ENHANCED MULTI VIEW CLUSTERING ALGORITHM FOR BIG DATA PROCESSING

Project Report submitted in partial fulfillment of the requirement

for the degree of

Master of Technology

in

Computer Science & Engineering

under the supervision of

Dr. Pardeep Kumar

By

Isha Aggarwal (152209)

Jaypee University of Information Technology

Waknaghat, Solan – 173234, Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled **"Enhanced Multi View Clustering Algorithms for Big Data Processing"**, submitted by **Isha Aggarwal** in partial fulfillment for the award of degree of Master of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been made under my supervision.

This report has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date:

Dr. Pardeep Kumar
Assistant Professor (Senior Grade)

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the dissertation entitled **"ENHANCED MULTI VIEW CLUSTERING ALGORITHM FOR BIG DATA PROCESSING"** in partial fulfillment of the requirements for the award of the degree of Master of technology and submitted in Computer Science and Engineering Department, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by **Isha Aggarwal** during a period from July 2016 to May 2017 under the supervision of **Dr. Pardeep Kumar,** Assistant Professor (Senior Grade), Computer Science and Engineering Department, Jaypee University of Information Technology, Waknaghat.

I have not submitted the matter embodied in this dissertation for the award of any other degree.

Date: **ISHA AGGARWAL**

Place: Waknaghat, Solan, H.P. **152209**

ACKNOWLEDGEMENT

I earnestly wish to express my heartfelt thanks and a sense of gratitude to my guide **‘Dr. Pardeep Kumar’**, Computer Science and Engineering Department, for his valuable guidance and constant inspiration in preparing this report. My frequent interactions with him in all aspects of the report writing have been a great learning experience for me. I shall always cherish his support and encouragement.

Last but not the least, I heartily appreciate all those people who have helped me directly or indirectly in making these task a success. In this context, I would like to thank all the other staff members, both teaching and non-teaching.

Date:

Isha Aggarwal

TABLE OF CONTENTS

ABBREVIATIONS

BD	-	Big Data
DB	-	Database
SDB	-	Stream Database
SD	-	Structured Data
SSD	-	Semi - Structured Data
UD	-	Unstructured Data
K-means	-	K-means Clustering Techniques
HS	-	Hadoop Structure
MRA	-	Map Reduce Algorithm
MATLAB	-	Matrix Laboratory
WS	-	Work Space
EW	-	Editor Window
CW	-	Command Window
GUI	-	Graphical User Interface
HW	-	Help Window
CF	-	Current Folder
PCA	-	Principal Component Analysis
UI		User Interface
SVM	-	Support Vector Machine

LIST OF FIGURES

ABSTRACT

Information is a noteworthy piece of the data i.e. it is appropriately organized utilizing some extraordinary procedure. Typically, information could lie in a few structures like numbers and content in the database, as bits/bytes in PC memory, as actualities in the mind of individual. In addition, data is at present at every last place on the planet. There is no-restriction on the data display everywhere throughout the world. Since the day when human improvement begins, we generally lie with some data. Everyday life manages some data with/with no reason. The data could be of any sort like memory of somebody's name/sparing a play-rundown of thousands of tunes.

Big data Streaming examination these days has separately turned out to be a standout amongst the most noteworthy subjects in the rundown of information experts since huge nature of information are delivered general by the few gadgets. Big data alludes to an accumulation of various datasets that are so immense and troublesome that it winds up noticeably hard to gather them utilizing conventional data giving out applications.

The problems add study, confine study, division, storage, and image, confidentiality violation. The features of big information are attributes and data types / units. Big data is the major information that manages use to analysis the attributes and data-types.

In this project, we work on databases that apply and compare two techniques namely k-means clustering and Multiview clustering such that we get to know how to categorize data and utilize the information extracted from it to predict future trends used further for analysis purposes.

We compare k means clustering approach with the hierarchical clustering also known as the multi view clustering in the aspects which tells us that the latter gives us the vision to analyze data from various perspectives and give better outcomes. The technique takes after a basic approach to group a given informational index done a specific number of groups stable a priority and calculate the performance parameters i.e. accuracy, precision and recall rate.

CHAPTER 1

1. Introduction

The sudden and intemperate development of information amid the start of the twenty first century has turned into a test for human progress. In the underlying stage that keeps an eye on around the last some portion of the twentieth century the idea of relational database came into the lime light [1]. In software engineering, the information is characterized as data that is crude (for example letters in order, numbers or images), which can be prepared by PC for certain importance. Today, digitalization is extremely normal on the planet. It talks simple data to advanced shape, which can be a great deal more effectively put away, handled and transmitted. Advanced data and information give a rich wellspring of vast scale collected information. Is advanced data quite recently used to share and perused? Can advanced data be changed over into helpful information with mindfulness esteem? Many research foundations have found that computerized data database will turn into an incredible fortune, and obviously, society needs are absolutely not fulfill with the constrained data sharing. So information mining turns into another important research course.

Figure 1. Overview of Big data

A colossal measure of information is delivered in varying backgrounds each day, and the unit of information estimation creates from Byte, KB, MB, GB, TB to PB, EB, ZB, YB. Separating profitable data and utilizing information for basic leadership are the center purposes that we yearning to achieve; just along these lines can information make esteem. Presentation of social database was progressive stride in the realm of information where information is put away in a table and can be essentially prepared by the need. At that point to make the life less demanding, some examination over the information was required which tackled a great deal of issues of the everyday life. Be that as it may, the genuine issue came into the photo shameful after the web was presented. Because of the boundless utilization of web, we got a few information having gigantic volume, high-speed, broad assortment [2]. The intuitive database couldn't ready to deal with and prepare that information. Consequently, another kind of information known as Large Information was presented with various ideas and diverse advances.

1.1 Types of data

Various types of big data i.e. structured data, semi-structure data etc. Briefly explained in below:

Structured Data

The information which can be put away in the social database table consecutively - section organize [3]. As the name proposes, organized information is having some particular structure and that structure is characterized by the information demonstrate which is delivered by the association. The association initially portrays the information model i.e., a model which will permit putting away, handle and getting to the hierarchical information. The model needs to characterize the properties of the information that will be stored..

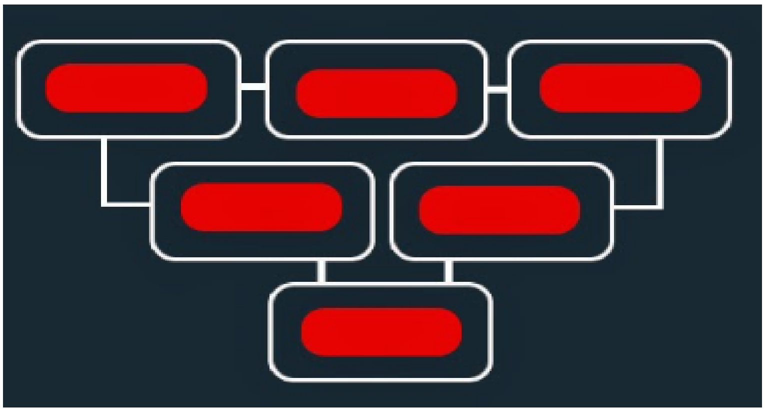


Figure 2.1 Structured data

The property of the information incorporates: information sort (numeric, alphabetic, name, date, and so forth.) and a few confinements on the information (size of the information, measure of characters, and so forth.). The significant advantage of organized information is that, it can be effortlessly put away and examined. Because of the high cost and restrictions of the storage room and preparing methods, social database is the best way to store and process the organized information adequately. Fundamentally organized information is overseen by utilizing Organized Inquiry Dialect (SQL). SQL is a programming dialect for dealing with the data in the RDBMS [6]. It was produced by IBM in the year 1970 and later it was created economically by Relational Software, INC (Presently Oracle Corporation).

Semi - Structured Data

The information which is organized information however that does not fit with the information models characterized for the organized information is known as semi-organized information [3]. The semi - organized information isn't put away in the social databases or different types of an information table, rather in some particular sort of records which fathom a few labels. The labels or the markers are isolated through some semantic principles and uphold the information to be put away with a pecking order.

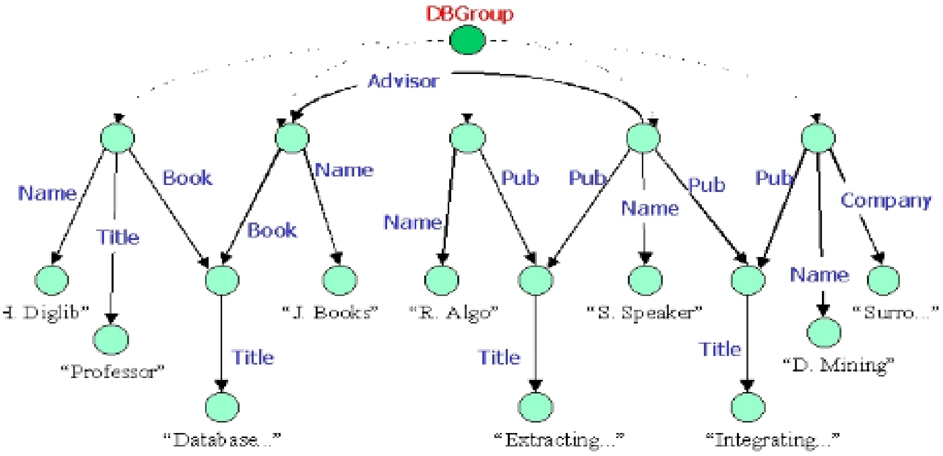


Figure 2.2 Semi Structured Data

This sort of information is combined quickly after the web is presented where unique types of the information and totally various types of uses need mediums for trading the data like XML and JSON.

Unstructured Data

The information which doesn't have a particular structure, consequently, can't be put away in succession section arrangement of a conventional database is known as unstructured information [3]. As the name proposes, the unstructured information is fundamentally the antonym of organized information. Henceforth, it can't be put away in fields of a database.

Illustrations: Content documents, picture records, sound records, video documents, website pages and so forth.

Figure 2.3 Unstructured Data

Presently a-days, the volume of unstructured information is developing so quickly that it is exceptionally hard to deal with and investigate the information. In this way, to break down the unstructured information it requires more learning with some propelled advancements.

Streamlining data

Various research associations are utilizing this kind of huge information examination to find new pharmaceuticals. A protection partnership may wish to analyze the examples of car crashes crossways a wide geographic zone by climate insights. In these cases, no advantage exists to deal with this data at continuous speed. Obviously, the investigation must be quick and handy. Also, associations will look at the information to see whether novel examples rise. Spilling information is a logical processing stage that goes for speedier speed. This is on account of these applications require a persistent stream of a considerable measure unstructured information to be handled. In this way, information is broke down at consistent interims and changed in memory before it is put away on a plate. The surge of information is handled by preparing "time windows" of data in memory over a bunch of servers.

This is like the approach while overseeing information very still utilizing Hadoop. The essential contrast is the issue of velocity. In the Hadoop group, information is gathered in cluster mode and then handled. Speed matters less in Hadoop than it does in information spilling. Some key standards characterize when utilizing streams is best fitting:

- When it is important to control a retail purchasing opportunity at the purpose of engagement, either by means of online networking or through authorization based informing.
- Collecting data about the development about a protected site.
- To be proficient to respond to an occasion that needs a quick reaction, for example, an administration blackout or an adjustment in a patient's medicinal condition.
- Real-time control of costs that are subject to factors, for example, use and accessible assets.

Gushing information is valuable when examination should be done progressively while the information is in movement. Truth be told, the cost of the investigation (and frequently the information) diminishes with time. For instance, on the off chance that you can't investigate and act instantly, a business opportunity may be lost or a danger may go undetected. [4]

Advantages and disadvantages of big data

Big data implies immense volumes of information that regularly have a confused structure. Present day instruments bolstered by effective PCs can prepare this sort of information, which was incomprehensible before. By the by, Big data has focal points, as well as inconveniences that ought to be considered by analysts.

Advantages of BIG data

1. There are practically boundless capacity potential outcomes for colossal information volumes

2. Big data are currently available from wherever and by means of different gadgets as they are typically put away in Clouds.

3. The speed of Big data transmission and handling is high attributable to cutting edge advances

4. Present day analytic techniques, advancements and instruments enable examiners to increase profound bits of knowledge into big data, which was outlandish in the past with constrained information volumes and weaker preparing apparatuses

Disadvantages of big data

1. Big data regularly have enormous commotion, i.e. there might be numerous good for nothing data points. The expert ought to endeavor to isolate the wheat from the tares.

2. Big data regularly infers protection issues, which can be seen, for example, from the analysis of social networks.

3. Big data likewise implies a significant low security level. It is normal as Clouds are dependably not as secure as on location information distribution centers.

Applications

Applications are all around used to examine enormous measure of data i.e. Enormous information and offer enhanced bits of learning to the end-customers. Taking after are a bit of the application fields

1. Advertisement Enhancement – Media Math is the essential demand side stage (DSP), changing the way modernized media is purchased, and making another, more capable course for marketing experts to accomplish customers, independently, at scale.

2. Distributer Apparatuses- Visual Pay is a steady farsighted examination organize developing a suite of instruments giving decision support to publication supervisor's substance

3. Energy Auto Network takes the information from savvy meters, voltage controllers, indoor regulators to help clients track the measure of energy utilized, downsize squander, adjust the framework, expands the framework operations and conjectures the future utilization.

1.2 K-Means Clustering Algorithm

K-means clustering is a calculation to order or to assemble your articles in view of traits/elements into K measure of gathering. K is certain number i.e. it refers to the number of clusters that are predefined as to hoe many clusters do we want our data to be classified into. The gathering is finished by limiting the whole of squares of separations in the midst of information and the relating group centroid. Therefore, the main purpose of K-mean clustering is to griup the data in the datasets.

K-means is one of the humblest unsupervised learning algorithm that tackles the outstanding clustering issue. The methodology takes after a direct approach to arrange a given informational collection done a specific number of clusters (accept k groups) stable apriority. The basic thought is to depict k centers, a single center for each group. These focuses ought to be put cleverly in view of various area causes disparate outcome. So, the improved choice is to change their location that is as away from each other as possible.

Advantages

1) Quick, robust and less demanding to get it.

2) Gives great outcome when data sets are distinguishable or all around isolated from each other.

Disadvantages

1) The learning procedure requires earlier specification of the quantity of group centers.

2) The utilization of Elite Task - If there are two exceptionally covering data then k-means won't have the ability to verify that there are two groups.

3) The learning count is not invariant to non-coordinate changes i.e. with dissimilar depiction of data we get differing results.

4) Euclidean partition measures can unequally weight concealed components.

5) The information calculation gives the neighborhood optima of the squared error work.

6) Arbitrarily choosing the group focus does not lead us to any productive outcome

7) Material just when mean is characterized i.e. comes up short for straight out information.

8) Unable to handle noisy data & exceptions.

9) Calculation falls flat for non-straight informational collection.

