



**ENHANCING THE CLUSTERING TECHNIQUE FOR
UNCERTAIN DATA IN DATA MINING**

DISSERTATION REPORT

Submitted in partial fulfillment of the
requirement for the award of the
Degree of

**MASTER OF TECHNOLOGY
IN
(Computer Science & Engineering)**

By

Avinash Kumar
Reg No: 11100856

Under the Guidance of

Mr. Robin Prakash Mathur
(Asst. Professor, Lovely Professional University)

(School of Computer Science & Engineering)
Lovely Professional University
Punjab, 144402
(May 2015)

PAC FORM



School of: Computer Science & Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Avinash Kumar Registration No: 11100.B56
Batch: 2011 Roll No. A22
Session: 2014-15 Parent Section: K2006
Details of Supervisor: Designation: Asst. Professor
Name: Robin Prakash Mathur Qualification: M.Tech (CSE)
U.ID: 14597 Research Experience: 4 yrs

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

PROPOSED TOPICS

- 1. Enhancing the clustering technique for uncertain data in data mining
- 2. Improving the efficiency of text clustering
- 3. Document clustering algorithm

R. P. Mathur (14597)
Signature of Supervisor

PAC Remarks:
First topic is approved, publication expected
llk

llk
Signature: llk Date: 30/9/14

APPROVAL OF PAC CHAIRPERSON: _____ Date: _____
*Supervisor should finally encircle one topic out of three proposed topics and put up for a approval before Project Approval Committee (PAC)
*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.
*One copy to be submitted to Supervisor.

ABSTRACT

To classify or clustering the “**Uncertain Data**” become more challenging in data mining. There are various algorithm are available to cluster or classify the certain data like Naïve Bayes Classification, Rule Based Classification, K-means and many more. Clustering of Uncertain data is bit difficult rather than clustering the certain data. The term “Uncertainty” means state of having limited knowledge where it is impossible to describe the current state, a future outcome. The Uncertainty mostly appears in data which is generated from the following sources like data generated by the sensor network, scientific result and so on. Data collected from these sources are referred to as uncertain data. To check the uncertainty in data a set of probabilities are assigned to each possible state or outcome. In this paper we are going to improve the efficiency of clustering algorithm in term of accuracy, time and space. We have used two algorithms to improve the efficiency. The first algorithm is UK-means and the second on is alpha beta distance pruning algorithm.

ACKNOWLEDGEMENT

I express my sincere gratitude towards my guide **Mr. Robin Prakash Mathur** for her constant help, encouragement and inspiration throughout the work. Without her guidance, this work would never be a successful one. My special thanks to my friend without his support it could be a dream. I would also like to thank those who supported me to accomplish the work. I am also thankful those who supported me directly, indirectly from my college and society. I express my sincere gratitude towards my parents, without their blessings this work could not be possible. At last but not the least I am thankful to **Mr. Rajeev Sobti** HOS of Computer Science & Engineering and **Mr. Dalwinder Singh** HOD of computer science & engineering for providing healthy academic environment.

Date.....

Avinash Kumar

Place: Lovely Professional University

Reg. No. 11100856

CERTIFICATE

This is to certify that **Avinash Kumar** bearing Registration no. 11100856 has completed objective formulation of thesis titled, “**Enhancing the Clustering Technique for Uncertain Data in Data Mining**” under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the thesis has ever been submitted for any other degree at any University.

The thesis is fit for submission and the partial fulfilment of the conditions for the award of M.Tech in Computer Science & Engineering.

Robin Prakash Mathur

(Asst. Professor, School of Computer Science & Engineering)

Lovely Professional University

Phagwara, Punjab (144402).

Date:.....

DECLARATION

I, **Avinash Kumar**, student of MCA – M.Tech under Department of Computer Science & Engineering of Lovely Professional University, Punjab, hereby declare that all the information furnished in this thesis report is based on my own intensive research and is genuine.

This thesis does not, to the best of my knowledge, contain part of my work which has been submitted for the award of my degree either of this university or any other university without proper citation.

Date:

Avinash Kumar

Place: Lovely Professional University

Reg No. 11100856

Table of Contents

PAC FORM	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT	iv
CERTIFICATE	v
DECLARATION	vi
Chapter 1.....	4
INTRODUCTION.....	4
1.1 Data Mining:	4
1.2 Clustering	5
1.3 Clustering Uncertain Data:.....	5
1.4 Challenges in Handling Uncertain Data	7
1.5 Scope of Study.....	7
1.7 Techniques used	8
Chapter 2.....	9
REVIEW OF LITERATURE	9
Chapter 3.....	11
PRESENT WORK	11
3.1 Problem Formulation	11
3.2 Objective	11
3.3 Research Methodology	12
3.3.1 Formulation of Research Problem:	12
3.3.2 Extensive Literature Survey:	12
3.3.3 Development of Working Hypothesis:	12
3.3.4 Preparing the Research Design:.....	13
3.3.5 Determining the Sample Design:.....	13
3.3.6 Collecting Data:.....	13
3.3.7 Execution of Project:.....	13
3.3.8 Analysis of Data:	13
3.3.9 Preparation of Report:.....	14
3.4 Working flow of our work	14
3.5 Algorithm Design	15

3.6 Tools Used	16
3.6.1 Software Used	16
3.6.2 Hardware Requirement:	18
Chapter 4.....	23
RESULT AND DISCUSSION	23
4.1 About Dataset Used	23
4.2 Result Discussion.....	40
Chapter 5.....	42
CONCLUSION & FUTURE WORK	42
References.....	43

List of Figures

Figure 1: Shows Steps involved in Data Mining Process.....	5
Figure 2: Shows Clustering Process.....	6
Figure 3 Shows Working of UK-means Algorithm	6
Figure 4 Shows Complete Research Process	14
Figure 5 Shows Working Flow	14
Figure 6 Shows Environment of R Studio	16
Figure 7 Shows Environment of R Console	17
Figure 8 Show Package available in R	18
Figure 9 Shows the Dataset	23
Figure 10 Shows RStudio Environment and Graph Plotting.....	24
Figure 11 Shows Group of Cluster with different Time Interval and Price	24
Figure 12 Shows RStudio Console with Result	25
Figure 13 Shows Group of Cluster with different Time Interval and Price	25
Figure 14 Shows Time taken to make a Group of Cluster	26
Figure 15 Shows Group of Cluster when changes the Cluster Value	27
Figure 16 Shows Group of Cluster when changes the Cluster Value	27
Figure 17 Shows Total Time taken to form a Cluster	28
Figure 18 Shows Total number of Observation and Function.....	29
Figure 19 Shows Installed Packages in R Library	30
Figure 20 Shows RStudio Console with Output	30
Figure 21 Shows Group of Cluster when changes the Cluster Value	31
Figure 22 Shows Group of Cluster when changes the Cluster Value	32
Figure 23 Show total Time Elapsed to Form a Cluster	32
Figure 24 Shows Group of Cluster when changes the Cluster Value	33
Figure 25 Shows Time elapsed to form a Cluster	33
Figure 26 Shows Group of Cluster when changes the Cluster Value	34
Figure 27 Shows total elapsed Time	34
Figure 28 Shows Group of Cluster when changes the Cluster Value	35

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

Figure 29 Shows Total Time Elapsed	35
Figure 30 Shows Group of Cluster when changes the Cluster Value	35
Figure 31 Shows Total Elapsed Time	36
Figure 32 Shows Group of Cluster when changes the Cluster Value	36
Figure 33 Shows Total Elapsed Time	37
Figure 34 Shows Group of Cluster when changes the Cluster Value	37
Figure 35 Shows Total Elapsed Time to form a Cluster with different size	38
Figure 36 Shows Group of Cluster when changes the Cluster Value	38
Figure 37 Shows Total Elapsed Time	39
Figure 38 Shows Group of Cluster when changes the Cluster Value	39
Figure 39 Show RStudio Console Output	40

Chapter 1

INTRODUCTION

1.1 Data Mining:

When we need to find any pattern or knowledge from the large volume of data then we need to perform Data Mining operation on the database. This process is referred to as data mining. The extracted knowledge is new to the user and is beneficial. As we all know that now day's huge amount of data is being generated from various source like Facebook, Google and YouTube. To find the useful information from that data we need to perform mining operation.

Discovery of knowledge is a different from classical approach of information retrieval from relational databases. In classical relational DBMS, database records are retrieved by processing a query unlike knowledge discovery process, what is retrieved is not explicit in the storage. Rather, it is an implicit approach. Data mining is finding of these hidden pattern. Data mining finds these patterns and relationships with various data analysis tools and techniques. Two different models are available; the first is predictive and other one is descriptive model. The predictive model is used to predict the value from the data which result is known earlier. The second model which is descriptive in nature used to generate the pattern. For example: the probability to selling the butter along with the bread.

To gather the knowledge we perform the ETL (Extract, Transform and Load) operation by integrating data from various sources. Following are the list of steps for the discovery of knowledge from the data:

- Data Cleaning: In the data cleaning process we remove the noise and the inconsistency of the data.
- Data Integration: In this phase we integrate the data from the various data sources.
- Data Selection: In this phase relevant data is retrieved for the analysis purpose.
- Data Transformation: In this phase data are transformed into the appropriate form.
- Pattern Evaluation: In this phase appropriate pattern is being generated.
- Knowledge Representation: In this phase visualization and knowledge representation technique are used to represent the mined data.

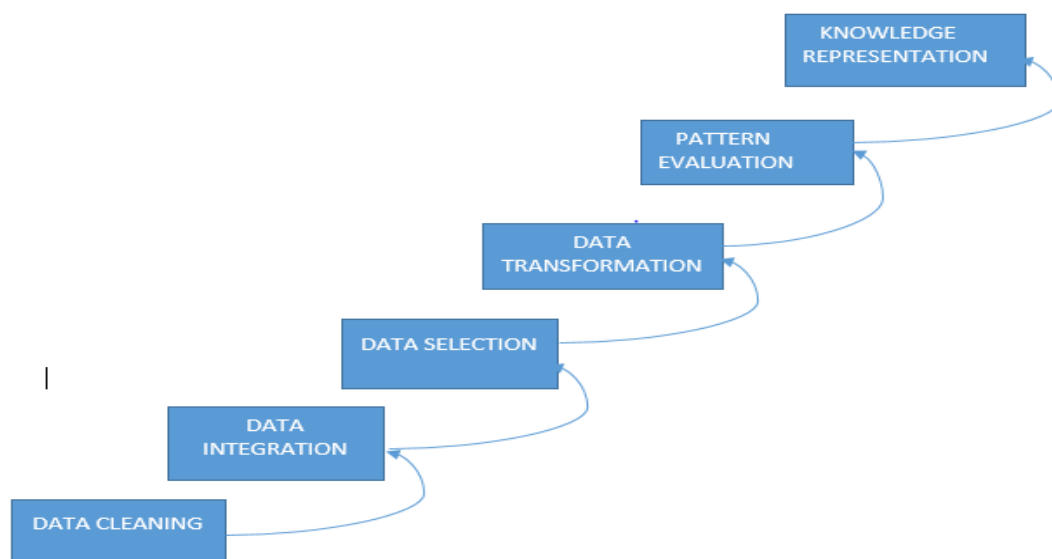


Figure 1: Shows Steps involved in Data Mining Process

1.2 Clustering

Clustering is the process in which we create the group of object which is similar. Cluster of data is widely known as group. For clustering we need to divide the data into groups depends upon the similarity.

1.3 Clustering Uncertain Data:

The clustering algorithm is mostly used in the real time scenario. There are lots of algorithms available for the clustering but not all the algorithm is used to cluster the uncertain data. For clustering the uncertain data UK means algorithm are available but single this algorithm is not able to produce the accurate cluster. It uses point value to find the distance among the object and its representative objects. The classical clustering algorithm deals with only that object whose location are known in earlier. The Data uncertainty mostly appears in case of the sensor network. For example: The price of Share Stock which is fluctuating in every time to time. In this research work we consider the problem of clustering uncertain data objects and try to improve the efficiency of that algorithm by adding the concept of alpha beta pruning algorithm which reduces the unnecessary distance calculation.

