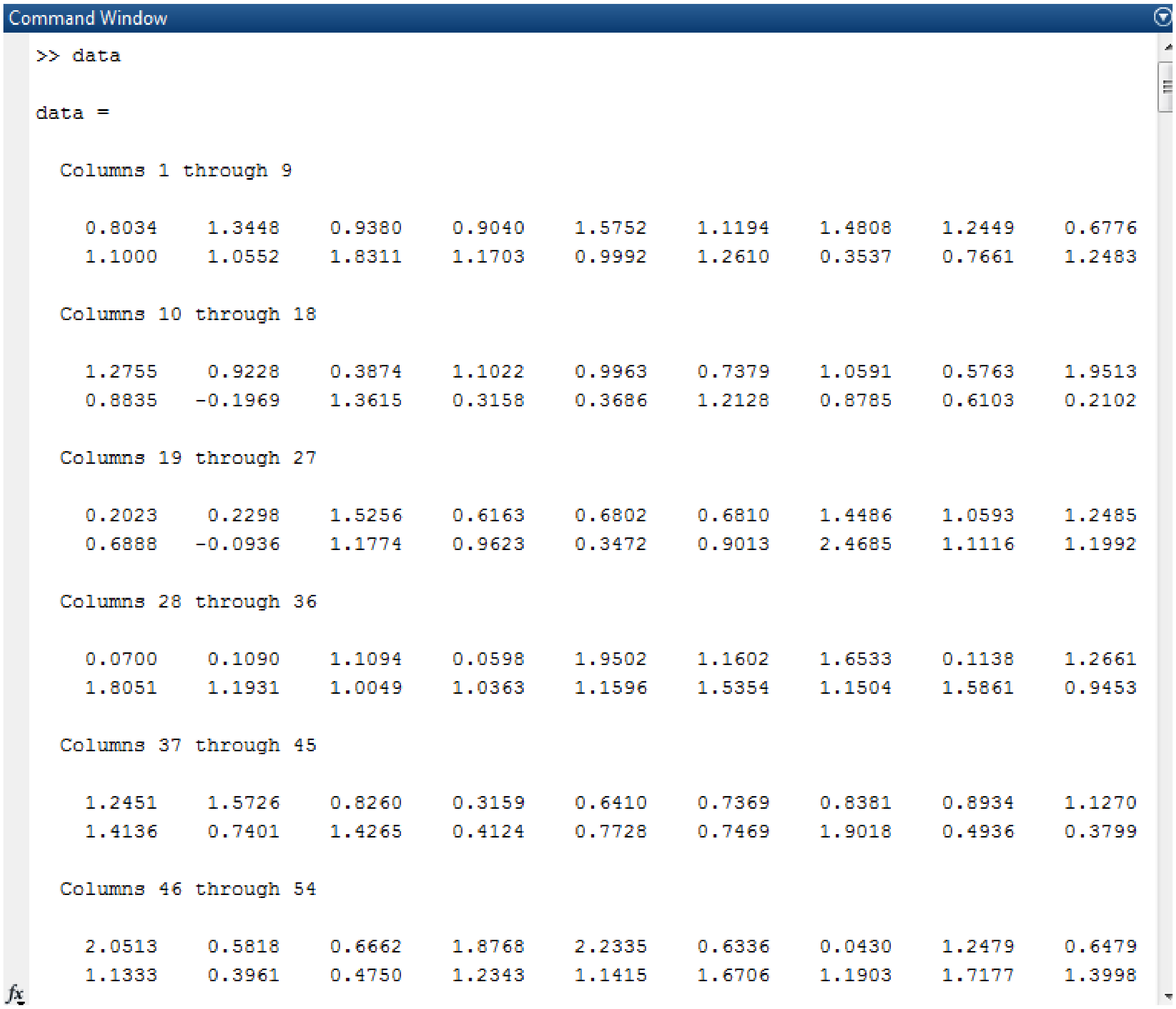
1. Because MATLAB is an interpreted language, it can be solved or not solved

2. Poor programming practices

CHAPTER-4

IMPLEMENTATION

Data set is taken from online UCI(Machine learning repository) which provides datasets. Figure 4.1 represents the dataset and its attributes. Some attributes of our dataset are date(April, May, June..),temp (lt-norm,norm,gt-norm,?.), crop-hist (diff-last-year, same-last-yr,same-last-two-yrs., same-last-sev-yrs,?.), plant-growth(norm,abnorm,?.), leaves (norm,abnorm.) etc..



```
>> data

data =

  Columns 1 through 9

    0.8034    1.3448    0.9380    0.9040    1.5752    1.1194    1.4808    1.2449    0.6776
    1.1000    1.0552    1.8311    1.1703    0.9992    1.2610    0.3537    0.7661    1.2483

  Columns 10 through 18

    1.2755    0.9228    0.3874    1.1022    0.9963    0.7379    1.0591    0.5763    1.9513
    0.8835   -0.1969    1.3615    0.3158    0.3686    1.2128    0.8785    0.6103    0.2102

  Columns 19 through 27

    0.2023    0.2298    1.5256    0.6163    0.6802    0.6810    1.4486    1.0593    1.2485
    0.6888   -0.0936    1.1774    0.9623    0.3472    0.9013    2.4685    1.1116    1.1992

  Columns 28 through 36

    0.0700    0.1090    1.1094    0.0598    1.9502    1.1602    1.6533    0.1138    1.2661
    1.8051    1.1931    1.0049    1.0363    1.1596    1.5354    1.1504    1.5861    0.9453

  Columns 37 through 45

    1.2451    1.5726    0.8260    0.3159    0.6410    0.7369    0.8381    0.8934    1.1270
    1.4136    0.7401    1.4265    0.4124    0.7728    0.7469    1.9018    0.4936    0.3799

  Columns 46 through 54

    2.0513    0.5818    0.6662    1.8768    2.2335    0.6336    0.0430    1.2479    0.6479
    1.1333    0.3961    0.4750    1.2343    1.1415    1.6706    1.1903    1.7177    1.3998
```

Figure 4.1: Attributes of data

K-mean algorithm will be implemented which the base of asymmetric clustering. According to clustering central points are defined and loaded data will be defined under certain classes according to their similarity. The similarity between the data points will be calculated using Euclidian distance.

