ISSN NO. 2320-5407

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/4661
DOI URL: http://dx.doi.org/10.21474/IJAR01/4661

RESEARCH ARTICLE

PERFORMANCE EVALUATION OF WEKA CLUSTERING ALGORITHMS ON LARGE DATASETS

Anju Parmar, Divya Chauhan and Dr. K.L. Bansal.
Department Of Computer Science, Himachal Pradesh University, Shimla, India.

Manuscript Info

Manuscript History

Received: 21 April 2017
Final Accepted: 25 May 2017
Published: June 2017

Key words:-
Data Mining, Clustering algorithms, Weka, Big Data

Abstract

Data Mining is the process of analyzing data from different viewpoints and summarizing it into useful information. By using Data mining tool, the user can analyze data from different dimensions or angles, categorize it, and process the relations recognized. Clustering is one of most widely used techniques in data mining. Clustering is the process of grouping data by finding similarities between data based on their features. Similar Items are grouped in one cluster and dissimilar in another.

In this paper, a comparative study of nine clustering algorithms is performed. For comparison three datasets are used. The main objective of the study is to observe the effect of size of different dataset on data mining tool and clustering algorithms. The dataset chosen for comparison are diverse in terms of number of attributes and instances. All the nine algorithms are compared according to the factors such as size of the dataset, number of clusters and time taken to form clusters. For performing comparison, data mining tool Weka is used. Also the performance of Weka for handling large datasets is analyzed.

Introduction:-

With the development of information technology, the large amount of databases and huge data in various areas has been generated. The research in the areas of databases and information technology has given rise to a new approach to store and manipulate this precious data which can be used for further decision making. Data mining is a process of extracting useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis [2]. Data mining is a powerful concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored at various sources such as data warehouse, World Wide Web, external sources, interesting pattern that is easy to understand, unknown, valid, potential useful. Data mining is a sorting technique which is mainly used to extract hidden patterns and data from large databases. The goals of data mining are fast retrieval of data or information, knowledge discovery from the databases, to identify hidden patterns and those patterns which are previously not explored, to reduce the level of complexity, time saving, etc [3]. Sometimes data mining is treated as knowledge in the databases (KDD) [1]. The ultimate goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge dataset of data and interpret them to useful knowledge and information [2].

Corresponding Author:- Anju Parmar.
Address**:-** Department Of Computer Science, Himachal Pradesh University, Shimla, India.

We are living in the 21st century, the digital age. Every day, people store large information and it is representing as data for further analysis and management. The amount of data in our world has been growing regularly. Company takes a million of bytes of data related to their consumers, dealers and their related operation, and trillions of networking sensors are used to establish (set) in the real world in storage spaces or devices like automobiles and mobile phone, for creating, and sensing and communicating data it needs smart phones or social networking sites or other devises that will used to maintain data exponential expansion. “Big data” can be defined as a large datasets whose size is so (too) large for the database software tools, that it is not easily capable to store data, capture data and handle data. Therefore, big data analytics can be good to impact business change and improve results, by applying advanced analytic techniques on big data, and discovering hidden insights and helpful information [4].

Clustering is considered as one of the most important issues in data mining and machine learning. Clustering is a task of discovering homogenous groups of the studied objects. Many researchers have a significant interest in developing clustering algorithms. The most important issue in clustering is that we do not have prior knowledge about the given data. Moreover, the (input) parameters choice like number, nearest neighbours, Kn, amount of clusters and some other factor in algorithms create the clustering is challenging process. One of the very effective ways of dealing with these data is to categorize or assemble that data into a set of classes. Nowadays clustering methods emerge as another influential meta-learning tool for correctly analysing the big volume of data created by some new applications. The Big data can also be define as the datasets which having big dimensions or they are in large variety as well as in large velocity so it is very hard to hold that datasets by applying conventional techniques and tools. Just because of some fast expansion of information, we require solutions that efficiently handles and extract knowledge from these datasets. Therefore analysis of clustering techniques with their some different available classes with big datasets provides an effective and useful conclusion [4].

