

A Comparative Study on the Performance of Enhanced Clustering Algorithm

A Dissertation Proposal submitted by

Sonia

(Registration No. 11310812)

To

Department of CSE/IT

In partial fulfillment of the requirements for the award of the Degree of

Master of Technology in Computer Science Engineering

Under the guidance of

Mrs. Alpana Vijay Rajoriya

Assistant Professor (CSE/IT)

Lovely Professional University, Punjab

May, 2015

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student : Sonia
Batch : 2013-2015
Session : 2014-2015

Registration No : 11310812
Roll No : RK2306B63
Parent Section : K2306

Details of Supervisor:

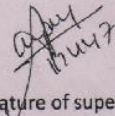
Name : Alpana Vijay Rajoriva
UID : 17447

Designation : Assistant Professor
Qualification : M.Tech
Research Exp. : 1 year

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

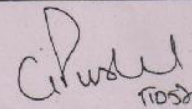
Proposed Topics:-

1. Research and improvement of clustering data mining algorithm on large data set.
2. An enhancement on clustering data mining.
3. An improvement on clustering data mining algorithm.


Signature of supervisor

PAC Remarks:

Total 1 approved. Paper expected.


11058

APPROVAL OF PAC CHAIRPERSON:


Signature:

Date:

*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

Clustering is a term used to divide objects into groups based upon their similarity. This research propose a enhanced k-mean clustering algorithm which improves the error function and this enhanced algorithm will apply on traffic dataset to analyze the major factors contributing to the accidents. The enhanced algorithm is compared with the existing k-mean algorithm using the software package weka.

CERTIFICATE

This is to certify that **Sonia** has completed M.Tech dissertation proposal titled **A Comparative Study on the Performance of Enhanced Clustering Algorithm** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature of Advisor

Name:Alpana Vijay Rajoriya

UID: 17447

ACKNOWLEDGEMENT

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

DECLARATION

I hereby declare that the dissertation proposal entitled **A Comparative Study on the Performance of Enhanced Clustering Algorithm** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 2 May 2015

Investigator: Sonia

Registration No. 11310812

TABLE OF CONTENTS

Chapter 1 INTRODUCTION..... 1

 1.1 Database 1

 1.2 Characteristics of Database..... 1

 1.3 Data Warehouse 3

 1.4 Data Mining 3

 1.5 Techniques of Data Mining 5

 1.6 Types of Data in data mining techniques:..... 6

 1.7 Cluster Analysis 7

 1.8 Example of clustering 10

 1.9 K-means clustering 10

 1.10 K-Means Algorithm Properties..... 11

 1.11 Advantages and Disadvantages of K-mean Clustering..... 11

Chapter 2 LITERATURE REVIEW 13

Chapter 3 PRESENT WORK 20

 3.1 Problem Formulation 20

 3.2 Scope..... 20

 3.3 Objectives 20

 3.4 Enhanced Algorithm 21

 3.5 Research Methodology 22

 3.6 Tool Used..... 24

Chapter 4 RESULTS AND DISCUSSION 31

Chapter 5 CONCLUSION 48

REFERENCES 49

APPENDIX..... 51

LIST OF TABLES

Table 1 Error Comparison of Both the Algorithms.....	41
Table 2 Number of Iterations.....	42

LIST OF FIGURES

Figure 1.1 Data mining as a step in the process of knowledge discovery.....	5
Figure 1.2 Process of clustering.....	8
Figure 1.3 Document clusters are formed from scattered documents.....	8
Figure 3.1 Flow chart of research.....	23
Figure 3.2 Preprocess panel of weka explorer.....	26
Figure 3.3 Weka classify panel of weka explorer.....	27
Figure 3.4 Clustered window panel of weka explorer.....	28
Figure 3.5 Associate window panel of weka explorer.....	28
Figure 3.6 Select attributes window panel of weka explorer.....	29
Figure 3.7 Visualize panel of weka explorer.....	30
Figure 4.1 Traffic dataset attributes.....	31
Figure 4.2 Some traffic dataset attributes.....	32
Figure 4.3 Some other Traffic dataset attributes.....	33
Figure 4.4 Shows weka GUI and how to open a file.....	33
Figure 4.5 Classify various attributes of dataset.....	34
Figure 4.6 Shows label and value of speed attribute.....	34
Figure 4.7 Shows label values of vehicle attribute.....	35
Figure 4.8 Shows no(blue) and yes(red) cluster.....	35
Figure 4.9 Shows numeric attribute in terms of mean and standard deviation.....	36
Figure 4.10 Shows Results of various attributes related to our dataset.....	36
Figure 4.11 Shows the plot matrix of the dataset.....	37

Figure 4.12 Classification of no (blue) and yes (red) clusters.....37

Figure 4.13 Classification of young (blue) and old (red) clusters.....38

Figure 4.14 Shows how to fetch an enhanced algorithm in weka tool.....38

Figure 4.15 Shows various attributes in enhanced algorithm in weka explorer.....39

Figure 4.16 Result of existing k-mean algorithm.....40

Figure 4.17 Result of enhanced k-mean algorithm.....41

Figure 4.18 Result of enhanced k-mean algorithm.....42

Figure 4.19 Bar graph shows performance comparison with same number of iterations...42

Figure 4.20 Shows probability of first cluster.....43

Figure 4.21 Shows probability of second cluster.....44

Figure 4.22 Shows probability of third cluster.....44

Figure 4.23 Shows probability of fourth cluster.....45

Figure 4.24 Shows probability of fifth cluster.....45

Figure 4.25 Shows probability of sixth cluster.....46

Figure 4.26 Shows each clustered instances.....46

Chapter 1

INTRODUCTION

1.1 Database

A database is a collection of related data. It is a piece of software or program that allows users to store retrieve and update collections of data. It is managed by some type of a database management system. A system that enables to access, organize and select data in a database is known as database management system. Data Mining and knowledge discovery in databases has several important application areas. Some application areas of database management system are banking, airlines, sales, finance etc. Database models are used to classify the database management system that they support.

1.2 Characteristics of Database

1. **Self-Describing Nature of a Database-** A database contains descriptions of data structure and Meta data along with the definition of database itself. This type of information is required by database management system if needed. This makes the DBMS completely different from traditional systems.
2. **Insulation between Program and Data-** In the traditional file based systems data file structures is viewed in the application programs because if the user wants any changes in the structure then all the application programs that access the file requires the changes as well. But the approach is different in database system, because the structure of data is stored in catalogues not in application programs..
3. **Support Multiple Views of Data-** A view is defined as a subset of database system that is defined for particular user. Different users have different views.
4. **Sharing of Data and Multiuser System-** A multiple database system at the same time allows multiple user access to the database. So the multiuser database systems have several strategies to ensure that the several users access the same data in a manner such that the data remains in integrated form.
5. **Control Data Redundancy-**In this DBMS approach every data item is kept only in one place. If any case redundancy occurs it is controlled to improve system performance and keeps it as minimum as possible.

- 6. Data Sharing-**From the huge amount of data, the integration of data from several sources requires the ability to produce large amount of information.
- 7. Enforcing Integrity Constraints-** Database management system should provide some constraints like data type, uniqueness etc to enforce integrity.
- 8. Restricting Unauthorized Access-**DBMS should provide authorized access to its users. All the users have not same set of privileges.
- 9. Data Independence-** Application programs and Meta data are not stored at the same positions. Both are stored separately. If any changes in the data structure occur then it must be handled by DBMS and these changes are not included in the program.
- 10. Transaction Processing-** The database management system should contains subsystems which ensure that many users who tries to update similar data item and performs these updates in the controlled manner.
- 11. Providing Multiple Views of Data-** A view is defined as a subset of database system that is defined for particular user. User should not have concern about where and how referred data is stored.
- 12. Providing Backup and Recovery Facilities-**When a computer aborts during a complex update process then it is the responsibility of system to make sure that the database should be restored to the stage before the process initialize executing.
- 13. Managing Information-**Managing information is basically used for taking care of the database for making it suitable to work for us and it can be useful for the work done by us.

Because of the databases are flexible enough every kind of project can be powered through them. So the database can link to:

- A site that can captures registered database users.
- An application by which clients can be tracked for various service organizations such as social service.
- It supports various medical record systems which provide health care facility.
- Ones personal memoranda in an email client.
- A series of documents of word processing.
- A system that issues reservation for airlines.

1.3 Data Warehouse

The concept of data extracted from various databases, operational systems and other resources for use as historical snapshots for schedule reporting and ad-hoc queries is known as data warehouse. Current and historical data which is used for creating reports for certain decision of higher management such as annual comparisons is stored in data warehouse. Basically in computing a system designed for reporting and data analysis is data warehouse. The characteristics of data warehouse are:

- **Subject Oriented-** Data mining is used for analysis of data. For example, to get information about monthly sale of your company, you can create a warehouse that contains the information of sales. Any type of question can be answered from this warehouse. This ability makes the data warehouse subject oriented.
- **Non-Volatile-** Non volatile means that there is no alteration in data once it enters into data warehouse. It is important because the concept of data warehouse enables us to analyze what has done.
- **Integration-** It is closely related to the concept of subject orientation of data warehouse. Data enters into the data warehouse from many sources. Data warehouse removes the problems of inconsistencies and naming conflicts. When this type of problems has been resolved it is said to be in integrated form.
- **Time Variant-** Analysts want huge amount of data in order to see the current trends in the business. Time variant means that the data warehouse should focus on the data over the time.

1.4 Data Mining

It is a process of finding patterns in data. Its goal is to find patterns that were previously unknown. Data mining is sometimes called Knowledge Discovery Process. Data mining is a powerful tool to find relevant information. Once you have right information, all you will need to do is apply it in the right manner. It is very easy to get information these days. But it is not so easy to obtain the relevant information that can help us to achieve a required goal. Hence the data mining becomes a powerful tool. It will give the power to predict certain behaviors within environment. It is used for a various kind of purposes in private as well as

in public sectors. The banking insurance and medicine industry uses data mining techniques for cutting down the costs and for increasing sales. For example- Both the insurance and banking sectors use data mining applications for risk assessment and detecting frauds. Data mining is also used by medical field for predicting the effectiveness of medicine. Data mining is used by various pharmaceutical firms to do research on treatments for various diseases. Data mining applications can be applied in the field of health care, immigration sectors and in business applications to solve specific problems. Various algorithms and techniques like clustering, classification ,decision trees, artificial intelligence, neural networks etc. are used for knowledge discovery from databases. But here we are going to explain clustering. The process of data mining includes analyzing data from several viewpoints and then summarizing into information which is helpful and is used to raise revenues and cut down the costs. Data mining software is one of the analytical tool that is used for analyzing data from so many different angles and summarizing the identified relationships. Converting data into knowledge helps in making quality decisions and process is known as knowledge discovery. Steps present in knowledge discovery are:

1. Data cleaning that includes removing the data that is inconsistent and noisy.
2. Data integration which involves integrating data from various data sources.
3. Data selection in which only task relevant data is extracted from the data warehouse.
4. Data transformation that involves converting the data into relevant forms.
5. Data mining step involves application of intelligent methods for extracting data patterns.
6. Pattern evaluation involves identifying the truly interesting patterns that represent knowledge.
7. Knowledge presentation step involves the usage of techniques of knowledge representation as well as of visualization for presenting the knowledge so mined to users.

There are numerous application of data mining —it has been utilized seriously and also widely by marketers, for direct marketing, cross-selling or up-selling; by financial

institutions, for credit scoring and fraud detection; by manufacturers in quality control, maintenance scheduling; and by retailers, for market segmentation and store layout.

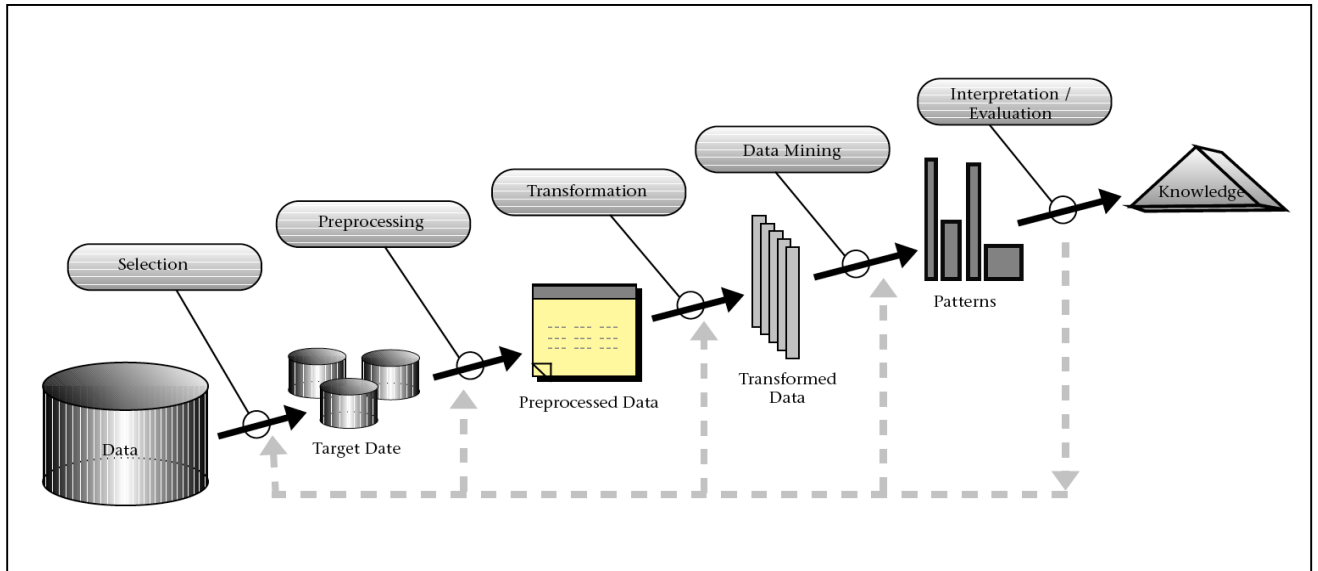


Figure 1.1: Data mining as a step in the process of knowledge discovery.

1.5 Techniques of Data Mining

The techniques of data mining are:

- 1. Clustering-** Clustering is the technique in which the objects of similar kind are grouped together into various classes which are termed as clusters. Various customer groups are discovered by the cluster analysis and the characteristics of each group is also analyzed. This is the common technique used for market analysis.
- 2. Association Analysis-** The analysis which shows the association rules discovery giving attribute value condition which are commonly occurred in a given dataset. The analysis is very much popular in transaction data analysis and market basket.
- 3. Classification-** The technique involved in predicting certain outcome based on any given input is known as classification. This approach involves certain processes of mining which are made to discover rules which are used to define the sub processes of the technique which bare model building and predicting. In this terms are belonged to class or particular subset of data.

4. Prediction-It is one of the data mining technique which gives relation between independent variables. It also defines relation between dependant variables and independent variables. For example if sales is taken as independent variable and profit as dependant variable then prediction technique is used for sales to predict the profit value for the future.

5. Sequential Pattern- The technique which identifies similar patterns based upon transactional data. It is one of the special cases of structured data mining.

6. Decision Tree- Decision tree basically us the hierarchal model of decisions and their consequences. The structure of decision tree includes branch, root node and leaf node. Attributes test is denoted on each interval node, the test outcome is denoted by branch and class labels are shown by leaf node. The uppermost node of the tree is known as the root node. The tree learning is done by dividing the source into set which are generally based on a test of attribute value. The top down approach of decision tree sets an example of greedy algorithm. Apart from this bottom-up approach is also common these days. Definition of decision trees can also be on the basis of combination of computational and mathematical techniques for getting the categorization, description and generalization of a given dataset.

There are mainly two types of data trees used in data mining.

1. Classification tree analysis- It is done when the class to which data depends in the predicted outcome.
2. Regression tree analysis-It is done when a real number can be taken as the predicted outcome example (The cost of a building)

1.6 Types of Data in data mining techniques:

Techniques of data mining can be applied to following different types of data:

1. Relational Database:

It contains form of relations known as tables that contain tuples (rows) and attributes (columns) that contains data.

2. Transactional Database:

It has data in transaction form. Each transaction is basically a record which has its own unique transaction id. Each transaction has a transaction id and list of items that are purchased in that transaction.

3. Spatial Database:

This includes data of maps. It stores data representing objects in geometric space.

4. Temporal and Time-Series Database:

Temporal database includes data such as stock exchange data that changes rapidly and time-series database includes biological sequence data, heartbeat of patients, etc.

5. World-Wide Web:

It includes hypertext, audio, video, text data. It is a widely distributed repository of information that is made available by internet.

