MULTIPLE VIEW CLUSTERING IN PROCESS MINING

A Dissertation Proposal

Submitted By

Renu Singh

TO

Department of Computer Science

In partial fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science

Under the guidance of

ADITYA BAKSHI

Asst.Professor

CERTIFICATE

This is to certify that **Renu Singh** has completed M.Tech dissertation titled **Multiple View Clustering in Process Mining** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has been submitted for any other degree or diploma

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

Date

Signature of Advisor

Aditya Bakshi

DECLARATION

I hereby declare that the dissertation entitled **Multiple View Clustering in Process** mining submitted for M. Tech Degree is entirely my original work and all the ideas and references have been duly acknowledged. It does not contain any work for the award of any degree or diploma.

Date

Investigator:Renu Singh

Reg No.11500903

ACKNOWLEDGEMENT

I would like to present my deepest gratitude to **Asst. Prof Aditya Bakshi** for his guidance, advice, understanding and supervision throughout the development of this dissertation study. I would like to thank to the **Project Approval Committee** members for their valuable comments and discussions. I would also like to thank to **Lovely Professional University** for the support on academic studies and letting me involve in this study.

TABLE OF CONTENT

TABLE OF FIGURES

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.

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 [1] 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 [2].

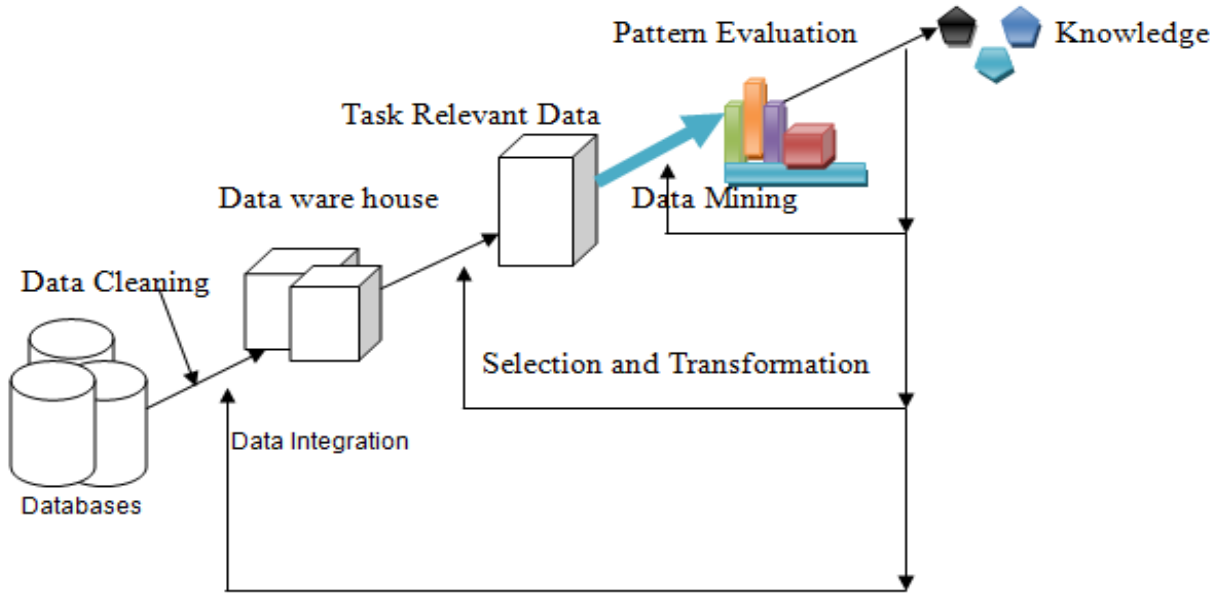


Figure 1.1: Data Mining Process

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 [3]:

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 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 [4].

1.2 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.

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

1. Association

2. Classification

3. Clustering

4. Prediction

5. Sequential Patterns

6. Decision Tree

1. Association: The most prominently utilized data mining technique is the association. On the basis of the relationships amongst various objects within the similar transaction, certain pattern is recognized between them. Thus it is also called the relation method. There are various sets of products that are to be identified for purchasing. This is done with the help of association method. As per the buying histories of various products by the customers, the retailers gather analysis and then provide sales to them as per the historical data [5].

ARM used to find out the interesting correlation between the items. After ARM many algorithm are generated that is apriori algorithm and improvement in 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 slow to show the result in reasonable time.

Association rule mining having two main steps:

- **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 satisfy 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.

2. Classification:

A data mining technique which relies on the machine learning is known as classification technique. For the purpose of classifying each object into the classes that are predefined basic classification is performed. The various mathematical methods such as decision trees, neural network, linear programming and statistics are utilized by the classification method. The data items can be classified into various groups by building software within the classification method. For instance, the records of the employees who left the company within certain duration can be classified and the number of employees who will leave the company in future can be predicted on the basis of it. The data of employees can further be divided into two distinct groups namely “leave” and “stay”. The data mining software helps in classifying them. On the basis of the input provided the results can be seen after the output is achieved [6].

3. Clustering: Clustering is a technique in objects of similar category is placed in one group and other are in different group using automatic technique. The defining of classes and placing similar objects within same groups to provide relation amongst them is known as clustering. The placing of objects is existing predefined classes as per their definition is known as classification.

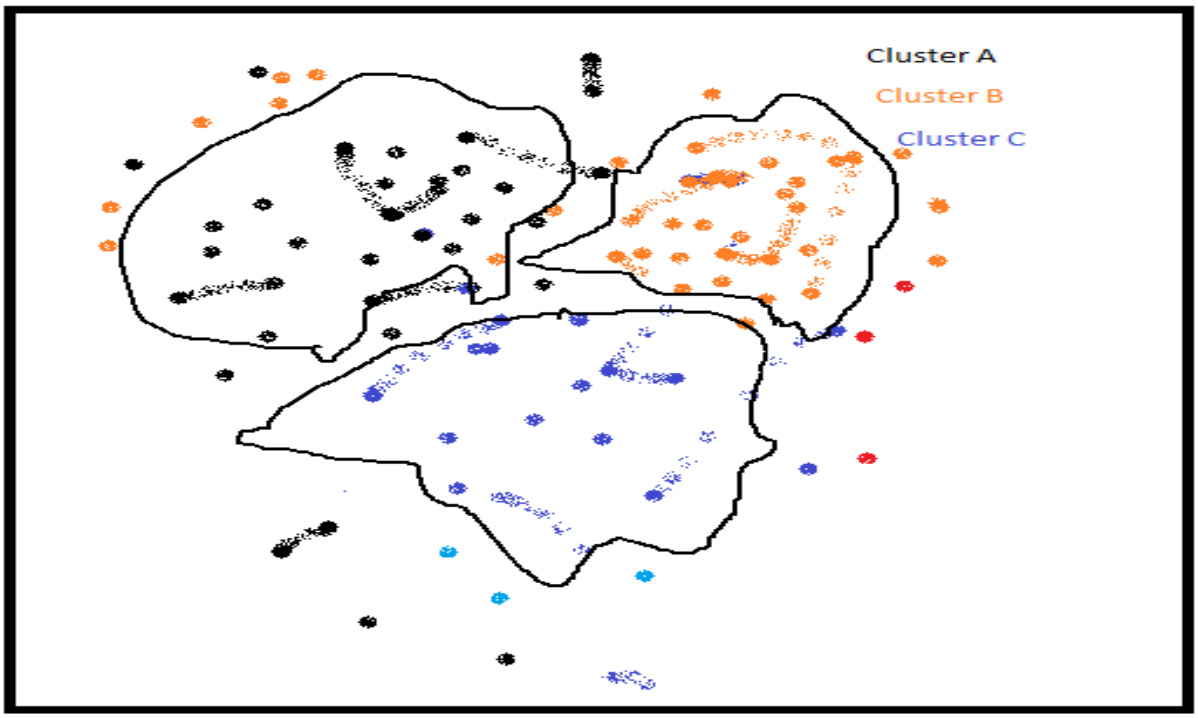
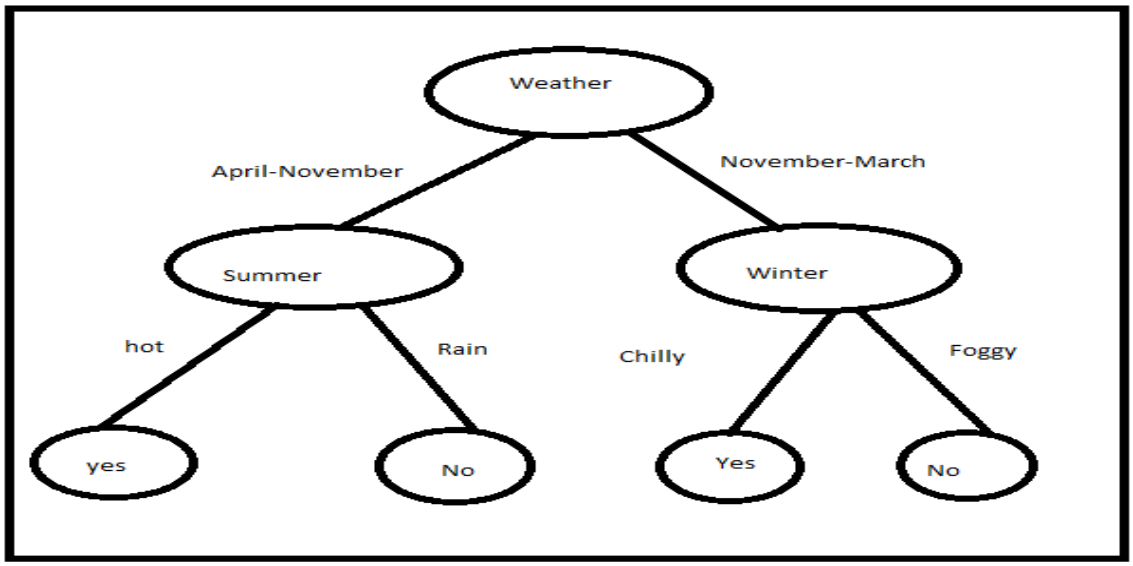


Fig.1.2 Clustering in Data Mining

4. Prediction: Providing association amongst the independent variables is done with the help of prediction method. Through this technique the relation amongst the dependent and independent variables is also provided. For example, on the basis of the records of a company the profits are predicted. Both the dependent as well as the independent variables are utilized for executing this technique. A fit regression curve can be plotted with the help of the historical sale and profit data achieved after analyzing the historic data. This can help in predicting the future of the respective organization.

5. Sequential Pattern: The identification of similar patterns within the transaction data within certain duration is done with the help of sequential pattern analysis of the data. Through a proper analysis of the historical data gathered over a period of time it can be analyzed that which products are bought by the customers. The recommendations can further be made on the basis of such analysis. Various deals are proposed for increasing the purchasing of certain products which have not been sold much previously [7].

6. Decision Tree: One of the most utilized data mining techniques is the decision tree which is very easy to be understood by the various users. There is a single question or condition present at the root node within the decision tree which holds various answers to it. The final decisions can only be made on the basis of the question or conditions which are presented within the decision tree. It is very similar to that of the hierarchical technique.



1.3.1 Clustering in Data Mining: The process which separates objects with similar properties within one group and objects with dissimilar properties into other is known as clustering process. For instance, the objects can be placed in separate groups on the basis of their threshold values. The objects with higher threshold values can be placed within same groups and the other with lower values into another. Thus, the separation of objects on the basis of their similar properties is known as clustering [8].

There are no classifiers and labels present within this method and so it is also known as an unsupervised learning technique. In various fields such as image processing, data analysis, and so on, the cluster analysis method can be utilizes. The outliers are the values that are placed outside the clusters. These outliers can be detected with the help of clustering technique as well.

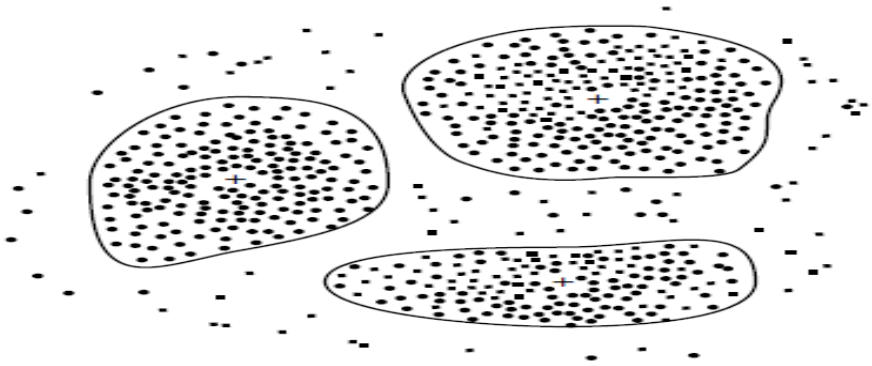


Figure 1.3 Clusters and Outliers

As shown in the figure 1.4, the dots placed outside the clusters are the outliers. Also the objects that have similar properties are placed within the clusters.