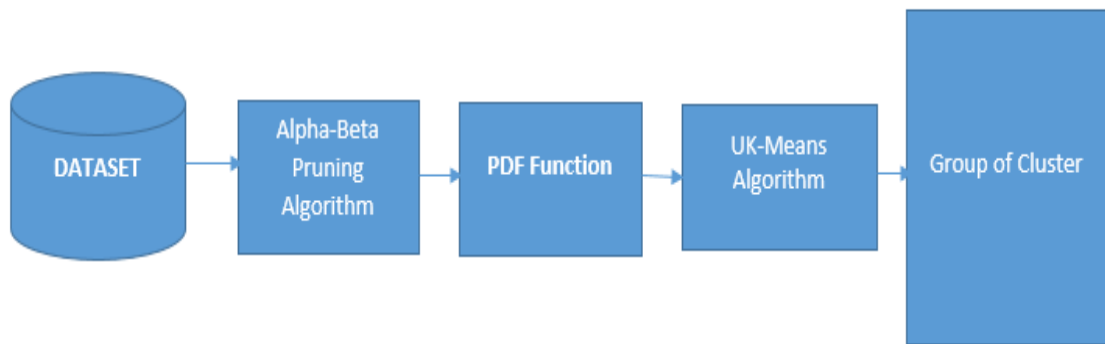


Figure 2: Shows Clustering Process

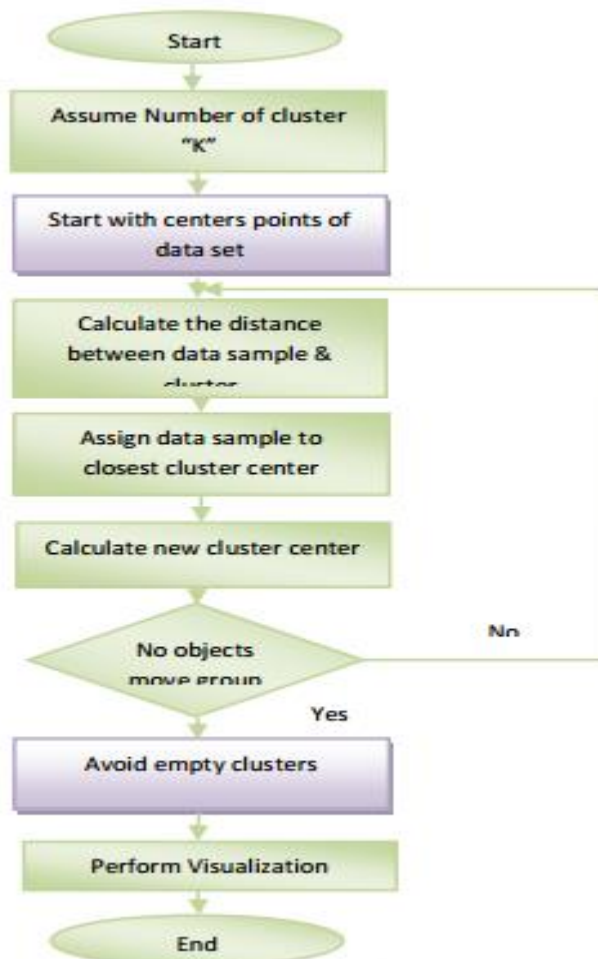


Figure 3 Shows Working of UK-means Algorithm

1.4 Challenges in Handling Uncertain Data

The traditional clustering algorithms are able to deal with those data object whose value is known in advance but it fails when it comes to cluster the uncertain data. The data uncertainty is categorised into two types. The first type of the data uncertainty known as existential uncertainty which works on the relational database. The second types of data uncertainty are known as value uncertainty in which records are known to exist but their values are not known.

As we all know that the uncertainty comes basically in the case of sensor network. In 1.3 we taken the example of Share Stock market in which the price of share is fluctuating between the different time intervals. Means we have to predict the region based on some previous value. So to predict the value we used probability density function (pdf) which provides the optimum result in dealing with the uncertain data.

1.5 Scope of Study

As we know that today's data is generated in the large volume. Huge amount of data is being generated by the sensor network, mobile device and many more sources are referred to as in the category of uncertain data. It is very difficult to cluster the uncertain data in terms to get quality of cluster. There are many algorithms are available to clustering the uncertain data. We have used UK-means algorithm along with the alpha beta distance pruning algorithm to improve the efficiency the clustering algorithm in terms of quality of cluster, time and space.

This works focus on improving the efficiency of clustering algorithm of the uncertain data. As we know that we cannot directly apply any clustering algorithm in case of uncertain data. So, we required some special treatment in terms of probability density function (pdf). The purpose of doing this is to find the region in which the probability of particular record is non zero. Our aim is to improve the efficiency by improving the cluster quality. The implementation of this work is done in R Language. The tools used to accomplish the task are R Language, R Studio. To improve the efficiency of the clustering technique of uncertain data we used two algorithms. The first algorithm consist Uncertain K-means. The second algorithm consist Alpha Beta distance pruning algorithm. The purpose of Alpha Beta distance pruning algorithm is reducing the unnecessary distance calculation which leads to improve the efficiency of the Algorithm.

1.7 Techniques used

Techniques used to accomplish this research work: To increase the efficiency of the clustering algorithm for uncertain data we have taken dataset of Stock details. This dataset contains the changes in price of stock over the different time intervals having ten different companies. We have applied Alpha-Beta pruning on dataset which is used to reduce the unnecessary ED calculation. After that we have applied pdf function to calculate the ED. After the calculation of ED we have applied UK-means algorithm to form the cluster. The alpha-beta pruning algorithm is equivalent to the min-max algorithm in that they both calculate the best move from its position and both method assign same value. The Alpha-Beta is faster than the min-max algorithm because it does not explore all the branch. Following are the condition for pruning in alpha-beta algorithm:-

For MAX node: If the value of $\beta \leq \alpha$ then its α cut off.

For MIN node: If the value of $\beta \leq \alpha$ then its β cut off.

About Probability Density Function: This function gives you the best optimum value among the set of random value. Its integral over the space is equal to one and it's always nonnegative.

Chapter 2

REVIEW OF LITERATURE

Cheng–Fa Tsai Introduced a new clustering method for mining in large dataset. They incorporated Ant Colony optimization algorithm into data clustering. Their proposed methods produce better result in data clustering than the FSOM with K-means and genetic K-means algorithm. The strategy they introduced in their paper is based upon Ant Colony optimization with different flavour which basically conceptualize the clustering problem into three different desirable approaches .The use of Ant Colony optimization with various favourable ants , stimulated annealing for ants to reduce number of visit to nodes to get local optimal solution , proper tournament selection, these three approaches for getting optimal solution in stimulation shows better performance of ACODF than other algorithms.

Hans-Peeter Kriegel Introduces new algorithm named FOPTICES which is used for hierarchical based clustering. It calculate the similarity between two fuzzy object with the help of probability density function. The FOPTICES algorithm provide you more accurate result.

Kanungo Introduced a filtering algorithm using Kd-tree based on Lloyd’s K-means clustering algorithm. The algorithm is based on storage of multidimensional dataset in different boxes and bounding boxes which consists of set of points. The binary tree, Kd-tree uses hyper-planes for hierarchical subdivision of boxes. Their approach shows determined Kd-tree for finding the data points and determining the weighted centroids, then formulating the vector sum of associated points. For each cell in filtering this approach always select the candidate closer to the centred of the cell thus having advantage over other implementations of Lloyd’s algorithm. This paper also presents an empirical analysis clustering algorithms to establish efficiency of filtering algorithm on large datasets.

Nazeeret Shown a modified approach of K-means in clustering for improving accuracy and efficiency .In this paper they divided the K-means clustering into various phases and applied modified approaches to determine centroid and forming cluster. They fine-tuned the clusters to in later stages for improving efficiency .Determining centroid and

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

optimizing the clusters are the main two phases of their approach where the first phase works as an input to the later. In later phase, where iterative process is involved to clustering the converged with nearest data points, uses heuristic based distance calculation among centroid and data points. This approach results entire clustering process in $O(n)$ without compromising accuracy of clusters.

W.K. Ngai - It uses two algorithms. The first algorithm concerns UK-means and the second algorithm used is min max distance pruning algorithm. The probability density function is used to find the similarity between the object and its representative. The min max algorithm is used here to reduce the unnecessary distance calculation. UK-means algorithm randomly selects k points as cluster representative. After that every object is assigned smallest expected distance. This is an iterative process. To compute the distance, require the computation of the integral $\int f(x)d(x, p_j)dx$ where $f(x)$ is probability density of a point x in the uncertainty region of o_i , and $d(x, p_j)$ is the distance between x and p_j .

Yu-Chen Song formulated a new clustering algorithm for arbitrary dataset, CADD (Clustering Algorithm based on object Density and Direction) from traditional K-means and DENCLUE algorithms. They applied the CADD algorithm on 2D graph and on a geochemical survey. Their result implies that it is robust and capable for determine clusters of different shape and size.

Chapter 3

PRESENT WORK

In this chapter, we are going to present the problem of our research work, its objectives, the methodology that we used for our purposed approach and the introduction of the developed tool. In the 3.1 section we explain how we formulated our problem and what the approach we are going to use. In the 3.2 and 3.3 section the objectives and the methodology of the work done. In the methodology the flow of our work with the help of flow chart is explained.

3.1 Problem Formulation

As we know that today's data is generated in the large volume. Huge amount of data is being generated by the sensor network, mobile device and many more sources are referred to as in the category of uncertain data. It is very difficult to cluster the uncertain data in terms to get quality of cluster. There are many algorithms are available to clustering the uncertain data. We have used UK-means algorithm along with the alpha beta distance pruning algorithm to improve the efficiency the clustering algorithm in terms of quality of cluster, time and space.

This works focus on improving the efficiency of clustering algorithm of the uncertain data. As we know that we cannot directly apply any clustering algorithm in case of uncertain data. So, we required some special treatment in terms of probability density function (pdf). The purpose of doing this is to find the region in which the probability of particular record is non zero. Our aim is to improve the efficiency by improving the cluster quality.

3.2 Objective

The main objective of this research is to improve the efficiency of the algorithm. During the literature of review the author [**Wang Kay Ngai**] introduces min-max algorithm to reduce the expected distance calculation but in the min-max algorithm we have to visit all the branch of the tree. So to reduce that we have introduces another algorithm named alpha-beta pruning algorithm which prune the some branches of tree. The main advantage of this algorithm is it reduce the number of expected distance calculation. To improve the efficiency of the clustering technique of uncertain data we used two algorithms. The first

algorithm consist Uncertain K-means. The second algorithm consist Alpha Beta distance pruning algorithm. The purpose of Alpha Beta distance pruning algorithm is reducing the unnecessary distance calculation which leads to improve the efficiency of the Algorithm.

3.3 Research Methodology

3.3.1 Formulation of Research Problem:

For formulating my research problem we contact our dissertation mentor. He has the 4yr experience in the area of research. We have also discussed with my senior who have completed her thesis earlier in the same domain. Formulation of Research is important and must be understandable because if you clearly understand the problem only then you can find remedy for that. So, the problem must be clearly stated and understandable. In this phase we have formulated the problem of Clustering of Uncertain Data.

3.3.2 Extensive Literature Survey:

Once the research problem is formalized the next steps is to review the literature. In this step we gather the information by studying the previous work done on the same topic. We have also studied various research paper which is needed in our Research work. We have reviewed the paper from the various standard journals like IEEE, ACM, and Springer so that we can get the quality of content. In this phase we have studied the problem of clustering in case of uncertain data means we have to make the cluster of uncertain data. The Data which is generated by the sensor network, GPS, mobile device etc. are referred to as uncertain data because you can only predict the next value.

3.3.3 Development of Working Hypothesis:

After the literature survey is done the next step is to develop the working hypothesis. In this phase we conduct meeting with our mentor about this problem, its origin and the objective of study. Our main aim is to improve the efficiency the clustering algorithm in terms of uncertain data so; we have taken the assumption that the alpha beta pruning algorithm with UK-means algorithm may improve the efficiency of the algorithm. The main purpose of adding the concept of alpha-beta pruning algorithm is to reduce the unnecessary ED calculation because in this algorithm we can prune some branches of tree according the value of alpha and beta. This pruning technique reduce the number of expected distance calculation.

3.3.4 Preparing the Research Design:

While preparing the research design we analyse the following thing:

- What kind of cost related to this research?
- What kind of help we can get from my mentor?
- How many hour we can spend on to my research work?

In this phase we gather the answer of the above question. Because to complete any research you must have time as well as resources to complete it.

3.3.5 Determining the Sample Design:

In this phase we determine what kind of sample suited for my research problem. This is done before the data collection. In this phase we have taken the sample of Stock data. We have chosen this dataset because, our problem is to cluster the uncertain data and the Stock data is a kind of uncertain data because we are not knowing the future value. We can only predict the value based of certain probability density function or cumulative density function.

3.3.6 Collecting Data:

While dealing with some real life problem it is very difficult to find the data so, it is necessary to find the appropriate data sample which is suited for the problem. In this phase we collected the data which is generated by Stock Market because these data are uncertain in nature means you cannot predict the future value, only you can find the probability of it.

3.3.7 Execution of Project:

This is very important step of the research process because the data which is collected earlier is fulfilling the requirement or not. We also check that our project is executed systematically or not. The implementation of this Research work done in R programming Language.

