IMPROVISING DBSCAN ALGORITHM BY INCREASING ACCURACY OF CLUSTERING

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

PAYAL GARG

11207094

Supervisor

JANPREET SINGH

(ASSISTANT PROFESSOR)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

January-May, 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : 1202D::B.Tech -M.Tech (Dual Degree) - CSE

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0285

Supervisor Name : Janpreet Singh **UID :** 11266 **Designation :** Assistant Professor

Qualification : ________________________ **Research Experience :** ________________________

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Payal Garg	11207094	2012	K1209	9779362333

SPECIALIZATION AREA : Software Engineering **Supervisor Signature:** ____________________

PROPOSED TOPIC : Data mining Algorithm (clustering).

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.20
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.80
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.80
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.40
5	Social Applicability: Project work intends to solve a practical problem.	6.60
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): NA
PAC Member 2 Name: Mandeep Singh	UID: 13742	Recommended (Y/N): Yes
PAC Member 3 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 6 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: **To increase accuracy of Density Based Clustering(DB Scan Algorithm) and reduce execution time by using back propagation technique.**

Overall Remarks: Approved (with major changes)

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 22 Nov 2016

4/26/2017 1:29:02 PM

ABSTRACT

In this report I have studied about Data Mining which is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining is also known as Knowledge Discovery of Databases which includes Selection, Pre-Processing, Transformation, Data Mining, and Evaluation. Then I have discussed about preprocessing techniques which includes data cleaning, data transformation, data reduction and data selection.

Next, I have discussed about clustering which is an unsupervised learning. In clustering objects within a class are more similar to each other in the meantime objects in separate class are more unlike. Various methods are includes like partitioning method, hierarchical method, density based method, grid based method etc.

Next, I have discussed about my problem which is DBSCAN that is Density Based Spatial Clustering of Application with Noise. The density based technique is the type of algorithm in which density of the whole dataset is calculated and most dense region is calculated to find similarity between the elements of the dataset. In the existing work, technique of density based clustering is applied in which density of whole dataset is calculated and dense region is calculated. On the Dense region EPS value is calculated to analyze similarity between the elements. The Euclidian distance is applied to analyze similarity between the elements. The EPS is calculated in the dynamic order to achieve maximum accuracy. The Euclidian distance is calculated in the static manner due to which accuracy is not achieved at the maximum point. In this work, improvement in DBSCAN algorithm has been implemented which calculate Euclidian distance in the iterative manner to increase accuracy of clustering

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "IMPROVISING DBSCAN ALGORITHM BY INCREASING ACCURACY OF CLUSTERING” in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Janpreet Singh. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Payal Garg

11207094

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**IMPROVISING DBSCAN ALGORITHM BY INCREASING ACCURACY OF CLUSTERING"**, submitted by **Payal Garg** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Mr. Janpreet Singh

Date: 27-04-2017

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

ACKNOWLEDGEMENT

I feel immense happiness for the completion of the Dissertation-II under the stipulated time duration to fulfill requirements and specifications under the guidance of guideposts which acted as lightening pillars to enlighten the way for carrying out Dissertation-II completion steps and due to which I was able to reach the heights of its completion and study.

I feel really grateful to my Mentor Mr. Janpreet Singh for providing me for all guidance and valuable time, care, support, sincere operation and involvement during all the phases of the Dissertation-II. I thank for the continuous efforts for completing my Dissertation-II within a fixed time.

Without thanking all who have been supporting and helping me throughout the overall reviewing and searching for the Research papers and the sites related to it, this acknowledgement could not be completed. I thank them for their support in ideas, creation and their help in maintaining and supporting. I am really thankful to my University, my Faculty Member’s guidance, without their help the Dissertation-II could not be taken to a feasible state.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

S. No.	Abbreviation	Expansion
1.	DBMS	Database Management System
2.	KDD	Knowledge Discovery in Database
3.	DBSCAN	Density- Based Clustering of Application With Noise
4.	MATLAB	Matrix Lab

Chapter 1

INTRODUCTION

1.1 ‘Data Mining

The computers, satellites and other technologies are present today which are source of deriving information. Large amount of information is gathered from surroundings and used by the people for their own profits. The information which can be further required is stored eventually. There are computers and other mass digital storage devices present in the technological world which provide the facility of storing the required information. There are however, varieties of devices which help in providing the facility to store different kinds of data. A structured database has been created in order to avoid all the chaos. For the purpose of its complete management, a Database Management System (DBMS) has been evolved which helps in proper arrangement of huge data in an effective manner. The DBMS also ensures that this data can be efficiently be retrieved from the huge collection as and when required by the users. The huge collection of all such data is thus possible mainly because of the proliferation of the DBMS. The data from all of the fields such as the business world, logical information, content reports, or the military insight is to be handled. For the purpose of decision making, the information retrieval method is not enough. For the purpose of making the management of data better, various new methods have been evolving. The activities which involve the programmed synopsis of information, the mining of important data stored, and the disclosure of examples in crude data are to be taken care of here. The analysis and interpretation of such huge amount of data is very important from the stored files and databases. This can also be required for the purpose of providing important related information which can help further in decision-making [1].

Another name for Data mining is the Knowledge Discovery in Databases (KDD). From the information present in the database, the KDD help in nontrivial extraction of implicit, new, as well as potentially useful information. Data mining is originally a part of KDD which is also now used as a synonym. There are various steps followed in the case of knowledge discovery from databases. The steps begin from identifying the raw material and gathering it to form new important information. The following steps are involved in the iterative process:

- Data Cleaning: The removal of noise data or the irrelevant data from the whole collection is known as data cleaning.
- Data integration: The combination of multiple data sources which are heterogeneous, into a common source is known as data integration.
- Data choice: The data which is suitable for the analysis being performed is identified and extracted from the accumulation in this step.
- Transformation of data: In this step, the gathered data is converted into forms which are useful for the extraction procedure.
- Data extraction: In this step, the patterns that are extremely important are extracted with the help of clever techniques.
- Pattern Evaluation: On the basis of provided measures, the interesting patterns which represent the knowledge are recognized within this step.
- Knowledge Representation: In this last stage, the user can view the discovered knowledge. Virtualization techniques are used which help the users to understand and interpret the results achieved from data mining [2].

There are various steps which can be combined here. An example can be given in which cleaning of data and data combination can be combined and called as a pre-processing stage. This phase generates a data warehouse in the system. The data consolidation can be given as a result from the combination of data selection and data transformation. The transformed data is used here for selection purposes. The KDD process is completely iterative and involves the presentation of knowledge to the user, the enhancement of measures, refining the mining, the selection or transformation of new data or integration of new data sources which can help in providing required outcomes. The valuable information is searched in large databases of the system and the important data is mined, which is a perfect definition for data mining. There can be sifting performed against the whole data present, or the material can be pinpointed towards where the values are present. Various data extraction systems are already present and are also being evolved gradually. There are also another which are more versatile and comprehensive.

There are various criteria according to which the data mining systems are categorized. Some of these are defined below:

- Type of Data Source Mined: In accordance with the type of data that is to be handled such as spatial information, sight and sound information, text information etc. the data mining systems are categorized.
- Type of Data Model drawn: On the basis of data model involved like social database, protest arranged data base, etc. the data mining systems are defined and distinguished.
- Ways of Knowledge Discovered: On the basis of the kind of knowledge which is identified such as the characterization, association, clustering, etc. or the data functionalities, the systems of data mining are distinguished.
- Types of Mining Techniques used: There are various techniques being involved for the data extraction systems. As per the information analysis examples such as machine learning, neural networks, genetic algorithms, etc. being used, the data mining systems are categorized. The level of user interaction involved such as the query-driven systems, autonomous systems, etc. can also be used for classification here. Various degrees of user interaction, using a comprehensive system are taken to provide data mining techniques for different situations and options [3].

1.1.1 Issues arising in Data Mining

There are various data mining techniques which have been evolved from earlier times and are being still used for reliable and scalable tools which involve some older statistical methods. Data mining applications have been evolving ever since due to the wide applications involved. There are various issues however, which have been identified in the data mining systems. There are some issues which have been enlisted below but are not always much exclusive:

- Security and Social Issues: Any type of data which is gathered to be shared or can be used for vital decision-making needs a proper secure measure in all possible ways. Along with the collection of data for customer profiling, the correlation of personal data, user behavior understanding, etc. there is other sensitive and private data related to organizations or individuals which is gathered and put away. The classified nature of information is sometimes revealed which causes problems within the system. There might be activities performed by the data mining, which would disclose the implicit knowledge of groups or individuals. This might cause the violation of privacy policies of system especially if the information is very crucial. The proper manner in which the data mining method is to be used is another major concern here. There might be some information here, which can be controlled and not shared with the other system and the other unimportant information which is not much of hiding, can be shared openly.

- Client interface issues: When the knowledge discovered by data extraction tools is interesting and understandable, the user finds it to be beneficial. The data mining results are to be better understood with the help of good data visualization. This also helps the user to understand better the applications.
- Mining Methodology Issues: The data mining approaches are to be applied and pertained which are related to these issues. The data mining methodology choices can be interrupted through dimensionality domain, appraisal of information found, the misuse of foundation learning and metadata, the control and treatment of clamor in information, and so on [4].
- Execution Issues: For the purpose of data analysis and interpretation, various artificial intelligence and statistical methods are present which are however, not available for larger datasets. The common size present is terabyte. The scalability and efficiency are the major issues of data mining methods which are being raised here when the processing of large data is considered.
- Information Source Issues: There are various issues which are concerned with the information sources. Some of them are involved with the practical applications such as the diversity of information types. There are also the philosophical issues such as the data glut issue. There is a lot of data present in the system which cannot be handled. It is gradually increasing at high rate at each instant. There is more data harvesting to be required where there is an increase in the collection of information through the database management systems. The gathering and processing of data at the instant or to process it later is the current issue. There are some issues which are to be taken care of. The collection of right information in exact amount, the

knowledge of what tasks to be performed using it, and the identification of important or useless data are some of them.

1.2 Clustering in Data Extraction

The arrangement of information into groups of similar objects is called clustering. When there are a less number of clusters involved in the system, it achieves a level of simplification. However, some fine details are lost within these less number of clusters. The information is modeled with the help of its clusters. According to the machine learning view, the clusters can be known as the hidden patterns in which the clusters are searched in an unsupervised manner. The system that comes as an outcome defines a data concept. The definition of clustering indicates that the clustering mechanism does not have only one-step. The clustering process is divided is into following steps according to the seminal texts provided on cluster analysis, Jain and Dubes:

- Data Collection: A proper process in which the relevant data objects are extracted from the available data sources is known as data collection. For certain set of attributes, the data objects are classified according to their respective values.
- Initial Screening: The data which is extracted from various sources is made to go through certain evaluations. The data cleaning process which is also executed in data warehousing is very similar.
- Representation: For the purpose of clustering algorithm, the data is well prepared in this step. The similarity measure is selected here and the characteristics and dimensionality of the data is observed and evolved [5].
- Clustering Tendency: The natural tendency of the data to cluster is tested in this step. In cases where large data sets are involved, this step is not performed.
- Clustering Strategy: The clustering algorithms as well as the initial parameters are chose after special considerations.
- Validation: on the basis of manual experiments and some visual methods, the validation is performed. There are no comparisons of data once the amount of data and its dimensionality grows with respect to the present methodologies or other various clustering methods.
- Interpretation: The clustering results are combined here with other technologies such as classification for providing conclusions and providing future analysis.

a. Partitional based clustering: There are certain partitions generated by the partitional clustering, for a given database of certain number of objects. A clustering criterion is followed by each cluster defined here. The criterion can be such as the reduction of the sum of squared distance from the mean within every cluster. There are chances that these clusters have groupings and also try to find out a worldwide optimum. So here, the complexity of these algorithms might arise. There are huge numbers of partitions even for a less number of objects present. The solution for this reason begins with an initial, basically random, partition. Further it proceeds by refining it in a proper manner. For the purpose of executing it for practice, various sets of initial points are made to run by partitional algorithm. All the solutions are further investigated which will lead one to the similar final partition. For the purpose of improving a certain criterion, the partitional clustering algorithms are used. At first, the values of similarity or the distance are calculated. Later the results are ordered and the one which optimizes the criterion is selected. This can result the majority of them to view as greedy-like algorithms [6].

