COMPARITIVE ANALYSIS OF CLUSTERING ALGORITHMS

Rashmi Sindhu[1], Rainu Nandal[2]

Computer Science and Engineering,
University Institute of Engineering and Technology,
Maharshi Dayanand University, Rohtak, Haryana, India

Abstract: - K-means is the simplest unsupervised learning algorithm. In this clustering algorithm k centers are defined for each cluster. These centers are placed in such a way that each center is away from each other as far as possible. Now, associate each point belonging to a data set to the nearest center. This process of assigning data set is continued until no point is left behind. In the next step k new centroids are re-calculated and a new binding is to be done again between the same data points. This step is repeated until there is no change in the position of data points.
EM algorithm is expectation-maximization algorithm which is an iterative method of find maximum likelihood of parameters in models, which depends on latent variables [1]. This algorithm estimates the current parameters and using these it creates a function for the expectation of log-likelihood to complete expectation step. The next step is the maximization step, in which those parameters are computed which is having maximum log-likelihood [1]. In this paper we will use the tool WEKA for performing clustering on the data points. In weka we will apply two different clustering algorithms on the real data set and try to analyse the two on the basis of no. clusters formed by them. At the end we will conclude the two as observed by the result of execution few data sets.

Keywords:- WEKA, K-means, EM, clustering.

1) Introduction

Data mining is the process of extracting information from a data set and transforms it into an understandable structure for future use. Data stored in a Data Warehouse is of a wide range that may or may not contain the relevant information that the user desire to use. Searching some information from the wider data space involves analysing of large data which probably may result in degrading of its efficiency. Thus to make the best use of the data of the data warehouse ,a tool called the data mining is introduced so that it can provide the required information to its user from the data pool [2].
Role of Clustering in data mining is that it provides the user with the benefit to understand that in a data set how we can do natural grouping of a structure. Use of Clustering can be done either as a tool that can stand alone to go in detail in the distribution of data or as a pre-processing step for any other algorithm.
Clustering is the most essential part of Data mining. Clustering is a technique in which we group different data based on their similarities and dissimilarities from the other. Data point in one data set will posses similar properties as compare to the other. Clusters can differ from each other in terms of shape size and density. The similarity between the objects is calculated by the use of a similarity function. This is mainly useful for organising documents, to improve recovery and support browsing [2].
K-mean algorithm is the most widely used algorithms for cluster analysis of large no. of data sets. K-mean algorithm is a type of Centriod-based clustering technique. K-mean clustering algorithm is efficient in generating clusters for many applications. In K-mean algorithm centroids are used to represent a cluster which is basically the mean of the points of the cluster [2].

Algorithm for K-means Clustering

K-means is the simplest unsupervised learning algorithm. In this clustering algorithm k centers are defined for each cluster. These centers are placed in such a way that each center is away from each other as far as possible. Now, associate each point belonging to a data set to the nearest center. This process of assigning data set is continued until no point is left behind. In the next step k new centroids are re-calculated and a new binding is to be done again between the same data points. This step is repeated until there is no change in the position of data points.

Let X be the set of data points where X = {x_1, x_2, x_3,……, x_p } and V has been considered as the set of centers as V = {v_1, v_2, v_3,……., v_c} [3].

a. Select ‘c’ centers for clusters randomly.

b. Calculate the distance between center of cluster and each data point.

c. After calculating the distance between center and data point assign each data point to that cluster center from which its distance is minimum of all cluster centers.

d. Cluster center is recalculated again using: -

$$V_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i \qquad [3]$$

where 'c_i' represents the number of data points in i^{th} cluster.

e. Now, the distance between each data point and new obtained cluster centers is recalculated again so that the data point is assigned to the nearest cluster.

f. The above 3 steps are repeated again and again until there is no change in the assignment of cluster. As this happens then stop.

Advantages of K-means algorithm

- If we keep value of K small and variables are huge than K-means is more faster algorithm than hierarchical clustering method.
- If the clusters are globular, than K-means produces tighter clusters.

Algorithm of EM clustering

EM algorithm is an iterative method of find maximum likelihood of parameters in models, which depends on latent variables [1]. This algorithm estimates the current parameters and using these it creates a function for the expectation of log-likelihood to complete expectation step. The next step is the maximization step, in which those parameters are computed which is having maximum log-likelihood [1].

Let **X** be a set of observed data and **Z** be a set of unobserved latent data or missing values and $\boldsymbol{\theta}$ is a vector of unknown parameters, along with likelihood function $\mathbf{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = \mathbf{p}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ [4].
Here the Maximum Likelihood Estimate (MLE) of the unknown parameters is calculated by the marginal likelihood of the observed data.

$$\mathbf{L}(\boldsymbol{\theta}; \mathbf{X}) = \mathbf{p}(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} \mathbf{p}(\mathbf{X}, \mathbf{Z}|\Theta) \qquad [4]$$

MLE of the marginal likelihood is calculated in two steps: [4]

i. Expectation step (E step) : - Calculate the expected value of the log likelihood function with respect to the conditional distribution of **Z** given **X** under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$.

$$\mathbf{Q}(\Theta | \Theta^{(t)}) = \mathbf{E}_{Z|X,\theta^{(t)}}[\log \mathbf{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})] \qquad [4]$$

ii. Maximization step (M step) : - Find the parameter that maximizes this quantity :-

$$\boldsymbol{\theta}^{(t+1)} = \arg\max \mathbf{Q}(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) \qquad [4]$$

Advantage of EM algorithm

- EM algorithm is best suited when the data is incomplete, has missing values or having unobserved latent variables. It works by choosing random values for the missing data points.

2) **Implementation**

In this paper we will apply K-means algorithm and EM clustering algorithm on the real data set of car data that contains details about various nominals like buyer, maint, doors nominal etc. [5]. The car data contains total 1728 instances with seven attributes. The total 1728 instances are divided into 6 different data sets. The first data set contains first top 100 instances; the second one contains top 200 instances, third one top 300 instances, fourth top 400 instances and fifth top 500 instances. Remaining instances are 1228 which is considered as one data set. As WEKA is a machine learning tool, it has its own learning capability based on its inbuilt features. Thus these 1228 instances are given as a training set to the WEKA tool through which it learn to operate on given data set. We will execute simple K-means algorithm and Expectation Maximization algorithm on these six data sets, and analyse on the cluster formation of the two algorithms.