3.3.8 Analysis of Data:

After the implementation is completed the next step is to analyse the data. In this phase we have analysed the result obtained from the implementation.

3.3.9 Preparation of Report:

It is the final step of the research process. In this phase we have written the report which consist complete process of doing the research work according to the standard given by our university.

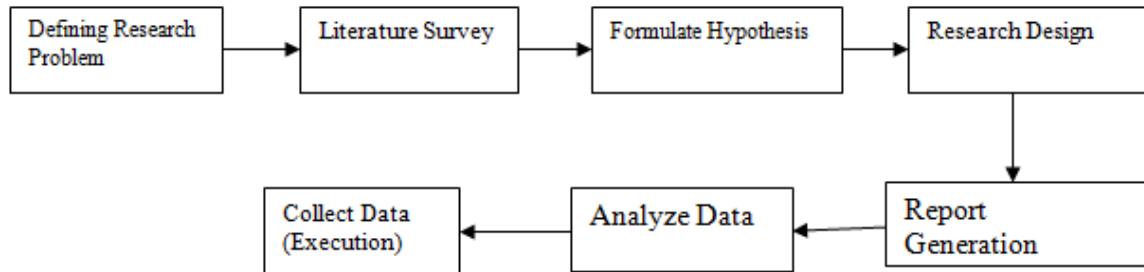


Figure 4 Shows Complete Research Process

3.4 Working flow of our work

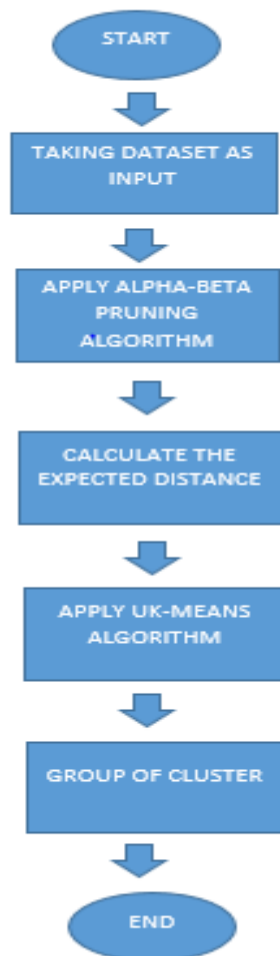


Figure 5 Shows Working Flow

3.5 Algorithm Design

1. For our research work we have taken the database of Stock details.

I. INPUT(DATASET)

II. [Initializing Variable]

2. Apply Alpha-Beta Pruning Algorithm

function alpha-beta(node, depth, α , β , max_val)

if depth = 0

return node value

if max_val

v := $-\infty$

for each child of node

v := max(v, alpha-beta(child, depth - 1, α , β , FALSE))

α := max(α , v)

if $\beta \leq \alpha$

break (*β cut-off*)

return v

else

v := ∞

for each child of node

v := min(v, alpha-beta(child, depth - 1, α , β , TRUE))

β := min(β , v)

if $\beta \leq \alpha$

break (*α cut-off*)

return v

alpha-beta (origin, depth, $-\infty$, $+\infty$, TRUE)

3. Calculating Expected Distance[ED] by applying Probability Density Function (pdf) on Dataset.

4. Apply UK-means algorithm to make a group of cluster.

3.6 Tools Used

3.6.1 Software Used

To implement this Research work we have used R Language. R programming is basically used for statistical computing and graphics. It comes under GNU project developed by John Chamber at AT&T Bell Lab. The R language is similar to S language but you can say that R is the next version of S. You can run and compile your code on various platform. The Version of R is available for various version of operation system. We have used 3.1.2 windows version of R. We have used RStudio-0.98.1091 version of Editor to implement this. You can easily plot any graph with the help of R. The version of R is free available for download.

Environment of R Language: - R is collection of software suit which provide you facility for data manipulation, calculation and graphics display. Its syntax and semantics is somehow different from the C programming language. It support function, loop, procedure etc. For effective use you can write your code on RStudio Editor. It is more flexible rather than the editor of R. Basic editor of R language look like as a command prompt. Extension of R file is .R.

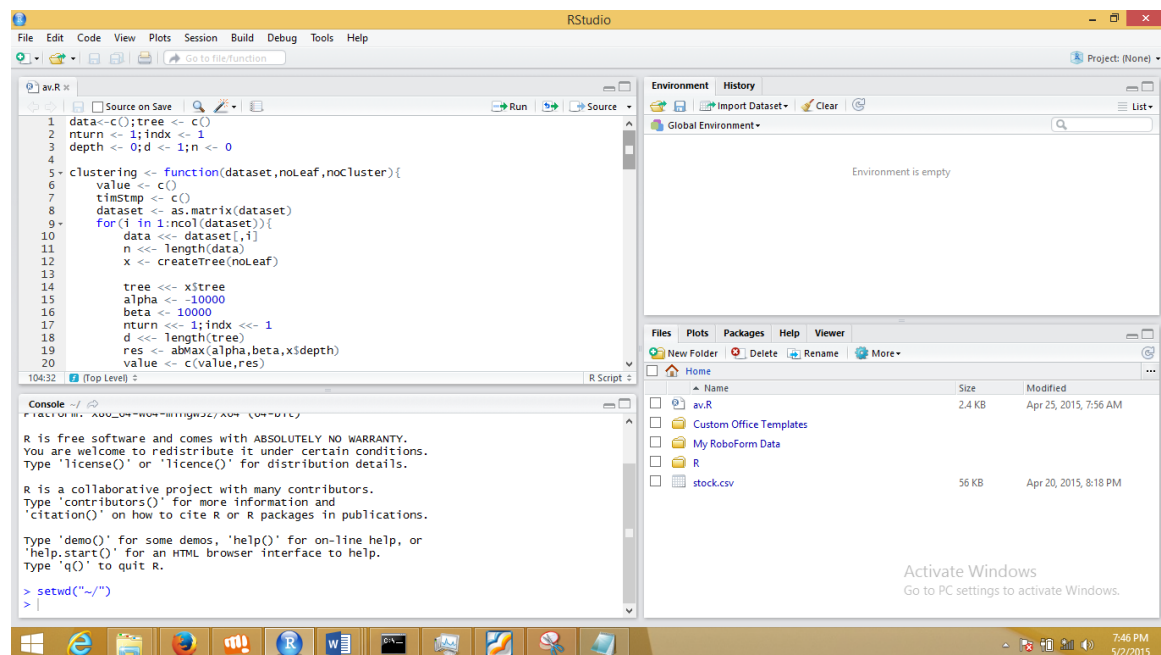


Figure 6 Shows Environment of R Studio

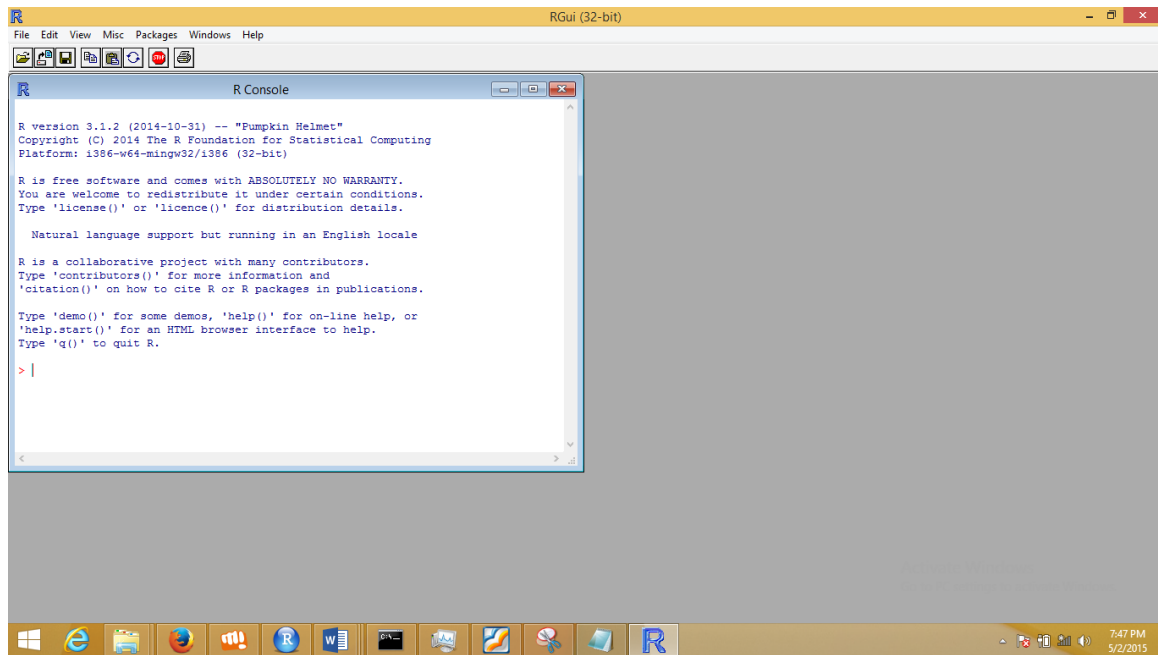


Figure 7 Shows Environment of R Console

Reason for choosing the R Language: As we all know that now the days there are various kinds of software available for data analysis purpose like MATLAB, SAS, SPSS, WEKA, Data Mining System etc. But following are the reasons for choosing R Language to implement this Research Work:-

- **R is free:** The R Language comes under the GNU project which is an open source community. There are no subscription charges for it. You can use it on various platforms like Windows, Mac PC, and Linux. There are different versions of R that are freely available for download.
- **Visualization of Data and its Graphics:** R is known for its data visualization power and its graphics. There are many packages available for plotting charts.
- **Flexible Statistical Analysis toolkit:** In R, all the useful data analysis tools are in the right place. They are so easy and efficient to use. They provide support for a wide range of data formats.
- **Access to Powerful, Cutting Edge Analytics:** All the academicians and researchers use R language to develop the latest methods in Statistics, Machine Learning, and Predictive Analysis. There are more than 2000 packages available for statistical analysis. They are all freely available for download.

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

- A Robust Community: There are more than five thousand contributor and millions of user available all around the world. If you found any difficulty while programming, you can post your problem to R community forum for quick result. There are lots of contributor are available for your help all around the world.

Package Available in R Language:

Available CRAN Packages By Date of Publication

Date	Package	Title
2015-05-02	coxinterval	Cox-Type Models for Interval-Censored Data
2015-05-02	EvalEst	Dynamic Systems Estimation - Extensions
2015-05-02	GDAdata	Datasets for the Book Graphical Data Analysis with R
2015-05-02	gplots	Various R Programming Tools for Plotting Data
2015-05-02	ig.vancouver.2014.topcolour	Instagram 2014 Vancouver Top Colour Dataset
2015-05-02	learnstats	An Interactive Environment for Learning Statistics
2015-05-02	mixtNB	DE Analysis of RNA-Seq Data by Mixtures of NB
2015-05-02	MLmetrics	Machine Learning Evaluation Metrics
2015-05-02	muir	Exploring Data with Tree Data Structures
2015-05-02	NPCD	Nonparametric Methods for Cognitive Diagnosis
2015-05-02	paleotree	Paleontological and Phylogenetic Analyses of Evolution
2015-05-02	pogit	Bayesian Variable Selection for a Poisson-Logistic Model
2015-05-02	soundecology	Soundscape Ecology
2015-05-02	survJamda	Survival Prediction by Joint Analysis of Microarray Gene Expression Data
2015-05-02	survJamda.data	Data for Package 'survJamda'
2015-05-02	TSmisc	'TSdbi' Extensions to Wrap Miscellaneous Data Sources
2015-05-02	W3CMarkupValidator	R Interface to W3C Markup Validation Services

Figure 8 Show Package available in R

3.6.2 Hardware Requirement:

We have implemented it on Windows 8.1 having Intel core i3 processor with 2.16GHz speed. Size of main memory 2GB. This implementation work running fine with this configuration. If you have large amount of dataset then we prefer you to increase the capacity of internal memory up to 4GB for efficient use.