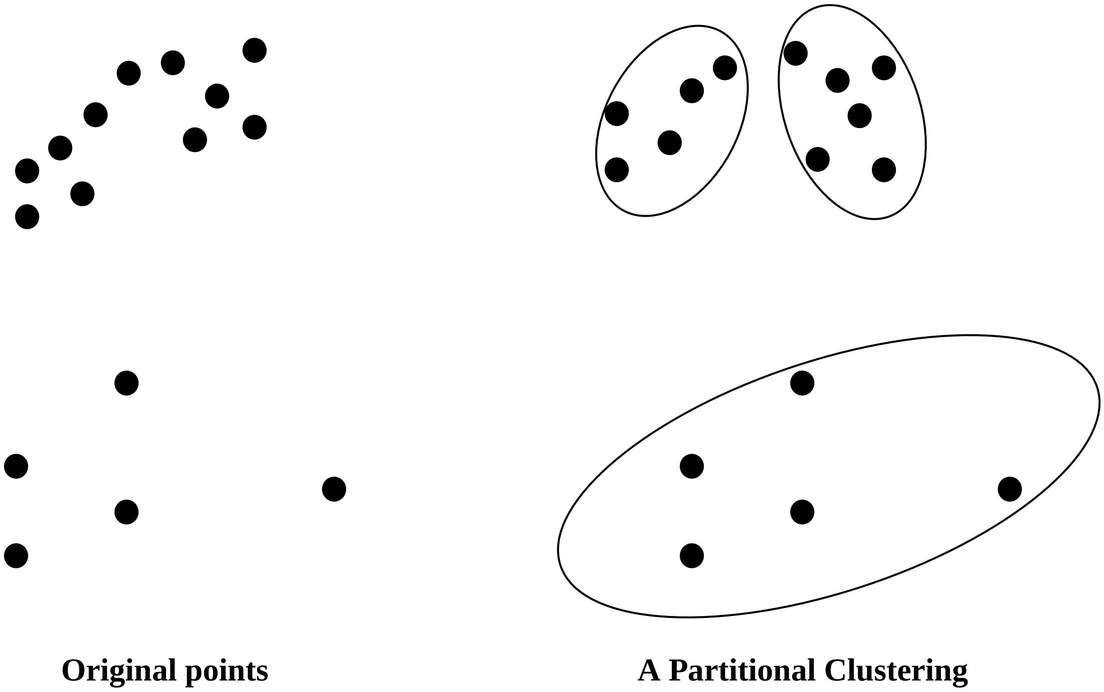


Fig 1.2.1 Partitional Clustering

A conceptual view is considered which recognizes the cluster with respect to certain model in one approach of data partitioning. This model does not have any defined parameters and its parameters are to be found yet. It is assumed by the probabilistic models that there is a blend of certain populations, the appropriation and priors of which are to be recognized. The interpretability of the constructed clusters is one major merit of the probabilistic methods. The global objective function is defined here due to the limited cluster representation which also results in inexpensive computation of intra-cluster measures. Another approach begins when the objective function depends on the partitioning of the data. The inter- as well as intra-cluster relations can be measured by the pair-wise distances or similarities. The expense of such pair-wise computations is very high during the iterative enhancements. The problem is resolved by involving unique cluster representatives who further result in linear computation of the objective function. On the basis of the way in which the representatives are formed, there are two categories for the optimization partitioning algorithms. They are k-medoids as well as k-means methods. The accurate information point present within a cluster is the K-medoid. There are two merits of this method. There are no boundaries on the types of attributes which can be considered as a benefit here. The second is the selection of medoids which depends on the location of the prevalent fraction of the points available within a cluster. This results in less sensitivity of the medoid towards the outliers. A centroid that is basically a mean of points within a cluster is used for representing the cluster within a k-means. A single outlier can result in affecting the complete work and so only the numerical attributes are involved. A clear geometric as well as statistical meaning is the merits of centroids [7].

b. Hierarchical based clustering: The hierarchical decomposition of objects is involved in the hierarchical abased clustering algorithms. There are two broader divisions of this type. One is the agglomerative (bottom-up) and the other is the divisive (top-down). They are further explained below:

- Agglomerative algorithms: The beginning of such algorithms is done by each object to be identified as a separate cluster. On the basis of the measurement of distances amongst them, these groups are merged gradually. When all the objects result in forming a single group, the clustering process stops. The process can also be stopped as and when the users demands. A greedy-like bottom-up approach is formed using these types of algorithms.

- Divisive algorithms: An opposite mechanism is followed by these algorithms. A group of all the objects is considered for beginning the process. Further, the groups are split and smaller groups are formed. This continues until each object falls into one cluster or into the cluster that it desires to. In each step, the data objects are divided into disjoint groups using the divisive approach. Unless all the objects fall into different clusters, the similar pattern in continued to be followed. The divide-and-conquer algorithms follow the similar approach. The major result of this approach is that once the merging or splitting of a cluster is done, it cannot be undone.

Merits of hierarchical clustering are: *ƒ*

- With respect to the level of granularity, the embedded flexibility is given *ƒ*
- The ease in which the various forms of similarity or distances can be handled.
- The application of these algorithms to any of the types of attributes.

Demerits of hierarchical clustering are:

- The termination criterion is not properly defined.
- For the improvement of clusters, the hierarchical algorithms do not revisit the clusters that are once created [8].

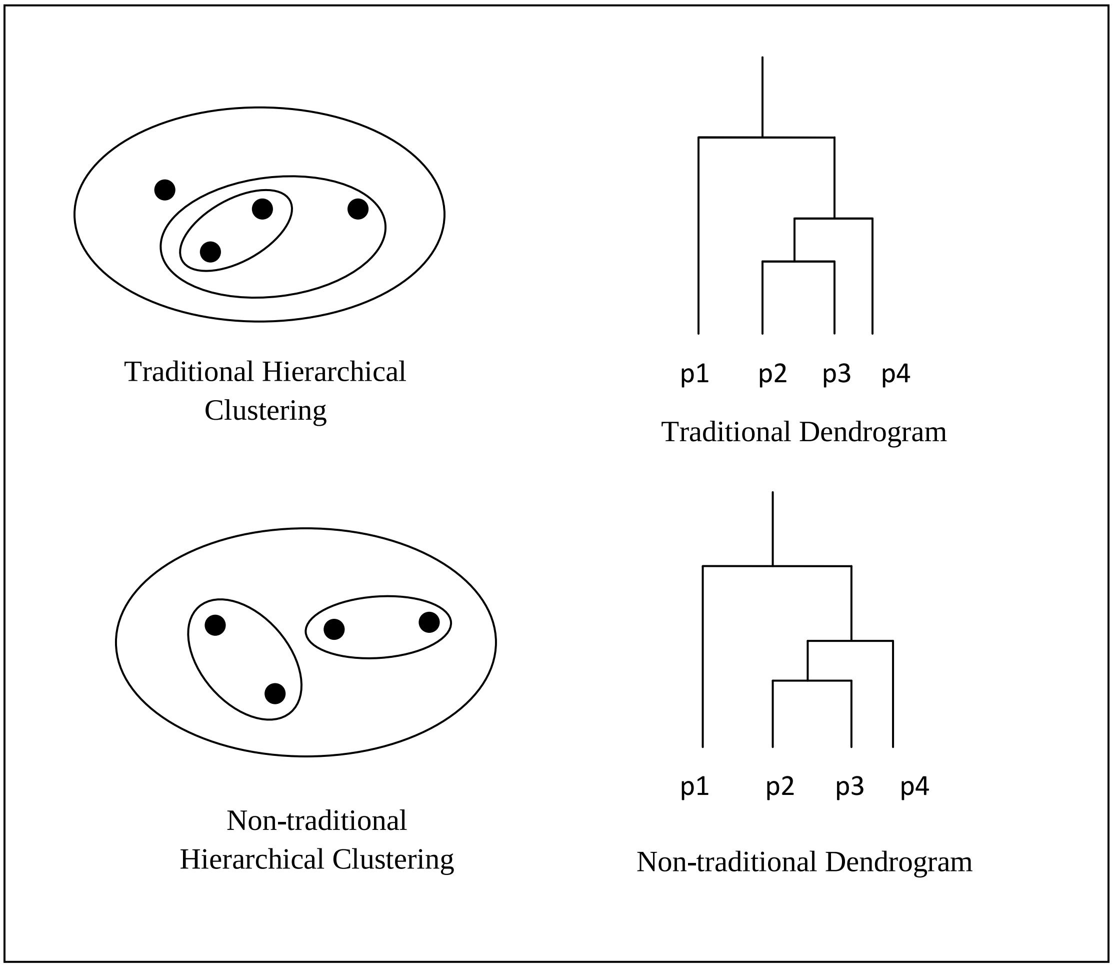

Traditional Hierarchical Clustering
p1
p2
p3
p4
Traditional Dendrogram
Non-traditional Hierarchical Clustering
p1
p2
p3
p4
Non-traditional Dendrogram

Fig 1.2.2 Hierarchical Clustering

1.3 Data Mining Clustering Techniques

For the purpose of clustering of data various techniques have been evolved apart from partitional and hierarchical clustering algorithms. On the basis of various data sets present, the various clustering techniques are to be implemented. The various methods are [9]:

a. Density-based Clustering: On the basic of certain density objective functions, the objects are grouped by these algorithms. The number of objects present in the neighbor of a data object is known as the density of that particular data object. According to the increment in the number of objects as per certain parameter, there is growth found in the given cluster.

The methodology used in the partitional algorithms is different from the ones applied here as the partitional algorithms use the relocation of points for a fixed number of clusters. Within the set of connected components, the Euclidean space of an open set is divided. There is a need of density, connectivity and boundary for partitioning of a fixed set of points during their implementation. A point's nearest neighbor is closely in relation to it. There is a growth of a cluster, also known as a dense component towards any direction in which the density leads. The clusters for arbitrary shapes are discovered by the density-based algorithms. This also helps in providing authentication against the outliers. The various clusters have problems related to the partitioning relocation clustering which can be solved by the density-based algorithms. Good scalability is also a property identified here. There are various problems also identified along with the advantages of these properties. There is need for a metric space within the density-based algorithms. Spatial data clustering in the natural setting provided for them. A metric space is required by the density-based algorithms and the natural setting here is done by the spatial data clustering. For the purpose of making computations feasible there is a need to construct the index of data. Only through reasonable low-dimensional data the classic indices were seen to be efficient. A part of density-based clustering and the grid-based preprocessing is known as the DENCLUE which is less affected by the dimensionality of the data. There are certain categories in which the density-based methods are classified. One of these is the one in which the training data point is pinned to the density. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD [10].

b. Grid-Based Clustering: The spatial data is the main highlight of these types of clustering algorithms. The spatial data is the one in which the geometric structure of articles in space, their connections, properties as well as their operations are provided. The quantization of information set into number of cells is the main aim of these algorithms. These cells that are created here work along with the objects which belong to these cells. The relocation of pints is not done here. Several hierarchical levels of groups of objects are created. These algorithms are thus in a way similar to hierarchical algorithms. However, the distance measuring factor is not the reason for selection of the merging of grids, and the clusters. The pre-defined parameter is used for deciding these parameters. Information parceling is instigated by focuses' participation in sections came about because of space dividing, while space apportioning depends on lattice qualities amassed from info information. One preferred standpoint of this backhanded taking care of (information lattice information, space-apportioning, information parceling) is that amassing of matrix information makes network based grouping methods autonomous of information requesting. Conversely, migration techniques and every single incremental calculation are exceptionally delicate as for information requesting. While thickness based parceling techniques work best with numerical traits, framework based strategies work with properties of various sorts. To some degree, the network based philosophy mirrors a specialized perspective. The class is mixed: it contains both parceling and various leveled calculations. The calculation DENCLUE from the past area utilizes lattices at its underlying stage.

c. Model-Based Clustering: Some better approximations of model parameters that best fit the data are searched with the help of this algorithm. On the basis of the structure or model they might have about the data set as well as the method through which they refine the model to recognize the partitioning, they can be either partitional or hierarchical. For the purpose of improving the preconceived model, various clusters are grown due to their nearness towards the density-based algorithms. The beginning is done by selecting only fixed number of clusters and also the similar concept of density is not utilized.

d. Categorical Data Clustering: The data on which the Euclidean and numerical distance measures cannot be used, these algorithms are used on that data. In the categorical data concept, the finite set of elements also known as items are related to variable size transaction from the common item universe. There are various types of representations of data available. The data which includes point-by-attribute format along with the transaction of binary attributes indicates that transaction is possible or not. There are very less items which are common and the representation is also sparse here. There are also many other examples which include point-by-attribute format for categorical information, critical measure of zero qualities and various others. The similarity based measures for conventional clustering methods are improper [11].

e. Constraint-Based Clustering: There are various unconstrained solutions which are used for real-world applications for customers. There are various problem-specific limitations within the clusters which make specific business actions. There are various individual objects and parameter values on which the taxonomy of clustering constraints is involved. These values can be denoted through preprocessing. For each cluster the taxonomy is also included which can provide constraints on individual clusters is terms of aggregate functions. There is a need of a new methodology for which all these constraints are necessary. For a specific subset of each cluster, the count of objects from an existing bound constraint is required. Within the partition clustering, the iterative optimization is used which depends on the mobility of objects to their nearest cluster representatives. The constraint might deplete due to this reason. A new technique is proposed related to how to resolve this conflict. The bounding of number of clusters from below is the very common necessity. There are only very few clusters provided by the k-means algorithm which is used very commonly. There is a need to modify or enhance the k-means objective function as well as their updates which hold the lower limits on cluster volumes. Along with the coefficients related to linear program needs, the soft tasks of data points are also involved. The researchers have also shown the need to enhance the k-means algorithm. An isotropic Gaussian mixture is related to here, which also has widths that are reversely proportional to the number of points present in the clusters. This results in achieving a frequency sensitive k-means algorithm. For the purpose of creating balanced clusters, another methodology is to convert certain task into graph partitioning issue. When there are any obstacles present, the major limitation based grouping application clusters 2D spatial data. For selecting the obstacle distance, the length of the shortest path among two points is used and not the Euclidean distance [12].

1.4 K-Means Clustering Algorithm

One of the measures which are used to solve the clustering problems is the k-means clustering algorithm. This is the easiest unsupervised learning algorithm. Certain numbers of clusters which include a fixed apriori are to be given classify provided data set using certain number of clusters which is simple and easy in this procedure. The objective here is basically to introduce k centers for each cluster present. A very careful placement of these centers is required. This is due to the fact that various results can be achieved as per the location variations. Due to this fact, the centers are to be placed far from each other. The next step involves selecting every point which is associated to a given data set and then relates it to the closest center. Once there is no point left, the completion of the initial step is done. During this stage, the clusters achieved from the previous steps are processed for calculating their k new centroids as barycenter. Once the k new centroids are achieved, the same data set points and the nearest new center are bound together. This result is generating a loop. The location of k centers is changed according to each step which is mainly due to this loop. This continues until no more changes are left which can also be seen in a way that there is no movement of the centers.

For the purpose of clustering analysis, the K-means is used as a simple learning algorithm. The main aim here is to identify the best division of n entities in k groups such that the total distance among the group's members and their related centroid is reduced. The centroids are the representatives of the group. The n entities are partitioned into k sets S_i, i=1, 2, ..., k. The within cluster sum of squares (WCSS) is reduced here. It is defined as

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_j^i - c_j\right\|^2$$

where term $\left\|x_j^i - c_j\right\|^2$ gives the distance between an entity point and the cluster's centroid.