Steps of K-mean Clustering Algorithm

Repeat until stable (= no centers should change further):

1. Determine the centroid.

2. Decide the separation of each object to the centroid.

3. Group the items in view of least space (locate the nearest centroid).

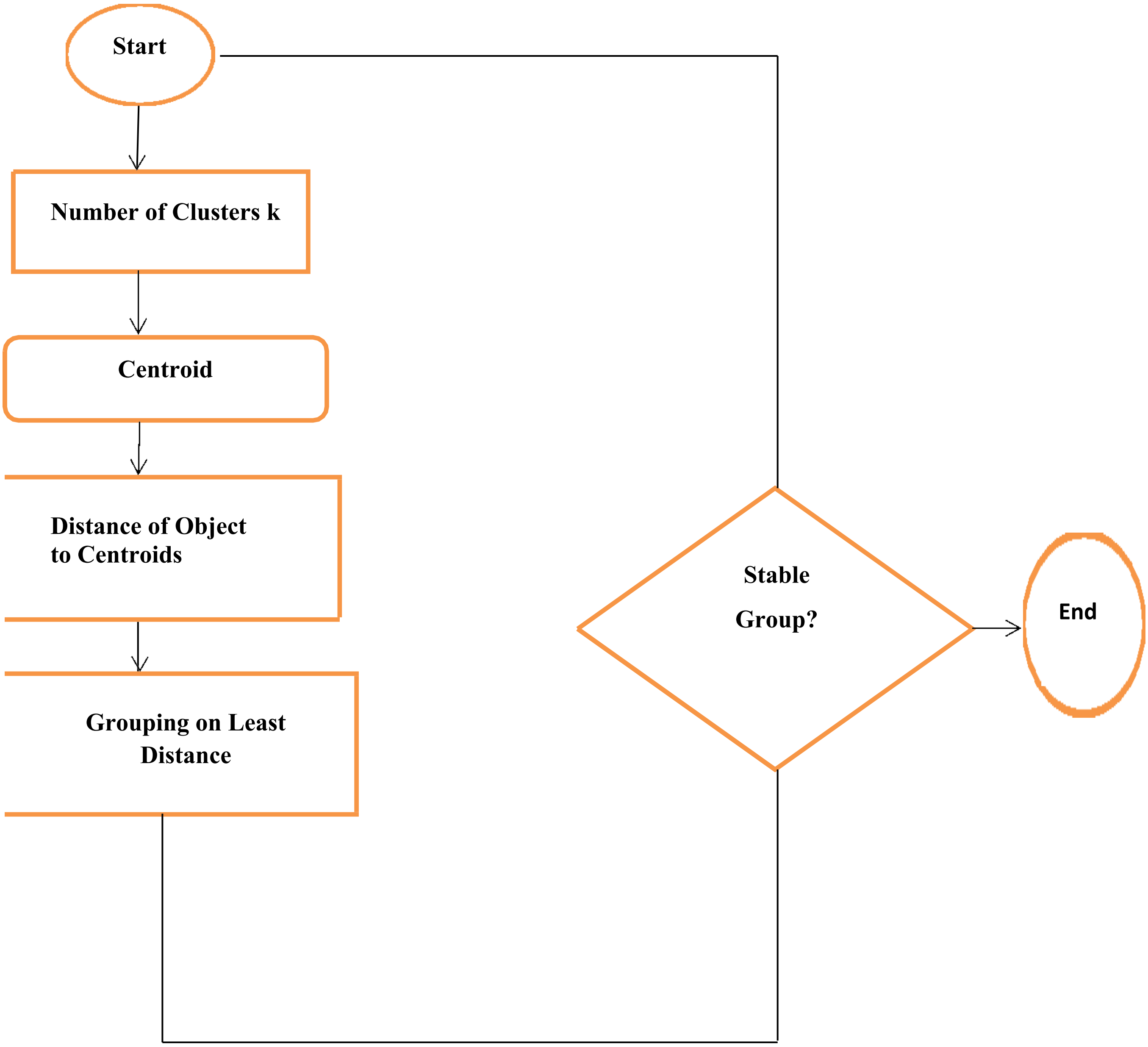


Figure 5.3 Flow Chart of K-Means Clustering Algorithm

1.3 Big data algorithm

Hadoop Structure

Hadoop is an open source stage sent on bunches of PCs with conveyed calculations. Keeping in mind the end goal to handle gigantic information, clients use Hadoop effectively to arrange PC assets. Apache Programming Establishment fabricates their own circulated processing stage and takes full preferred standpoint of cluster registering and substantial capacity limit. In February 2006, Hadoop project propelled formally with the principle support of two center innovations: Outline and HDFS. In this piece of article, we will disclose how to assemble a dispersed application with Hadoop, which depends on Guide Diminish execution component and base stockpiling of HDFS (Hadoop Appropriated Document Framework). HDFS has the benefits of solid adaptation to internal failure and high superfluity which enables clients to convey Hadoop circulated framework in mid-range quality equipment. In enormous information, HDFS grouping framework is running as Ace Slave mode, and there are two principle sorts of hubs; A Name Hub (i.e., Ace Hub) and various Information Hubs (i.e., Slave Hubs). Name Hub deals with the record framework's Namespace, which keeps up the document framework tree and metadata in all records and envelopes. HDFS can't work without Name Hub. Actually, if the machine's Name Hub running is broken, records in the framework will be totally lost, in light of the fact that there is no other approach to remake document obstructs on various Information Hubs.

Hence, the fault-tolerance mechanism of Name Node is overwhelmingly vital [5].

Map Reduce Algorithm

Map Reduce is a sort of dispersed processing model, which tackles the issue of huge information calculation. It consists of two phases: Map and Reduce. Users only need to apply map () and reduce () two functions according to their requirements, which lead to achieve distributed computing.

The parameters of these two functions are < *key* > and < *value* >, which stand for function input. Outline utilizes the Master Slave structure. Master (Occupation Tracker) is the main chief of worldwide bunching, which has components of employment administration, state checking and errand booking. Slave (Errand Tracker) is in charge of the input of assignment usage. There is just a single Employment Tracker, yet heaps of Assignment Trackers. The Employment Tracker is responsible for getting client demands and allotting errands to Assignment Trackers. Work Tracker perpetually tunes in and gets pulse data which is sent from each Undertaking Tracker, including asset state and errand data [6].

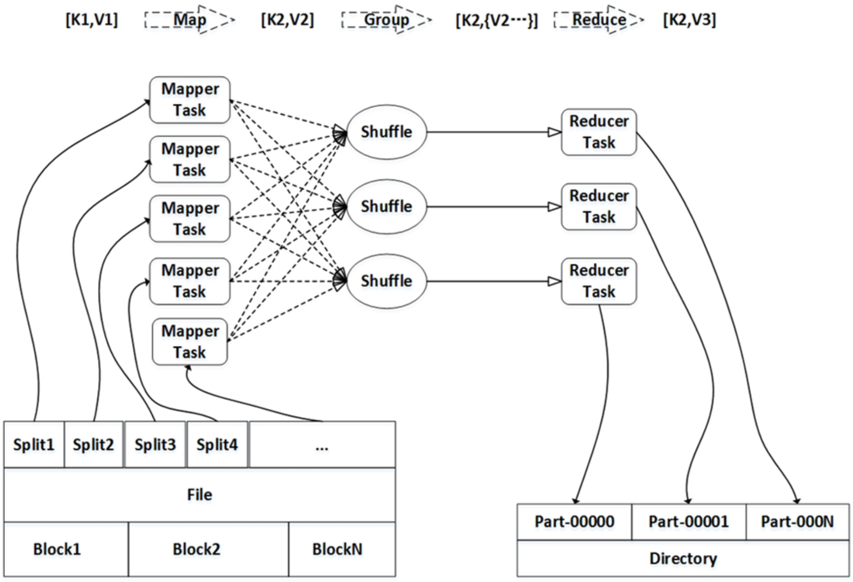


Figure 6.2. Map Reduce Algorithm Step

MRNB algorithm

MRNB calculation [7] is a classifier used to figure precipitation which depends on the stage Hadoop. Test comes about demonstrate that the MRNB classifier has focal points of good precision, high usage proportion of bunch asset and solid expansibility, and it is a remarkable answer for precipitation informational indexes characterization. Administrations like MRNB can be utilized to improve the precipitation accuracy by managing meteorological huge information.

KNN algorithm

KNN [8] grouping utilizes day by day meteorological information and Guide Lessen programming model to lead an investigation on Hadoop stage for precipitation forecast and finish the CVKNN parallelization, which accomplishes larger amount of expectation. It can hold gigantic of meteorological information, yet doesn't examine the connection research of Information Hubs number and accelerate proportion.

CHAPTER 2

LITERATURE SURVEY

[9] Fei Sui et al. (2016) depicts a testing stage for BDP to test the execution of huge information stacking and big data examination. The aftereffects of experiment demonstrate that the execution of information stacking and analysis of BDP is superior to anything conventional data distribution center.

[10] Konstantin's F. Xylogiannopoulos et al. (2016) another philosophy in view of our past work in regards to the covering of every single rehashed example in a string so as to break down a genuine enormous information stream with 1 Trillion digits, made from 1 thousand subsequences of 1 billion digits every one. All the more uniquely, utilizing the novel information structure, LERP Finish Addition Exhibit, and the inventive ARPaD calculation which permits the location of every single rehashed example in a string we figured out how to investigate each one of the 1 billion information focuses, utilizing 10 PCs through standard equipment arrangement, in 33 minutes which beats to the best of our insight some other existing approach, which is equal to information point era each 2 microseconds.

[11] Yuan-Chih Yu et al. (2016) characterize a theoretical system of protection weaving pipeline devoted for creating open and huge information while safeguarding security. Contained by the handling pipeline, each progression of the procedure stream considers the protection confirmation to control datasets. Be that as it may, the many-sided quality of process stream is the comparative as would be expected information pipeline. The untried model affirms the practicality of system outline.

[12] K.VijiyaKumar, et al. (2016) providing privacy for clients' reviews while advancing the authorities of social media by using the LDA algorithm for examining client's reviews.

[13] Priyanka Dhaka et al. (2015) depict the thoughts and procedure of utilizing information mining strategies, for example, hereditary calculation that can be useful on the information which was obtained from world insights of emotional wellness and sent this information into a major information instrument like Mongo DB and expelled organized informational collections that could help us to enhanced comprehend what treatment work for which sort of patient.

[14] D.Franklin Vinod et al. (2015) characterize Highly Correlated Feature Set Selection (HCFS) process is proposed to join through the various leveled inclining way to deal with enhance its execution. This calculation recognizes the great element subsets which will enhance the accuracy.

[15] Jie Gao et al. (2016) gaining and breaking down the estimation report (ER) and counters (database of the execution indicators of systems) from existing systems, an obstruction administration calculation (OAC) is proposed in view of huge information examination.

[16] Haina Ye et al. (2016) depict the current innovative work on safety and protection in big data is reviewed. To start with, the impacts of qualities of enormous information on data security and protection are portrayed. At that point, themes and issues on security are talked about and assessed. Encourage, security saving direction information distributing is concentrated because of its future use, particularly in telecom operation.

[17] Xin lu et al. (2016) depict the framework execution of SQLon-Hadoop innovation and MPP innovation in an OLAP and OLTP blended situation, which is turning into the most widely recognized enormous information application situation in telecom industry. Up to this point, most SQL-on-Hadoop frameworks just concentrate on the speed of information investigation yet ignore the support for exchanges. MPP columnar databases, then again, can manage both logical and value-based errands, yet have a tendency to have poor value-based execution because of the column-store mechanism.

[18] Yuwei Jai, et al. (2016) characterize utilizes the telecom information and proposes a client bunching and influences the power positioning plan. The plan is executed through three phases, i.e. the client picture investigation arrange, the client bunching examination organize and the positioning phase of client impact control.

CHAPTER 3

PROBLEM DESCRIPTION

Presently we have what it takes to handle a lot of information. The information of this kind is known as big data. Big data is the term utilized for an accumulation of informational indexes that are so huge, substantial, novel and conceivably valuable that it winds up noticeably hazardous to process it utilizing all the regular information preparing devices. Big data can be described in three ways:

(i)Volume (a lot of information),

(ii) Assortment (incorporates distinctive sorts of information), and

(iii) Speed (continually aggregating new information).

When managing big data, information grouping issue is a standout amongst the most vital issues. Every now and again informational collections, especially enormous informational collections, include of a few gatherings (bunches) and it is important to discover the gatherings. Clustering techniques have been connected to numerous vital issues, for instance, to decide human services slants in patient records, to evacuate copy passages in address records, to recognize new classes of stars in cosmic information and so on. To address these applications and numerous others an assortment of bunching calculations has been created. There exist a few downsides in the current bunching approaches; most calculations include checking the informational index for a few times, along these lines they are unacceptable for huge information grouping.

The standard k-means grouping was intended for taking care of single-view information clustering issue. Another strong multi-view grouping strategy was proposed to integrate heterogeneous features for clustering.Some k-means changes for information are presented. The streaming k-means calculation for well-clustered capable information is distributed in. The primary k-means issue is place to store the information is excessively extraordinary, making it impossible to be put away in the fundamental memory and must be recovered in a sequence.

CHAPTER 4

PROPOSED SOLUTION

1. To upload the dataset in simple dataset and stream dataset.
2. Pre-processing of the error reports by using text mining method and create Term document matrix in binary form
3. Cataloguing of the velocity, volume and variety of groups in big data by exercise the dataset consuming K-mean clustering algorithm. Analysis the competence of the planned work by assessing the accuracy of the system.

CHAPTER 5

METHODOLGY

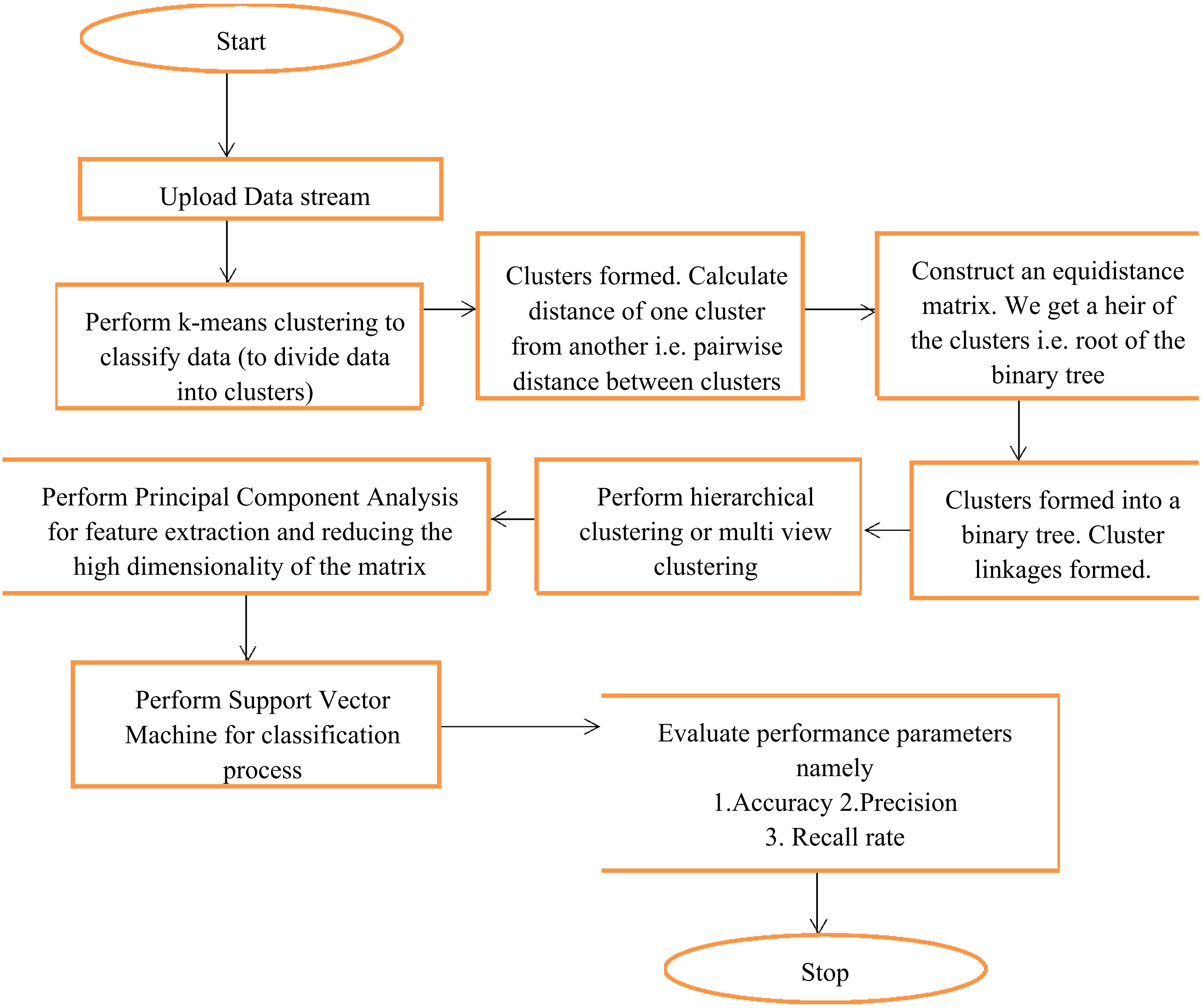

Start
Upload Data stream
Perform k-means clustering to classify data (to divide data into clusters)
Clusters formed. Calculate distance of one cluster from another i.e. pairwise distance between clusters
Construct an equidistance matrix. We get a heir of the clusters i.e. root of the binary tree
Clusters formed into a binary tree. Cluster linkages formed.
Perform hierarchical clustering or multi view clustering
Perform Principal Component Analysis for feature extraction and reducing the high dimensionality of the matrix
Perform Support Vector Machine for classification process
Evaluate performance parameters namely
1.Accuracy 2.Precision
3. Recall rate
Stop

Figure 5.1 Proposed Methodology Flow Chart

CHAPTER 6

IMPLEMENTATION AND EXPERIMENT

6.1 Simulation Tool

The ensuing Improvement Apparatuses has been utilized as a part of the development of this work. There might be different devices which can be utilized as a part of this improvement as it depends individual to individual and his advantage. Hence the tools used are:

1. Minimum of 3 GB RAM
2. Intel Pentium III Processor or over
3. MATLAB R2010a

Table 6.1 Tool used

Computer	Core 2 Duo or higher
RAM	4 MB
Platform	Windows 7,8,10
Other hardware	Keyboard, mouse
Software	Matlab 2013a

6.2 What is MATLAB?

The expansion of MATLAB is MATRIX LABORATORY. MATLAB was images right off the bat to give simple access to grid programming created through the LINPACK (Straight Framework Bundle) MATLAB is a high appearance dialect for innovative registering. It consolidates figuring, nebulous vision, and modified condition. Also, MATLAB is current programming dialect environment: it has complex information structures, contains worked in altering and investigates apparatuses, and bolsters protest arranged programming. These issues fabricate MATLAB an estimable instrument for instructing and research.