1228 instances taken as training data to WEKA

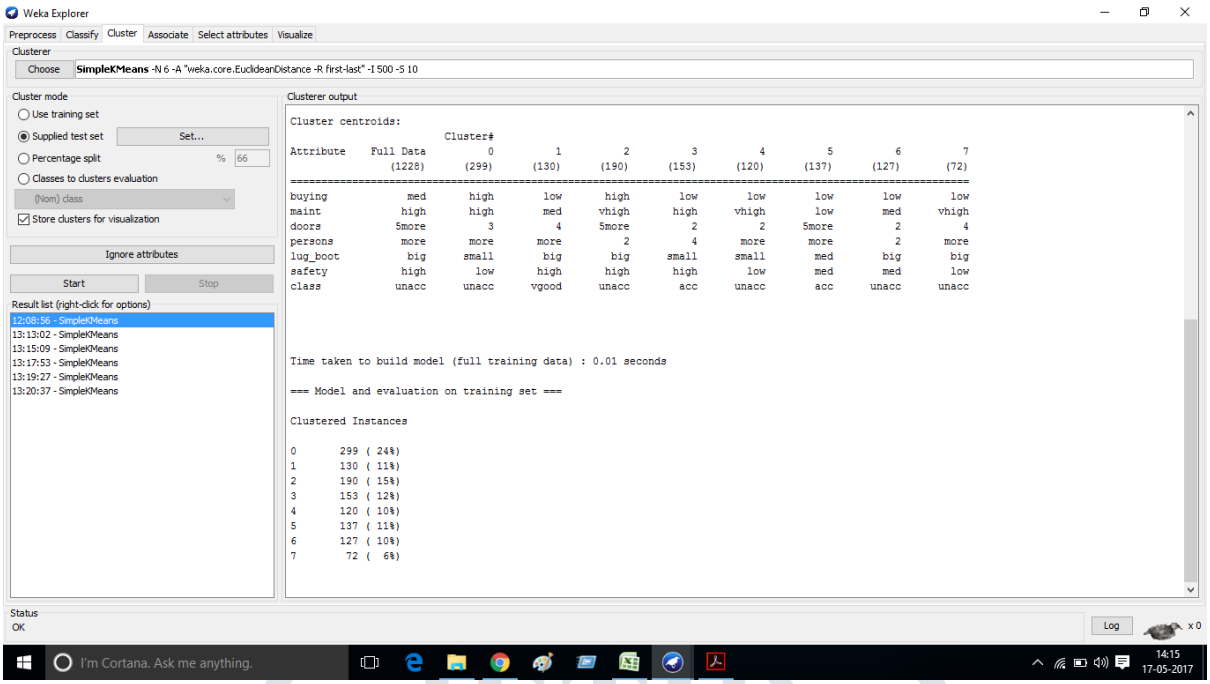


Figure 1:- K-means algorithm on 1228 instances used as training set.