In this paper, section 1 gives the introduction about data mining and section 2 gives the introduction about clustering. Section 3 describes the literature review related to the study. Section 4 describes the comparative analysis of the clustering algorithms using weka and section 5 describes the conclusion.

Clustering:-

Clustering is a process of dividing the data elements into groups which are similar to each other. Each group is referred to as a cluster that consists of objects that are similar to one another and dissimilar to objects of another group. It is a technique that recognizes different patterns of data. Good clustering techniques will produce a good or a high quality cluster [5]. It helps to analyse large volumes of data visually thus assists in making quick decisions. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups and the groups are called as clusters. [6]. The Clustering techniques can be classified into the following categories:

A. Partition Clustering
B. Hierarchical Clustering
C. Density Based Clustering
D. Model Based Clustering
E. Grid Based Clustering

Partition Based Clustering:-

All objects are considered initially as a single cluster. The objects are divided into no of partitions by iteratively locating the points between the partitions. The partitioning algorithms like K-means, K-medoids (PAM, CLARA, CLARANS, and FCM) and K-modes. Partition based algorithms can found clusters of Non-convex shapes [7].

Hierarchical-Based Clustering:-

There are two techniques of performing hierarchical based clustering Agglomerative (top-bottom) and Divisive (Bottom-up). In Agglomerative approach, initially one object is selected and successively merges the neighbour object is selected and successively merges the neighbour objects based on the distance as minimum, maximum and average. The process is continuous until a desired cluster is formed. The Divisive approach deals with the set of objects as a single cluster and divides the cluster into further clusters until desired number of clusters are formed. BRICH CURE, ROCK, Chameleon is some of the clustering algorithm in which clusters of non convex, Arbitrary Hyper rectangular are formed[7].

Density-Based Clustering:-

Data objects are categorized into core points, border points and noise points. All the core points are connected together based on the densities to form cluster. Arbitrary shaped clusters are formed by various clustering algorithms such as DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU and SUBCLU [7].

Grid-Based Clustering:-

Grid based algorithm partitions the data set into no number of cells to form a grid structure. Clusters are formed based on the grid structure. To form clusters Grid algorithm uses subspace and hierarchical clustering techniques. STING, CLIQUE, Wave cluster, BANG, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, FC and STIRR. Compare to all Clustering algorithms Grid algorithms are very fast processing algorithms. Uniform grid algorithms are not sufficient to form desired clusters. To overcome these problem Adaptive grid algorithms such as MAFIA and AMR Arbitrary shaped clusters are formed by the grid cells [7].

Model-Based Clustering:-

Set of data points are connected together based on various strategies like statistical methods, conceptual methods, and robust clustering methods. There are two approaches for model based algorithms one is neural network approach and another one is statistical approach. Algorithms such as EM, COBWEB, CLASSIT, SOM, and SLINK are well known Model based clustering algorithms [7].

Literature Review:-

A number of surveys on clustering are present in the literature. Some researchers have proposed new algorithms for clustering. Other researchers have improved the existing clustering algorithms overcoming the drawbacks of the algorithms while some have performed the comparative study of the various clustering algorithms.

Narendra Sharma et al. [8] studied various clustering algorithms and compared the various clustering algorithms of the WEKA tool for data mining. This repository used provides the past project data for analysis. The aim of comparison is to show which algorithms is most suited for the users. The paper provides a detailed introduction of the WEKA clustering algorithms with the advantages and disadvantages of each algorithm.

Sunita B Aher, LOBO L.M.R.J [9] surveyed the application of data mining in the field of education and also presented the result analysis using Weka tool. They discovered classification using ZeroR algorithm. Also they clustered the student into group using DBSCAN-clustering algorithm. Finally, noisy data was detected.

Garima et al. [10] provides a comparative analysis based upon the similarity criterion and the complexity. This paper discusses the various clustering algorithms like partitioned clustering, hierarchical clustering, density based clustering, grid based clustering and their time and space complexities.

Prakash Singh, Aarohi Surya [11] presented the comparison of 9 clustering algorithms in terms of their execution time, number of iterations, sum of squared error and log likelihood using Weka.
Sapna Jain et al. [12] have evaluated the performance of the K-Means clustering using Weka tool. Rupali Patil et al. [13] have also evaluated the performance of K-Means Clustering algorithm for multiple dimensions using Weka tool. The various dimensions considered are Time taken to build the model, number of attributes, number of iterations, number of clusters and error rate.