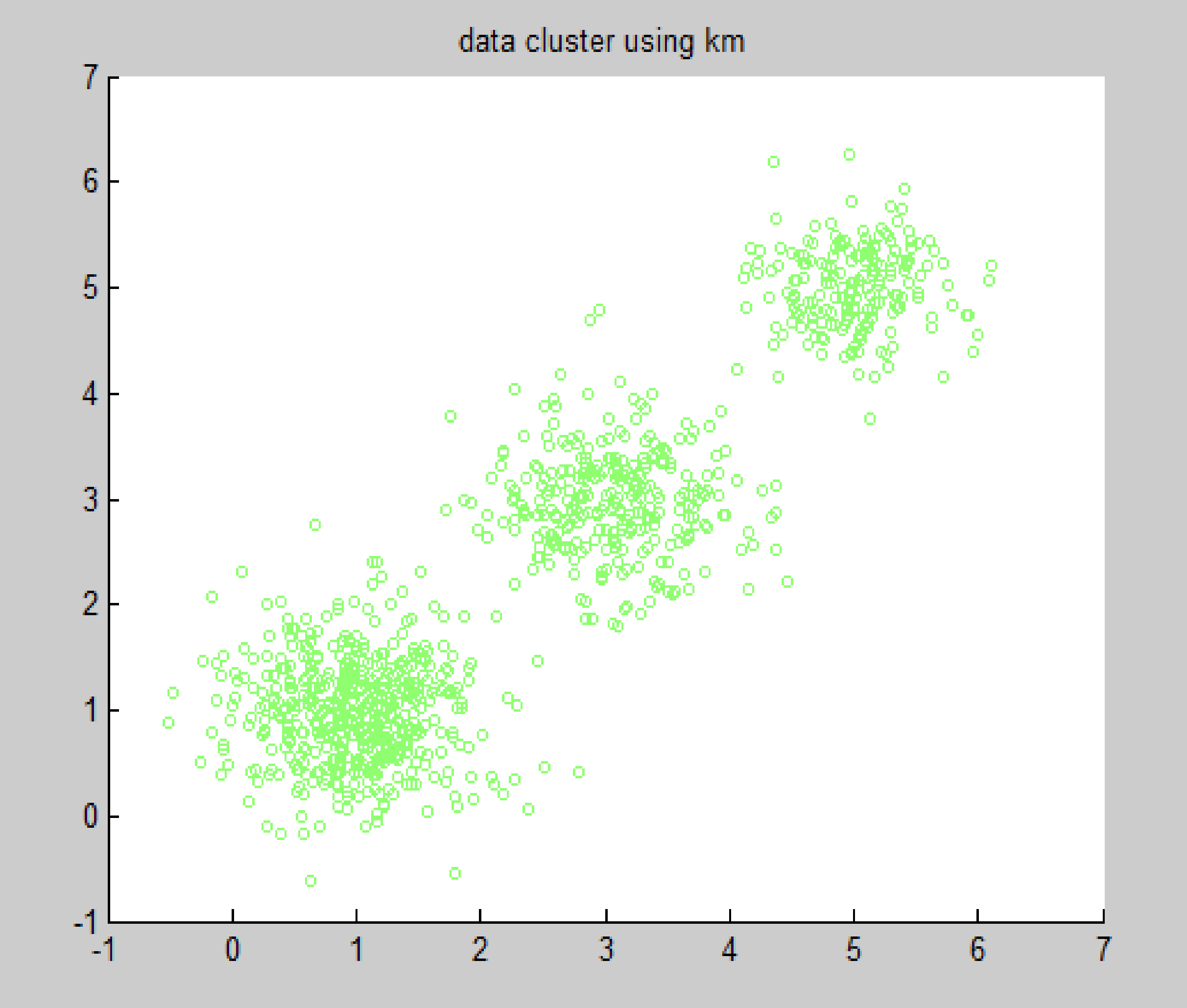


Figure 4.2: Scattering of data

As shown in figure 4.2, the dataset which is loaded will be scattered and plotted on the 2D plane.

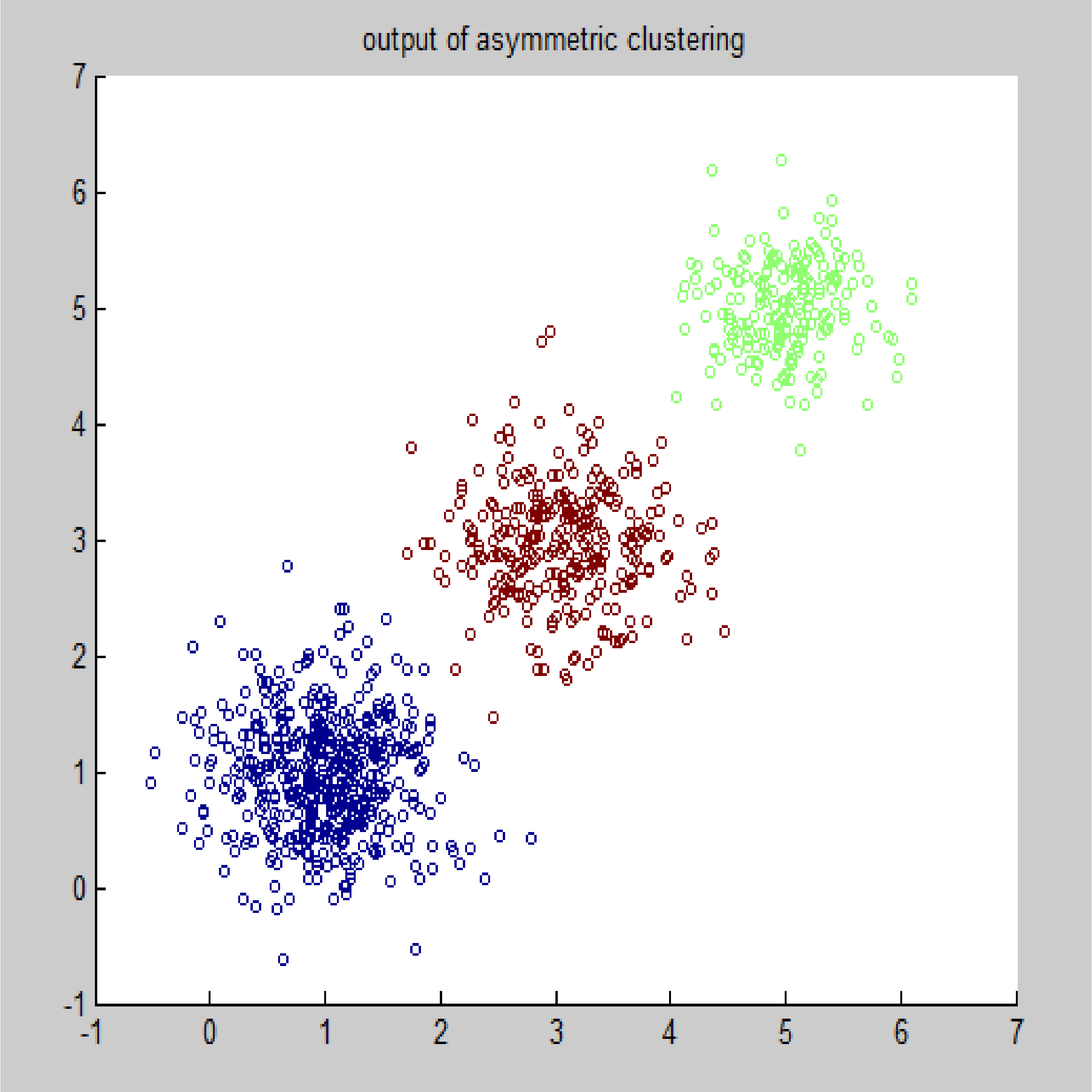
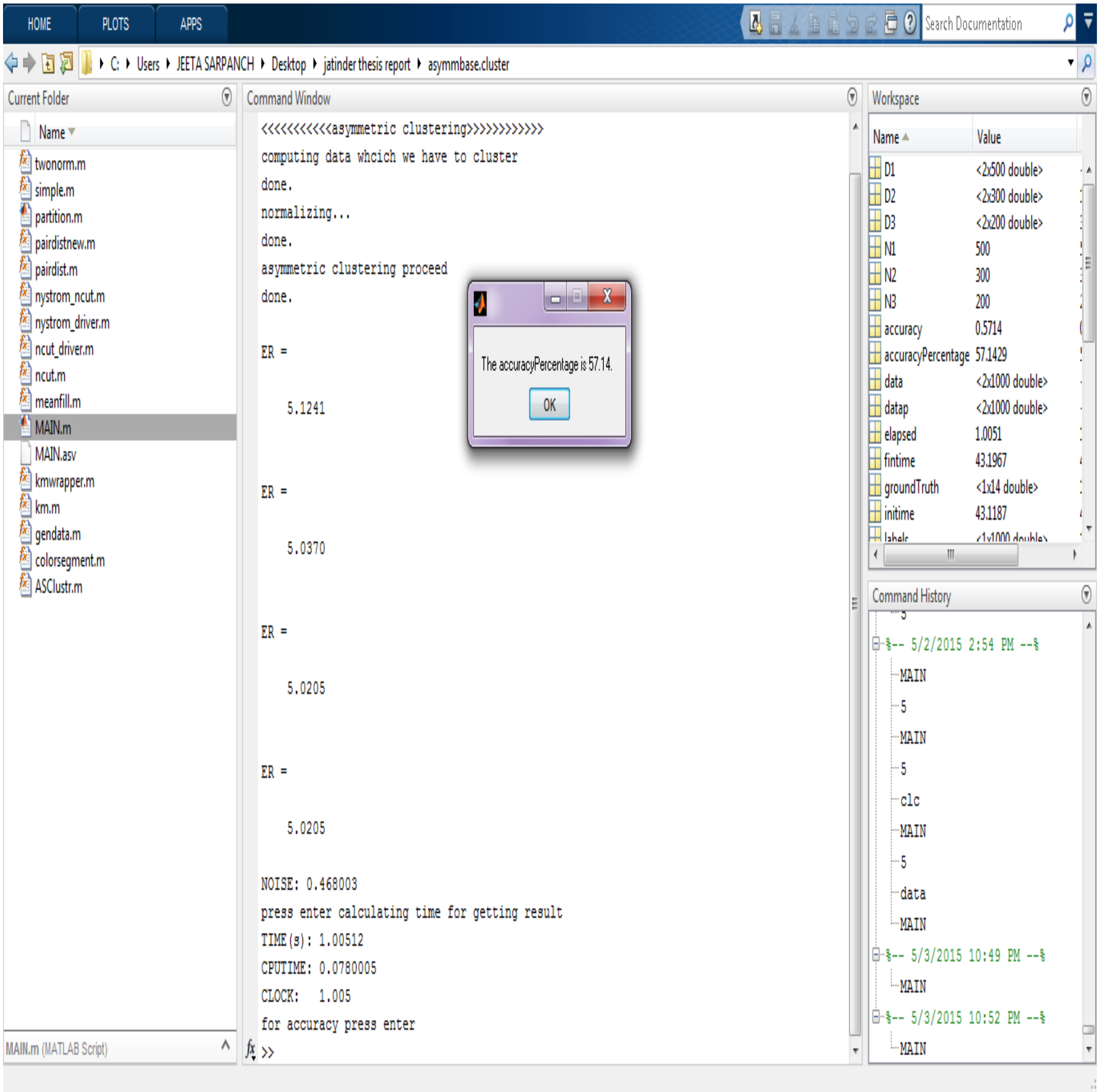