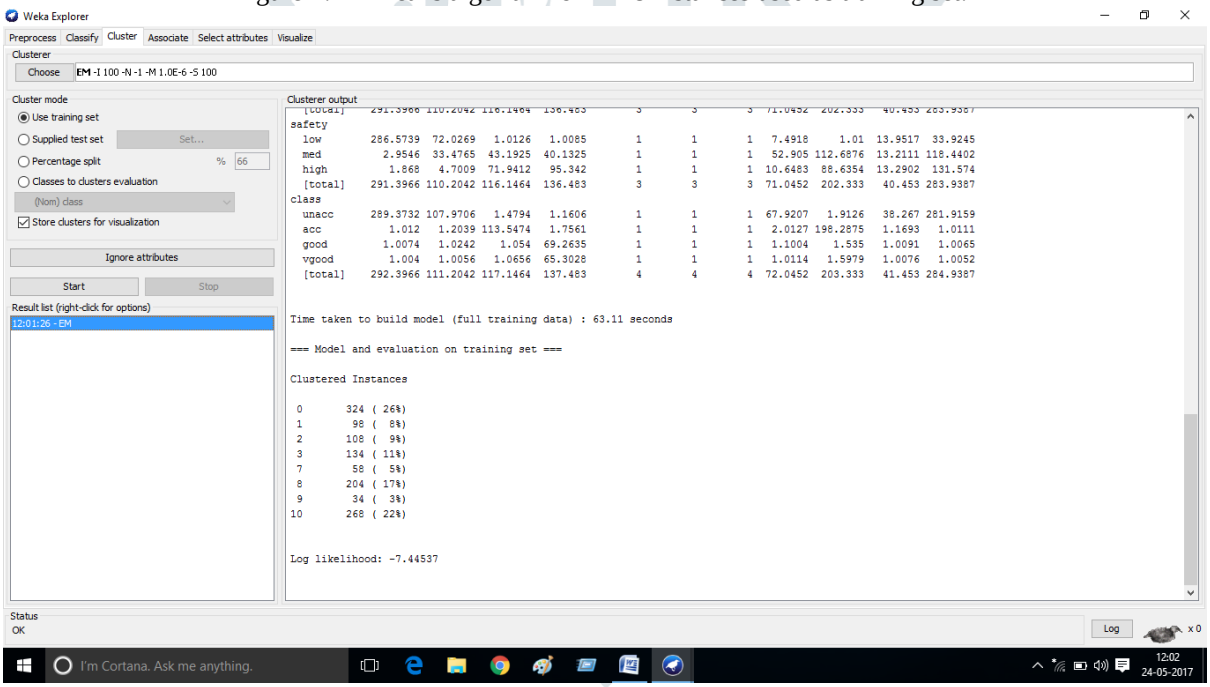


Figure 2: - EM algorithm on 1228 instances used as training set

Top 100 instances supplied as a test set to WEKA

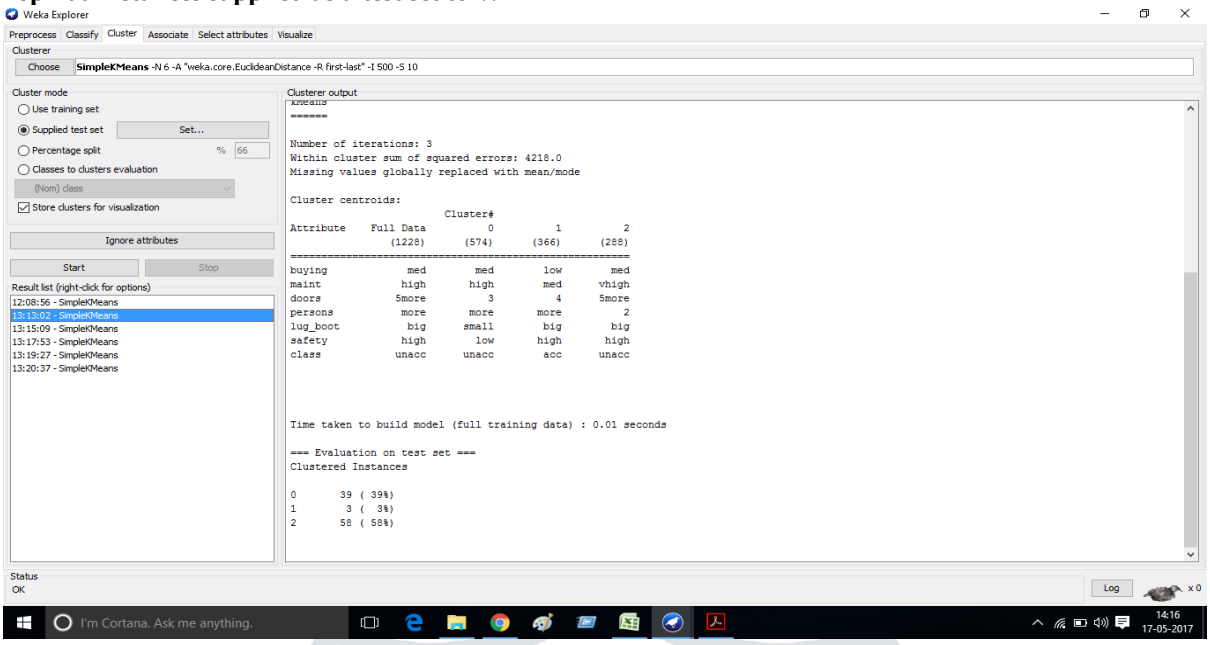


Figure 3: - K-means algorithm on top 100 instances of car data set

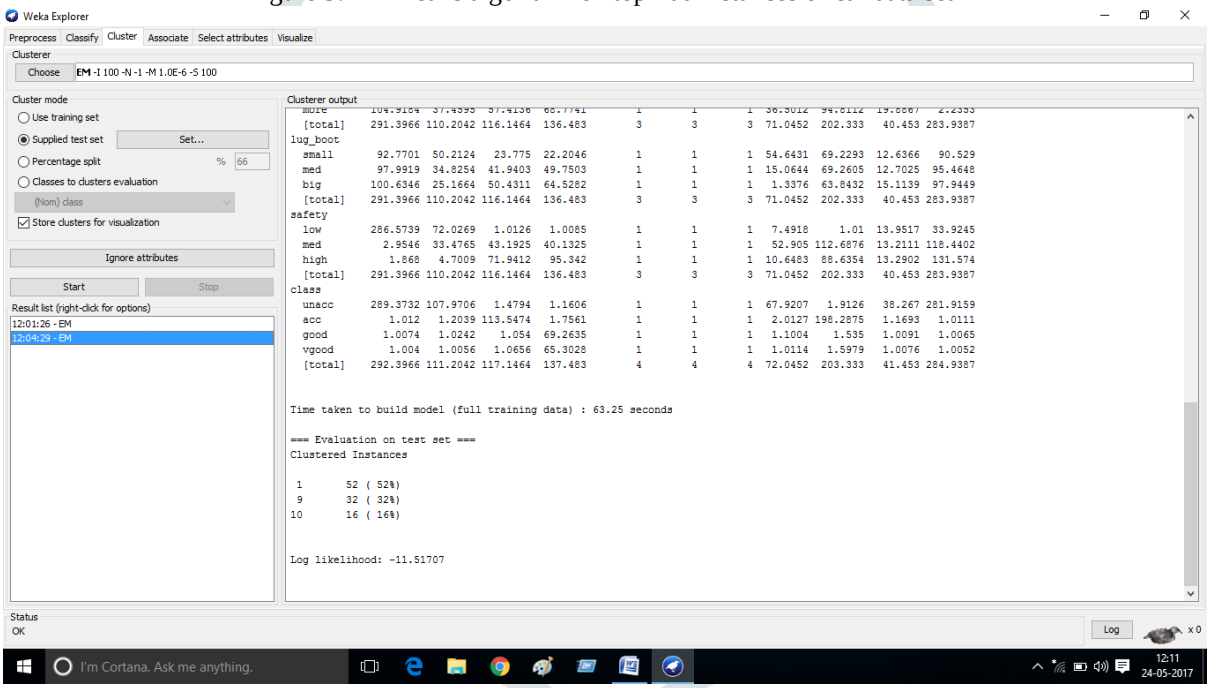


Figure 4: - EM algorithm on top 100 instances of car data set

Top 200 instances supplied as test set to WEKA