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

Implementation Code

```
data<-c();tree <- c()

nturn <- 1;indx <- 1

depth <- 0;d <- 1;n <- 0

clustering <- function(dataset,noLeaf,noCluster){

  value <- c()

  timStmp <- c()

  dataset <- as.matrix(dataset)

  for(i in 1:ncol(dataset)){

    data <<- dataset[,i]

    n <<- length(data)

    x <- createTree(noLeaf)

    tree <<- x$tree

    alpha <- -10000

    beta <- 10000

    nturn <<- 1;indx <<- 1

    d <<- length(tree)

    res <- abMax(alpha,beta,x$depth)

    value <- c(value,res)

    timStmp <- c(timStmp,which(data==res)[1])

  }

  ds <-cbind("StockValue"=value,"TimeStamp"=timStmp)

  clust <- kmeans(ds, noCluster)

  plot(ds)

  points(clust$centers,col=1:noCluster,pch=8,cex=2)
```

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

```
    clust
  }

createTree <- function(nChild){
  x <- n
  tree <- c()
  depth <- 0
  while(x > 1){
    y <- as.integer(x/nChild)
    i <- 1
    while(i < y){
      tree <- c(nChild,tree)
      i <- i+1
    }
    if(y == 0 )
      tree <- c((x %% nChild),tree)
    else
      tree <- c((nChild + x %% nChild),tree)
    depth <- depth +1
    x <- y
  }
  x <- list("tree"=tree,"depth"=depth)
  x
}

evaluate <- function(){
  total <- 0; indx <<- indx+1
```

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

```
for( i in (nturn-1):d){total <- total+tree[i]}
x<-data[n-total+indx-1]
x
}

abMax<- function(alpha, beta, depthleft){
  if ( depthleft == 0 ) {
    rtn <-evaluate()
  }
  else{
    nt <- nturn; nturn <<- nturn+1;indx <<- 1
    for(i in 1:tree[nt]){
      score <- abMin( alpha, beta, depthleft - 1 )
      if( score >= beta ){
        rtn <- beta
        break
      }
      if( score > alpha ){
        alpha <- score
        rtn <- alpha
      }
    }
  }
  rtn <- rtn[1]
  rtn
}
```

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

```
abMin <- function(alpha, beta, depthleft){  
  if ( depthleft == 0 ) {  
    rtn <- evaluate()  
    print(rtn)  
  }  
  else{  
    nt <- nturn;nturn <<- nturn+1;indx <<- 1  
    for(i in 1:tree[nt]){  
      score <- abMax( alpha, beta, depthleft - 1 )  
      if( score <= alpha ){  
        rtn <- alpha  
        break  
      }  
      if( score < beta ){  
        beta <- score  
        rtn <- beta  
      }  
    }  
  }  
  rtn <- rtn[1]  
  rtn  
}
```

4.1 About Dataset Used

In our Research work we have used Stock Data as Input. This dataset contain the price of ten different companies. We have observed approximate 20 sample with the different cluster size.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	Company1	Company2	Company3	Company4	Company5	Company6	Company7	Company8	Company9	Company10
2	17.219	50.5	18.75	43	60.875	26.375	67.75	19	48.75	34.875
3	17.891	51.375	19.625	44	62	26.125	68.125	19.125	48.75	35.625
4	18.438	50.875	19.875	43.875	61.875	27.25	68.5	18.25	49	36.375
5	18.672	51.5	20	44	62.625	27.875	69.375	18.375	49.625	36.25
6	17.438	49	20	41.375	59.75	25.875	63.25	16.5	47.5	35.5
7	18.109	49	19.5	41.875	59.625	26.625	66.25	17.125	47.75	34.375
8	18.563	49.375	19.125	42.5	60.75	27.25	65.75	16.875	47.875	34
9	18.672	50.125	19.25	43	61.75	28	66	16.875	47.25	34.625
10	18.563	49.75	19	43.25	61.75	29	65.75	17.125	47	34.875
11	19.063	50.5	19.125	43.875	61.875	29.625	66.875	17.75	47.375	36
12	19	50.25	19.625	44	62.125	30	66.5	17.375	47.75	35.625
13	19.063	49.75	20	44.375	61.25	29.875	66.5	16.875	48	35.375
14	18.719	49.25	19	43.5	60.375	29	65.875	16.5	48	34.5
15	18.438	49.25	18.375	43.375	60.375	29	65	16.5	47.5	34.875
16	19.063	50.25	18.375	43.5	60.375	29.125	65.75	16.375	47.875	36.625
17	20	50.25	18.125	44	60.75	30	67	16.75	49	37
18	19.891	50.125	18.25	44.625	60.875	30	66.25	17	48.125	37.375
19	19.563	50.125	18.625	46	61.25	29.75	66.5	16.875	48.75	37.75
20	19.891	51	18.75	46.5	61.875	31.375	67.375	17.625	49	37.875
21	20.328	52.25	18.875	47	63.5	32.125	67.625	17.875	49.25	38.375
22	20.563	52.625	18.875	46.5	63.375	32.125	67	18	49	38.625
23	20.438	53.25	19.25	46.125	63.625	32.125	66.375	18.25	48.875	38.5
24	20.5	53.75	19.25	46	63.25	30.75	66.5	18	47.75	37.5
25	20.563	53.75	19.125	45.75	63.25	30	67.375	18.25	47.5	37.125
26	20.328	53.5	19	45.5	62.375	30	67.375	18.375	46	37.375
27	19.891	52.875	18.875	45	61.375	29.25	67.375	17.625	44.5	36.375
28	20.391	52.5	19	45.125	61.625	29.25	67.5	18	46	36.375

Figure 9 Shows the Dataset

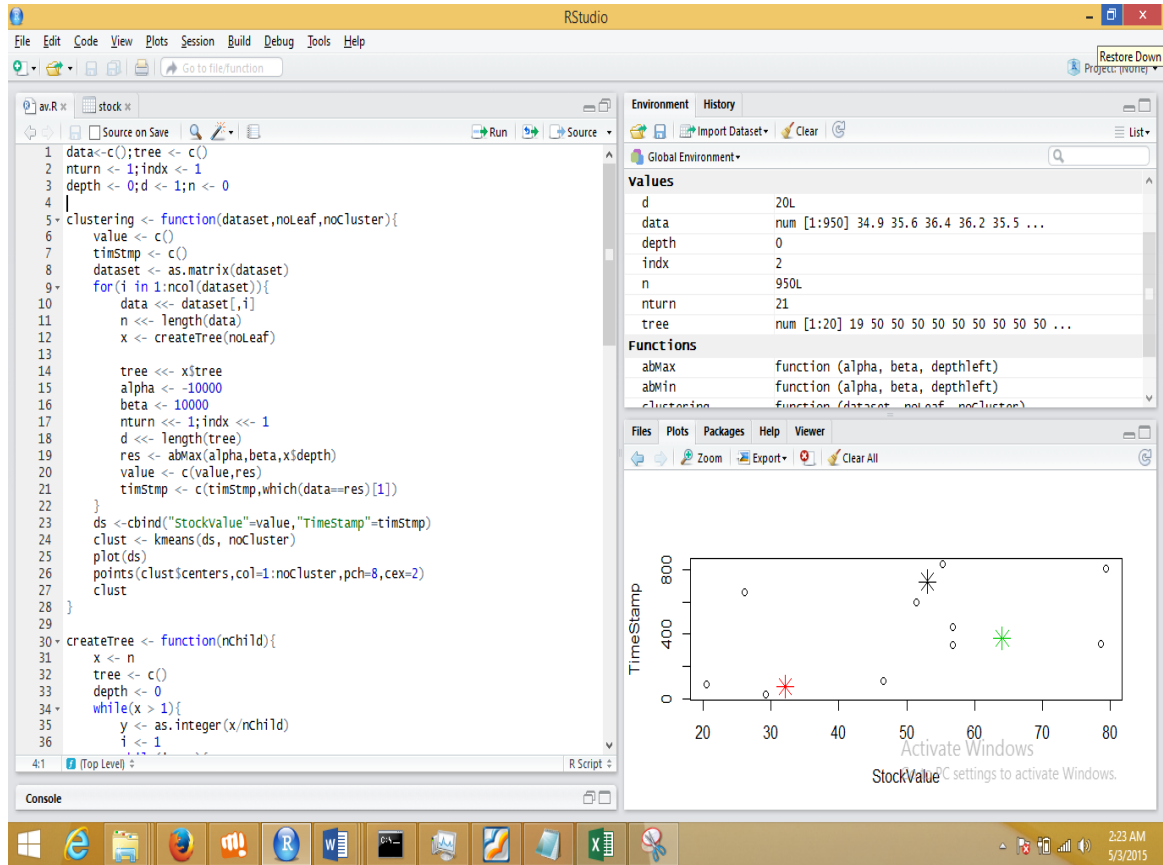


Figure 10 Shows RStudio Environment and Graph Plotting

In the above figure we have plotted the Cluster having the three cluster size. Further we will discuss total time elapsed to form a Cluster.

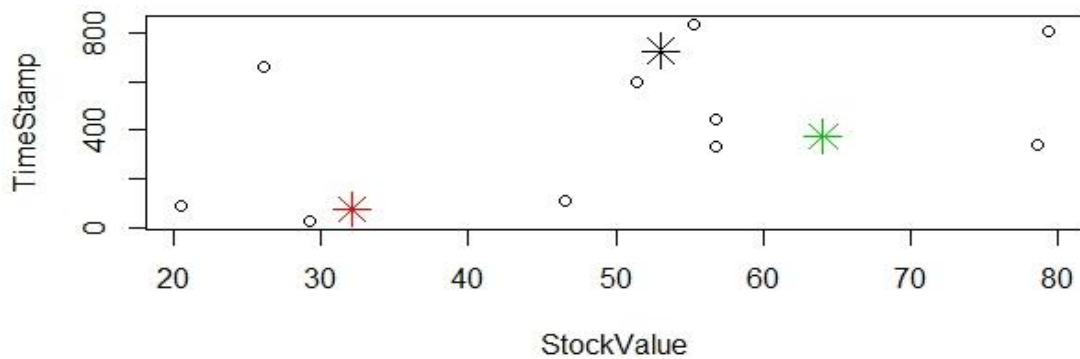


Figure 11 Shows Group of Cluster with different Time Interval and Price

```

Console ~/
> source('~/.R')
> X<-read.csv("~/stock.csv")
> clustering(X,200,3)
K-means clustering with 3 clusters of sizes 2, 6, 2

Cluster means:
  StockValue Timestamp
1   24.37500  327.00000
2   46.47917   48.66667
3   46.37500  540.50000

Clustering vector:
[1] 3 2 2 2 2 1 2 1 2 3

within cluster sum of squares by cluster:
[1] 12170.5312 10556.0339  132.0312
(between_ss / total_ss = 94.6 %)
    
```

Figure 12 Shows RStudio Console with Result

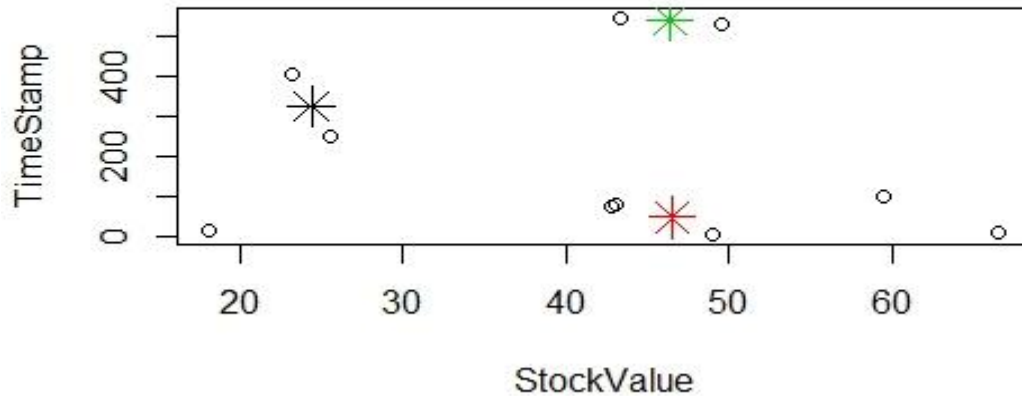


Figure 13 Shows Group of Cluster with different Time Interval and Price

The above figure shows that clustering result when we pass stock dataset to clustering function having 200 observation with three cluster size it will take 0.14 User Time, 0.02 System Time and 0.02 as Elapsed Time. Which is more efficient than the previous algorithm which uses min-max technique.