Fig 4.3: Clustering of data

As illustrated in the figure 4.3, the dataset which is loaded had been scattered and scattered data is clustered asymmetric according to asymmetric between the loaded data . When the data points are plotted according to their similarity, normalization technique will be applied on the points. The normalization will assign classes to that points which are not assigned to any of the defined class. The storm function is the function which is applied after k-mean clustering.. In the functions random pick function will chose any random point from the dataset and apply mapping with pairing and exponential function to choose asymmetric points. In the term of efficiency of this algorithm noise and execution time is calculated. In performing asymmetric clustering, normalization is applied on the scattered data and the some functions are applied to find clustering results of asymmetric clustering.



```
<<<<<<<<<<asymmetric clustering>>>>>>>>>>>
computing data whcich we have to cluster
done.
normalizing...
done.
asymmetric clustering proceed
done.

ER =

    5.1241


ER =

    5.0370


ER =

    5.0205


ER =

    5.0205

NOISE: 0.468003
press enter calculating time for getting result
TIME(s): 1.00512
CPUTIME: 0.0780005
CLOCK:   1.005
for accuracy press enter
>>
```

The accuracyPercentage is 57.14.

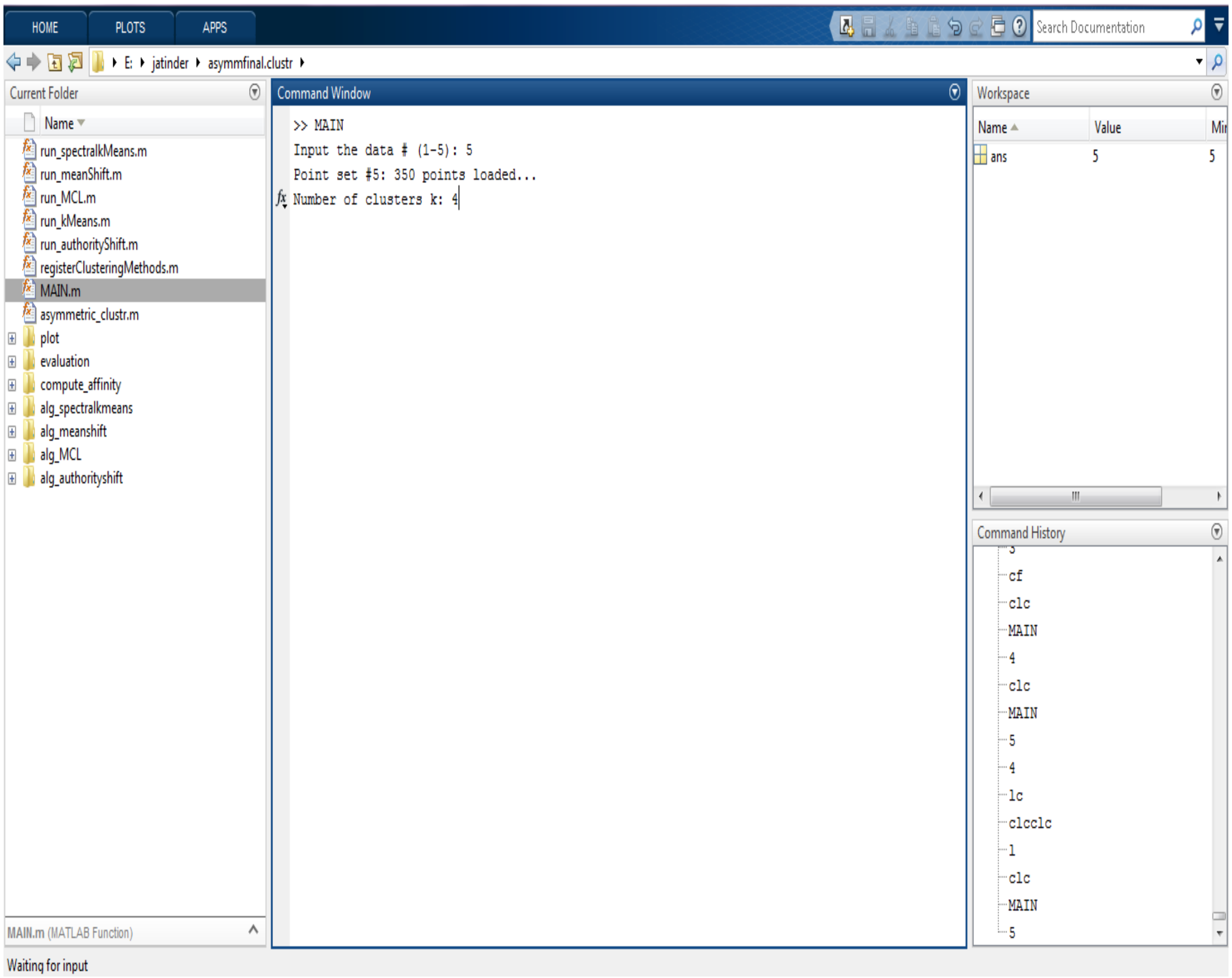
Name	Value
D1	<2x500 double>
D2	<2x300 double>
D3	<2x200 double>
N1	500
N2	300
N3	200
accuracy	0.5714
accuracyPercentage	57.1429
data	<2x1000 double>
datap	<2x1000 double>
elapsed	1.0051
fintime	43.1967
groundTruth	<1x14 double>
initime	43.1187
labels	<1x1000 double>

```
  5
%-- 5/2/2015 2:54 PM --%
  MAIN
  5
  MAIN
  5
  clc
  MAIN
  5
  data
  MAIN
%-- 5/3/2015 10:49 PM --%
  MAIN
%-- 5/3/2015 10:52 PM --%
  MAIN
```