An iterative refinement method is utilized here which follows certain steps. These steps are given below [13]:

1. The first step involves defining the centroids of initial groups. Various approaches can be used to perform this step. A commonly used way is to provide randomly values for centroids placed in each group. The utilization of values of K distinct entities as centroids is an anther way which is utilized.

2. The cluster that is closest to the centroid is assigned an entity. The algorithm calculates the distance among all the entities and centroid for finding the cluster which has most similar centroid.

3. The value of centroids are recalculated. The updating of the values of centroid fields is done. This is done by taking the average values of the properties of the entities which are included in a cluster.

4. Until the entities do not change the groups the 2 and 3 steps are to be repeated.

The K-Means is a greedy, moreover computationally efficient technique and is growing popularity as a representative-based clustering algorithm.

Advantages

- The k-mean algorithm is fast and robust. It also is very easy to understand.
- This algorithm is efficient relatively as compared to other algorithms when calculated through certain defined parameters.
- In the case where the data set are different or separated from each other, the algorithm given the best results.

Disadvantages

- Apriori specification is required for the number of cluster centers present in the learning algorithm.
- The major factors can be weighted unequally using the Euclidean distance measures.
- The local optima of the squared error function are given by this algorithm.
- Required results cannot be achieved by selecting cluster center randomly.
- The noisy data and outliers are not able to be handled here.
- In case where non-linear data set is provided, the algorithm results in causing failure [14].

1.5 DBSCAN algorithm

There are various types of clustering techniques present in data mining which involve partitioning, hierarchical, density, grid, model and constraint based clustering. On the density based parameters, the density based clustering algorithm is applicable. The regions which are different from a thin region are formed as thick regions or areas. Until the density in the neighbors rises above certain threshold, the identified cluster is increased here. DBSCAN (Density Based Spatial Clustering of Applications with Noise) DBSCAN is a well known density based clustering algorithm. This algorithm identifies the arbitrary shaped clusters which also includes separating the noise from large spatial databases. There are two parameters which are accepted by it which are Eps (radius) and MinPts (minimum points-a threshold). The numbers of points within a specific radius Eps are counted for the purpose of estimating the density at a specific point of the data set. This is known as a center-based approach which is applied here. There are various points which are classified in this approach in the categories such as core point, border point and noise points. The important step here is to provide less number of points (MinPts) for the neighbourhood of a given radius (Eps) of each point of a cluster. In other words, the density of the neighbor is to be more than the predefined threshold value. There are three input parameters of this algorithm:

- k, the neighbour size;
- Eps, the radius that delimitate the neighbourhood area of a point (Epsneighbourhood);
- MinPts, the minimum number of points which exist in the Eps-neighbourhood.

There are various parameters on which the clustering process depends. The arrangement of points in the information set as core points, border points and noise points as well as the usage of density relations among the points are involved [15].

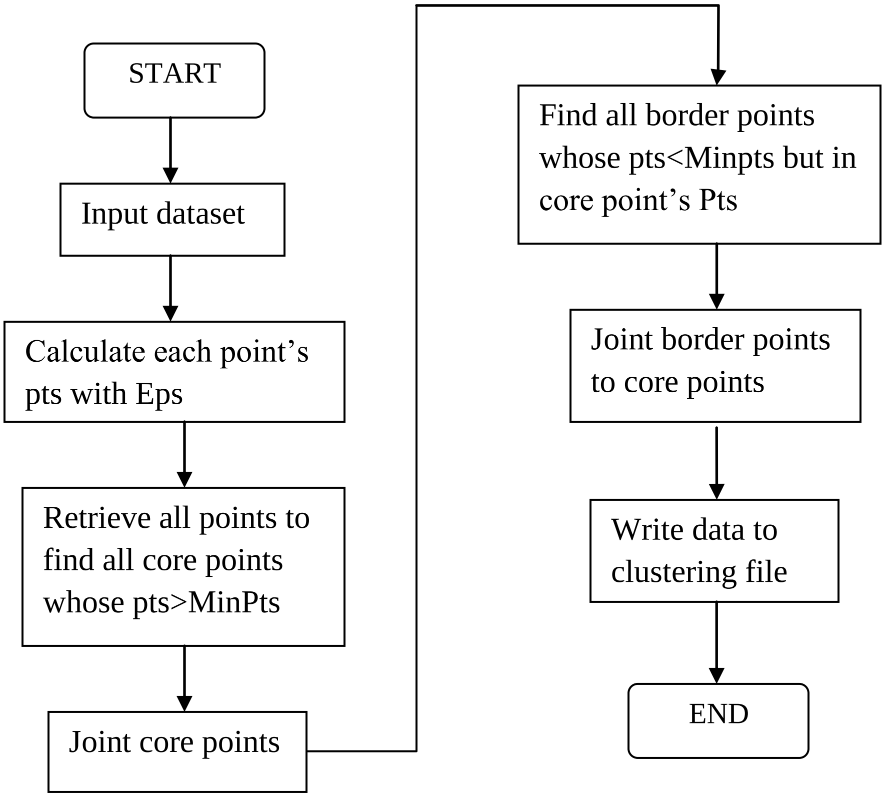

START
Input dataset
Calculate each point's pts with Eps
Retrieve all points to find all core points whose pts>MinPts
Joint core points
Find all border points whose pts<Minpts but in core point's Pts
Joint border points to core points
Write data to clustering file
END

Fig.1.5.1 Flowchart for DBSCAN algorithm

1.5.1 Complexity of DBSCAN

Every point of a database is visited numerous times by the DBSCAN algorithm (ex. Candidates to various clusters). The time complexity of the algorithm is calculated by the number of regionQuery invocations as per the practical algorithms are involved. For each point, one query is executed by the DBSCAN. A neighborhood query is executed as O(log n) when the indexing structure is utilized. An inclusive average runtime complexity of O(nlogn) is obtained. The worst case run time complexity is $O(n^2)$ which occurs when the accelerating index structure is not utilized or when the degenerated data is involved [16].

Advantages

1. The number of clusters needs not to be specified in the data apriori in the case of DBSCAN which is different from that of k-means.

2. The arbitrarily shaped clusters are identified by the DBSCAN. A cluster which is completely surrounded by other cluster is also identified here. The single-link effect is reduced with the help of MinPts parameter.

3. The notion of noise is involved in the DBSCAN. It is also robust to outliers present in the data set.

4. There are only two parameters required in the DBSCAN. The ordering of points within a database is insensitive in nature.

5. The databases which can accelerate the region queries are designed with the help of DBSCAN.

6. If the data is well understood, the domain expert can be used for setting the minPts and ε parameters [17].

Disadvantages

- The DBSCAN is not deterministic completely. On the basis of the order in which the data is processed, the border points which are near to the numerable clusters are involved as a part of any one cluster. However, not all the situations involve this type of issue to arise. Here, the DBSCAN is deterministic. Here, the border points are considered as noise and the completely deterministic result is achieved through this method.
- The distance measure which is used in the regionQuery(P,ε) function is utilized for calculating the quality of DBSCAN. Euclidean distance is the most common distance metric involved.
- As there are huge difference is densities, the DBSCAN is not able to cluster data. The combination of minPts-ε cannot be selected in an appropriate manner for all the clusters present.

Chapter 2

Literature Review

Ahmad M. Bakr, et.al," Efficient incremental density-based algorithm for clustering large datasets", 2015. An improvement in the incremental DBSCAN algorithm is proposed for building and updating shaped clusters in huge datasets in an incremented manner in this paper. An incremental clustering process is enhanced using this proposed algorithm[19] .The search space is limited to partition instead of the whole dataset which further provides certain enhancements in the performance. The enhancements can be seen when this method is compared to the other similar incremental clustering algorithms. According to the experimental results it is seen that when the proposed algorithm is compared with already existing incremental algorithms when being used on various size and dimensions of datasets, there is an increase in runtime speed of the incremental clustering process by factor up to 3.2. The accuracy of the algorithm has also improved here and has shown better results when implemented on large datasets with higher dimensions. There are other enhancements also which are to be proposed in the future work. To make an algorithm efficient enough to be working in a parallel manner is also to be proposed. The incremental DBSCAN algorithm can be applied here within each partition in parallel manner. This can only be done when the independence of partitions is seen. The parallel version of the newly proposed algorithm has shown enhancements which are better than thte already existing ones.

Iyer Aurobind Venkatkumar, et.al," Comparative study of Data Mining Clustering algorithms", 2016. There is a large amount of data generated in today's era. There are certain advantages of the hidden information within the data which include the patterns as well as the correlations of data. There are many constructive fields in which this information can be beneficial especially in the cases where huge data are involved. Data mining is one such field where there is a requirement of handling such huge amount of data in a proper manner. The techniques such as clustering, prediction, classification, association, and so on are involved in the data mining. The process of dividing the data set into related groups in such a manner that no two groups hold common set properties is known as clustering. According to the present data set, the predictions are set which is known as the prediction technique [20]. However, the prediction provided here might not be correct for sure and so no guarantee is provided by it. With the help of certain mathematical models, the data sets are classified into certain predefined sets. This technique is known as classification. Within the large amount of data given, the hidden correlations are determined with the help of association technique. Further, on the basis of these relationships within the objects certain pattern is recognized within the transaction. Certain data mining clustering algorithms are also studied within this paper according to their merits and demerits when used against certain data. There is a proper comparative analysis made on these four classic clustering algorithms which are k-means, BIRCH, DBSCAN, and STING. Certain outcomes provided by them highlight their properties.

Qi Xianting, et.al," A density-based clustering algorithm for huge-dimensional information with feature selection", 2016. A sub-part of the density-based representative algorithms is the density-based spatial clustering of applications with noise (DBSCAN) which has been utilized in certain fields because of its property of detecting the clusters which are of various shapes as well as sizes. When high dimensional data is present in certain applications that is when the algorithm stays no more stable. For the purpose of resolving this issue, an enhanced DBSCAN algorithm which is based on feature selection (FS-DBSCAN) is put forth [21]. This algorithm is provided on various real world datasets and the various series of simulations are achieved. This helps in testing the performance of this newly proposed algorithm. The results depict that this new algorithm is more efficient as compared to the already existing ones. The high-dimensional data also is very easy to deal with through this algorithm proposed.

Kuan-Teng Liao, et.al," An Effective Clustering Mechanism for Uncertain information extraction Using Centroid Boundary in UKmeans", 2016. There is uncertain data clustering also present within the applications which results in causing errors. Due to these errors, the time cost as well as the effectiveness of the system is affected gradually. The time cost needs to be reduced and the effectiveness needs to be increased for providing a proper clustering algorithm. In this paper the centroid based clustering and the UKmeans algorithm is proposed [22]. The similarity of the application is improved through the first mechanism. The time cost and effectiveness are affected through this similarity factor. For instance, the time cost is ignored by the similarity calculations along with a special attention on the effectiveness of the clustering. In contrast, the cost time issue is a major concern for the similarity calculations along with simplified approaches while the effectiveness property is ignored. So, for taking time cost and effectiveness measures as a major concern equally, an enhancement is proposed. A simplified similarity is utilized to reduce the cost time and add additional two factors which are density of clusters and intersection. These two factors will further help in increasing the effectiveness of the clustering mechanism. During the overlapping of a cluster on the object, the degree of the object belongingness is increased with the help of the intersection factor. The range can be decreased here by providing square root boundary method for limiting the upper bound of positions of centroids which will further help in increasing the effectiveness of clustering. It is seen through the experimental results that the mechanisms provide better results in terms of time cost and effectiveness.

Cheng-Fa Tsai, et.al," A Newly Data Clustering Approach for information extraction in Large Databases", 2002. A new data clustering method is proposed in this paper which also includes the data mining to be performed in applications which involve huge databases [23]. There are many issues which arise during the clustering with respect to many aspects. There is a need of providing proper data analysis to reflect the broader appeal of it. The main objective of clustering techniques is to partition a set of data points into classes in such a manner that properties of points within a similar class are different from the ones belonging to other classes. There are various experiments conducted and the simulation results achieved show that the newly proposed clustering algorithm is efficient as compared to the Fast SOM combined with K-means algorithm as well as the genetic k-means algorithms. The errors given by this algorithm are also very less as compared to the other methods. There are three enlisted methodologies proposed by the ACODF algorithm. The first is to utilize ACO for the purpose of solving the clustering issue. The second is to adopt simulation based method for the ant in order of reducing the amount of cities which will further result in giving optimal results. The third one is to use tournament selection strategy for the purpose of selection a path. The ACODF method is made to compare with the FSOM+K-means approach and GKA which shows that the accurate results are achieved in the case of high dimensional datasets.

Wenbin Wu et.al," A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting",2016. For this purpose of enhancing the forecasting accuracy and handling the training sample dynamics, a new approach has been proposed in this paper. In this approach, the k-means clustering algorithm is used along with the neural network for the cases where short term WPF is involved [24]. The samples are classified into various categories on the basis of the similarities provided from earlier concepts which involve the k-means clustering. These categories hold the information of meteorological conditions and historical power data. For resolving the over-fitting and instability problems involved in conventional networks, the integration of bagging-based ensemble approach is done into the back propagation neural network. On the basis of real wind generated data traces, the effectiveness of the method is ensured by performing certain tasks on it. As compared to the baseline and previous short-term WPF methods, the simulation results of this proposed algorithm are more efficient. There should be a proposed research related to the effective meteorological forecasting which will help in enhancing the forecasting accuracy. There should be a design proposed for the relative optimal method which involves the BDNN method.

Vadlana Baby, et.al," Distributed threshold k-means clustering for privacy preserving information extraction",2016. An effective distributed threshold privacy-preserving k-means clustering algorithm is proposed in this paper. In this method, the code based threshold secret sharing is utilized as a privacy-preserving method. There is a code based approach involved here which allows the division of data into various shares which is further processed at various servers [25]. There is less number of iterations in the newly proposed protocol as compared to the previous ones. There is no trust required from the end of servers or users. There are certain comparisons made with respect to various techniques. The security analysis of this proposed method is also given here. The code based threshold secret sharing scheme along with secure addition and comparison protocols is utilized for privacy preserving k-means clustering algorithm. The clustering mechanism is performed in a collaborative manner and the third party's trust is avoided. A perfect preservation of the user data is provided through this newly proposed method.