1.7 Cluster Analysis

Cluster analysis organizes data into groups that is meaningful. To maintain the natural structure of data, resulting clusters should be meaningful. Clustering is a term used to define the grouping of similar data items or objects. Items or objects which are similar in nature are grouped in one cluster and items or objects which are dissimilar in nature are far away are grouped together in another cluster. Clustering is used for analysis of data and also resolves the classification problem. The process of clustering is shown in figure 1.2. In this process of clustering raw data is transformed into appropriate clustering algorithm. After applying appropriate algorithm cluster of data is obtained. The main objective of the clustering technique is that the objects resides in one group should be close to one another For example cluster analysis has been used to identify groups of books according to their topic, author, and their core areas.

By using Classification we can distinguish groups or classes of objects. But the general question faced by the researchers is that how to classify the data into meaningful clusters. An

object belonging to same groups have maximal degree of association and remain minimal if they belong to different groups.

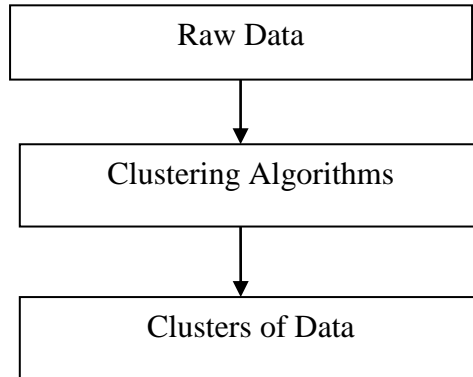


Figure 1.2: Process of clustering.

Clustering is the demanding field in the research. Different requirements of the clustering are:

- Scalability
- Have the Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shapes
- High dimensionality
- Incremental clustering
- Having the ability to deal with the noisy data

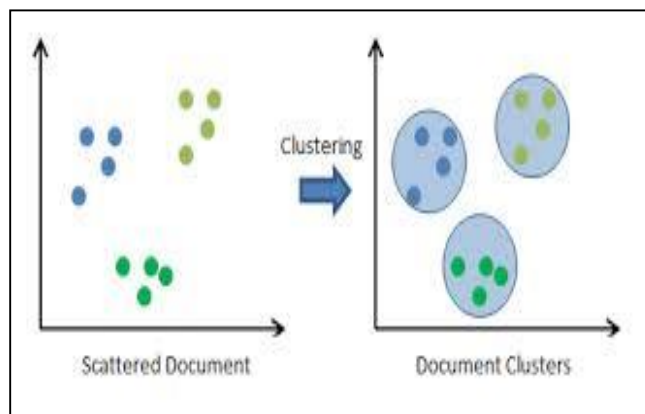


Figure 1.3: Document clusters are formed from scattered documents.

Cluster analysis is an important activity undertaken by humans. Using clustering technique, one is able to classify between cats, rats and various animals. In the machine learning, clustering is very suitable example of unsupervised learning. So we can use clustering used to learn from observations than learning by examples. Several attempts have been taken to find methods for efficient and effective cluster analysis in large number of databases. Various clustering models are as follows:

1. **Connectivity Models:** These are also known as hierarchical clustering. These types of models make clusters based upon their distance. These are classified by the way which the distance is calculated. In addition to this, user also knows about linkage criteria means that the clusters have multiple objects to calculate the distance.
2. **Centroid Models:** k-mean clustering is one of the examples of centroid model or centroid clustering. In this type of model clusters are defined by the mean vector. For k-mean clustering the number of clusters is known in advance.
3. **Distribution models:** In this type of model, clustering can be defined as the objects belong to same distribution. This type of model suffers with the problem known as over fitting. Distribution models give complex clustering models.
4. **Density Models:** In this type of clustering clusters are defined on the basis of the higher density. For this purpose noise points and border points are defined. A noise point is not a border point.
5. **Group models:** These types of models just provide grouping information. These algorithms do not provide any kind of model for clustering.
6. **Graph Base Models:** A Graph base model connects each two nodes by an edge. It is also called clique.

Clustering can be hard or soft. Hard clustering means whether the object belongs to cluster or not. It may also lies outside the cluster. In soft clustering each object has a tendency by which they belong to each other up to a certain extent.

1.8 Example of clustering

Consider a library system in which the books related to a huge amount of topics are available. The books are always arranged in the manner that forms the clusters group. The books which are similar in nature are placed in one cluster group and the books which do not have any kind of similarity are placed in another cluster group. For example, the operating system books are placed in one shelf and the Networking books are placed in other shelf, and so on. The complexity can be further minimized by keeping the books which covers same type of topics are placed in that same shelf and then these shelves are given a specified name. Whenever a person wants a book of a particular topic, the user will go to that particular shelf and take the book from that shelf only rather checking in the whole library.

1.9 K-means clustering

K-means clustering comes under centroid models of clustering algorithms. The k-means algorithm comes under the family of algorithms called as optimization algorithms of clustering. That is, the examples are divided into clusters groups in this way that the cluster gives good optimal results according to criteria defined. The name of the algorithm has been derived such that the k clusters are formed from the data set where the cluster centre is the arithmetic mean of all objects within that type of cluster. The number of the clusters k is known in advance. The first step is to find the initial centroids for each cluster. The next step is to associate each data object to its nearest centroid. Early grouping is done by assigning each data object to centroid which is so close to it and the first iteration is completed. The algorithm works in iterations until the objects do not change their cluster centre. Centroids move their positions until the convergence criteria have reached. Pseudocode for algorithm is as follows:

Algorithm 1: The k-means clustering algorithm

Input: Dataset objects containing $D = \{d_1, d_2, d_3, \dots, d_n\}$

- Number of clusters k

Output:

- A set of k clusters obtaining from dataset objects

Steps:

1. Randomly choose k objects from the dataset containing n objects as the initial centroids.
2. Repeat
 - a. Assign each object to its nearest centroid;
 - b. For each cluster, new mean is calculated;

Until convergence point is reached.

The k-mean when compare to SOM has a disadvantage that it cannot perform vector quantization, which means naturally it, is not in a form that can be easily visualized. K-mean has an advantage over SOM is that it is more computationally efficient.

1.10 K-Means Algorithm Properties

- K-Means algorithm always contains k-clusters.
- At least one data item is always present in each cluster.
- In this the clusters do not overlap and they are non-hierarchical in nature.
- In a cluster each member is close to its cluster than another cluster as the closeness not always involve center of clusters.

1.11 Advantages and Disadvantages of K-mean Clustering

Advantages:

1. If we keep k smalls and variables are large, then many times k-mean is faster than the hierarchical clustering.
2. If the clusters are globalised in nature, k-means produce tighter clusters than hierarchical clustering.
3. It will best results when the data set are distinct.
4. It gives good results than classification.
5. It is fast and easy to understand.

Disadvantages:

1. It is difficult to predict k-Value.
2. It did not work well with the globalised cluster.

3. If initial partitions are different then it will produce different final clusters.
4. Different size and Different density based clusters cannot perform well with the K-mean algorithm.
5. It is applicable only when mean is defined.
6. It does not handle noisy data and outliers.
7. Algorithm is not applicable for non-linear data set.

Chapter 2

LITERATURE REVIEW

Pooja Mittal et al (2014): has done comparative study on the role of data mining techniques in education. This paper highlights various data mining techniques which can help in education environment to improve the performance of existing data and helps to create new data. Data mining techniques like clustering, decision tree, classification, prediction and many others can be used in every field of science like medical, agriculture, marketing etc. This paper represents various classification and clustering techniques which can help in education sector to improve the growth of education sector. Clustering is a term used to define the grouping of similar data items or objects. Items or objects which are similar in nature are grouped in one cluster and items or objects which are dissimilar in nature are far away are grouped together in another cluster. This paper applies clustering techniques on the group of some schools to find which are similar in nature, and which differs from each other and in which aspect. Clustering techniques are also applied to group the students based on their behavior. Each technique is associated with different algorithms. Clustering algorithms are partition based, hierarchical based, grid based etc. Second technique classification is used to find a model that describes how to classify test data. It can be type of both unsupervised learning and supervised learning. Different classification techniques are decision tree, associations rule mining, neural networks etc. This paper examines various techniques to be used in data mining and also in the education section.

Amandeep Kaur Mann et al (2013): has done survey on various clustering techniques. This paper highlights different clustering algorithms to be used in data mining. Each technique is associated with different types of algorithms. Various clustering algorithms are Partitioning based or centroid based, density based, grid based and hierarchical based algorithms. Portioning based algorithms divides or partitions the data points into number of partitions i.e. number of k-partitions and that number of partitions represents number of clusters. Different partition based clustering algorithms are: k-mean, CLARA, Medoid based and PAM. K-mean

is the popular partition based clustering. Hierarchical based algorithms are viewed as the series of steps that forms the hierarchy of the clusters. Hierarchical algorithms can be divisive analysis or DIANA and agglomerative Nesting or AGNES. Density based algorithm determine the cluster that grow with the high density. Grid based clustering algorithms deals with the space that covers the data points. The problem arises in these algorithms is how to choose grid size. Different grid base algorithms are sting, clique, wave and cluster. Grid based algorithm differs from other algorithm in terms of fast processing time. The main theme of this paper was to give basic understanding of types of clustering algorithm used in data mining context.

Saurabh Shah et al (2012): has done comparative study on clustering algorithm to make the algorithm better in terms of number of clusters and execution time comparisons with k-Mean and k-Medoid algorithms. Proposed algorithms make the use of real data set and results are compared with both these algorithms which show that the proposed algorithm takes less time in computation and gives better performance as compared to k-mean and k-medoid algorithms. The proposed algorithm worked as: From original dataset, draw some sub samples. On these extracted subsamples, apply k-mean algorithm and then compute the results. To choose initial centroid choose minimum of minimum distance and Apply k-mean algorithm again on dataset. Then Join two clusters into one cluster and find the new cluster centre until the clusters are reduced to sample. Performance of proposed algorithm is highly focused on the choice of initial cluster centre taken in each step. Comparative studies show that if number of clusters is less, only then k-mean and k-medoid shows better results. If cluster size is increased, then k-medoid takes less computation time than k-mean algorithm. The proposed algorithm given by this literature takes less execution time than both the k-mean and k-medoid algorithms even if the number of cluster size increases.

Malay K. Pakhira (2009): This paper eliminates the problem of empty clusters due to bad initialization and provides the modified view of the k-mean algorithm. This empty cluster problem leads to performance degradation of the k-mean algorithm and also may produces

anomalous behavior. The proposed algorithm is almost similar to the original algorithm, but there is no performance degradation due to any modifications. The basic algorithm starts by assigning data to the centre based on the minimum distance and then following the computation of new clusters centre. The process stops when cluster centre become stable. The proposed algorithm starts set of the old centre vectors and with the distribution of data elements among the clusters by selecting the Euclidean distance and new clusters are generated by averaging the data elements. The execution steps of the proposed algorithm are similar to the basic k-mean algorithm. Experiments are performed under normal and extreme initial conditions. Normal condition produces cluster centre randomly and in extreme condition the centre vectors are identical. If random method is used then it does not produces empty clusters even if the number of clusters varies from 2 to 10. But when clusters are initialized to some value then empty clusters are not formed until it has value ranges between 2 to 6. As the number of clusters varies than specified range, empty clusters may form. It can say that modified algorithm handles the empty cluster problem very efficiently.

Raed T. Aldahdooh *et al* (2013): This paper describes the method to identify initial clustering centroid of k-mean clustering. A previous approach uses random selection method for identifying initial centroid. This paper enhances the performance of initialization method over many datasets by taking into consideration different observations, number of clusters, groups and clusters complexity. The experimental results shows that the proposed initialization methods lead to better clustering results than random method and are very effective. The paper works on following strategy to find out the initial centroid: Firstly, select the initial centroid by using random method. After selecting initial centroid, some calculations are performed to determine the initial centroid and the points which are closest to the centroid. The choice of the calculations performed is based on finding the Euclidian distance or some other points found on the dataset. To find the number of nearest points of the initial centroid divide the total no. of objects used in the given dataset by the total number of the clusters given by the user. If the first selected centroid points contain the noise, then it was ignored. Another point is selected until the first centroid point is not found.

Jyoti Agarwal et al (2013): has performed crime analysis using k-mean clustering using Rapid miner tool. The main objective of doing crime analysis involved in this paper is to predict the crime causes based on the existing crime data and apply data mining techniques in an efficient manner. The tool used in this paper mine the dataset according to user requirement and applies k-mean clustering to compute the distance matrix. After then crime analysis is applied on the resultant cluster. From the clustered result, it is observed that the crime rate decreases over the years and this approach is useful for finding the new precautions method for future.

K. A. Abdul Nazeer et al (2009): This paper works on the k-mean algorithm to improve the accuracy and efficiency of simple k-mean algorithm. In order to make cluster less complex, this paper represents a method and making a cluster more efficient and accurate. The modified approach works in phases: In the first phase, in order to make cluster accurate initial centroid are determined systematically rather than randomly. The output is given to the next phase which redistributes the entire process. The second phase starts by calculating the relative distance for each data points from the initial centroid, thereby forming the initial clusters. Next phase is the iterative phase which uses heuristic approach. This saves the time to assign the data points to initial centroid by calculating the nearest distance, therefore enhances the efficiency of the algorithm. The limitation of this approach is finding the number of clusters and given it as input. The future scope of this paper is to find some statistical methods to compute the value of k.

M. Goyal et al (2014): has done study on k-mean algorithm and presents a improved mean based algorithm, which is easy to implement. This paper focuses on making k-mean algorithm globally optimum. For achieving this, initial centroids must be selected carefully. The paper described an improved algorithm to determine initial centroids, which results in better clusters equally for uniform and non-uniform data sets. The proposed strategy in this paper is as follows: From each data point in the dataset, measures the distance from the origin by using Euclidian distance. Then sort the original data points based on distance

calculated in above step and divide the sorted data into equal number of partitions. For each partition find the mean of data points. These means will be selected as initial centroids. Both the midpoint based and the proposed algorithm was applied on different datasets and observed that the algorithm would produce better clustering results. To check the effectiveness of the proposed algorithm firstly, the real dataset of compressor Chiller machines used in Chiller industry is taken, secondly the real dataset of internal marks obtained by the students of a class in a semester in engineering have been considered and finally a dataset of hourly temperature of Delhi for year 2004 has been studied. The future scope of this paper is to use mixed type of data, as k-mean algorithm can be applied only on numerical data only. But in day to day life, scenarios with mixed type of data set has also encountered.

Sapna Jain et al (2010): provides the comprehensive view of using clustering data mining techniques to be used in weka 3.7 interface. Weka is available freely on the internet and has many powerful features in the context of data mining. This literature review provides the history of weka workbench and throws a light of using K-mean clustering execution in weka workbench. Weka has many features like explorer, experimenter, knowledge flow, and simple CLI. This paper shows the implementation of k-mean clustering used in weka to cluster the customers in the bank data set. The weka interface along with the data set work as follows: The dataset is used to be taken in .csv format, as it uses bank based dataset. So bank-data.csv file is used in this technique and algorithm to be used in weka automatically normalizes numerical attributes when distance is calculated. Euclidian distance measure is used to find the distance of clusters. From the weka explorer tab, select the cluster. Available clustering algorithms are shown there. Select the desired algorithm from the window. Enter the value in the number of the cluster and do not fill in the seed value. In the cluster model panel, start the training set. From the result list panel window, check the results. Weka interface supports more algorithms of clustering than association rule mining. It provides clustering algorithms to be executed in java.

Azhar Rauf et al (2012): provides the enhanced k-mean algorithm which is helpful in reducing the number of iterations and time complexity. The paper proposes a new method of calculating the initial centroid rather than choosing the random selection method. By choosing this the number of iterations is reduced and also the time complexity is improved. The algorithm proposed by this paper has worked in two phases. In the first phase size of the clusters is fixed and initial clusters are formed by output of first phase. During the second phase the size of the cluster is not fixed, it varies and as the result of the second phase final clusters are formed. The output of the first phase is given as the input to the second phase. The whole process continues until cluster reaches the convergence points. During the second phase the data objects change their positions to find the appropriate cluster. The number of data elements chosen is 1000. The algorithm proposed by this paper is more efficient than the existing one because it improves the quality of cluster. This enhanced algorithm uses any type of integer data. The future scope of this paper is that it can be tested on text based clustering.

H.S. Behera et al (2012): has given a hybridized k-mean clustering algorithm which is based on outlier detection technique. This approach is used for effective data mining. This paper presents a technique that uses k-mean clustering algorithm and outlier finding technique for the detection of outliers. Outliers detection technique is divided into three categories, these are: density based outlier detection, distance based outlier detection and third is outlier finding technique. Outlier detection is one of the major issues in data mining. The effectiveness of the proposed algorithm is reviewed by taking 2 dimensional data set. After then high dimensional data set is chosen. The paper proposes a new approach that can improve the k-mean clustering algorithm over the noisy data and this approach also used to detect the outliers.

Narendra Sharma et al (2012): has done comparison through weka tool on various clustering algorithms. Weka has many inbuilt facilities like knowledge flow, explorer, simple CLI and experimenter. In order to do the comparison this paper chooses past project datasets from the repositories. The dataset is loaded in the weka in .arff file format in order to perform

cluster analysis. In case if dataset is not in .arff file format then first it needs to be converted. Weka gives good clustering results without having deep knowledge of data mining techniques. Farthest first algorithms, Optics, k-mean, EM clustering algorithms are compared with weka tool. According to this paper each clustering algorithms have own advantages and disadvantages.