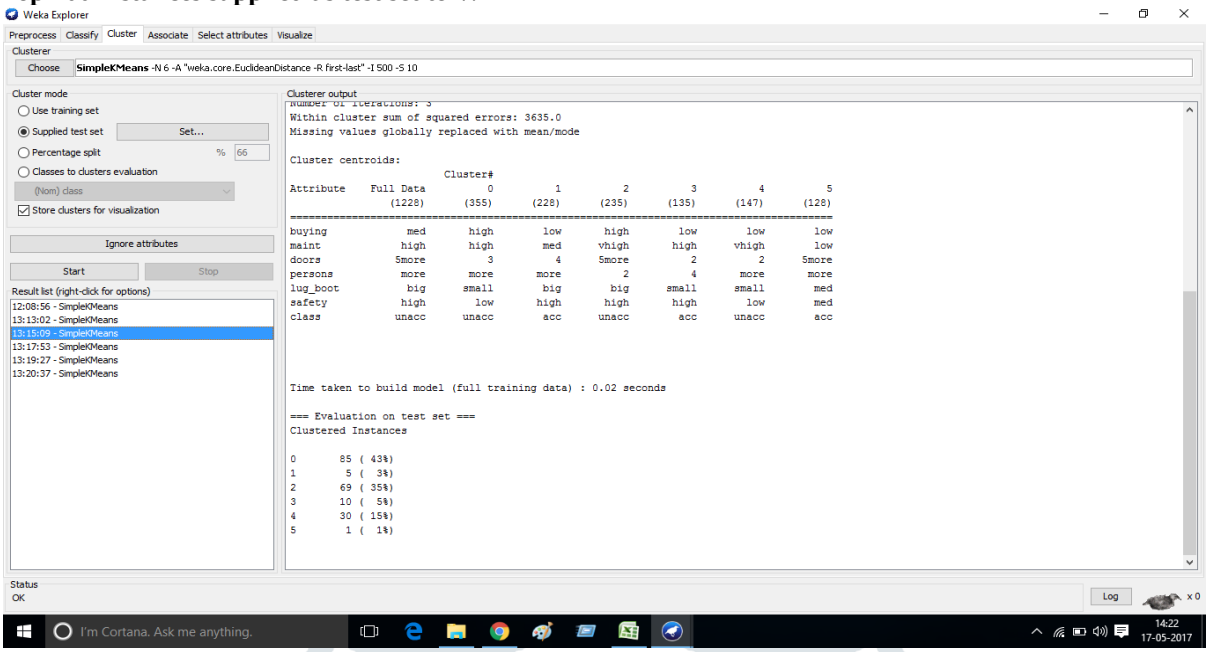


Figure 5:- K-means algorithm on top 200 instances

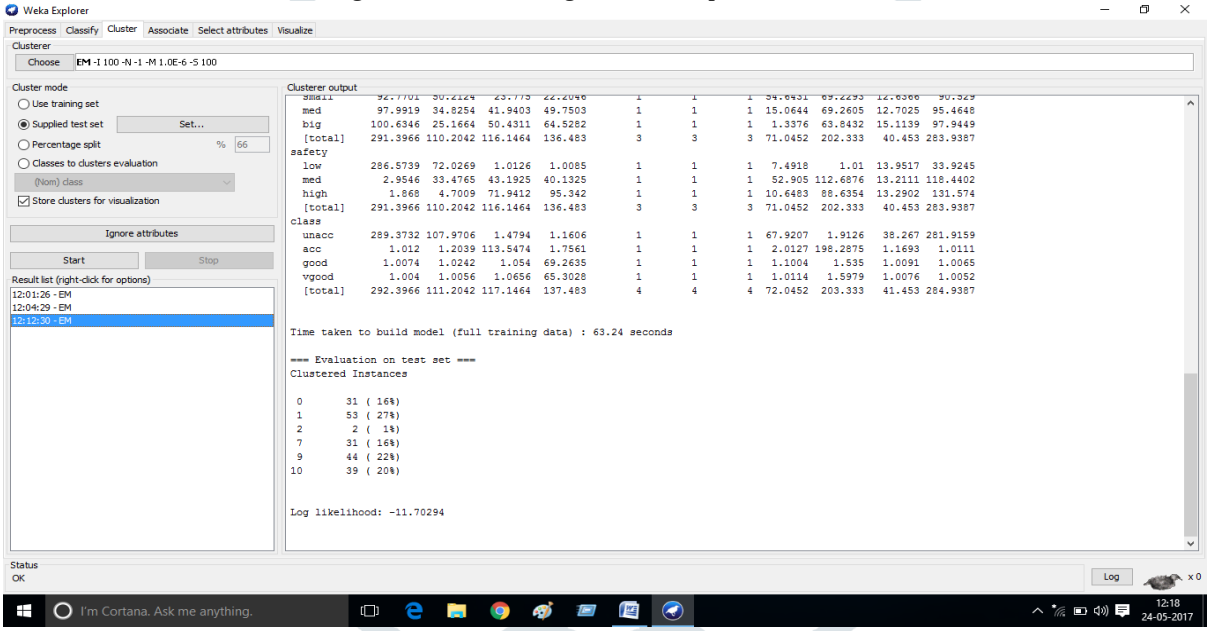


Figure 6: - EM algorithm on top 300 instances

Top 300 instances supplied as test set to WEKA

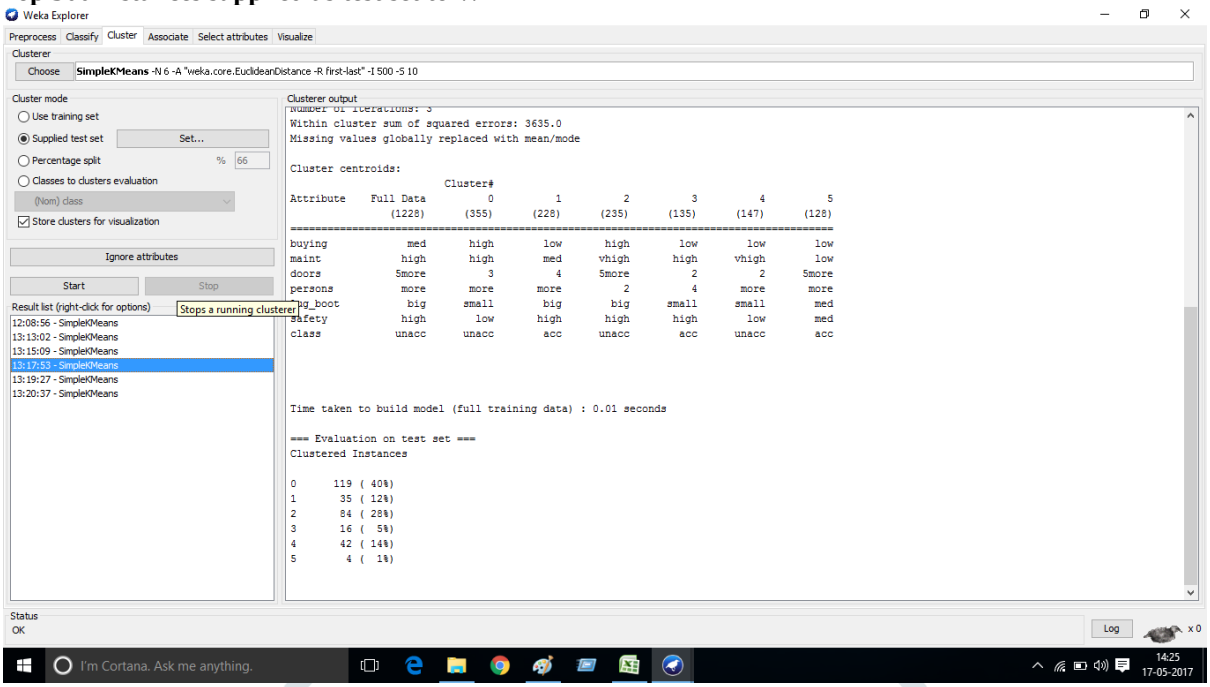


Figure 7: - K-means algorithm on top 300 instances

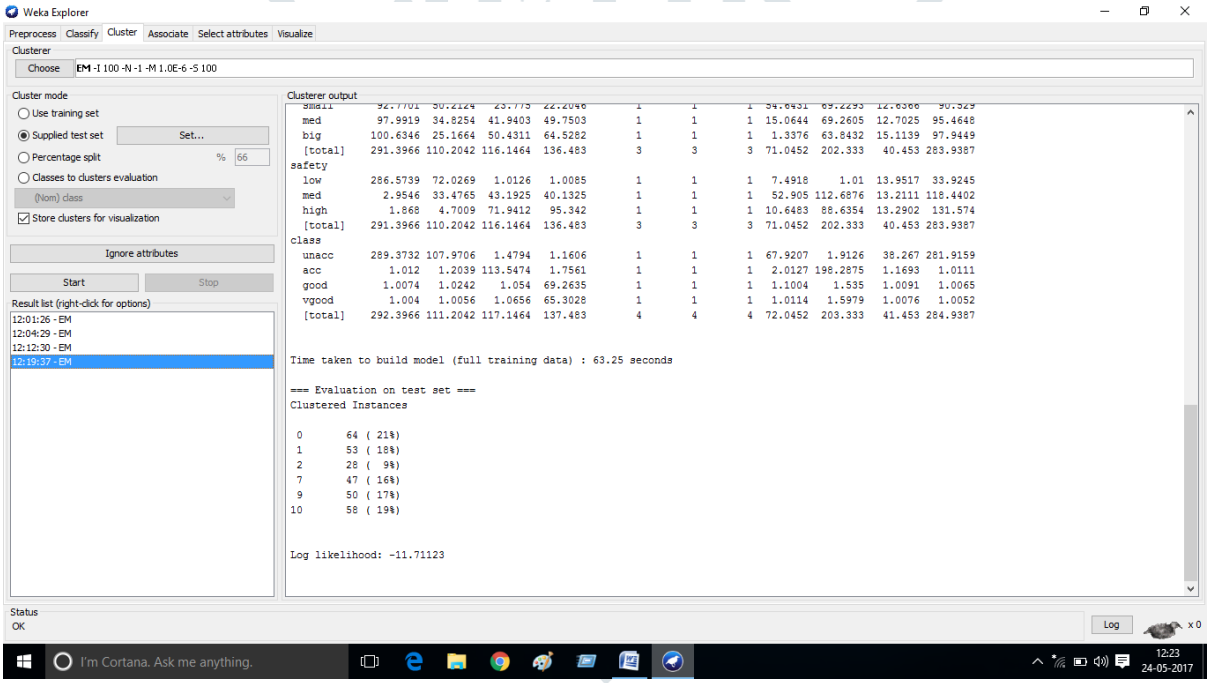


Figure 8: - EM algorithm on top 300 instances

Top 400 instances supplied as test set to WEKA

