Enhance K-Mean Clustering Algorithm for Rough Sub-spaces

A Dissertation report submitted

By

Harjinder Kaur

to

Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the

Award of the Degree of

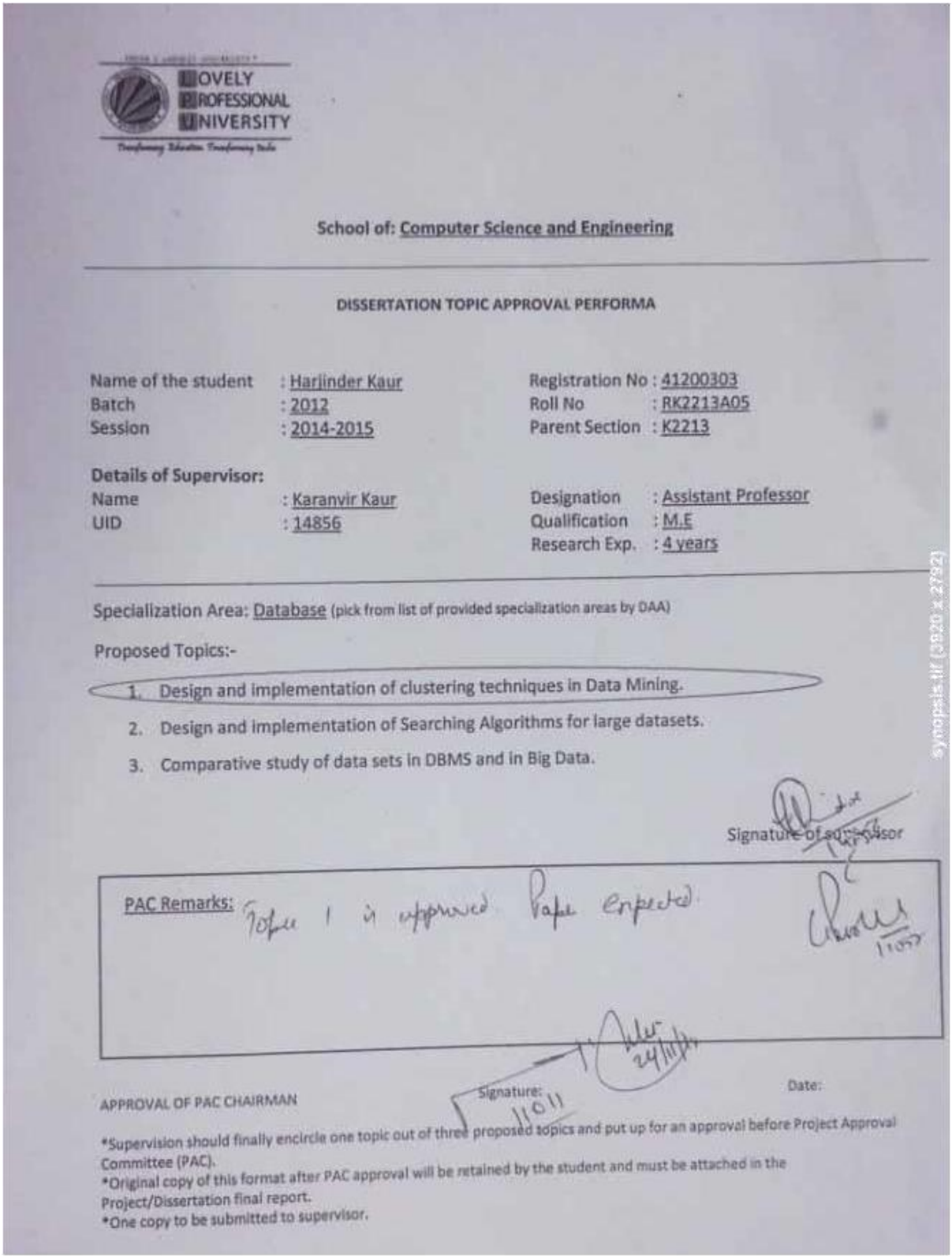
Master of Technology in Computer Science and Engineering

Under the guidance of

Assistant Professor Karanvir Kaur

(June 2015)

PAC APPROVAL FORM



LOVELY PROFESSIONAL UNIVERSITY

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	: Harjinder Kaur	Registration No	: 41200303
Batch	: 2012	Roll No	: RK2213A05
Session	: 2014-2015	Parent Section	: K2213

Details of Supervisor:

Name	: Karanvir Kaur	Designation	: Assistant Professor
UID	: 14856	Qualification	: M.E
		Research Exp.	: 4 years

Specialization Area: Database (pick from list of provided specialization areas by DAA)

Proposed Topics:-

1. Design and implementation of clustering techniques in Data Mining.
2. Design and implementation of Searching Algorithms for large datasets.
3. Comparative study of data sets in DBMS and in Big Data.

Signature of supervisor

PAC Remarks: Topic 1 is approved. Paper expected.

APPROVAL OF PAC CHAIRMAN

Signature: 24/11/14

Date: 11/11

*Supervision should finally encircle one topic out of three proposed topics and put up for an 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

In K-Mean clustering algorithm is that we could not ensembles the sub rough spaces from categorical data due to which mining of such type of data slow down the processing speed of clustering process.

So, we have proposed an algorithm which extracts sub spaces from categorical data the result of which increases the speed of creation and searching of rough sub spaces including which required huge amount of time to mine the information which contain the rough sub spaces. Our algorithm helpful in reducing time and space requirement by eliminating the rough sub spaces from categorical data.

CERTIFICATE

This is to certify that **Harjinder Kaur** has completed her M.Tech dissertation proposal titled **"Enhance K-Mean Clustering Algorithm for Rough Sub-spaces"** under my supervision and direction. To the best of my knowledge, the present work is the result of her novel exploration and study. No part of the dissertation proposal has ever been submitted to any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of the degree of Master in Technology in Computer Science and Engineering.

Date: ---------- Signature of Advisor:

Acknowledgement

I owe my thanks to great people who helped and support me during the preparation of my dissertation synopsis report.

My earnest thanks to **Assistant Professor, Ms. Karanvir Kaur**, The mentor of dissertation, for guiding and correcting me at every stage of the literature review with attention and care. She has taken pain to go through the topic and make necessary corrections as when needed.

My sincerest thanks to **Assistant Professor, Mr. Dalwinder Singh**, Head of Department for allowing me to take up this topic.

DECLARATION

I hereby declared that the dissertation proposal entitled **"Enhance K-Mean Clustering Algorithm for Rough Sub-spaces"** submitted for the M.Tech Degree is exclusively my novel work and all notions and references have been duly acknowledged. It does not contain any work for the award of any degree and diploma.

Date: Investigator

Registration Number:

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

We are living in the era of data age. The size of data is expanding day by day for collection and also for retrieval. It is not easy to maintain large volume of data sets electronically but now days there will be good methods have been generated through which we can store and access the data very easily. There are some figures for data count per minute which is increasing from kilobytes to megabytes, gigabytes to terabytes and so on. There is research is going on nearest neighbor, K-means for search in highest dimensional spaces. By his process data is to be stored in cluster form multiple source and then segmented into groups which is stored on various disks. This created a indexing for accessing the file from stored place and the processing speed is increases for query.

So there's a lot of data out there. But you are probably wondering how it affects you. Most of the data is locked up in the largest web properties (like search engines) or in scientific or financial institutions, isn't it?

The problem is simple: although the storage capacities of hard drives have increased massively over the years, access speeds—the rate at which data can be read from drives have not kept up. One typical drive from 1990 could store 1,370 MB of data and had a transfer speed of 4.4 MB/s,4 so you could read all the data from a full drive in around five minutes. Over 20 years later, one terabyte drives are the norm, but the transfer speed is around 100 MB/s, so it takes more than two and a half hours to read all the data off the disk. This is a long time to read all data on a single drive—and writing is even slower. The obvious way to reduce the time is to read from multiple disks at once. Imagine if we had 100 drives, each holding one hundredth of the data. Working in parallel, we could read the data in less than two minutes. Using only one hundredth of a disk may seem wasteful. But we can store one hundred datasets, each of which is one terabyte, and provide shared access to them. We can imagine that the users of such a system would be happy to share access in return for shorter analysis times, and, statistically, that their analysis jobs would be likely to be spread over time, so they wouldn't interfere with each other too much.

So there several indexing methods is proposed for resolving this problem such as NV-TrEE and LSH and Clustering etc. By creating this kind of index which helps in querying the data from collection, comparison if data points. By using good technique of clustering the disk reading is minimizing. When we were working with large data then, indexing helps in disk reads from multiple data source of data collection at an average time.

Large amount of data was collected which leads to rich data but poor information situation. The problem is that the decisions are not based on rich information data that are stored on different data repositories rather they are based on decision makers intuitions because they don't have the appropriate tools to extract the knowledge embed in various data repositories. Data mining is the process of mining knowledge from the data. Intelligent methods were applied to extract the data patterns from massive amount of data.

The following are the different kinds of data that can be mined:

1) Database Data: Database data can be mined using various aggregate functions.
2) Data Ware House data: This type of data is helpful in mining multidimensional data.
3) Transactional Data: This type of data is mined based on the frequent item sets.
4) Web Mining: Web mining is the process of integrating the information gathered by tradition mining techniques along with the data gathered from the web
5) Text Mining: It is the process of extracting quality information from text.

1.1 Data Mining

We are now in the era where large amount of data needs to be collected on every day. This data without analysis of no use. The data mining we can say is the process of extracting knowledge or meaningful patterns from the stored data that can be further used for decision making. The data that can be used for mining process can be from any data source. Depending upon the data to be mined the following are the functions which are involved in mining process:

Data mining functionality can be of following types:

1) Characterization: Data characterization is the summarization of general characteristics of a particular target class of data. These features can be represented using pie charts, bar graphs etc.

2) Discrimination: The process of comparing general features of target class of data to other similar class of data based on some discrimination rules.

1.2 Technologies used for data mining

1) Statistics: Study of collection, analysis an interpretation of data.
2) Machine Learning: The process of investigation how computers can learn based on the stored data.
3) Information retrieval: The process of searching data from the various sources it can be in the form of text or it can be retrieved from the web.

1.3 Data Mining Issues

1. Mining methodology: Various kinds of data can be mined including mining different kinds of knowledge, knowledge from different disciplines, handling incomplete and noisy data and mining of data in different formats.
2. User interaction issues: These issues include interactive mining at multiple levels of abstraction, assimilation of background knowledge, Knowledge of data mining query languages and ad hoc data mining is required, user friendly interfaces are required to present the data in different formats.
3. Performance issues: efficiency and scalability of data mining algorithms, parallel, distributed and incremental algorithms.
4. Database diversity issues: handling of relational and complex data, heterogeneous databases and global information systems.