KM Archana Patel et.al," The Best Clustering Algorithms in Data Mining",2016. The most genuinely utilized unsupervised learning method in the case of data mining is the clustering mechanism. The similar data objects are located within same clusters based on any particular type of similarity amongst them. There are seven various groups in which the clustering algorithms are categorized. As per their conditions, various results are given by these different clustering algorithms. There are certain techniques which help in clustering data present in huge data sets. There are other techniques which provide better results for the purpose of finding cluster with arbitrary shapes. In this paper, various data mining clustering algorithms are learned and related [26]. There are some clustering algorithms such as k-means algorithm, k-medoids, and distributed k-means clustering algorithm which are discussed. There are various factors which are kept as base for providing comparisons in between these algorithms. There are certain specifications which are enlisted after comparisons of these algorithms which define which algorithm will be beneficial at certain conditions. The clustering algorithms are to known well for providing better results. There is therefore, no such algorithm which can be applied at all different scenarios.

ZHANG Ke, et.al" An Algorithm to Adaptive Determination of Density Threshold for Density-based Clustering",2016. For the purpose of classifying huge data whose prior knowledge is not known, the density-based clustering algorithm is utilized. The main highlight of this paper is to measure the distance of dense regions which are adaptive in nature. In this paper a novel density-based clustering algorithm is proposed which is based on the basis of determination of density threshold [27]. This involves the translation of a density threshold selection issue into a determination issue related to sperality radius. The partial cluster is known as the radius threshold which is comprised of one object. There are various partial clusters created within the dense regions which are based on the analysis of radius threshold range from the dense transitive closure. The regions which are divided are also to be decided as to be declared as clusters or not. On the basis of the present datasets, the clustering results are given. In this paper, the future implications to be provided are also discussed. The proposed technique has resulted in providing better outcomes as compared to the previous methods. The original cluster methods are also enhanced here using the distance measurements.

Guangchun Luo, et.al," A Parallel DBSCAN Algorithm On the Basis Of Spark", 2016. With the tremendous growing of data, the method has entered the period of big data. With a specific end goal to filter through masses of data, various information mining calculations using parallelization are being actualized [28]. Group investigation involves an essential position in information mining, and the DBSCAN calculation is a champion among the most comprehensively used calculations for bunching. In any case, when the current parallel DBSCAN calculations make information parcels, the first database is typically isolated into a few disjoint allotments; with the expansion in information measurement, the part and solidification of high-dimensional space will expend a ton of time. To tackle the issue, this paper proposes a parallel DBSCAN calculation (S_DBSCAN) in view of Spark, which can rapidly understand the parcel of the first information and the blend of the bunching comes about. It is separated into the accompanying steps: 1) parceling the crude information in view of an irregular example, 2) registering nearby DBSCAN calculations in parallel, 3) combining the information allotments in light of the centroid. Contrasted and the customary DBSCAN calculation, the trial result exhibits the proposed S_DBSCAN calculation gives better working effectiveness and adaptability. This paper assesses the S_DBSCAN calculation by managing yearly outpatient information. The trial result exhibits the proposed S_DBSCAN calculation can viably; and productively; creates bunches and recognize clamor information. So, the S_DBSCAN calculation has predominant execution when managing huge information, when contrasted with existing parallel DBSCAN calculations.

Dianwei Han, et.al," A novel scalable DBSCAN algorithm with Spark", 2016. DBSCAN is an outstanding clustering algorithm that is based on density and can identify arbitrary shaped clusters and also eliminates noise data. Be that as it may, parallelization of DBSCAN is a testing work on the grounds that based on MPI or OpenMP environments, there exist the issues of absence of adaptation to non-critical failure and there is no assurance that workload is adjusted. Also, programming with MPI requires information researchers to have a propelled involvement to deal with correspondence between hubs which is a major test [29]. DBSCAN algorithm has been extremely famous since it can identify arbitrary shaped clusters and additionally handle noise data. Be that as it may, parallelization of DBSCAN based on MPI and OpenMP acognize from lacking of fault–tolerance. Also, for implementation of parallelization with MPI or OpenMP, data scientists need to deal with implementation in detailing, for example, to handle communication, dealing with synchronization, et cetera, which can represent a test for a few customers. This paper presented another Parallel DBSCAN algorithm with Spark. It maintains a strategic distance from the communication amongst executors and in this way prompts to a better scalable performance. The results of this analyses concludes that the new DBSCAN algorithm with Spark is scalable and outflanks the execution in view of MapReduce by an element of more than 10 regarding productivity.

Nagaraju S, et.al," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016. This paper present a productive approach for grouping investigation to recognize implanted and settled adjoining bunches using thought of thickness based idea of groups and neighborhood contrast. Fundamentally the proposed calculation is enhanced variant essential DBSCAN calculation, proposed to address the bunching issue with the usage worldwide thickness parameters in essential DBSCAN calculation and issue of identifying settled nearby groups in EnDBSCAN calculation. The test comes about that recommended that proposed calculation is more successful in distinguishing inserted and settled adjoining bunches thought about both DBSCAN and EnDBSCAN without including any extra computational unpredictability [30]. Furthermore the paper has preset strategy to assess the worldwide thickness parameters using sorted k-separate plot and first request subsidiary. Through this paper the idea of thickness based methodologies for information bunching and considered neighborhood distinction is used adequately distinguish installed and settled contiguous groups. Our exploratory outcomes recommended that proposed calculation powerful in identifying settled contiguous groups contrasted with DBSCAN and EnDBSCAN calculation with computational intricacy as same as DBSCAN calculation.

Jianbing Shen, et.al," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016. This paper proposes a constant picture superpixel division strategy with 50fps by using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) calculation. Keeping in mind the end goal to diminish the computational expenses of superpixel calculations, the strategy got a brisk two-organize structure. In the main bunching stage, the DBSCAN calculation with shading similitude and geometric restrictions is used to rapidly group the pixels, and a while later little bunches are converged into superpixels by their neighborhood through a separation estimation characterized by shading and spatial elements in the second blending stage [31]. A powerful and clear separation capacity is characterized for improving superpixels in these two phases. The test comes about exhibit that our continuous superpixel calculation (50fps) by the DBSCAN grouping beats the cutting edge superpixel division strategies as far as both exactness and proficiency. The paper proposed a novel picture superpixel division calculation using DBSCAN grouping as a piece of this paper. The DBSCAN superpixel division calculation produces normal formed superpixels in 50fps. Our proposed superpixel division initially creates the underlying superpixel comes about with the comparative hues by playing out the DBSCAN grouping calculation, and after that joins the little introductory superpixels with their closest neighbor superpixels by considering their shading and spatial data. Assessment was led on the general population Berkeley Segmentation Database with using three assessment measurements. This calculation accomplishes the best in class execution at an extensively littler estimation cost, and essentially outflanks the calculations that require more computational expenses notwithstanding for the photos including complex items or complex surface areas.

Ilias K. Savvas, et.al," Parallelizing DBSCAN Algorithm Using MPI", 2016. The latest years, gigantic packs of data are removed by computational systems and electronic devices. To misuse the inferred measure of information, new inventive calculations must be utilized or the set up ones may be changed. A champion among the most entrancing and beneficial strategies, with a specific end goal to find and concentrate data from information storage facilities is grouping, and DBSCAN is a fruitful thickness based calculation which bunches information agreeing its qualities [32]. In any case, its key weight is its serious computational multifaceted nature which demonstrates the method astoundingly deficient to apply on huge datasets. Notwithstanding the way that DBSCAN is an outstandingly especially contemplated procedure, a totally operational parallel rendition of it, has not been acknowledged yet by standard scientists. In this work, a three stage parallel adaptation of DBSCAN is exhibited. The acquired test results are particularly encouraging and exhibit the accuracy, the versatility, and the viability of the method. In this work, a parallel form of the eminent DBSCAN was introduced and executed using MPI. The outcomes acquired from different solid cases demonstrated that were indistinguishable with the outcomes conveyed by the utilization of the first consecutive strategy. The time intricacy diminished significantly and the exploratory outcomes showed that the calculation scales in a particularly effective way.

Ahmad M. Bakr, et.al," Efficient incremental density-based algorithm for clustering large datasets", 2014. 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 [33]. 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. The algorithm incrementally partitions the dataset to reduce the search space to every partition as opposed to filtering the whole dataset. After that the algorithm incrementally forms and updates dense regions in every partition. Following identifying possible dense regions in every partition, the algorithm utilizes an inter-connectivity measure to merge dense regions to shape the final number of clusters. Experimental results demonstrate that the proposed algorithm has a comparable accuracy compared to related incremental clustering 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," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015. Information mining procedure is to get data from an informational index and after that change over it into justifiable and significant data for further use. DBSCAN, a thickness based grouping calculation, distinguishes bunches of moving shape and exceptions. DBSCAN depends on bivalent rationale. Subsequently it can simply identify questions as totally having a place with a specific bunch or not entirely having a place with it [34]. In this paper, a system of technique of DBSCAN calculation with the incorporation of fluffy rationale is proposed. How much a protest has a place with a specific group will be settled using participation values. The enhanced rendition of DBSCAN calculation will be the hybridization of DBSCAN calculation with fluffy if-then guidelines. To improve the level of participation, multivalent rationale will consider in which the enrollment qualities are to be used. Empower, the current procedures have not concentrated on the hybridization of DBSCAN with fluffy if then standards. This calculation will be solidified with fluffy if then principles for bosom tumor recognition. With this enhanced half and half DBSCAN calculation, certain parameters, for instance, precision, geometric exactness, bit blunder rate, detail, and affectability and mistake rate will be assessed and the outcomes will be thought about over the DBSCAN calculation. The hybridization will enable DBSCAN to pick the group in more productive way.

Karlina Khiyarin Nisa, et.al," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework",2014. Backwoods flames are a huge issue that occurs on and on in Indonesia. Fire occasions can be anticipated by checking the datasets of hotspots which are recorded through remote detecting satellite. This review arrangements to build a web application that performs bunching on the hotspots information [35]. This application actualizes DBSCAN calculation using Shiny web system for R programming dialect. Bunching is performed on a dataset of hotspots on Kalimantan Island and South Sumatra Province in 2002-2003. The spread case of hotspots came about by this bunching can be used as a prescient model of timberland flames occasion and can be gotten to through the web program. This examination built up an online application bunching with DBSCAN calculation using the R programming dialect with Shiny structure. DBSCAN needs minPts and Eps parameter. The greater estimations of minPts will make less, however more the quantity of commotions. While the greater estimation of Eps will bring about less bunches. MinPts parameter assurance is done by looking measurements of the information and plot the diagram of minPts and the quantity of bunches and commotion. While Eps parameter assurance is acquired from k-dist chart perception and the slant contrast computations.

Negar Riazifar, et.al,” Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm”,2015. Retinal vessel division used for the early finding of retinal maladies, for instance, hypertension, diabetes and glaucoma. There exist a few strategies for sectioning veins from retinal pictures. The purpose of this paper is to investigate the retinal vessel division in view of the grouping calculation DBSCAN depending on a thickness based idea of bunches which is intended to discover bunches of subjective shape. DBSCAN requires one and just info parameter and an incentive for this parameter is recommended to the customer [36]. The execution of calculation is thought about and broke down using various measures which fuse affectability and specificity. The specificity and affectability of this strategy is ٩5.36 and ٧3.82 independently. Vein division is a basic preprocessing venture for the early analysis of retinal infections. Though various changes and adjustments are proposed to explore retinal vessel division, only a bit of the calculations are important. The DBSCAN calculation takes care of each one of the issues when using grouping techniques, finds the correct information parameters, limits bunches of subjective shapes and does the entire procedures in a sensible time. The execution of the calculation in this paper is preferred as a rule over the past ones. The new arrangement diminishes the sit still time and upgrades the speed and exactness of division.

Yumian Yang, et.al,” Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm”, 2014. With the fast advancement of E-trade, how to assess the E-business destinations precisely has transformed into a basic issue. In any case, assessment record of E-trade locales has qualities of high measurements and uneven thickness, which prompts dreadful execution of the assessment result [37]. To break down 100 E-trade indicate endeavors in 2013-2014 named by the Ministry of Commerce People's Republic of China, this paper lessens dimensionality by component examination strategy initially, then executes an enhanced DBSCAN calculation to handle the uneven thickness information, at long last offers proposals to these 100 Ecommerce attempts in light of examination results. Since DBSCAN calculation disregards weights while computing the Euclidean separation, the aftereffect of the comparability estimation is not exact, while consider investigation is a decent intends to manage weights. This paper enhances the bunching precision and sensibility of the assessment by joining element investigation and DBSCAN. In any case, the information handled by element examination have qualities of uneven thickness. The customary DBSCAN is enhanced to segment the information with different densities and bunch these locales. This paper propels another preparing thought on E-business locales assessment: another DBSCAN calculation consolidating element examination with different densities. Contrasted and the conventional DBSCAN calculation, the aftereffects of assessing sites are more sensible and interpretable with the enhanced DBSCAN calculation. Later on work, the size of the assessment protest will be additionally extended and more research ought to be done.

XiaoqingYu, et.al," Explore Hot Spots of City Based on DBSCAN Algorithm", 2014. Spatial grouping is one of the standard techniques for information mining and learning disclosure. DBSCAN calculation can be found in space with "commotion" database bunching of discretionary shape, is a kind of good grouping calculation. This paper presents the fundamental idea and standard of DBSCAN calculation, and applies this calculation to perform grouping investigation dispersions of weibo area data [38]. The article differentiate k-implies calculation and DBSCAN calculation so as to exhibit the viability of DBSCAN calculation. The DBSCAN calculation will create much clamor focuses, and there are two or three information focuses in its each group. In any case, the k-implies calculation doesn't shape clamor focuses. So it can be influenced by a few focuses. This paper looks at the essential standard of DBSCAN calculation and its usage procedure. It applies this calculation in the field of city wanting to find the hot domain in the city. Furthermore, it looked at DBSCAN calculation and k-implies calculation, and show its adequacy. Later on, it can be used to break down city open workplaces or city open workplaces to give a logical introduce and direction for city arranging.