There is a major difference that exists between the clustering and the nearest neighbour technique. The clustering method belongs to the unsupervised learning techniques category. However, the nearest neighbour can be categorized within the prediction or the supervised learning techniques. There are no supervisions available within the unsupervised techniques whereas, the supervisors such as teachers are available within the supervised learning methods. The predictions that are identified within the database and are depicted in the model within the prediction method are the most prominent patterns which can be utilized for predictions within the database. There is no clarity of the idea for which the few records are closer to each other or are placed within the similar cluster in the clustering method [9].

1.4 Types of Clustering: There are different types of clustering in data mining. These are as follow:

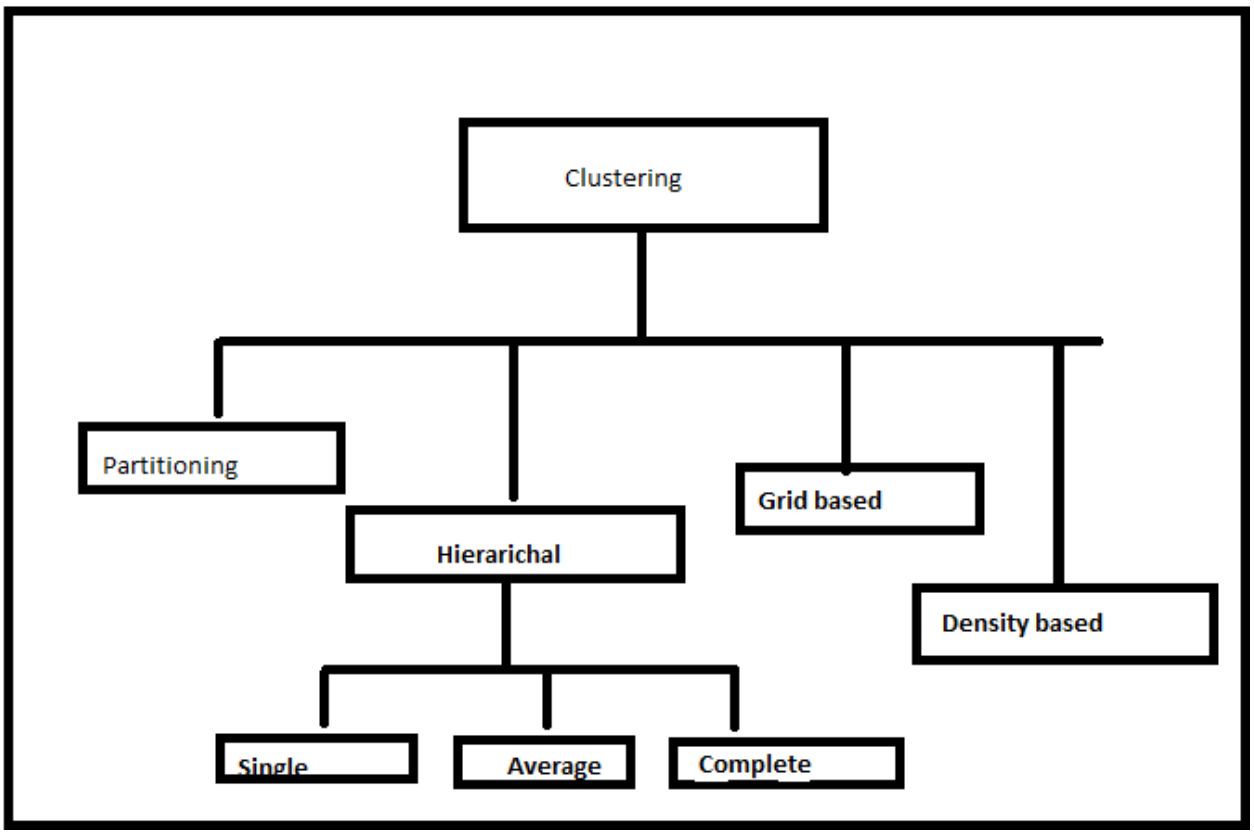


Fig.1.4 Clustering in Data Mining

1.4.1 Description of clustering's as follow:

Type of Clustering	Description
Partitioning Clustering	The combination of data that is highly similar within the clusters and separating it from the data which is highly dissimilar is known as partitioning clustering technique. The data is placed within distinct clusters on the basis of their properties. The distance-based techniques are the most partitioning based techniques. Let, k is the number of partitions that are to be provided for the data. An initial partitioning is generated by the partitioning method. Further, an iterative relocation method is utilized which enhances the partitioning method by moving the

	objects from one group to the other. The objects that are placed within the same clusters have stronger relationships within a good partitioning method. However, the objects that are placed in various clusters are distant from each other. The greedy methods such as k-means and k-medoid algorithms are some of the widely utilized heuristic methods which can be used within the applications. The clustering quality of the technique can be enhanced with the help of these techniques and further the local optimum can be achieved. The spherical-shaped clusters are identified within small to medium sized databases within the clustering techniques.
Density Based Clustering	On the basis of the distance between the objects, various partitioning methods are proposed. With the help of these methods, the spherical shaped clusters are recognized. The clusters of the arbitrary shapes are difficult to be discovered and thus the spherical shaped clusters can be recognized with the help of these methods. Thus, the new methods known as the density-based methods are proposed here which are based on the density factor of the clusters. The growths of the clusters keep growing as the density of the neighbourhood increases as per the threshold [10].
Hierarchal Clustering	Within this method the set of objects are hierarchically decomposed as being either agglomerative or divisive on the basis of the decomposition formed. Each object forms a different group within the agglomerative approach which is a bottom-up approach. The groups that are closer to one another are merged until only one group is left. The top-down approach known as divisive approach begins from the objects that lie within same cluster. A cluster is then split into smaller clusters after each iteration step. This process continues until each object is present

	within a separate cluster.
Grid Based Clustering	The quantization of object space into finite number of cells such that a grid structure is formed is known as a grid based technique. There is no dependency of the number of data objects present within this method and has high speed for execution. The number of cells that are present within each dimension of the quantized space is the factor on which this method relies.

1.5 Asymmetric Clustering: There is a vigorous partition within the asymmetric clustering technique. At certain duration, a single cluster is to be processed within this method. The task is identified as per its properties and is placed within the appropriate cluster. There are various applications that utilize the asymmetric clustering methods such as the electronic trading systems present within the banks include such techniques. 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 result, in each node where the cluster implements its own local cache results in high performance reading and writing without the need to maintain a distributed cache between cluster nodes. There are various differences between the symmetric and asymmetric clusters. Within the asymmetric clusters [11]:

- During the execution, the named partitions can be declared by the various applications at any instant.
- The names of various partitions are given in a unique manner as singletons. They are further made to run on a single cluster member.
- The cluster member that hosts the partition is made to route on the input of the partition.

There is a specific lifecycle followed by the partitions within this technique. There can be various threads or alarms set at the background and the incoming events can also be responded here. The asymmetric clustering as well as partitioning enhances the performance of the systems

by providing low level primitives in them. A custom application can be built with the help of these primitives which further provides the features of application server. For the application developer, many high level services are provided by the J2EE specification. If the respective service does not match the requirements of the user, no alternate services are provided to replace them. So, the choice here is to utilize the J2SE which provides the facility of adding new features from the commercial application server's next release.

1.5.1 Clustering in Data Mining

Here are various applications that have been utilizing the cluster analysis technique for providing proper classification of data. Some of these techniques are data analysis, image processing, and pattern recognition and so on. On the basis of the patterns, the interests of the customers are discovered by the marketers within the business applications. On the basis of the derived patterns, there are various groups characterized amongst them. Various plant and animal taxonomies are derived within the biology applications, and the genes are further categorized on the basis of similar functionalities. There are various structures which are different for each of the populations and the structures derived are compared with these standard structures to know which structure is inherited by which population. Within the geology applications, the identification of similar lands, houses etc. can be done with the help of clustering technique. Further, the classification of documents on the internet for discovering the information is done with the help of data clustering technique [12].

The process of creating groups of objects or clusters by ensuring that the objects within same clusters are similar and the objects within different clusters are dissimilar is known as data clustering method. The method is an unsupervised type of classification method. The initial step within the direction of providing knowledge discovery is the cluster analysis process. Clustering is the process grouping data objects into a set of disjoint classes. After the clustering process the similar objects are grouped together and the other dissimilar objects are kept separately.

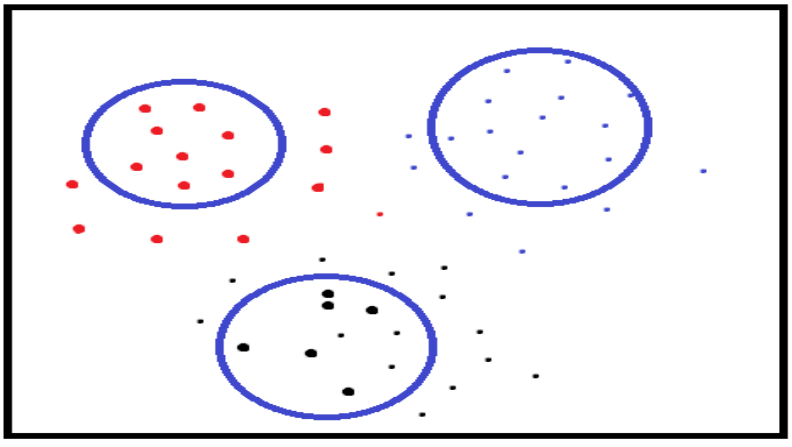


Fig.1.5 Output of Clustering

For the purpose of grouping similar objects into one cluster and providing dissimilar objects into another, the clustering technique has been derived. Instead of utilizing the predefine objects, the documents are clustered in a different manner within the classification process which proves it is different from clustering process. When the documents can be seen within various subtopics, the clustering can be useful in not misplacing the indexed lists which proves to be another advantage for using clustering. The vector of topics for each document is designed with the help of fundamental clustering algorithm and the weights are measured to determine whether the cluster fits according to the requirement or not [12].

Within the clustering process, the unsupervised classification process is followed. The assigning of data objects to a set of classes is known as classification process. There is no dependence of the clustering process on the predefined classes and training which means it follows the unsupervised clustering process.

Unsupervised clustering is not the same as pattern reorganization in the area of statistics known as discriminate analysis and decision analysis which arrange the objects from a given set of object.

There are numerous clustering algorithms utilized for clustering. The major fundamental clustering methods can be classified into taking after categories.

1. **Partitioning Methods:** The combination of objects that have higher similarity amongst the clusters and separating it from dissimilar objects by keeping it in separate clusters is known as the partitioning method. The distance-based method is utilized for almost all of the partitioning methods. Let, k is the quantity of number of partitions to be built. On the basis of these partitions, the initial partitioning strategy is built. Further the iterative relocation system is utilized which provides enhancement in partitioning the objects through the mobility process. The objects are moved as per their similarities within various groups one by one. The objects within the similar cluster are closer to each other as compared to the objects that are present within the separate clusters. There are various heuristic methods such as the greedy methods which include k-means and k-medoid algorithms utilized by the applications. The clustering quality of the applications is

enhanced and a local optimum approach is achieved. The spherical-shaped clusters are recognized with the help of clustering methods as per the sizes. Here is a partitioning of the data set which contains n objects into a set of k clusters. His helps in minimizing the Ɵ criterion. He main objective here is to partition the k clusters provided such the partitioning criterion is optimized. The input parameter is known as k [13].

Fig. 1.6 Partitioning Clustering [13]

2. **Hierarchical Methods:** The hierarchical decomposition of a given set of data objects is done. He classification of data can be done wither in the agglomerative or in divisive based hierarchical decomposition. A different group is generated by each object in the agglomerative type of approach. Further, the groups that are closely similar to each other are merged on the basis of properties. This is opposite to the divisive approach that follows the top down mechanism. Here, all the objects are placed in one cluster and then after each iteration step the clusters are split on the basis of the similarities found among the objects. This results in forming smaller clusters and at the last each object is found to be placed in different cluster. For each of the provided data set of data objects, the hierarchical algorithms create a hierarchical decomposition which is represented by a tree structure also known as a dendrogram. The clusters that are provided as inputs are not disturbed here. At different levels of granularities various types of clustering methods can be achieved where various typed of k are used.