Impact of data mining on society: The first issue is what type of technology is benefitted to the society, How to ensure the privacy of data that can be mined and the last and the most important is issue of invisible data mining which means mining of data without having the information of mining algorithms. (Jiawei Han, 2011)

The following diagram summarizes the various design issues:

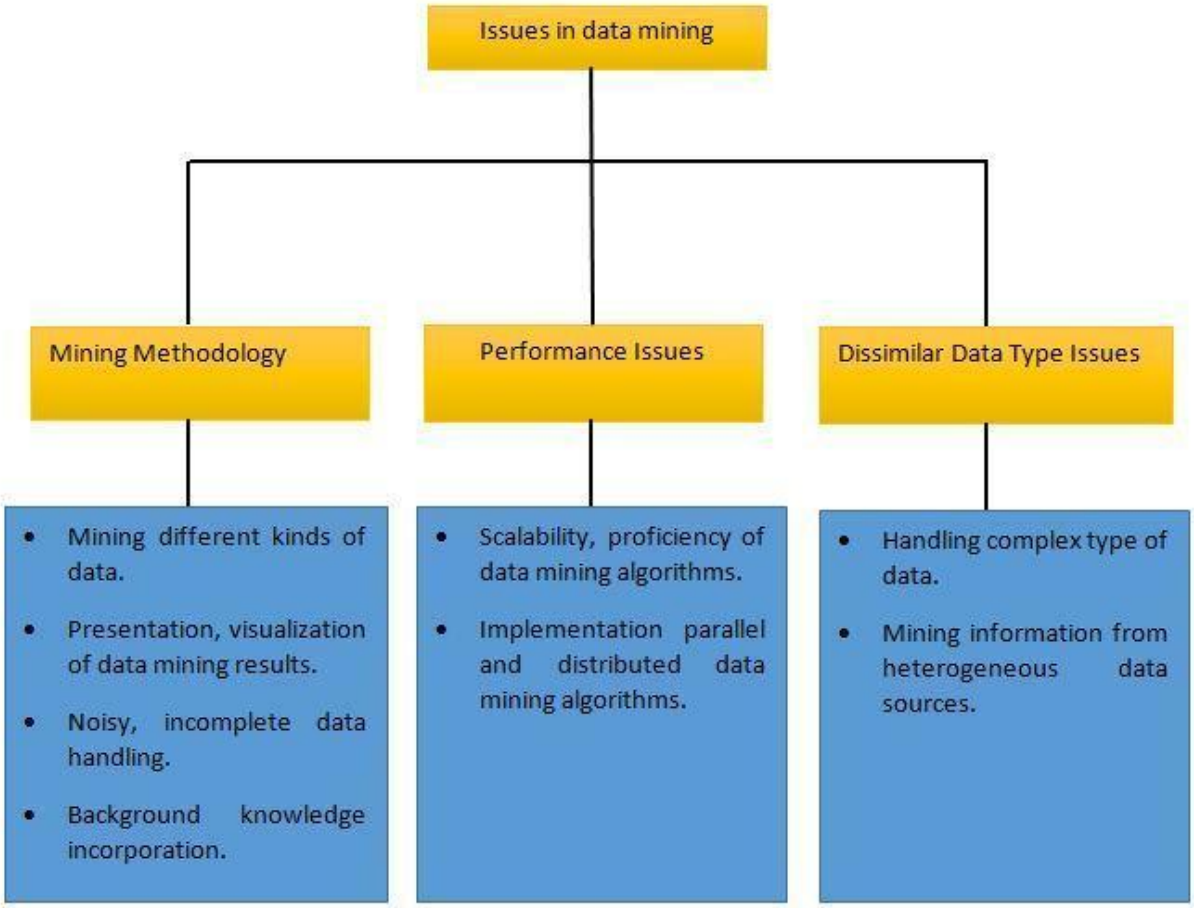


Figure 1.1: Design issues in data mining

1.4 Clustering

Clustering is the unsupervised process of separating a set of data elements into meaningful groups known as clusters. It is known as unsupervised learning because datasets are assigned to a cluster without knowing to which this dataset belong or there are no predefined classes available. The better quality clusters will be produced by the clustering algorithms which satisfy the following conditions:

- The Similarity factor within same cluster must be high which is also known as intra cluster similarity.
- The similarity factor between different clusters must be low which is also known as inter -cluster similarity.

1.5 Types of attributes

There are different types of attributes based on which we can classify the various clustering techniques. These attributes can be as follows:

- Interval variables.
- Ratio variables
- Nominal or ordinal variables.
- And variables of mixed types.

1.6 Design Issues

There are certain issues that we need to keep in mind while designing a clustering algorithm:

- Type of attributes on which the algorithm is going to work.
- Scalability of dataset
- Ability to handle multidimensional data.
- Ability to support multi shape clusters

1.7: Types of clustering algorithms

Portioning based: Build various partions and then evaluate them based on some criteria

- K-Means: Every cluster is represented by center of the cluster.
- K-Medoids: Every cluster is represented by one of the object in the cluster.

Hierarchal based: Create a hierarchal breakdown of data objects.

- Agglomerative: Start with single cluster and at each step join two closest clusters.
- Deglomerative: Start with one cluster and then divide that cluster into sub-clusters and progress recursively on each sub-set..

Grid based: Distribute the data elements into various data cells.

- STING: A Statistical information grid approach for spatial data mining.
- CLIQUE: Used to cluster high dimensional data stored in large tables.

Density based: Collect the data objects based on some density function.

- DBSCAN: It is a density based algorithm creates clusters based on the density distribution of agreeing nodes.

- OPTICS: density based clustering algorithm which finds clusters in spatial data.

1.8 Tools Used

1.8.1 WEKA

WEKA is an open source data miming software developed in New Zealand by the university of Waikato that implements data mining algorithms for data preprocessing, classification, clustering, association rules. Using WEKA the data mining algorithms are directly applied to dataset provide by the user. The following picture showed how the WEKA home screen looks like.

Figure 1.2: Front screen of WEKA

E*xplorer:* This interface provide graphical front end environment for exploring data.

Experimenter: This interface provides an atmosphere for conducting experiments between different learning patterns and analysis of results.

Knowledge Flow: It is an alternate to Explorer for WEKA core algorithms written in java.

Simple CLI: It provides command line interface to user for the direct execution of WEKA commands. WEKA does not support all the data files it supports only ARFF file format. The data file which we need to use in WEKA first it needs to be converted into ARFF format.

The following are the steps to convert the .XLS file format into ARFF:

1) Save the excel file with .csv comma delimited format.
2) Open the file in MS-word and do the following changes.
3) Precede the file name with @relation and all the attributes with @attributes.
4) After doing the above changes we need to save the file by choosing the file type as plain test with line break.
5) Now the file is in ARFF format now he file is ready to use with WEKA.

	A	B
1	City	Temperature
2	New York, NY	77.8
3	Washington, DC	81.3
4	Dulles, VA	77.8
5	Richmond, VA	81.3
6	Atlantic City, NJ	77.5
7	Philadelphia, PA	79.6
8	Trenton, NJ	77.7
9	Wilmington, DE	77.8
10	Baltimore, MD	79.2
11	Norfolk, VA	81.1
12	Tampa, FL	84.5
13	Lakeland, FL	84.6
14	St. Petersburg, FL	85.6
15	Asheville, NC	75.4
16	Greenville, SC	81.0

Figure 1.3: Data Set in CSV comma delimited format

The following is the corresponding ARFF file:

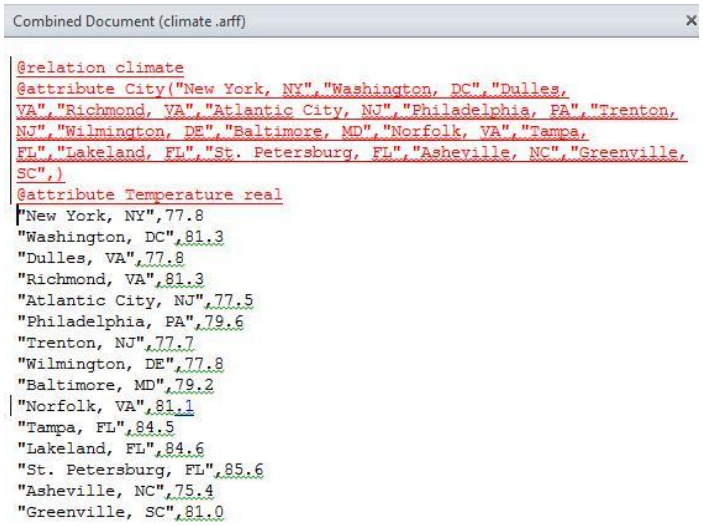


Figure 1.4: Data set in ARFF format

1.8.2 Eclipse

When exploring system source code many of the developers have reported the disorientation of source code. This problem occurs when developers lose the context of their recent action towards overall goal.

A study has been conducted in order to find out the answer for the following questions:

1. Whether the expert developers become disoriented.
2. Identification of various factors that can prevent disorientation.

Disorientation is not so frequent but it has great impact on the developers because it requires lots of time and effort to improve disorientation. Identification of the causes of the disorientation and the methods to remove the disorientation will improve the effectiveness of software developers. Correlating information gathered from different presentations require by the software developers so that the gathered content can be revealed on the screen at some specific point of time. Using visual momentum which is a qualitative quota of user's ability to extract pertinent information from different displays. Using this momentum user has identified the various factors in Eclipse UI that are responsible for introducing disorientation.

The following are the major factors for dis-orientation:

1. During program exploration absence of connecting navigation context.
2. Thrashing.
3. Isolated sub-task searching.

Integration of various collaborative features with IDE provides various benefits to software developers. Collaboration into IDE provides various challenges so collaborative features needs to be selected very carefully by considering the various issues of collaborative work over individual work. Here e we are focusing on Eclipse a collaborative tools used by developers to collaborate their work in IDE environment. (C.Murphy)

(Pradeep Kumar,1997) K-mean clustering method is used to generate the clusters and describing the member of each cluster whereas the technique based on rough set theory create s the clusters and describe the features of each cluster. Rough clusters produce more clusters as compared to traditional cluster analysis. More cluster means the object has a higher chance of being in more than one cluster by moving from lower estimate to boundary region and reducing the size of lower calculation.

Tian Zhang and Raghu Ramakrishnan et al (1997) discussed that the existing clustering algorithms are not capable of handling large datasets. As the size of dataset increases they do not scale up with well in terms of memory, running time and quality. In their paper they have proposed a new clustering algorithm named as BIRCH which is scalable enough to handle large datasets which was implemented a data structure named as CF-trees. The BIRCH algorithm architecture provides parallel and concurrent clustering which was prove successful when implemented on large datasets.