CHAPTER-3

Present Work

3.1. **Problem Formulation**

The density based technique is the type of algorithm in which density of the whole dataset is calculated and most dense region is calculated to find similarity between the elements of the dataset. In the existing work, technique of density based clustering is applied in which density of whole dataset is calculated and dense region is calculated. On the Dense region EPS value is calculated to analyze similarity between the elements. The Euclidian distance is applied to analyze similarity between the elements. The EPS is calculated in the dynamic order to achieve maximum accuracy. The Euclidian distance is calculated in the static manner due to which accuracy is not achieved at the maximum point. In this work, improvement in DBSCAN algorithm has been proposed which calculate Euclidian distance in the iterative manner to increase accuracy of clustering

3.2. **Objectives**

1. To propose the improvised DB-SCAN algorithm to increase the accuracy of clustering by studying and analyzing various density based algorithm for it.

2. The proposed improvement is based on back propagation algorithm to calculate Euclidian distance in the dynamic manner, and to scrutinize the outcomes regarding exactness and execution time.

Density based clustering is applied to calculate similarity from the most dense region which can define clusters on the basis of similar and dissimilar type of data. The Euclidian distance is calculated in the static manner which leads to reduction in accuracy of clustering. The proposed improvement will calculate Euclidian distance is dynamic manner which increase accuracy of clustering. When the Euclidian distance will be calculated in dynamic manner, it will leads to increase accuracy and also reduce execution time of the algorithm.

3.3. Research Methodology

In the DBSCAN algorithm the most dense region is calculated from the dataset. The central point is calculated from the most dense region which is the called EPS value of the dataset. To calculate similarity between the data points of the data Euclidian distance is calculated from central point to all other points. The elements which are similar is clustered in one dataset and other are in the second dataset. In the base paper, to improve accuracy of clustering EPS values is calculated in the dynamic manner which leads to the clustering of the points which are remained unclustered. The basic DBSCAN algorithm is the static algorithm which the EPS value is given by the user. The EPS value is the radius value which can be the covered in the dataset to part the larger cluster. The Euclidian distance is calculated which cluster the similar and dissimilar values from the defined radius. The incremental DBSCAN algorithm is the further improvement in the DBSCAN algorithm in which the EPS value is not given statically by the user by it the calculated according to the input dataset. To calculate EPS value according to the dataset, the most dense region is calculated from the dataset and arithmetic mean of the dense region is calculated which is the central point of the dataset. The EPS value defines the class of the input dataset. The Euclidian distance is calculated from the EPS point which cluster similar and dissimilar values from the dataset. In this work, further improvement in the incremental DBSCAN algorithm is done which calculates the Euclidian distance dynamically. 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

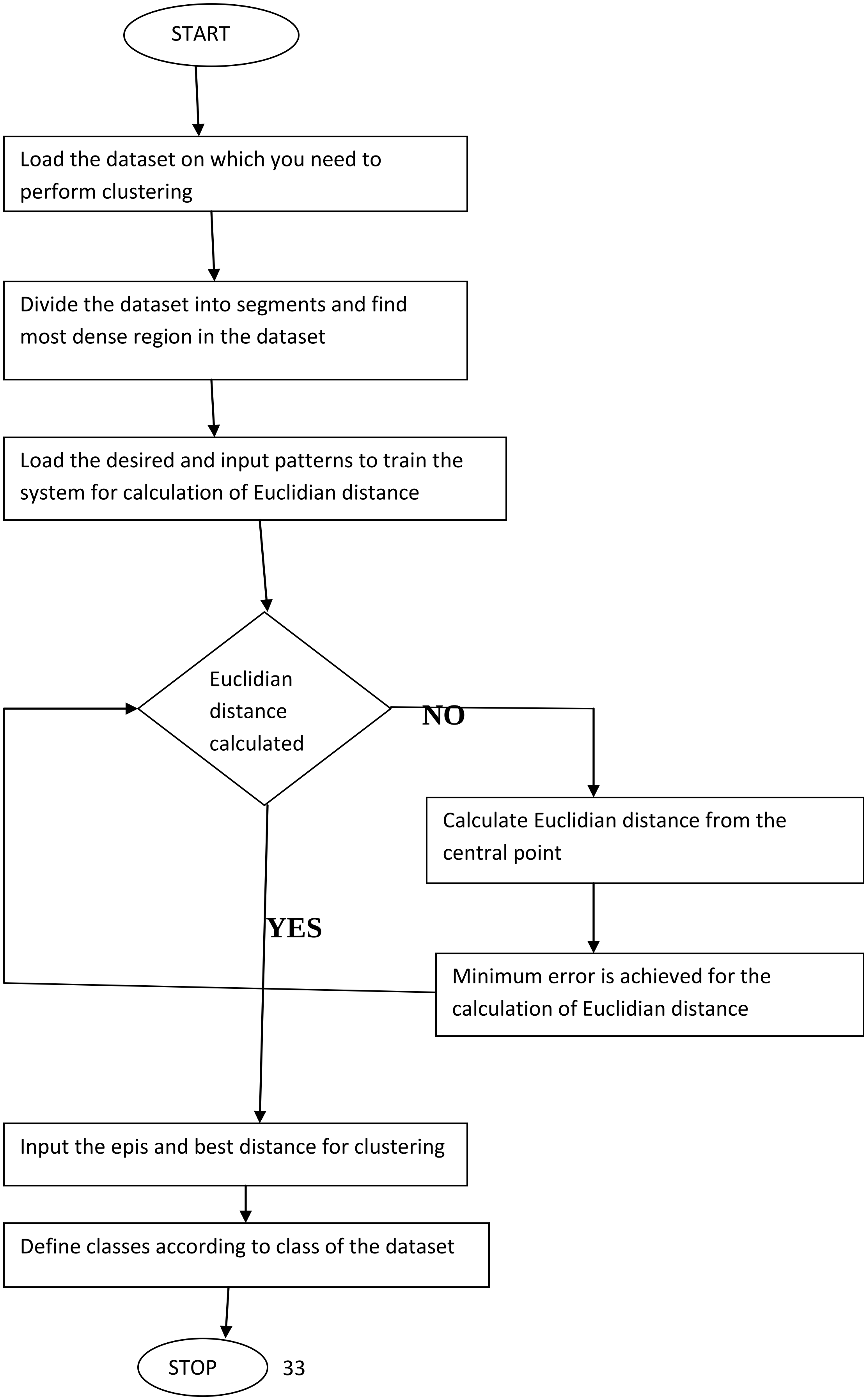
iii) Back propagation to the hidden layer

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:

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

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


START
Load the dataset on which you need to perform clustering
Divide the dataset into segments and find most dense region in the dataset
Load the desired and input patterns to train the system for calculation of Euclidian distance
Euclidian distance calculated
NO
Calculate Euclidian distance from the central point
Minimum error is achieved for the calculation of Euclidian distance
YES
Input the epis and best distance for clustering
Define classes according to class of the dataset
STOP

CHAPTER-4

Results and Discussion

4.1. Tool Description

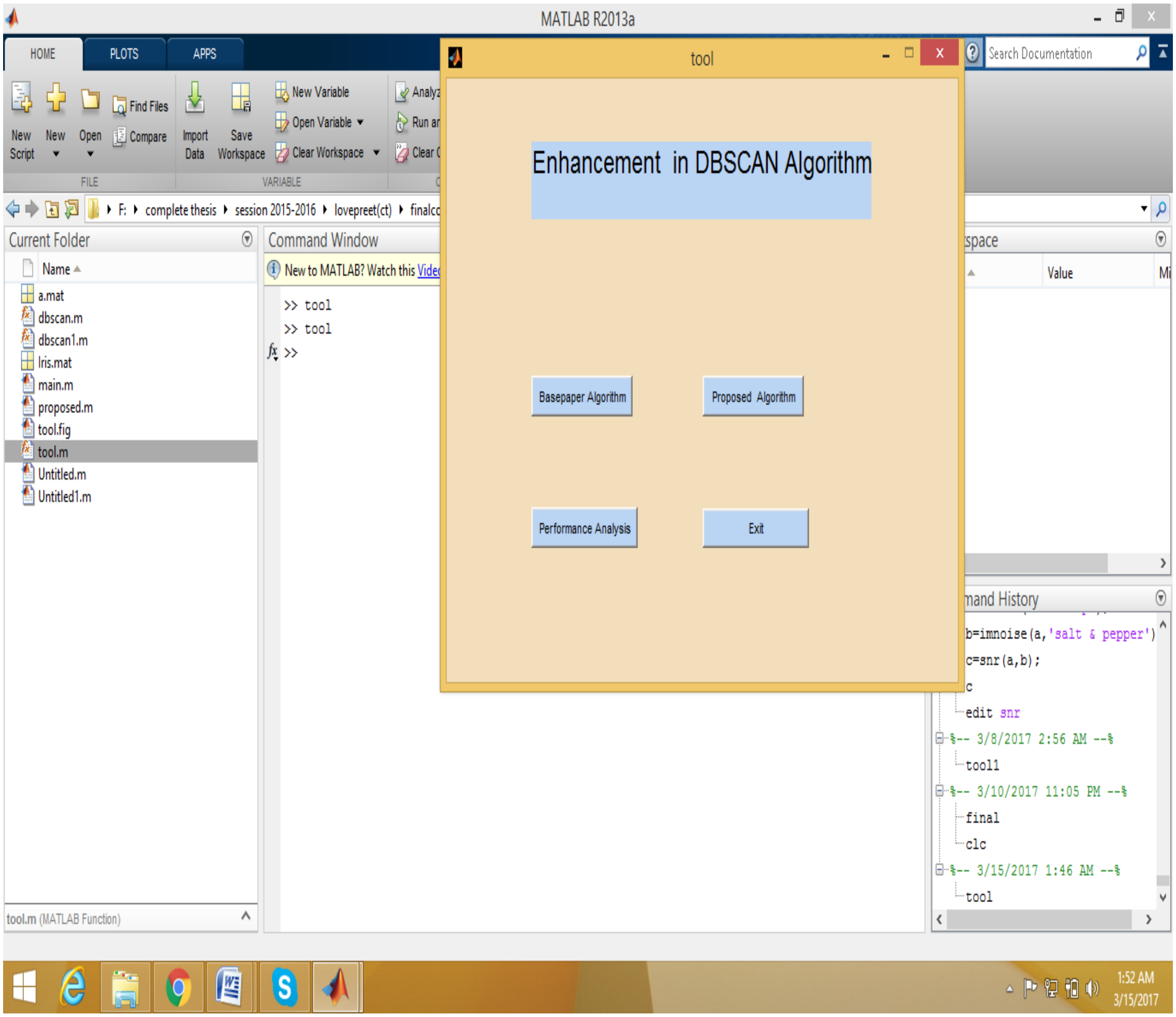
The MATLAB is the tool which is used to perform mathematical complex computations. In this MATLAB simplified C is used as the programming language. The MATLAB has various inbuilt toolboxes and these toolboxes are mathematical toolbox, drag and drop based GUI, Image processing, Neural networks etc. The 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

