ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

A Comparative Analysis between K –Means & EM Clustering Algorithms

Y.Naser Eldin [1], Hythem Hashim [2], Ali Satty [3], Samani A. Talab [4]

P.G. Student, Department of Computer, Faculty of Sciences and Arts, Ranyah, Taif University, Saudi Arabia [1]

Assistant Professor, Faculty of Sciences and Arts, Blaqarn, Bisha University, Saudi Arabia [2]

Assistant Professor, School of Statistics and Actuarial Sciences, El Neelain University, Sudan [3]

Professor, Faculty of Computer Sciences and Technology, El Neelain University, Sudan [4]

ABSTRACT: This thesis is concerned with comparative analysis of the K-means and EM algorithms as clustering approaches in order to cluster the students and their results to evaluate the department's performance. Furthermore, the thesis turned to investigate the use of the so-called association rules task, with particular emphases on the K-means algorithm.

KEYWORDS: K-means, EM, Clustering, Algorithm.

I. INTRODUCTION

The term clustering refers to the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. In other words, cluster is a collection cluster of data objects, such that the objects within a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. The unsupervised assignment of elements, which is a problem of any given set to group or cluster points, is the objective of cluster analysis. There are many approaches to deal with this problem, including statistical issues, machine learning, and mathematical programming. The importance of clustering is that it is a useful task for any position where objects or events have defined with respect to attributes, and there are useful reasons to identify groups of objects by those attributes. In this thesis, we focus on clustering normalized data sets, in the sense that data is often normalized before clustering in order to remove.

UN important distinctions of scale. In clustering a corpus of text, for example, each document often represented as a point where each dimension represents a words frequency when clustering data in this format. On the other hand, the term "cluster centers" are often used to represent prototypical points, so it is usually desirable to normalize cluster centers that arise from normalized data. This fact has been well documented, for instances, the spherical K -means algorithm is well suited to this task. It is desirable to use 1-norm distance (also known as Manhattan distance, denoted here as $\|*\|_1$) to measure the distance between points. This is because the cluster center that minimizes 1-norm distance to all points within that cluster is the median of that cluster, and using the median instead of the mean tends to be more robust to outliers. Actually, most of the clustering algorithms try to generate clusters with small distances among the cluster elements, intervals, dense areas of the data space or particular statistical distributions.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

II. RELATED WORK

K -means algorithm

This algorithm is a well-known clustering technique. It is widely and frequently used when we need to find cluster centers that minimize the total of the squared 2-norm distance (also known as Euclidean distance), denoted here by $\| * \|_2$ from each point to its closest cluster center. Since finding globally optimal cluster centers is an NP-hard problem, *K* -means may be used to find a local solution. In order to carry out this algorithm, the number of clusters should be chosen to find and an initial set of cluster centers [1,5,4,7].

This is because the *K* -means algorithm splits objects in a data set into a fixed number of *K* disjoint subsets. Moreover, for each cluster, the splitting algorithm maximizes the homogeneity (Raviya and Dhinoja, 2013)[3,6,8,9]. Hand et al., 2001 noted that there are many diff t routes for selecting an initial cluster centers. However, in this article the work holds regardless of which technique is used. In what follows, we utilize the *K* -means framework, following the mathematical justifications given by Barros and Verdejo (2000).. Now, let $x_1, x_2, \ldots, x_n$ be a set of points. We can split them into K disjoint clusters $\pi_1^*, \pi_2^*, \pi_3^*, \ldots \pi_k^*$ that minimize the objective function:

$$\Phi\{\pi_j\}^k \; j=1 = \sum_{j=1}^{k} \sum_{x \in \pi_j} \left\| x - c_j \right\|_2^2 \tag{1}$$

Where for each cluster *j*, c_j is the center of each cluster, which is defined as the point for which $\sum_{x \in \pi_j} \left\| x - c_j \right\|_2^2$ is minimized, this has been proven to be accomplished by choosing *cj* to be the centroid of cluster *j*, expressed by:

$$c_j = \frac{1}{n_j} \sum_{x \in \pi_2} x, \tag{2}$$

Where n_j denotes the number of points in cluster *j*. Therefore, we can find the clusters that fix the following minimization problem:

$$\left\{\pi_i^*\right\}_{j=1}^{k} = \arg\min Q\left(\pi_j{}_{j=1}^{k}\right) \tag{3}$$

In order to find these clusters, we start with an intermittent splitting of the data $\left\{\pi_j^{(0)}\right\}_{j=1}^{k}$, We additionally define

the cluster centers associated with this partitioning as Thereafter, the following $\left\{c_j^{(0)}\right\}_{j=1}^{k}$, then the index of iteration is specified as $t = 1$.