(Emmanuel Müller, 2011) Scalable density based sub space clustering method is proposed that directs mining to limited and selected sub space clusters. This method reduces the sub space processing by identifying sub spaces from categorical data and maintaining the accuracy by narrow down the search apace. Information is gathered from sub space regions and based on that information sub space clusters are generated the issue with this approach is that repeatable database scans and exponential search space. In this proposed algorithm which solves the various issues faced by the apriori based sub space clustering algorithm that were exponential search space and repeated database scans. At the end the authors conclude that the steering enables the density based clustering technique to provide better scalability and high quality results as compared to the existing subspace clustering techniques

(Shreya Jain, 2012) discussed and compare the various measuring parameters used for K-Mean clustering technique-Means clustering try to find user specified number of Clusters specified by their centroids. Authors in their paper discussed about the two major phases of clustering. The working of K-means algorithm and analyzed different sets of K-

values to enhance the performance of algorithm. Authors have discussed Silhouette plot techniques helps in finding the better value of K and arbitrarily chosen the preliminary centroid values.

Anoop Jain and Aruna Bajpai et al. (2012) discussed about the need of text clustering. The concept of descriptive clustering was introduced which group the semantically related documents and present it to the user in a compact way. Authors introduce the DCF (Description Comes First) algorithm which is used to implement descriptive clustering which increases the searching speed of a document. The problem of searching a document on search engine is that sometimes we issued queries that are short and confusing which results in large number of documents so it's very difficult to search from that large hit list. So the author proposes a clustering algorithm that is known as document clustering or text clustering algorithm which makes the cluster of documents based on the user query passed to the search engine which increases the searching speed of a document.

Er.Arpit Gupta ,Er. Ankit Gupta et al. (2012) discussed the impact of various clustering techniques on variety of data. The clusters are considered to be meaningful clusters if the resulting clusters capture the same structure. The major problem in handling high dimensional data is the number of dimensions. Various methodologies used to handle high dimensional data is discussed. No single approach can handle different types of data that's why different approaches were discussed to handle different types of data.

Narendra Sharma and Aman Bajpai et al. (2012) focused on the comparison of clustering algorithms available in WEKA. WEKA tool is used to analyze the data by considering different dimensions and after analysis represent the results in concise form. The main objective is to compare the results of different clustering algorithms of WEKA and let the end user know which clustering is better and why. The authors conversed that they have use WEKA because it can be used without knowing the deep knowledge of data mining techniques and this is the only tool that provides graphical user interface. In this paper they have discussed the advantages and disadvantage of clustering algorithms available in WEKA and after analysis of results they have concluded that K-means is the best and simplest algorithm from all others.

Can Gao and Witold Pedrycz et al. (2013) Clustering of categorical data arise the problem in data mining process. The proposed algorithm deal with the problem of categories data by decomposing the attributes of categorical data into rough sub-spaces. A novel clustering algorithm is then proposed which is based on rough sub spaces to deal with categorical data. For categorical data the results of clustering algorithm is evaluated by introducing cluster index. The algorithm in this paper combines only the rough sub apace portioning's and the diversity of individual algorithms also retained. The future scope of this algorithm is to speed up the process of searching rough sub spaces and refining the strategy for selecting rough sub spaces with improved quality.

(Joshi Aastha, 2013) have discussed about the comparison of various clustering techniques. They have done the comparative study of six clustering techniques. Clustering is used as data mining tool which group similar data objects into one cluster and group dissimilar data objects into other clusters. At the end they have conferred that in K-mean from partitioning method is the simplest and better than any hierarchal method but the problem with K-mean is that it works only for numerical values. Further the problem with partitioning and hierarchal clustering is that they generate the clusters of only spherical shapes and this problem is resolved by density based methods which creates the clusters of random shapes. In density based we need not to specify the number of clusters in advance as it was needed in K-means. The problem with hierarchal lustering is that once the sub-clusters were made they cannot be changed.

(Jagadeeswaran V.S., 2013) conversed that use of BIRCH algorithm for noise detection in large dataset. From large and multidimensional dataset it is very difficult to obtain cluster from such data. The authors take into consideration agriculture dataset for the outlier detection using BIRCH algorithm which gives improved results as compared to other algorithms when worked on large dataset. Outlier detection is a mechanism in data mining to detect values that deviates from other values. Handling outliers may alter the normal objects and blur the distinction between normal objects and outliers.

(Mittal, 2014) have made the analysis various clustering algorithms by obtaining different clustering when applied to education dataset by using WEEKA tool. Clusters can be made based on various parameters that can be a distance, density and intervals. Based on these parameters they have classified and compare K-means, Hierarchal, and density based algorithms. At the end they conclude that by removing the limitations of these algorithms we can further improve the clustering algorithms.

(S, 2014) discussed about the current review of K-means clustering algorithm. The work carried out by different researchers using K-means is recognized. Raw data cannot be used as such for analysis and decision making so by using one of the clustering algorithms of analysis of data. The emphasis is also on various challenges and restrictions of K-means algorithm. From the last few years' lots of improvements had made in K-mean but still lots of efforts are required to improve the performance of this algorithm. [9]

Yogita Rani and Manju et al. (2014) discussed about the comparative analysis of BIRCH and CURE algorithm using WEKA tool. Both the algorithms are classified under hierarchal clustering and used for large dataset. Comparisons on both the algorithms were made by considering complexity, geometry, noise and running time as parameters based on which comparisons were made. At the end it was concluded that Cure is better than BIRCH because BIRCH is restricted to produce only spherical shape clusters having uniform size whereas CURE can produce clusters of non-spherical shapes with patchy sizes.

CHAPTER 3

PRESENT WORK

3.1 Problem Definition

Speeding the creation and searching of the rough spaces, as the process required huge amount of time to mine the information data while creating the rough space.

3.2 Objectives

- Modify the existing algorithm to speed up the process.
- Reduce the rough sub spaces data in categorical data so that processing time and space will be reduced.
- Simulate the new Algorithm in java and weka.
- Compare the existing algorithm with new algorithm to evaluate the efficiency in terms of time and space

3.3 Research Methodology

Our aim is to make our technique speedy in all respect of collecting sub spaces from different data sets. Our proposed clustering technique is very useful for the stability and robust point of view for this algorithm. According to K-mean clustering algorithm we could not ensembles the rough sub spaces from categorical data because of which the mining process take too much of time and space for such data to create the clusters of such data. Here, we proposed an algorithm which can search and ensembles the rough sub space data at very high speed with low latency. When we focused on the previous algorithms they all are very slow in processing the data, which affects the different parameters of clustering algorithm like, buffer capacity, Agility, Multi-tenancy, Peak-load capacity, reliability and Utilization and efficiency. In our proposed algorithm we are separating the rough sub spaces from the categorical data which increases the processing speed.

The following algorithm is divided into 3 steps .The first step scans the database, searching algorithm starts working for finding the rough sub spaces to form the clustered data. Rough subspaces are identified on the basis of their attributes and

cardinality. So, we are focusing on performance and significance of rough sub spaces which means higher the significance of rough sub space is directly proportional to higher its suitability for clustering ensemble. In the second phase scanning of database for categorical data has been done which can improve the efficiency because of the extraction of rough sub spaces from the categorical data. Finally the large data set will be generated which will be output.

3.4 Algorithm

Input String-> (U, A, V,F)

Output-> All rough subspaces of IS

Step 1 Set discernibility matrix M=Ø, core set Core= Ø and all rough subspace set REDS= Ø;

Step2 Compute the discernibility matrix M and exclude unnecessary elements of discernibility matrix M with the law of absorption;

Step 3 Add the singlection in M to Core set Core, M=M-Core; push Core and M to data stack L for rough surface;

Step 4 If L is null, then goto step 10;

Step 5 Pop the last attribute set RED and candidate information M';

Step 6 If M' is null, REDS= REDS ∪ RED, goto Step4;

Step 7 Select Maximum Frequency attribute a, M'=M"-$RS_{M'}$({a}), exclude unnecessary elements of

$CS_{M'}$({a}) with the law of absorption;

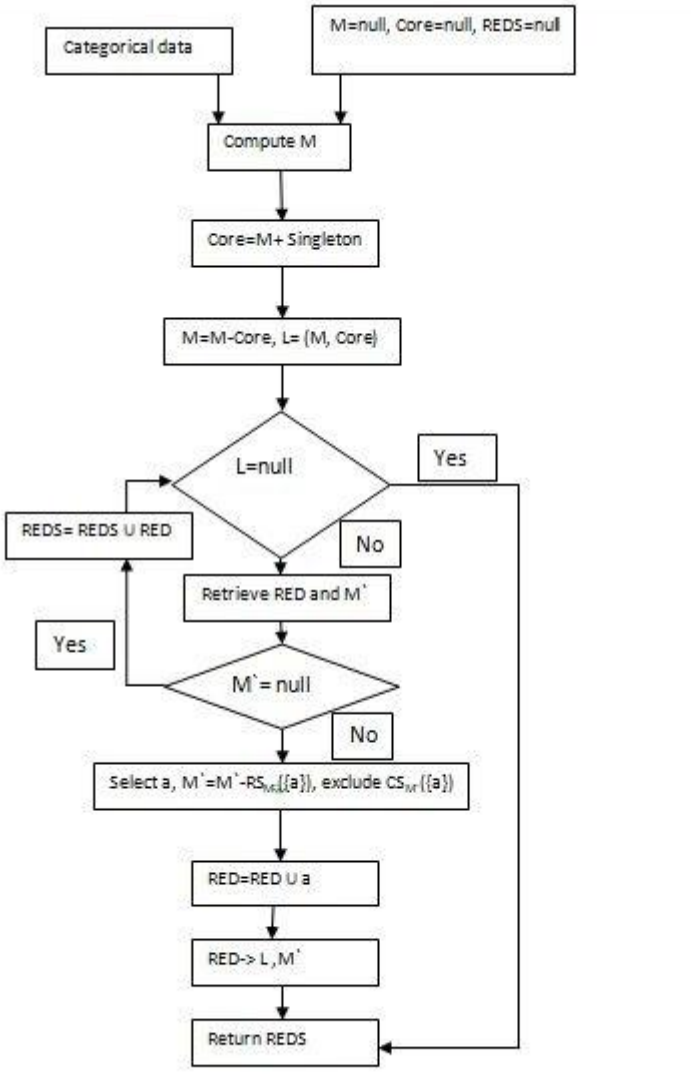
Step 8 RED=RED ∪{a}; push RED and M' to L;

Step 9 Return REDS.

Step 10 Output Values with rough subspaces (IS).

In input sting U is the fine set of objects which are non-empty ,A is finite set of attributes which are non-empty, V is the attribute domain union and F is the function which associate an object a unique value of each attribute belonging to U.