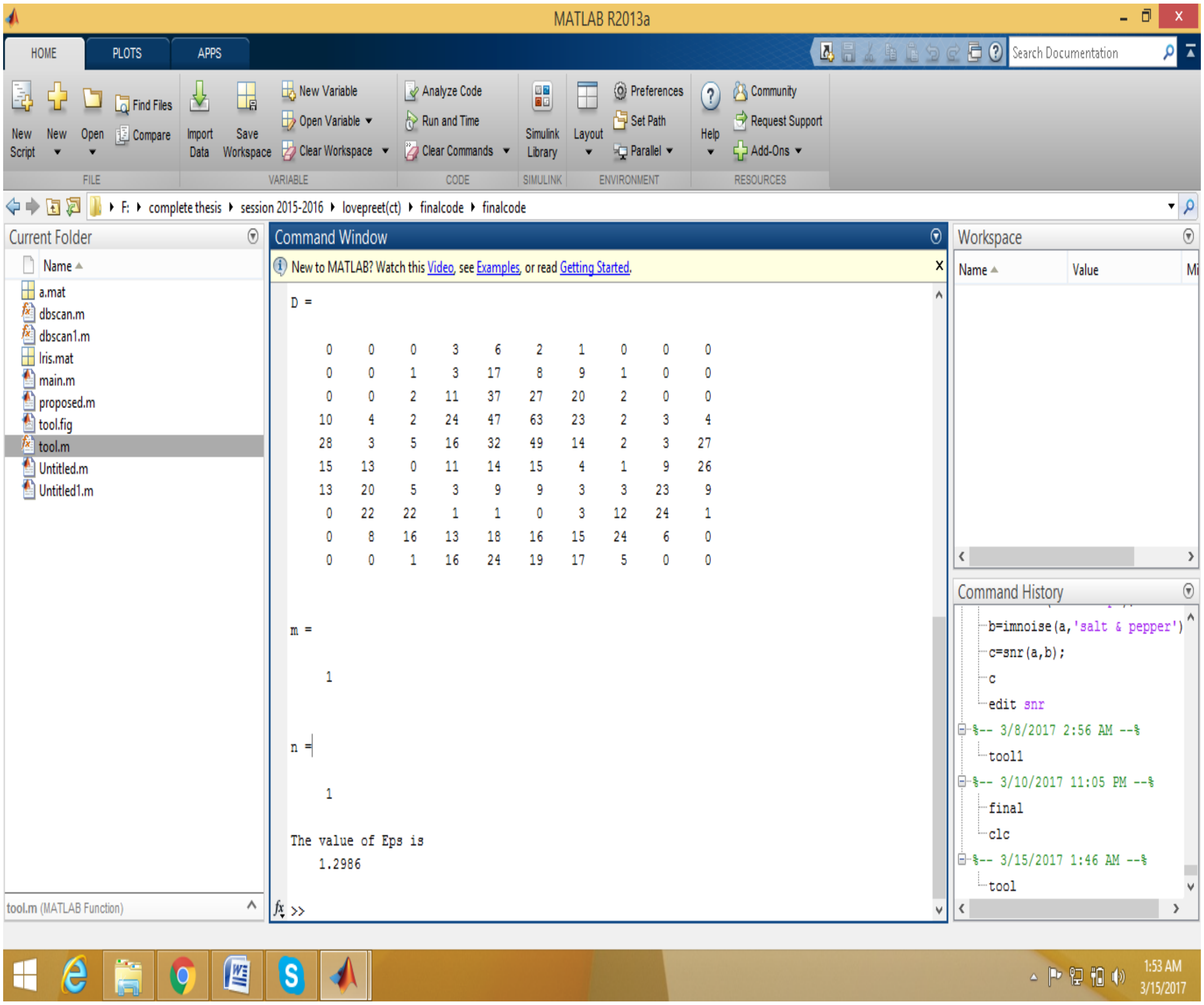


```
>> tool
>> tool
fx >>
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.1: Default Interface of Tool

As illustrated in figure 1, the tool is designed in the MATLAB which shows the execution of incremental DBSCAN algorithm and proposed DBSCAN algorithm

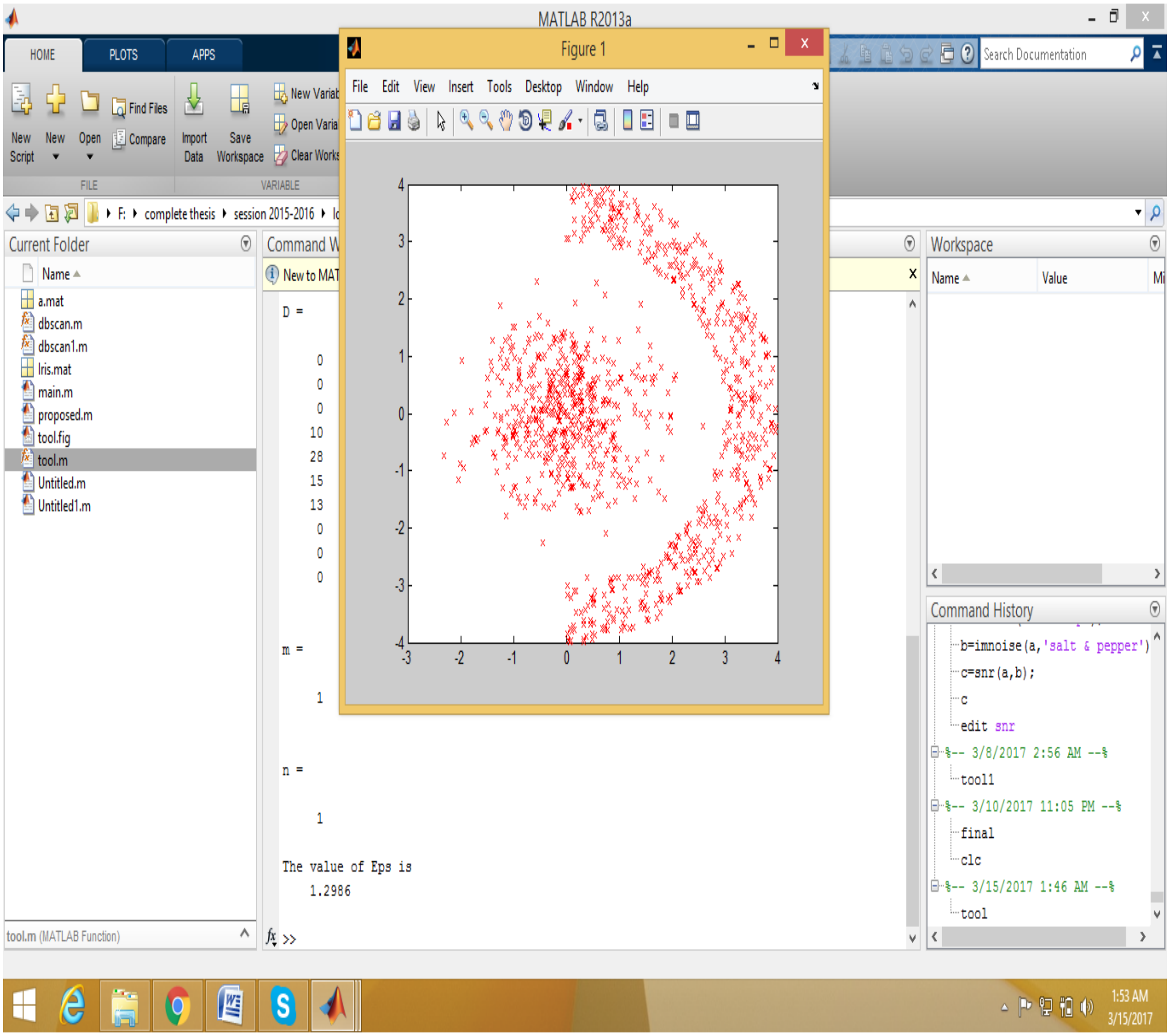


```
D =

     0     0     0     3     6     2     1     0     0     0
     0     0     1     3    17     8     9     1     0     0
     0     0     2    11    37    27    20     2     0     0
    10     4     2    24    47    63    23     2     3     4
    28     3     5    16    32    49    14     2     3    27
    15    13     0    11    14    15     4     1     9    26
    13    20     5     3     9     9     3     3    23     9
     0    22    22     1     1     0     3    12    24     1
     0     8    16    13    18    16    15    24     6     0
     0     0     1    16    24    19    17     5     0     0


m =

     1


n =

     1

The value of Eps is
    1.2986
>>
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.2: Calculation of Dense Region

As shown in figure 2, the incremental DBSCAN algorithm is implemented in which the most dense region is calculated and calculated dense region is shown in the snapshot. According to the most dense region the EPS value is calculated which define radius of the cluster



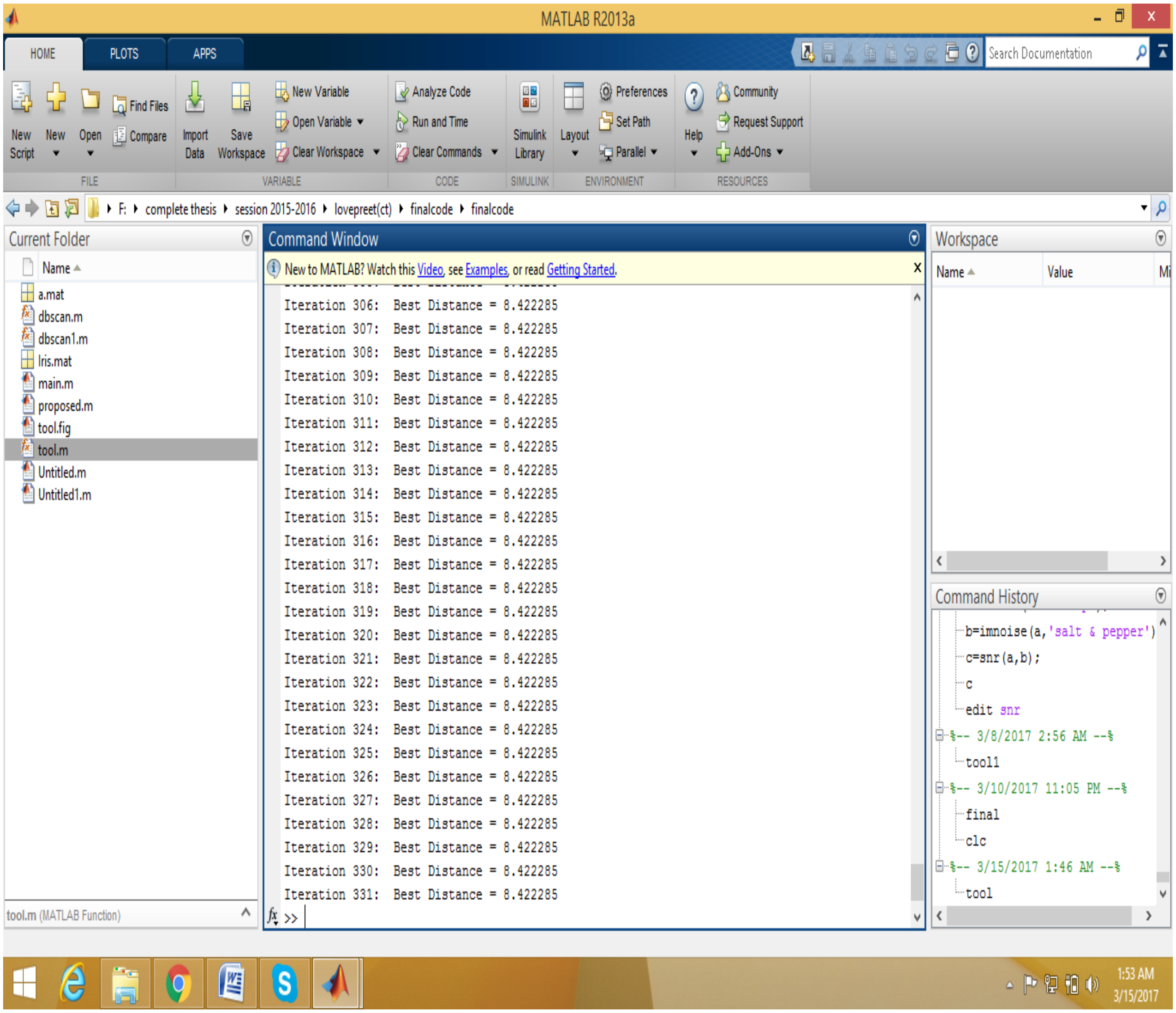
Name
a.mat
dbscan.m
dbscan1.m
Iris.mat
main.m
proposed.m
tool.fig
tool.m
Untitled.m
Untitled1.m

```
D =

     0
     0
     0
    10
    28
    15
    13
     0
     0
     0


m =

     1


n =

     1

The value of Eps is
    1.2986
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.3: Generation of clusters

As shown in figure 3, the EPS value is defined according to the input dataset. The EPS value defines the class of the dataset and Euclidian distance from the central point is calculated according to that similar and dissimilar values are clustered

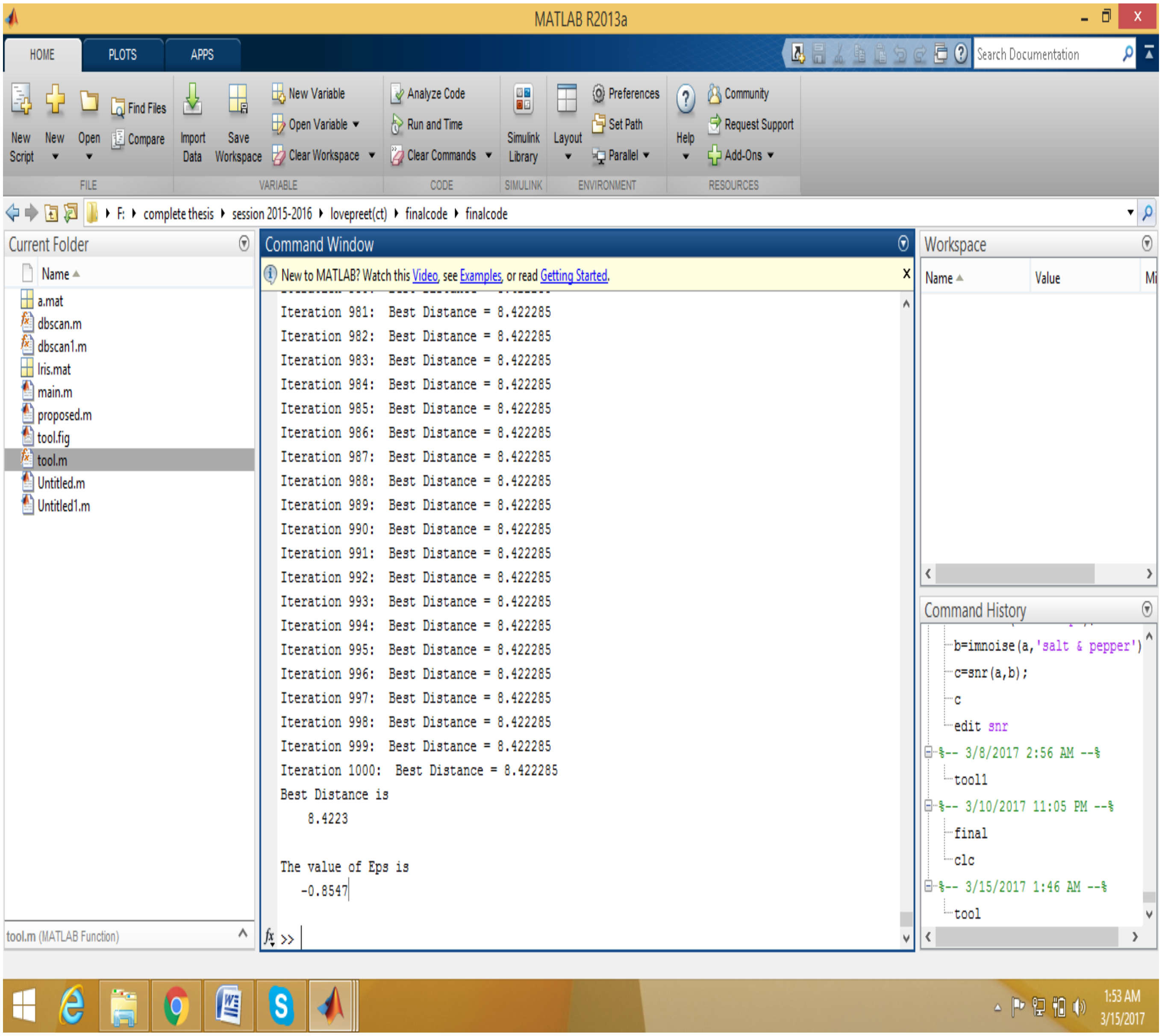


```
Iteration 306:  Best Distance = 8.422285
Iteration 307:  Best Distance = 8.422285
Iteration 308:  Best Distance = 8.422285
Iteration 309:  Best Distance = 8.422285
Iteration 310:  Best Distance = 8.422285
Iteration 311:  Best Distance = 8.422285
Iteration 312:  Best Distance = 8.422285
Iteration 313:  Best Distance = 8.422285
Iteration 314:  Best Distance = 8.422285
Iteration 315:  Best Distance = 8.422285
Iteration 316:  Best Distance = 8.422285
Iteration 317:  Best Distance = 8.422285
Iteration 318:  Best Distance = 8.422285
Iteration 319:  Best Distance = 8.422285
Iteration 320:  Best Distance = 8.422285
Iteration 321:  Best Distance = 8.422285
Iteration 322:  Best Distance = 8.422285
Iteration 323:  Best Distance = 8.422285
Iteration 324:  Best Distance = 8.422285
Iteration 325:  Best Distance = 8.422285
Iteration 326:  Best Distance = 8.422285
Iteration 327:  Best Distance = 8.422285
Iteration 328:  Best Distance = 8.422285
Iteration 329:  Best Distance = 8.422285
Iteration 330:  Best Distance = 8.422285
Iteration 331:  Best Distance = 8.422285
>>
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.4: Apply of back propagation algorithm

As shown in figure 4, the back propagation algorithm is been applied which will calculate the Euclidian distance dynamically to cluster similar and dissimilar values.