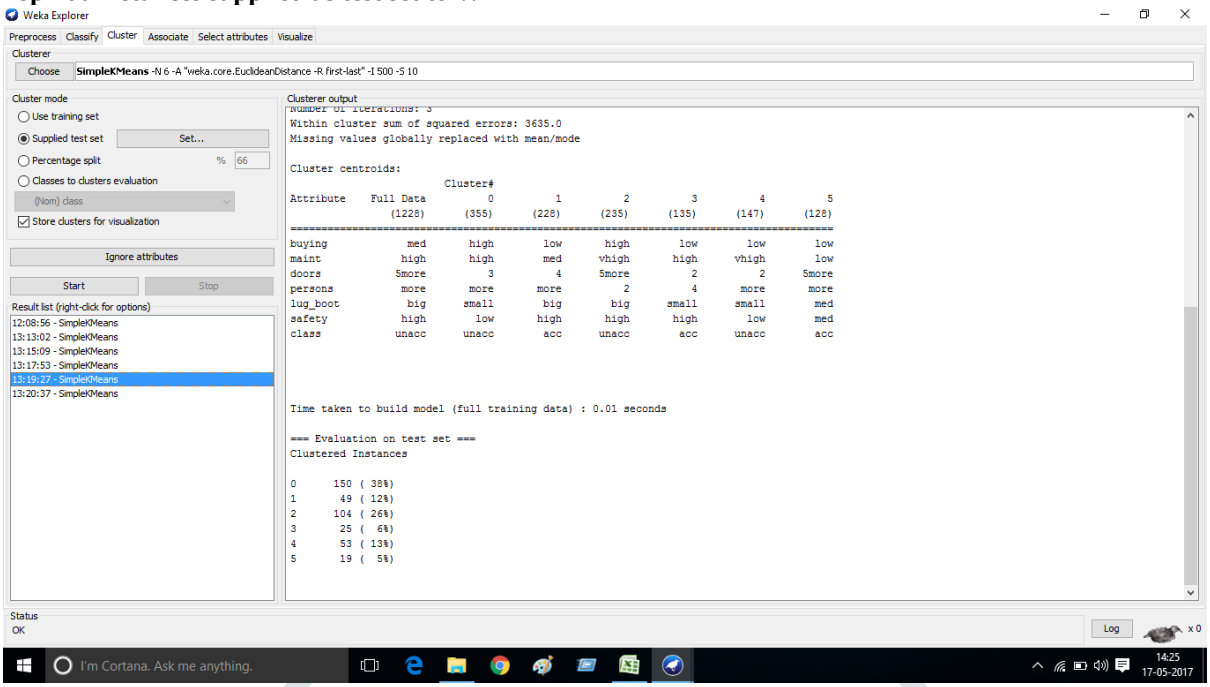


Figure 9: - K-means algorithm on top 400 instances

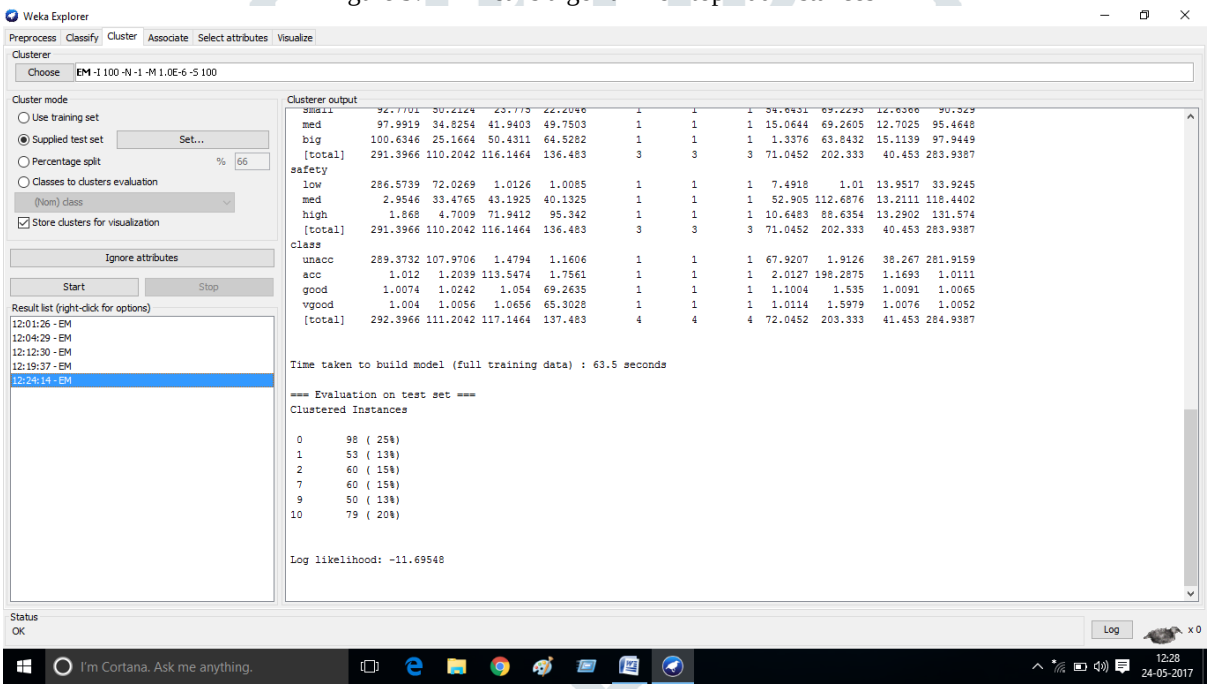


Figure 10:- EM algorithm on top 400 instances

Top 500 instances supplied as test set to WEKA

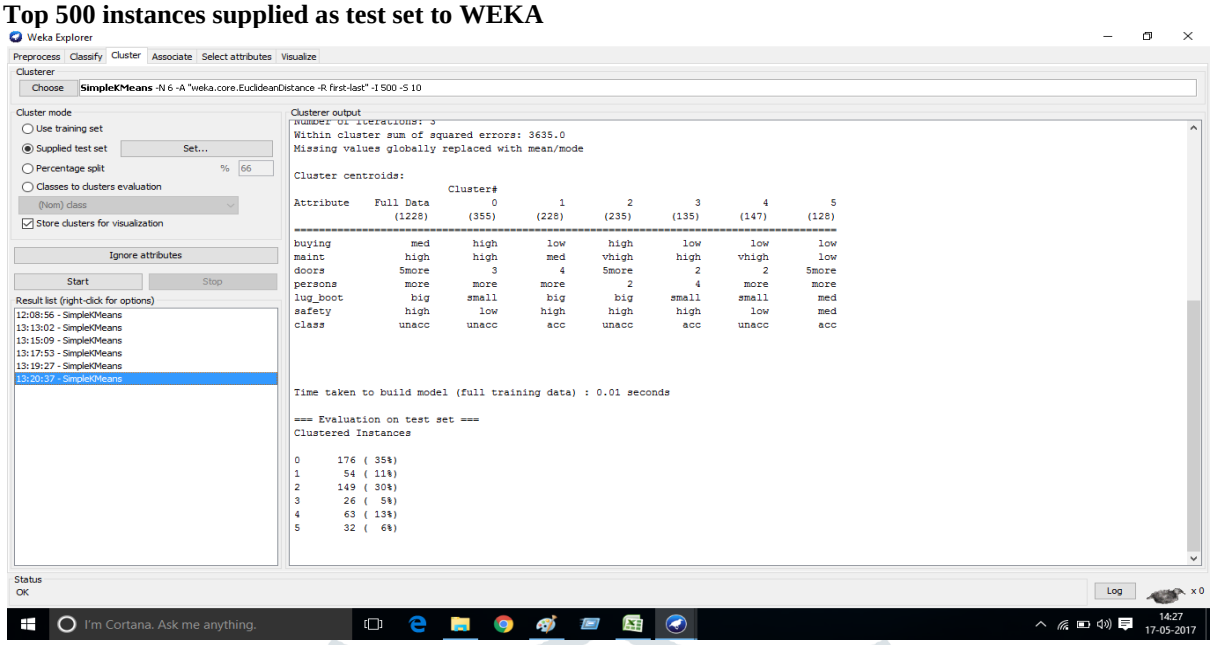


Figure 11: - K-means algorithm on top 500 instances

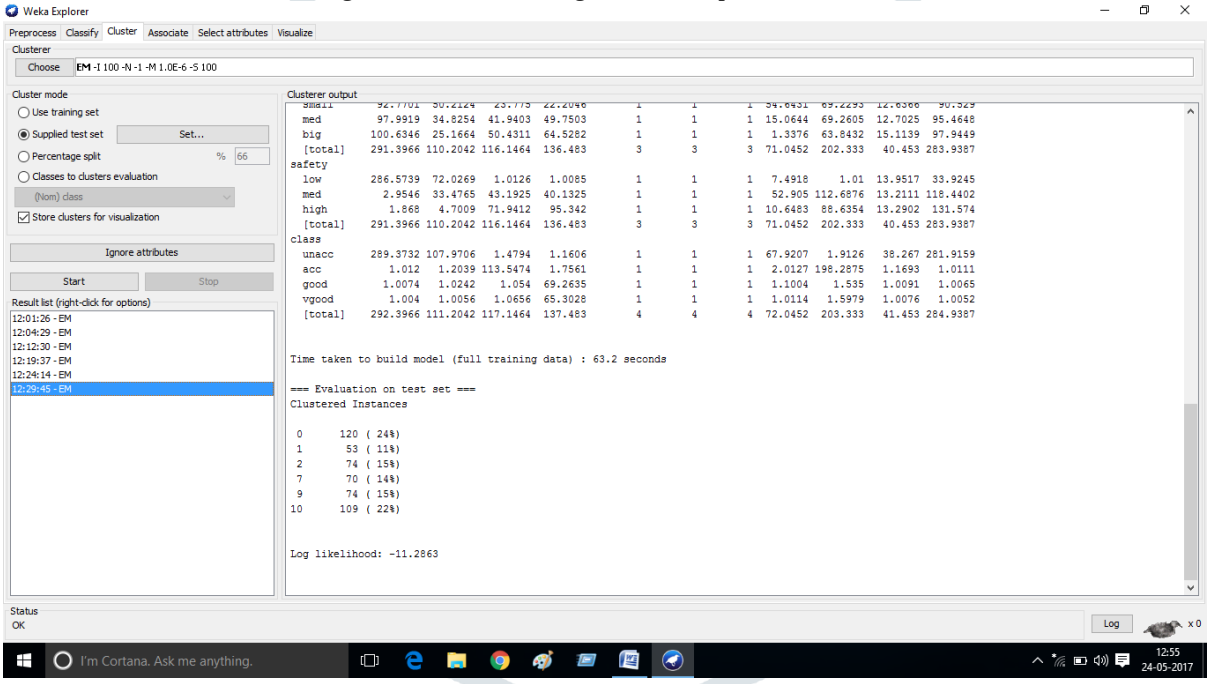


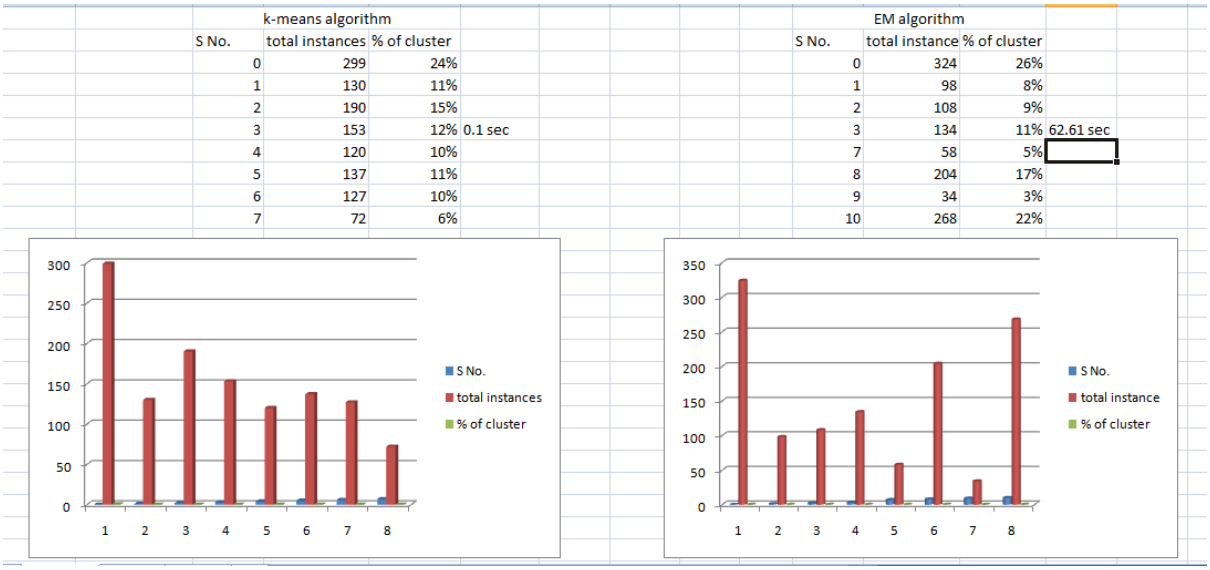
Figure 12: - EM algorithm on top 500 instances

3) Result and Discussion

After implementing simple K mean and Expectation maximization algorithm on the six different data sets we reach to a conclusion that K-means algorithm is much faster algorithm as its execution time is near about 0.1 sec in execution of all the six data sets, wherelse the EM algorithm execution time is approx. 64 sec which is comparatively very large as compared to that of K-means algorithm.