3.5 Flow Chart



CHAPTER 4

RESULTS AND DISCUSSION

4.2 Results

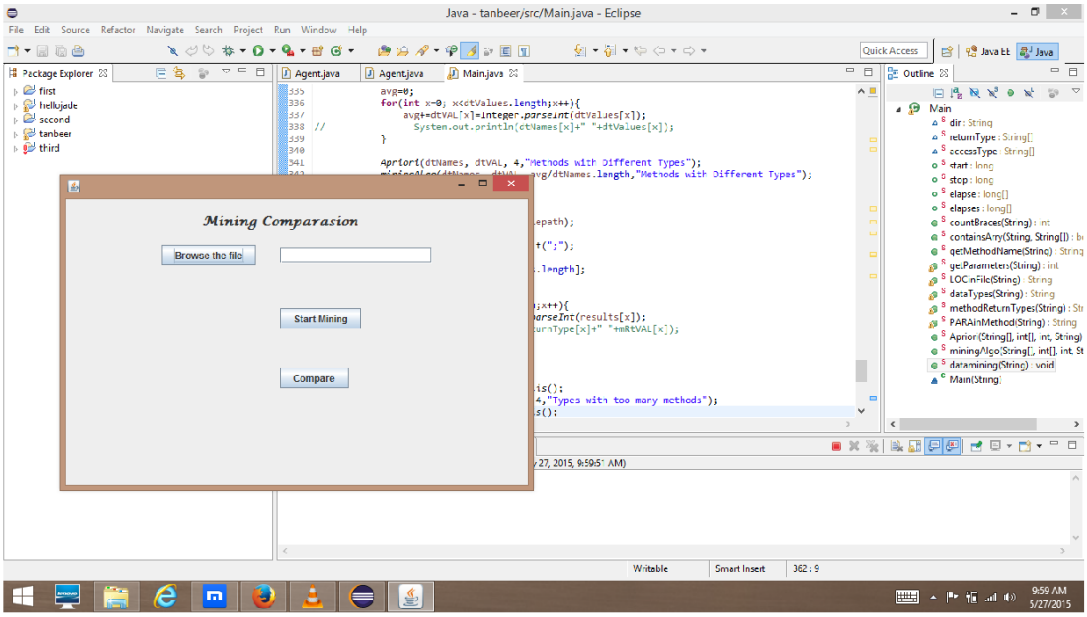


Figure 4.4: Data mining interface

The above figure shows the interface of our implemented algorithm application in java eclipse. This is our first implementation where we can start searching of different data and after that it is classify on the basis of categorical data and subspaces.

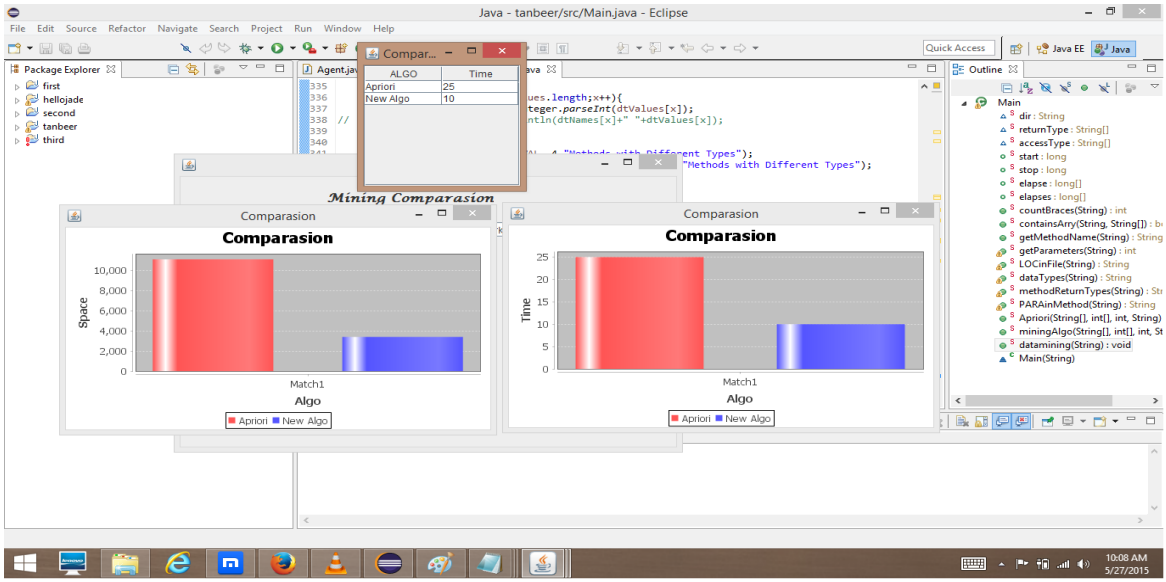


Figure 4.5: Space and time comparison

The figure shows the comparison between time and space complexity of algorithm for two different files. This is done here because in Weka there we cannot show the graph chart and table that's why it is needed to implemented in java first.

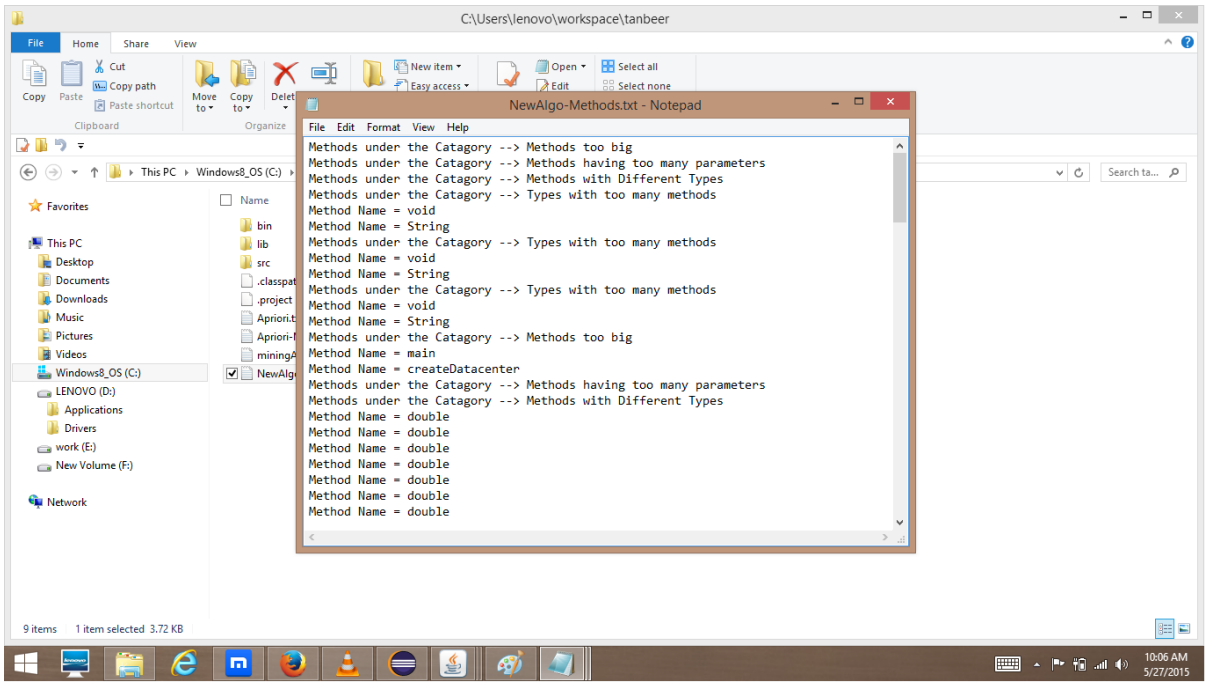


Figure 4.6: Result analysis file

The above figure shows result analysis of the algorithm in eclipse. Data sets for this file have selected which is .java extension.

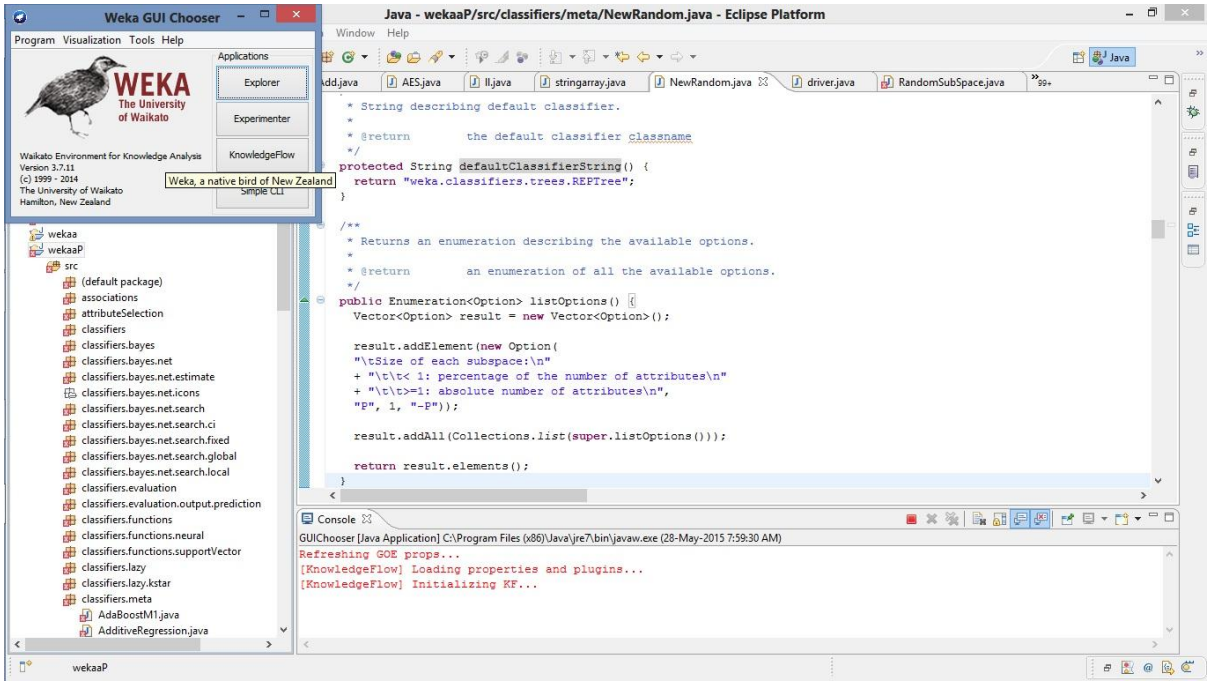


Figure 4.7: Weka Implementation

The above figure shows the Weka running GUI from here we are going to explorer menu and there we select our data sets on which we are going to apply our algorithm and visualize the cluster formation for subspaces.