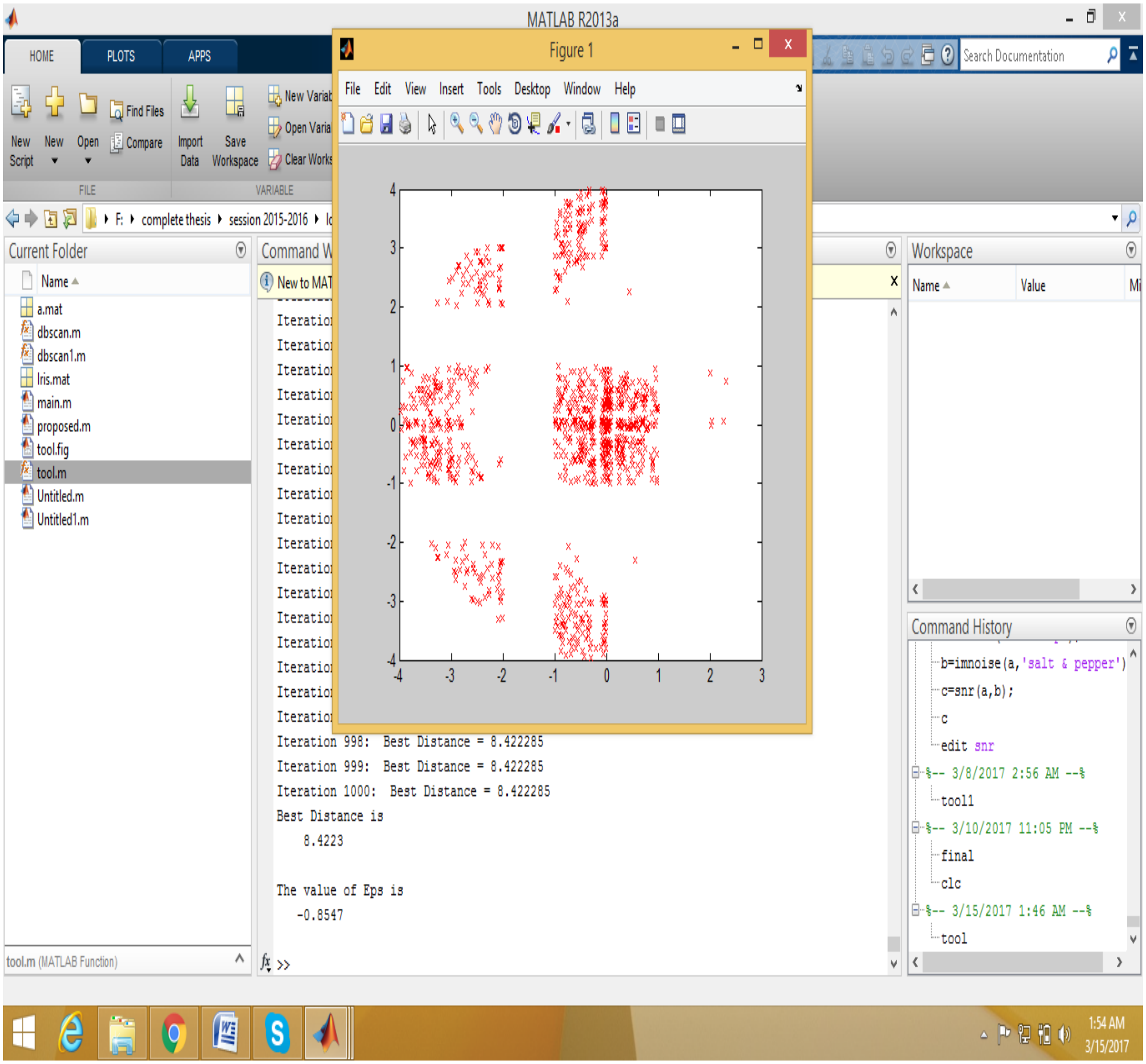


```
Iteration 981:  Best Distance = 8.422285
Iteration 982:  Best Distance = 8.422285
Iteration 983:  Best Distance = 8.422285
Iteration 984:  Best Distance = 8.422285
Iteration 985:  Best Distance = 8.422285
Iteration 986:  Best Distance = 8.422285
Iteration 987:  Best Distance = 8.422285
Iteration 988:  Best Distance = 8.422285
Iteration 989:  Best Distance = 8.422285
Iteration 990:  Best Distance = 8.422285
Iteration 991:  Best Distance = 8.422285
Iteration 992:  Best Distance = 8.422285
Iteration 993:  Best Distance = 8.422285
Iteration 994:  Best Distance = 8.422285
Iteration 995:  Best Distance = 8.422285
Iteration 996:  Best Distance = 8.422285
Iteration 997:  Best Distance = 8.422285
Iteration 998:  Best Distance = 8.422285
Iteration 999:  Best Distance = 8.422285
Iteration 1000:  Best Distance = 8.422285
Best Distance is
    8.4223

The value of Eps is
   -0.8547
>>
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.5: Euclidian distance value calculation

As shown in figure 5, the Euclidian distance is calculated using back propagation algorithm and average of the distance is taken to cluster similar and dissimilar values



```
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteration 998:  Best Distance = 8.422285
Iteration 999:  Best Distance = 8.422285
Iteration 1000:  Best Distance = 8.422285
Best Distance is
    8.4223