Figure 4.4: Calculating accuracy

The figure 4.4 shows the efficiency of algorithm which is calculated after applying asymmetric clustering.

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. K-means and normalization is used to make clusters.



```
>> MAIN
Input the data # (1-5): 5
Point set #5: 350 points loaded...
Number of clusters k: 4
```

Figure 4.5: Selection of number of Clusters

In figure 4.5, the number of clusters (figure 4.7) are required are entered as input by the user.

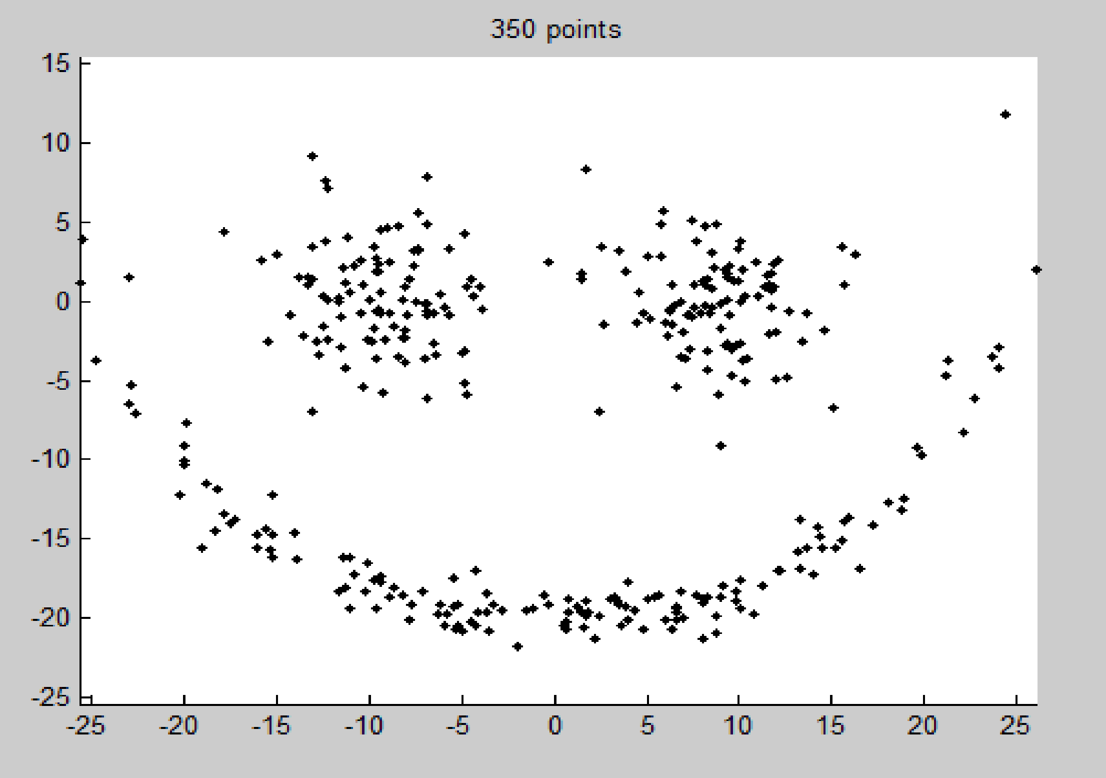


Figure 4.6: Loading of data

As illustrated in figure 4.6, 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