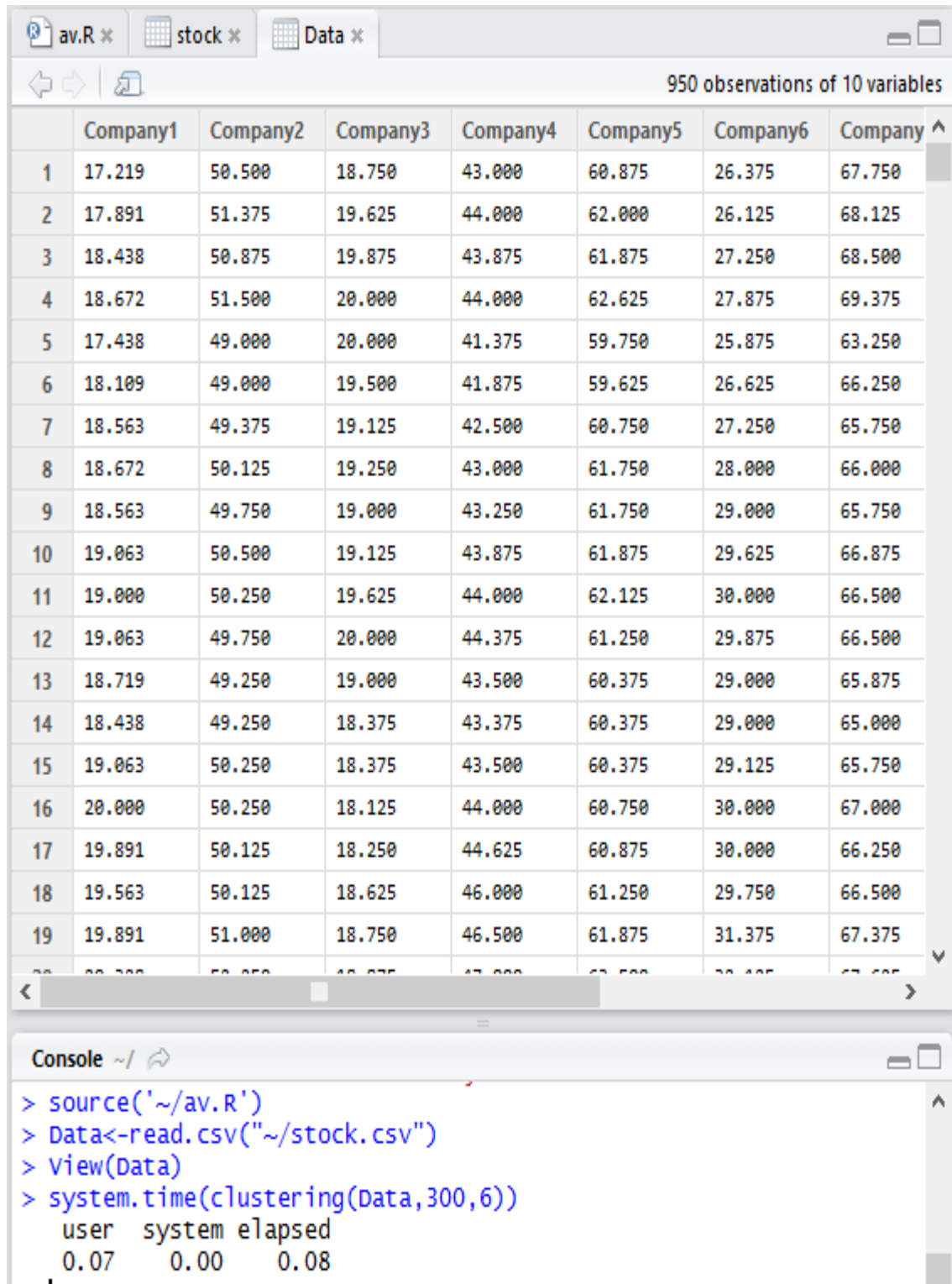


Figure 14 Shows Time taken to make a Group of Cluster

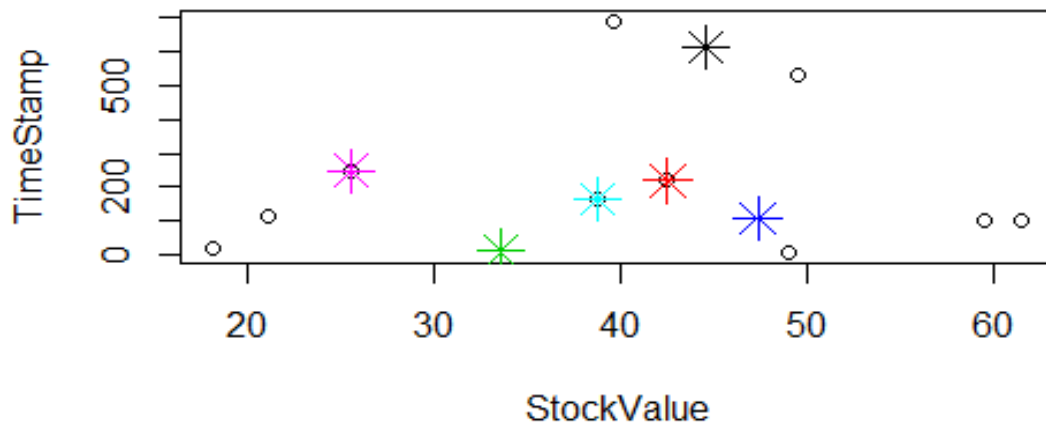


Figure 15 Shows Group of Cluster when changes the Cluster Value

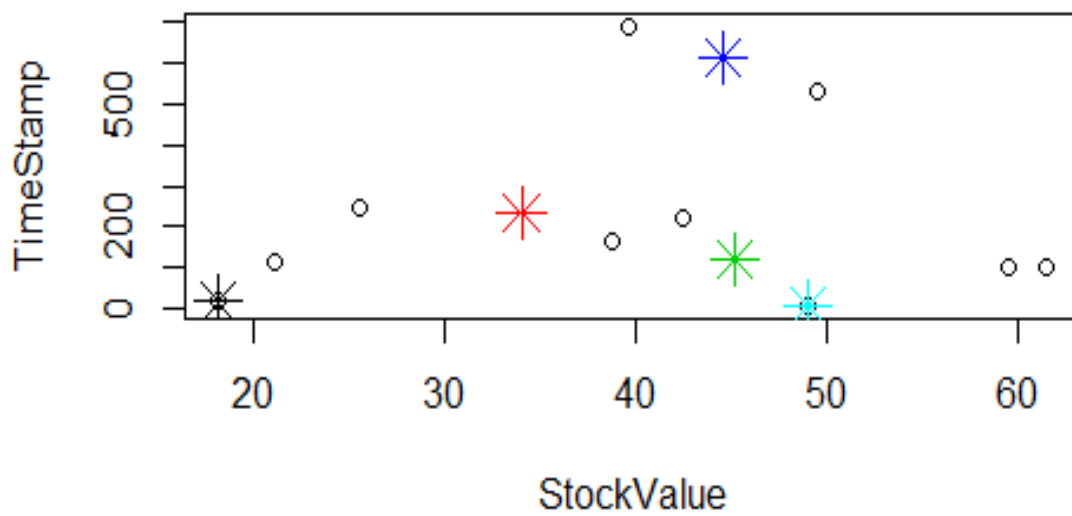


Figure 16 Shows Group of Cluster when changes the Cluster Value

The above figure shows that clustering result when we pass stock dataset to clustering function having 250 observation with five cluster size it will take 0.11 User Time, 0.02 System Time and 0.12 as Elapsed Time. Which is more efficient than the previous algorithm which uses min-max technique.

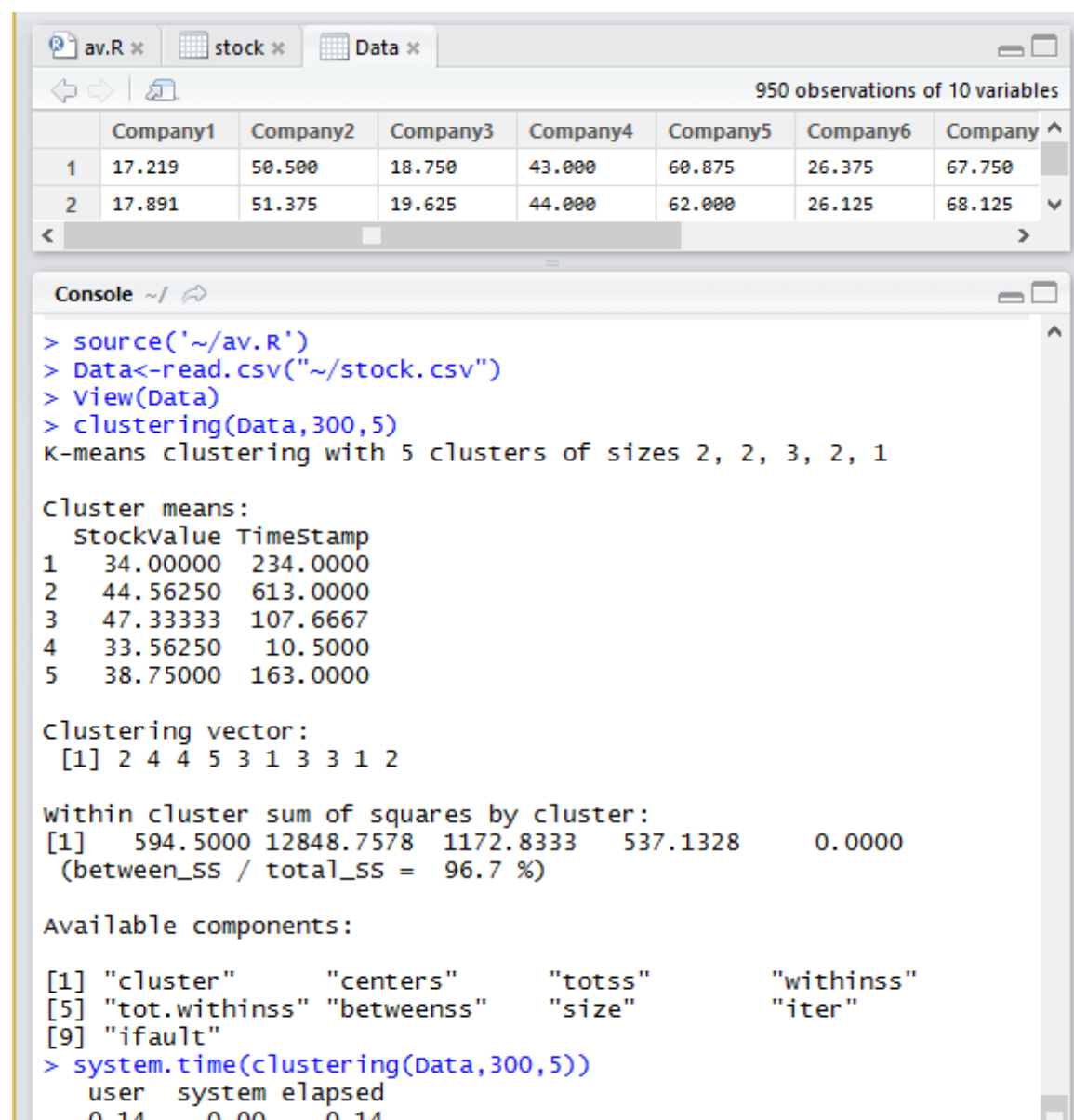


Figure 17 Shows Total Time taken to form a Cluster

The screenshot shows a software interface with tabs for 'Environment' and 'History'. Below the tabs is a toolbar with icons for file operations and a search bar. The main content area is divided into sections: 'Data', 'values', and 'Functions'. The 'Data' section shows '950 obs. of 10 variables'. The 'values' section lists variables and their values. The 'Functions' section lists five functions with their signatures.

Data	
Data	950 obs. of 10 variables
values	
d	1
data	NULL (empty)
depth	0
indx	1
n	0
nturn	1
tree	NULL (empty)
Functions	
abmax	function (alpha, beta, depthleft)
abmin	function (alpha, beta, depthleft)
clustering	function (dataset, noLeaf, noCluster)
createTree	function (nChild)
evaluate	function ()

Figure 18 Shows Total number of Observation and Function

In the above figure the Data variable contains the stock dataset value after reading from the stock dataset. To implement this Research work we have made five different function.

