International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

Comparative Analysis of Different Clusteirng Techniques for Data Analytics

Dr. Gagandeep
Assistant Professor,
Department of Computer Science,
Punjabi University, Patiala
(PUNJAB)

Navneet Kaur
M.Phil Scholar,
Department of Computer Science,
Punjabi University, Patiala
(PUNJAB)

***Abstract*- The basic principle of data mining is to analyze the data and find out the useful information from it. Data mining is the basic stage of knowledge Discovery in Databases process. Clustering or cluster analysis is part of data mining. Data mining software allows users to analyze the data. This paper introduces the key principle of clustering of WEKA tool. WEKA is data mining tool used to analyze the data. It provides the facility to clustering the data through various algorithms.**

Keywords— Cluster Analysis, Clustering, Data Mining.

I. INTRODUCTION

We live in the world of data. A large number of users are generating the data everyday by various sources including cell phones, E- mails, attaching files, online chat, uploading pictures or videos, online shopping, e-banking etc. Big data can also refers to a collection of large data set that is increasing day by day and it cannot be handled by traditional data managing techniques like DBMS, RDBMS etc. Big data has five characteristics. These are also known as 5 V's of big data. The characteristics are used to understand the nature of big data. These are volume, variety, velocity, veracity and value. Along with these characteristics, big data also faces some challenges. These challenges are privacy and security. It means that it is a challenge to secure the data. Next challenge is information sharing. Third one is analytical challenge, means how the data can be analyzed if it increases continuously. What type of techniques and tools are used to analyze the data. Next challenge is human resource and manpower. Technical challenge includes quality of data, fault tolerance and heterogeneous data.

If we are talking about an analytical challenge then it is very necessary to analyze the data so that the useful information can be retrieved from the data. Various analyzing techniques are available like data mining, data clustering, machine learning, text analytics association rule learning and classification but clustering technique helps to dividing the structured data into the groups. WEKA software tool is the data analyzing tool that helps in classification, clustering and visualization of the data. WEKA can be considered as one of the best method of data analyzing tool.

II. CLUSTERING ANALYSIS

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in same group (cluster) is more similar to each other than to those in other groups (clusters). Or clustering is a technique for finding similarity groups in data, called cluster. In this process objects with same features are collected in one group (cluster). Special property is the objects are more similar to one another with in the same cluster and dissimilar to the objects in other clusters. It is a main task for statistical data analysis, used in many fields including machine learning, pattern recognition, image analysis, information retrieval etc.

Clustering analysis itself is one specific algorithm, but general task is to be solved. By using clustering techniques we can identify even denser and sparse regions in object space and can also discover overall distribution pattern and correlations between data attributes.

The simple example of clustering is library system. In which books are related to different subjects. Each subject has its own section (cluster) and particular subject is further categorized into subgroups. Like computer applications subject is further subdivided into networking, software engineering, programming languages etc. and according to each subgroups books are arranged.

III. TYPES OF CLUSTERING

The clustering analysis methods or clustering may be categorized into following types:

1. Centroid Based Clustering: In the centroid based clustering, clusters are represented by a central vector, which may not necessarily of the dataset. It is assumed that each cluster has at least one object and each object belongs to only one cluster. Its main article is K-Means clustering.
1. Distribution Based Clustering: The clustering model most related to statistics is based on distribution model. Clusters can then easily be defined as objects belonging most likely to the same distribution. Example: Expectation Maximization.
2. Density Based Methods: In density based clustering, clusters are defined as area of higher density than the remainder of the data set. Objects in these sparse areas- that are required to separate clusters are usually considered to be noise and border points.

The most popular density based clustering method is DBSCAN.

3. Hierarchical Clustering: Hierarchical method obtains a nested partition of the objects resulting in a tree of clusters. These methods either, start with one cluster and then split into smaller clusters or start with each object in an individual cluster and try to merge similar clusters into large and large clusters. Two types of hierarchical methods are Divisive and Agglomerative.
4. Grid Based Clustering: In this, the object space rather than the data is divided into grid. Grid partition is based on characteristics of the data and such methods can deal with non-numeric data more easily and not affected by data ordering.
5. Model Based Clustering: This model is based on a probability distribution. The algorithm tries to build clusters with a high level of similarity within them and a low level of similarity between them. Similarity measurement is based on the mean values and the algorithm tries to minimize the squared- error function.