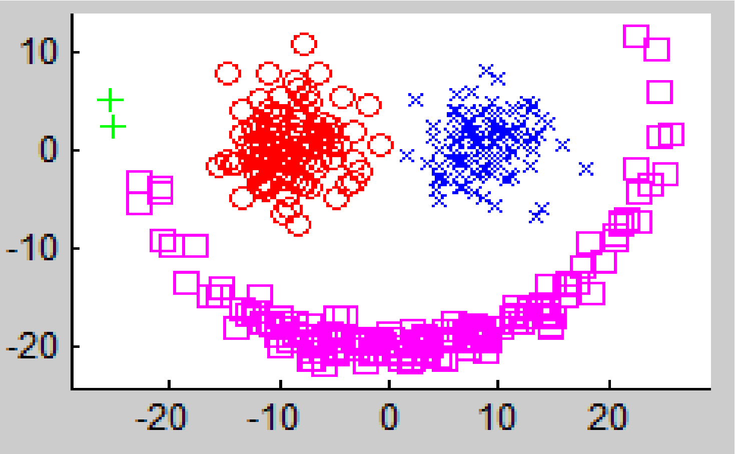
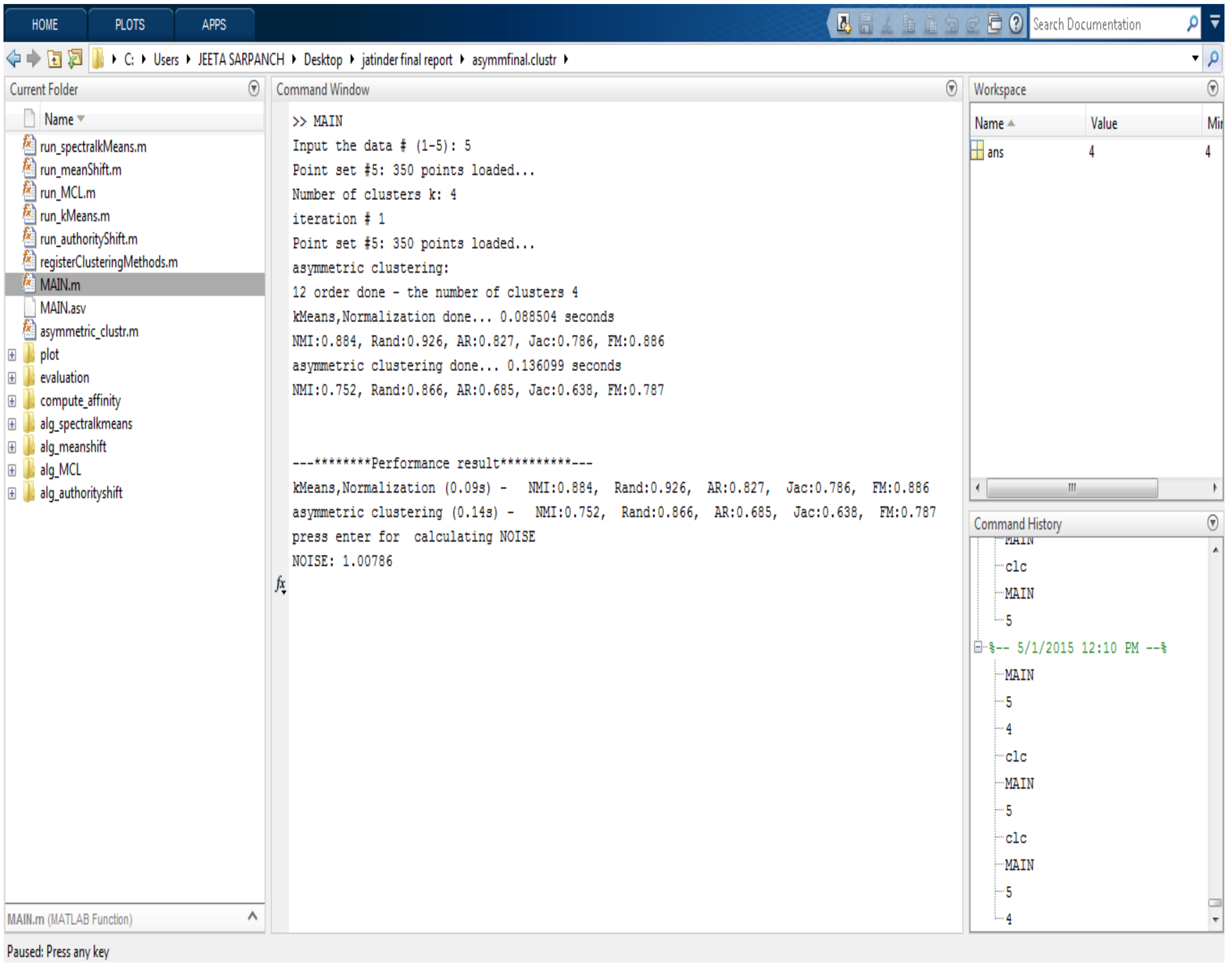


Figure 4.7: Clustered data

In figure 4.7 shows the number of cluster which we want to make. Suppose if want to make four clusters than whole data which is initialized is divided into four clusters which are identified by different color for each cluster. Clusters are made on the basis of k-means and normalization functions.



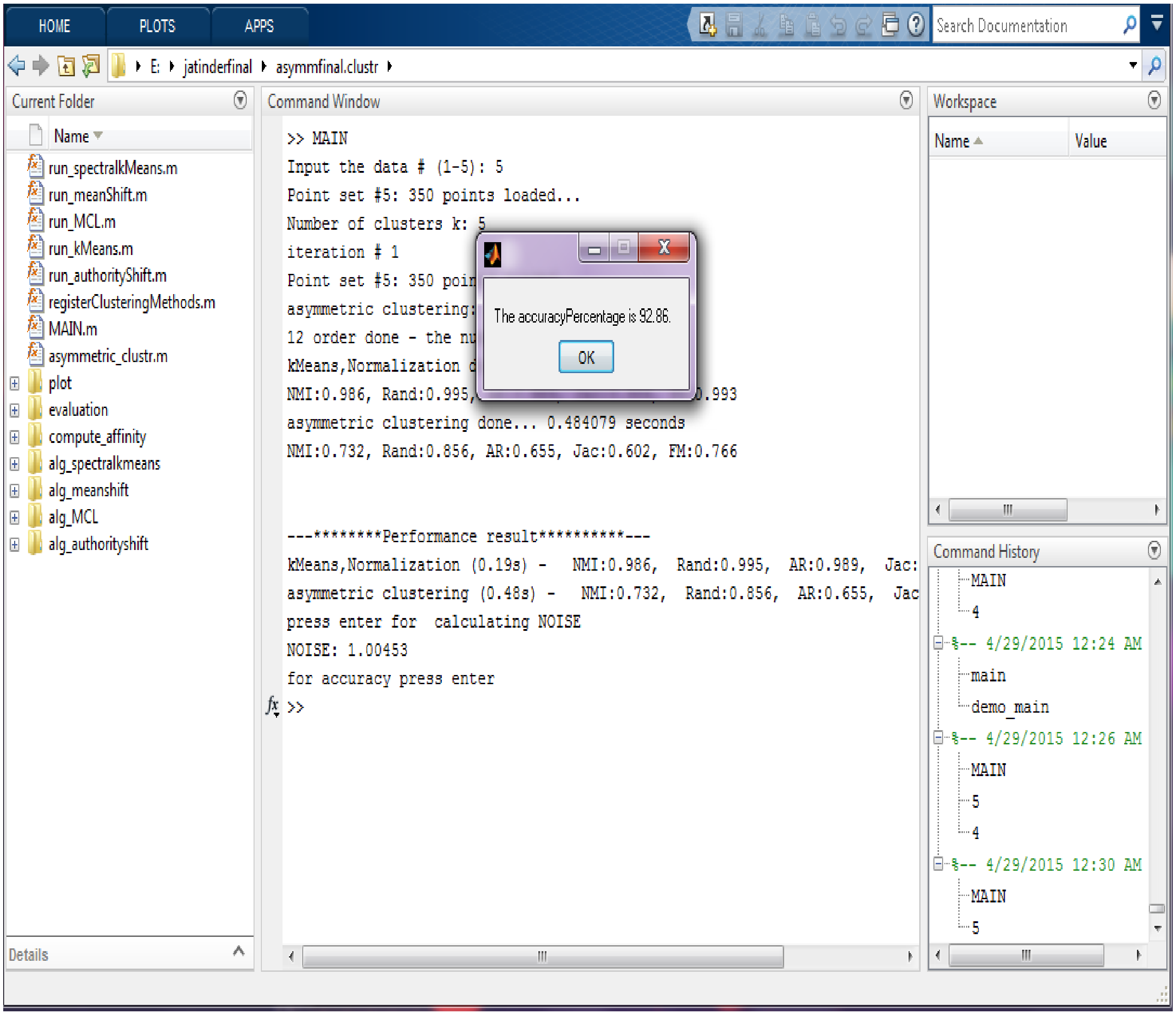
```
>> MAIN
Input the data # (1-5): 5
Point set #5: 350 points loaded...
Number of clusters k: 4
iteration # 1
Point set #5: 350 points loaded...
asymmetric clustering:
12 order done - the number of clusters 4
kMeans,Normalization done... 0.088504 seconds
NMI:0.884, Rand:0.926, AR:0.827, Jac:0.786, FM:0.886
asymmetric clustering done... 0.136099 seconds
NMI:0.752, Rand:0.866, AR:0.685, Jac:0.638, FM:0.787


---********Performance result**********---
kMeans,Normalization (0.09s) -   NMI:0.884,  Rand:0.926,  AR:0.827,  Jac:0.786,  FM:0.886
asymmetric clustering (0.14s) -   NMI:0.752,  Rand:0.866,  AR:0.685,  Jac:0.638,  FM:0.787
press enter for  calculating NOISE
NOISE: 1.00786
```

```
MAIN
clc
MAIN
5
%-- 5/1/2015 12:10 PM --%
MAIN
5
4
clc
MAIN
5
clc
MAIN
5
4
```