1. MATLAB is contrasted with preservationist coding languages (e.g., C, FORTRAN) for taking care of modern issues.
2. MATLAB is an intuitive framework whose straightforward information component is a cluster that does not require dimensioning.
3. The product bundle has been industrially realistic since 1984 and is presently measured as a model apparatus at almost all colleges and industries around the world.
4. It has telling inherent schedules that empower a wide assortment of calculation. It additionally has simple to utilize design guidelines that make the fantasy of result potentially useful.
5. Correct applications are gathered in bundle alluded to as tool kit. There are tool kits for sign handling, delegate calculation, control hypothesis, proliferation, advancement and numerous different areas of connected science and designing.

6.3 Basic Features

The course is to make us friendly with the basics of MATLAB.

6.4 Advantages of MATLAB

- It gives a major database of inherent calculations for picture preparing.
- It permits testing calculations instantly without recompilation.
- Ability to prepare both still pictures and recordings.
- Ability to call inside libraries.
- Ease of utilization of tool makes it less demanding for new clients.

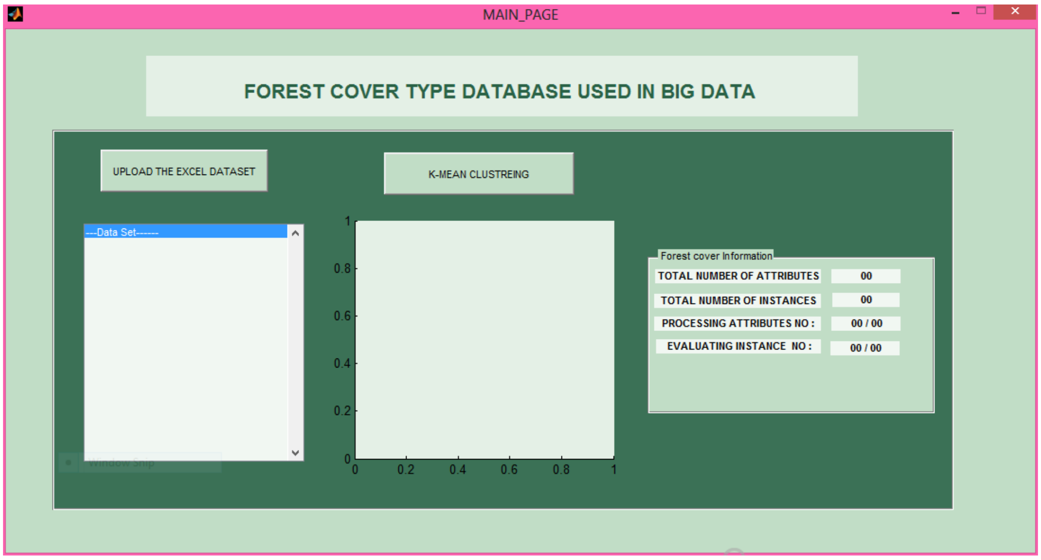
6.5 MATLAB characteristics

- Imitative from FORTRAN
- Developed to get to LINPACK and EISPACK grid
- Rewritten in C
- The Math Works incorporates was framed by 1984 to commercial centre and go ahead with extension of MATLAB.

6.6 Strength of MATLAB

- Act a calculator or as a programming dialect
- Used for estimation and representation plotting
- Easy to learn
- The fault is easy when performing grid operations.
- Contains few OOPs ideas and terminologies.

6.7 Results and Discussions

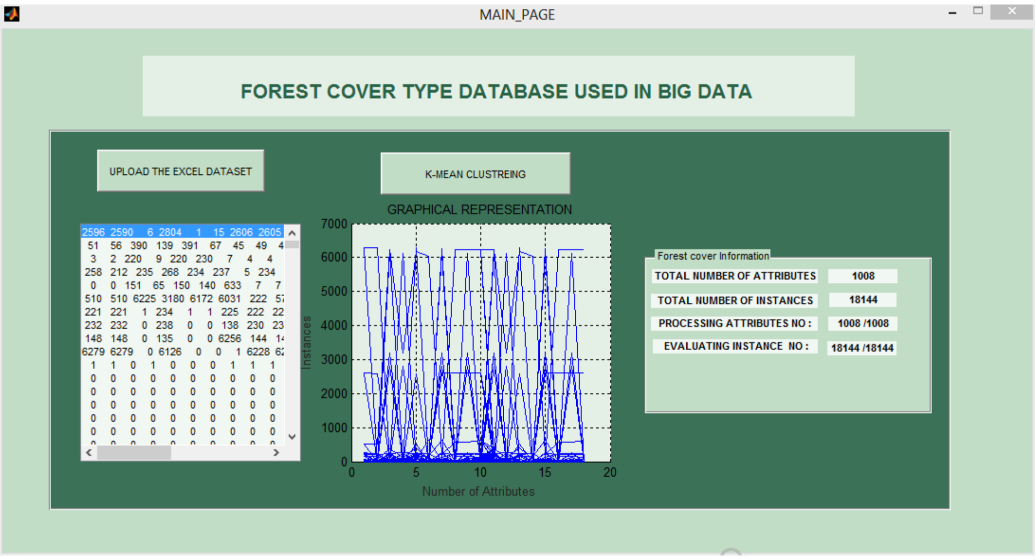

MAIN_PAGE

FOREST COVER TYPE DATABASE USED IN BIG DATA

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	00
TOTAL NUMBER OF INSTANCES	00
PROCESSING ATTRIBUTES NO :	00 / 00
EVALUATING INSTANCE NO :	00 / 00

Figure 6.3 Main Page


MAIN_PAGE

FOREST COVER TYPE DATABASE USED IN BIG DATA

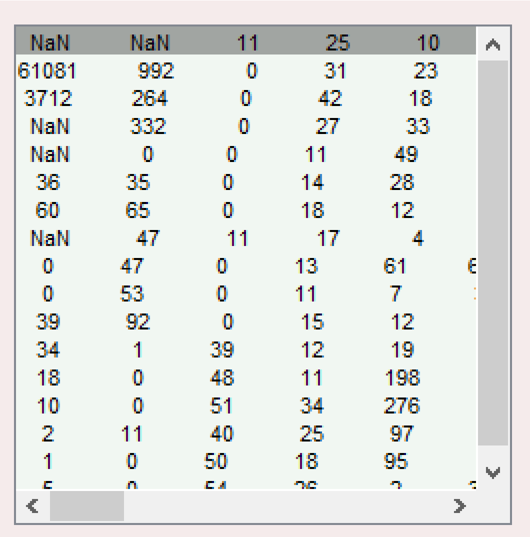
```
2596 2590    6 2804    1   15 2606 2605
  51   56  390  139  391   67   45   49  4
   3    2  220    9  220  230    7    4    4
 258  212  235  268  234  237    5  234
   0    0  151   65  150  140  633    7    7
 510  510 6225 3180 6172 6031  222  57
 221  221    1  234    1    1  225  222  22
 232  232    0  238    0    0  138  230  23
 148  148    0  135    0    0 6256  144  14
6279 6279    0 6126    0    0    1 6228  62
   1    1    0    1    0    0    0    1    1    1
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
```

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	1008
TOTAL NUMBER OF INSTANCES	18144
PROCESSING ATTRIBUTES NO :	1008 /1008
EVALUATING INSTANCE NO :	18144 /18144

Figure no: 6.4 Upload Dataset with 1000 tuples

The above figure shows that the simple data set read form the excel sheet. The above figure shows that the main page of the graphical user interface. First, we upload the simple database and second one is stream database in the data analysis. Apply the k-mean clustering approach for divide the data in the cluster forms .After that calculate the performance parameters i.e. accuracy.



NaN	NaN	11	25	10
61081	992	0	31	23
3712	264	0	42	18
NaN	332	0	27	33
NaN	0	0	11	49
36	35	0	14	28
60	65	0	18	12
NaN	47	11	17	4
0	47	0	13	61
0	53	0	11	7
39	92	0	15	12
34	1	39	12	19
18	0	48	11	198
10	0	51	34	276
2	11	40	25	97
1	0	50	18	95

Figure no. 6.5 Data values

The above figure represents that the simple dataset values in the list box.

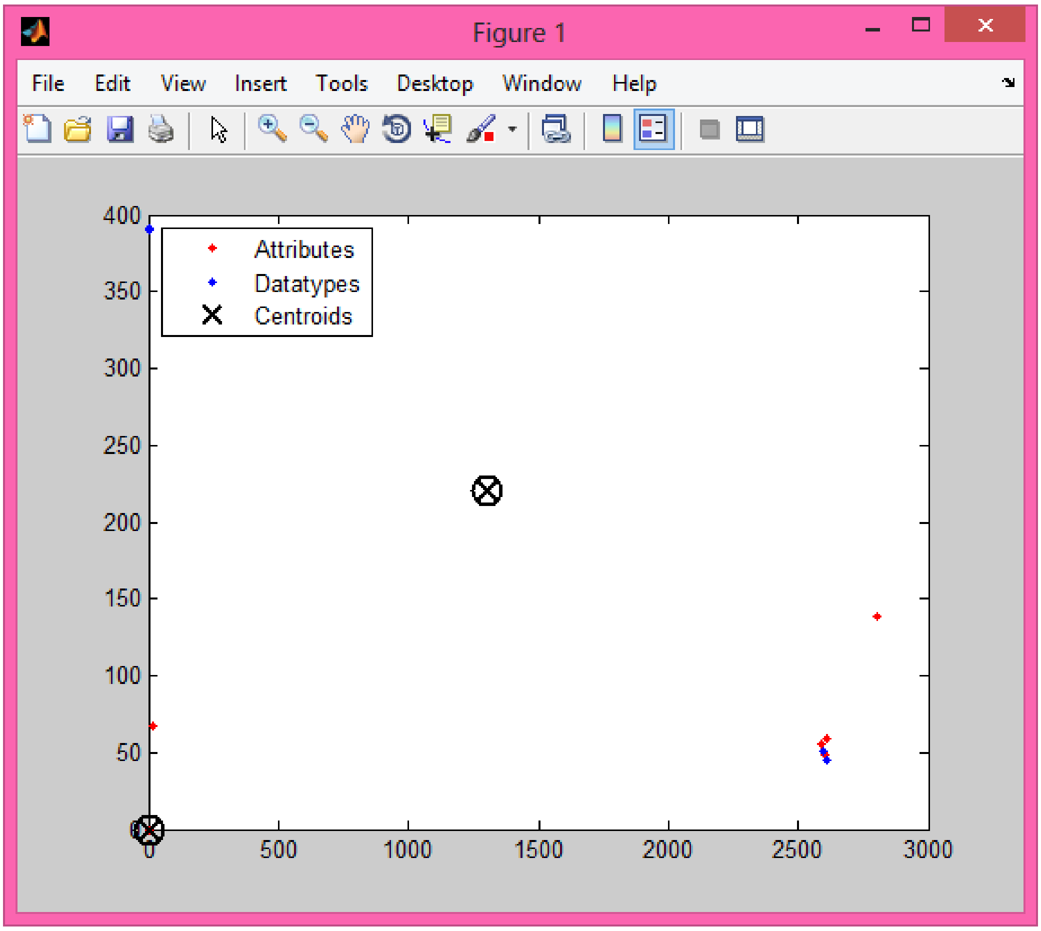


Figure no: 6.6 smaller dataset with 1000 tuples using k-means clustering

The above figures 6.6 show the cluster output of simple data set. The primary ideology is to signalize k centers, one for each group. These centers should be situated cleverly since of various area causes diverse outcome. Cover type data originated from US Forest Service inventory information, while the mapmaking factors used to conjecture cover type comprised of height, angle, and other data got from standard advanced information about space handled in a geographic data framework.