Ritu Sharma et al (2012): has worked on k-mean clustering in spatial data mining. The tool used is the weka interface. The dataset used in this paper is the crop yield records that are taken from the agriculture website of India. Association rule mining is used for finding relationship between the items. This paper extracts the patterns from the spatial database and the spatial association mining is used, so the process is quite costly. This paper also highlights some weaknesses of the k-mean clustering and gives possible solutions of k-means clustering.

Chapter 3

PRESENT WORK

Present work is mainly focus on developing an enhanced k-mean clustering algorithm. This approach takes the advantage of clustering technique. The traffic dataset is taken and is used for checking the error of the enhanced k- mean algorithm. The same dataset is used for the enhanced k-mean and the existing algorithm. In order to do the analysis, dataset is loaded into the tool where it reads the file format. After then clustering algorithm is applied and traffic analysis is performed.

3.1 Problem Formulation

This research proposes a technology based on data mining algorithms for the induction of clustering. It is well suited in our context for various reasons.

1. Enhance the accuracy with a new Algorithm of clustering.
2. To Bring advantage of clustering approach in traffic analysis
3. Reduce error rate as minimum as possible.
4. Analysis of various factors of traffic on NH-1 highway that are the main reason for accidents.

3.2 Scope

The scope of the study is to bring clustering technique in traffic management system. In this clustering technique, k-mean algorithm is enhanced to improve the accidental analysis. In this study, attempt is made to reduce the error of the existing algorithm. The basic idea is to improve the performance of existing algorithm and to improve the traffic safety. This will benefit to the society and reduces the accident rate, which also help the traffic management to take effective decisions.

3.3 Objectives

The objective of this research focuses on the data analysis for important attributes of accidents on highways by implementing a clustering model. Thus, the problems taken for this research work is divided into some objectives which are as follows:

1. The accuracy is increased by comparing self developed program with the existing algorithm.
2. The objective is to show the advantage of clustering approach for examining the accidental analysis.
3. The error rate is reduced.
4. The objective of this project is to preprocess highway accidental data.

3.4 Enhanced Algorithm

The original k-mean algorithm starts by choosing the initial centroid randomly and works in three iterations. The number of clusters used in this algorithm is two. The next step is to associate each object to its nearest centroid. The process is iterated until it reaches to a convergence point. But the algorithm proposed by this paper works as follows:

Algorithm 2: The enhanced k-mean algorithm

Steps:

1. Choose arbitrary point p from a dataset.
2. Find all the points which lie in the neighborhood of point p.
3. All the points which lie in the same area make a cluster.
4. Continues the whole process until all the remaining points have been processed.
5. At each iteration calculate the standard deviation and probability of each cluster formed.

Until cluster don't reached its convergence point.

In order to calculate standard deviation of data objects lies in the cluster, calculate the variance of the dataset. The variance is calculated by using the following formula:

$$\frac{\sum(x - \mu)^2}{n}$$

Where μ represents the mean and n is number of items. By taking the square root of the variance gives the standard deviation. It is denoted by σ notation. It is the square root of the variance and calculated by the formula as:

$$\sigma = \frac{\sqrt{\sum(x - \mu)^2}}{n}$$

3.5 Research Methodology

Solving the research problem systematically is known as research methodology. It is scientific way of studying how research is done. Various steps that that are followed by the researcher while carrying out the research are studied and also the logic behind using those steps are analyzed. The researcher must know not the research methods/techniques as well as the methodology. Algorithms, procedures that are adopted to solve the research problem are explained. The procedure is described as: Firstly, we take traffic dataset required.

1. The required dataset is filtered according to the user requirement.
2. After filtering, the dataset file is read by the desired tool.
3. After reading the dataset file, apply the clustering algorithm on dataset objects.
4. After then, merge the closets clusters lies in same area.
5. If there is a single cluster then the algorithm ends and if there is more than one cluster the closets clusters will be merged.
6. If the number of clusters is more than one then calculate the mean/standard deviation is calculated.
7. After then calculate the probability of the cluster formed. At last, the traffic analysis on resultant cluster is performed.

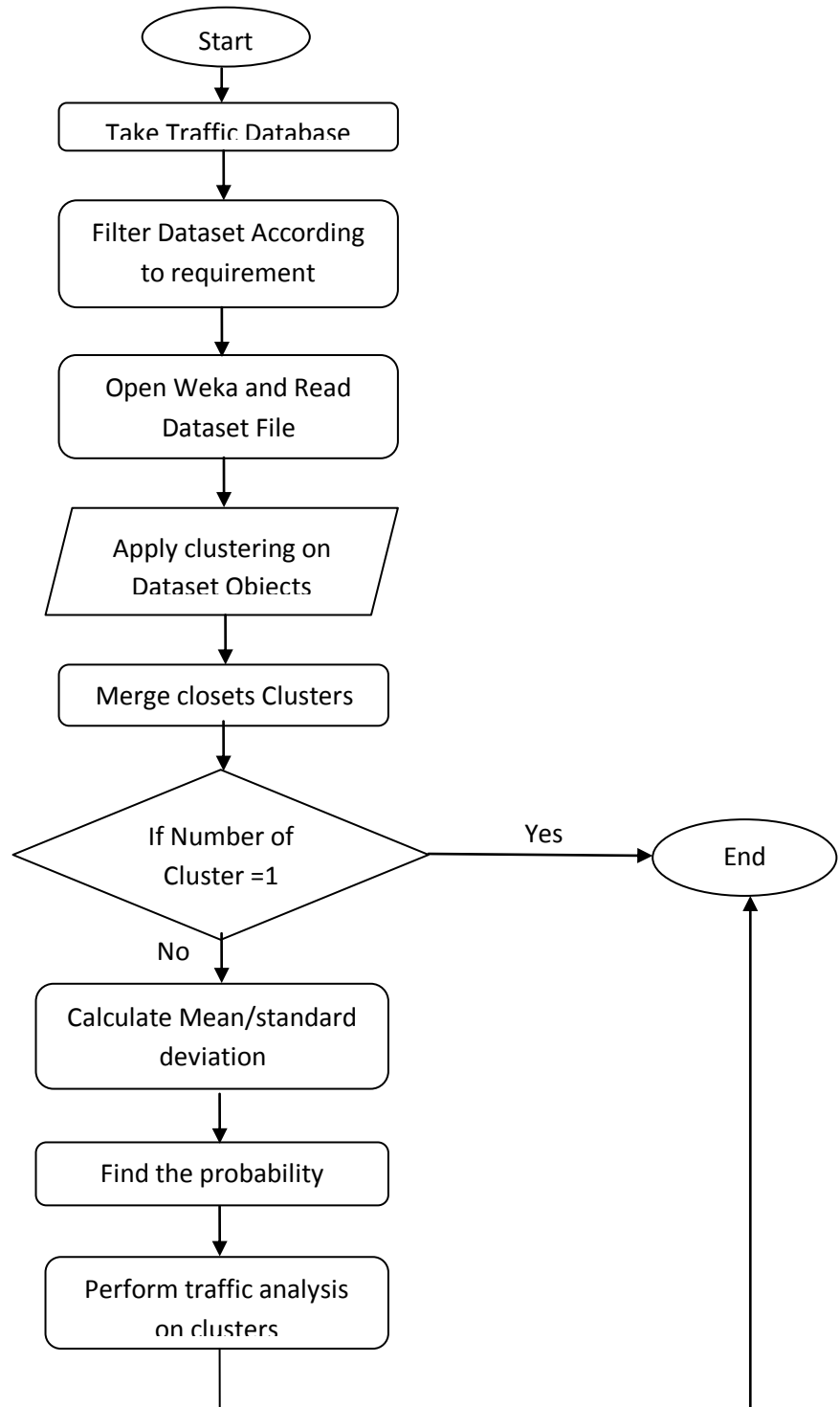


Figure 3.1: Flow chart of research.

3.6 Tool Used- JAVA is a high level computer programming language which is based on classes, object-oriented and also concurrent. Java can be used to write many computer applications that can store data, and also many more things which normal computer software can do. Java is platform independent which means a user or programmer can run on any other machine and that code does not need to be recompiled to run on another. It is developed by Sun Microsystems of USA in 1991. The syntax of java is somewhat similar to C and C++. The applications which are built in java are firstly converted to byte code or class file that are portable which means these applications can be run on any JVM (Java Virtual Machine). It is one of the most accepted programming languages which is mainly developed for client server web applications. There are compilers which are available for byte code like Ruby, Python, and Ada etc. Java is utilized within large types of the platforms ranging from low level devices like mobile phones and embedded devices to large scale super computers. The java applets are used for secure browsing on World Wide Web. There are some constructs which are at low level such as pointers are removed from java and the memory model is very simple in which object is allocated on the heap and references are all variables of object types. Features of Java are as follows:

1. Object-Oriented: Java is an object-oriented language. The program code and the data all are in the form of objects and classes. Java is available with large set of classes which are managed in the form of packages and they can be used in the program by inheritance. The model of object is problem free and it can be easily extended.

2. Robust and secure: Java is a powerful language and also provides securities which can be used to make trustworthy code. It also includes the exception handling concept which minimizes the problem of system crashing. Java provides strong security and due to this reason programmer uses this language for browsing and internet programming. The lack of pointers in java gives surety that programs cannot get right of going into memory location without proper confirmation.

3. Compiled and Interpreted: Usually any other language is either interpreted or compiled but both these are present together in java which makes the java two-stage system. Java

compiler makes an interpretation of Java code to byte code instructions and java interpreter produce the machine code that can be directly executed by machine which is running the program.

4. Portable and Platform Independent: Java provides the powerful feature of portability which means that the programs of java can be run on any machine irrespective of the computer architecture. If there is some changes or upgrades in hardware, operating system and software, it will not force the alterations in the programs implemented in java. It is the only reason that java is most accepted language for programming which also helps in interconnecting different people world-wide. Java provides portability in two ways. Firstly java compiler creates the byte code which can be run on any other JVM irrespective of the architecture. Secondly primitive data type sizes are machine independent.

5. Small and simple: Java is a very small and simple language. It does not use header files and pointers, goto statements. Multiple inheritance and operator overloading is excluded from java.

6. Interactive and Multithreaded: Java supports multithreading which means handling multiple tasks at the same time. It also preserves multithreaded programming. This means we do not need to wait for one application to complete to run some another application. This feature is very helpful in computer graphics.

7. High Performance: Java provides high performance. Its performance is extraordinary for interpreted the language only because it uses intermediate byte code. Architecture of java is designed to reduce the overheads which appear during runtime. The fusion of multithreading enhances the execution speed of program.

8. Extensible and Dynamic: Java is a dynamic language. It performs dynamic linking in new classes, objects and libraries. Through the query building, java also creates the class type. Because of query building feature, it became also possible to abort the program or link dynamically which depends on the reply. Java program supports functions written in other language such as C and C++, which are known as native methods.

Weka: It is known as a collection of several machine learning algorithms and is used for performing various tasks of data mining. The algorithms can be run in weka either directly to a dataset or invoke through java code. It is also suitable for the emergence of new techniques of machine learning. Weka GUI is one of the best features of the weka tool. When weka GUI is launch then four options are available. Weka GUI is the power of the weka tool. These are:

1. **Simple CLI** -It provides commands to be executed directly and is simple command line interface.
2. **Explorer** -It is other good option provided by weka GUI. It is used for exploration of data.
3. **Experimenter** –It provides environment for conducting experiments and also used for performing different statistical test among learning schemes.
4. **Knowledge Flow**- It is an interface of java beans and used for preparing and running different experiments.

Weka knowledge explorer: It has six panels which are as follows:

Preprocess: - It is the starting point of the weka explorer. This panel is used for loading of datasets, displaying the various characteristics of different attributes and applying data to combination of unsupervised filters.

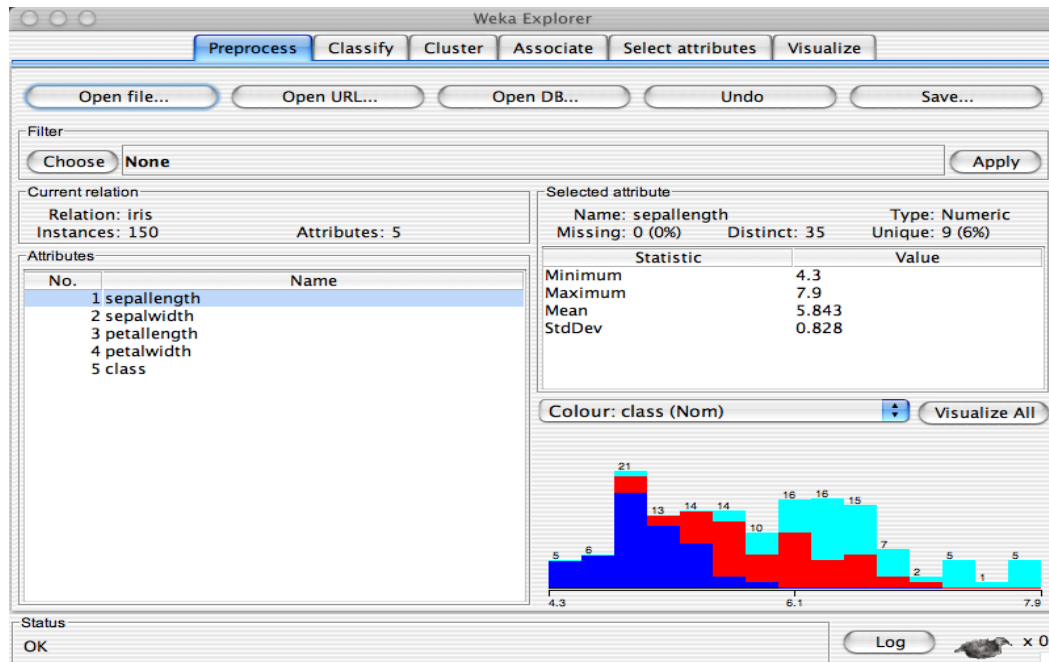


Figure 3.2: Pre-process panel of weka explorer.

Classify: - This panel is used for the configuration and execution of weka classifier on the given dataset. It allows us to conduct test on different dataset. The errors are shown in visualization tool. If it produces a decision tree it is displayed in tree visualize. The percentage can also check based upon the entered data. It is used for predicting nominal and numerical data. Different test options are available to run the algorithm. These are training set, cross validation, supplied test and percentage split. The results are checked by the result list option on the left of classifier window.

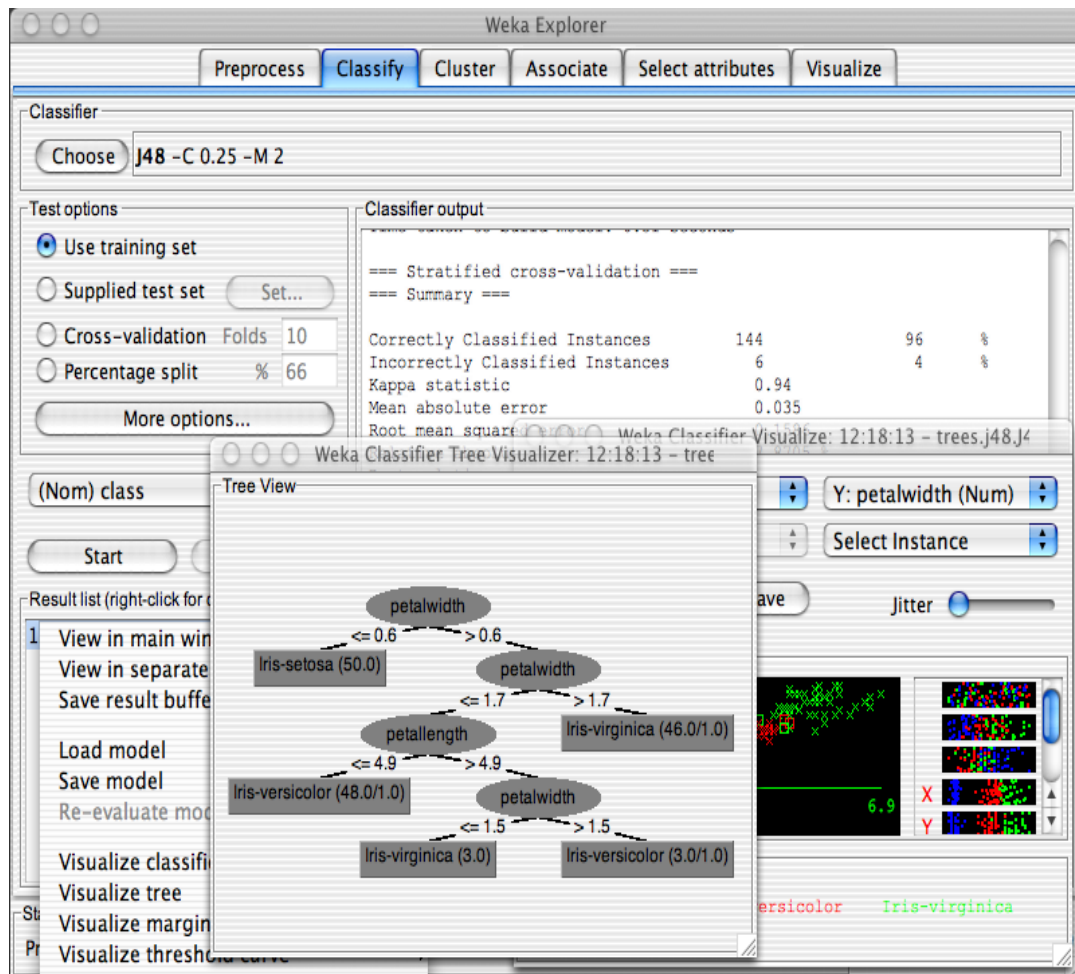


Figure 3.3: Weka classify panel of weka explorer.