Mugdha Jain, Chakradhar Verma [14] has proposed a new approximate algorithm based on the K-means algorithm. They tried to improve the efficiency of the K-Means algorithm when dealing with big data. The algorithm proposed by them overcomes the drawbacks of the K-Means algorithm of uncertain number of iterations by fixing the number of iterations, without losing precision.

Olga Kurasova et al. [15] overviewed the various methods and technologies which can be used for big data clustering. They paid the great attention to the K-Means clustering and its modifications and are implemented in innovative technologies for big data analysis.

Bhagyashri S. Gandhi, Leena A. Deshpande [16] combined approach of supervised and unsupervised learning with an ensemble approach in distributed computing environment is proposed for reducing high dimensional data to improve the overall efficiency in mining big data. Aris-Kyriakos Koliopoulos et al. [19] discusses DistributedWekaSpark in their paper.

Venkateswara Reddy Eluri et al. [18] compared the various clustering algorithms on big data sets using Apache Mahout. In this paper three clustering algorithms are described: K-means, Fuzzy K-Means (FKM) and Canopy clustering. Also, they have underlined the clustering algorithms which are best suited for big data.

Implementation, Result and Discussions:-

In this paper, for the comparison of various clustering algorithms, Weka tool is used. Weka is a data mining tool which consists of a set of machine learning algorithms. Weka consists of tools for preprocessing, classification, regression, clustering, association rules, and visualization of the data. Three datasets of different sizes are chosen for comparing the algorithms. Each dataset is described by the number of attributes and instances. Table 1 shows the name of the dataset, number of attributes, and number of instances in each dataset.

Table 1:- Description of the Datasets.

Name of the dataset	Number of attributes	Number of instances
Gesture_Phase_dataset	33	1069
Semeion Dataset	257	1116
Convex Dataset	785	8000

In this paper for comparing the clustering algorithms, parameters chosen are the time taken to build model and number of clusters formed. The table 2 shows the results for the gesture_phase_dataset..The dataset has 33 attributes and 1069 instances.

Table 2:- Experiment Results for Gesture_Phase_dataset.

Name	*Clustering Type*	*No. Of clusters*	*Cluster distribution*	*Time taken to build the model(sec)*
K-means	*Partitioning based*	*2*	*637 (24%) 432(76%)*	*0.16*

EM	*Model based*	*12*	*159(15%)
			48(4%)
			40(4%)
			31(3%)
			166(16%)
			45(4%)
			111(8%)
			25(2%)
			121(11%)
			132(12%)
			106(10%)*
				776.55

Hierarchical	*Hierarchical based*	*2*	*1068(100%)
			1(0%)*
				7.47

Farthest First	*Partitioning Based*	*2*	*1068(100%)
			1(0%)*
				0.03

LVQ	*Hierarchical based*	*2*	*1064(100%)
			5(0%)*
				13.15

DBSCAN	*Density Based*	*5*	*70(7%)
			217(20%)
			411(39%)
			280(26%)
			81(8%)*
				3.57

cobweb	*Model Based*	*7*	*412(39%)
			217(20%)
			83(8%)
			70(7%)
			287(27%)*
				1.31

MakeDenistyBasedCluster	*Density Based*	*2*	*435(41%)*	*0.16*

			634(59%)	

CascadeSimpleKMeans	*Partitioning based*	*2*	*462(43%)*
			607(57%)
				22.71

The table 3 shows the results for the Semeion dataset..The dataset has 257 attributes and 1116 instances.

Table 3:- Experiment Results for Semeion dataset.