Thereafter, the following steps are needed:

1. For each point, find the cluster center with closest Euclidean distance. This gives a new partitioning as follows:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

$$\pi_j^t = \left\{ x : \left\| x - c_j^t \right\|_2^2 \le \left\| x - c_j^t \right\|_2 2,1 \le i \le k \right\}, j = 1,\ldots k, \quad (4)$$

Where ties between clusters are resolved by random assignment to one of the optimal centers. Note that this is guaranteed not to increase the objective Q, since each point is assigned to its closest center.

2. Compute the new set of cluster centers $\left\{ c_j^t \right\}_{j=1}^{k}$ by computing the mean (centroid) of each cluster.

Since the centroid is that minimizes the total distances from all points, this step is also guaranteed not only to increase the objective *Q*.

3. If a stopping criterion is met, report $\left\{ \phi_j^t \right\}_{j=1}^{k}$ as the final partitioning and $\left\{ c_j^t \right\}_{j=1}^{k}$ as the final cluster centers.

Otherwise, increment *t* by 1, and go to step 1 above. A variety of stopping conditions are available.

The expectation maximization (EM) algorithm

The EM algorithm is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of *K* -means methods. According to Demp-ster, et al. (1997), an expectation maximization (*EM*) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori MAP) estimates of parameters in statistical models, where the model depends on unobserved hidden variables. The *EM* iteration alternates between performing an expectation (*E*) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization (*M*) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter- estimates have used to determine the distribution of the hidden variables in the next *E* step[3,4,5,7,9]. *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 generate by cross validation, or you may specify a priori how many clusters to generate. Each iteration of the EM algorithm consists of two steps, namely the expectation step and the maximization step (Little and Rubin, 2002). Each step has completed once within each algorithm cycle, which is to say that cycles have repeated until a suitable convergence criterion is satisfied. Further theoretical justification of these steps can be found in Demp-ster et al. (1977) and Little and Rubin (2002). According to Demp-ster et al. (1977), the fitted parameters (on convergence) are equal to a local maximum of a likelihood function, which is the maximum likelihood estimate in the case of a unique maximum. The EM algorithm has two disadvantages: firstly, it is typically very slow to converge, and secondly, it lacks direct provision of a measure of precision for the maximum likelihood estimates. Several proposals have been made to overcome these drawbacks, and we refer to the techniques as provided by Louis (1982), McLachlan and Krishnan (1997), Rubin (1991) and Baker (1992). Details overview of the EM algorithm are given in Bin and Hancock (2001), Pal and Mitra (2002) as well as in Pernkopf and Bouchaff (2005)[1,3,4].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

For clustering data mining, this algorithm has some advantage to be under taken into accounts (Ruoming and Agrawal, 2002). On of these advantage is that it has strong statistical properties, in the sense that it is robust to noisy data. The second advantage associated with algorithm is that it can accept the desired number of clusters as input. The third benefit is that this algorithm can provide a cluster membership probability per point, it can deal with specific circumstances such as handling high dimensionality and it converges fast given a good initialization. These advantages have been well documented in Roweis and Ghahramani (1999) as well as Ordonez and Cereghini (2000)[2,5].