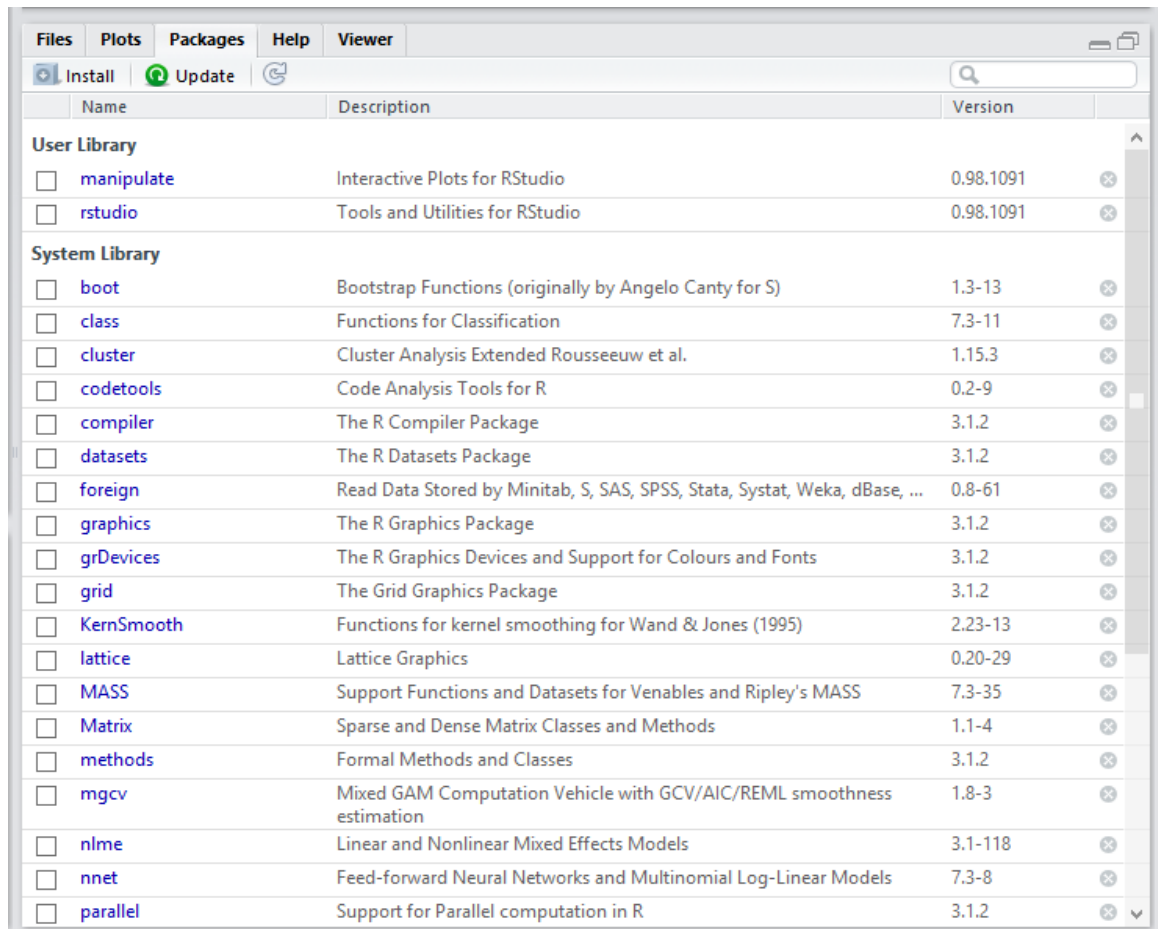


Figure 19 Shows Installed Packages in R Library

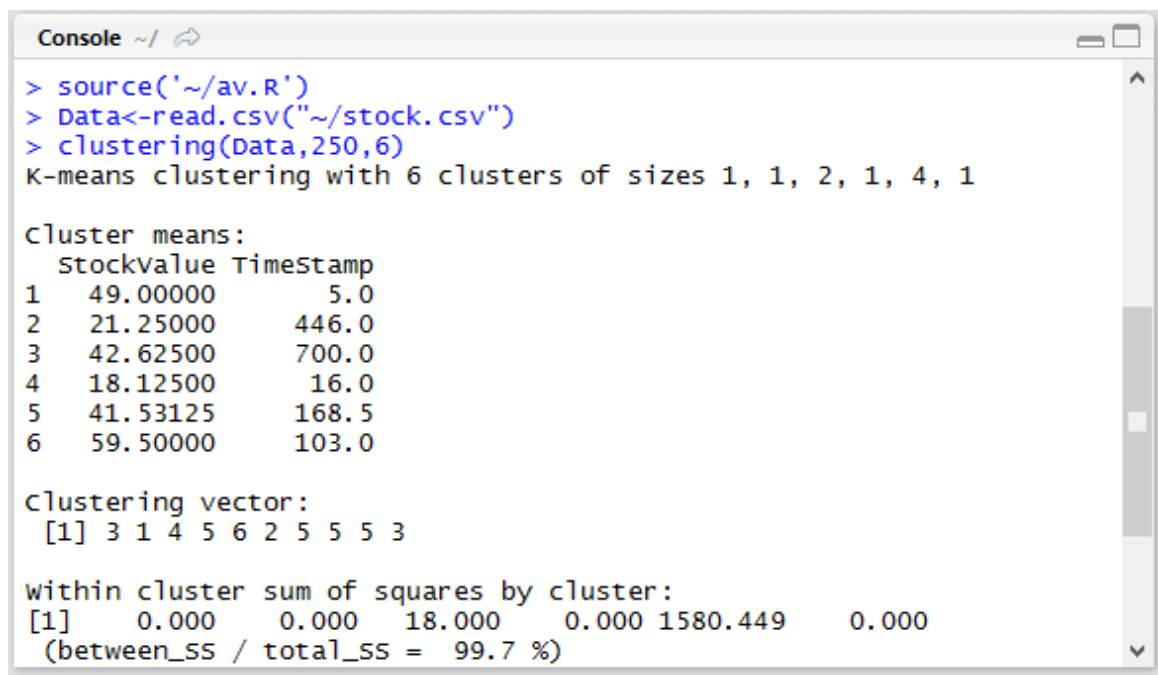


Figure 20 Shows RStudio Console with Output

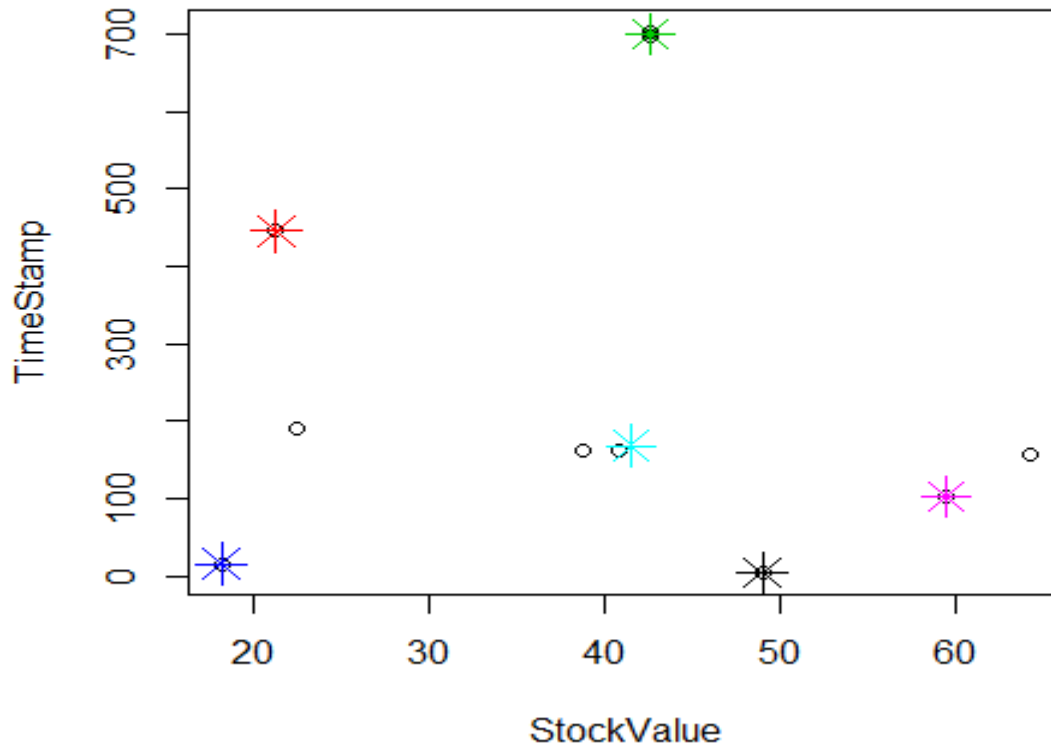


Figure 21 Shows Group of Cluster when changes the Cluster Value

The above figure shows that clustering result when we pass stock dataset to clustering function having 250 observation with six cluster size it will take 0.16 User Time, 0.02 System Time and 0.17 as Elapsed Time. Which is more efficient than the previous algorithm which uses min-max technique.

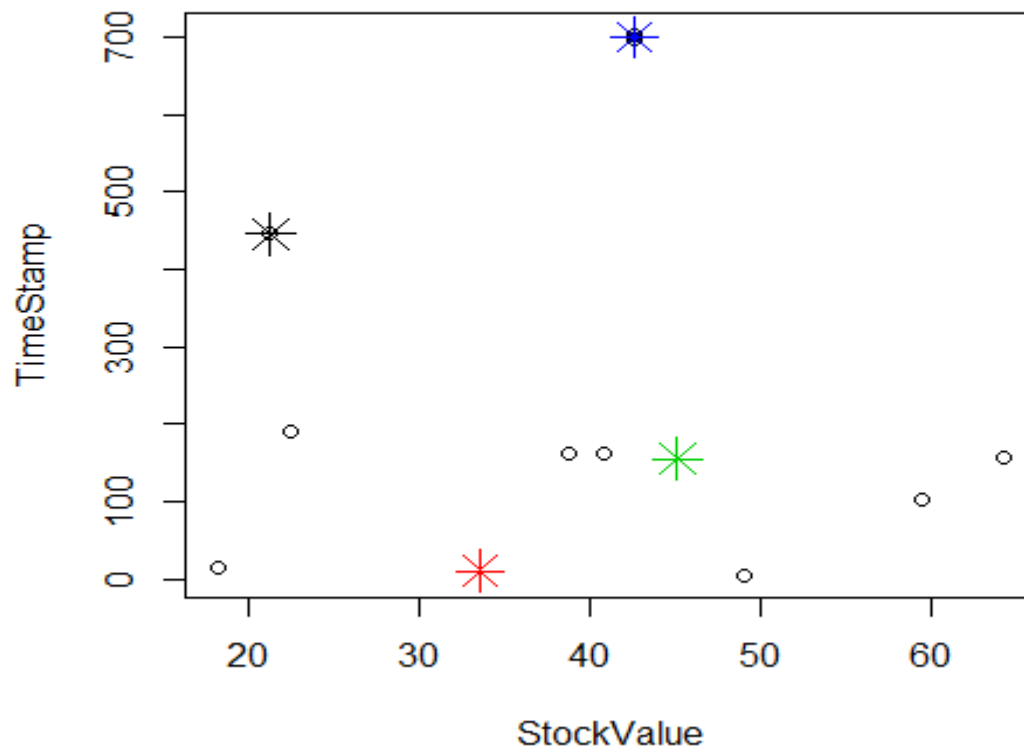


Figure 22 Shows Group of Cluster when changes the Cluster Value

```

Console ~/
> system.time(clustering(Data,250,4))
  user  system elapsed
 0.11   0.00   0.11
>
    
```

Figure 23 Show total Time Elapsed to Form a Cluster

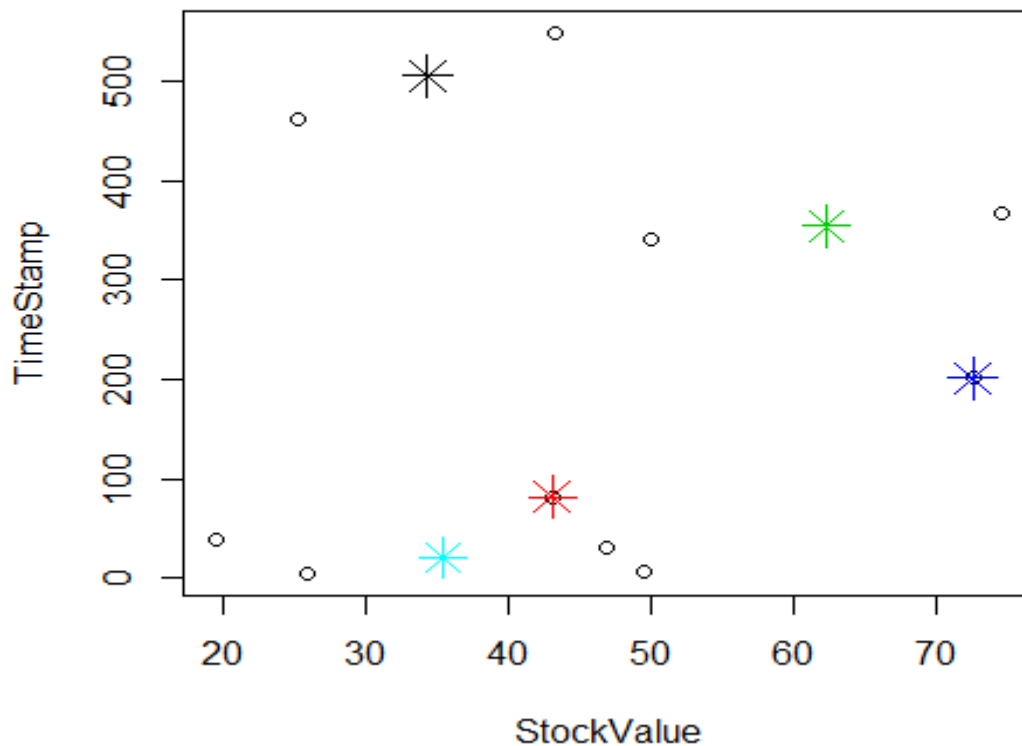


Figure 24 Shows Group of Cluster when changes the Cluster Value

```

Console ~/
> system.time(clustering(Data,150,5))
  user  system elapsed
 0.13   0.00   0.12
>
    
```

Figure 25 Shows Time elapsed to form a Cluster

The above figure shows that clustering result when we pass stock dataset to clustering function having 150 observation with five cluster size it will take 0.13 User Time, 0.00 System Time and 0.12 as Elapsed Time.