3. **Density Based Methods:** On the basis of various data objects, the partitioning methods help in clustering the objects. With the help of various methods, spherical shaped clusters are recognized. However, it is difficult to recognize the clusters that have arbitrary shapes. Further, the density based methods are utilized for recognizing such arbitrary shapes on the basis of notion of density. Along with the increment in the threshold of density within an area, the clusters are formed with the help of these methods. As long as the density of the area increases certain threshold value, the provided cluster keeps growing. For each data point provided within a cluster, there are certain numbers of points that are to be made sure are achieved. This helps in providing arbitrary shape within the clusters. Any issues arising within the data are also to be handled. Only one time scan is done here and there is a need of additional density parameters within this process [14].
4. **Grid Based Methods:** The quantization of object space into finite number of cells such that they form a grid type of structure is known as a grid based method. There is no independence of the method on the number of data objects and it only depends on the number of cells provided within each dimension within the quantized space. Ill the shape grid is achieved, the objects are met. A grid structure is formed further by quantizing the object space into finite number of cells. It assigns to the object grids cells and process density of every cell. After that wipe out whose density is beneath threshold t. Presently shape cluster as per group of dense clusters. In this no distance computations so it is fast process. In this it is likewise simple to figure out which cluster is neighboring. Here shapes are restricted to the union. On the basis of the groups formed by the cells, the quality of the various sided clustering is dependent. He space is quantized with the help of grid based algorithms into finite number of grids. On the quantized space available, the operations are made to run. On the basis of number of segments present within each dimensions of the quantized shape, the dataset size is made to be independent due to fast processing merits of the discussed approaches.

1.6 DBSCAN

There is a huge applicability of density based clustering algorithms within the data mining techniques. To certain group of objects the local criteria is applied where the clusters are seen as

a region within the data space. Here within this data space, the objects are placed more densely and are separated by the regions which have low object density [14].

Among the density based clustering algorithms DBSCAN is exceptionally well known due both to its low complexity and its capacity to detect clusters of any shape, which is a desired characteristics when one doesn't have any knowledge of the possible clusters' shapes, or when the objects are circulated heterogenously, for example, along paths of a graph or a road network. In any case, to drive the process, this algorithm needs two numeric input parameters, minPts and which together characterize the desired density characteristics of the generated clusters. In particular, minPts is a positive integer determining the minimum number of objects that must exist inside a maximum distance of the data space all together for an object to have a place with a cluster. Since DBSCAN is extremely sensible to the setting of these input parameters they should be picked with incredible accuracy by considering both the scale of the dataset and the closeness of the objects all together not to affect an excessive amount of both the speed of the algorithm and the effectiveness of the outcomes. To settle the right values of these parameters one by and large engages an exploration phase of trials and errors in which the clustering is run several times with distinct values of the parameters [15].

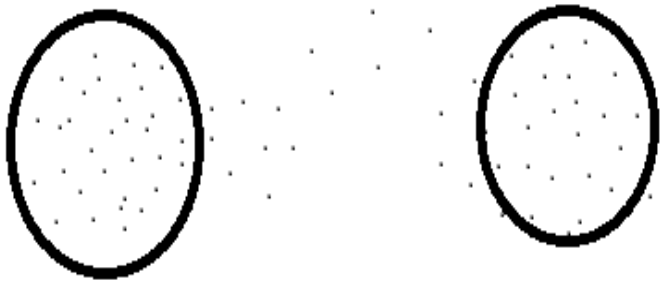


Fig. 1.7 Density based clustering

Chapter 2

REVIEW OF LITERATURE

Hao Huang *et al* (2014) proposed in paper [16] that within the data mining methods, the mining of arbitrary shaped clusters is a great challenge. As there is higher time complexity, there are various solutions proposed to solve such major issue. There are various algorithms that try to reduce the size of the dataset which further help in minimizing the computational costs of the systems. However, the clustering performance can be affected with the help of user-defined reduction ratios provided within this method. In this paper, an effective and efficient algorithm known as CLASP is proposed which mines the arbitrary shaped clusters present within the data. This process helps in reducing the overall size of the dataset along with the preservation of the information of the shapes of the clusters. The information is saved within the dataset through the

representative data examples. The positions of the representative data examples are modified for improving the internal relationships within this method. This helps in providing more clear and distinct clusters within the clustering technique. Here, a Pk metric is utilized for recognizing the clustering structures within the agglomerative clustering. This is done on the basis of k-nearest neighbours based metrics. There are various experiments performed on the synthetic as well as real datasets. The simulation results achieved show that the newly proposed technique helps in providing effective and efficient mechanisms to solve the problems.

Gunnar Carlsson *et al* (2014) proposed in this paper [17] that the asymmetry of the input data can be preserved through a hierarchical quasi clustering method. This is a generalized hierarchical clustering method which is utilized for the output structure within the asymmetric networks. There is a great similarity between the finite quasi ultra-metric space and the output of the asymmetric network whose admissibility related to the two distinct properties of the system. The only admissible quasi-clustering technique is the enhanced version of the single linkage method. There are various invariance properties that are achieved through this process and the method thus provides stability within the system.

R. Jensi Dr.G.Wiselin Jiji *et al* (2013) proposed [18] 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. 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 [19] 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 behaviour of customer that what they want or what are their requirements. For e-commerce conventional methods are no longer useful to find customers behaviour. 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 behaviour in online shopping strategy of attracting customers with good offers and combos is done by seeing their profiles. Age, gender and behaviour are main attributes for analysing the customers marketing in e-commerce.

Satoshi Takumi*et a l*(2012)explains [20] 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. [21] 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 [22] 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, 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) proposed in this paper [23] that there are various techniques being introduced for clustering the diverse data which is to be stored by the applications within groups. There are two various ways utilized by the clustering algorithms. Either the pure numeric data or the pure categorical data can be performed here on both the mixed categorical as well as the numeric data types within this system. In this paper, a new two-step clustering method is proposed where the items present within the categorical attributes are processed such that relationships amongst them are recognized on the basis of various similar properties. The co-occurrence of various objects is proposed on the basis of similarity amongst the objects. The categorical attributes are converted into numeric attributes on the basis of relationships amongst them within this process. There are various clustering algorithms within the dataset that can be applied to convert the categorical data into numeric. There are various demerits of these already existing technique and for avoiding such issues the two-step method is required that adds attributes to the clusters along with the integrated hierarchical and partitioning clustering algorithms within the system. As per the simulation results achieved it can be seen that the accuracy is improved here and enhanced results are gathered which can help to cluster the mixed numeric and categorical methods.

Wilhelmiina Hamalainen *et al* (2008) introduced [24] 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.

Jiawei Han J and Kamber M (2012) proposed in this paper [25] that there is a need of execution of huge parallel computer programs to achieve the simulation of complex scientific systems. Across the spatio-temporal space, there are large-scale datasets present which provide simulation programs. In this paper, a simple however effective multivariate clustering algorithm is presented which provides simulations for huge datasets. A linking algorithm is also utilized within this system for connecting the clusters to their appropriate nodes within the dataset of the topology tree. According to the simulation results achieved, the value of our multivariate clustering and linking algorithm determined from the two huge simulation datasets given.

Hui Xiong et.al (2013) proposed in this paper [26] that within the data cleaning process, the removal of noises present within the data is an important task. The analysis of data is interrupted due to the presence of noise within it. So, it is important to remove the noise present here. There are various noises that are present due to the low-level data errors present within the data for the removal of which various algorithms already exist. However, there are various data objects that are irrelevant or weakly relevant to each other which can also degrade the analysis process. If the analysis is to be done at a very higher level, the objects present within the data should also be considered as noise as per the underlying analysis. It is seen through the experimental results that the proposed methods provide the performance of clustering to be enhanced. There are many quality association patterns proposed here which have higher quality. This is due to the removal of noise in higher amount with the help of this technique. There are various other techniques that involve the binary data but have less efficient results as compared to the technique proposed within this paper.

Yu Qian and Kang Zhang (2005) proposed in this paper [27] that it is important to use the visualization techniques for assisting the conventional data mining tasks. A major issue within the visualization process is to select appropriate parameters for spatial data cleaning methods. To resolve this method the performance of visualization technique is enhanced in this paper. Further, the characteristics and properties of the methods and various features of the data are to be presented so that the user can get a feedback regarding its own data. Waterfall, a 3-D

visualization model is proposed in this paper for assisting the spatial data cleaning with the four important measures. They are dimension-independent data visualization, visualization of data quality, algorithm parameter selection and measurement of noise removing methods on parameter sensitiveness.

Sumit Garg and Arvind K. Sharma (2013) proposed in this paper [28] that the there have been many recent advancements made in the data mining techniques for growing its efficiency. Various new patterns are to be discovered within these huge datasets through this process. There are various algorithms being proposed by the researchers in the recent times for providing enhancements in the techniques. There are variety of data types available and so each of them cannot utilize one single algorithm. There have been varieties of algorithms proposed for various types of datasets present. Therefore the compatibility of the dataset is an equally important factor to be considered such as the end goal to be achieved by the application for choosing an appropriate data mining algorithm. The main objective here is to provide an appropriate algorithm on the educational dataset provided by an application. A comparative analysis has been done and various data mining algorithms have been compared with each other to provide results that would be beneficial during the selection of algorithms.

K. Krishna and Raghu (2001) proposed in this paper [29] that there are various subsets of dataset which satisfy the various criterions for identifying the data from the datasets. This can be done with much ease through the relations that exist among the data present within datasets. The relation identified should be symmetric in nature. For instance, the inclusion relation which is mainly the block of text within the meaning of another block is an asymmetric relation present within the text analysis. The asymmetric data is related to each other through the newly proposed algorithm such that it can be clustered easily. There are two applications which utilize such technique. First is the summarization of short documents and the second is the creation of a hierarchy from the set of documents present within the dataset. The simulation results achieved determine that the performance of this algorithm is efficient than the already existing techniques.

Ahmad M. Bakr, et.al (2015) proposed in this paper [30] that in dynamic information conditions, for example, the web, the amount of information is rapidly increasing. Along these lines, the need to organize such information in an efficient manner is more important than any time in recent memory. With such dynamic nature, incremental clustering algorithms are

constantly preferred compared to traditional static algorithms. In this paper, an enhanced version of the incremental DBSCAN algorithm is presented for incrementally building and updating arbitrary shaped clusters in substantial datasets. The proposed algorithm upgrades the incremental clustering process by constraining the inquiry space to partitions as opposed to the entire dataset which results in significant improvements in the performance compared to relevant incremental clustering algorithms. Exploratory results with datasets of various sizes and dimensions demonstrate that the proposed algorithm speeds up the incremental clustering process by factor up to 3.2 compared to existing incremental algorithms. Other conceivable improvements to the proposed algorithm are gotten ready for future work. Likely a major upgrade is designing the algorithm to work in a parallel manner. Given the independence of the partitions, incremental DBSCAN for each segment can be connected in parallel. It is normal that the parallel version of the proposed algorithm will accomplish better performance with comparable accuracy.

Guangchun Luo, et.al, (2016) proposed in this paper [31] that there has been advancement in the big data which is mainly due to the increased growth of data in all the fields. There are various parallelization methods used for the processing or extracting of data as per the requirements of the user. The cluster analysis is an important task being implemented within the data mining and among all the techniques derived within it, the DBSAN algorithm is the most prominently utilized algorithm. There are various divisions of the database generated into disjoint partitions with the help of the already existing parallel DBSAN algorithm. The data dimensions are also increased here as the splitting and consolidating the high-dimensional space will take more duration. Hus, to solve all such issues arising in the existing algorithm, a parallel DBSAN algorithm known as the S_DBSAAN which utilizes Spark in proposed. The partitions of the original data can be done in a very easy manner through this process. Further, the clustering results are mixed. There has been some data provided on the annual basis on which this proposed algorithm has been applied. It has been seen through the experimental results that there can be an effective and efficient generation of the clusters and the noise data present within the data set can also be recognized.

Dianwei Han, et.al, (2016) proposed in this paper [32] that for the purpose of identifying the arbitrary shaped clusters and eliminating the noise data the DBSCAN clustering algorithm is

utilized. On the basis of the MPI or OpenMP environments, the parallelization of DBSCAN algorithm is utilized. There is an absence of fault tolerance within this method. The workload is balanced within this algorithm and the process is enhanced in such prominent manners. There is a need of much experience for providing enhancements within such algorithms for handling the communication amongst the nodes. There have been a lot of applications that have utilized the DBSCAN algorithm for their own needs. So, this algorithm has proven to be more efficient in terms of performances and the experience is also huge. Also, this algorithm has been utilized for detecting the arbitrary shaped clusters and so it is very helpful in providing such efficient results which remove the noise within the data easily. The Spark is utilized within the DBSCAN algorithm for providing enhancement in the DBSCAN algorithm. As per the simulation results achieved, the proposed algorithm has been more efficient in providing required results.