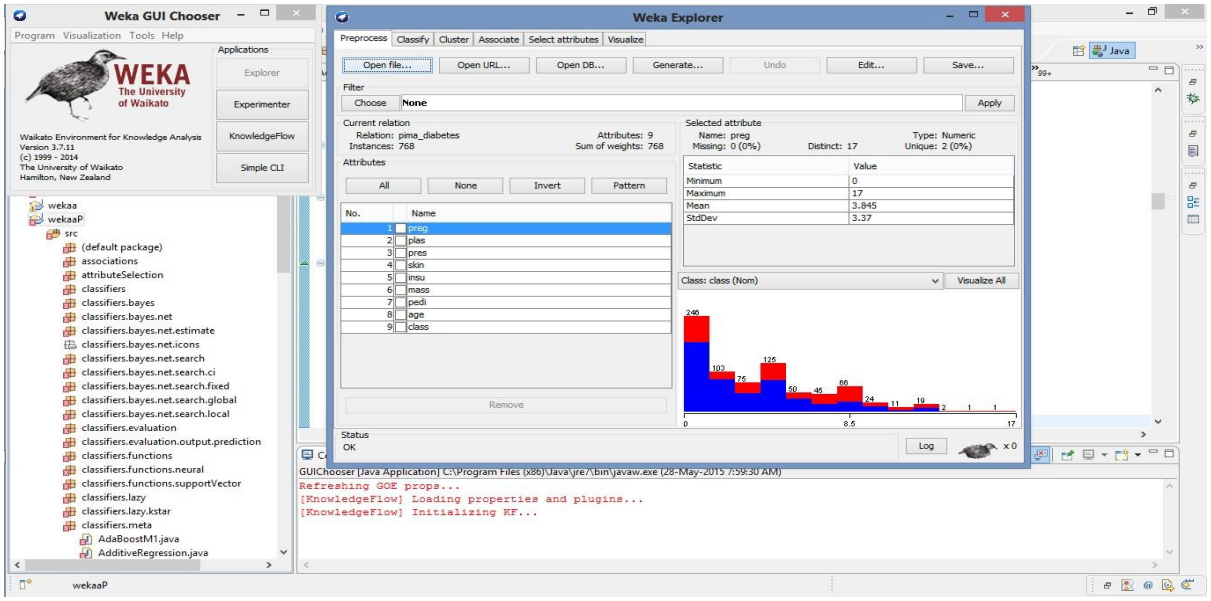


Figure 4.8: Loading of data sets

Figure 4.8 shows the data sets are uploaded in the form of bar charts by using different methods and visualization of data sets is shows in left hand side.

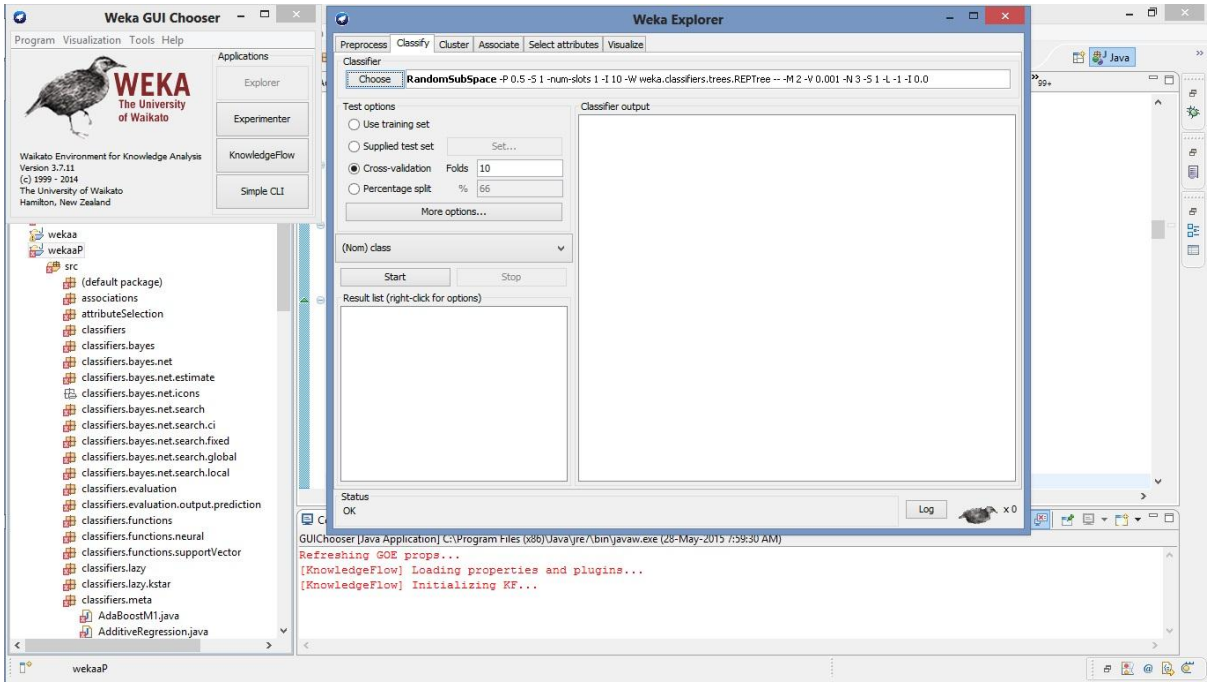


Figure 4.9: Classification of rough space clustering

The above figure shows the various classifications of rough space clustering algorithms.

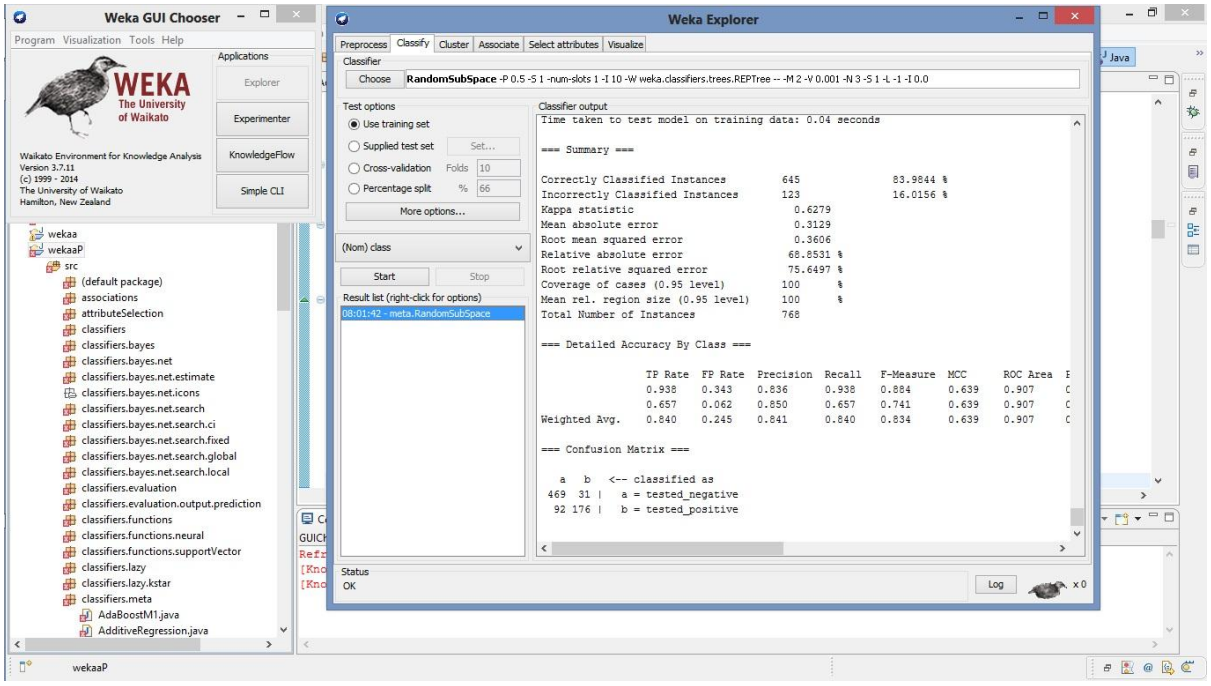


Figure 4.10: Results generated for rough subspace

The above figure shows the reports generated for the file which is uploaded as a data sets in Weka and it shows negative tested and positive tested values as it classified a=469/92 and b=31/176 respectively. Correctly classified instances are 83%, mean absolute error is 0.3129 all these results shows the stability of algorithm.

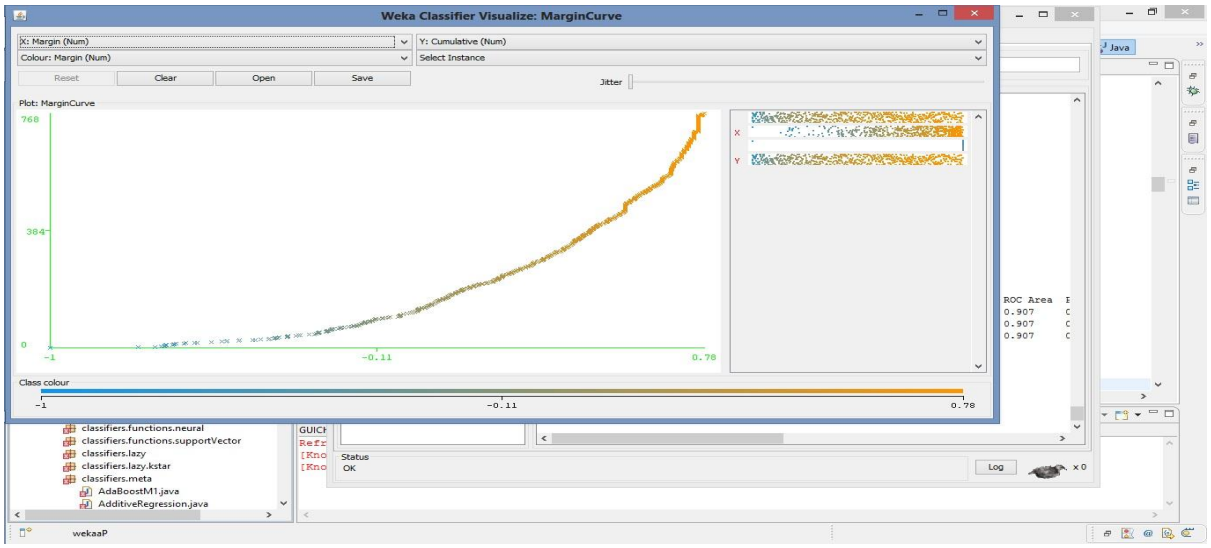


Figure 4.11: Curve generated on basis of our algorithm for multiple cluster formation

Figure 4.11 describes about the margin curve for multiple cluster formation. Multiple cluster formations means when large number of data sets is analyzed then many clusters are formed for different values. That's why this graph shows the variety of clusters in same rough subspaces. When subspaces in data sets are found, it will classify it along with the same clusters of those data sets which help in analysis.