The other point which we conclude form this implementation is that the simple K means algorithm perform the clustering process on data set based on the given no of clusters, which may be specified by the user according to its requirement and choice, by default it is set to 2 in WEKA 3.6. All data sets are centred to the number of clusters specified by the user which will change according to the number of clusters. Where else the EM algorithm works on the basis of cross-validation. This is the mechanism in which as we provide the data set to the EM algorithm it will form number of clusters on its own. We have no need to specify the number of clusters in the EM algorithm. The distribution of data set to form clusters is based on the same. The number of clusters formed in EM algorithm basically based on the number of instances in a particular data set.

Another point that has been noted is that the simple K means algorithm form clusters on the calculation of only mean value of the data values in the data set. But the EM algorithm form clusters by evaluating both mean and standard deviation of the values in data set.



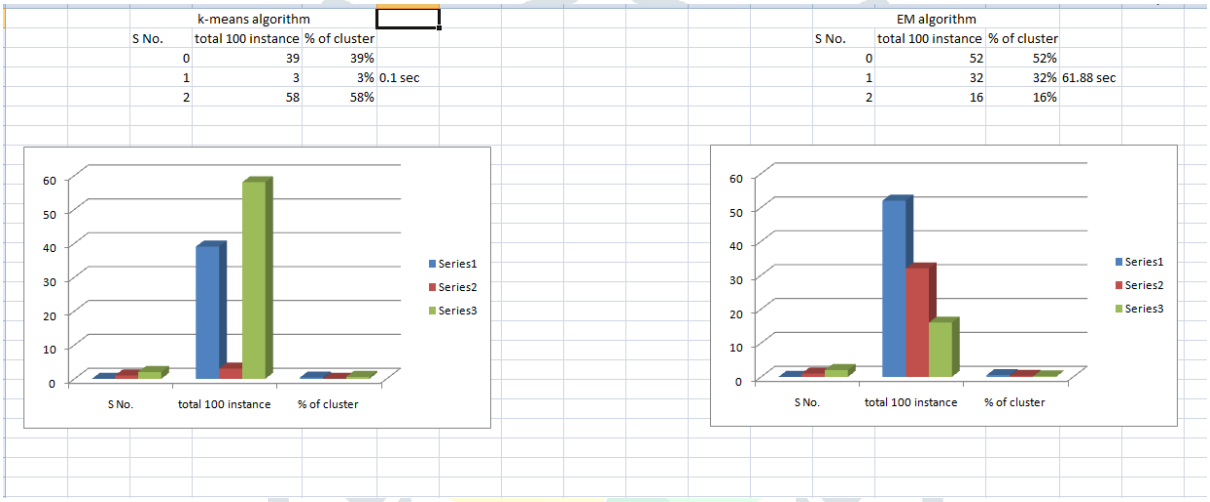
k-means algorithm

S No.	total instances	% of cluster	
0	299	24%	
1	130	11%	
2	190	15%	
3	153	12%	0.1 sec
4	120	10%	
5	137	11%	
6	127	10%	
7	72	6%	

EM algorithm

S No.	total instance	% of cluster	
0	324	26%	
1	98	8%	
2	108	9%	
3	134	11%	62.61 sec
7	58	5%	
8	204	17%	
9	34	3%	
10	268	22%	

Figure 13: - graph showing number of clusters formed by both algorithms when data set has 1228 instances.



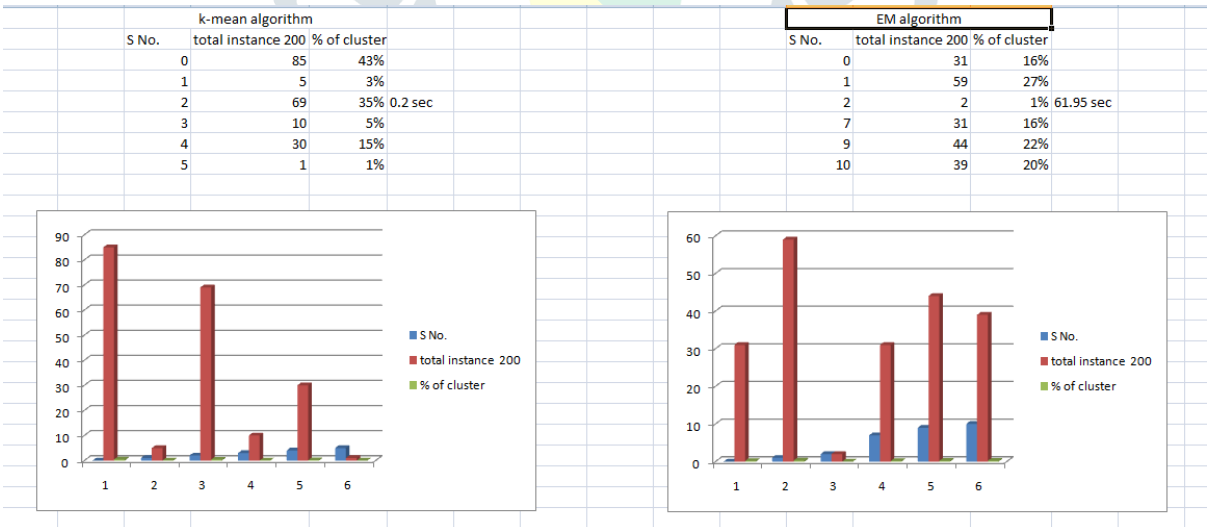
k-means algorithm

S No.	total 100 instance	% of cluster	
0	39	39%	
1	3	3%	0.1 sec
2	58	58%	

EM algorithm

S No.	total 100 instance	% of cluster	
0	52	52%	
1	32	32%	61.88 sec
2	16	16%	

Figure 14: - graph showing number of clusters formed by both algorithms when data set has 100 instances.



k-mean algorithm

S No.	total instance 200	% of cluster	
0	85	43%	
1	5	3%	
2	69	35%	0.2 sec
3	10	5%	
4	30	15%	
5	1	1%	

EM algorithm

S No.	total instance 200	% of cluster	
0	31	16%	
1	59	27%	
2	2	1%	61.95 sec
7	31	16%	
9	44	22%	
10	39	20%	

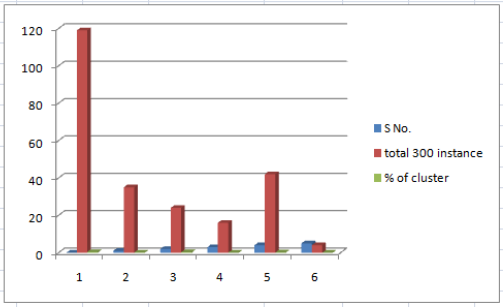
Figure 15: - graph showing number of clusters formed by both algorithms when data set has 200 instances.

k-means algorithm

S No.	total 300 instance	% of cluster	
0	119	40%	
1	35	12%	
2	24	28%	0.1 sec
3	16	5%	
4	42	14%	
5	4	1%	

EM algorithm

S No.	total 300 instance	% of cluster	
0	64	21%	
1	53	18%	
2	28	9%	61.84 sec
7	47	16%	
9	50	17%	
10	58	19%	