Figure 4.8: Calculation of Noise

Above figure 4.8 shows that as already described input the data and after inputting the data points of dataset and the number of clusters are defined which are to be made according to the k-means and normalization techniques. Asymmetric clustering in applied to these clusters to make these clusters more distinct and clear. 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.



```
>> MAIN
Input the data # (1-5): 5
Point set #5: 350 points loaded...
Number of clusters k: 5
iteration # 1
Point set #5: 350 poi
asymmetric clustering:
12 order done - the nu
kMeans,Normalization d
NMI:0.986, Rand:0.995,                    .993
asymmetric clustering done... 0.484079 seconds
NMI:0.732, Rand:0.856, AR:0.655, Jac:0.602, FM:0.766


---********Performance result**********---
kMeans,Normalization (0.19s) -   NMI:0.986,  Rand:0.995,  AR:0.989,  Jac:
asymmetric clustering (0.48s) -   NMI:0.732,  Rand:0.856,  AR:0.655,  Jac
press enter for  calculating NOISE
NOISE: 1.00453
for accuracy press enter
>>
```

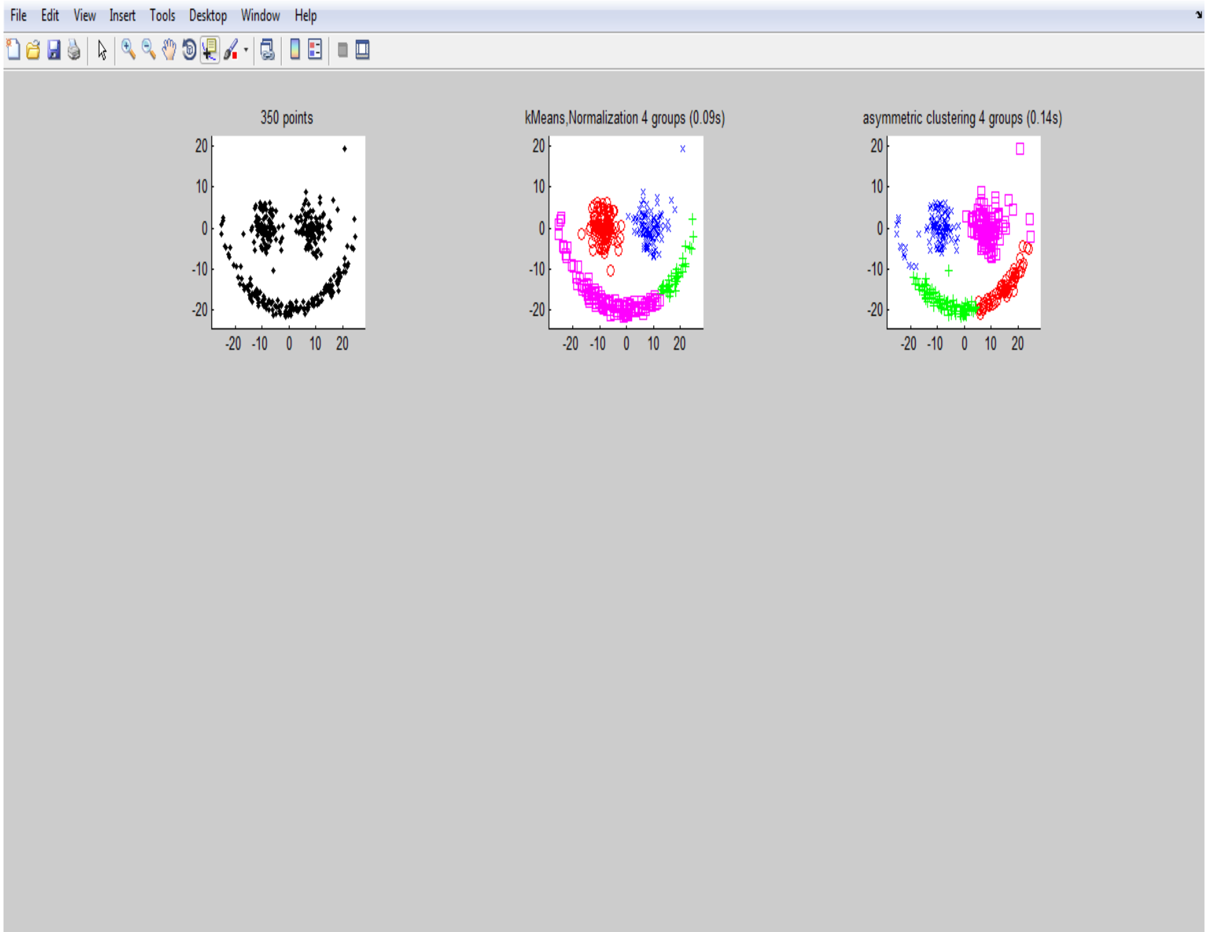
The accuracyPercentage is 92.86.

OK

```
MAIN
4
%-- 4/29/2015 12:24 AM
main
demo_main
%-- 4/29/2015 12:26 AM
MAIN
5
4
%-- 4/29/2015 12:30 AM
MAIN
5
```

Figure 4.9: Calculation of accuracy

This figure 4.9 shows that the efficiency of our enhanced algorithm in terms of accuracy. The MCL is the markov clustering 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



350 points

kMeans,Normalization 4 groups (0.09s)

asymmetric clustering 4 groups (0.14s)

Figure 4.10: Clustering of data

This figure 4.10 shows the final visual output of our enhanced algorithm. 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 asymmetric clustering is applied for data clustering .