Nagaraju S, et.al, (2016) proposed in this paper [33] that for the detection of embedded and nested adjacent clusters an efficient algorithm has been utilized within the cluster analysis method. This is done on the basis of the density based notion of the clusters as well as the difference between the neighborhood clusters. There has been enhancement made within the already existing DBSCAN algorithm with the help of the global density parameters. This also provides the identification of nested adjacent clusters within the EnDBSCAN algorithm. there are various parameters to be utilized within this proposed method for enhancing the performance of the algorithms. The detection of embedded and nested adjacent clusters is done with the help of density based notion parameters and the difference between the neighbors. It is seen through the experimental results that the enhanced DBSCAN algorithm has provided better results as compared to the earlier provided algorithm. The nested adjacent clusters have provided comparisons between the both of the algorithms. The processes included here do not add any computational complexity within the algorithms and the procedure has provided enhanced results. The global density parameters are also achieved within this paper with the help of sorted k-distance plot and the first order derivative processes.

Jianbing Shen, et.al, (2016) proposed in this paper [34] with the help of DBSCAN algorithm a real-time image super-pixel segmentation method. A faster two-stage framework is proposed in this paper for decreasing the computational costs within the super-pixel algorithms. For the purpose of clustering the pixels at higher rate, the various factors such as color similarity and

geometric confinements are used at the principal clustering stage. Further, the neighborhood clusters help in merging the smaller clusters into super-pixels on the basis of the similarity between color and spatial features. For the purpose of achieving better super-pixels within the two mentioned stages, a robust and straightforward distance function is proposed here. As per the experimental results achieved, the proposed algorithm has outperformed the existing algorithm in terms of accuracy and efficiency. The calculation cost within this method is also reduced and the results are also enhanced as compared to the algorithms that have more computational costs. There are also algorithms that have complex objects or texture areas. These algorithms have also been easily calculated and the computational costs have also been less as compared to when the existing DBSCAN algorithms are utilized.

Ilias K. Savvas, et.al, (2016) proposed in this paper [35] that through the computational frameworks and the electronic devices, a large amount of information is being extracted every day. There is a need of new algorithms for managing and extracting all such data from the datasets. For the purpose of allocating and extracting the required data from the data warehouses, various algorithms have been proposed. The most prominently utilized methods here is the clustering process. The clustering of the data according to its characteristics is done with the help of DBSCAN algorithm. The computational complexity within these processes is higher due to which the applications of such algorithms within the large datasets is not possible. There have been numerous enhancements made within the DBSCAN algorithm. However, the changes made have not yet been up to the satisfaction of the researchers and there has been no fixed algorithm that has met all the needs of the researchers. A three phase parallel version of DBSCAN is proposed in this paper. The accuracy, scalability as well as the effectiveness of the results has however been achieved with the help of the algorithm proposed in this paper. The parallel version of the DBSCAN is proposed here and the implementation is done with the help of MPI. Here were similar results achieved within the original sequential technique within this process. However, there has been a reduction in the time complexity within this paper. The performance provided has been improved to huge extent within this paper.

Ahmad M. Bakr, et.al, (2014) proposed in this paper [36] that in dynamic information environments, for example, the web, the amount of information is quickly increasing. Therefore, the need to organize such information in an efficient manner is more essential than any other

time in recent memory. With such dynamic nature, incremental clustering algorithms are constantly preferred compared to traditional static algorithms. In this paper, an enhanced version of the incremental DBSCAN algorithm is introduced for incrementally building and updating arbitrary shaped clusters in extensive datasets. The proposed algorithm enhances the incremental clustering process by limiting the search space to partitions as opposed to the whole dataset which results in significant improvements in the performance compared to relevant incremental clustering algorithms. Experimental results with datasets of various sizes and dimensions demonstrate that the proposed algorithm speeds up the incremental clustering process by factor up to 3.2 compared to existing incremental algorithms. In any case, the proposed algorithm has significant improvements on the runtime with a speedup factor of 3.2. The proposed algorithm is additionally proved to perform better in expansive datasets with higher dimensions compared to related algorithms.

Saefia Beri, et.al, (2015) proposed in this paper [37] that the process of acquiring the required data from the dataset is known as the data mining technique. The important part within this process is also to convert the data achieved into an understandable and meaningful manner for utilizing it further as well. The arbitrary shapes as well as the outliers are recognized with the help of DBSAN algorithm which is based on the bivalent logic. With the help of this algorithm it can be seen that the objects belong to a specific cluster or not. Within this paper, a new DBSAN algorithm is proposed which uses the fuzzy logic method within it. With the help of the membership values present within this algorithm, the degree to which the object belongs to a specific cluster can be determined. There are fuzzy if-then rules utilized within the DBSAN algorithm for hybridizing it. The multivalent logic is utilized for improving the membership values to certain degree. He simulation results achieved have shown improvement as per certain aspects such as bit error rate, specification, sensitivity as well as accuracy he results are also compared with the results achieved through previous algorithms. This technique has been helpful in selecting the cluster in a more appropriate manner.

Karlina Khiyarin Nisa, et.al, (2014) proposed in this paper [38] that there have been various areas located across the globe that have been facing the forest fire related issues. There are remote sensing satellites that have been recording the information attained from the hotspots present within the specified areas. The analysis of the datasets gathered can provide the helpful

information related to the forest fires and also can help one know the chances when they can occur. Within the R programming language, the Shiny web framework is utilized for implementing the DBSAN algorithm. The data sets of the hotspots present in Kalimantan Island and South Sumatra Province in years 2002-2003 is provided within this paper for analyzing the clustering performance. There is a need of the minPts and Eps parameters within the DBSAN algorithm in this work. There will be increase in number of noises which are less within the minPts. Here will be less number of clusters generated when the value of Eps is bigger. Various operations are performed for calculating both of these values and the acquired values are utilized within the DBSAN algorithms to enhance the performance of the results.

Negar Riazifar, et.al, (2015) proposed in this paper [39] that for the purpose of detecting various serious diseases such as hypertension, diabetes and glaucoma, the retinal vessel segmentation method has been utilized over the years. The retinal images within the algorithms have provided various methods through which the segmentation of blood vessels can be done. On the basis of the clustering DBSAN algorithm, the retinal vessel segmentation method is to be analyzed. Within this algorithm, the clusters of arbitrary shapes are to be recognized with the help of the density-based notion of the clusters. There is a need of only one parameter within this algorithm which is also to be provided with the help of the suggestions made by the client. There are numerous measures that are to be utilized for comparing the performance of the proposed algorithm with the already existing algorithms. There have already been made various advancements within this blood vessel segmentation process. However, not all the enhancements provided have been useful to the process. And so, each of the problems have been solved with the help of DBSAN algorithms which include the identification of correct input parameters, localizing the arbitrary shaped clusters and the completion of the process within the time limit given. It has been seen through the experimental results that this algorithm has given better performance results as compared to the other algorithms.

Yumian Yang, et.al, (2014) proposed in this paper [40] that the evaluation of E-commerce sites in an accurate manner has been required as its growth has been increased. There are huge dimensions of characteristics and uneven density involved within the E-commerce sites. These diversions result in decreasing the performance accuracies of the results that are evaluated. In this paper, the data which involves the 100 E-commerce sites of the Ministry of Commerce

People's Republic of China in the year of 2013-2014. At the initial step, the dimensionality of the data is reduced with the help of factor analysis. On the basis of various investigation results, the enhanced DBSAN algorithm is utilized for processing the uneven density present within the data available within these 100 E-commerce sites. When the Euclidean distance is to be calculated, the weights of the data are ignored by the previous algorithm. This will result in causing major issues within the accuracy of the results achieved. However, the factor analysis process eliminates all such problems arising within the already existing algorithms. There are various comparisons made among the new proposed DBSAN algorithm as well as the already existing ones. The simulation results achieved show that there has been an improvement in the results achieved by the new proposed algorithm.

Xiaoqing Yu, et.al, (2014) proposed in this paper [41] that the spatial clustering is a very important factor affecting the data mining and knowledge discovery processes. For the purpose of removing noise within the arbitrary shaped clusters, the DBSAN algorithm has been considered as a good method. The clustering analysis distribution process is to be performed on the weibo location information and within this paper, the DBSAN algorithm is utilized for this. The k-means clustering algorithm is compared with the DBSAN algorithm which can show that the DBSAN algorithm is really beneficial in providing the required results. Numerous noise points and data points are generated within each cluster through this process. There is no shaping of the noise points provided within the clusters with the help of k-means clustering algorithm. There is thus, various numbers of points that can affect it. The main objective of the DBSAN algorithm is identified and is implemented here. The current city of the territory can be located with the help of this algorithm when it is applied to the data that has the complete city planning. The effectiveness of this algorithm can be seen with the help of its comparisons made with various other algorithms. Further, more operations can be performed within this process which can also differentiate the data on the basis of various different properties.

Chapter 3

PROBLEM FORMULATION

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 doneon 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. This work, is based on the multi-view clustering in which the data is clustered on the basis of their density. In the multi-view clustering input dataset is pre-processed and most dense region is calculated from the data. In the second phase, the EPS value is calculated which will be the central point for the clustering. In the last phase, Euclidian distance is calculated from the central point. The points which has similar distance is clustered in one cluster and other in the second. The accuracy of the multi-view clustering less because distance is calculated statically for the clustering.

Chapter 4

OBJECTIVES OF STUDY

Following are the various objectives of this research :-

1. To study and analyse various asymmetric clustering technique to cluster relevant and irrelevant data
2. To propose enhancement in the multi-view clustering to improve accuracy of the algorithm
3. The proposed enhancement will be based on the neural network to cluster the un-clustered data points

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

Chapter 5
RESEARCH METHODOLOGY

The muti-view clustering is also called the density based clustering. In the technique of density based clustering the density of the dataset is calculated which define dense and non-dense regions. The most dense region is considered to generate final clustered data. The EPS value is calculated from the dense region which will be the central point for the generation of final clustered. In the second phase, the Euclidian distance will be calculated from the central point and calculated distance will be analyzed for their similarity and dissimilarity. The points which has similar distance will be clustered in one cluster and other are in the second cluster. In this work, algorithm of back propagation is applied which will calculate the Euclidian distance in the dynamic manner. The back propagation algorithm is one of the most utilized Neural Network

algorithms. This method is used for training the artificial neural networks and also utilizes the two phase cycle which involves the propagation and weight updates. When an input network enters the network, it is propagated forward through the network across each layer until it reaches the output layer. The comparisons are made using the output achieved as well as the desired output. This is done utilizing a loss function. For every neuron in the output layer, an error value is calculated. The propagation of the error values is then done in backward manner which starts from the output. Here, each neuron has its own error value which also shows its contribution to the originally achieved output.