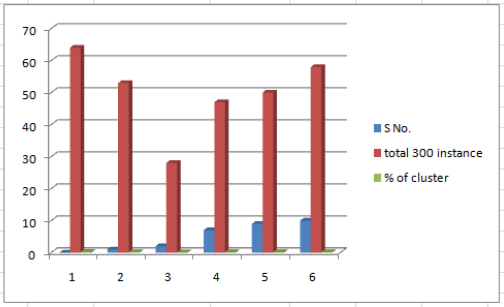


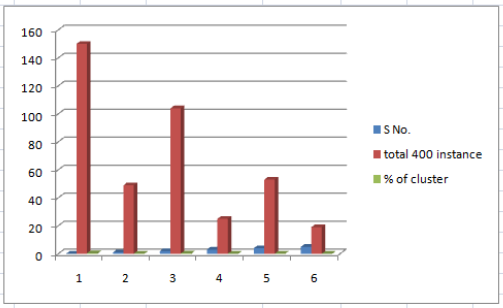
Figure 16: - graph showing number of clusters formed by both algorithms when data set has 300 instances.

k-means algorithm

S No.	total 400 instance	% of cluster	
0	150	38%	
1	49	12%	
2	104	26%	0.1 sec
3	25	6%	
4	53	13%	
5	19	5%	

EM algorithm

S No.	total 400 instance	% of cluster	
0	98	25%	
1	53	13%	
2	60	15%	62.2 sec
7	60	15%	
9	50	13%	
10	79	20%	



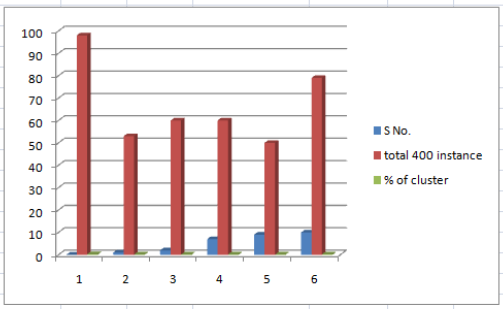


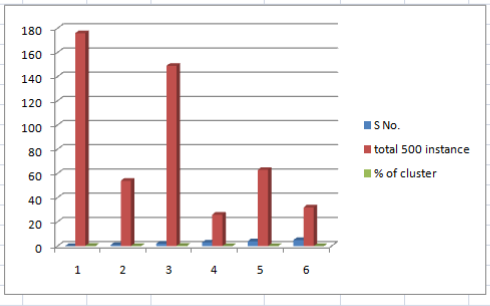
Figure 17: - graph showing number of clusters formed by both algorithms when data set has 400 instances.

k-means algorithm

S No.	total 500 instance	% of cluster	
0	176	25%	
1	54	11%	
2	149	30%	0.1 sec
3	26	5%	
4	63	13%	
5	32	6%	

EM algorithm

S No.	total 500 instance	% of cluster	
0	120	24%	
1	53	11%	
2	74	15%	62.03 sec
7	70	14%	
9	74	15%	
10	109	22%	



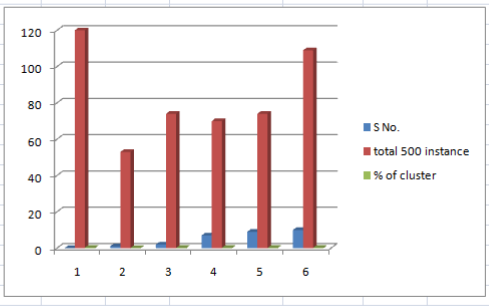


Figure 18: - graph showing number of clusters formed by both algorithms when data set has 500 instances.

4) Conclusion and future scope

After analysing the result of the two clustering algorithms it has been concluded that the mechanism of K-means algorithm is based only on the mean of the values of data set. The EM algorithm works with the calculation of mean value as well as standard deviation. When the same data sets are supplied to both the algorithms based on the same training set than to form the equal number of clusters simple K-means algorithm takes almost negligible time wherelse the time taken by the EM algorithm is comparatively greater than K-means algorithm.

This paper is intended to compare only two algorithms, which can be further extended to more algorithms comparing their result with these two. Also the instances and attributes of the data set can be increased and then the performance of the algorithm can be analysed.

5) References

1. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
2. Rashmi Sindhu et al. "*A review on K-mean algorithm and its different distance matrices"* International Journal of Engineering and Technology, Vol 9 No 2 Apr-May 2017, ISSN (Print) : 2319-8613.
3. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
4. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
5. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
6. Narender Kumar, Vishal Verma, Vipin Saxena "*CLUSTER ANALYSIS IN DATA MINNING USING K-mean METHOD*", International Journal of Computer Applications, volume 76, No.- 12, August(2013).
7. Miss. Vrindakhairnar and Miss. Sonal Patil "*EFFICIENT CLUSTERING OF DATA USING IMPROVED K-means ALGORITHM: A REVIEW*", Imperial Journal of Interdisciplinary research, volume 2, Issue 1, ISSN- 2454-1362, (2016).
8. Sudesh Kumar and Nancy *"EFFICIENT K-mean CLUSTERING ALGORITHUM FOR LARGE DATASETS USING DATA MINNING STANDARD SCORE NORMALIZATION*", International Journal on Recent and Innovation trends and Computing and Communication, volume 2, issue 10, ISSN – 2321-8169, 3161-3166.
9. Jyoti Yadav and Monika Sharma "*A REVIEW OF K-mean ALGORITHM*". International Journal of Engineering Trends and Technology, volume 4, Issue 7, July (2013)
10. Akanksha Choudhary "*SURVEY ON K-mean AND ITS VARIENTS*", International Journal of Innovative Research in Computer and Communication Engineering, volume 4, Issue 1, January (2016).
11. Archana Singh et al "*K-MEANS WITH THREE DIFFERENT DISTANCE MATRICS*", International Journals of Computer Applications (0975-8887), volume 67, No.10, April(2013).
12. Ketan Sarvakar "*A COMPARATIVE STUDY OF CLUSTERING DATA MINNING: TECHNIQUES AND RESEARCH CHALLENGES*", International Journal of Latest Technology in Engineering, management & Applied Science, volume 3, Issue 9, September (2014).
13. Brijesh Kumar Bhardwaj and Saurabh Pal "*DATA MINING: A PRIDICTION FOR PERFORMANCE IMPROVEMENT USING CLASSIFICATION*", International Journal of Computer Science and Information Security, volume 9, No. 4, April (2011).
14. Mihika Shah and Sindhu Nair "*A SURVEY OF DATA MINNING CLUSTERING ALGORITHMS*", International Journal of Computer applications (0975-8887), volume 128, No.-1, October (2015).
15. Rishikesh Suryawanshi and Shubha Puthran "*A NOVEL APPROACH FOR DATA CLUSTERING USING IMPROVED K-means ALGORITHM*". International Journal of Computer Application (0975-887), volume 142, No.-12, May (2016).
16. Sachine Shinde and Bharat Tidke "*IMPROVED K-MEANS ALGORITHM FOR SEARCHING RESEARCH PAPER*". International Journal of Computer Science & Communication Networks, volume 4(6), 197-202 197, ISSN: 2249-5789.