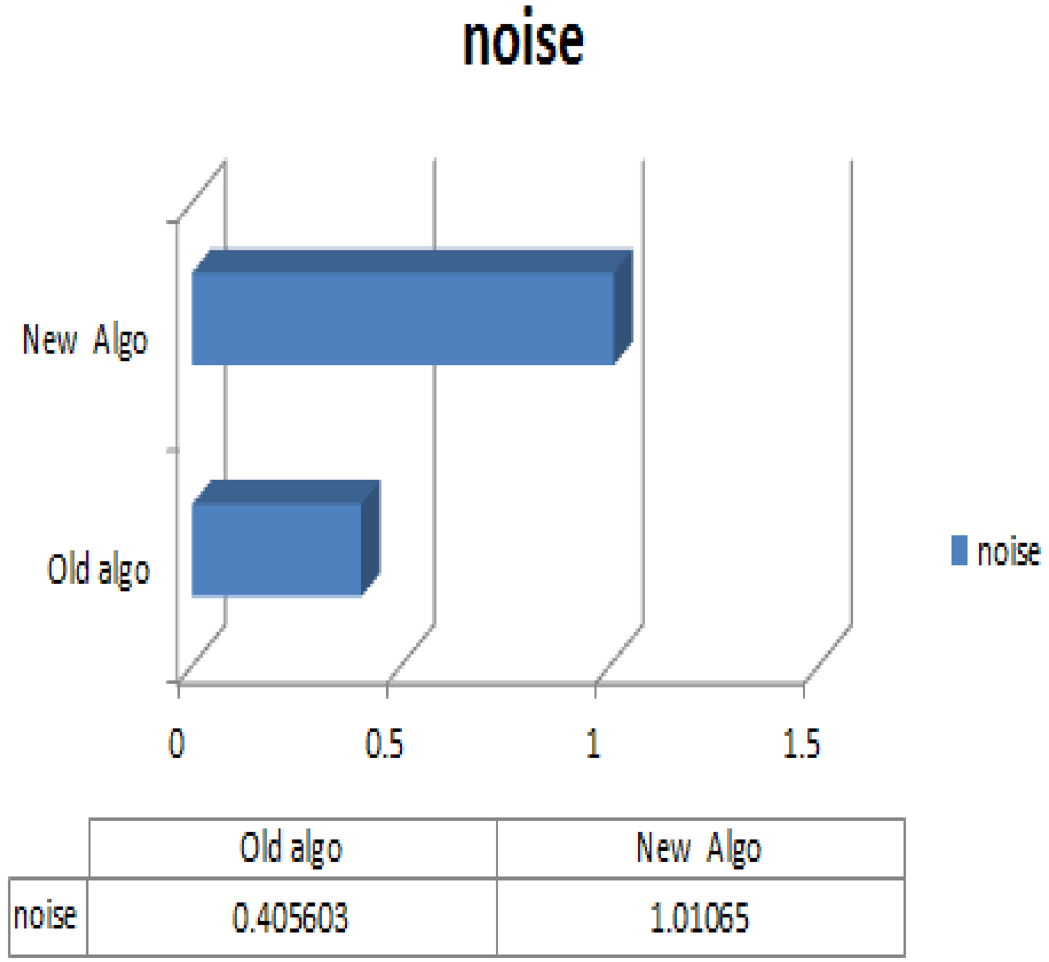
CHAPTER-5

CONCLUSIONS & FUTURE SCOPE

To extract useful or interested information from large set of databases data mining techniques are used. KDD (knowledge discovery from databases) is data mining method to extract information from data warehouses. Association rule is method to place the frequent item sets together to ado analysis like in basket analysis, retail stores and stock market etc. Asymmetric clustering is unsupervised technique of data mining. Clustering is technique in which large datasets are divide in to small datasets in this way that objects and items with having similar properties into one group and objects having dissimilar properties into another.

There are number of algorithms that work well with simple datasets in the term of accuracy and performance but, when these algorithms has to work with mixed and tightly coupled different data sets their performance in the term of accuracy is decreased. Asymmetric algorithms along with techniques such as k-means normalization, mean shift, Markov Cluster Algorithm, S-Cluster techniques are enhanced in the previous asymmetric algorithm to improve the efficiency of asymmetric clustering in terms to reduce noise and improve the accuracy. In the previous technique noise is calculated about 0.405603(fig. 5.1) and time taken for execution is 1.00129 sec. In enhanced version of this asymmetric clustering algorithm calculated noise is 1.01065 and time for execution is .09 sec. There is quite large difference is seen in the terms of accuracy of both previous and enhanced asymmetric clustering. In previous asymmetric clustering accuracy is only 57.14% and in enhanced asymmetric clustering accuracy is seen 92.86%.

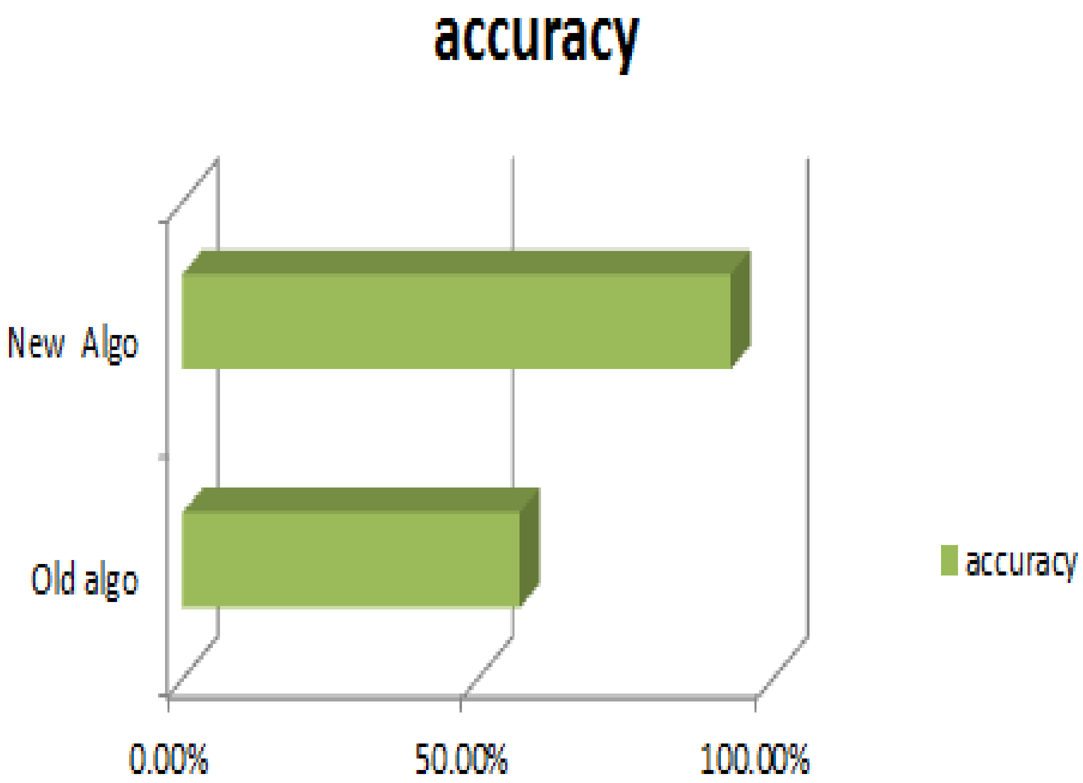
In future, this asymmetric algorithm may use for classification instead of clustering to improve the noise, accuracy and escape time.