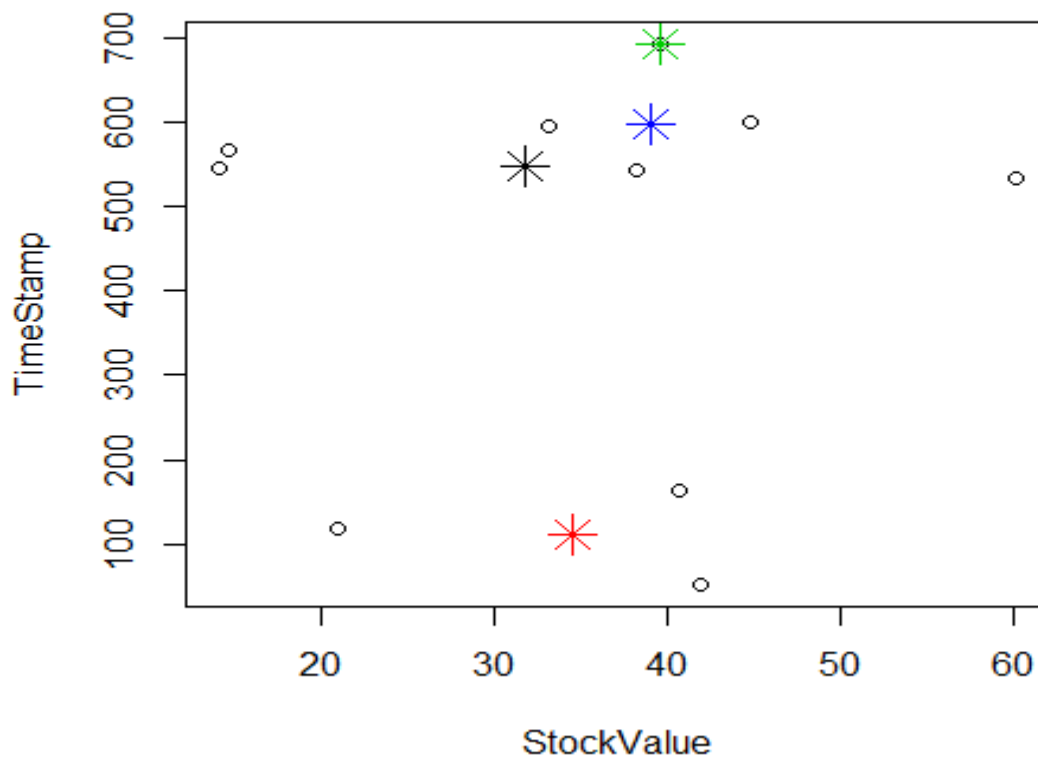


Figure 26 Shows Group of Cluster when changes the Cluster Value

```

Console ~/
> system.time(clustering(Data,350,4))
  user  system elapsed
 0.11   0.01   0.13
> |
    
```

Figure 27 Shows total elapsed Time

The above figure shows that clustering result when we pass stock dataset to clustering function having 350 observation with four cluster size it will take 0.11 User Time, 0.01 System Time and 0.13 as Elapsed Time.

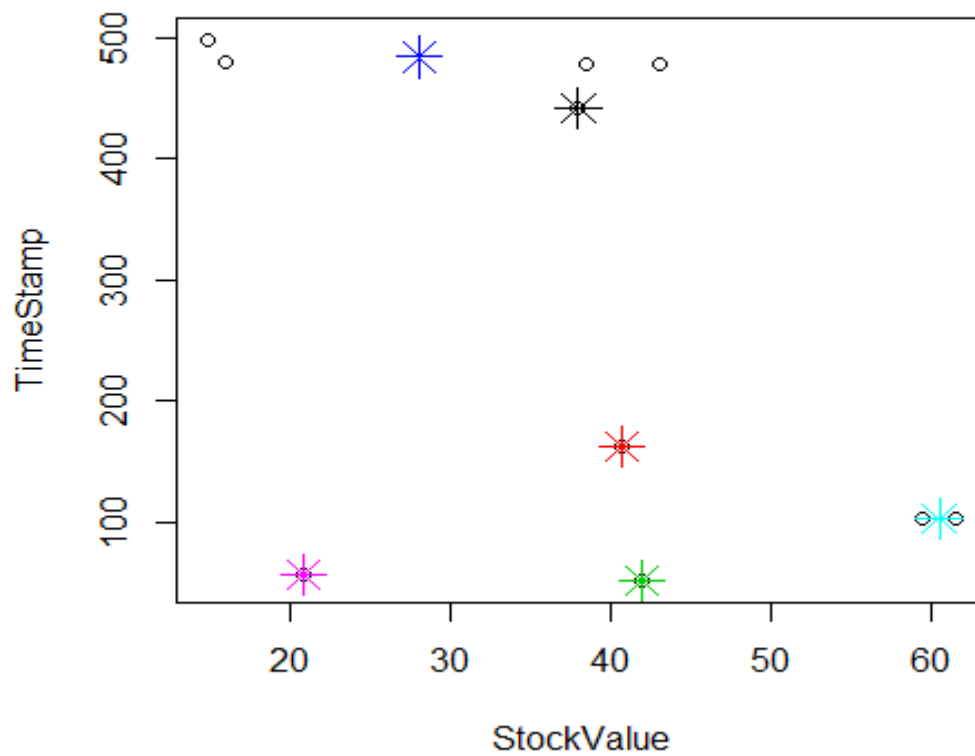


Figure 28 Shows Group of Cluster when changes the Cluster Value




```
Console ~/     
> system.time(clustering(Data,450,6))  
user system elapsed  
0.12 0.00 0.12  
> |
```

Figure 29 Shows Total Time Elapsed

Figure 30 Shows Group of Cluster when changes the Cluster Value


```
Console ~/   
[1] 43.375  
user system elapsed  
1.60 0.08 1.61  
> |
```

Figure 31 Shows Total Elapsed Time

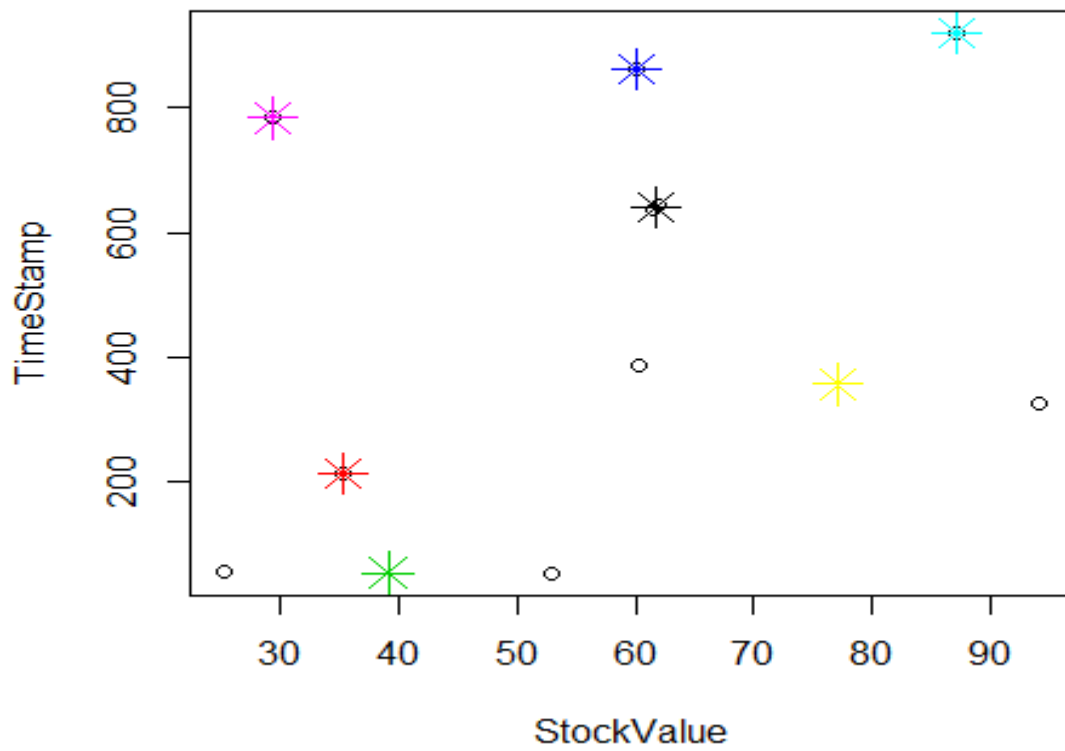


Figure 32 Shows Group of Cluster when changes the Cluster Value


```
Console ~/   
[1] 43.375  
user system elapsed  
1.51 0.03 1.52  
> |
```

Figure 33 Shows Total Elapsed Time

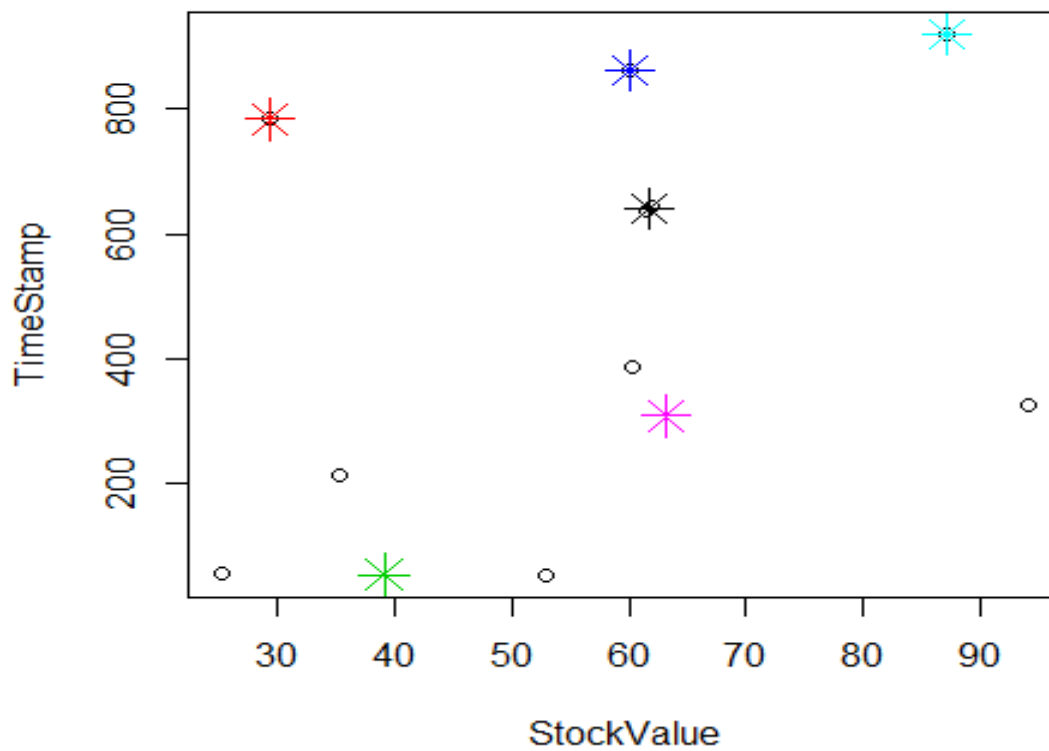


Figure 34 Shows Group of Cluster when changes the Cluster Value


```
Console ~/ 
[1] 43.375
user system elapsed
1.72 0.06 1.72
> |
```