There are mainly four steps in which this algorithm can be executed. The required corrections are to be computed only once the weights of the network are selected randomly. The following are the steps in which the algorithm is decomposed:

i) Feed-forward computation

ii) Back propagation to the output layer

iii) Back propagation to the hidden layer

iv) Weight updates

At the time when the values of error function become small, the algorithm is stopped. This is just an overview of the basic BP algorithm. However, various changes are proposed by researchers with time. The algorithm for back propagation is mentioned below:

$$\text{Actual Output: } \sum_{\substack{x=0 \\ w=0}}^{\substack{x=n \\ w=n}} x_n w_n + bias$$

$$Error = Desired\ Output - Actual\ Output$$

In the algorithm, the data point and their value is considered as the input for the calculation of actual output. The actual output will be the input for the next iteration unless the error get minimized. The error is inversely proportional to accuracy. The algorithm has maximum accuracy of clustering when error get minimized

Pseudo code of Proposed Algorithm

Input: Data set $P = \{p1, p2, \ldots, pn\}$, > 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 $i = 1, 2, \ldots, n$, sort the ith row of W, then calculate $pi\ K$, which is the Kth neighbor of pi , $K = 2, \ldots, K$max;

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

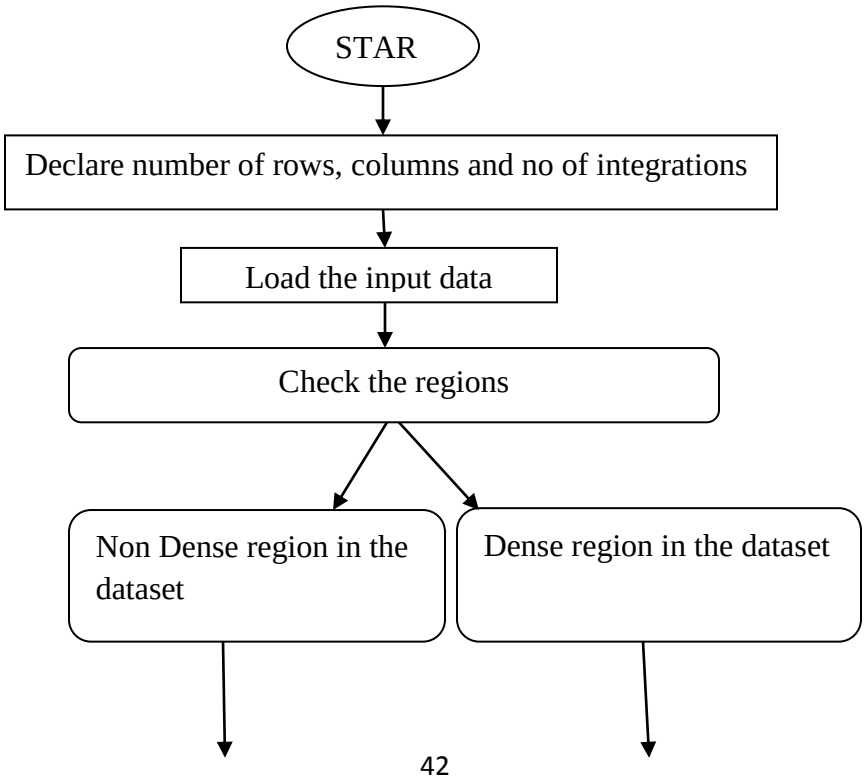
Step 4 Calculate the similarity matrix S, where $S(i, j)$ = exp( );

Step 5 For every $k = 2, \ldots, C$max, make use of the Meil˘a–Shi spectral clustering algorithm to cluster the data set P into k clusters and calculate the value of index Ratio(k) for obtained clusters;

Step 6 To determine whether the candidate cluster number $2 \leq k \leq C$max is an - 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.

In Fig.5.1



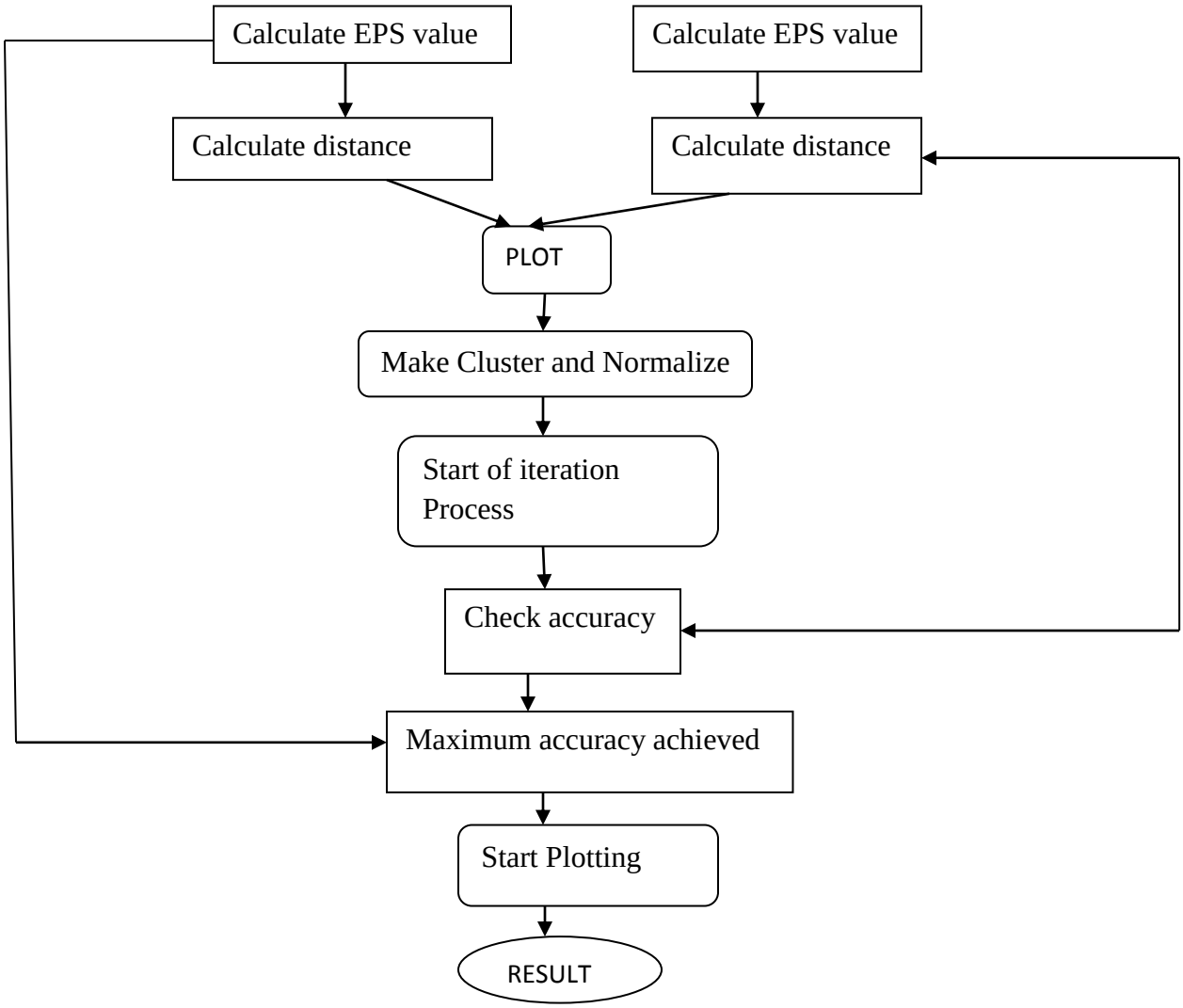


Fig5. 1: Flowchart Of proposed Algorithm

In the figure 1, proposed flowchart is illustrated in which input data is pre-processed and dataset is analyzed in which the regions which is dense and which are non-dense is analyzed. The dense region is analyzed and EPS value is calculated from the dense region which define central point of the dataset. The Euclidian distance is calculated as each phase and point at which maximum accuracy is achieved as considered as the final clustering result which will be plotted on the 2-D plane .

Chapter 6
IMPLEMENTATION

The matlab is the tool which is used to perform mathematical complex computations. In this MATLAB simplified C is used as the programming language. The MATLAB has various inbuilt toolboxes and these toolboxes are mathematical toolbox, drag and drop based GUI, Image processing, Neural networks etc. The MATLAB is generally used to implement algorithms, plotting graphs and design user interfaces. The MATLAB has high graphics due to which it is used to simulate networks. The MATLAB has various versions by current MATLAB version is 2015. The MATLAB process elements in the form of MATRIXs and various other languages

like JAVA, PYTHON and FORTAN are used in MATLAB. The MATLAB default interface has following parts

1. **Command Window:-** The Command Window is the first importance part of MATLAB which is used to show output of already saved code and to execute MATLAB codes temporarily
2. **WorkSpace** :-The workspace is the second part of MATLAB which is used to show allocation and deallocation of MATLAB variables. The workspace is divided into three parts. The first part is MATLAB variable,variable type and third part is variable value
3. **Command History** :- The command history is the third part of MATLAB in which MATLAB commands are shown which are executed previously
4. **Current Folder Path** :- The current Folder path shows that path of the folder in which MATLAB codes are saved
5. **Current Folder Data**: - The Current Folder Data shows that data which is in the folders whose path is given in Current Folder Path

The MATLAB has three Command which are used frequently and these commands are :-

1. CLC= The ‘clc’ stands for clear command window
2. Clear all:- The ‘clear all’ command is used to de-allocate the variable from the workspace
3. Close all:- The close all is the command which is used to close all the interfaces and return you to default MATLAB interface

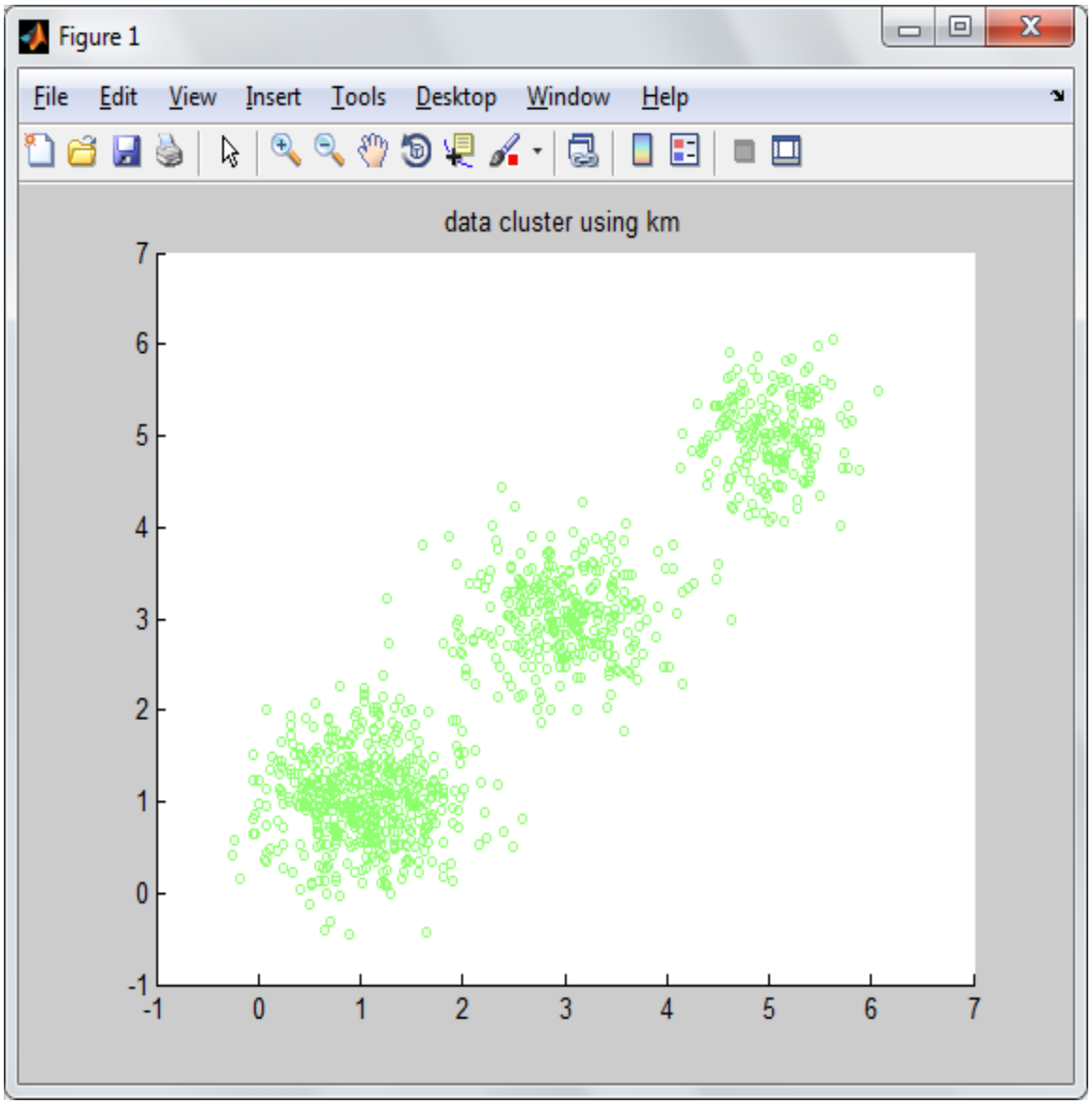


Fig 6.1: Scattering of data

As shown in figure 6.1, as in the previous chapter, the dataset which is loaded will be scattered and plotted on the 2D plane