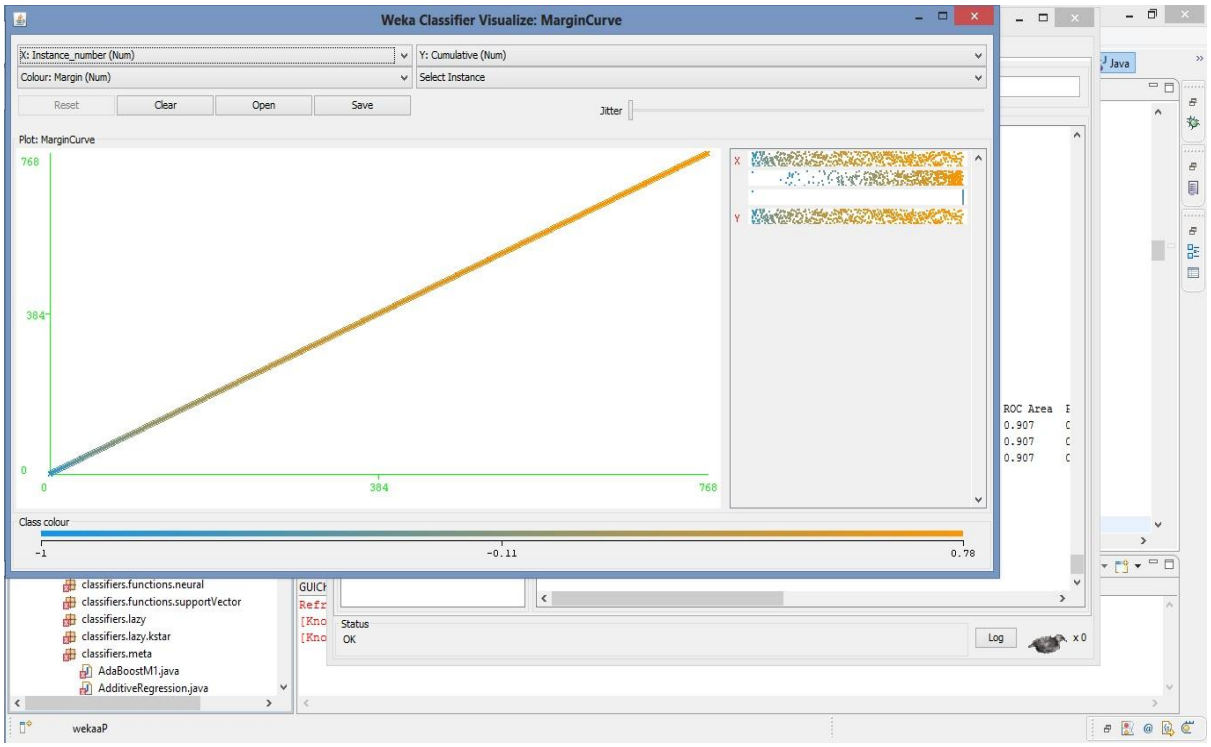


Figure 4.12: Curve with single cluster formation

The above figure shows the margin curve with single cluster formation.

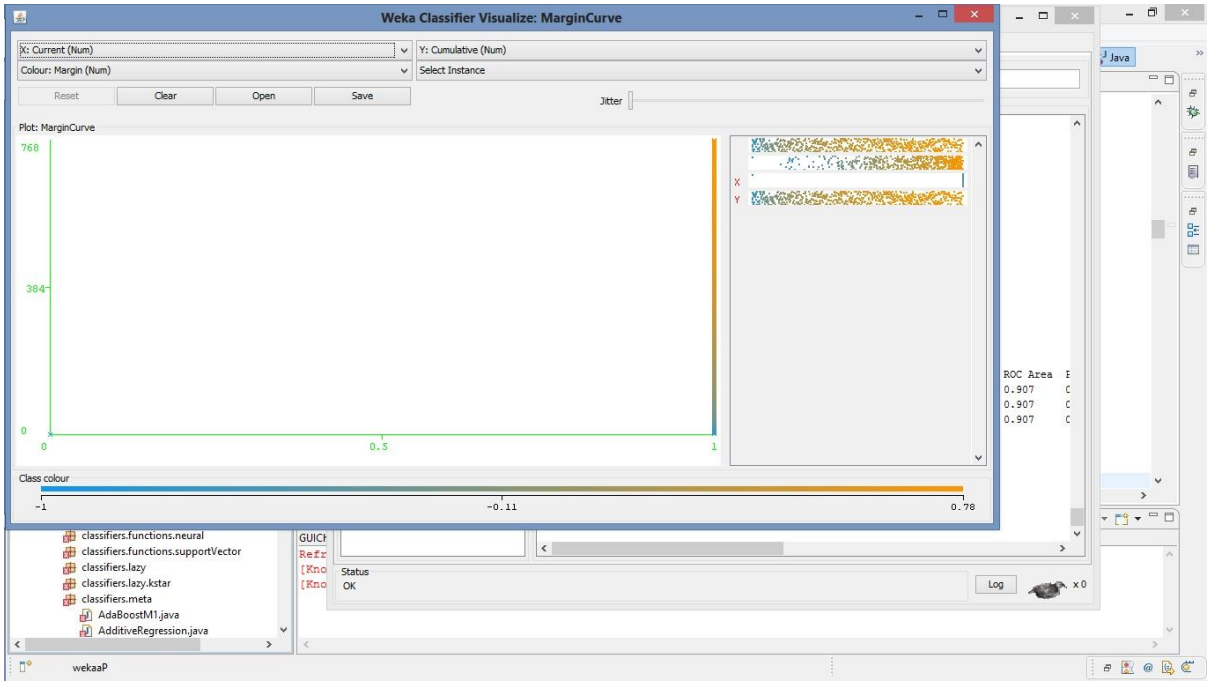


Figure 4.13: Curve with non-cluster formation

The above figure shows the margin curve for non-cluster formation under this a slant line and straight bar line is generated respectively.

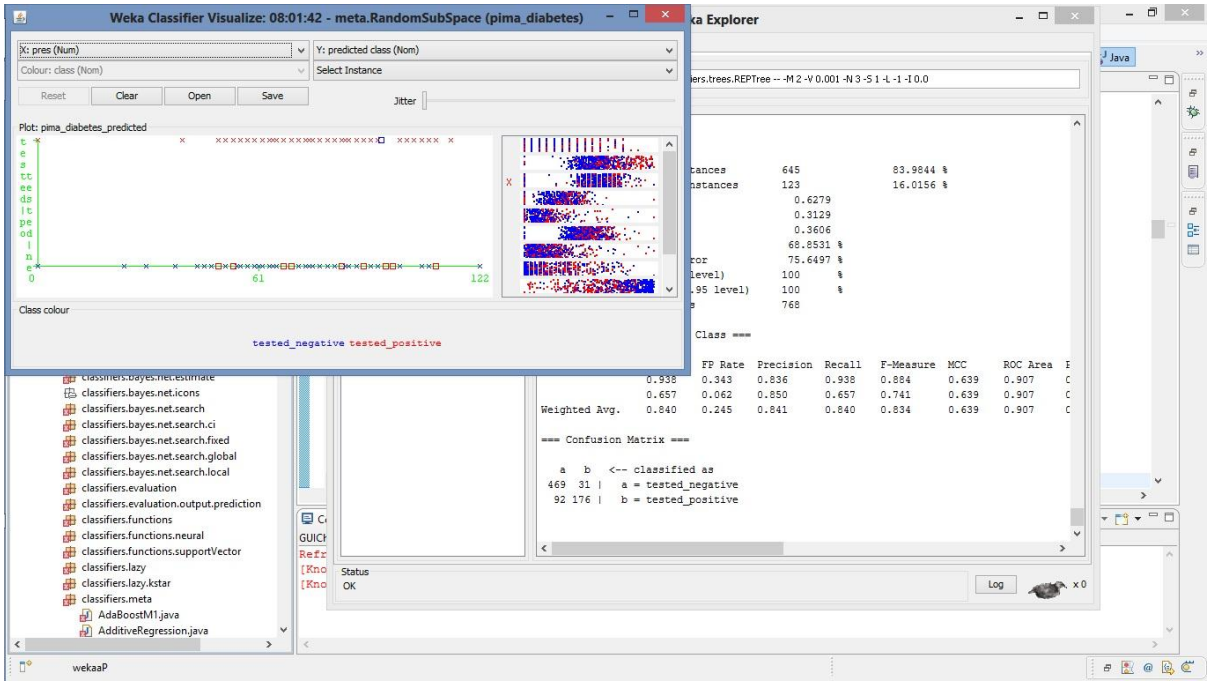


Figure 4.14: Error Classifier visualization of clusters

The above figure shows the graph of error classifier of whole data sets. This figure explains that according to each cluster how much error rate is generated which is very small or negligible.