Table (1) illustrates the parameters of the EM algorithm. EM algorithm first initializes the parameters of the mixture model according to the approximate clusters, which has obtained through GCEA. In the first iteration (t = 1), $\pi_j^{(1)}$ and $\mu_j^{(1)}$ are computed by Equations. (5) and (6), respectively, while the entries of $\sum_j^{(1)}$ are evaluated based on Equation (7) as follows

$$\pi_j^{(1)} = \frac{|c_j|}{\sum_{j=1}^{n_c}(|c_j|)} \tag{5}$$

$$\mu_j^{(1)} = \frac{\sum_{x \in c_j} x}{|c_j|} \tag{6}$$

$$Cov(k,\ell) = \frac{\sum_{x \in c_j}(x_k - \bar{x}_k)(x_1 - \bar{x}_1)}{|c_j|} \tag{7}$$

Parameter	Meaning
xi	The th data point
|cj |	The cardinality of the cluster *cj*
Nc	The number of the clusters in the initial step
Πj	Mixing proportion of the cluster *cj*
μj	The mean vector of the cluster *cj*
$\sum_j$	Covariance matrix of the cluster *cj*
$Cov(k,\ell)$	Covariance between *k* and *l* dimension of the cluster *cj*
Xk	The average value of the data points in the *k*th dimension
Θ	The parameters to be estimated
t	The *t* th iteration
N	The cardinality of the point set *P*

Table (1) The parameter of the EM algorithms

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

III. PERFORMING THE COMPARISON ANALYSIS BETWEEN ALGORITHMS

The data description

The student's data set has been used for clustering in this paper consists of 108 records divided into three classes that represent three departments in the School of statistic, Faculty of Mathematical Sciences and Statistics. The data includes all the academic information regarding the student in the school. Particularly, the analyzed data set consists of the following fi e numeric variables and nominal column.

Deg1 : represents the students degree in first year,
Deg2 : represents the students degree in second year,
Deg3 : represents the students degree in third year,
Deg4 : represents the students degree in forth year,
Deg5 : represents the students degree in final year,
Depart : factor with three levels as the class label.

A summary of numeric columns is displayed in Table (2). Figure (1) (a) shows the distribution of students in departments and (b) shows the distribution of students marks in the departments.

Measurement	Deg1	Deg2	Deg3	Deg4	Deg5
Min.:	52.18	52.41	52.077	52.10	54.34
1stQu.:	59.81	59.96	61.34	59.28	61.80
EM:	67.06	64.77	66.69	64.60	67.14
K-Mean:	65.57	65.35	66.73	66.59	68.24
3rd Qu.:	70.38	70.19	72.38	73.12	72.92
Max.:	78.77	87.62	80.92	89.07	92.29

Table (2): Dataset summary

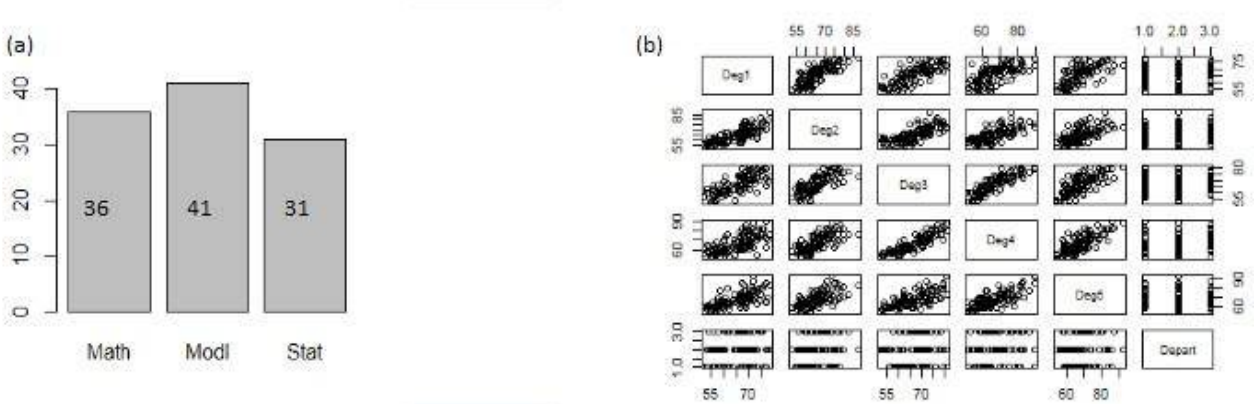

Figure (1) Students distribution

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

For performing the comparison analysis regarding the above-mentioned clustering algorithms, we used an education data set that comes from Faculty of Mathematical Science and Statistics. This data set is very helpful for the researchers, particularly for those who are considering most of the clustering algorithms. We directly apply this data set in order to predict the result. The data has been applied using diff t student records in order to detect useful results that can be very helpful for the new users, new researchers and enhancing the degrees for the current students. In order to perform the analysis, we convert the data set to the CSV data set format. This data set contains 457 and 8 attributes. These attributes are year, sex, department (Dep), and the students' degrees for 5 years, denoted by Degree 1 (Deg1), Degree 2 (Deg2), Degree 3 (Deg3), Degree 4 (Deg4) and Degree 5 (Deg5). Figure (2) explains the description of the considered data set.