Name	***Clustering Type***	***No. Of clusters***	***Cluster distribution***	***Time taken to build the model(sec)***

K-means	*Partitioning based*	*2*	*792(71%)*
			324(29%)
				1.15

EM	*Model based*	*8*	*133(12%)*
			34(8%)
			138(12%)
			126(11%)
			114(10%)
			121(11%)
			137(16%)
			213(19%)
				2191.8

Hierarchical	*Hierarchical based*	*2*	*1114(100%)*
			2(0%)
				18.1

FarthestFirst	*Partitioning Based*	*2*	*844(76%)*
			272(24%)
				0.11

LVQ	*Hierarchical based*	*2*	*547(49%)*
			569(51%)
				456.94

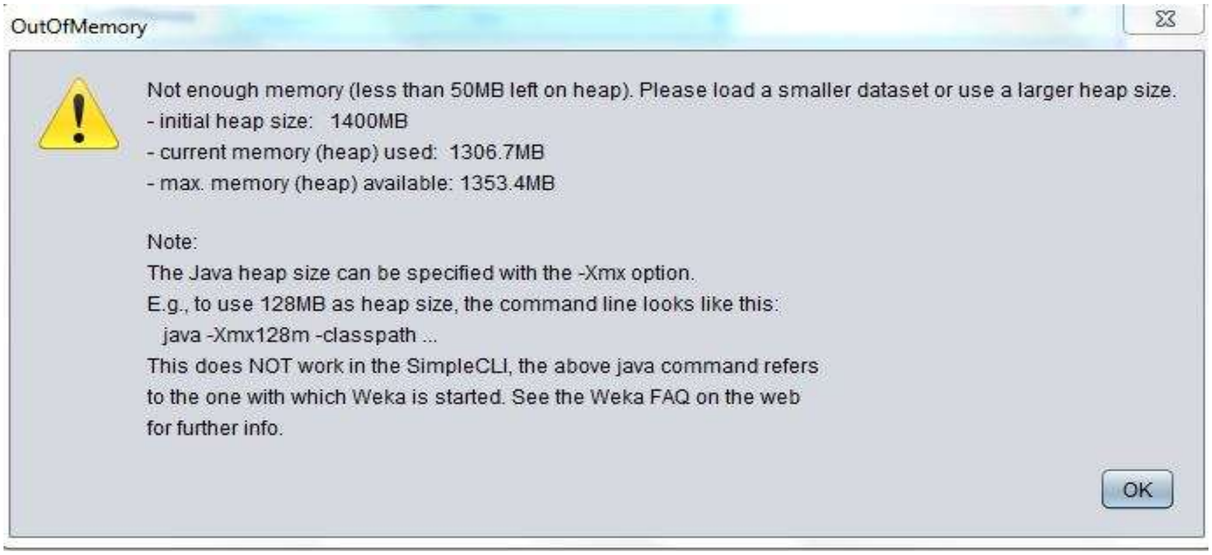
DBSCAN	*Density Based*	*0*	*noise*	*25.54*

cobweb	*Model Based*	*11*	*118(11%)*
			117(10%)
			109(10%)
			106(9%)
			118(11%)
			107(10%)
			105(9%)
			122(11%)
			105(9%)
				11.12