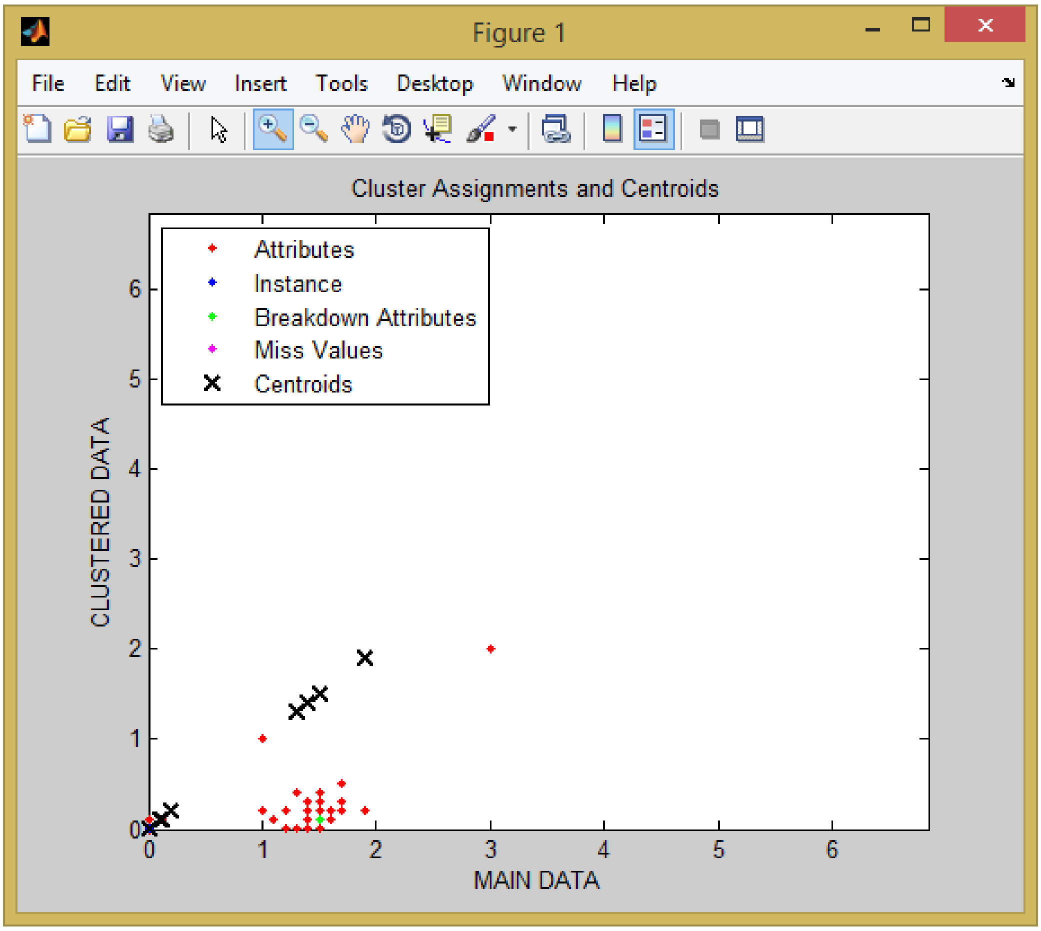
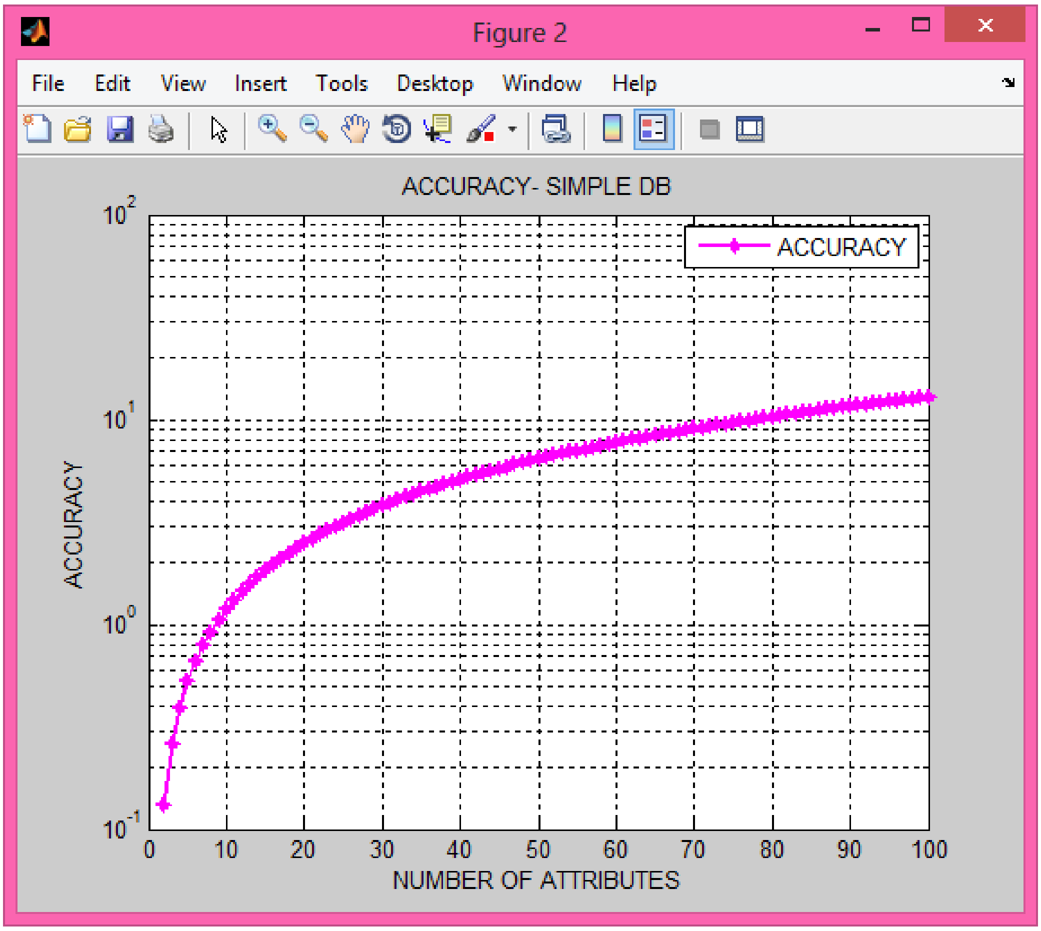
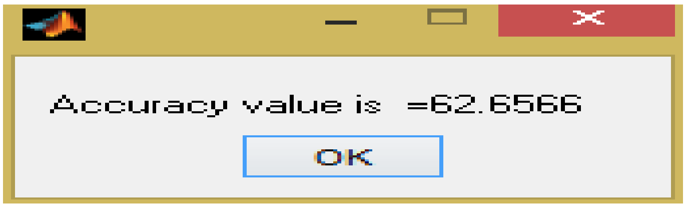


Figure 6.7: Clusters in smaller dataset with 1000 tuples using k-means clustering

Figure 6.7 represents the clusters that are obtained as a result of applying k means clustering to a set of data of forest cover. The different colored clusters represent different attributes or properties the clusters possess.



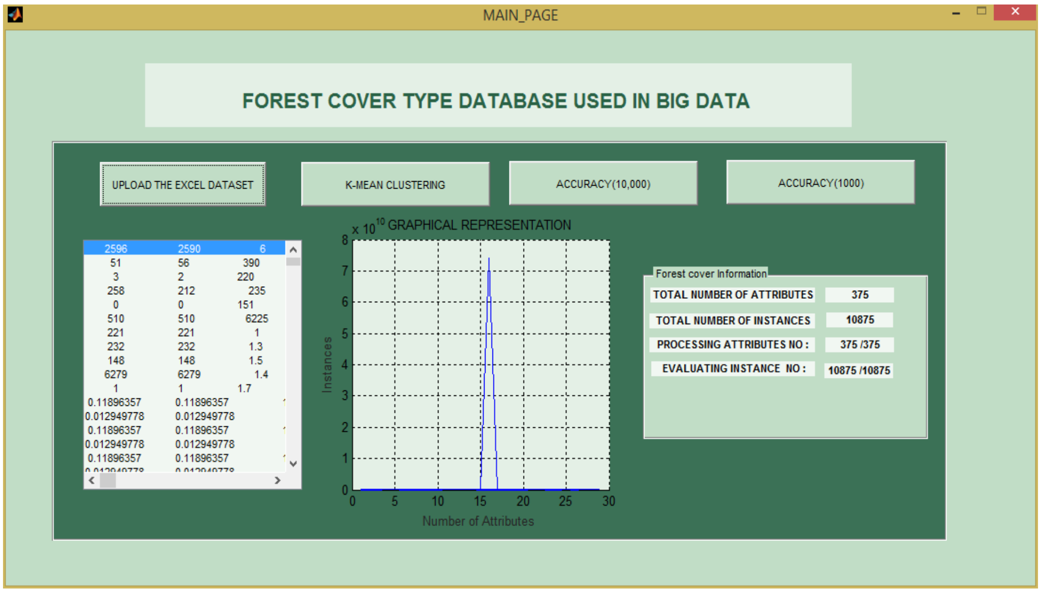


Accuracy value is =62.6566

Figure no: 6.8 Accuracy in Simple dataset with 1000 tuples through a graph

The above figure described that the accuracy define the existing work like 62% accuracy achieved. Accuracy is used to describe the closeness of a measurement to the true value.

Similarly, we have carried out the procedure for 10,000 tuples as well. Here, the accuracy is known to have increased by an appreciable amount since the number of tuples increased, so did the data and the attributes of the data provided. Hence, with the increase in the number of tuples, the result obtained is better.



2596	2590	6
51	56	390
3	2	220
258	212	235
0	0	151
510	510	6225
221	221	1
232	232	1.3
148	148	1.5
6279	6279	1.4
1	1	1.7
0.11896357	0.11896357	
0.012949778	0.012949778	
0.11896357	0.11896357	
0.012949778	0.012949778	
0.11896357	0.11896357	

TOTAL NUMBER OF ATTRIBUTES 375

TOTAL NUMBER OF INSTANCES 10875

PROCESSING ATTRIBUTES NO : 375 /375

EVALUATING INSTANCE NO : 10875 /10875

Figure 6.9: Upload dataset of 10,000 tuples

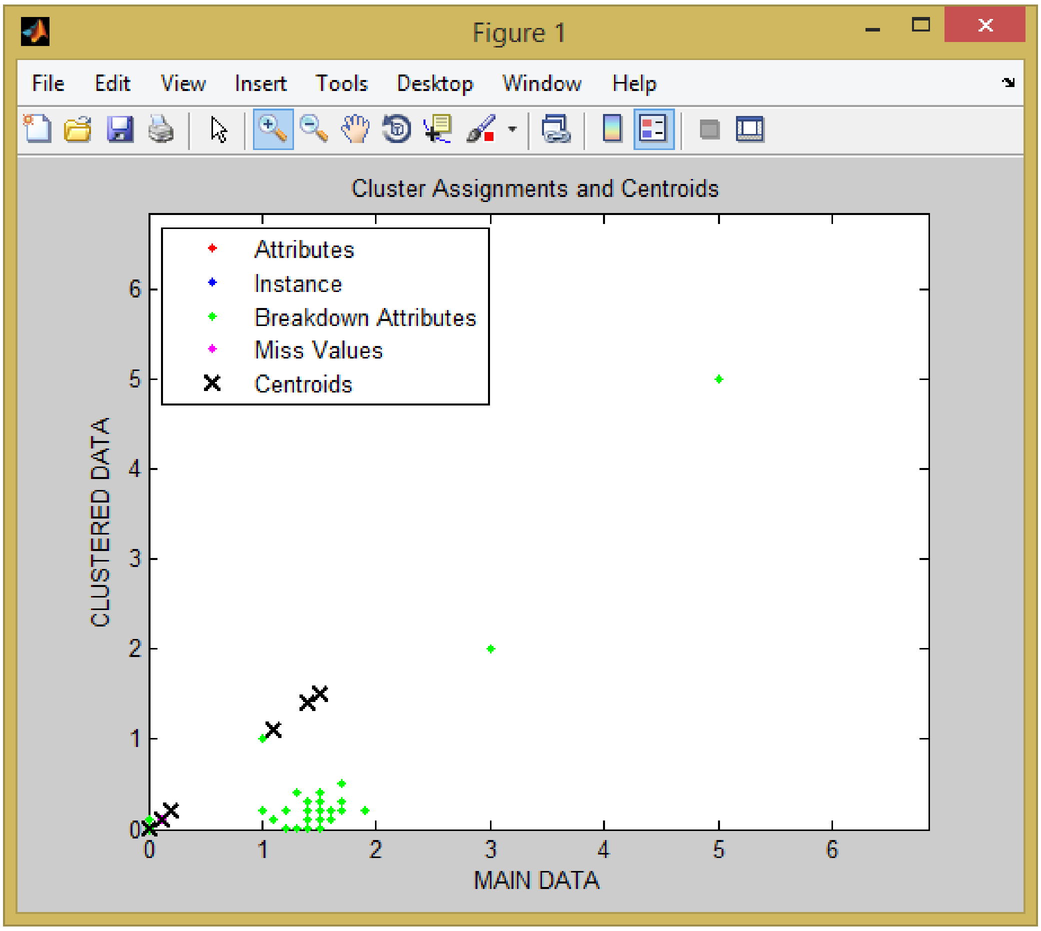
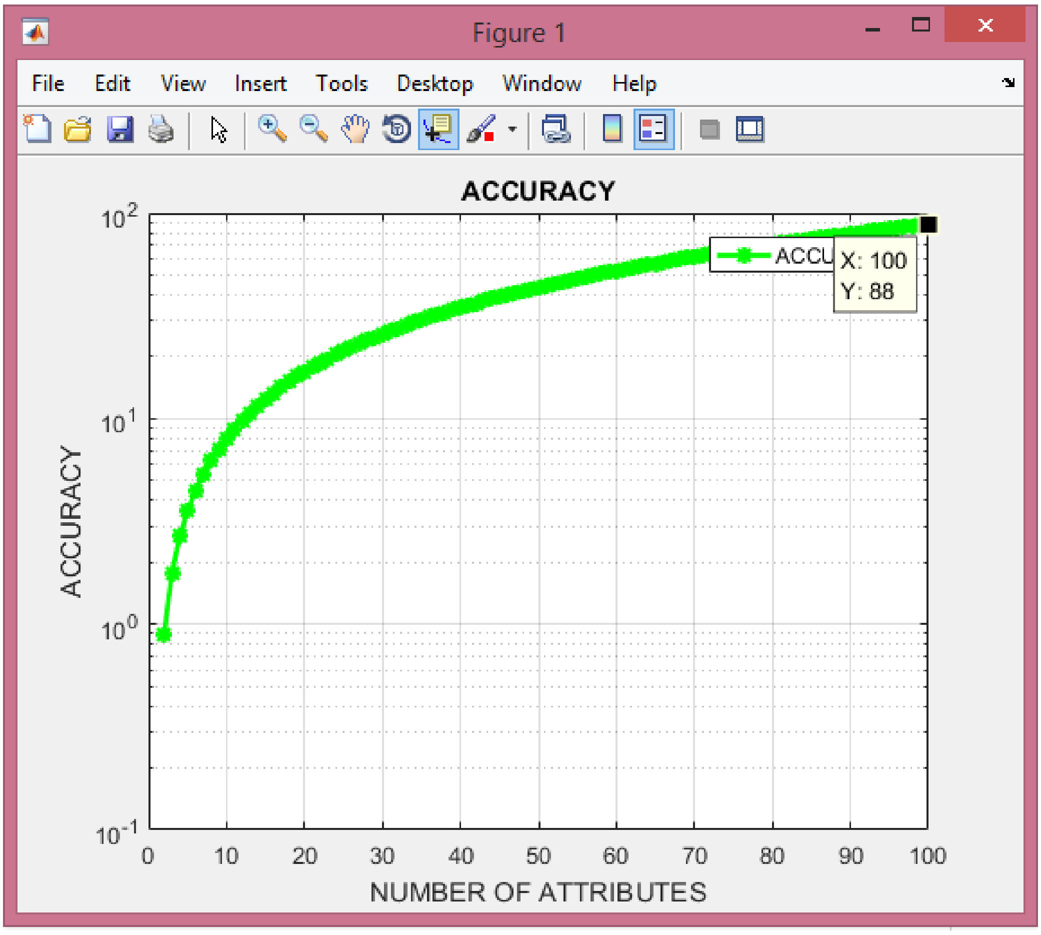
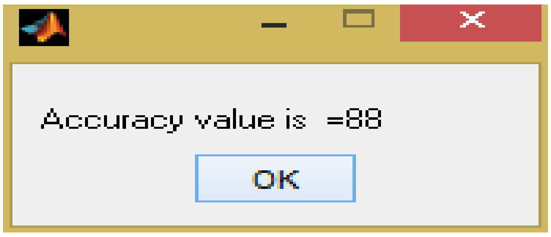


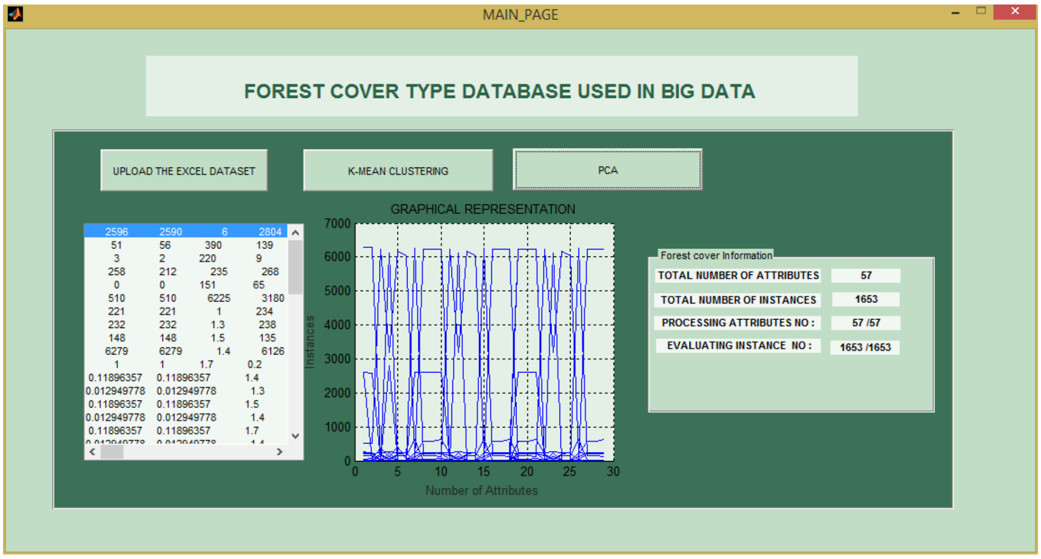
Figure 6.10: Clusters in larger dataset with 10000 tuples using k-means clustering





Accuracy value is =88

Figure 6.11: Accuracy in Simple dataset with 10000 tuples



2596	2590	6	2804
51	56	390	139
3	2	220	9
258	212	235	268
0	0	151	65
510	510	6225	3180
221	221	1	234
232	232	1.3	238
148	148	1.5	135
6279	6279	1.4	6126
1	1	1.7	0.2
0.11896357	0.11896357	1.4	
0.012949778	0.012949778	1.3	
0.11896357	0.11896357	1.5	
0.012949778	0.012949778	1.4	
0.11896357	0.11896357	1.7	