Cluster: - Weka contains dataset in the form of clusters of similar data. This panel is used for the configuration and execution of weka clustered of current dataset. The errors are shown in visualization tool. The file is saved using .arff extension and analysis is done using any

clustering scheme. The result can be seen by choosing the result list option. The algorithm is run by using start option.

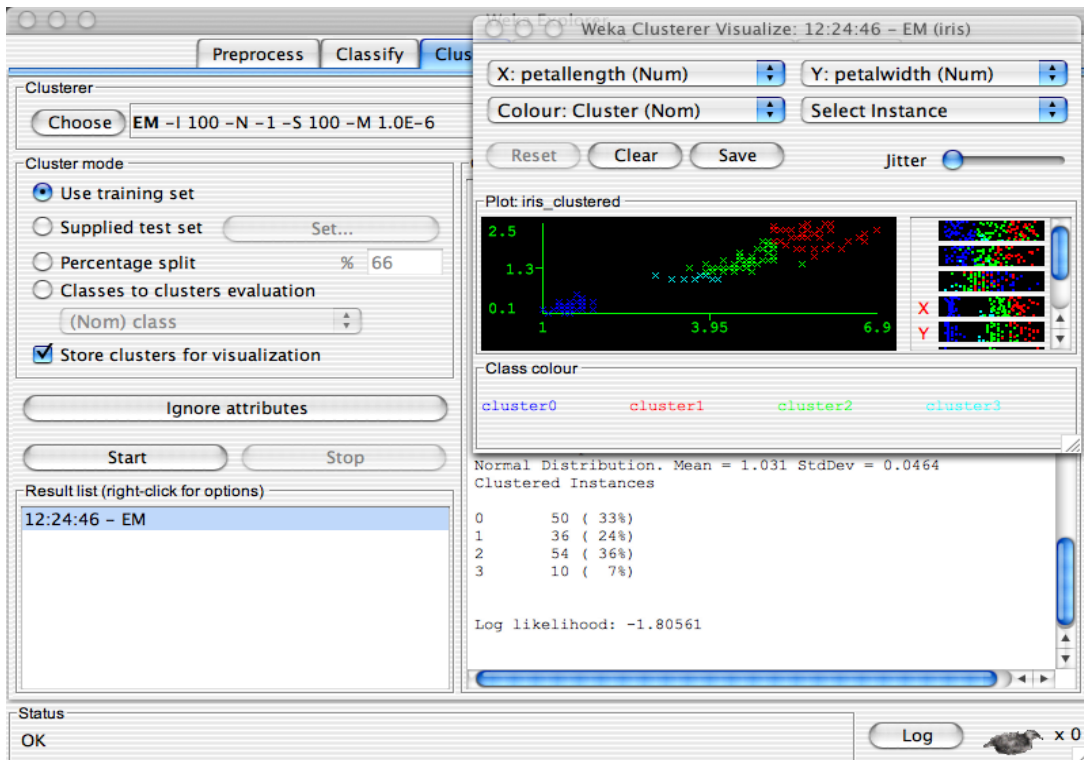


Figure 3.4: Clustered window panel of weka explorer.

Associate: - This panel allows mining the dataset with the help of weka association. Association rules are used for this purpose. Numerical values are not handled by this scheme. Only nominal values are used. The association mining algorithm is run by start option at the bottom.

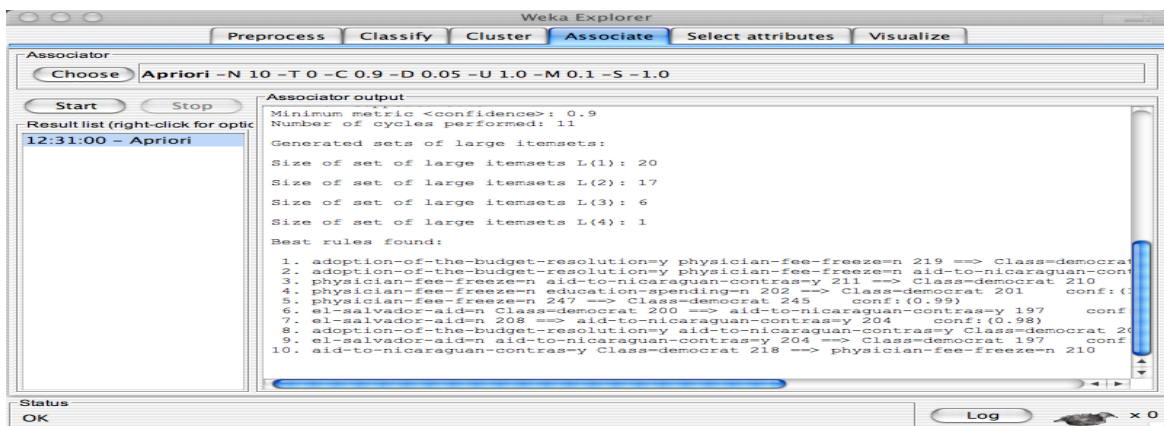


Figure 3.5: Associate window panel of weka explorer.

Select attributes: - This panel is used for searching all the possible combination of attributes of data objects lies in dataset. It finds the best attributes for prediction. It has two parts that is search method and evaluation method. There are two options available under this scheme for the selection purpose. Results are visualized by the visualization window.

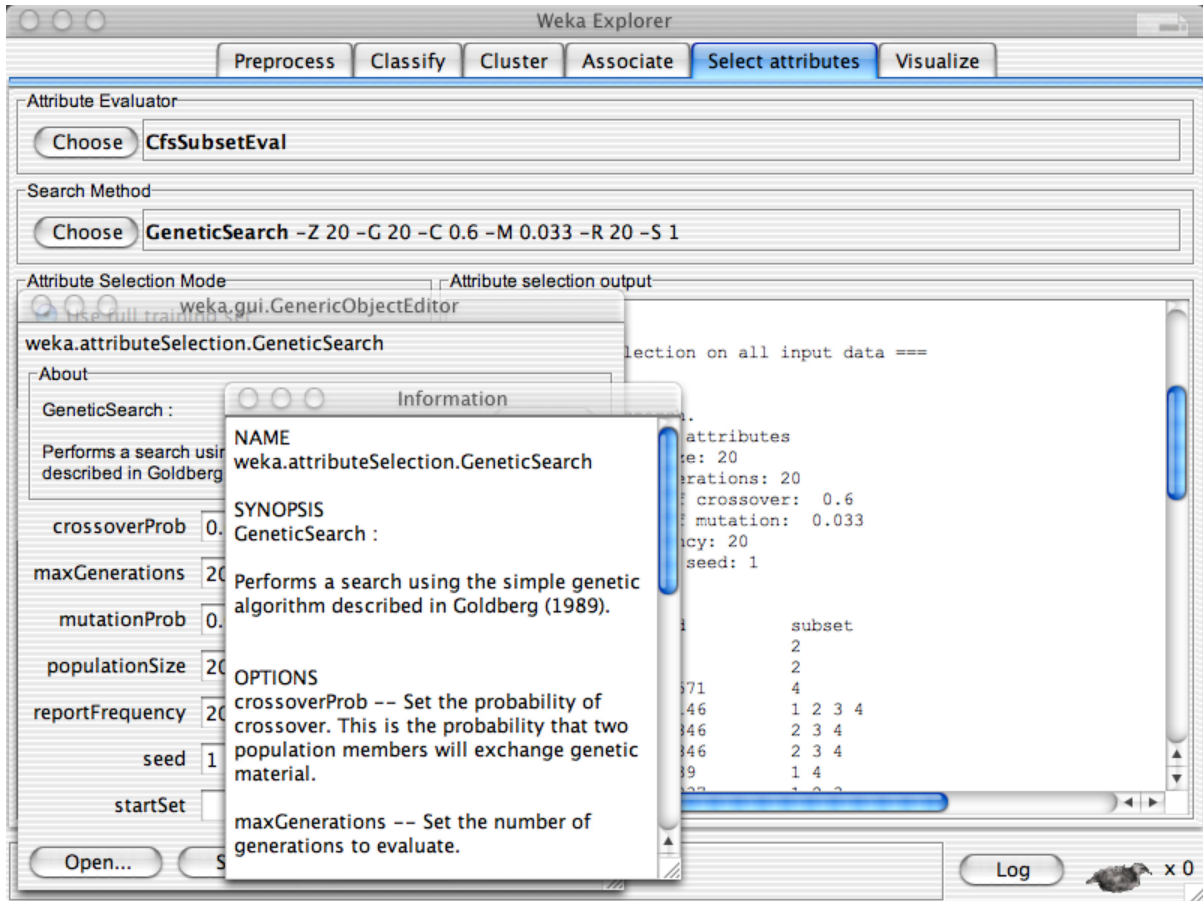


Figure 3.6: Select attributes window panel of weka explorer.

Visualize Panel: With the help of this window scatter plot matrix is displayed. Using the slider control panel number of cells and size of points can be adjusted. The number of the cells can be changed by using select attributes option. Then the attributes are displayed. If dataset is huge then performance is enhanced by showing subset of attributes of the given dataset. View of the cell is displayed by clicking that cell in the matrix. This panel is helpful in showing two dimensional views and three dimensional views of the desired dataset. If the attribute is discrete, then each value is shown by different color and if the attribute is continuous, a spectrum is formed which is used to indicate the value. Summary of the

attributes are shown by the attribute bars which is located at the right hand side of the panel window.

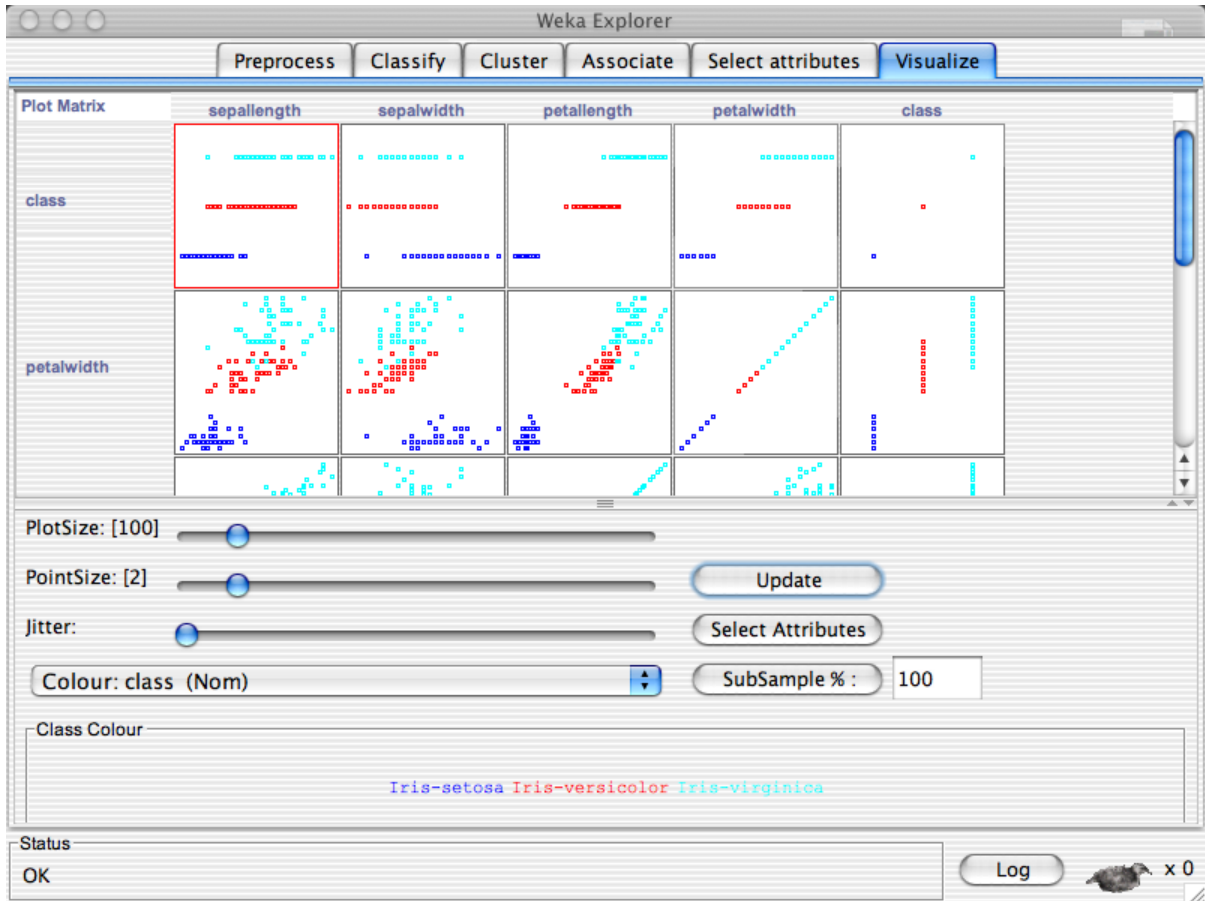


Figure 3.7: Visualize panel of weka explorer.

Chapter 4

RESULTS AND DISCUSSION

To test the performance of the enhanced algorithm traffic dataset is taken. For effective performance the experiment result of the enhanced algorithm is compared with the existing k-mean algorithm. The number of cluster is 6 in case of enhanced k-mean algorithm. But in the existing k-mean the number of cluster is 2. Each cluster contains different clustered instances. The total number of instances of dataset is 866. Some of the objects of traffic dataset is as shown in figure 4.1:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Attention	Drink	Physical	Inexperie	Rule	Mistake	Speed	Vehicle	Weather	Road	Sight	AGE	Accident
2	YES	NO	YOUNG	5	FOLLOWE	NO	100	CAR	DRY	GOOD	GOOD	25	NO
3	NO	YES	YOUNG	9	NOT FOLL	YES	90	CAR	DRY	AVERAGE	GOOD	30	YES
4	YES	NO	OLD	19	FOLLOWE	NO	90	CAR	rainy	GOOD	GOOD	80	NO
5	YES	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
6	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
7	NO	YES	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
8	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
9	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
10	NO	yes	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
11	YES	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
12	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
13	YES	YES	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
14	NO	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
15	YES	YES	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
16	NO	NO	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
17	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
18	NO	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
19	YES	YES	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
20	NO	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
21	YES	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
22	YES	yes	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
23	YES	NO	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
24	NO	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
25	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO

Figure 4.1: Traffic dataset attributes.

The dataset contains thirteen attributes. Different attributes are shown in the above figure.

These are:

- Attention
- Drink
- Physical
- Inexperience

- Rule
- Mistake
- Speed
- Vehicle
- Weather
- Road
- Sight
- Age
- Accident

And some of the instances of the dataset is shown in figure 4.2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
482	NO	NO	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
483	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
484	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
485	YES	YES	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
486	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
487	NO	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
488	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
489	YES	YES	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
490	YES	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
491	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
492	NO	NO	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
493	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
494	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
495	YES	YES	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
496	NO	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
497	YES	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
498	YES	YES	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
499	NO	NO	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
500	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
501	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
502	NO	YES	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
503	NO	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
504	YES	NO	YOUNG	9	NOT FOLL	YES	90	CAR	DRY	AVERAGE	GOOD	30	YES

Figure 4.2: Some traffic dataset attributes.

The output of the enhanced algorithm is displayed using the software package known as weka. Weka is freely available on the internet. Weka GUI has following facilities like:

- Simple CLI
- Explorer

- Experimenter
- Knowledge Flow

Weka explorer has many other options available like classify, cluster, associate, preprocess, select attribute and visualize.