MakeDenistyBasedCluster	*Density Based*	*2*	*810(73%)*
			306(27%)
				1.26

CascadeSimpleKMeans	*Partitioning based*	*2*	*552(49%)*
			564(51%)
				250.01

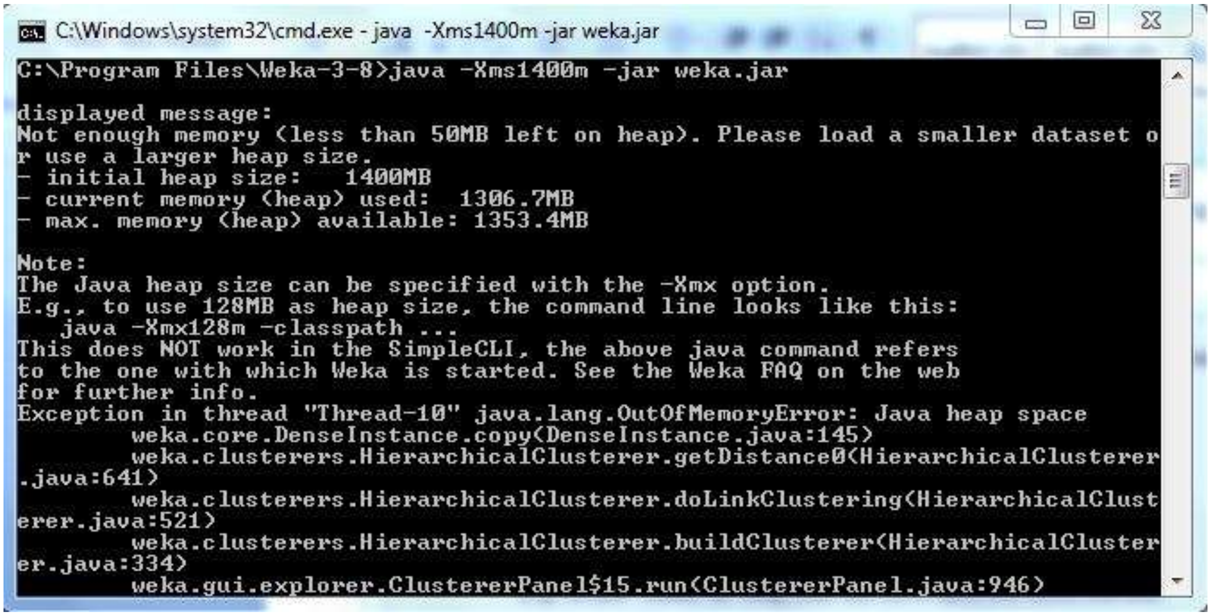
The Figure 1 and Figure 2 show the experimental results for the convex dataset. The dataset has 785 attributes and 8000 instances.



```
Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heap size.
- initial heap size:   1400MB
- current memory (heap) used:  1306.7MB
- max. memory (heap) available: 1353.4MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
   java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.
```

Figure 1:- Error message for the convex Dataset.



```
C:\Program Files\Weka-3-8>java -Xms1400m -jar weka.jar

displayed message:
Not enough memory (less than 50MB left on heap). Please load a smaller dataset o
r use a larger heap size.
- initial heap size:   1400MB
- current memory (heap) used:  1306.7MB
- max. memory (heap) available: 1353.4MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
   java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.
Exception in thread "Thread-10" java.lang.OutOfMemoryError: Java heap space
        weka.core.DenseInstance.copy(DenseInstance.java:145)
        weka.clusterers.HierarchicalClusterer.getDistance0(HierarchicalClusterer
.java:641)
        weka.clusterers.HierarchicalClusterer.doLinkClustering(HierarchicalClust
erer.java:521)
        weka.clusterers.HierarchicalClusterer.buildClusterer(HierarchicalCluster
er.java:334)
        weka.gui.explorer.ClustererPanel$15.run(ClustererPanel.java:946)
```

Figure 2:- Error message at command prompt for the convex dataset.

The results related to time taken to build model and number of clusters formed have been noted for the datasets. Figure 3 shows the graphical representation of the results for the time taken to build model and Figure 4 shows the graphical representation of the results for the number of clusters formed.