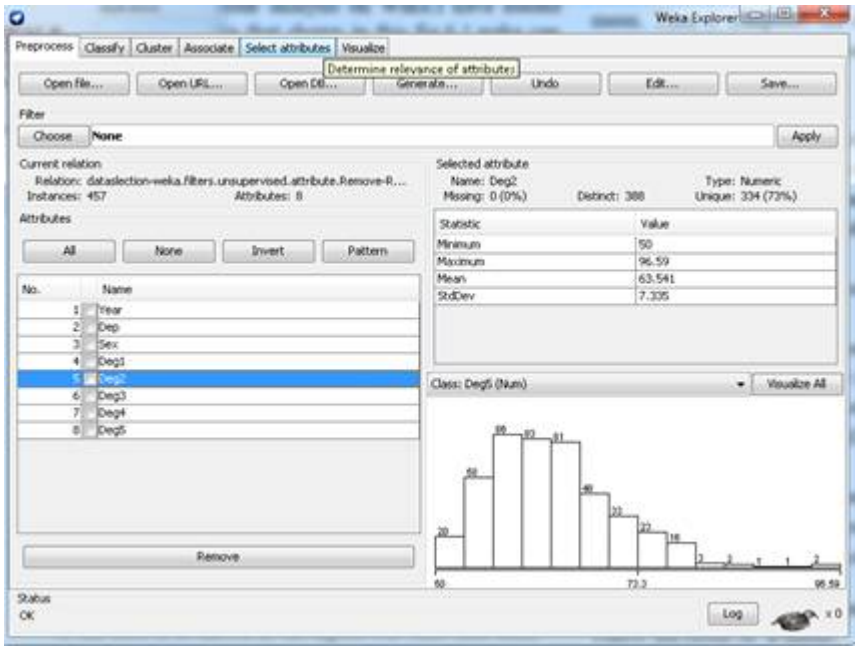

Figure (2) Loaded data set in to the WEKA software

1) The *R* software

As noted by Development Core Team (2012), this is a free software environment for statistical computing and graphics. This software presents a wide variety of statistical and graphical techniques. Also, it can be extended easily via packages. As discussed in Hand et al., (2001), there are around 4000 packages available in the CRAN pack- age repository. Comprehensive details with respect to this software have been given by Venables et al., (2010) and R Language Definition (R Development Core Team, 2010a, 2010b) at the CRAN website. R is widely used in both academia and in- dustry. Note that this software would be a primary issue in the current analysis for implementing both, the K -means and EM algorithms [2,4,7]. In order to assist users getting suitable R packages to implement, the CRAN Task Views can be considered as a good guidance. Particularly, these views give collections of packages for diff t tasks. Below, we outline several task views linked with data mining are:

- Machine Learning & Statistical Learning;
- Cluster Analysis & Finite Mixture Models;

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

- Time Series Analysis;
- Multivariate Statistics
- Analysis of Spatial Data.

IV. RESULTS

This article is devoted to a comparison of different data mining clustering techniques, namely K -means clustering, EM. Once the clustering task is chosen, the number of clusters (K) must be decided. Therefore, the main question when dealing with the clustering algorithms is that what number of clusters or means should be specified. Each algorithm uses different ways, so the final results are different depending on the way. The main aim of this article is to compare some of the clustering algorithm under different number of clusters. Performance of these techniques is presented and compared using a clustering tool provided in WEKA. We compare the performance of these major clustering algorithms on the aspect in terms of:

- Number of clusters,
- Clustered instances,
- Time taken to build model, and

Algorithm	Number of clusters	Clustered instances	Time taken
K -means	2	32% 68%	0.03 second
EM	2	37% 63%	0.16 second
K -means	4	23% 32% 18% 27%	0.06 second
EM	4	20% 28% 44% 08%	0.3 second