	Old algo	New Algo
noise	0.405603	1.01065

Figure 5.1 Noise Comparison

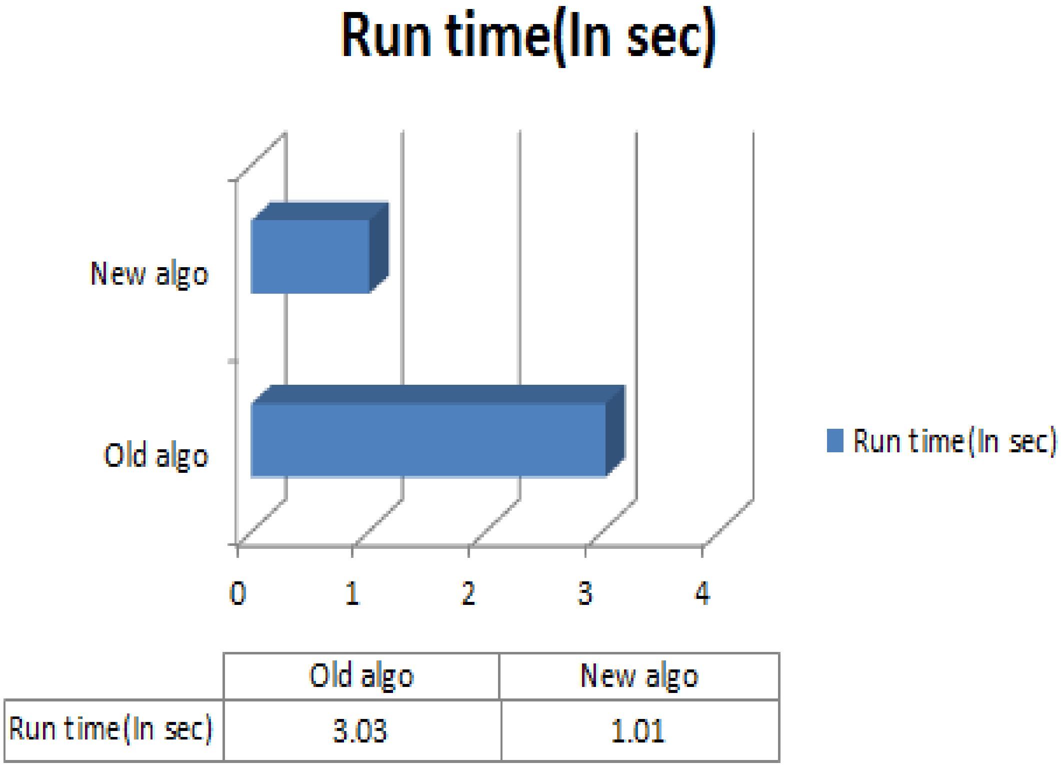
Noise comparison of previous and enhanced asymmetric algorithm is made. In figure 5.1, it is found that in the enhanced asymmetric algorithm more noise is calculated than previous asymmetric algorithm.



	Old algo	New Algo
accuracy	57.14%	92.86%

Figure 5.2 Accuracy Comparison

Accuracy comparison of previous and enhanced asymmetric algorithm is made. In figure 5.2, it is found that in the enhanced asymmetric algorithm more noise is calculated than previous asymmetric algorithm.



	Old algo	New algo
Run time(In sec)	3.03	1.01

Figure 5.3 Run time comparison

Run time comparison of previous and enhanced asymmetric algorithm is made. In figure 5.3, it is found that in the enhanced asymmetric algorithm more noise is calculated than previous asymmetric algorithm

REFERENCES

[1] Cloud Computing Principlesand Paradigms,Edited by, Rajkumar Buyya, James Broberg, Andrzej Goscinski, by John Wiley & Sons,Inc publications 2011.

[2] Hao Huang† Yunjun Gao‡,_ Kevin Chiew§ Lei Chen# Qinming He " Towards Effective and Efficient Mining of Arbitrary Shaped Clusters" *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China,* ICDE Conference 2014

[3] Gunnar Carlsson et.al , "Hierarchical Quasi-Clustering Methods for Asymmetric Networks",*Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR:W&CP volume 32, 2014*

[4]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*

[5] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav (2012*) "Mining the customer behavior using web usage mining In e-commerce"* Coimbatore, India. IEEE-201S0

[11]Neelamadhab Padhy , Dr. Pragnyaban Mishra and and Rasmita Panigrahi "*The Survey of Data Mining Applications And Feature Scope*"International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[6] Satoshi Takumi and Sadaaki Miyamoto,"*Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering*", 2012 International Conference on granular Computing

[7] S.R.Pande, Ms..S.S.Sambare, V.M.Thakre,"*Data Clustering Using Data Mining Techniqes*", IJARCCE Vol. 1, issue 8, October 2012

[8] 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*

[9] Wilhelmiina Hamalainen, Matti Nykanen (2008) "*Efficient discovery of statistically significant association rules*", Eighth IEEE International Conference on Data Mining.

[10]Jiawei Han J and Kamber M, Data Mining: Concepts and Techniques (3rd ed.). *Morgan Kaufmann, San Francisco*, CA, 2012.

[11] **Hui Xiong, Gaurav Pandey, Michel and Vipun, "Enhancing Data Analysis with Noise Removal",** IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 13, 2013

[12] Yu Qian and Kang Zhang, "The Role of Visualization in Effective DataCleaning", SAC'05,March 13-17,2005,Santa Fe,New Mexico,USA

[13] Sumit Garg and Arvind K. Sharma, **"**Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications (0975 –8887) Volume 74–No.5 , July 2013

[14] K.Krishna and Raghu, "A clustering algorithm for asymmetrically related data with applications to text mining", ACM, New York, USA, 2001

[15] Batagelj,V., Mrvar, A.,andZaversnik,M., ",Partitioning approaches toclusteringin graphs, Pr Drawing'1999, LNCS, 2000, pp. 90-97.

[16] Ertoz, L., Steinbach, M., and Kumar, V., "Finding clusters of different sizes, shapes, and densitie dimensional data", In Proc. of SIAM DM'03.

[17] Ester, M., Kriegel, H.P., Sander,J., and Xu, X., " A density-based algorithm for discovering clusters databases with noise", in Proc. of 2nd Int. Conf. on Knowledge Discovery and DataMining(KDD-96),AAAI Press, 1996, pp. 226-231.

[18] Fayyad, U., Piatetsky-Shapiro,G.,Smyth,P., and Uthurusamy,R. (eds.), "A and Data Mining, AAAI/MIT press, 1996.

[19] Fayyad, U. and Grinstein,G.,Information Visualization in Data Mining and Knowledge Discovery, M 2001, pp. 182-190.

[20] Fayyad,U. and Uthurusamy,R., "Evolving data mining intosolutions for insight pp. 28-31.

[21] Han, J., Kamber, M., and Tung, A. K. H., "Spatial clustering methods in (eds.), Geographic Data Mining and Knowledge Discovery, TaylorandFrancis, 2001.

[22] Harel, D.andKoren, Y., “Clustering spatial data using random walks”, In Proc. 7th and Data Mining(KDD-2001),ACM Press, New York, pp. 281-286

[23] K.Rajkumar “*Dynamic Web Page Segmentation Based on Detecting Reappearance and Layout of Tag Patterns for Small Screen Devices”,IJSET,2011*

[24] Shuang Lin, Jie Chen, Zhendong Niu, “*Combining a Segmentation-Like Approach And A Density-Based Approach In Content Extraction”* TSINGHUA SCIENCE AND Technologyissnll1007-0214ll05/18llpp256-264 Volume 17, 2012

[25] Yan Gu , “*ECON: An Approach to Extract Content from Web News Page”* 12th International Asia-Pacific Web Conference,2010

[26] Chaw Su Win, Mie Mie Su Thwin , “ *Informative Content Extraction By Using Eifce”* International Journal Of Scientific & Technology Research Volume 2, Issue 6, 2013

[27] Jan Zeleny, “Web Page Segmentation And Classification” Journal of Data and Knowledge Engineering, 2010

[28] K.S.Kuppusamy, “*A Model for Web Page Usage Mining Based on Segmentation”* International Journal of Computer Science and Information Technologies, Vol. 2 issue 3, 2011