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES 57

TOTAL NUMBER OF INSTANCES 1653

PROCESSING ATTRIBUTES NO : 57 /57

EVALUATING INSTANCE NO : 1653 /1653

Figure 6.12: Upload dataset with 1000 tuples

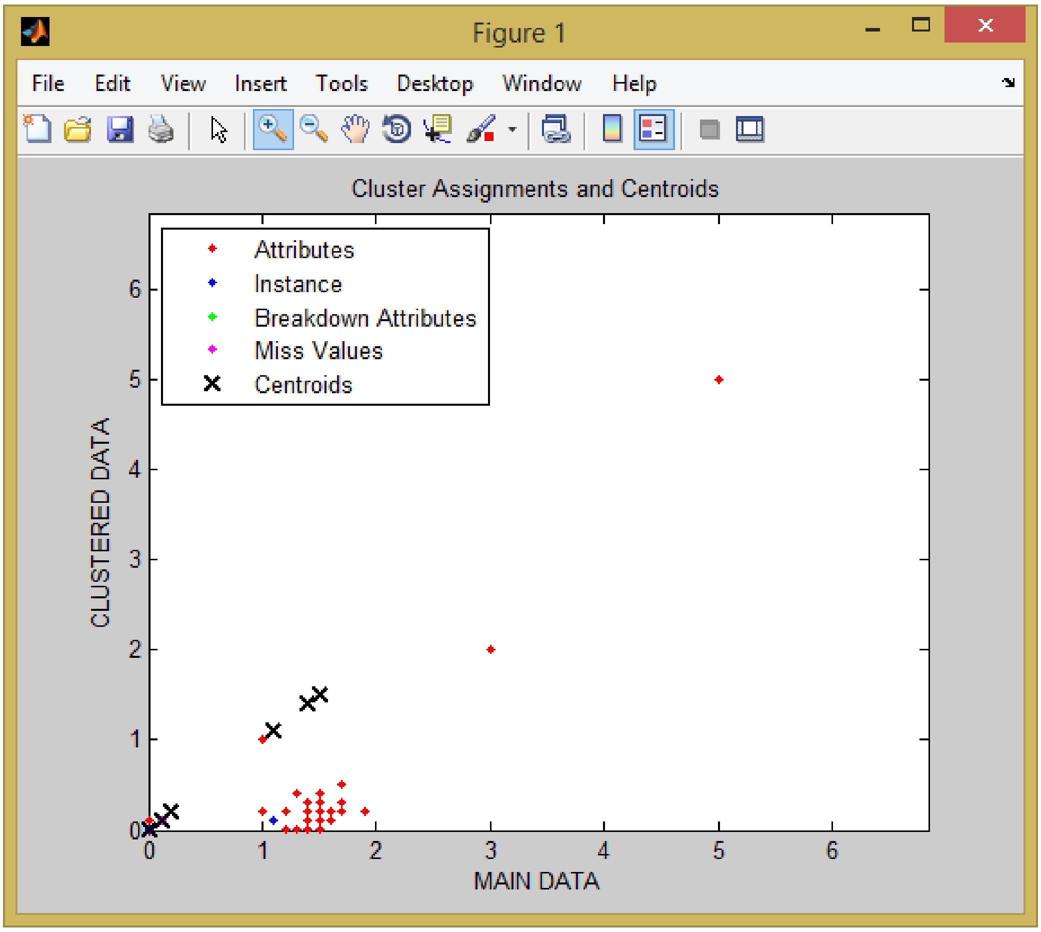


Figure 6.13: Clusters in 1000 tuples using k means clustering

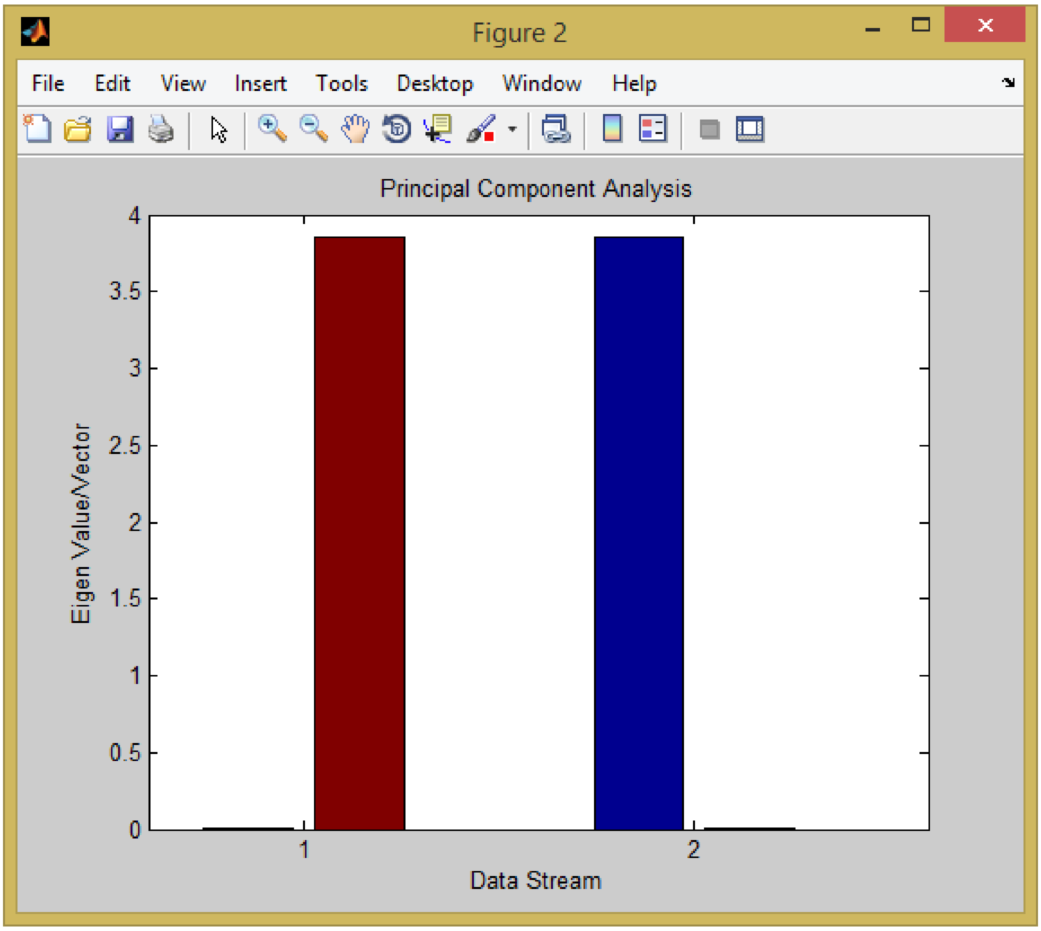
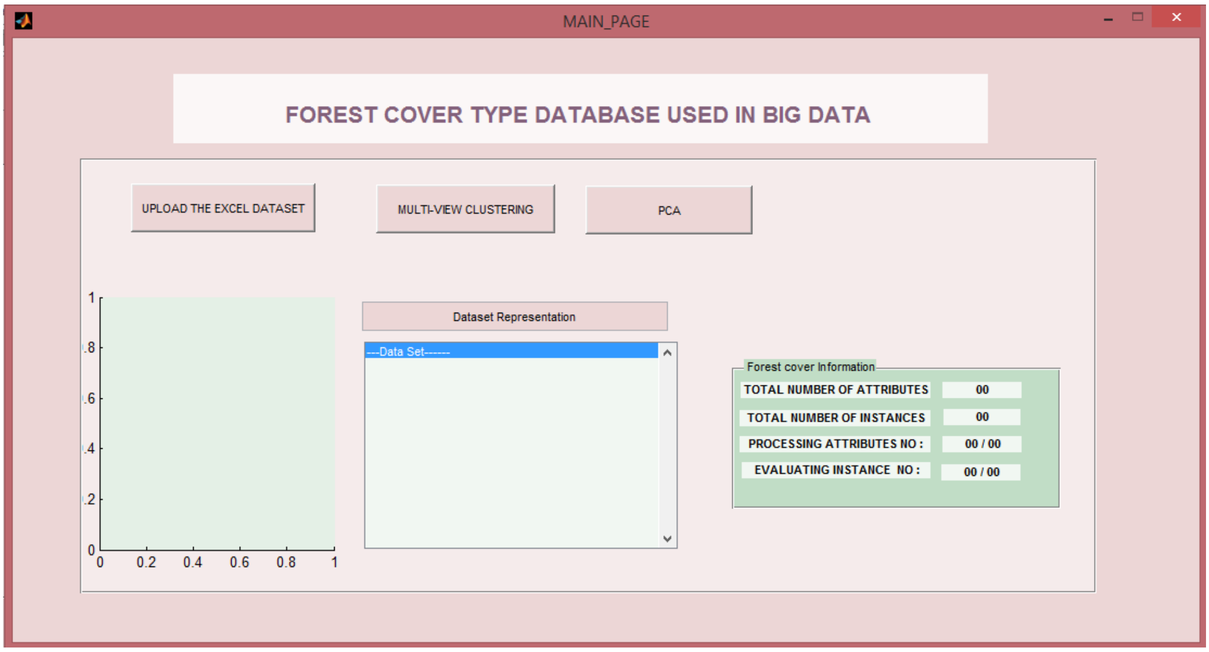


Figure 6.14: Bar graph representing eigen vectors/values w.r.t. the data stream using PCA algorithm

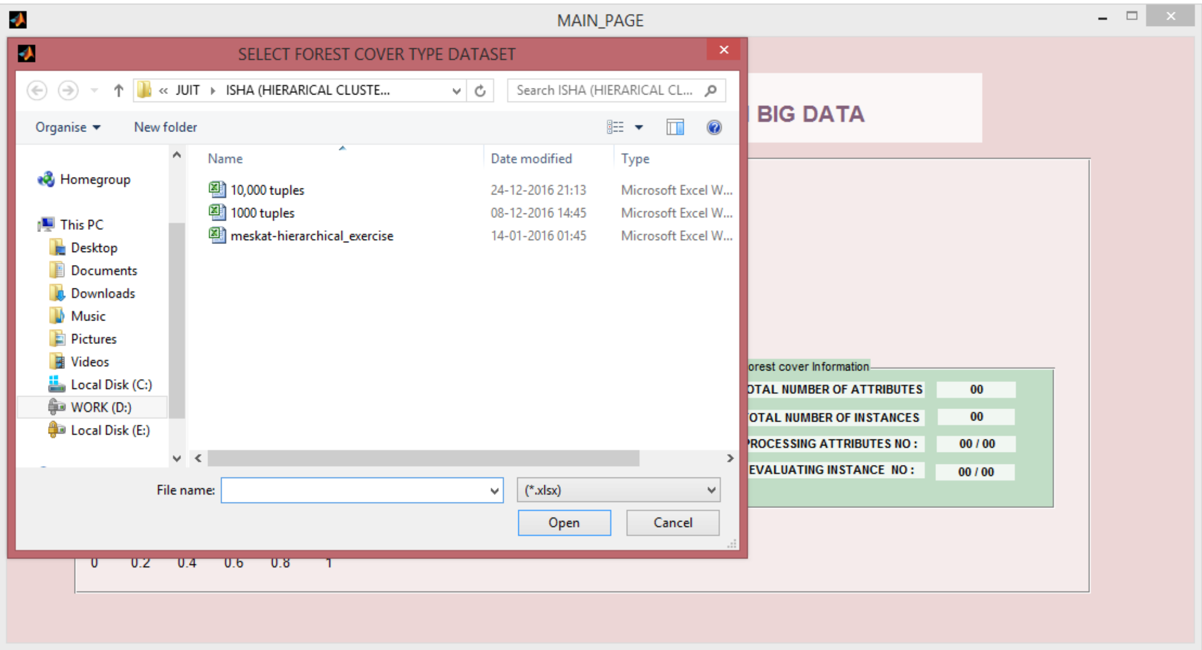
Now implementing the Principal Component Analysis algorithm on hierarchical or multiview clustering.



TOTAL NUMBER OF ATTRIBUTES	00
TOTAL NUMBER OF INSTANCES	00
PROCESSING ATTRIBUTES NO :	00 / 00
EVALUATING INSTANCE NO :	00 / 00

Figure 6.15: Main page

The above figure 6.15 discusses the uploading of the data stream dataset in excel sheet. In list box define the dataset values and axes shows the histogram to identify the high and low frequency of the data in the dataset in big data. Moving to next button, i.e. multi-view clustering and applied the PCA algorithm for component analysis.

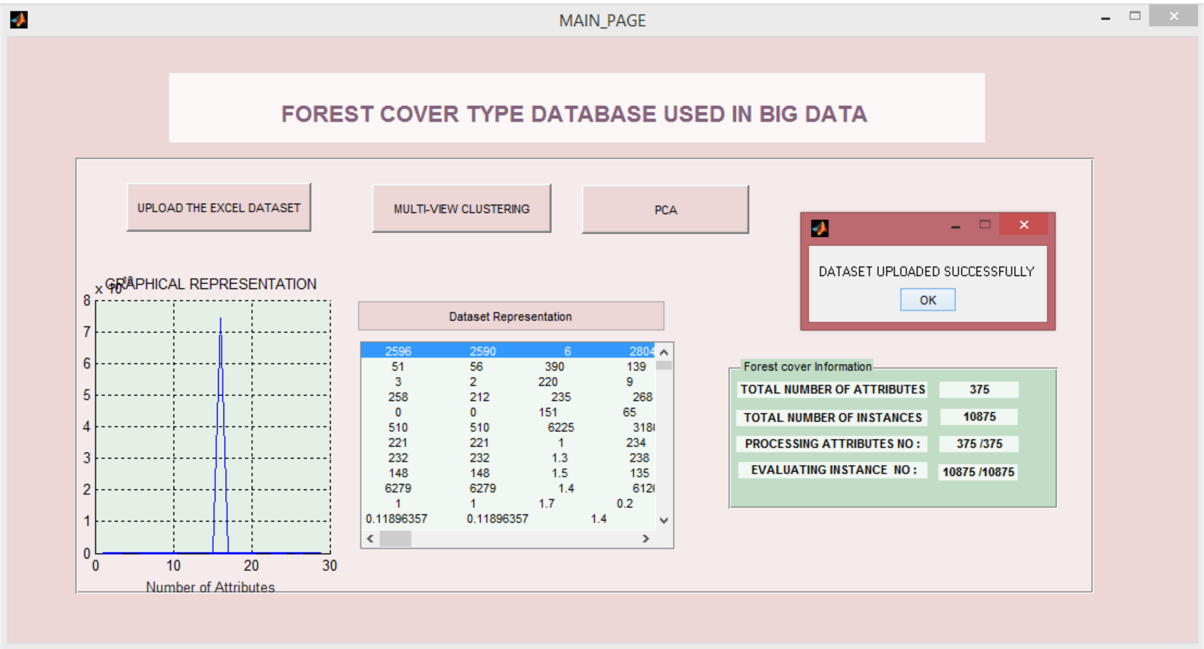


Name	Date modified	Type
10,000 tuples	24-12-2016 21:13	Microsoft Excel W...
1000 tuples	08-12-2016 14:45	Microsoft Excel W...
meskat-hierarchical_exercise	14-01-2016 01:45	Microsoft Excel W...

OTAL NUMBER OF ATTRIBUTES	00
OTAL NUMBER OF INSTANCES	00
ROCESSING ATTRIBUTES NO :	00 / 00
EVALUATING INSTANCE NO :	00 / 00

Figure 6.16: Upload the dataset

The above figure 6.16 shows that the upload the dataset from the excel sheet and read dataset i.e. file name and path name.