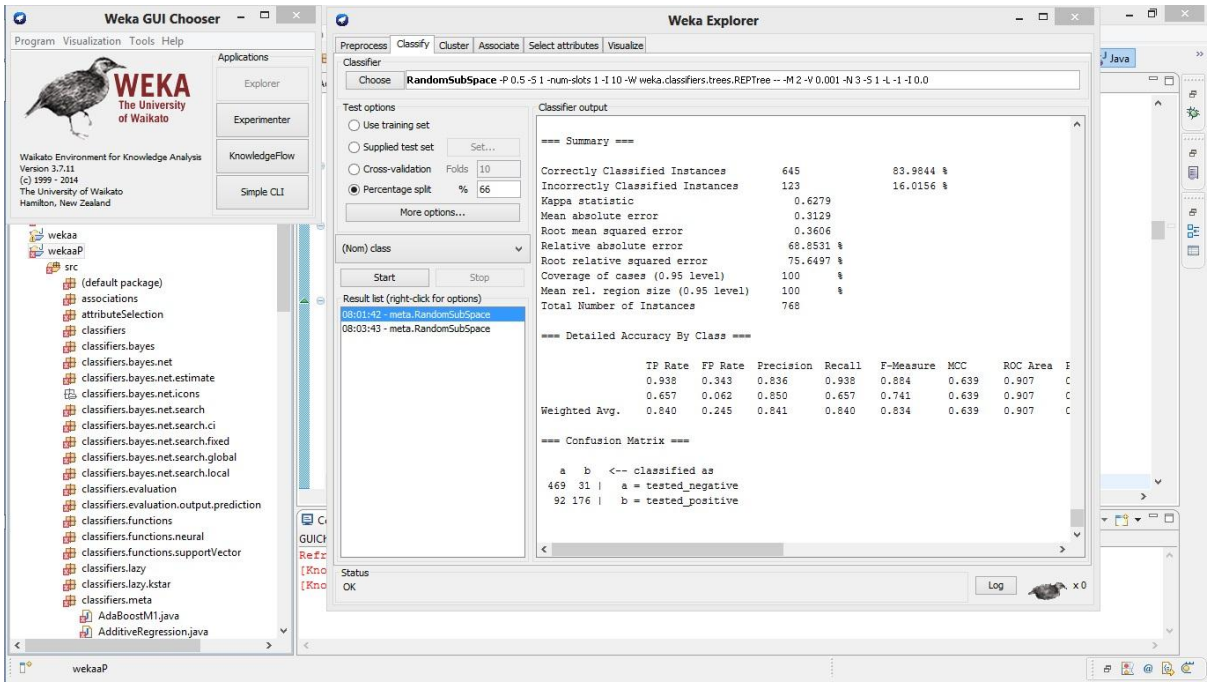


Figure 4.15: Summary report for random subspace 1

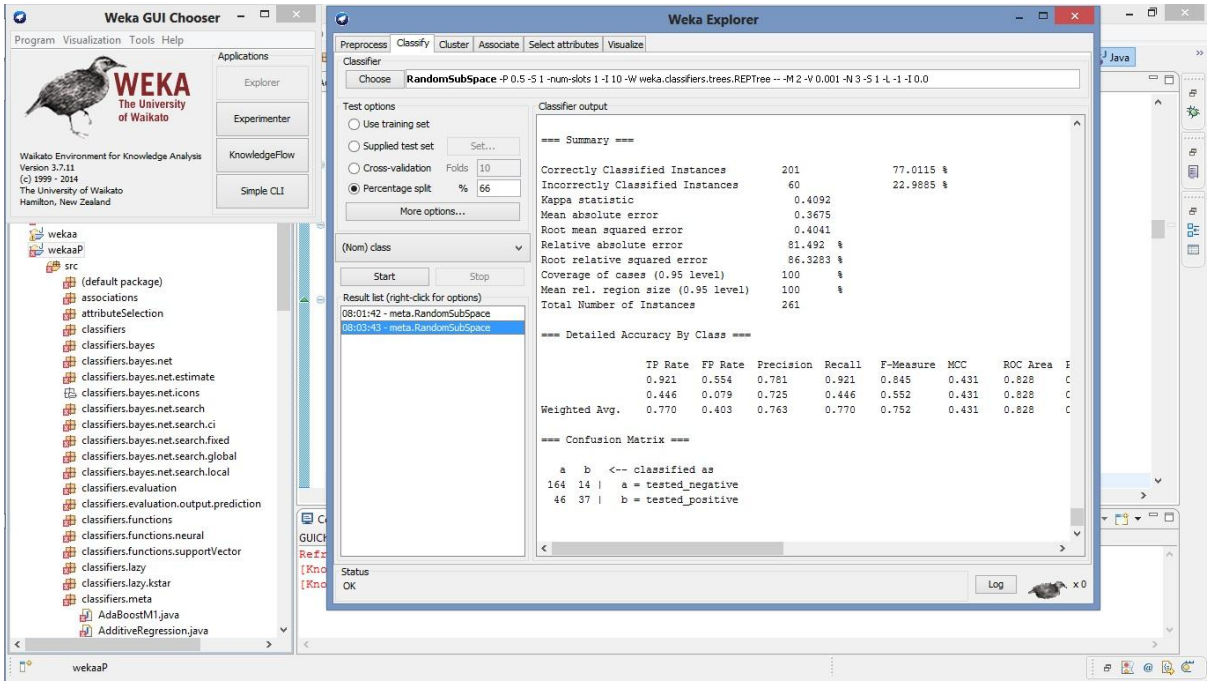


Figure 4.16: Summary report for random subspace 2

Figure 4.15 and figure 4.16 shows the summary reports of different algorithm on same data sets.

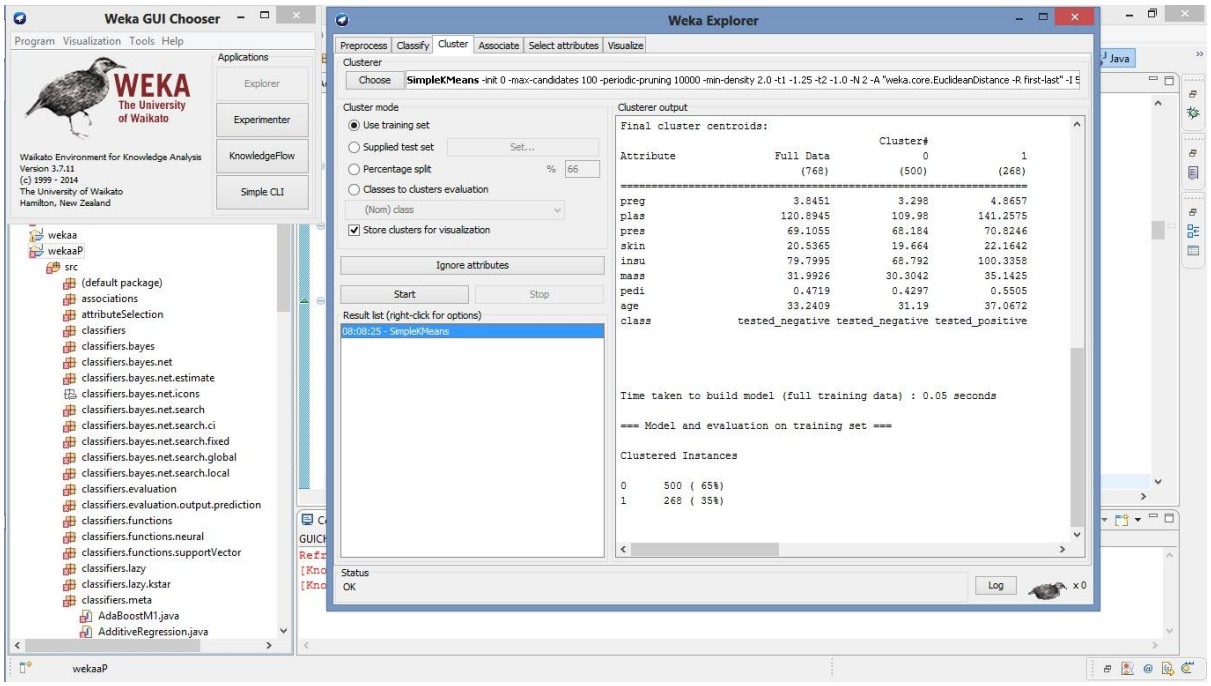


Figure 4.17: Cluster report information

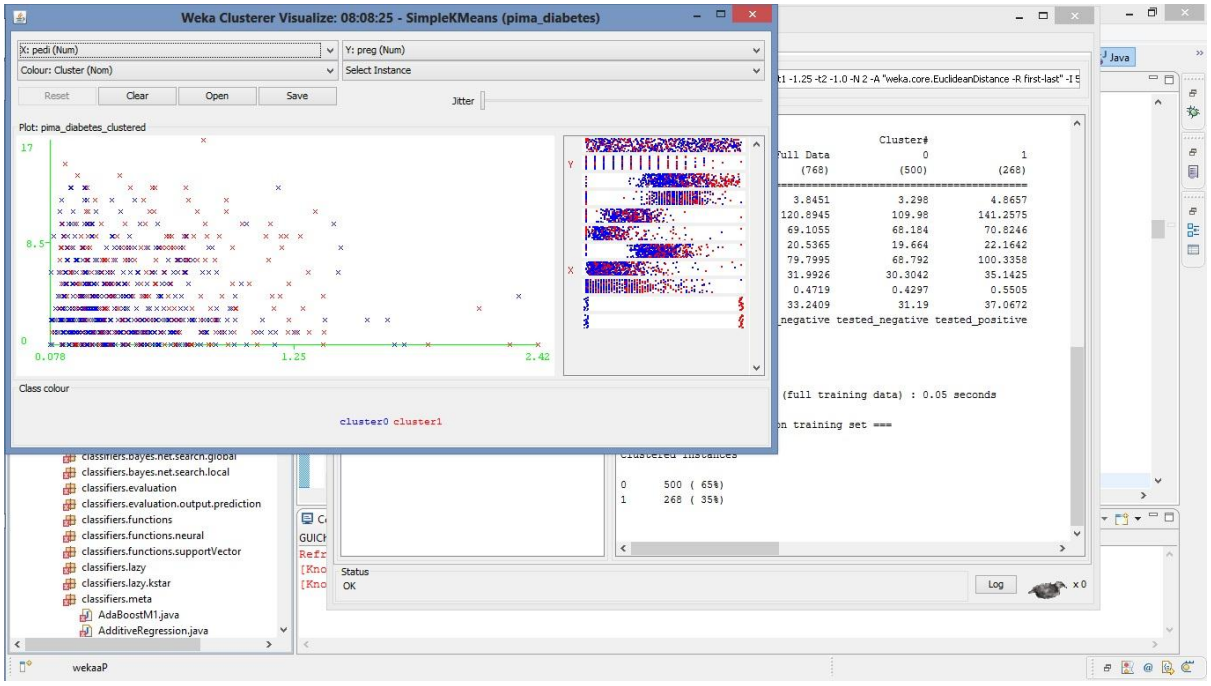


Figure 4.18: K mean cluster formation

Figure 4.18 describes about the k means clustering algorithm information that how much it takes the time and how many types of attributes are present in that data sets.

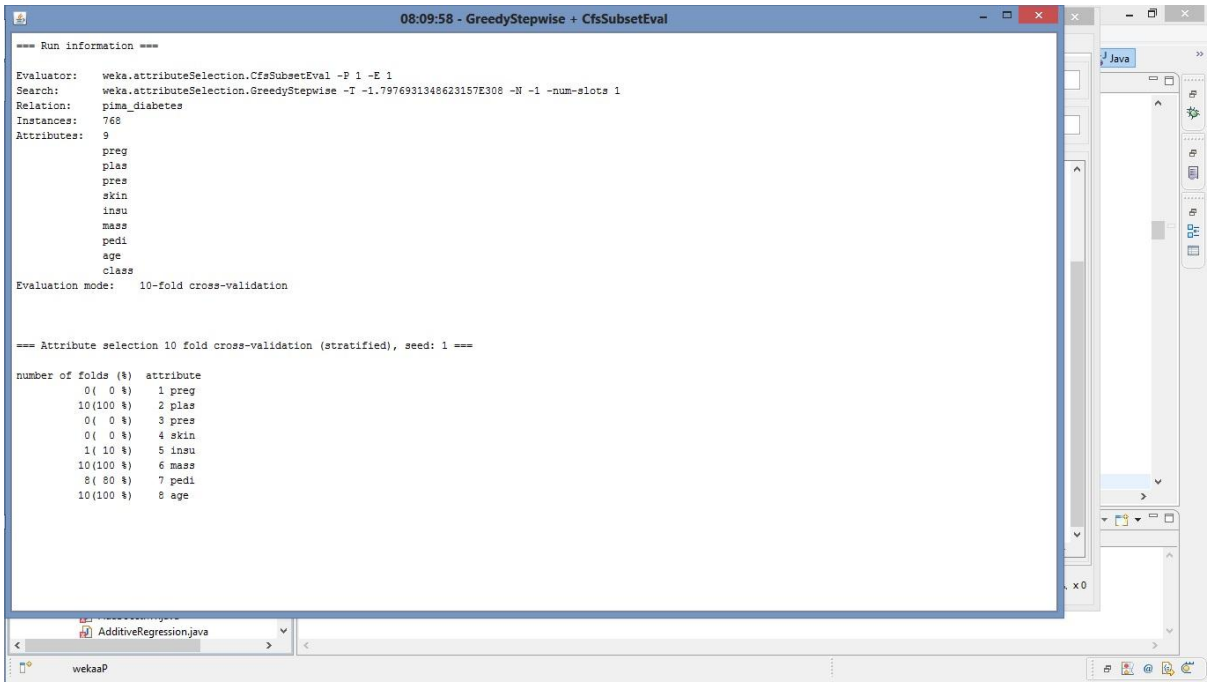


Figure 4.19: Attribute results

The above figure shows the attributes presents in the data sets.