IV. WEKA : DATA MINING TOOL

WEKA (Waikato Environment for Knowledge Analysis) is popular suite of machine learning software written in Java, developed at the University of Waikato, New Zeeland. The WEKA is an endemic bird of New Zeeland. The timelines of the various stages of Weka developed history are as follows:

- Late 1992 –funding was applied by Ian Witten
- 1993- developed of the interface and infrastructure
- Sometimes in 1994- first internal release of Weka
- October 1996- first public release of Weka (v2.1)
- July 1997- Weka 2.2
- Early 1997- decision was made to rewrite Weka in Java.
  - Originated from code written by Eibe Frank for his Ph. D.
  - Originally codenamed JAWS (Java Weka System)
- May 1998- Weka 2.3
- Mid 1999- Weka 3(100% Java) released

The main reason why selected to use WEKA was because of its visibility. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from user's own Java code. WEKA is open source and freely available.

Fig.1 Interface of WEKA Tool

Weka's four Application Interaface:

- Explorer: An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail). Its functions are pre-processing, attribute selection, visualization, classification and clustering.
- Experimenter: it is an environment for performing experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag and drop interface. One benefit is that it supports incremental learning. It explains visual design of data flow.
- Simple CLI: it is command line interface allows direct execution of Weka commands for operating systems that do not provide their own command line interface.

Features of Weka Tool:
Weka is a GUI for where various applications are available. Some of the features of Weka tool are described as:

- Data Pre-processing
- Data Classification
- Data Clustering
- Attribute Selection
- Data Visualization

Advantages of Weka

- ✓ They have freely availablity under the GNU(General Public Licence)
- ✓ Due to its portability, it is fully implemented in Java programming language and runs on almost any architechture.
- ✓ It is easily use due to its graphical user interface.
- ✓ Large collection of data pre-processing and modelling techniques.
- ✓ It is easy usable by people who are not data mining specilist.

V. DATASET

For performing the comparison analysis we need the past project dataset. For this research Pima _ diabetes dataset has been taken from officially website of WEKA. Dataset includes 768 instances and 8 attribute with class. The class has two values class 0 and class 1. No missing value is available in dataset. Three clustering methods have been applied on this dataset. We can directly apply this data in data mining tool. Each instance of dataset belongs to a class. On the basis of this class the cluster are generated by applying algorithms using WEKA interface. WEKA is a landmark system in the history of the data mining and machine learning research communities.

Pima_diabetes Dataset The diagnostic, binary-valued investigated is whether the patient shows the sign of diabetes (1 is interpreted as tested positive and 0 is interpreted as tested negative)

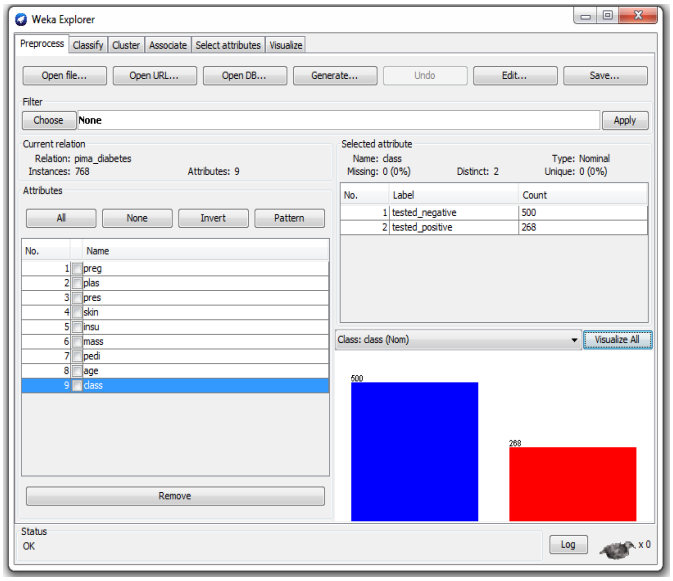

Fig 2. Weka Explorer With Dataset Results

Weka explorer results of dataset in figure 2. In this the Relation, Instances Attribute and visualization results have been shown. 500 and 268 are two values of Class attribute. In above result Class attribute is selected and its information has been shown in selected attribute part of result. In this Distinct value shows that class attribute has 2 distinct values i.e. 500 (Tested Negative) and 268 (Tested Positive). Type of the selected attribute has been defined by Type.