FOREST COVER TYPE DATABASE USED IN BIG DATA

DATASET UPLOADED SUCCESSFULLY

```
  2596        2590          6        2804
    51          56        390         139
     3           2        220           9
   258         212        235         268
     0           0        151          65
   510         510       6225        318(
   221         221          1         234
   232         232        1.3         238
   148         148        1.5         135
  6279        6279        1.4        612(
     1           1        1.7         0.2
0.11896357  0.11896357        1.4
```

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	375
TOTAL NUMBER OF INSTANCES	10875
PROCESSING ATTRIBUTES NO :	375 /375
EVALUATING INSTANCE NO :	10875 /10875

Figure 6.17: Dataset successfully uploaded

The above figure 6.17 defines the uploaded dataset values shown in edit text and static text UI-control tool. It shows the total number of attributes, instances and missing values. Histogram defines the dataset values in higher and lower frequency values in axes UI-control tool.

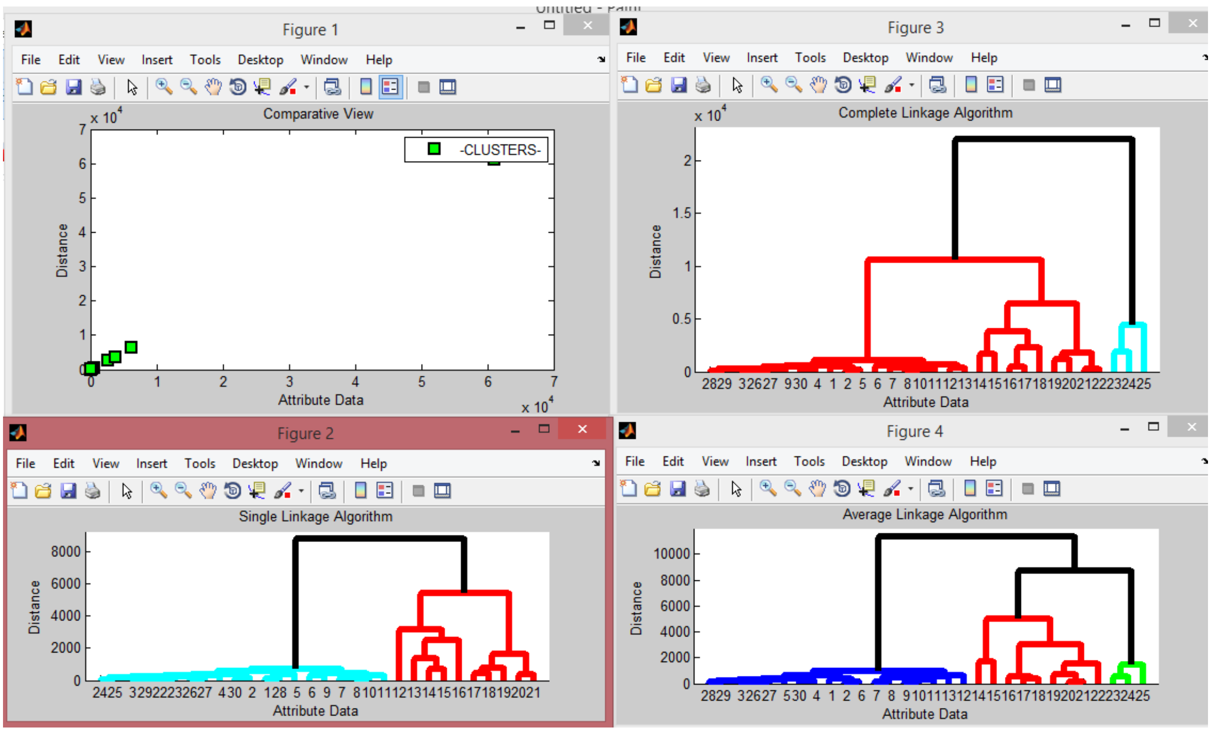


Figure 6.18: Multi-view clustering

The above figure defines that the multi-view clustering results in four graphs.

(i) First one shows the clusters in the data stream
(ii) Single linkage algorithm
(iii) Complete linkage algorithm
(iv) Average linkage algorithm

The multi-view clustering consists of four steps:

1. **Finding the same or dissimilar between every set of objects in the data set:** In this step, we calculatet the separation between objects using the pdist function. The pdist work keeps up various approaches to figure this amount.
2. **Grouping the items into a binary, multi-view cluster tree.** In this progression, we interface sets of items that are in nearness utilizing the linkage function. The association function uses the distance data created in step 1 to decide the closeness of those articles to each other. As occasions are matched into parallel bunches, the oddity shaped groups are assembled into bigger groups until a multi-view tree is framed.
3. **Determining where to split the multi-view tree into clusters.** In this progression, we utilize the cluster function to prune branches out the base of the multi-view tree, and allocate every one of the items underneath each slice to a solitary group. This makes a segment of the information. The group capacity can make these groups by recognizing characteristic groupings in the multi-view tree or by removing the multi-view tree at a discretionary point.

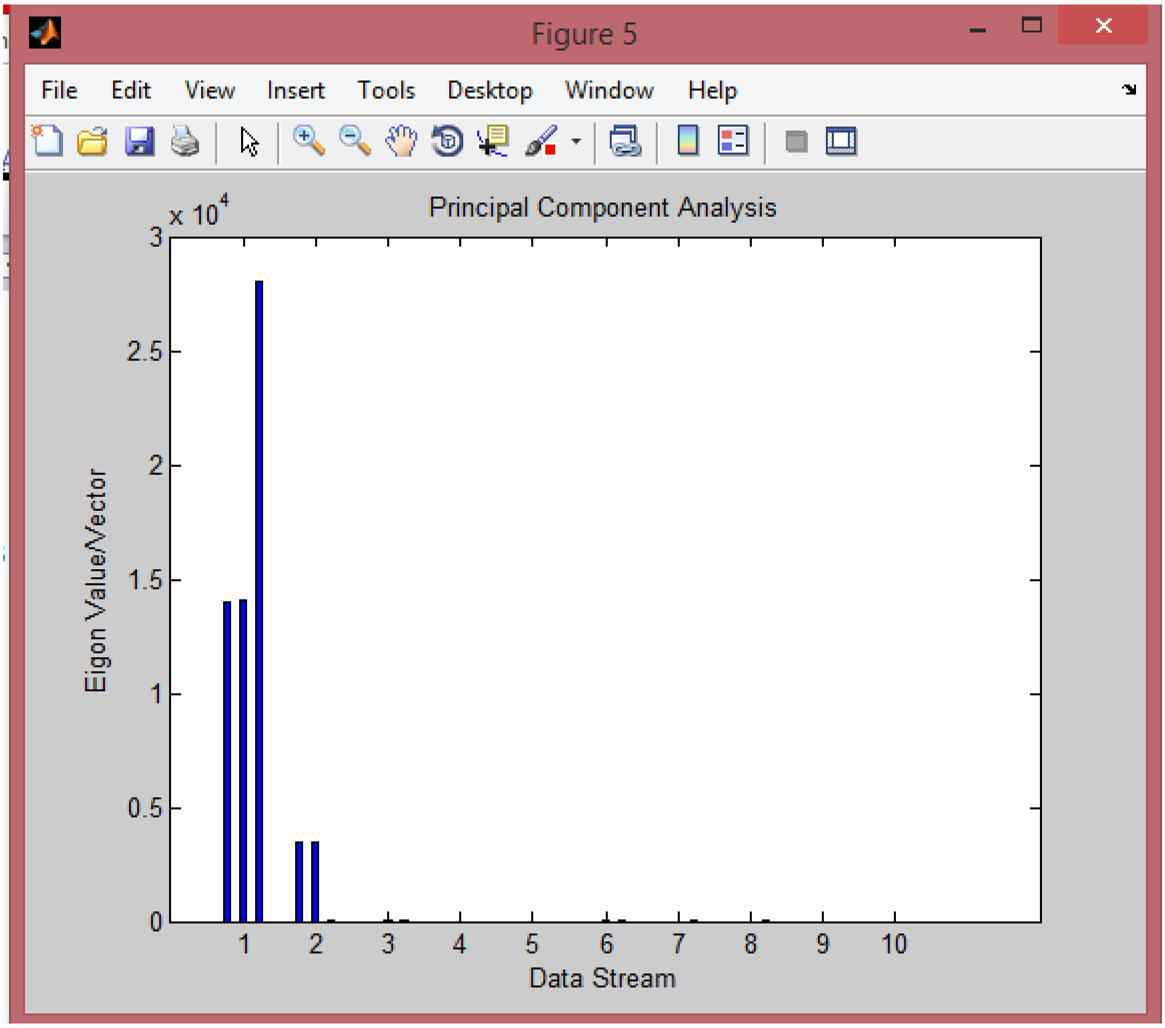


Figure 6.19: PCA Algorithm on hierarchical clustering

The above figure defines the feature extraction using principle component analysis algorithm used for identify the unique data i.e. called features in data stream dataset in big data. The feature extraction technique is used to recover the most revealing terms from amount of matrix. This study uses Component Analysis technique to calculate and study the Eigen vector and values to find the feature values and then to direct individual data with its principal components / Eigen Vectors.

In previous work of applying the PCA algorithm to k-means clustering, we implement the principle component analysis algorithm used for identify the feature extracted values. We identify the huge values to extract in 10,000 and 1000 tuples. In proposed work, we implement the PCA algorithm after multi-view clustering approach for minimum extracted component found.

The thick bar graph in figure 6.14 represents the huge values in extracted feature that implies that less accuracy is achieved. Component analysis is difficult to predict the k type of values in K-means clustering algorithm. It didn't work properly and is difficult to initialize division and could result in dissimilarity in final clusters. It doesn't work completely with clusters of different sizes and densities.

In proposed work we actualize the Multi-view grouping to deliver a sequence of the items, which might be educational for information show. Little subgroups are created, which might be useful for removing valuable data and social affair helpful bits of knowledge. In PCA calculation implies absence of excess of information given the orthogonal parts and decreased multifaceted nature in dataset gathering with the utilization of PCA. It produces the thin diagram in figure 6.19 that implies exactness has enhanced considerably. It generates the thin graph in figure 6.19 that means accuracy has improved appreciably.

Further, we implement Support Vector Machine (SVM) as a classifier in order to classify the data that is uploaded and make a detailed comparison when SVM is applied to both k means clustering and hierarchical clustering and compile the results.

The data i.e. training data when entered is represented as a set of points in space and these points are segregated on the basis of their categories or groups. The gap should be as wide as possible and should be as significantly visible as possible for better results to come in. The separation is done either with a line in 2D or a plane in 3D space. When the new data flows in it is characterized by as to on which side of the line or plane of the gap the newly entered points fall in. Support Vector Machine can be split into two types-

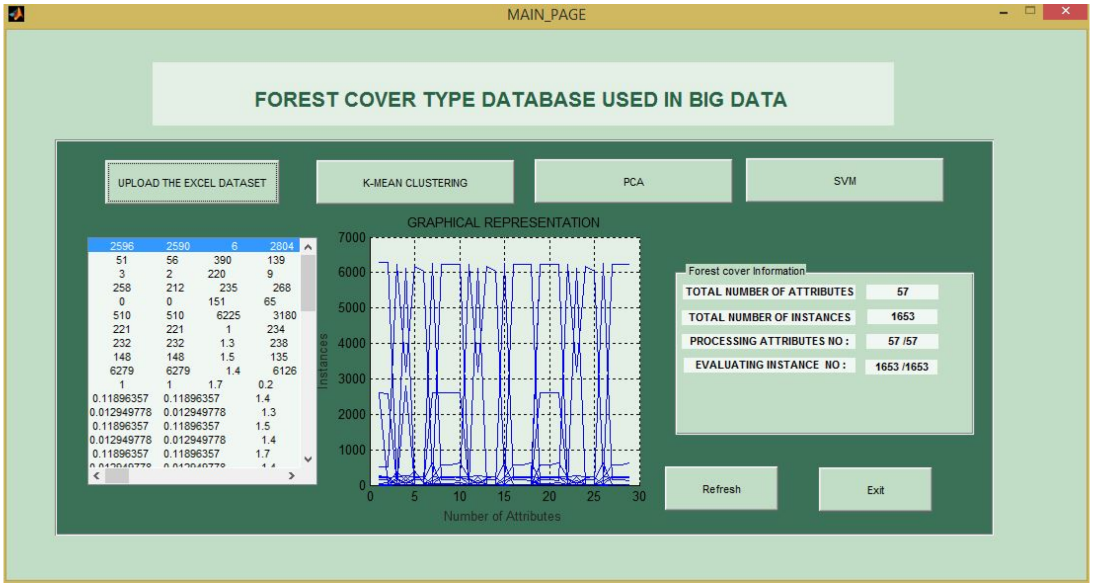
1. Linear classification
2. Non-linear classification

The function 'SVM struct' is used to train the data and verify it using kernel information. The training phase is divided into two groups namely-

Group 0 and Group 1

The data that is classified using SVM is re-verified as to whether and by what amount is the data properly trained based on which a report is compiled and the parameters are evaluated as the output.

The application of SVM on multi view clustering is better than application of
k- means clustering in the ways mentioned below-
The k -means clustering approach just determines the attributes of the various entities and how the grouping takes place in the dataset whereas SVM on multi view clustering involved various phases like normalization, clustering, and complete linkage and then followed by classification and comparison of parameters.



2596	2590	6	2804
51	56	390	139
3	2	220	9
258	212	235	268
0	0	151	65
510	510	6225	3180
221	221	1	234
232	232	1.3	238
148	148	1.5	135
6279	6279	1.4	6126
1	1	1.7	0.2
0.11896357	0.11896357	1.4	
0.012949778	0.012949778	1.3	
0.11896357	0.11896357	1.5	
0.012949778	0.012949778	1.4	
0.11896357	0.11896357	1.7	

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	57
TOTAL NUMBER OF INSTANCES	1653
PROCESSING ATTRIBUTES NO :	57 /57
EVALUATING INSTANCE NO :	1653 /1653

Figure 6.20: Upload a data set of 1000 tuples

The above figure depicts the uploading of the dataset with 1000 tuples from the forest cover type dataset used for analysis and classification.

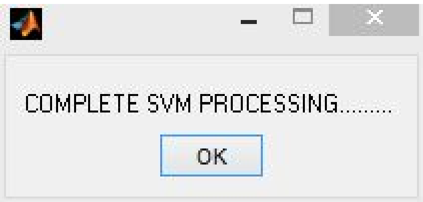

COMPLETE SVM PROCESSING.........

Figure 6.21: Apply SVM

The message box represents the classification used in support vector machine. The processing of the SVM algorithm that consists of two phases namely training and testing phase is complete.