	A	B	C	D	E	F	G	H	I	J	K	L	M
842	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
843	YES	YES	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
844	NO	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
845	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
846	YES	YES	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
847	NO	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
848	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
849	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
850	YES	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
851	NO	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
852	NO	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
853	YES	YES	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
854	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
855	YES	NO	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
856	NO	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
857	YES	NO	YOUNG	12	NOT FOLL	YES	100	JEET	rainy	AVERAGE	LOW	25	YES
858	YES	YES	YOUNG	9	NOT FOLL	YES	90	CAR	DRY	AVERAGE	GOOD	30	YES
859	YES	NO	OLD	19	FOLLOWE	NO	90	CAR	rainy	GOOD	GOOD	80	NO
860	YES	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
861	NO	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO
862	YES	YES	YOUNG	5	NOT FOLL	YES	100	JEET	cloudy	BAD	GOOD	25	YES
863	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	GOOD	30	NO
864	YES	YES	OLD	19	FOLLOWE	NO	90	TRUCK	DRY	GOOD	GOOD	80	NO
865	NO	NO	YOUNG	5	FOLLOWE	NO	100	JEEP	DRY	GOOD	GOOD	25	NO
866	YES	NO	YOUNG	9	FOLLOWE	NO	90	CAR	DRY	GOOD	LOW	30	NO

Figure 4.3: Some other Traffic dataset attributes.

The weka explorer facilities are shown in figure 4.4. It also describes that how to open a file in the weka explorer.

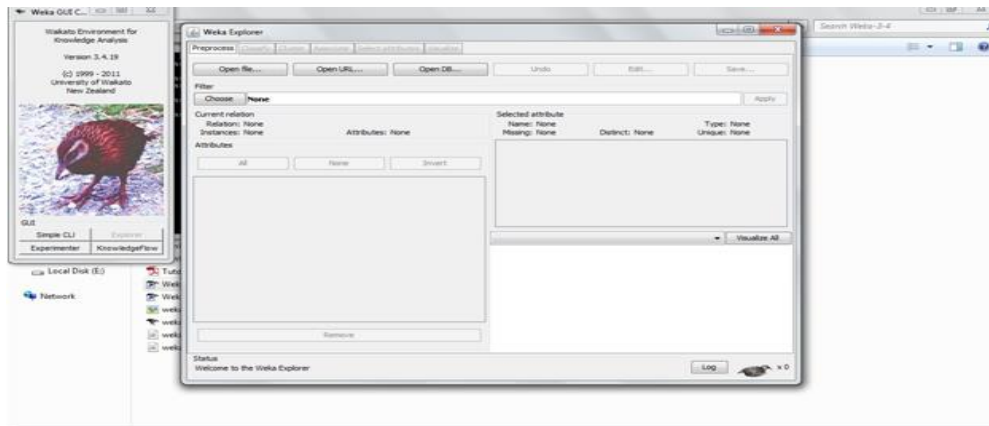


Figure 4.4: Shows weka gui and how to open a file.

The numbers of attributes are thirteen. Figure 4.5 shows different attributes of traffic dataset. It specifies how to classify various attributes in weka explorer.

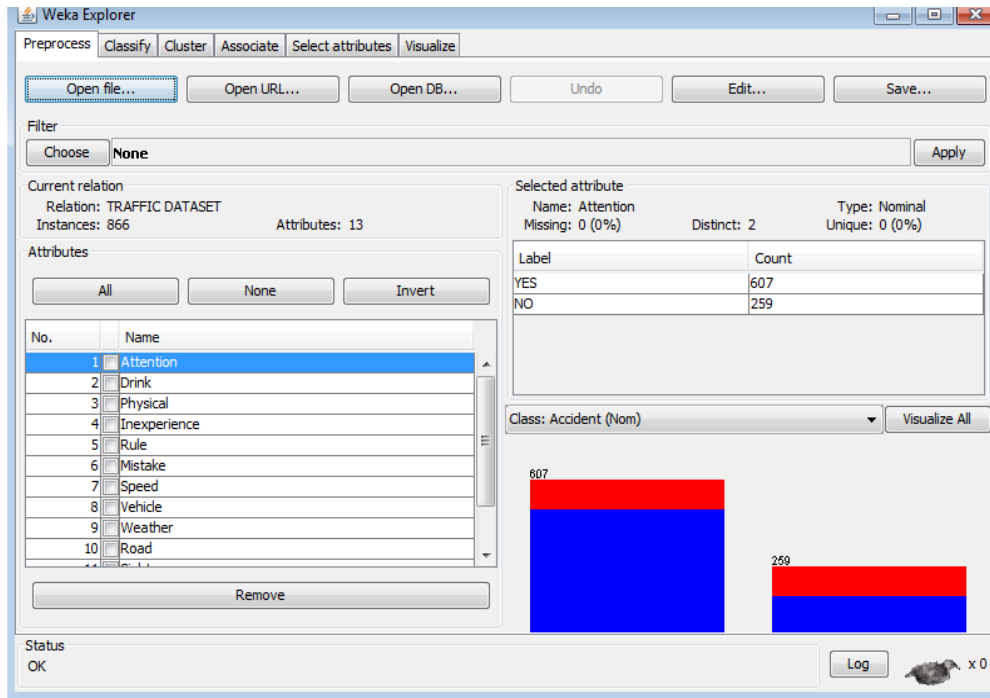


Figure 4.5: Classify various attributes of dataset.

Figure 4.6 classify no label value is in blue color and yes label value is in red color. The accident class attribute is used for this purpose.

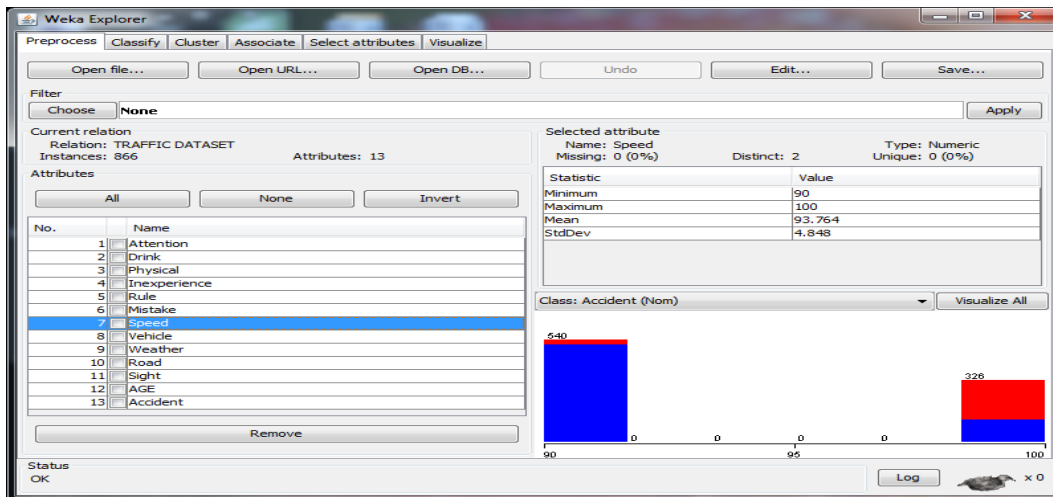


Figure 4.6: Shows label and value of speed attribute.

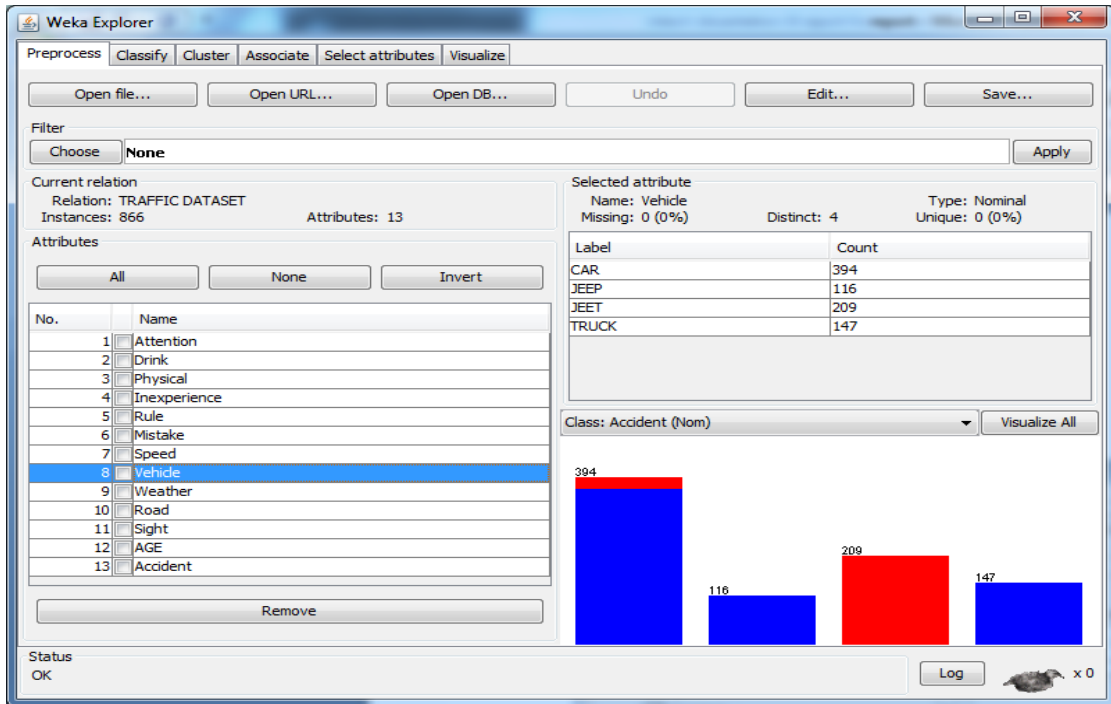


Figure 4.7: Shows label values of vehicle attribute.

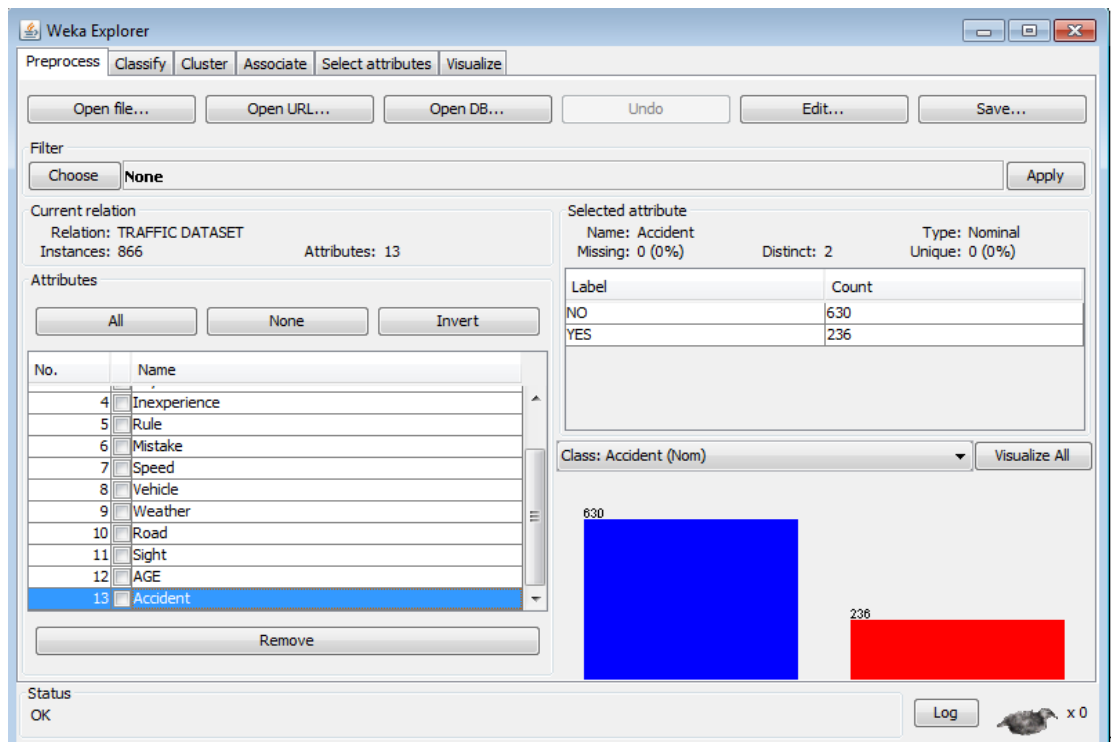


Figure 4.8: Shows no(blue) and yes(red) cluster.

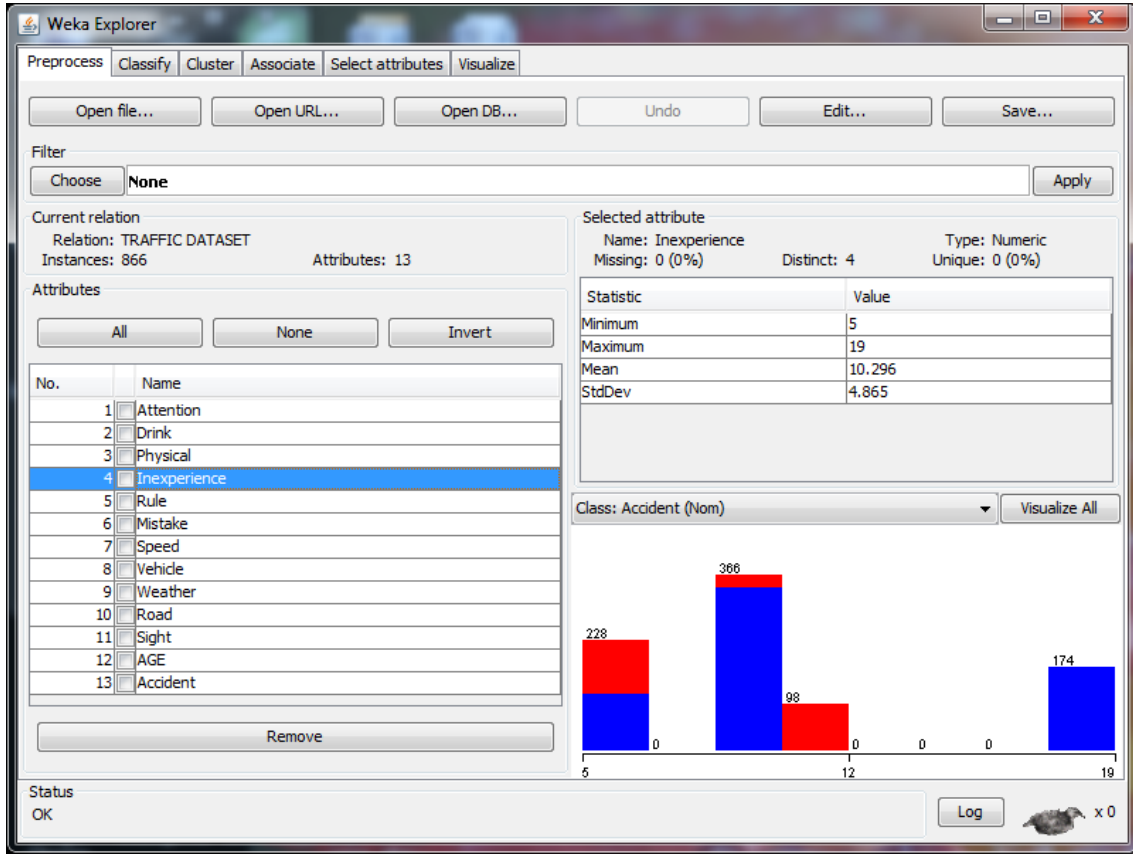


Figure 4.9: Shows numeric attribute in terms of mean and standard deviation.

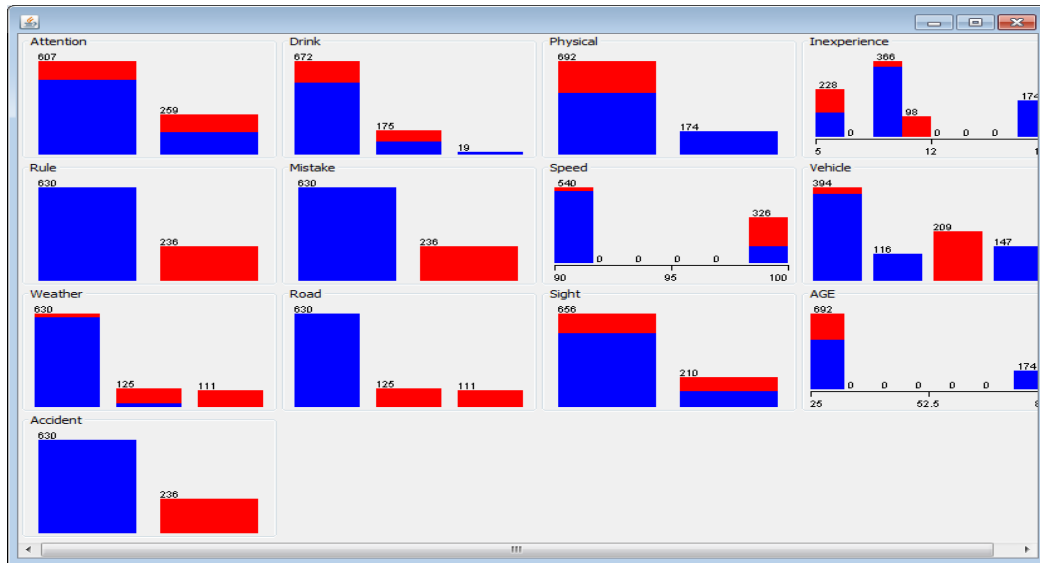


Figure 4.10: Shows Results of various attributes related to our dataset.