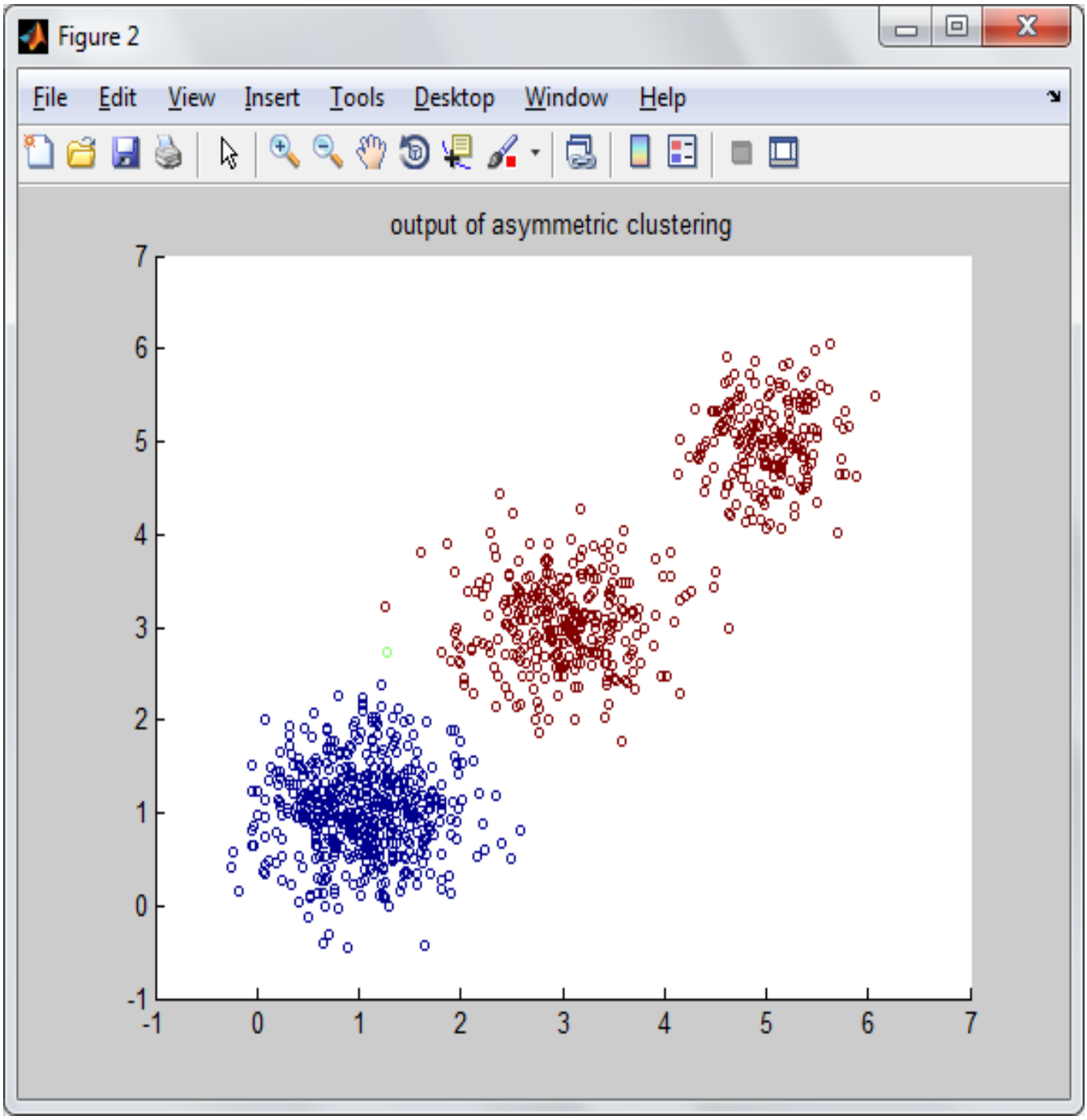


Fig 6.2: Clustering of data

As illustrated in the figure 1, the dataset which is loaded had been scattered and scattered data is clustered asymmetric according to asymmetric between the loaded data

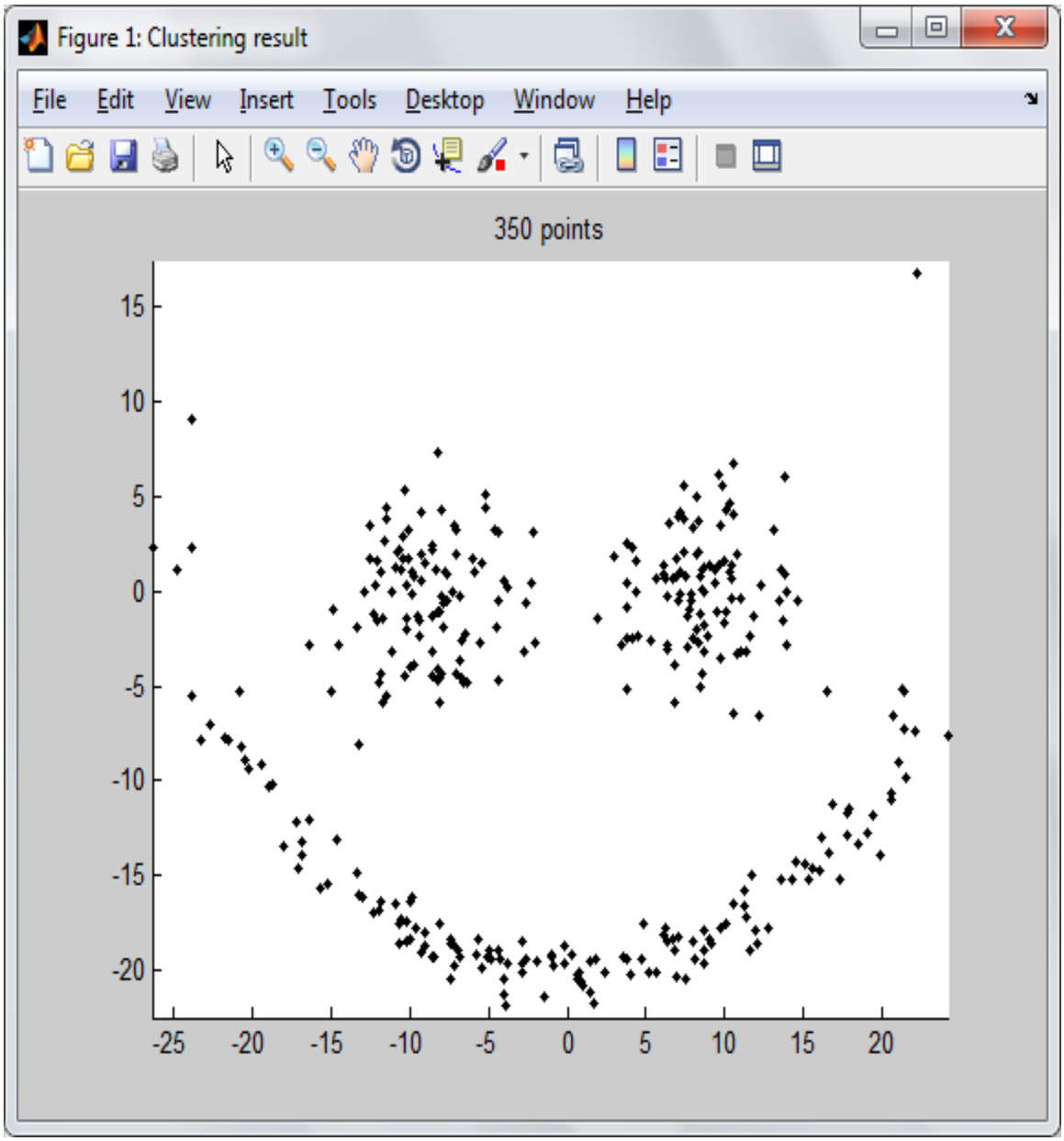


Fig 6.3: Loading of data

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

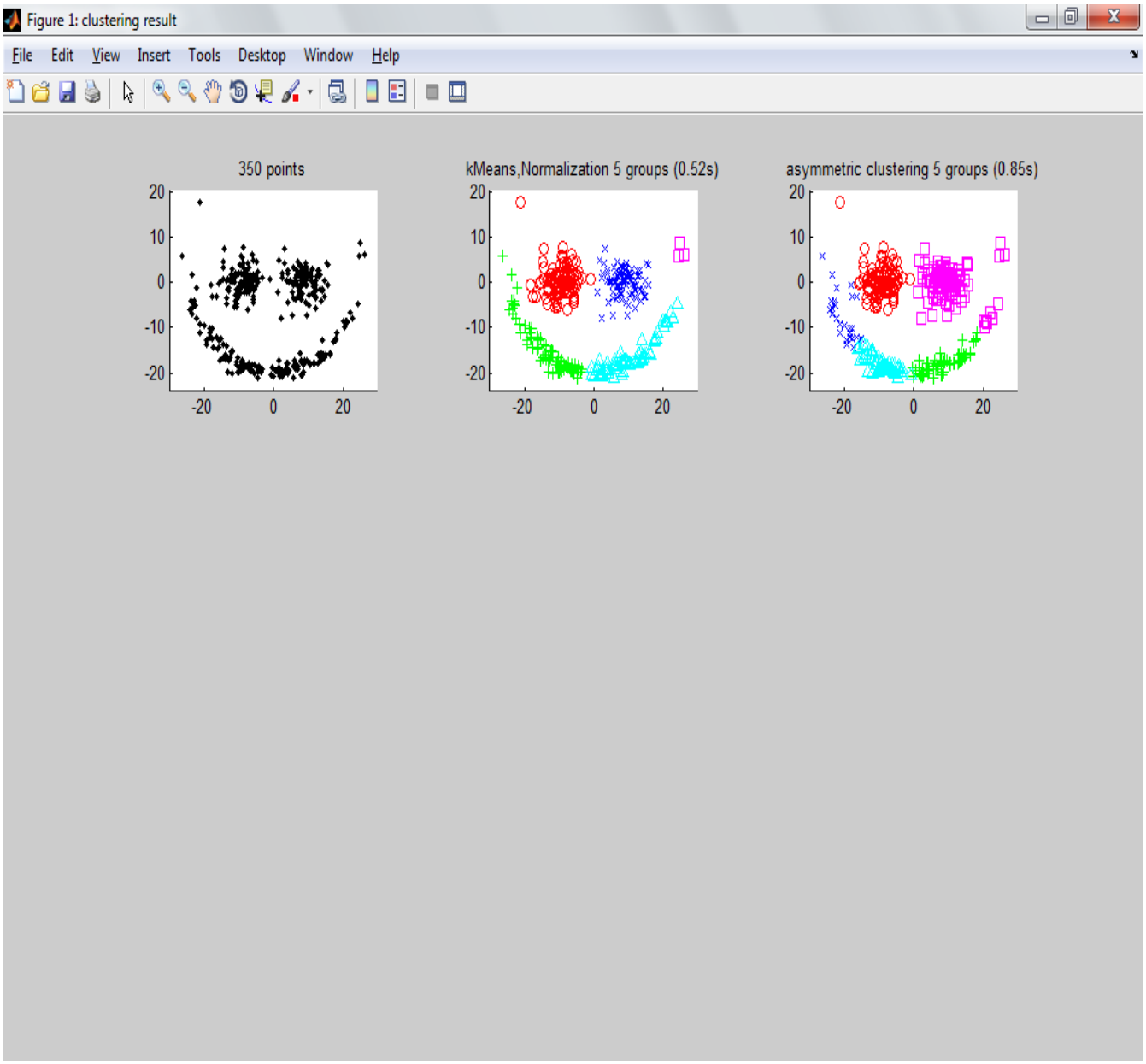


Fig 6.4: Clustering of data

As shown in the figure 3, the data is clustered using the multi-view clustering in which the most dense is calculated and Euclidian distance is calculated with the back-propagation algorithm. The output of the final clustering is shown in the sub-plot