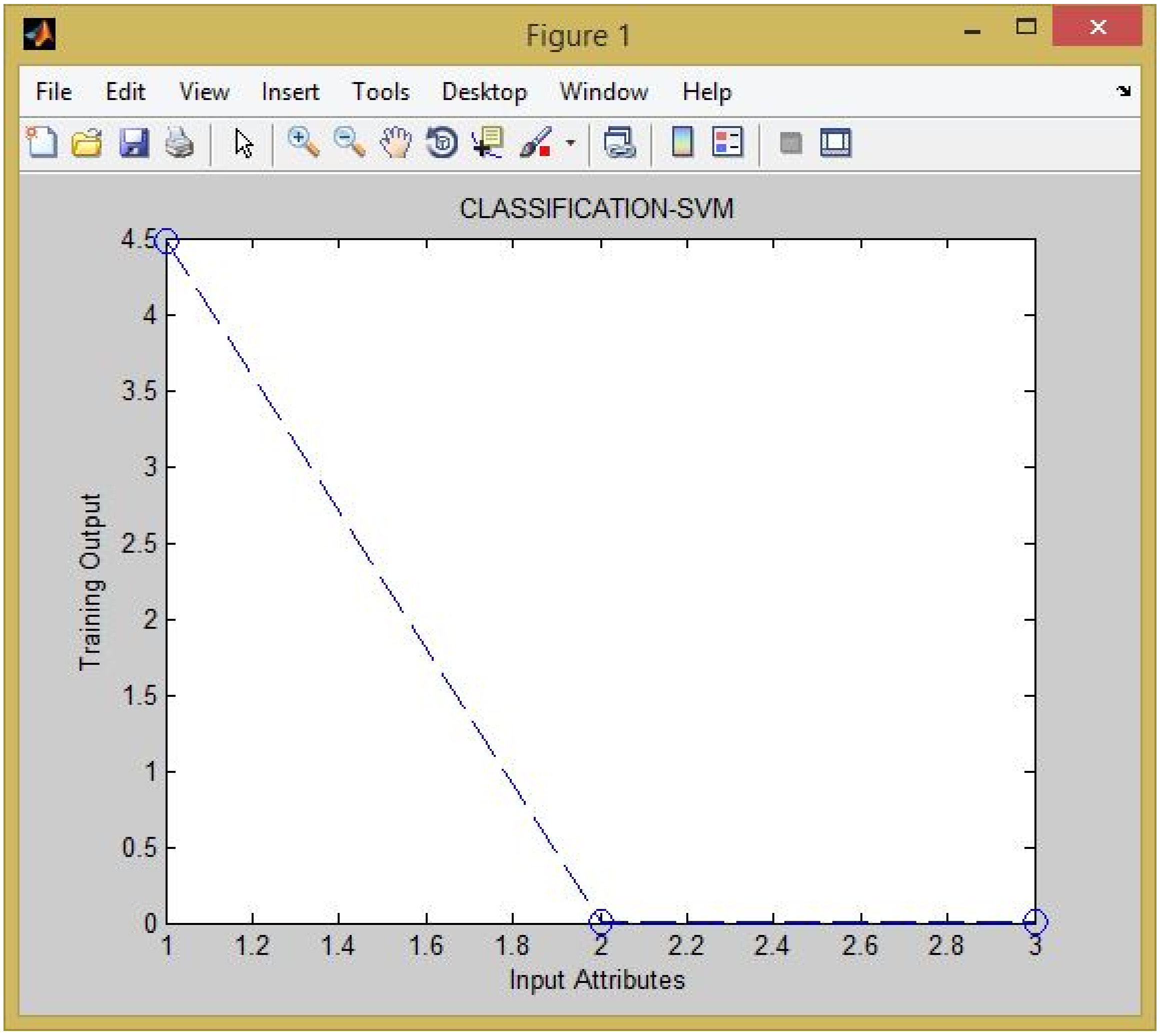
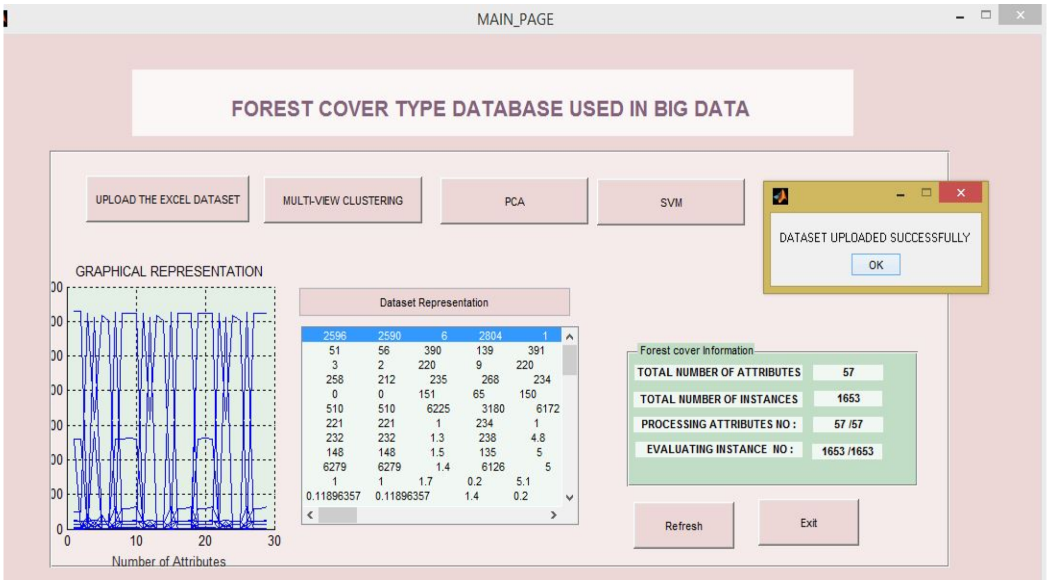


Figure 6.22: SVM application on K means clustering

The above figure depicts the support vector machine algorithm on K-mean clustering algorithm. Earlier it was applicable for linear two-class classification with some gap, where the gap depicted the minimum separation from the segregating hyperplane to the nearest data points.



FOREST COVER TYPE DATABASE USED IN BIG DATA

DATASET UPLOADED SUCCESSFULLY

Dataset Representation

2596	2590	6	2804	1
51	56	390	139	391
3	2	220	9	220
258	212	235	268	234
0	0	151	65	150
510	510	6225	3180	6172
221	221	1	234	1
232	232	1.3	238	4.8
148	148	1.5	135	5
6279	6279	1.4	6126	5
1	1	1.7	0.2	5.1
0.11896357	0.11896357		1.4	0.2

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	57
TOTAL NUMBER OF INSTANCES	1653
PROCESSING ATTRIBUTES NO :	57 /57
EVALUATING INSTANCE NO :	1653 /1653

Figure 6.23: Upload data set of 1000 tuples

The above figure represents the uploading of the dataset with 1000 tuples from the forest cover type dataset used with multi-view clustering algorithm.

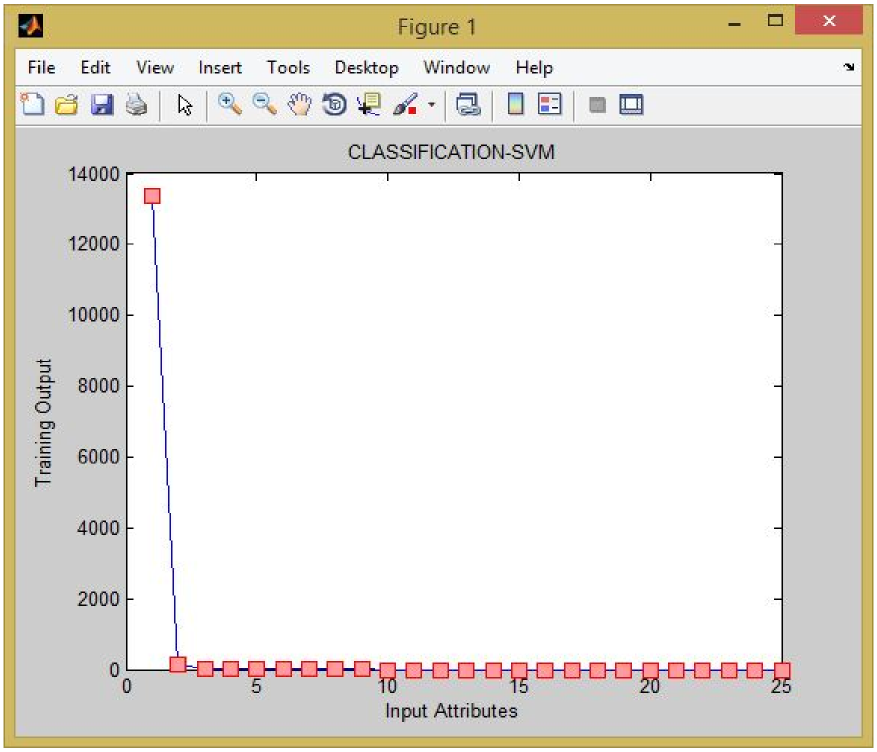
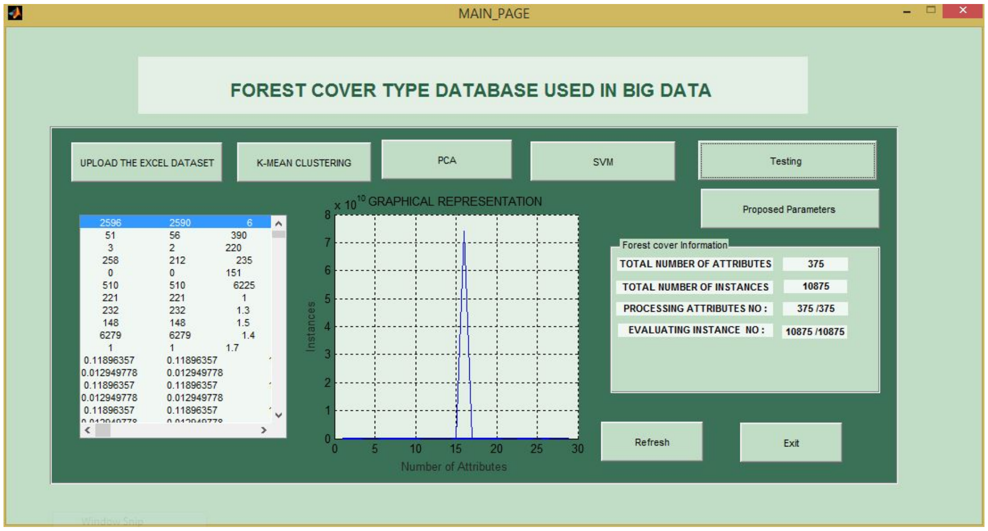


Figure 6.24: Application of SVM on Heirarchical clustering

The above figure 6.24 defines the application of support vector machine on hierarhical clustering. SVM learning machine looks for an ideal isolating hyper-plane, where the edge is most extreme with the end goal that the margin isolating both the sets is as substantial as could be expected under the circumstances. An incomparable and unmistakable component of this approach is that the arrangement is constructed just with respect to those information focuses, which are at the fringe. These focuses are known as support vectors. The line machine learning with SVM can be stretched out to non-straight one when first the issue is changed into a trademark space utilizing an arrangement of non-linear premise capacities. In the component space which can be high dimensional - the information focuses can be isolated straightly. A huge plus point of the SVM is that it is not required to execute this change and to decide the isolating hyperplane in the conceivably high dimensional element space.

Instead of this, a bit portrayal can be utilized, where the arrangement is composed as a weighted entirety of the estimations of certain piece of kernel function assessed at the support vectors.



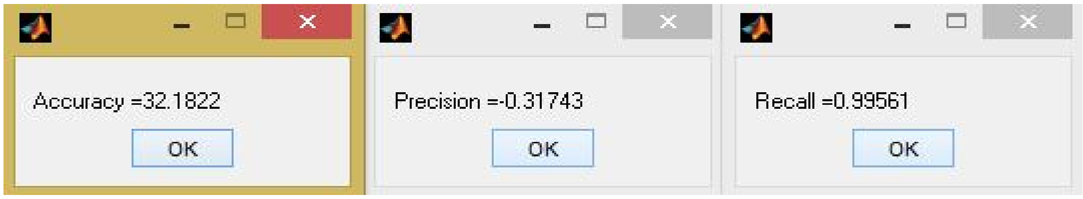
```
2596        2590        6
51          56          390
3           2           220
258         212         235
0           0           151
510         510         6225
221         221         1
232         232         1.3
148         148         1.5
6279        6279        1.4
1           1           1.7
0.11896357  0.11896357
0.012949778 0.012949778
0.11896357  0.11896357
0.012949778 0.012949778
0.11896357  0.11896357
```

Forest cover Information

TOTAL NUMBER OF ATTRIBUTES	375
TOTAL NUMBER OF INSTANCES	10875
PROCESSING ATTRIBUTES NO :	375 /375
EVALUATING INSTANCE NO :	10875 /10875

Figure 6.25: Upload 10000 tuples for testing and calculating parameters

The above figure defines the upload of the dataset with 10000 tuples from the forest Cover type dataset used for multi-view clustering algorithm.


Accuracy =32.1822

Precision =-0.31743

Recall =0.99561

Figure 6.26: Parameters evaluated: Accuracy, Precision, and Recall Rate

The above figure defines the k-mean clustering algorithm used with performance parameters that is accuracy, precision and recall rate.

The result obtained by calculation of performance parameters and their values are: Accuracy value is 32.188, Precision value is 0.317 and Recall Rate value is 0.99561.

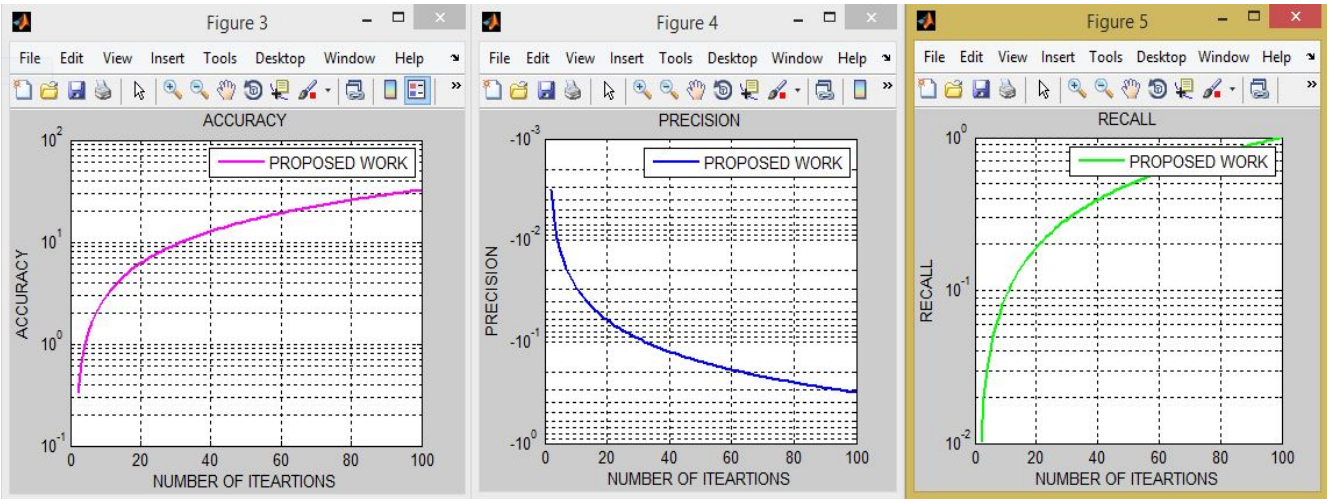
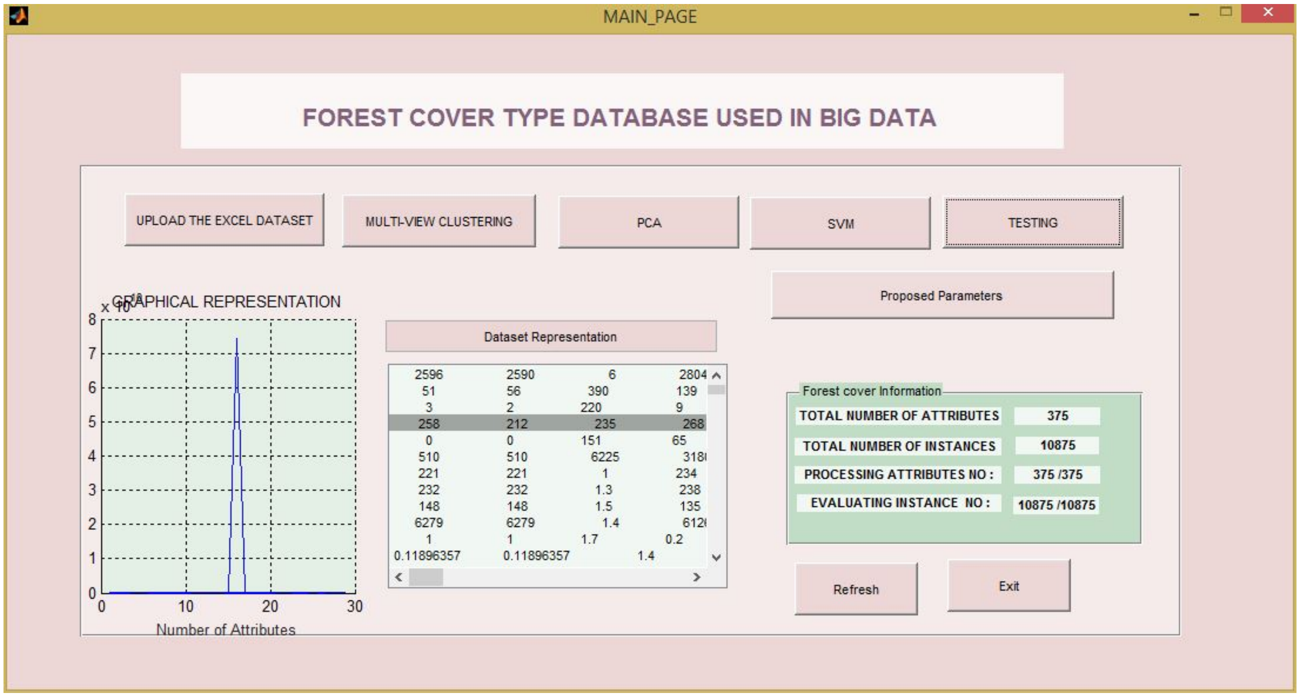


Figure 6.27: Graphs plotted for Accuracy, Precision and Recall Rate

The above figure represents that the number of iterations which evaluates the performance with accuracy, precision and recall with K-mean clustering algorithm.



```
  2596        2590           6        2804
    51          56         390         139
     3           2         220           9
   258         212         235         268
     0           0         151          65
   510         510        6225        318
   221         221           1         234
   232         232         1.3         238
   148         148         1.5         135
  6279        6279         1.4        612
     1           1         1.7         0.2
0.11896357  0.11896357         1.4
```

TOTAL NUMBER OF ATTRIBUTES: 375

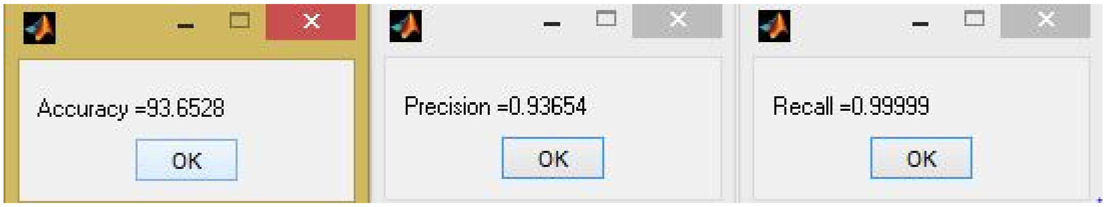
TOTAL NUMBER OF INSTANCES: 10875

PROCESSING ATTRIBUTES NO : 375 /375

EVALUATING INSTANCE NO : 10875 /10875

Figure 6.28: Upload data set of 10000 tuples for testing and evaluating parameters for multiview clustering

The above figure represents that the testing phase with Multi-view clustering and support vector machine. In the testing phase, we upload testing phase in to identify the feature to compare with the training phase in the feature extraction algorithm. If feature compared with training and testing phase, then evaluate the performance parameters.