Table (3) Results of comparison of tow clustering algorithms

Comparing the time taken to build model for the used algorithms, Table (3) shows the results calculating the time taken to execute each algorithm. Under K =2, the results of these algorithms revealed that K -means, has less time (0.03) to execute as compared with EM algorithm. Thus, it was more efficient algorithms compared to EM algorithm in this article. In comparison with results for EM (0.16). Consequently, the Hierarchical algorithm being asymptotically less efficient than EM. For K =4, the times taken to build model for the K -means was 0.06, which is again less time to execute. Therapy, the benefits of K -means, based over EM are clearly evident. Overall, the performance for these algorithms were good as that based on K =2. Figure (3) shows the performance of applying the aforementioned clustering algorithms.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

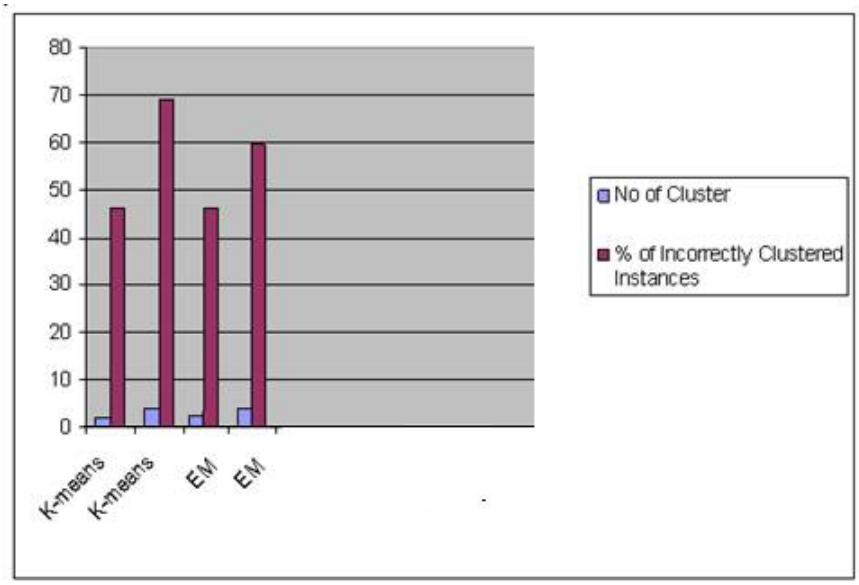

Figure (3) Clustering performance

V. CONCLUSION

By looking at the results in Table (2) we see that the K -means has shown a better performance compared with EM algorithms considered in this study. The is because the fact that as the percentage of the incorrectly classified attribute is low, the performance of the clustering algorithm could give a satisfied performance.

REFERENCES

[1] Abonyi, J. and Feil, B., Cluster analysis for data mining and system identification. Birkhauser, Basel, 2007.
[2] Berry, M. W. , Survey of Text Mining: Clustering, Classification, and Retrieval. New York: Springer 2003.
[3] Bradley, P., Fayyad, U. and Reina, C. (). Scaling clustering algorithms to large databases. In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 1998.
[4] Dhillon, I. S. and Modha, D. S., Concept decompositions for large sparse text data using clustering. Machine Learning, vol.42, pp.143-175 , 2001.
[5] El-Darzi, E., Abbi, R., Vasilakis, C., Gorunescu, F., Gorunescu, M., Millard, P. , Length of stay-based clustering methods for patient grouping. In: McClean, S., Millard, P., El- Darzi, E., Nugent, C. (eds.) Intelligent Patient Management. Studies in Computational Intelligence, Springer , vol. 189, pp. 39-56, 2008.
[6] Gonzalez, T. F. (). Clustering to minimize the maximum inter cluster distance. Theoretical Computer Science, vol.38, pp.293-306, 1985.
[7] Zhang, C. and Zhang, S. , Association Rule Mining. Models and Algorithms. LNCS (LNAI), vol. 23, pp.170-189,2002.
[8] Tan, P., Steinbach, M. and Kuman, V. Introduction to Data Mining Addison Wesley 2006
[9] Vishal, S. and Prem, N. A. , A study of various algorithms on retail sales data. International Journal of Computations and Networking, vol.1, pp.2319-2720, 2012.