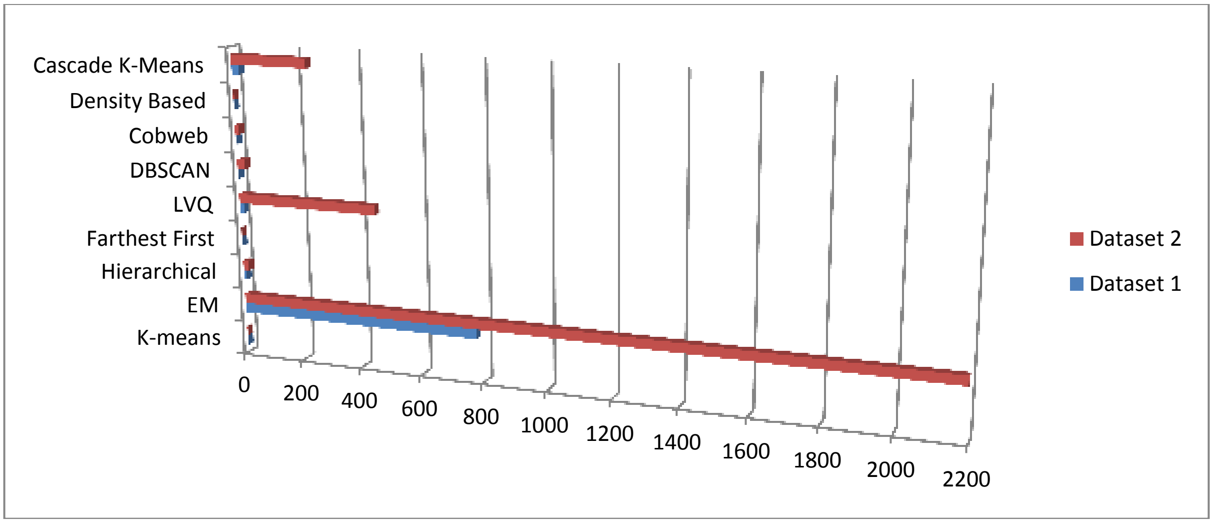


Figure 3:- Graphical representation of time taken to build model for different datasets

It has been observed that the Farthest First clustering algorithm took least time in forming clusters whereas Expectation Maximization took maximum time for forming the clusters and the time taken to form clusters increases with the increase in the size of the dataset.

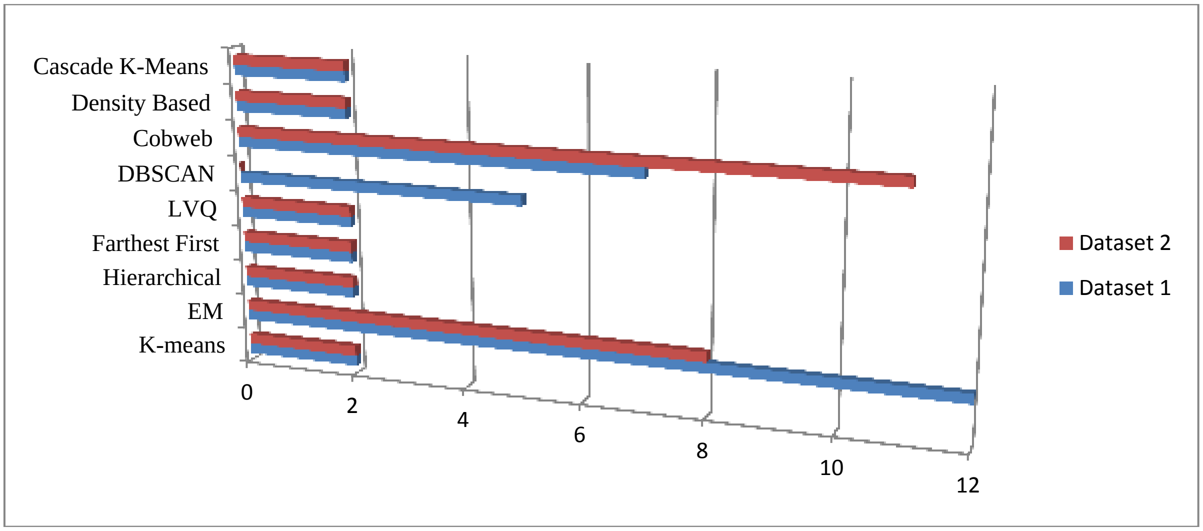


Figure 4:- Graphical representation of number of clusters formed for different datasets

It has been observed that the K-means, Hierarchical Farthest First, LVQ, Density based and cascade k-means formed the same number of clusters. Number of clusters formed by EM and cobweb were quite high.

Conclusion:-

Data mining is the computer assisted process of digging through and analyzing enormous set of data and then extracting the meaningful data. The overall goal of the data mining process is to extract information from large dataset and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that the objects in the same group are more similar to one another than to those in other groups (clusters).In this paper, data mining tool Weka is used to perform clustering. Different datasets have been taken to analyze the performance of the data mining tool and the clustering algorithm. The dataset varies in number of attributes and instances. Nine clustering algorithms have been compared on the parameters time taken to build model, number of clusters formed. It has been observed that k-means, Farthest First, Hierarchical, LVQ, Density based and cascade K-means formed the same number of clusters. In terms of time taken, Farthest First clustering algorithm took least time for forming the clusters while Expectation Maximization took maximum time for forming the clusters. Also, with the increase in the size of the dataset the time taken to form clusters increases.