Accuracy =93.6528

Precision =0.93654

Recall =0.99999

Figure 6.29: Parameters evaluated: Accuracy, Precision, and Recall Rate

Here we have evaluated the performance parameters with Multi-view clustering algorithm and Support vector machine. The proposed parameters and their values are as follows: Accuracy value is 93.65, Precision values is 0.93654 and Recall Rate value is 0.999.

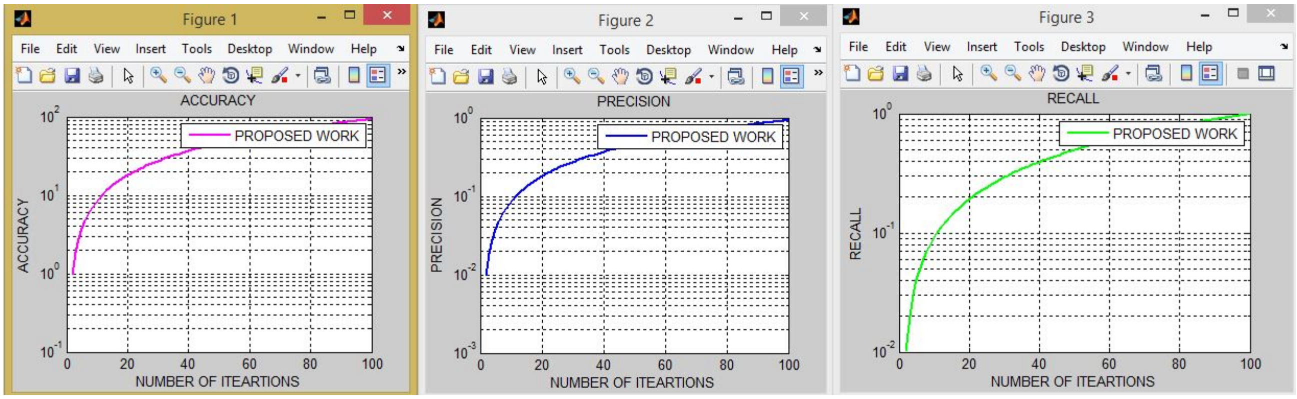


Figure 6.30: Graphs plotted for Accuracy, Precision and Recall Rate

The above figure represents the graph plotted for accuracy, precision and recall rate. The precision esteem is the portion of recovered instances that are applicable, while recall rate (otherwise called sensitivity) is the division of significant cases that are recovered. Both precision and recall are in this way in light of overseeing and measure of importance. Given an arrangement of information focuses from a progression of estimations, the set can be said to be exact if the qualities are near the normal estimation of the amount being measured, while the set can be said to be precise if the qualities are near the genuine estimation of the amount being measured.

The two ideas are individual of each other, so a specific arrangement of information can be said to be either accurate, or precise, or both, or not one or the other.

CHAPTER 7

CONCLUSION

The research and the experimentation concluded the ability of a multi view clustering algorithm in contrast to a k means clustering algorithm when applied to a forest cover to predict the various classes in a particular dense forest. For instance, the forest comprises of various tress like oak, deodar, teak etc. This algorithm helps us to classify the various trees. In addition to it, we also get to know the attributes of the distinct objects in a class. Here, we introduce the advantages of multi view clustering i.e. hierarchical clustering in comparison to k means clustering where we get to know the various perspectives in a heterogeneous system. Based on this approach, we calculate our performance parameters. The k-means clustering model produced appreciable clustering or classification results, with accuracy way more than those produced by the DB analysis. Next, we applied the Principal Component Analysis both on k-means clustering and the hierarchical clustering. The results obtained show us the drastic improvements that occur in case of Multiview clustering over k-means clustering by the means of a bar graph and evaluated it using accuracy, precision and recall rate.

CHAPTER 8

FUTURE SCOPE

Besides, the dataset utilized for trial purposes is far bigger than anything exists in writing, to the best of our insight. In spite of the fact that the strategy has been utilized for consecutive successive and non-visited things, we have demonstrated that it can likewise be utilized as a part of big data streams progressively, something that to the best of our insight does not exist in writing up until now.

REFERENCES

[1] Y. Chen , S. Alspaugh, and R. Katz,"Interactive analytical processing in big data systems: A cross-industry study of map reduce workloads", Proc. VLDB Endowment, vol. 5, no. 12, pages-1802-1813, 2012.

[2] Y. Demchenko, P. Membrey,"Defining Architecture Components of the Big Data Ecosystem", In *Collaboration Technologies and systems (CTS)*, pages-104- 112, 2014.

[3] Rider, Fremont, The Scholar and the Future of the Research Library. Page-236. New York, Hadham Press, 1944.

[4] Cheng, Otto KM, and Raymond Lau. "Big Data Stream Analytics for Near Real-Time Sentiment Analysis." *Journal of Computer and Communications*3, no. 05 (2015): 189.

[5] Guo, Xi. "Application of meteorological big data." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 273-279. IEEE, 2016.

[6] Vinay Kumar Jain and Shishir Kumar, "Big Data Analytic Using Cloud Computing", in *2015 Second International Conference on Advances in Computing and Communication Engineering*, pp.667- 672, 2015

[7] Xue Shengjun, Liu Yin, "Establishment and Test of Meteorological Data Warehouse Based on Hadoop", in *2015 Computer Measurement and Control*, vol.20, no.4, pp.926-928, 2012

[8] Yan Yonggang,"Research of KNN Classification Based on Hadoop in Precipitation", Technical Research Report, Nanjing University of Information Science and Technology, P457.6, 2013.

[9] Su, Fei, Yi Peng, Xu Mao, Xinzhou Cheng, and Weiwei Chen. "The research of big data architecture on telecom industry." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 280-284. IEEE, 2016.

[10] Xylogiannopoulos, Konstantinos F., Reda Alhajj, and Panagiotis Karampelas. "Frequent and non-frequent pattern detection in big data streams: An experimental

simulation in 1 trillion data points." In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 931-938. IEEE, 2016.

[11] Yu, Yuan-Chih, and Dwen-Ren Tsai. "A privacy weaving pipeline for open big data." In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 997-998. IEEE, 2016.

[12] VijiyaKumar, K., V. Govindasamy, and T. Esther. "An online big data take oution using latent dirichlet allocation." In *Communication and Signal Processing (ICCSP), 2016 International Conference on*, pp. 2278-2283. IEEE, 2016.

[13] Dhaka, Priyanka, and Rahul Johari. "Big Data Application: Study and Archival of Mental Health Data, using MongoDB 2015."

[14] Vinod, D. Franklin, and V. Vasudevan. "A Filter Based Feature Set Selection Approach for Big Data Classification of Patient Records", 2015.

[15] Gao, Jie, Xinzhou Cheng, Lexi Xu, and Haina Ye. "An interference management algorithm using big data analytics in LTE cellular networks." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 246-251. IEEE, 2016.

[16] Ye, Haina, Xinzhou Cheng, Mingqiang Yuan, Lexi Xu, Jie Gao, and Chen Cheng. "A survey of security and privacy in big data." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 268-272. IEEE, 2016.

[17] Lu, Xin, Fei Su, Haozhang Liu, Weiwei Chen, and Xingzhou Cheng. "A unified OLAP/OLTP big data processing framework in telecom industry." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 290-295. IEEE, 2016.

[18] Jia, Yuwei, Kun Chao, Xinzhou Cheng, Mingqiang Yuan, and Mingjun Mu. "Big data based user clustering and influence power ranking." In *Communications and Information Technologies (ISCIT), 2016 16th International Symposium on*, pp. 371-375. IEEE, 2016.

[19] Matthew A. Waller and Stanley E. Fawcett, “Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management,” in *Journal of Business Logistics,* 2013, pp. 77-84.

[20] Amir Gandomi and Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics," in *Elsevier,* 2015, pp. 137-144.

[21] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, "Big Data: Issues and Challenges Moving Forward" in *IEEE,* 2012, pp. 1530-1605.

[22] Jens Dittrich JorgeArnulfo and Quian´eRuiz, "Efficient Big Data Processing in Hadoop MapReduce" in *International Conference on Very Large Data Bases*, Vol. 5, No. 12, 2012, pp. 469-494.

[23] Hai Wang , Zeshui Xu, Hamido Fujita and Shousheng Liu, "Towards Felicitous Decision Making: An Overview on Challenges and Trends of Big Data," in *Information Sciences*, 2016, pp. 367-394.

[24] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, "Data Mining with Big Data" in *IEEE,* Vol. 26, No. 1, 2016, pp. 1041-4347.

[25] Chunhe Shi, Chengdong Wu, Xiaowei Han4, Yinghong Xie5 and Zhen Li, "Machine Learning under Big Data," in *International Conference on Electronic, Mechanical, Information and Management,* 2016, pp. 1061-1064.

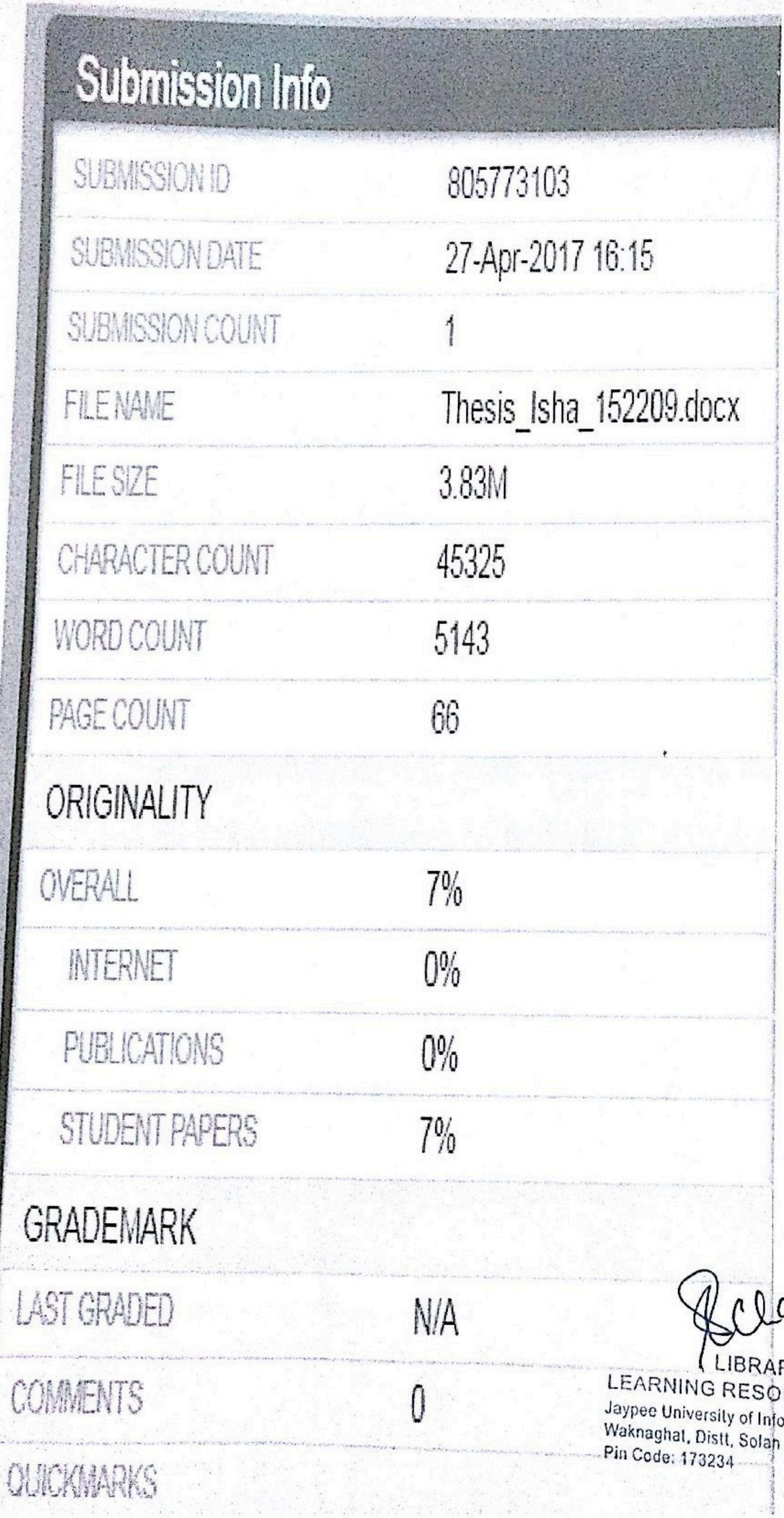
[26] Shan Suthaharan,"Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning," in *ACM,* Vol. 41, No. 4, 2014, pp. 70-73**.**

[27] Jaseena and Julie M. David, "Issues, challenges and solutions: Big Data Mining," in *Elsevier,* pp. 131–140, 2014.

[28] Tyson Condie, Paul Mineiro, Neoklis Polyzotis, Markus Weimer, "Machine Learning for Big Data" in *ACM,* 2013, pp. 939-942.

[29] Xiaohui Huang , Yunming Ye, Liyan Xiong , Raymond Y.K. Lau, Nan Jiang, Shaokai Wang, "Time series *k* -means: A new *k* -means type smooth subspace clustering for time series data," in *Elsevier*, 2016, pp. 645-654.

[30] Xiao Cai, Feiping Nie, Heng Huang, "Multi-View K-Means Clustering on Big Data," in *International Joint Conference on Artificial Intelligence,* pp. 2598-2604.



Submission Info

SUBMISSION ID	805773103
SUBMISSION DATE	27-Apr-2017 16:15
SUBMISSION COUNT	1
FILE NAME	Thesis_Isha_152209.docx
FILE SIZE	3.83M
CHARACTER COUNT	45325
WORD COUNT	5143
PAGE COUNT	66

ORIGINALITY

OVERALL	7%
INTERNET	0%
PUBLICATIONS	0%
STUDENT PAPERS	7%

GRADEMARK

LAST GRADED	N/A
COMMENTS	0
QUICKMARKS	