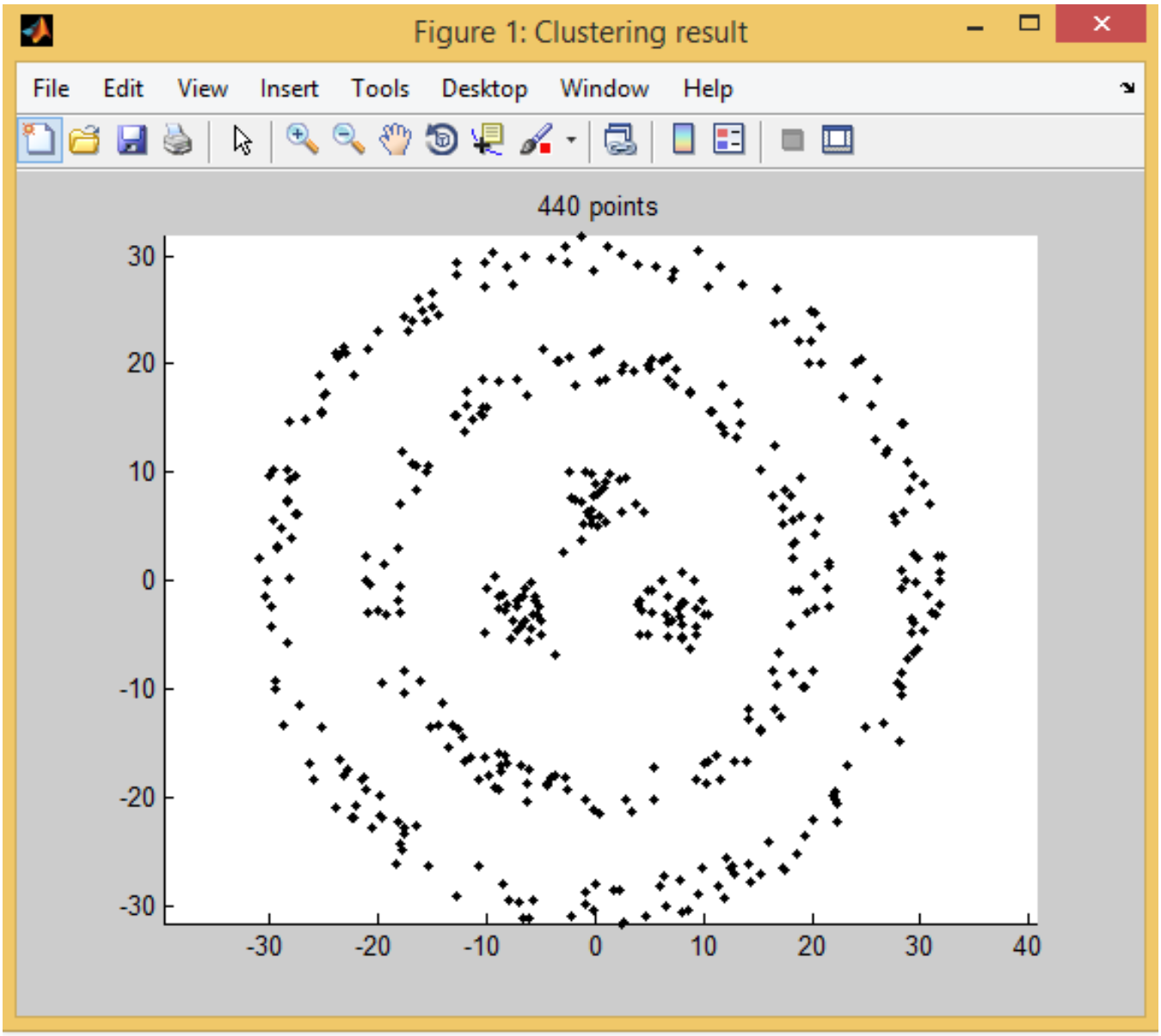


Fig6.5: Loading of data

As shown in the figure 4, the data points which need to be considered are 440 points. These points are given as input to generate final clusters

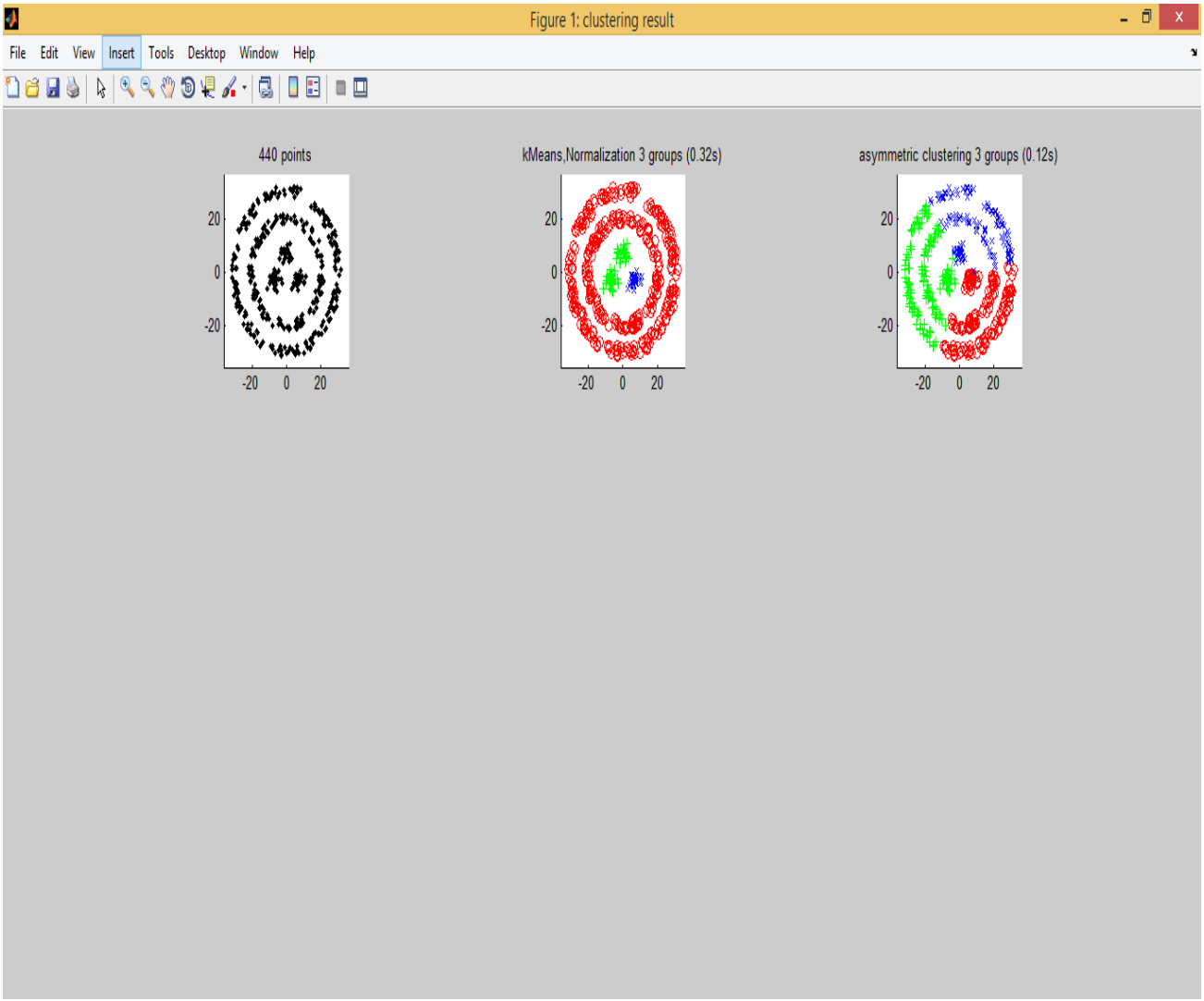


Fig 6.6: Clustering of data

As shown in the figure 5, the data is clustered using the multi-view clustering in which the most dense is calculated and Euclidian distance is calculated with the back-propagation algorithm. The output of the final clustering is shown in the sub-plot

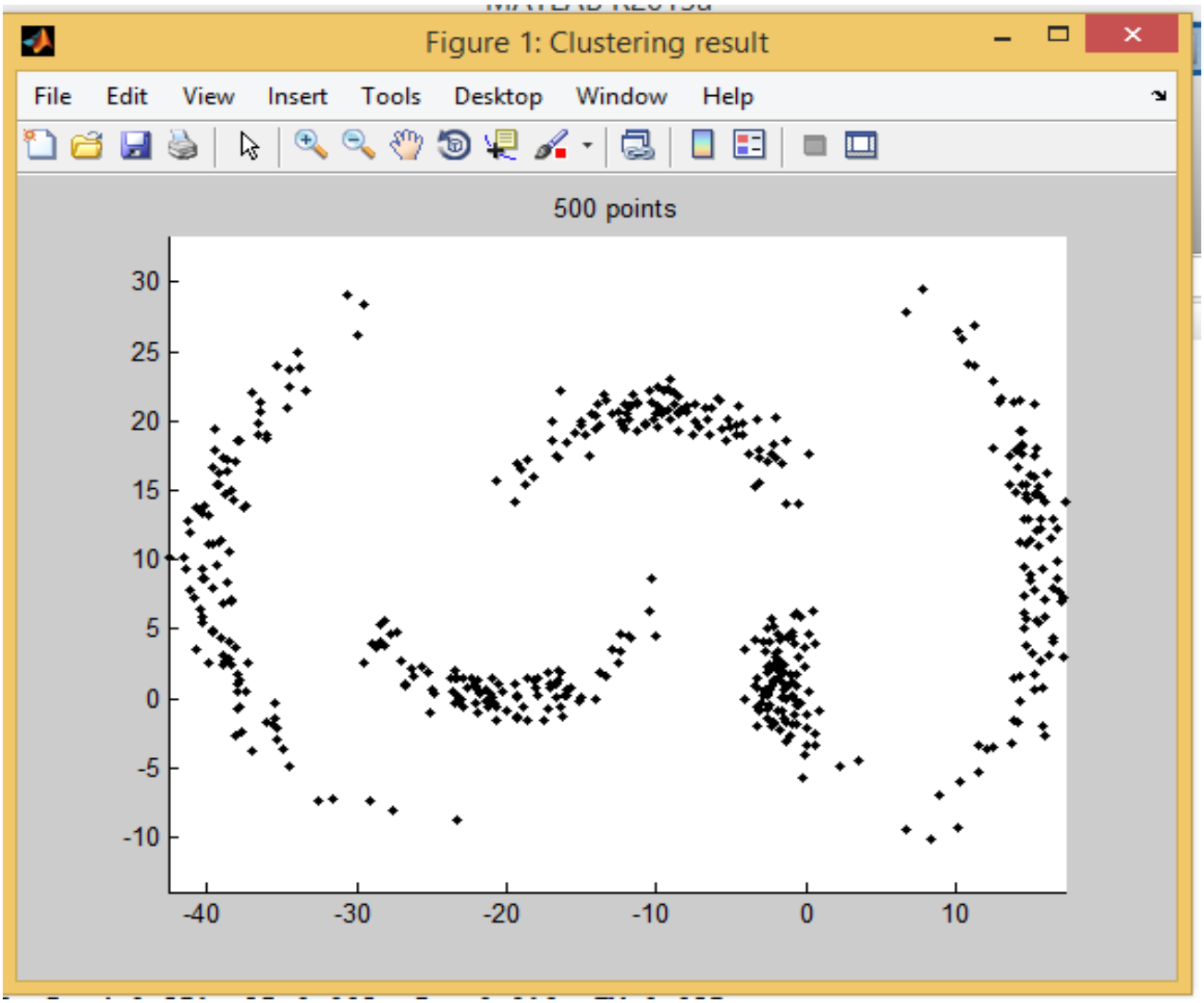


Fig 6.7: Loading of data

As illustrated in figure 6, the dataset is loaded and no of rows and columns are defied. The second step is to ask for iterations. According to number of iterations defined data is shown into the 2 D plane