Using the visualize panel one can see the plot matrix of a dataset. The two dimensional and three dimensional view of the dataset can be viewed by this panel.

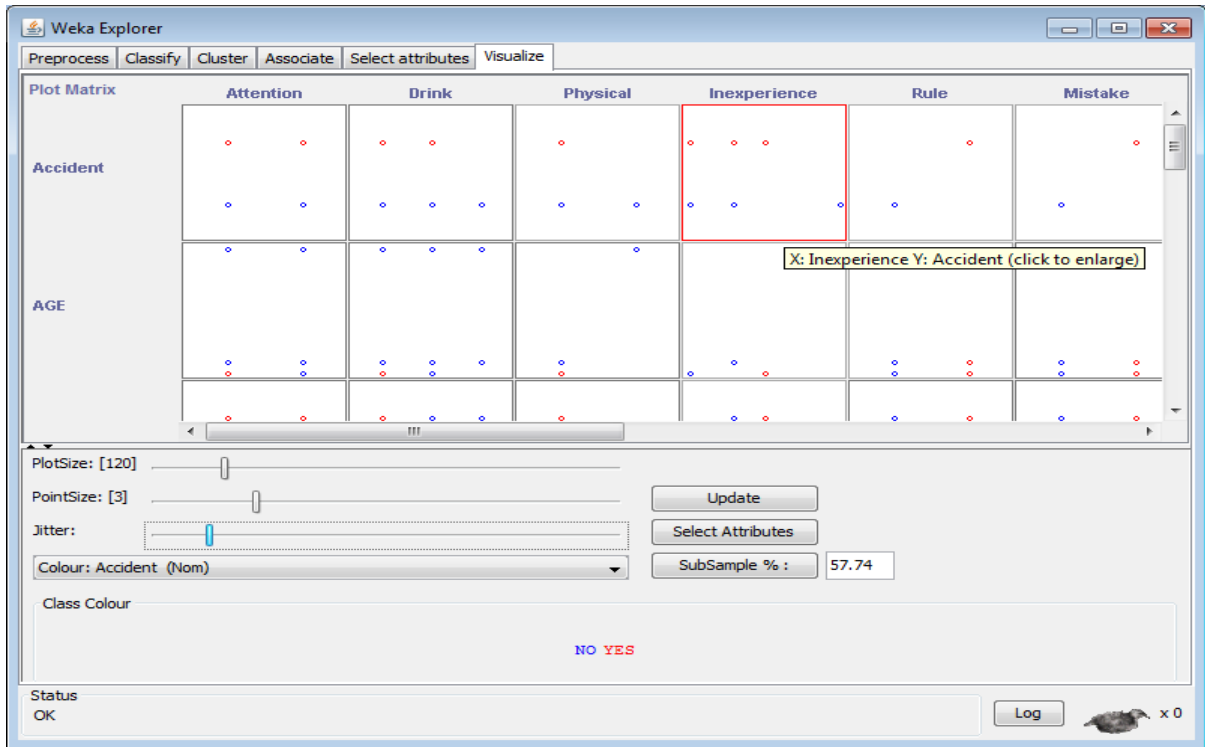


Figure 4.11: Shows the plot matrix of the dataset.

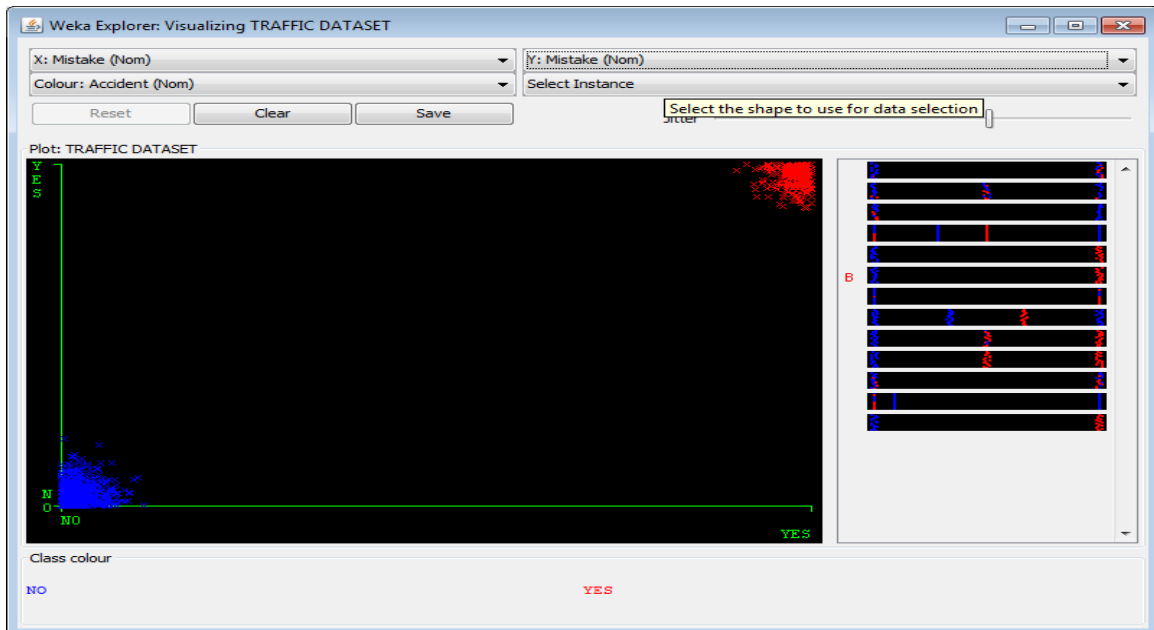


Figure 4.12: Classification of no (blue) and yes(red) clusters.

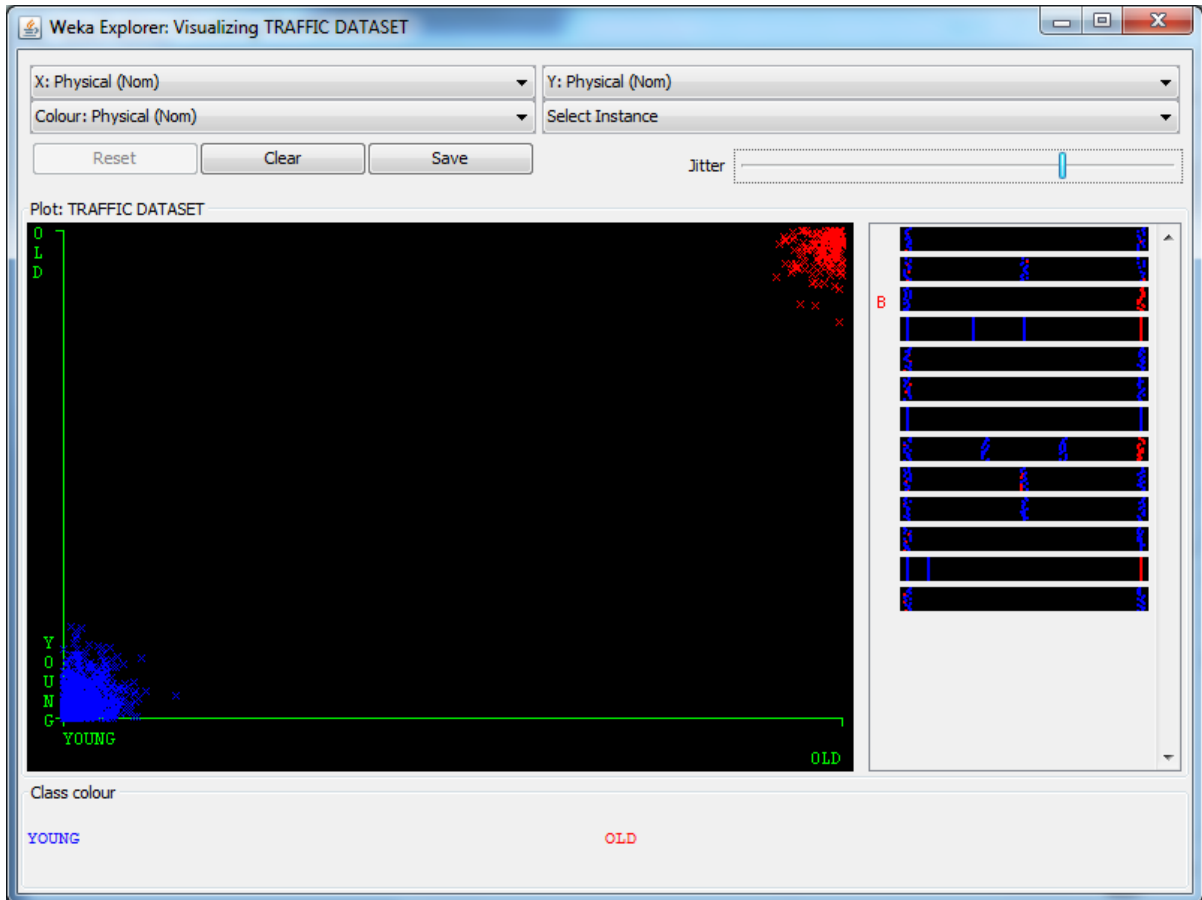


Figure 4.13: Classification of young(blue) and old(red) clusters.

After opening the dataset in the weka explorer using the preprocess panel select the cluster panel to fetch an enhanced algorithm in weka explorer.

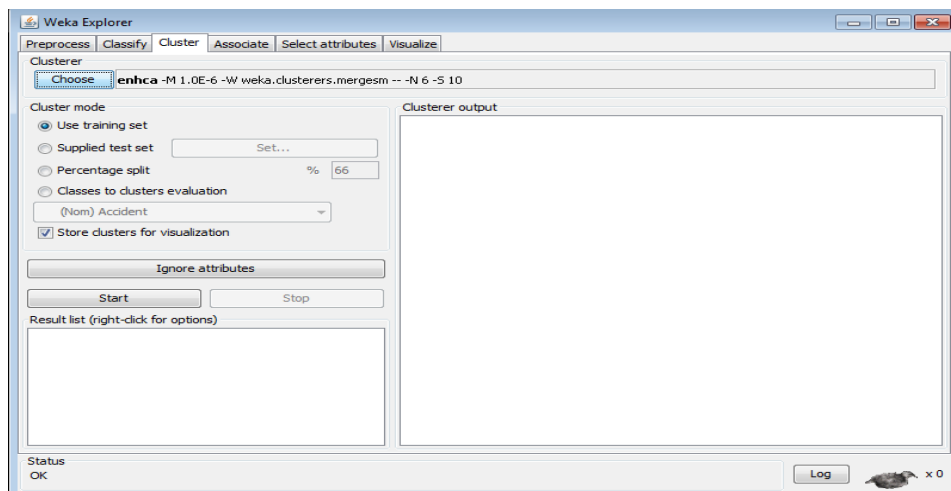


Figure 4.14: Shows how to fetch an enhanced algorithm in weka tool.

Figure 4.15 shows various attributes of dataset run by the enhanced k-mean algorithm. the cluster output gives run information . It shows scheme , number of instances, Relation with respect to our dataset and attributes etc.

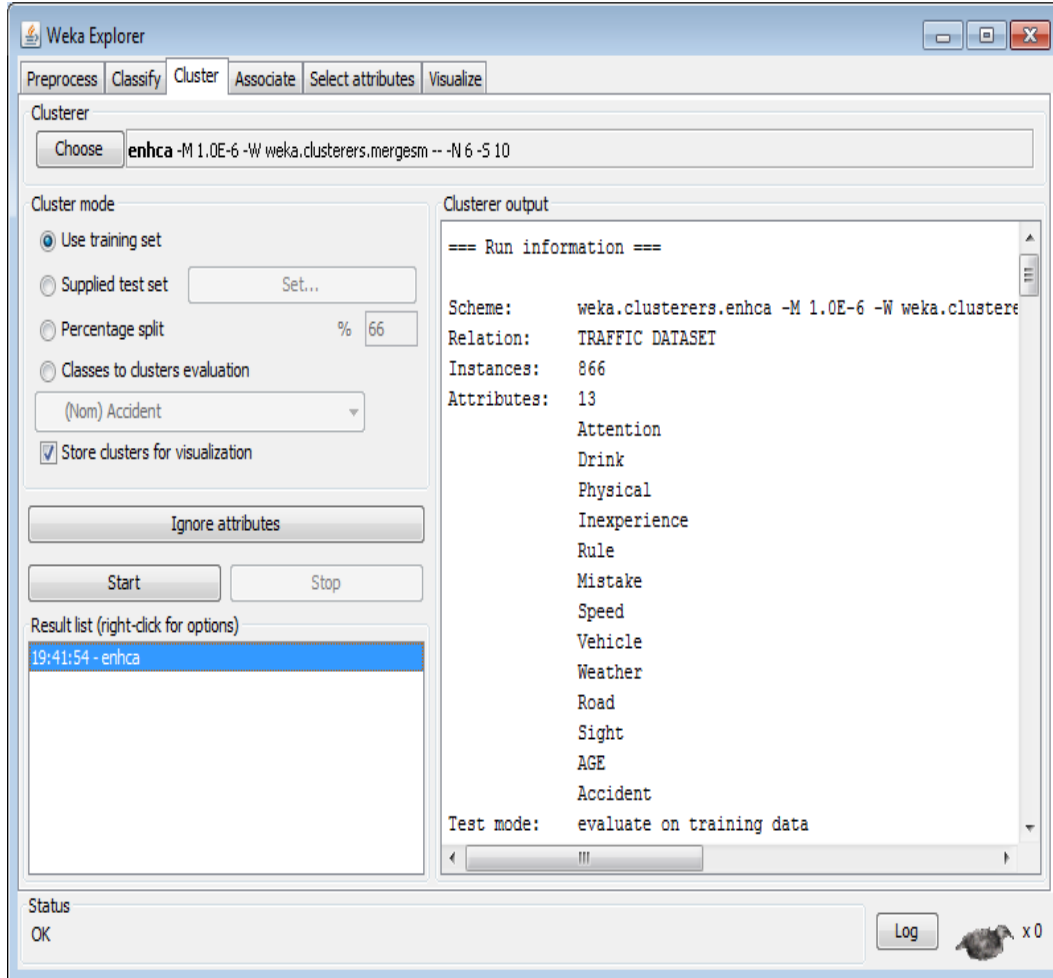


Figure 4.15: Shows various attributes in enhanced algorithm in weka explorer.

The result of the k-mean is as shown in figure 4.16. It has been observed that the number of clusters are two here. The cluster sum of squared error is calculated as output. The first cluster has instances 630 and the second cluster contain 236 instances. According to each numerical attribute its mean value and standard deviation is calculated.