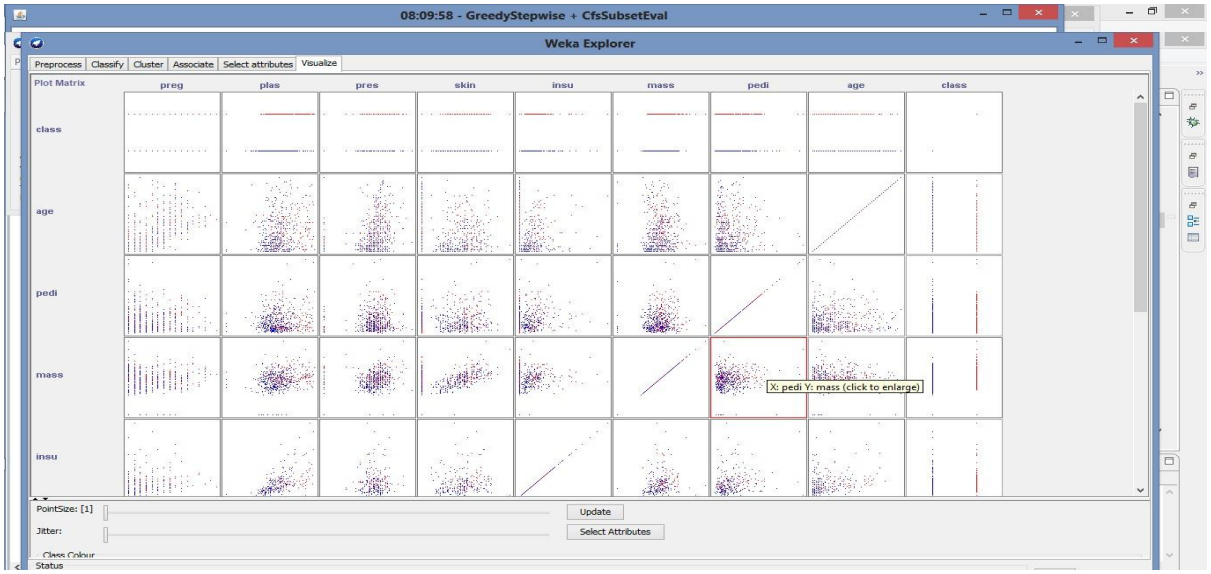


Figure 4.20: Cluster classification results

The above figure shows the cluster formation based on different parameters which is visualizes the clusters of data sets, which helps in understanding the data classification.

This dissertation proposed an algorithm which works on rough subspaces cluster formation. Now a day there is lots of data is presence around all of you. So clustering is most important thing for classification of data in different field. Just take an example of let suppose we want fruits, as we move to market then there is variety of fruits are present and they all are kept on single bucket. So it is impossible to select out banana, mango, and grapes in easy way. But as we make the clusters of mango, banana, grapes, then it is very easy to select the fruits according to our need from basket. This is the same concept should be applied for data mining. If we want any data of any file then we have to mine that area very well, where it takes too much time and if we make clusters of that file then it is easy to handle it.

Table 1.1: Summary report for Data set1

Classified as	A	B
Tested Negative	469	31
Tested Positive	92	176

Summary	
Correctly Classified Instances	645
Incorrectly Instances	128
Kappa Statistics	0.6279
Mean Absolute error	0.3129
Root Mean Square error	0.3606
Relative absolute error	68.8531
Root relative square error	75.6497

Table 1.2: Summary report for Data set2

Classified as	A	B
Tested Negative	164	14
Tested Positive	46	37

Summary	
Correctly Classified Instances	201

Incorrectly Instances	60
Kappa Statistics	0.4092
Mean Absolute error	0.3676
Root Mean Square error	0.4041
Relative absolute error	81.492
Root relative square error	86.3283

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In this section conclusion of our dissertation is mentioned which is as below:
The proposed work was the clustering algorithm, where each searching and classification of data is done. It is done in the way where worker creates its own index and clustered file, and both files are stored. By this proposed work the speed of cluster formation for categorical rough subspaces is formed. Here we reduces the subspaces formed in data sets, this is done by selecting the attributes of same type and formed a cluster of these same type attributes within it.

5.2 Future work

By this proposed work the processing time and space is reduced to 60 %. Hence it would be more beneficial for future if we implement this with more efficient parameters and it would be more accurate in classification of data for data analysis. We can further enhance the speed and reduce the time and space complexity by using high class and different sub spaces.

CHAPTER 6

REFERENCES

Reference to a book

Jiawei Han, M. K. (2011). *Data Mining: Concepts and Techniques.* The Morgan Kaufmann .

Reference to Journals

C.Murphy, B. d. (n.d.). Using Visual Momentum to Explain Disorientation in the Eclipse IDE.

Can Gao, W. P. (2013). Rough subspace-based clustering ensemble for categorical data . *Springe*.

Er. Gupta Arpit, E. G. (2012). Research paper on cluster techniques of data variations. *International Journal of Advance Technology & Engineering Research (IJATER)* , ISSN NO: 2250-3536.

Jagadeeswaran V.S., P. (2013). Detection of noise by efficient hierarchical birch algorithm for large data sets. *International Journal of Advanced Research in Computer and Communication Engineering*, ISSN 2319-5940.

Jain Anoop, B. A. (2012). Efficient Clustering Technique for Information Retrieval in Data Mining. *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459.

Joshi Aastha, K. R. (2013.). Comparative Study of Various Clustering Techniques in Data Mining . *International Journal of Advanced Research in Computer Science and Software Engineering* , ISSN: 2277 128X.

Mittal, S. a. (2014). Comparison and Analysis of various Clustering Methods in Data Mining on Educational data set using WEEKA tool. *International Journal of Emerging Trends & technology in Computer Science(IJETTCS). Volume3,Issue2,* ISSN 2278-6856 .

Mittal, S. a. (2014). Comparison and Analysis of various Clustering Methods in Data Mining on Educational data set using WEEKA tool. *International Journal of Emerging Trends & technology in Computer Science(IJETTCS). Volume3,Issue2,* ISSN 2278-6856.

P. Indira Priya, a. D. (2014). A Survey on Different Clustering Algorithms in Data Mining Technique. *International Journal of Modern Engineering Research (IJMER)* , ISSN: 2249-6645 267-274 .

Pradeep Kumar, P. R. (n.d.). Rough clustering of sequential data. *ELSEVIER*.

S, S. S. (2014). A Review ON K-means DATA Clustering APPROACH. *International Journal of Information & Computation Technology*, ISSN 0974-2239.

Sharma Narendra, B. A. (2012). Comparison the various clustering algorithms of wekaTools. *International Journal of Modern Engineering Research*, ISSN 2250-2459.

Shreya Jain, S. G. (2012). Comparing and Selecting Appropriate Measuring Parameters for K-means Clustering Technique. *International Journal of Soft Computing and Engineering (IJSCE)* , ISSN: 2231-2307.

Tian Zhang, R. R. (1997). BIRCH : A new Data Clustering Algorithm and its Application. *Data Mining and knowledge Discovery.Volume1,Issue2* , 141-182.

Xumin, L., & Guan Yong, S. N. (2010). An Improved k-means Clustering Algorithm. *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium* (pp. 63 - 67). Jinggangshan: IEEE.

Yogita Rani, M. R. (2014). Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9 . *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1*.

C.Murphy, B. d. (n.d.). Using Visual Momentum to Explain Disorientation in the Eclipse IDE.

Er. Gupta Arpit, E. G. (2012). Research paper on cluster techniques of data variations. *International Journal of Advance Technology & Engineering Research (IJATER)* , ISSN NO: 2250-3536.

Jagadeeswaran V.S., P. (2013). Detection of noise by efficient hierarchical birch algorithm for large data sets. *International Journal of Advanced Research in Computer and Communication Engineering*, ISSN 2319-5940.

Jain Anoop, B. A. (2012). Efficient Clustering Technique for Information Retrieval in Data Mining. *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459.

Jiawei Han, M. K. (2011). *Data Mining: Concepts and Techniques.* The Morgan Kaufmann .

Joshi Aastha, K. R. (2013.). Comparative Study of Various Clustering Techniques in Data Mining . *International Journal of Advanced Research in Computer Science and Software Engineering* , ISSN: 2277 128X.

Mittal, S. a. (2014). Comparison and Analysis of various Clustering Methods in Data Mining on Educational data set using WEEKA tool. *International Journal of Emerging Trends & technology in Computer Science(IJETTCS). Volume3,Issue2*, ISSN 2278-6856 .

Mittal, S. a. (2014). Comparison and Analysis of various Clustering Methods in Data Mining on Educational data set using WEEKA tool. *International Journal of Emerging Trends & technology in Computer Science(IJETTCS). Volume3,Issue2*, ISSN 2278-6856.

P. Indira Priya, a. D. (2014). A Survey on Different Clustering Algorithms in Data Mining Technique. *International Journal of Modern Engineering Research (IJMER)* , ISSN: 2249-6645 267-274 .

S, S. S. (2014). A Review ON K-means DATA Clustering APPROACH. *International Journal of Information & Computation Technology*, ISSN 0974-2239.

Sharma Narendra, B. A. (2012). Comparison the various clustering algorithms of wekaTools. *International Journal of Modern Engineering Research*, ISSN 2250-2459.

Shreya Jain, S. G. (2012). Comparing and Selecting Appropriate Measuring Parameters for K-means Clustering Technique. *International Journal of Soft Computing and Engineering (IJSCE)* , ISSN: 2231-2307.

Tian Zhang, R. R. (1997). BIRCH : A new Data Clustering Algorithm and its Application. *Data Mining and knowledge Discovery.Volume1,Issue2* , 141-182.

Xumin, L., & Guan Yong, S. N. (2010). An Improved k-means Clustering Algorithm. *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium* (pp. 63 - 67). Jinggangshan: IEEE.