A. K-MEANS CLUSTERING

K-Means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It is most popular classical clustering methods and easy to implement. Its aim is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

The method is called K-Means since each of the K clusters is represented by the mean of the objects (called the centroid) within it. It is also called the centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it.

Once allocation is completed, the centroid of the clusters are recomputed using of simple means and process of allocating points to each cluster is repeated until there is no change in the clusters.

K-Means clustering method has two distance functions:

1. Euclidean Distance Function: In mathematics, the Euclidean distance is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula.
2. Manhattan Distance Function: tha Manhattan distance function computes the distance that would be travelled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corrsponding components.

K-Means Algorithm

1. Select the number of clusters. Let the number be k.
2. Pick 'k' seeds as centroid of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.
3. Compute the Euclidean distance of each object in the data set from each of the centroid.
4. Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.
5. Compute the centroid of the clusters by computing the means of the attribute values of the objects in each cluster.
6. Check if the stopping criterion has been met (e.g. the cluster membership is changed). If yes, go to step 7 and if no, go to step 3.
7. [optionl] one may decide to stop at this stage or to split a cluster or combine two clustres heuristically until a stopping criterion is met.

Advantage of K-Means

- Simplest method and easily understandable
- If clusters are globular, K-means may produces tighter clusters
- It may comsume lesser time to build the model

Disadvantage of K-Means

- It is not suitable for different size of clusters.
- It does not work properly with noise and outlier data points.