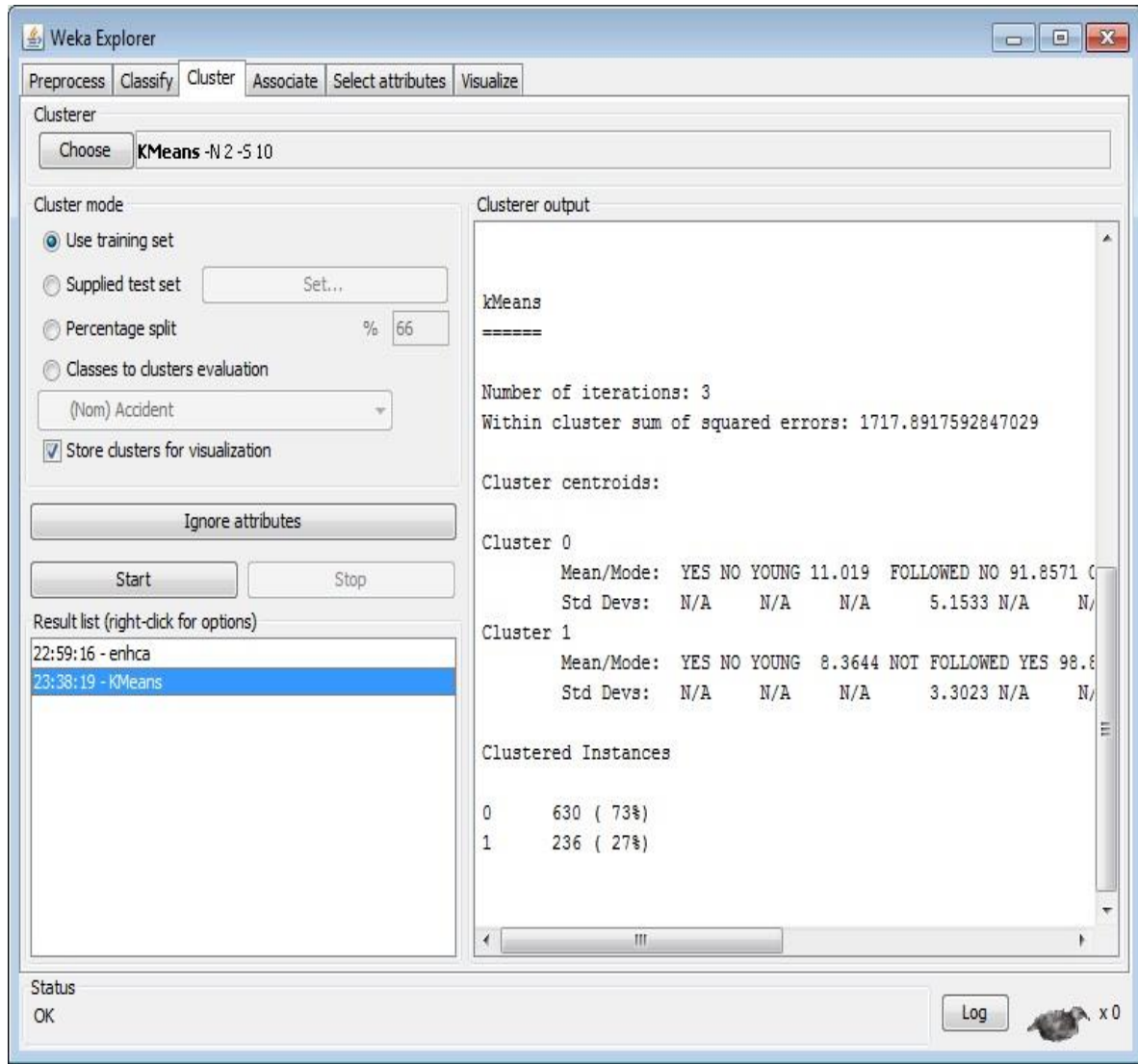


Figure 4.16: Result of existing k-mean algorithm.

The traffic dataset is taken and is used for checking the error of the enhanced k-mean algorithm. The enhanced clustering algorithm is used for mining high dimensional dataset. The same dataset is used for the enhanced k-mean and the existing algorithm. The number of clusters is six. Figure 4.17 depicts that the error obtained from the enhanced algorithm. In this case number of iterations is same but number of clusters is different. The means and standard deviation is calculated for each numerical value lies in the cluster. The enhanced mean algorithm is much better in terms of both numbers of clusters and in error.

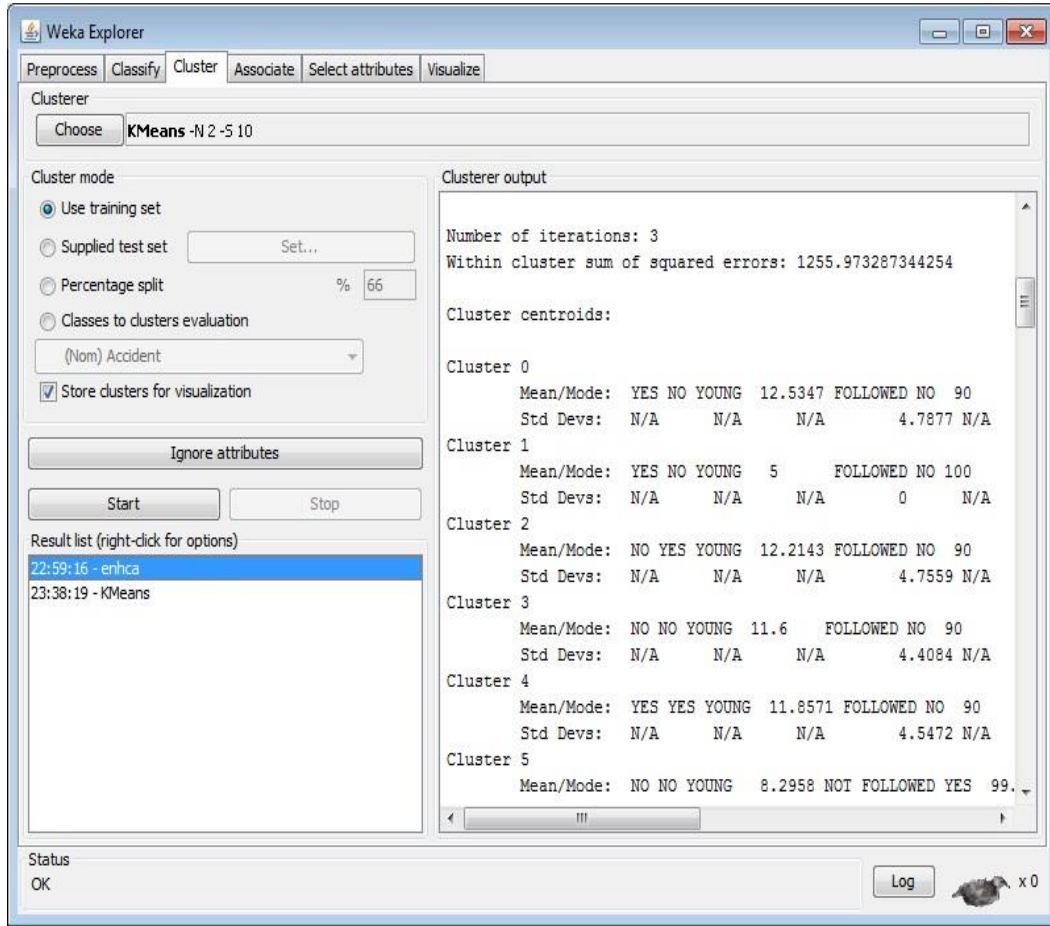


Figure 4.17: Result of enhanced k-mean algorithm.

The error is less in this case. The mean and the standard deviation are calculated for each cluster. The enhanced k-mean algorithm is much better in number of clusters as compared to the existing k-mean algorithm. The error obtained from the enhanced algorithm is less than the existing k-mean algorithm. The results of the experiments are shown in the table 1.

Table 4.1: Shows error comparison of both algorithms.

	K-means	Enhanced Algorithm
Error	1717.89	1255.97

Figure 4.18 shows the error bar graph for the performance of both the algorithms. It is clear

from the graph that the enhanced algorithm generates less error as compared to the existing k-mean algorithm.

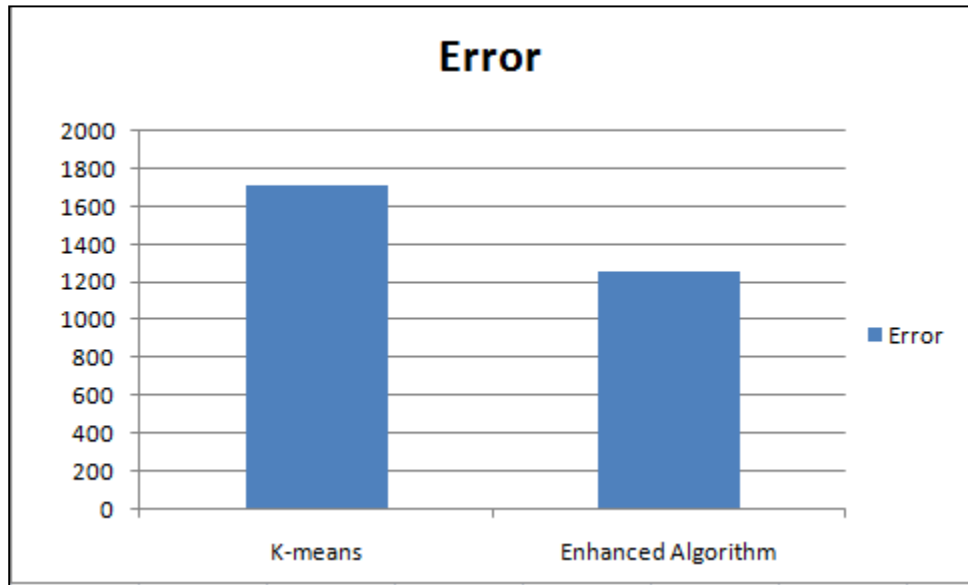


Figure 4.18: Error graph shows performance of both the algorithms.

It is clear from the table 2 that the numbers of iterations are same in both the cases. Same numbers of iterations are used for the comparison of both the algorithms.

Table 4.2: Number of iterations.

	K-means	Enhanced Algorithm
No. of iterations	3	3

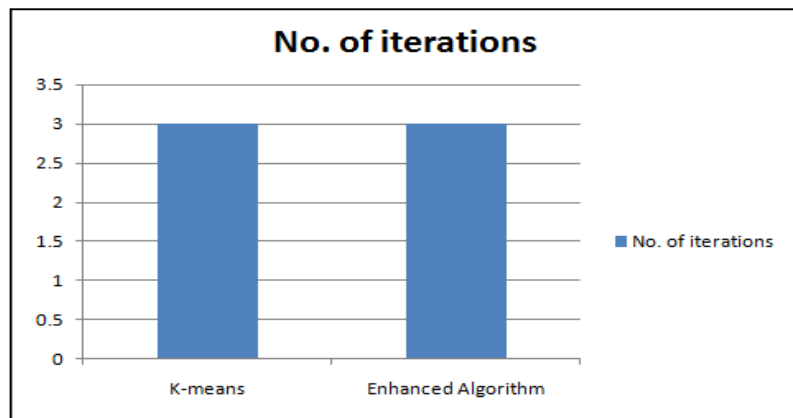


Figure 4.19: Bar graph shows performance comparison with same number of iterations.

According to each cluster formed it also calculates the probability. The figure 4.20 shows prior probability of first cluster for each attribute. Prior probability is that probability which occurs before observing the data. The revised probability of an event after getting new information is called its posterior probability. Posterior probability is calculated by updating the prior probability. Normal distribution often differs in terms of their mean and standard deviation.

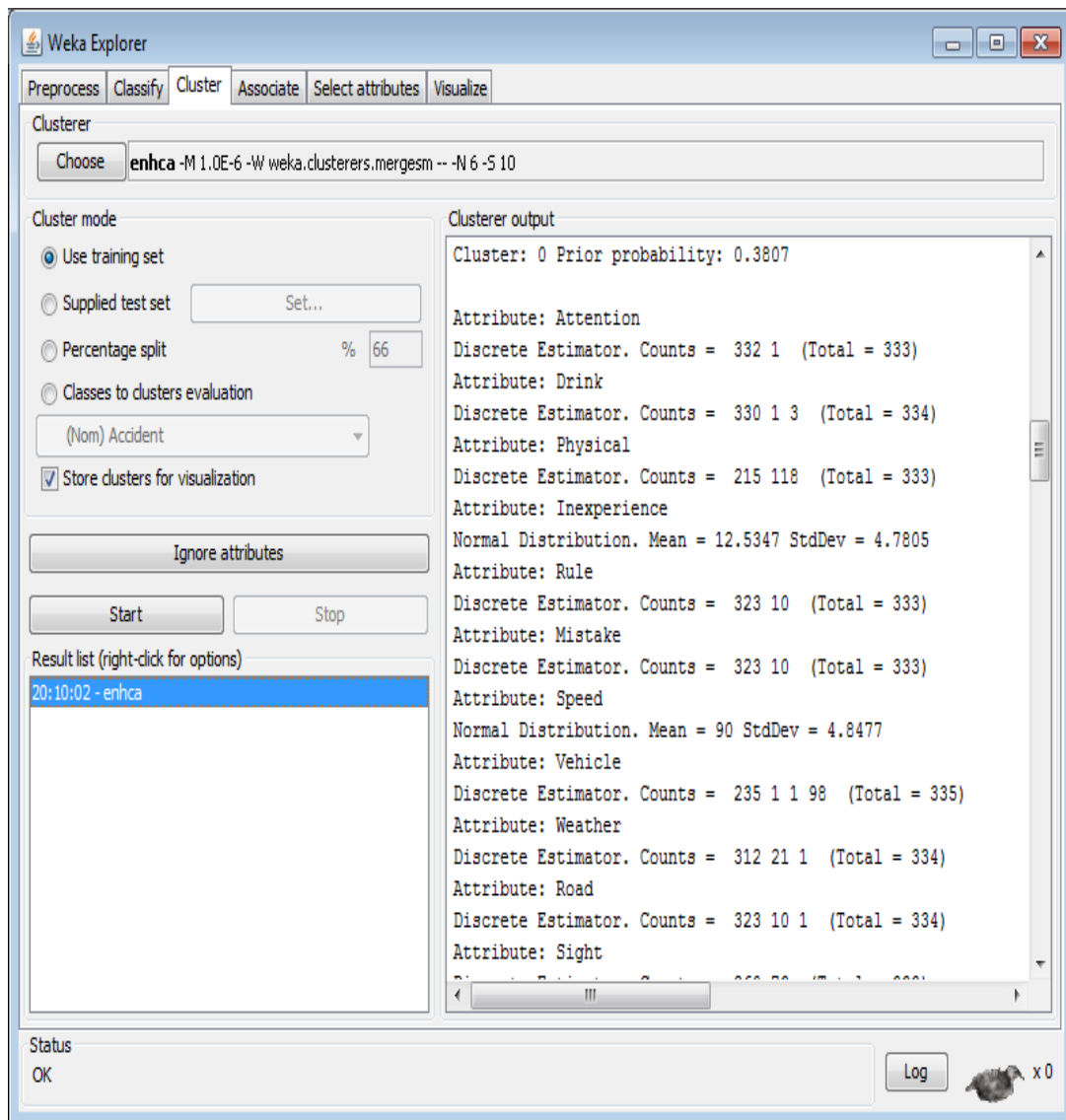


Figure 4.20: Shows probability of first cluster.

The figure 4.21 shows probability of second cluster for each attribute. For non-numeric attribute discrete estimator and its count value is calculated.

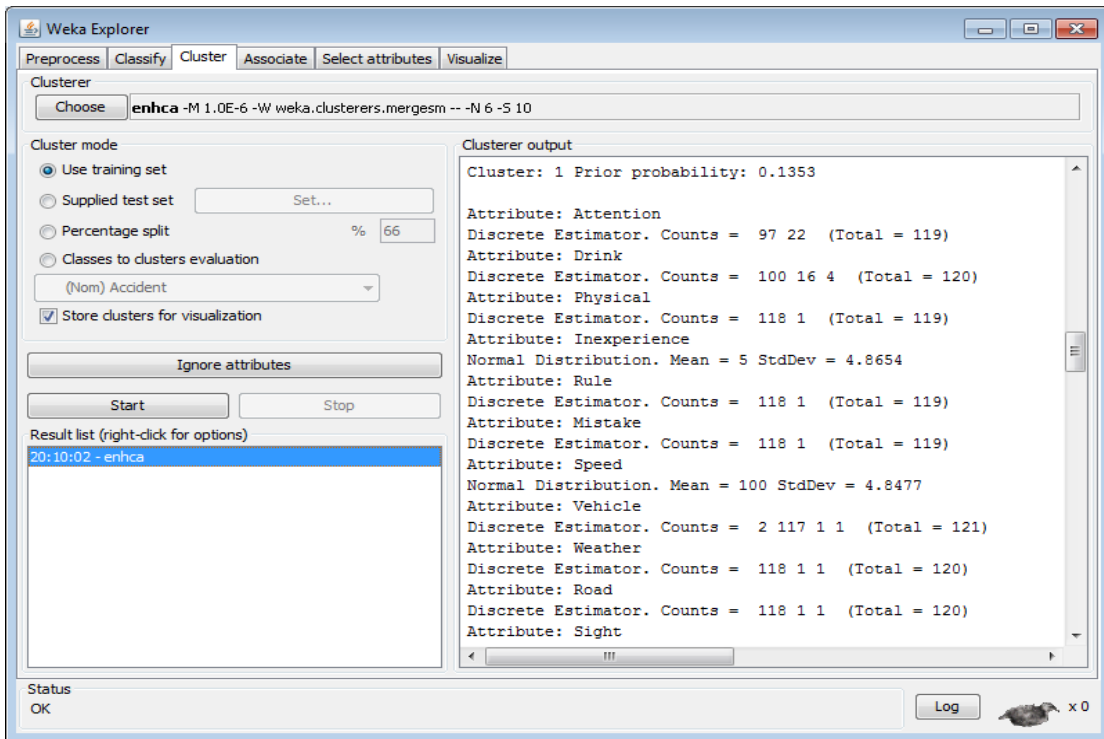


Figure 4.21: Shows probability of second cluster.