The value of Eps is
   -0.8547

fx >>
```

```
b=imnoise(a,'salt & pepper')
c=snr(a,b);
c
edit snr
%-- 3/8/2017 2:56 AM --%
tool1
%-- 3/10/2017 11:05 PM --%
final
clc
%-- 3/15/2017 1:46 AM --%
tool
```

Fig 4.1.6: Generation of final clusters

As shown in figure 6, the final clusters are generated according to Euclidian distance value and results shows that generated clusters are different from the existing clusters .

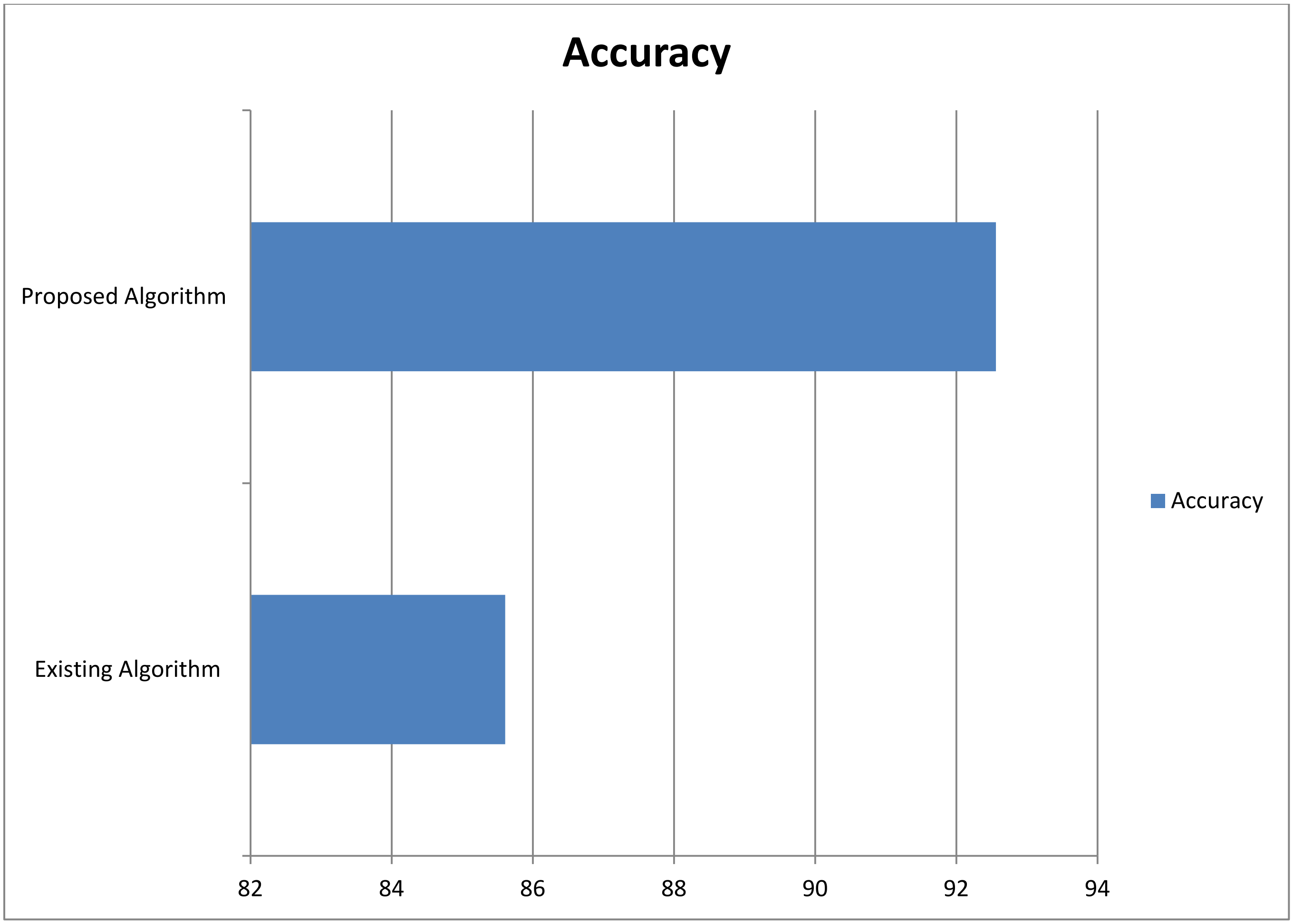


Fig 4.1.7: Accuracy of clustering

As shown in figure 7, the accuracy of proposed and existing algorithm is compared to check reliability of the algorithms and it is been analyzed that accuracy of proposed algorithm is more as compared to existing algorithm

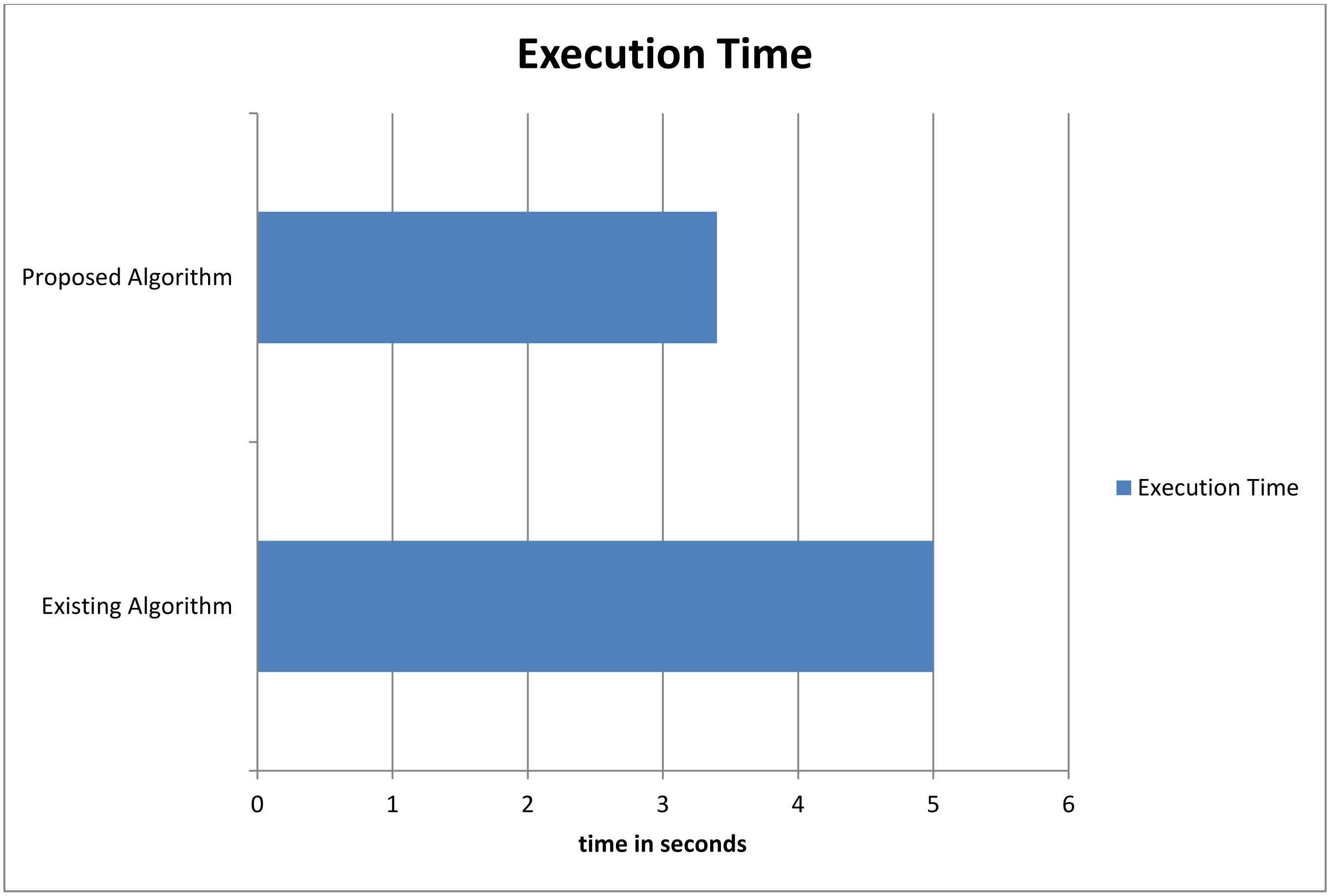
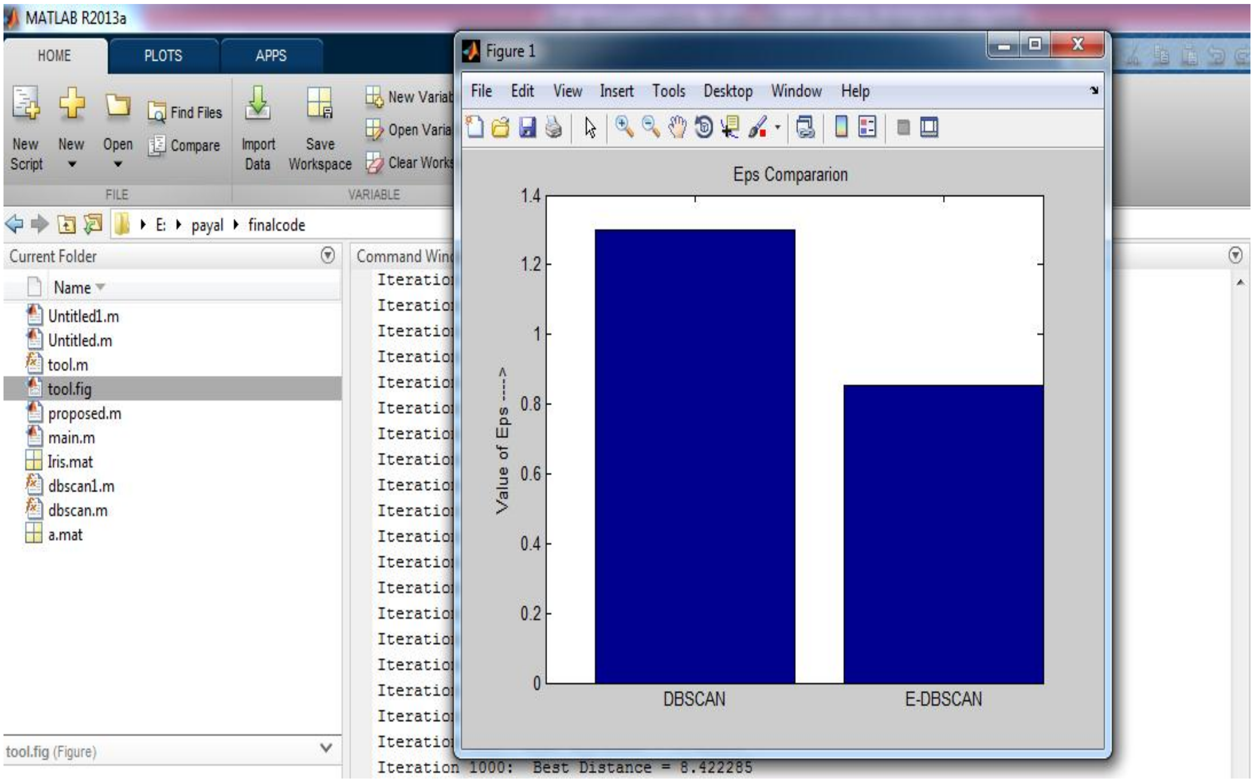


Fig 4.1.8: Execution time Comparison

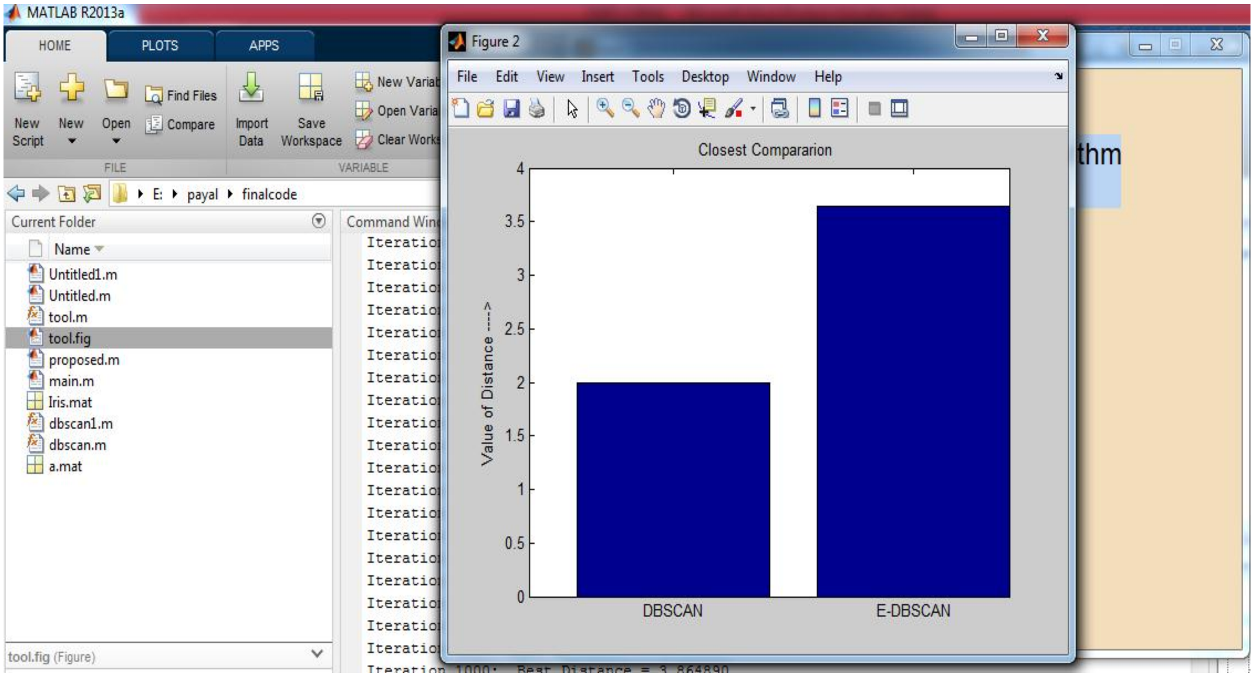
As shown in figure 8, the execution time of proposed and existing algorithm is compared and it is been analyzed that due to dynamic calculation of Euclidian distance execution time is reduced in the DBSCAN algorithm



```
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteration 1000:  Best Distance = 8.422285
```

Fig 4.1.9: Reduction of Eps value

As shown in figure 9, the Eps value of proposed and existing algorithm is compared and it is been analyzed that due to dynamic calculation of Euclidian distance Eps value is reduced in the E-DBSCAN algorithm .



```
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteratio
Iteration 1000:  Best Distance = 3.864890
```

Fig 4.2.0: Increase in value of distance

As shown in figure 10, the value of distance of proposed and existing algorithm is compared and it is been analyzed that due to dynamic calculation of Euclidian distance the value of distance is increased in the E-DBSCAN algorithm .

CHAPTER 5:

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

It is been concluded that density based clustering is type of clustering which is applied to cluster the data according to data density. The DBSCAN algorithm is the efficient algorithm which used for the density based clustering. This algorithm identifies the arbitrary shaped clusters which also includes separating the noise from large spatial databases. There are two parameters which are accepted by it which are Eps (radius) and MinPts (minimum points-a threshold). The numbers of points within a specific radius Eps are counted for the purpose of estimating the density at a specific point of the data set. This is known as a center-based approach which is applied here. There are various points which are classified in this approach in the categories such as core point, border point and noise points. The incremental DBSCAN algorithm is the variant of DBSCAN algorithm in which EPS value is calculated according to input dataset. To improve accuracy of clustering, improvement in the incremental DBSCAN algorithm is proposed in which based on back propagation algorithm. The testing results shows that accuracy is increased upto 8 percent and execution time is reduced to 16 percent.

5.2 Future Scope

1. The proposed technique will be enhanced further to increase the security of the clustered data.
2. The proposed algorithm can be compared with other algorithms of clustering in terms of execution time and accuracy by varying datasets.

References

[1] C. Bahm, K. Haegler, N.S Maller, C. Plant,” CoCo: coding cost for parameter-free outlier detection”, 2009, 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 149–158

[2] D. Wang, S. Zhu, T. Li, Y. Chi, Y. Gong,” Integrating clustering and multi document summarization to improve document understanding,” 2008, 17th ACM CIKM Conference on Information and Knowledge Management

[3] H.-P. Kriegel, M. Pfeifle,” Effective and efficient distributed model-based clustering,” 2005, 5th International Conference on Data Mining (ICDM’05), pp. 285, 265

[4] K.M. Hammouda, M.S. Kamel,” Efficient phrase-based document indexing for web document clustering,” 2004, IEEE TransKnowledge and Data Eng., vol. 16, no. 10, pp. 1279–1296

[5] Zhe Zhang, Junxi Zhang, Huifeng Xue,” Improved K-means clustering algorithm,” 2008, Congress on Image and Signal Processing CISP, vol. 5, May pp. 169–172

[6] L. Li, J. You, G. Han, H. Chen, Double partition around medoids based cluster ensemble, 2012, International Conference on Machine Learning and Cybernetics, vol. 4, pp. 1390– 1394

[7] H. Du, Y. Li,” An improved BIRCH clustering algorithm and application in thermal power,” 2010, International Conference on Web Information Systems and Mining (WISM), vol. 1, Oct pp. 53–56

[8] R.T. Ng, J. Han,” CLARANS: a method for clustering objects for spatial data mining,” 2002, IEEE Trans. Knowl. Data Eng. 14 (5) 1003–1016

[9] S. Guha, R. Rastogi, K. Shim,” CURE: an efficient clustering algorithm for large databases,” 1998, Proceedings of the ACMSIGMOD International Conference Management of Data (SIGMOD’98), pp. 73–84

[10] G. Karypis, H. Eui-Hong, V. Kuma,” Chameleon: hierarchical clustering using dynamic modeling”, 1999, Computer 32 (8) 68–75

[11] M. Ester, H. Kriegel, J. Sander, X. Xu," A density-based algorithm for discovering clusters in large spatial databases with noise," 1996, Proc. 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231

[12] Z. Wang, Y. Hao, Z. Xiong, F. Sun," SNN clustering kernel technique for content-based scene matching," 2008, 7th IEEE International Conference on Cybernetic Intelligent Systems, pp. 1–6

[13] E. Achtert, C. Bhm, A. H. Kriegel, P. KrAger, I. Maller- Gorman, A. Zimek," Detection and visualization of subspace cluster hierarchies", 2007, Advances in Databases: Concepts, Systems and Applications, Lecture Notes in Computer Science, pp. 152–163

[14] G. Tzortzis, A. Likas," The global kernel K-means algorithm for clustering in feature space", 2009, IEEE Trans. Neural Netw. 1181–1194

[15] D. Widyantoro, T. Ioerger, J. Yen," An incremental approach to building a cluster hierarchy", 2002, ICDM Proceedings IEEE International Conference on DataMining, pp. 705–708

[16] S.A.L. Mary, K.R.S. Kumar," A density based dynamic data clustering algorithm based on incremental dataset", 2012, J. Computer Sci. 8 (5) 656–664

[17] K.M. Hammouda, M.S. Kamel," Incremental document clustering using cluster similarity histograms", 2003, IEEE/WIC Proceedings International Conference on Web Intelligence, pp. 597–601

[18] M. Ester, H.P. Kriegel, J. Sander, M. Wimmer, X. Xu," Incremental clustering for mining in a data warehousing environment", 1998, Proceedings of the 24th VLDB Conference, Institute for Computer Science, University of Munich, Germany, New York, USA

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

[20] Iyer Aurobind Venkatkumar, Sanatkumar Jayantibhai, Kondhol Shardaben," Comparative study of Data Mining Clustering algorithms", 2016, IEEE

[21] Qi Xianting, Wang Pan,” A density-based clustering algorithm for high-dimensional data with feature selection”, 2016, IEEE

[22] Kuan-Teng Liao, Chuan-Ming Liu,” An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans”, 2016, IEEE

[23] Cheng-Fa Tsai†, Han-Chang Wu, and Chun-Wei Tsai,” A New Data Clustering Approach for Data Mining in Large Databases”, 2002, IEEE

[24] Wenbin Wu and Mugen Peng,” A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting”, 2016, IEEE

[25] Vadlana Baby, Dr. N. Subhash Chandra,” Distributed threshold k-means clustering for privacy preserving data mining”, 2016, IEEE

[26] KM Archana Patel and Prateek Thakral,” The Best Clustering Algorithms in Data Mining”, 2016, IEEE

[27] ZHANG Ke, HUANG Lei, CHAI Yi,” An Algorithm to Adaptive Determination of Density Threshold for Density-based Clustering”, 2016, IEEE

[28] 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

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

[30] 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

[31] 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

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

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

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

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

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

[37] 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

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