Results of K-Means

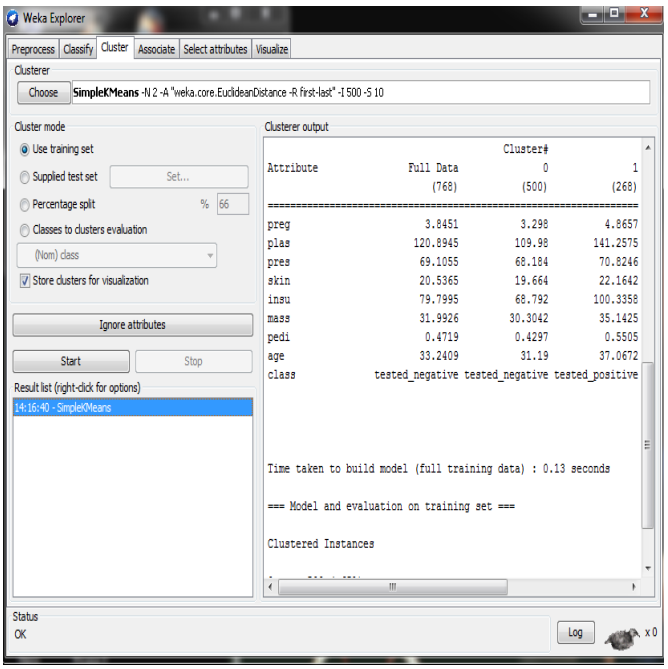

Fig 2. K-Means Algorithm

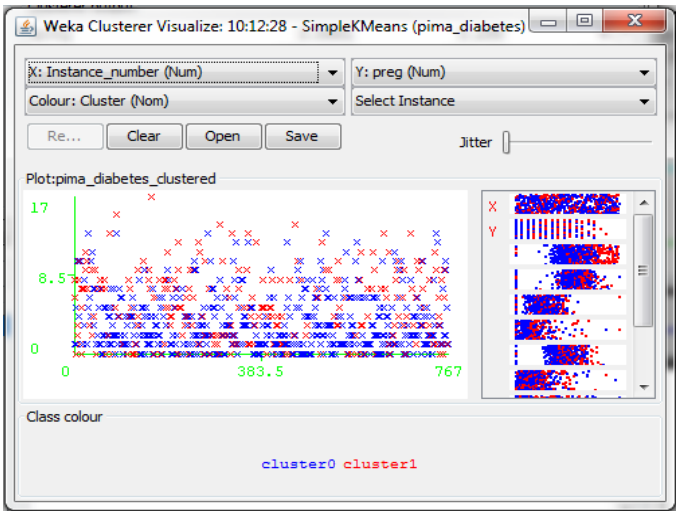

Fig. 3 Result of K-Means Algorithm

The visualize Cluster Assignment tab shows the 2-D plot of the attribute values on the x-axis, y-axis and give the detailed information about very instance in figure 3. The instances in the blue color classified as the tested negative instances and the instances in the red color classified as the tested positive instances. By fixing the one axis attribute and keep on changing the other attribute, different graphs have been shows.

Jitter function in the graph just adds artificial random noise to the coordinates of the plotted points in order to spread the data out a bit (so that you can see points that might have been obscured by others). The blue color instances are correctly classified instances and the red color instances are the incorrectly classified instances.

Table 1: Values of K-Means

No. of Iterations	4
Time taken to build model	0.02 Seconds
Within cluster sum of squared errors	149.52

Table 2: Clusters Values

Clusters Values	
0	500(65%)
1	268(35%)

B. Expectation Maximization Algorithm (EM) EM algorithm is also an important algorithm of data mining. An Expectation Maximization (EM) algorithm is an iterative method for finding maximum likelihood of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes the expectataion of the log-liklehood evaluated using the current estimate for parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameters –estimates are then used to determine the distribution of the latent variables in the next E step.

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or user may specify apriori how many clusters to generate.

1. The number of cluster is set to 1.
2. The training set is split randomly into 10 folds.
3. EM is performed 10 times using the 10 folds the usual CV way.
4. The log likelihood is averaged over all 10 results.
5. If log likelihood has increased the number of clusters is increased by 1 and the program continues at step 2.

Advantage of using EM Technique

- Extremely useful for real world dataset.

Disadvantage

- Highly complex in nature.

Results of EM

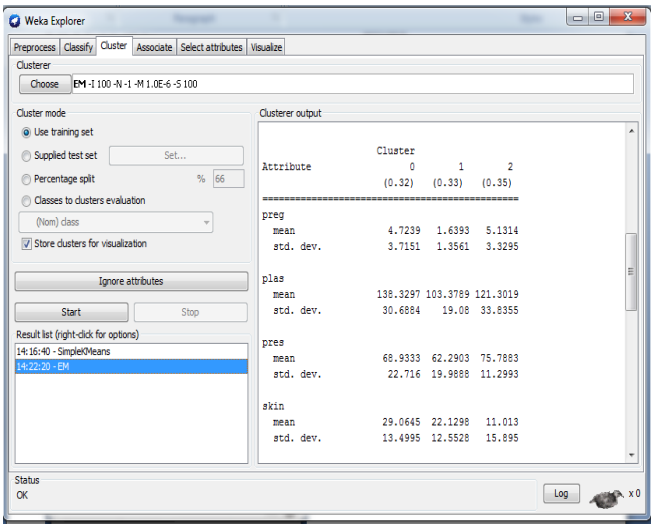

Fig. 3 EM Technique

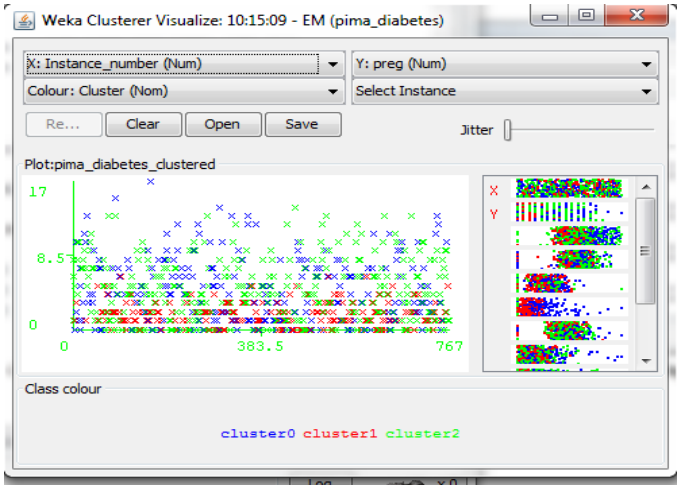

Fig. 4 Results of EM Technique

The visualize Cluster Assignment tab shows the 2-D plot of the attribute values on the x-axis, y-axis and give the detailed information about very instance in fig 4.