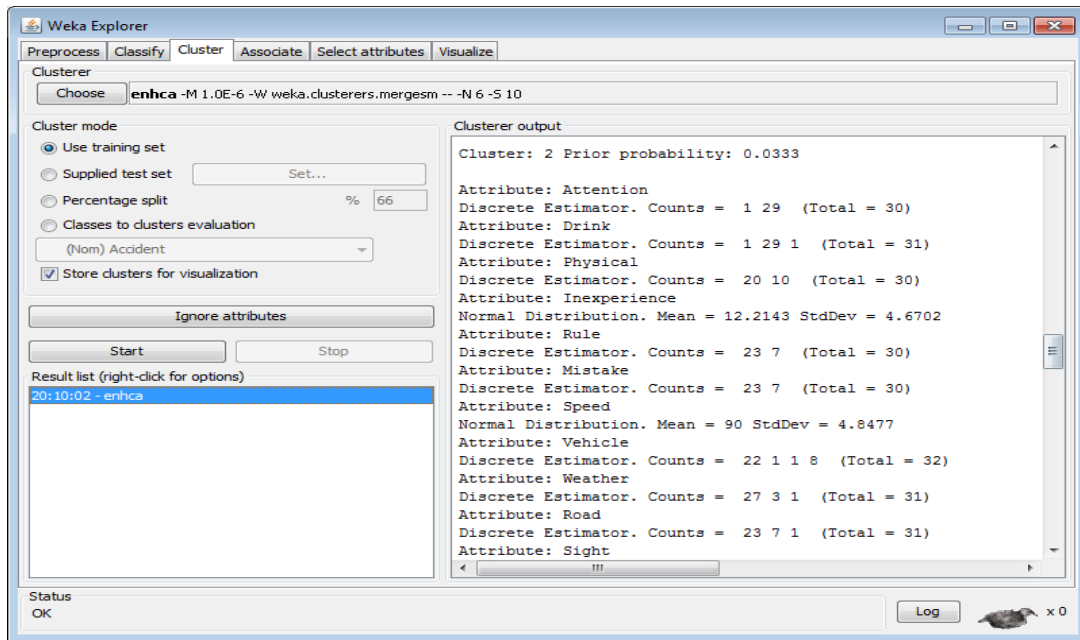


Figure 4.22: Shows probability of third cluster.

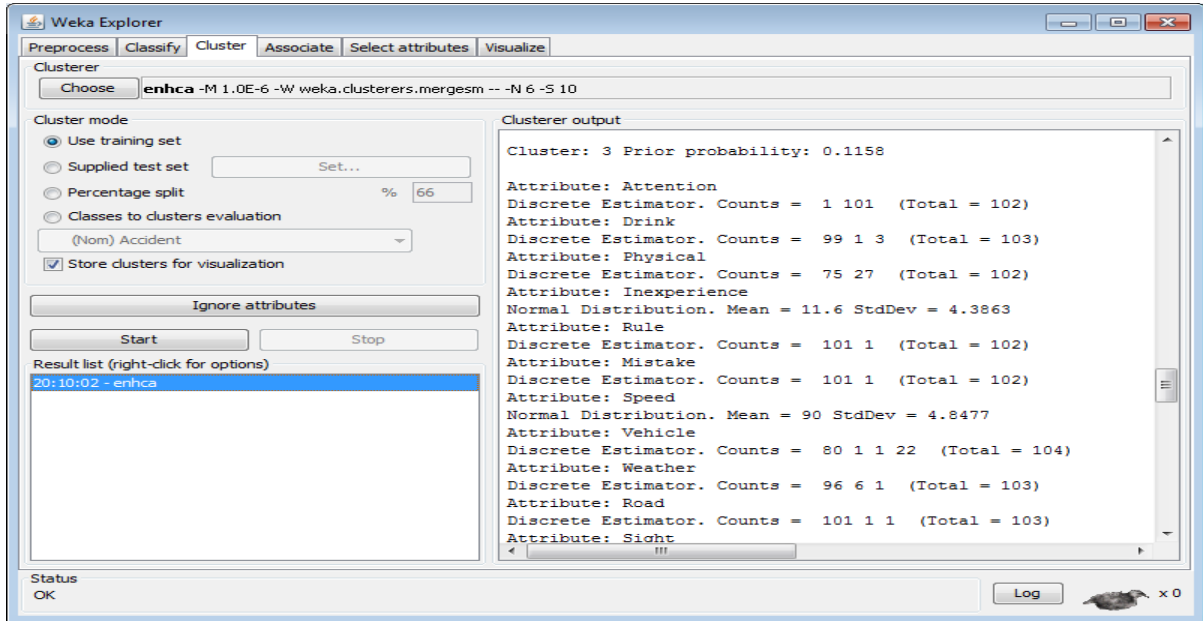


Figure 4.23: Shows probability of fourth cluster.

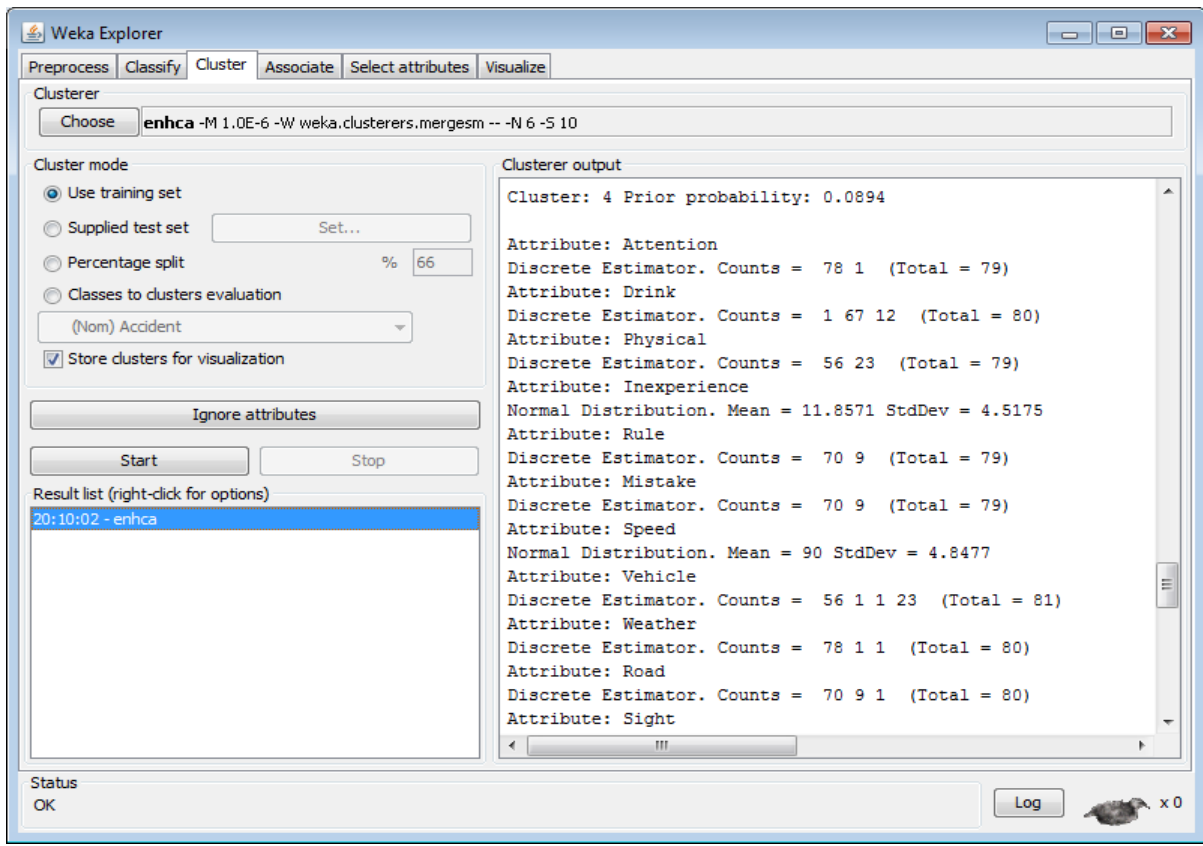


Figure 4.24: Shows probability of fifth cluster.

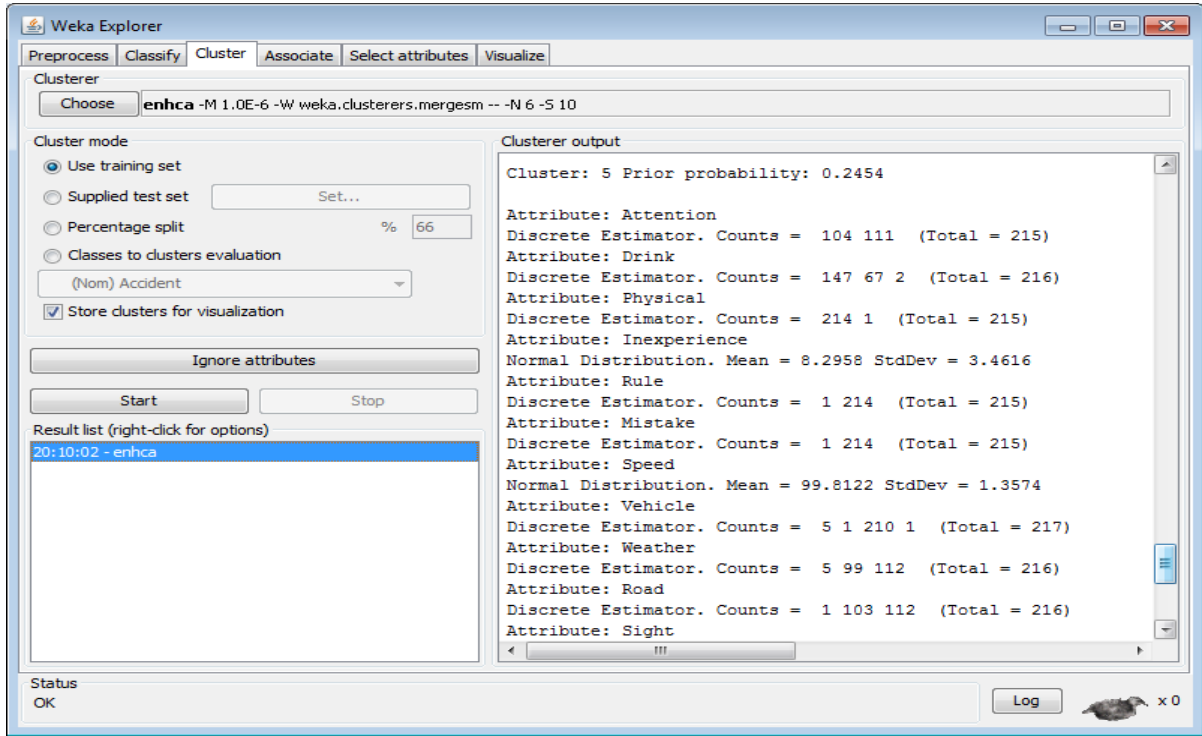


Figure 4.25: Shows probability of sixth cluster.

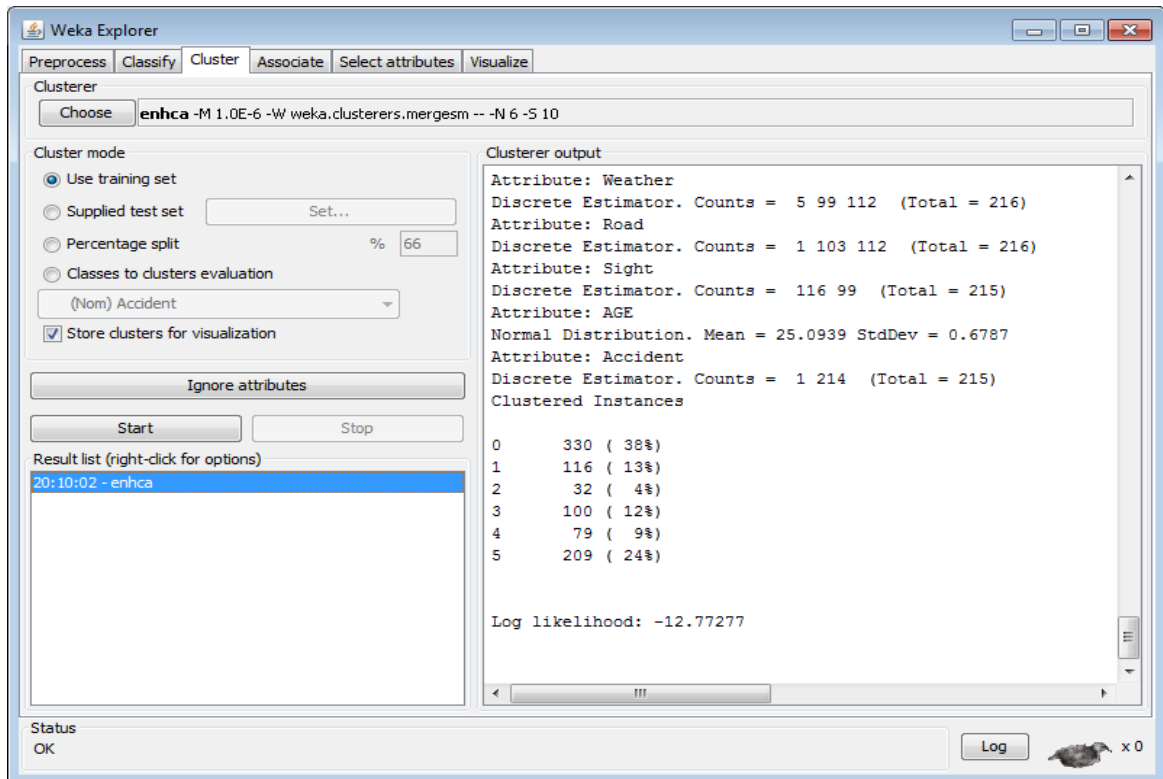


Figure 4.26: Shows each clustered instances.

The figure 4.26 shows each clustered instances. First cluster contains 330 instances. Second contains 116 instances. Third contains 32 instances. Fourth contains 100 instances. Similarly fifth and sixth cluster contains 79 and 209 instances respectively.

Chapter 5

CONCLUSION

The k-means clustering algorithm is used to mine high dimensional dataset. But the existing algorithm produces more error as compared to the enhanced algorithm. Thus it does not provide good accuracy results. The accuracy obtained by enhanced algorithm is much better than that of existing one. The paper provides an enhanced algorithm which performs better in number of clusters and the method for finding the centroid. The analysis is done by taking the traffic dataset which considers several attributes. Thus listing various attributes, this is the main reason for the accidents.

REFERENCES

-
- [1] Suman, Pooja Mittal “A Comparative Study on Role of Data Mining Techniques in Education”, International Journal of Emerging Trends & Technology in Computer Science , Vol 3, Issue 3, May – June 2014.
 - [2] Amandeep Kaur Mann, Navneet Kaur "Survey Paper on Clustering Techniques."International Journal of Science, Engineering and Technology Research 2.4 (2013): pp-0803.
 - [3] Saurabh Shah, Manmohan Singh "Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm."Communication Systems and Network Technologies (CSNT), 2012 International Conference on. IEEE, 2012.
 - [4] Malay K. Pakhira "A modified k-means algorithm to avoid empty clusters."International Journal of Recent Trends in Engineering 1.1 (2009).
 - [5] Raed T. Aldahdooh, Wesam Ashour “Distance-based Initialization Method for K-means Clustering Algorithm”, I.J. Intelligent Systems and Applications, 2013, 02, 41-51.
 - [6] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal "Crime Analysis using K-Means Clustering." International Journal of Computer Applications 83.4 (2013): 1-4.
 - [7] K. A. Abdul Nazeer, M. P. Sebastian,"Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." Proceedings of the World Congress on Engineering. Vol. 1. 2009.
 - [8] M. Goyal, S. Kumar "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability" Journal of The Institution of Engineers (India): Series B: 1-6.
 - [9] Sapna Jain, M. Afshar Aalam, M. N DOJA. “K-means Clustering Using Weka Interface” Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010 Bharati Vidyapeeth’s Institute of Computer Applications and Management, New Delhi .
 - [10] Rauf, Azhar, et al. "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity." Middle-East Journal of Scientific Research 12.7 (2012): 959-963.
 - [11] H.S. Behera, Abhishek Ghosh., Sipak ku. Mishra “New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining”

- International Journal of Advanced Research Computer Science and Software Engineering”Volume 2, Issue 4, pp 287-292, April 2012.
- [12] Narendra Sharma, Aman Bajpai, and Ratnesh Litoriya "Comparison the various clustering algorithms of weka tools." Facilities 4 (2012): 7.
- [13] Ritu Sharma, M. Afshar Alam, Anita Rani "K-Means Clustering in Spatial Data Mining using Weka Interface" International Conference on Advances in Communication and Computing Technologies (ICACACT) Proceedings published by International Journal of Computer Applications, pp 26, 2012.

APPENDIX

LIST OF ABBREVIATIONS

1. DBMS – Database Management System
2. SOM– Self Organizing Map
3. CLARA-Clustering Large Applications
4. PAM – Partition Around Medoid
5. DIANA –Divisive Analysis.
6. AGNES–Agglomerative Nesting
7. CLI-Command Line Interface
8. EM-Expectation Maximization Algorithm
9. JVM-Java Virtual Machine
10. USA-United States of America
11. GUI- Graphical User Interface