It is also observed that Weka cannot handle very large datasets because it supports only sequential single node execution. Hence, the size of datasets and processing tasks that Weka can handle within its existing environment is limited both by the amount of memory in a single node and by sequential execution. For handling large datasets, parallel and distributed computing- based systems and technologies are required. Hadoop based technologies and libraries are the most popular solution or handling large datasets.

References:-

1. J. Han and M. Kamber,*"Data Mining, Concepts and Techniques",*Second Edition, Morgan Kaufman Publishers
2. Kalyani M Raval, "Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering* , Volume 2, Issue 10, October 2012
3. Smita, Priti Sharma, "Use of Data Mining in Various Field: A Survey Paper", *IOSR Journal of Computer Engineering (IOSR-JCE*) ,Volume 16, Issue 3, Ver. V (May-Jun. 2014)
4. Prachi Surwade, Prof. Satish S. Banait, "A Survey on Clustering Techniques For Mining Big Data*", International Journal Of Advanced Research in Science And Management*, Volume 2, Issue 2, Feburary 2016
5. Miss. Harshada S. Deshmukh, Prof. P. L. Ramteke, "COMPARING THE TECHNIQUES OF CLUSTER ANALYSIS FOR BIG DATA",*International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Volume 4 Issue 12, December 2015
6. KeshavSanse, Meena Sharma*, "*Clustering methods for Big data analysis*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Volume 4 Issue 3, March 2015
7. T. Sajana, C. M. Sheela Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining ",*Indian Journal of Science and Technology*, Vol 9(3), DOI:10.17485/ijst/2016/v9i3/75971, January 2016
8. Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", *International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459*, Volume 2, Issue 5, May 2012)
9. Sunita B Aher, Mr. LOBO L.M.R.J, "Data Mining in Educational System using WEKA*", International Conference on Emerging Technology Trends (ICETT) 2011 Proceedings published by International Journal of Computer Applications® (IJCA)*
10. Garima, Hina Gulati, P.K.Singh, "Clustering Techniques in Data Mining: A Comparison*",2nd International Conference on Computing for Sustainable Global Development, 2015*
11. Prakash Singh, Aarohi Surya, "PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS IN DATA MINING IN WEKA", *International Journal of Advances in Engineering & Technology*, Jan., 2015
12. Sapna Jain, M AfsharAalam, 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
13. Rupali Patil, Shyam Deshmukh, K Rajeswari, "Analysis of Simple K-Means with Multiple Dimensions using WEKA", *International Journal of Computer Applications (0975 – 8887)* ,Volume 110 – No. 1, January 2015
14. Mugdha Jain, Chakradhar Verma, "Adapting k-means for Clustering in Big Data", *International Journal of Computer Applications (0975 – 8887)* ,Volume 101– No.1, September 2014
15. Olga Kurasova, VirginijusMarcinkevicius, Viktor Medvedev, AurimasRapecka, and Pavel Stefanovic , "Strategies for Big Data Clustering", *2014 IEEE 26th International Conference on Tools with Artificial Intelligence* , DOI74110.1109/ICTAI.2014.115
16. Bhagyashri S. Gandhi, Leena A. Deshpande, *"The Survey on Approaches to Efficient Clustering and Classification Analysis of Big Data",International Journal of Engineering Trends and Technology (IJETT) –* Volume 36 Number 1- June 2016
17. Dr.Venkateswara Reddy Eluri, MS. Amina Salim Mohd AL-Jabri, Dr.M.RAMESH, Dr. Mare Jane, "A Comparative Study of Various Clustering Techniques on Big Data Sets using Apache Mahout", *2016 3rd MEC International Conference on Big Data and Smart City*
18. Aris-Kyriakos Koliopoulos, Paraskevas Yiapanis, FiratTekiner, Goran Nenadic, John Keane, "A Parallel Distributed Weka Framework for Big Data Mining using Spark*", IEEE International Congress on Big Data,2015.*