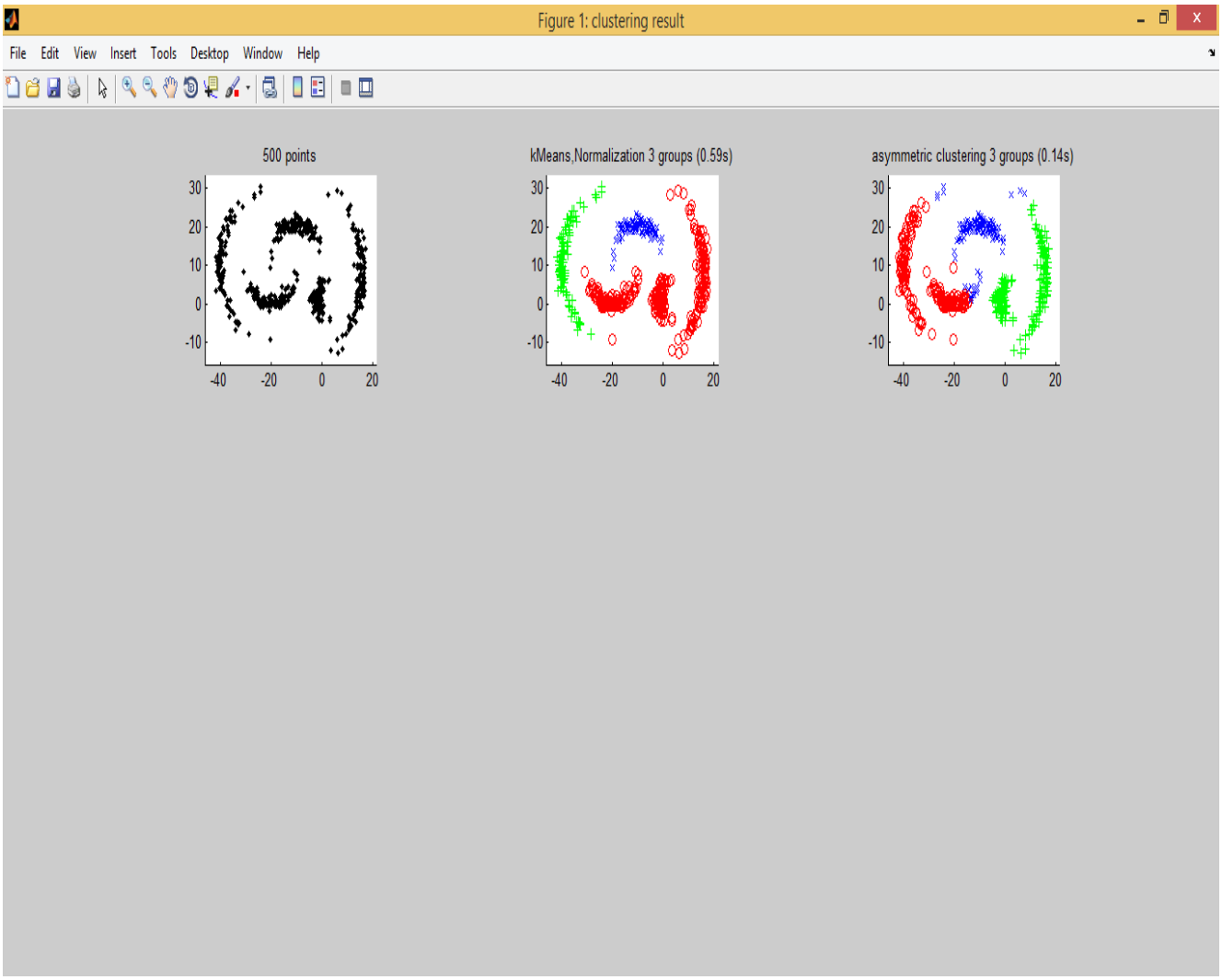


Fig 6.8: Clustering of data

As shown in the figure 7, the data is clustered using the multi-view clustering in which the most dense is calculated and Euclidian distance is calculated with the back-propagation algorithm. The output of the final clustering is shown in the sub-plot

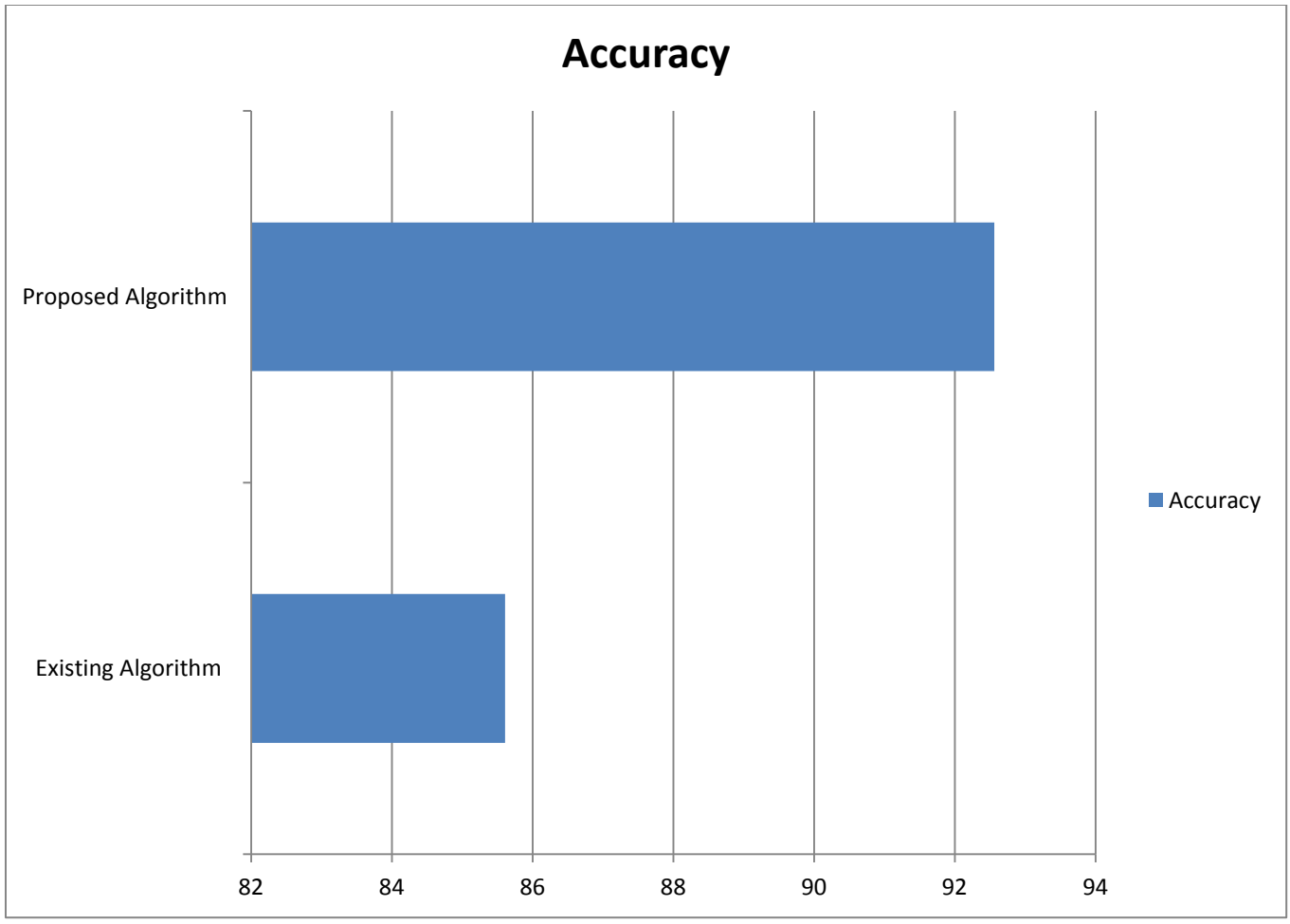


Fig6. 9: Accuracy of clustering

As shown in figure 8, the accuracy of proposed and existing algorithm is compared to check realibility of the algorithms and it is been analyzed that accuracy of proposed algorithm is more as compared to existing algorithm

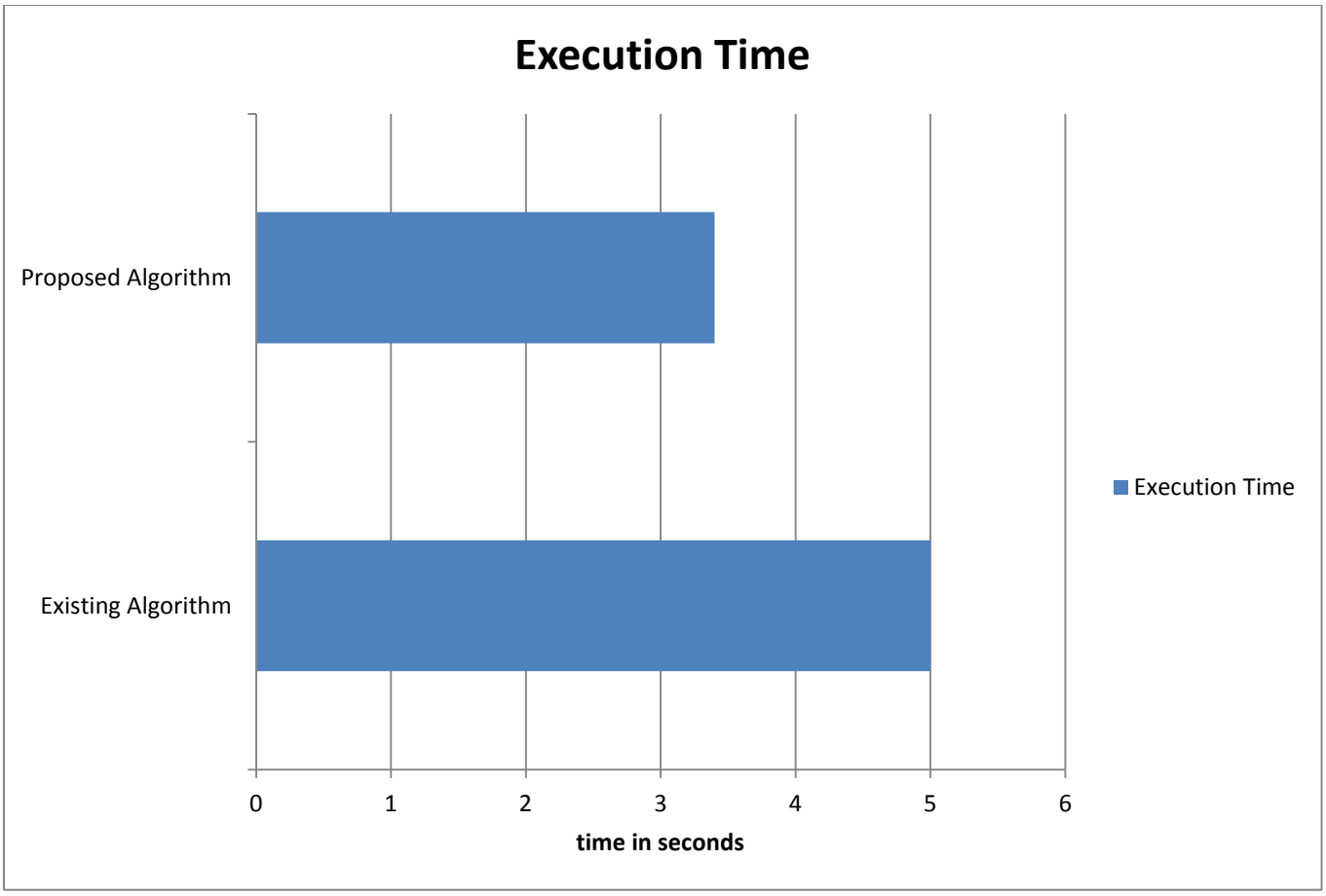


Fig6. 10: Execution time Comparison

As shown n figure 9, the execution time of proposed and existing algorithm is compared and it is been analyzed that due to dynamic calculation of Euclidian distance execution time is reduced in the DBSCAN algorithm

Chapter 7

SUMMARY AND CONCLUSIONS

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 dividing 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. Neural networks can be combined with these existing asymmetric algorithms to improve and accuracy and reduce escape time.

References

[1] Rajkumar Buyya, James Broberg, Andrzej Goscinski, "Cloud Computing Principles and [1] Rajkumar Buyya, James Broberg, Andrzej Goscinski, "Cloud Computing Principles and Paradigms", 2011, John Wiley & Sons, Inc publications

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

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

[4] 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.

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

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

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

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

[9] 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

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

[11] 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

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

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

[14] Jan Zeleny, "Web Page Segmentation and Classification" Journal of Data and Knowledge Engineering, 2010

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

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

[21] 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

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

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

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

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

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

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

[28] 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

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

[30] Ahmad M. Bakr , Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2015, Elsevier B.V.

[31] Guangchun Luo, Xiaoyu Luo, Thomas Fairley Gooch, Ling Tian, Ke Qin," A Parallel DBSCAN Algorithm Based On Spark", 2016, IEEE, 978-1-5090-3936-4

[32] Dianwei Han, Ankit Agrawal, Wei−keng Liao, Alok Choudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4

[33] Nagaraju S,Manish Kashyap, Mahua Bhattacharya," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, IEEE, 978-1-4673-9197-9

[34] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE, 1057-7149

[35] Ilias K. Savvas, and Dimitrios Tselios," Parallelizing DBSCAN Algorithm Using MPI", 2016, IEEE, 978-1-5090-1663-1

[36] Ahmad M. Bakr , Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2014, Elsevier Pvt. Ltd.

[37] Saefia Beri, Kamaljit Kaur," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015, IEEE, 978-1-4799-8433-6

[38] Karlina Khiyarin Nisa, Hari Agung Andrianto, Rahmah Mardhiyyah," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework", 2014, IEEE, 978-1-4799-8075-8

[39] Negar Riazifar, Ehsan Saghapour," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015, IEEE, 978-1-4799-8445-9

[40] Yumian Yang, Jianhua Jiang," Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm", 2014, IEEE, 978-1-4799-6543-4

[41] XiaoqingYu, Yupu Ding, Wanggen Wan, Etienne Thuillier," Explore Hot Spots of City Based on DBSCAN Algorithm", 2014, IEEE, 978-1-4799-3903-9