Figure 35 Shows Total Elapsed Time to form a Cluster with different size

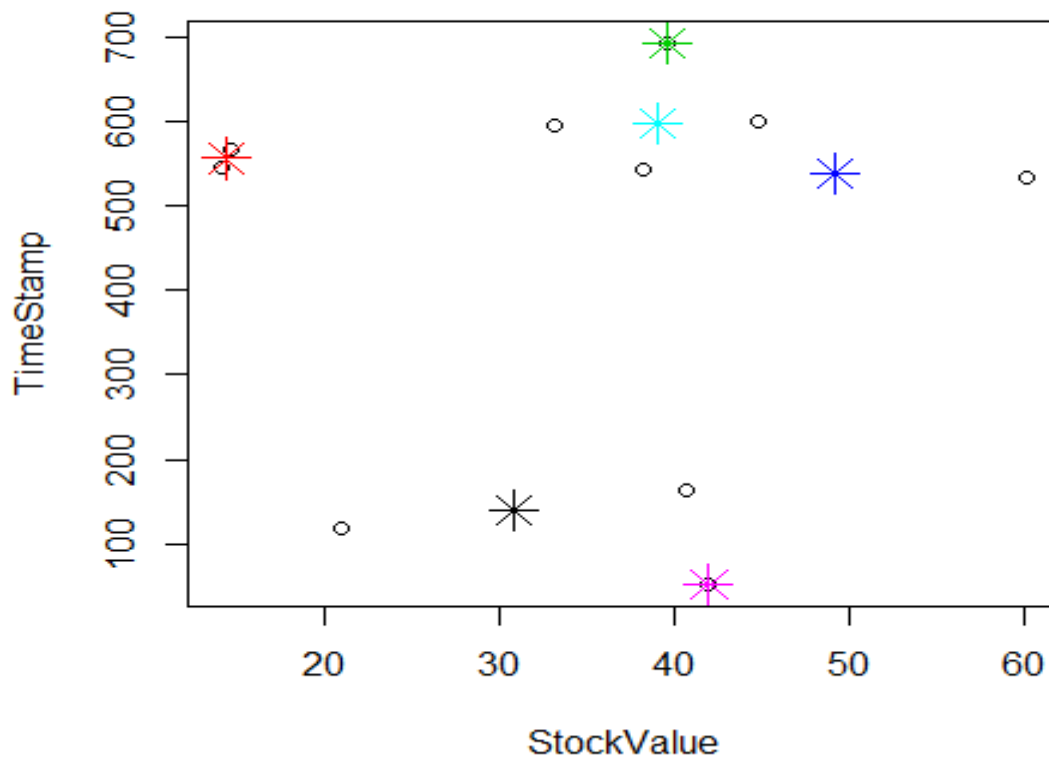


Figure 36 Shows Group of Cluster when changes the Cluster Value

```

Console ~/
> clustering(Data,350,6)
K-means clustering with 6 clusters of sizes 2, 2, 1, 2, 2, 1

Cluster means:
  Stockvalue Timestamp
1    30.8750    139.5
2    14.4375    556.5
3    39.6250    693.0
4    49.1875    539.0
5    39.0625    598.0
6    42.0000     52.0

Clustering vector:
[1] 3 5 2 4 5 2 4 1 1 6

within cluster sum of squares by cluster:
[1] 1207.53125 264.69531 0.00000 271.25781 75.57031
[6] 0.00000
(between_ss / total_ss = 99.6 %)
    
```

Figure 37 Shows Total Elapsed Time

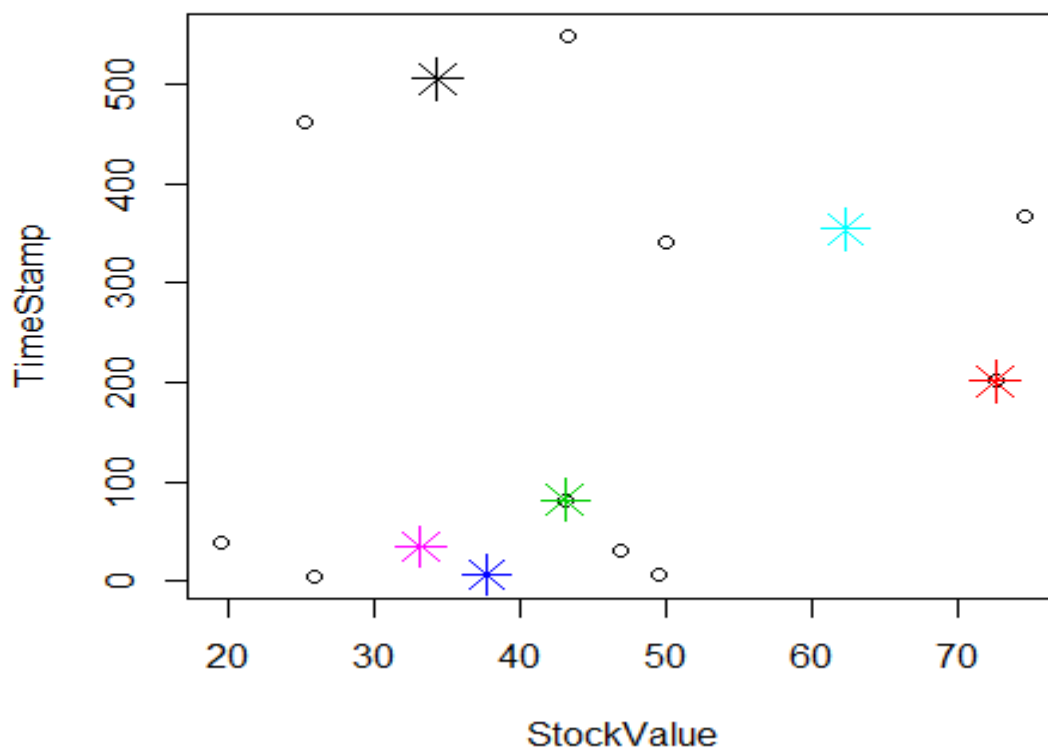


Figure 38 Shows Group of Cluster when changes the Cluster Value

```

Console ~/
> clustering(Data,150,6)
K-means clustering with 6 clusters of sizes 2, 1, 1, 2, 2, 2

Cluster means:
  StockValue TimeStamp
1    34.250      505
2    72.500      201
3    43.000       82
4    37.625        6
5    62.250     354
6    33.125       35

Clustering vector:
[1] 1 4 6 6 2 4 5 1 3 5

within cluster sum of squares by cluster:
[1] 3860.000  0.000  0.000  278.125  638.125  410.125
(between_ss / total_ss = 98.6 %)
    
```

Figure 39 Show RStudio Console Output

4.2 Result Discussion

OBSERVATION	USER TIME	SYSTEM TIME	ELAPSED TIME
clustering(Data,850,6)	1.47	0.04	4.3
clustering(Data,100,6)	0.11	0.01	0.17
clustering(Data,120,6)	0.14	0	0.14
clustering(Data,150,3)	0.08	0.03	0.12
clustering(Data,200,6)	0.11	0.02	0.12
clustering(Data,250,6)	0.13	0	0.12
clustering(Data,300,6)	0.11	0	0.11
clustering(Data,350,6)	0.13	0	0.17
clustering(Data,400,6)	0.09	0.01	0.11
clustering(Data,450,6)	0.11	0	0.11
clustering(Data,500,6)	1.48	0.05	1.51
clustering(Data,550,6)	1.64	0.03	1.73

To cluster the uncertain data we have used UK-means algorithm to form a cluster. There are lots of algorithm available for clustering but simply with the help of that algorithm you cannot form a cluster because here data is uncertain in nature. So, to calculate the expected the ED you have to apply pdf function. As we all know that to calculate the expected distance is very much costly operation in terms of time and resources. So, to reduce the unnecessary ED calculation we have applied alpha-beta pruning algorithm

over the dataset. The alpha-beta pruning algorithm is equivalent to the min-max algorithm in that they both calculate the best move from its position and both method assign same value. The Alpha-Beta is faster than the min-max algorithm because it does not explore all the branch. In the above observation we found that the alpha-beta pruning algorithm taking less time and the space while forming the cluster with different size.

Chapter 5**CONCLUSION & FUTURE WORK**

In our work we have studied the clustering problem with respect to uncertain object which region is defined by the probability density function (pdf). To cluster the uncertain data we have used UK-means algorithm. The other concept added into this is alpha beta pruning algorithm. Here the purpose of Alpha-Beta pruning algorithm is used to reduce the unnecessary expected distance calculation. The Alpha-Beta pruning algorithm provide facility to prune some branches of tree. Due to the pruning technique we saved lots of time. When we apply UK-means algorithm directly to the dataset then we have to calculate the large number of expected distance calculation, which is costly operation. Here the alpha-beta pruning algorithm perform the great job which reduce the unnecessary expected distance calculation.

We described the basic Alpha-Beta distance pruning method and showed that it was very much effective in pruning the expected distance calculation.

In future we can more improve this algorithm by adding the concept of MBR. So, that we can apply different bound estimation method which will more increase the accuracy as well it reduce the clustering time also.

REFERENCES

Biao Qin, Yuni Xia, Sunil Prabhakar, Yicheng Tu A Rule Based Classification Algorithm for Uncertain Data, IEEE International Conference on data engineering 2009.

Charu C. Aggarwal, P. S. Yu A Survey of Uncertain Data Algorithms and Applications in IBM Research report October 31, 2007

Charu C. Aggarwal, Philip S. Yu A Framework for Clustering Uncertain Data Streams IEEE international conference on data mining 2008. pp. 150-159.

Cheng-Fa Tsai, Chun-Wei Tsai, Han-Chang Wu. "A new data clustering approach for data mining in large databases", Inproc. to ISPAN'02, IEEE Computer Society, 2002.

Dilhan Perera, Judy Kay, Irena Koprinska, "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data", IEEE Transaction on Knowledge and Data Engineering.

Fahim A.M, Salem A.M, "An efficient k-means Clustering Algorithm", Journal of Zhejiang University of SCIENCE A, ISSN 1009-3095

K A Abdul Nazeer; MP Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm , In Proceedings of the World Congress on Engineering, 2009 Volume I ,WCE, July1 - 3, 2009, London.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. & Wu, A. Y. An Efficient k-Means Clustering Algorithm: Analysis & Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (7), page 881-892, 2002.

Kiri Wagstaff, Claire Cardie, Seth Rogers," Constrained Clustering with Background Knowledge", International Conference of Machine Learning, 2001, P. 577-584

Mario Cannataro, Andrea Pugliese, "Distributed Data Mining on Grids: Services, Tools, and Application", IEEE Transaction on Systems, Man, and Cybernetics-Part-B: Cybernetics, Vol. 34, No 6 December 2004

N.D Nilesh and D. Suci. Efficient query evaluation on probabilistic databases. In Proc. Of VLDB Conference, pages 864-875, 2004.

ENHANCING THE CLUSTERING TECHNIQUE FOR UNCERTAIN DATA IN DATA MINING

P.S. Bradley, Usama Fayyad, Cory Reina, "Scaling Clustering Algorithm to Large Databases", Copyright © 1998 American Association of Artificial Intelligence.

Rui Xu, "Survey of Clustering Algorithms", IEEE Transaction on Neural Networks, Vol. 16, No. 3, May 2005

Sanpawat Kantabutra and Alva L. Couch, "Parallel K-means Clustering Algorithm on News", Technical Journal Vol.1, No. 6, January 2000/43

Wang Kay Ngai, Ben Kao, Chun Kit Chui, Reynold Cheng, Michael Chau, Kevin Y. Yip, "Efficient Clustering of Uncertain Data", Proceedings of the sixth international conference on data mining (ICDM 06)

Wang Kay Ngai; Ben Kao; Chun Kit Chui; Cheng, R.; Chau, M.; Yip, K.Y., "Efficient Clustering of Uncertain Data," Data Mining, 2006. ICDM '06. Sixth International Conference on, vol., no., pp.436, 445, 18-22 Dec. 2006.

Woo-Sung Jung, Keun-Woo Lim, Young-Bae Ko, "A Hybrid Approach for Clustering Based Data Aggregation in Wireless Sensor Network", IEEE Computer Society, 2009, Third International Conference on Digital Society.

Xiangyang Li, Nong Ye, "A Supervised Clustering and Classification Algorithm for Mining Data with Mixed Variable", IEEE Transaction on Systems, Man, and Cybernetics-Part-A: System and Humans, Vol. 36, No 2 March 2006

Yu-Chen Song, J.O'Grady, G.M.P.O'Hare, Wei Wang, "A Clustering Algorithm incorporating Density and Direction", IEEE Computer Society, CIMCA 2008.

Website Reference

http://en.wikipedia.org/wiki/K-means_clustering

<http://www.r-project.org/about.html>

<http://www.revolutionanalytics.com/what-r>

<http://www.inside-r.org/what-is-r>

<http://www.inside-r.org/why-use-r>

http://en.wikipedia.org/wiki/R_%28programming_language%29

http://www.academia.edu/8489421/PraatR_An_architecture_for_controlling_the_phonetics_software_Praat_with_the_R_programming_language

<http://www.r-project.org/posting-guide.html>

<https://stat.ethz.ch/mailman/listinfo/r-help>

<http://www.r-statistics.com/tag/r-parallel-architecture/>

APPENDICES

Abbreviation

ED- Expected Distance.

UK-Means- Uncertain K-means

PDF- Probability Density Function.