Table 3: Statistical Value

No. of Cluster selected by cross validation	3
Time taken to build model(full data model)	10.02 Seconds
Log likelihood	24.97

Table 4: Clusters Instances

0	228(30%)
1	203(26%)
2	337(44%)

C. DBSCAN

The most popular density based clustering method is the DBSCAN. It is desity based clustering. In density based clustering, clusters are defined as area of higher density than the remainder of data set. Objects in these sparse areas- that are required to separate clusters are usually considered to be noise and border points.

In contrast to many newer methods, if features a well defined cluster model called "density rach ability". Another interesting property of DBSCAN is that its complexity is fairly low. It requires on the database and it will discover essentially the same results in each run, therefore there is no need to run it multiple times.

DBSCAN Algorithm

1. Arbitrary select a point p.
2. Retrieve all points density-reachable from p w.r.t. Eps and Minpts.
3. If p is a core point, a cluster is formed.
4. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.

Advantages

- DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to K-means.
- DBSCAN can find arbitrary shaped clusters. It can even find a cluster completely surrounded by a different cluster.
- DBSCAN has a notion of noise, and is robust to outlier.

Disadvantage

- DBSCAN does not deal with high dimensional data.

Results of DBSCAN

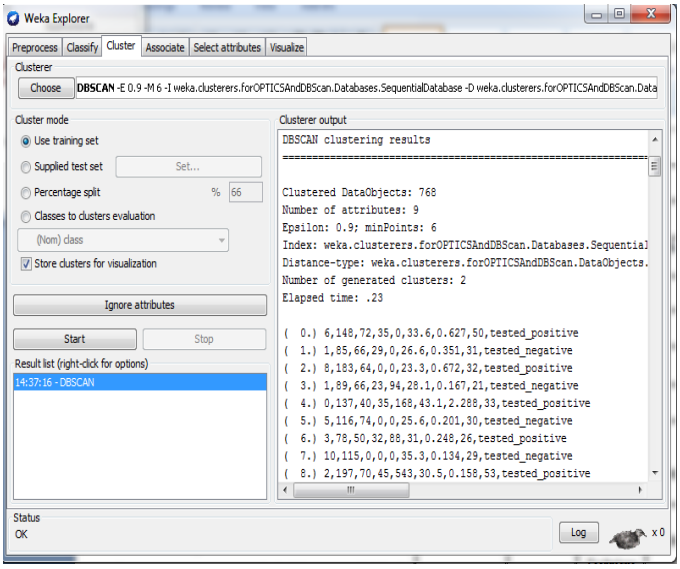

Fig. 5 DBSCAN Algorithm

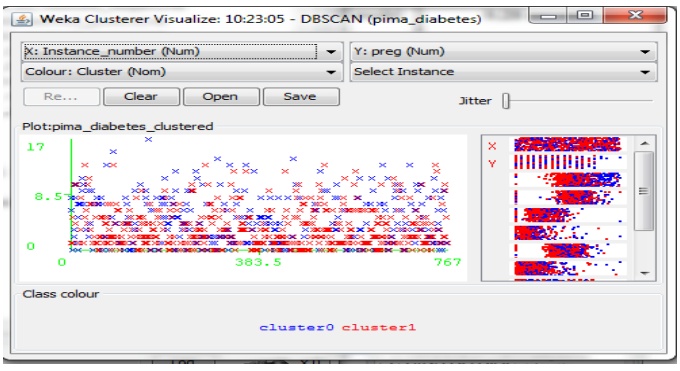

Fig. 6 Result of DBSCAN Algorithm

Table 5: Calculations of DBSCAN

Epsilon	0.09
Min Points	6
No. of Generated clusters	2
Time taken to build model	0.28 seconds

Table 6: Clusters Values

Clusters Values	
0	268(35%)
1	500(65%)

VI. COMPARISON RESULTS OF DIFFERENT ALGORITHMS

Parameters	K-Means	EM Technique	DBSCAN
TIME	0.02 Sec.	10.02 Sec.	0.28 Sec.
RATIO	65:35	30:26:44	35:65
SCAN / ITERATION	4	No Scan is defined	One Scan
CLUSTER	Centroid	Statistical	Density
NUMBER OF CLUSTERS	2	3	2
NOISE HANDLING	NO	NO	YES

VII. CONCLUSION

In recent few years data mining technique for analyzing data covers every area in our life. Data mining techniques are used in medical, banking, insurances, education etc. The main aim of this paper is to provide a detailed introduction of Weka clustering algorithms. Weka is the data mining tool mainly used for data analyzing. With the help of figures working of Weka clustering algorithms has been shown. Every clustering algorithm has its own importance. Various parameters have been used for the comparison of clustering algorithm. If we conclude the results, K-Means clustering algorithm takes lesser time to build the model than EM and DBSCAN. Ratio is correctly classified in K-Means and DBSCAN algorithm than EM Technique. Total number of scans in DBSCAN algorithm is minimum i.e. 1. Number of clusters is selected by algorithm itself but user can also select it. If we are talking about the sample dataset i.e. pima_ diabetes, results of K-Means algorithm is more accurate than EM and DBSCAN because K-Means takes less time and correctly classified the ratio. Moreover, there is no missing value which creates difficulty in clustering technique. Weka is more suitable tool for data mining applications because there is no need of deep knowledge of algorithms for Weka. This paper shows only clustering operations and the comparison of clustering algorithms of Weka.

REFERENCES

[1] A.Laud and R.Shah,"An Enhance Approach For Cluster Analysis For Large datasets",*(IJERT)* vol.2 Issue 6,June 2013.

[2] A.Katal, M.Wazid and R H Goudar,"Big Data: Issues, Challenges, Tools and Good Prctices", *IEEE (2013)* pp. 404-409.

[3] A. Kumar Jain and Satyam Maheshwari,"Survey of Recent Clustering Techniques in Data Mining",(*IJCSMR)* vol.1, issue 1,pp.72-78. Augut 2012.

[4] Han, Wen Yonggang and Xuelong Li," Towards Scalable systems for big data analytics: A Technology Tutorial". IEEE Access. Vol. 2,pp 652- 687, July 2014.

[5] N.Bhan and D. Mehrotra,"Comparative Study of EM and K- Means Clustering Techniques in Weka interface".(*IJATER)* Vol.3, Issue 4, July 2013.

[6] N.Choudhary and P.Singh,"Cloud Computing and Big Data Analytics", (*IJERT)* vol. 2 Issue 12, pp 2700-2704, December 2013.

[7] N. Sharma and A. Partap Singh,"A Comparative Study of Data Clustering Techniques",*(IJERT)* vol.2 Issue 6 , June 2013.

[8] P. Chandarana and M. Vijayalakshmi,"Big Data Analytics Frameworks"IEEE. pp 430-434, 2014

[9] R.Jindal, S. D. Sharma and M.Sharma,"Survey Paper on Different Techniques of Measuring Efficiency of Clustering", *(IJERT)* vol.2 , Issue 6. pp 1373-1376, June 2013.

[10] S.Sarangi and V. Jaglan,"*Performance Comparison of Mchine Learning Algorithms on integration of Clustering and Classification techniques.(IJETCAS,2013)* pp.251-257.

[11] S. Siddiqui and D. Gupta ,"Big Data Analytics: A Srvey", *International Journal of Emerging Research in Management & Technology*, vol. 3, issue 7. pp 117-123, July 2014.

[12] S. Singhal and M.Jena, "*A* Study of Weka Tool for data Processing, Classification and Clustering",*IJERT*,vol.2, Issue 6 pp.250-253. August 2013.