

# An Overview of Expectation Maximization and K-Means family Clustering Algorithms in Data Mining Applications

Mehrsa Moghadaszadeh<sup>1</sup> and Hamid Shokrzadeh<sup>2</sup>

<sup>1</sup> Department of Computer Engineering Pardis Branch, Islamic Azad University, Pardis, Iran

<sup>2</sup> Department of Computer Engineering Pardis Branch, Islamic Azad University, Pardis, Iran  
moghadaszadeh.6mehrsa@gmail.com , shokrzadeh@gmail.com

**Abstract:** One of the important issues in data mining is data clustering that it is a type of unsupervised learning. Data sets are divided into sub-groups where the existing data are similar in each cluster. During the past years, many algorithms have been proposed for clustering applications such as Expectation Maximization (EM) algorithm and popular versions of K-Means. In this paper, we introduce these methods in detail and finally, compare them based on parameters such as convergence speed, efficiency in data set size and accuracy.

**Keywords:** Expectation Maximization, K-Means, Data mining, Clustering.

## 1. Introduction

Data mining is an approach defined as Knowledge Discovery of data in which clustering (subsets of points are similar) is a process for data examination. Here, we introduce some significant clustering algorithms such as the EM and the K-Means family.

## 2. K-Means algorithm

K-Means algorithm [1] was introduced by MacQueen in 1967. It is one of the simplest unsupervised learning algorithms in data mining in which cluster formation is explained as follows:

In the first step,  $k$  points are selected randomly as initial centers. In the second step, the distance between each point to all of these central points is calculated according to Euclidean distance. Then, clusters are formed and the mean of each cluster is calculated as a novel center for the clusters. Finally, the cycles is repeated until it is converged (the K-Means algorithm steps are shown in Figure 1).

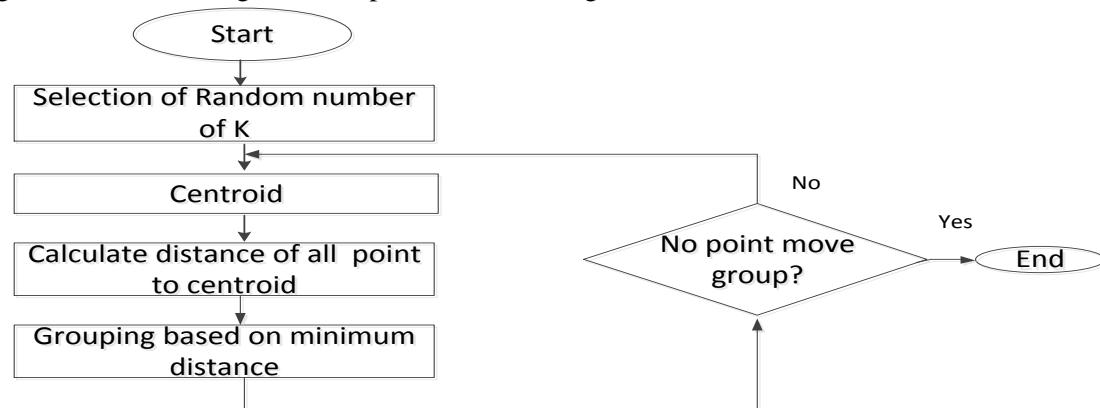


Fig. 1: Flowchart of the K-Means algorithm

## 3. EM algorithm

The EM [2] was introduced by Dempster in 1977. The EM is used in mixture models one of the most important of which is the Gaussians mixture model. The Gaussian Mixture Model (GMM) [3] is used to model multifaceted probability distributions where each model includes parameters such as mean, variance and weight.

The EM is a popular iterative approach which estimates the aforementioned parameters according to maximum likelihood estimate (MLE). It consists of two steps: Expectation (E-step) and Maximization (M-step). In the E-step, membership probability of each point to a cluster is computed and in the M-step, the parameters are obtained for the next iteration using the parameters in the E step [4]. These steps are executed until it is converged, i.e., there is no change (as shown in equation 1).

$$\log l(\Theta) = \sum_{i=1}^N \log p(x_i | \Theta) = \sum_{i=1}^N \left( \log \sum_{k=1}^K \alpha_k p_k(x_i | z_k \Theta_k) \right) \quad (1)$$

Where parameter vector  $\Theta$  presents the above-mentioned parameters and  $p(x_i | \Theta)$  is a mixture model. Here, a data set  $X = \{x_1, x_2, \dots, x_N\}$  is introduced where  $x_i$  is a d-dimensional vector measurement  $1 \leq i \leq N$  and  $p_k(x_i | z_k \Theta_k)$  is the Gaussian density function for the  $k^{th}$  mixture component where  $z_k$  is K-ary random variable representing the identity of the mixture component that generated  $x_i$ . It is worth noting that  $\alpha_k$  is introduced as the membership weight for  $x_i$  and  $k$  is all mixture components  $1 \leq k \leq K$ .

#### 4. Enhanced K-Means Algorithm using Elbow Method

The K-Means algorithm problems are explained in [5] one of which is the determination of an optimal number of clusters. To overcome this problem, many methods have been introduced to improve K-Means. One of the newest K-Means family algorithms has been proposed in [6] named Elbow K-Means. The Elbow is a smart method used to find the number of clusters where percentage of variance is introduced as a function of the number of clusters [6]. As shown in Figure 2, curve credits in the chart and an existing angle in the chart are indicated as Elbow criteria, i.e., the number of clusters is chosen at this point.

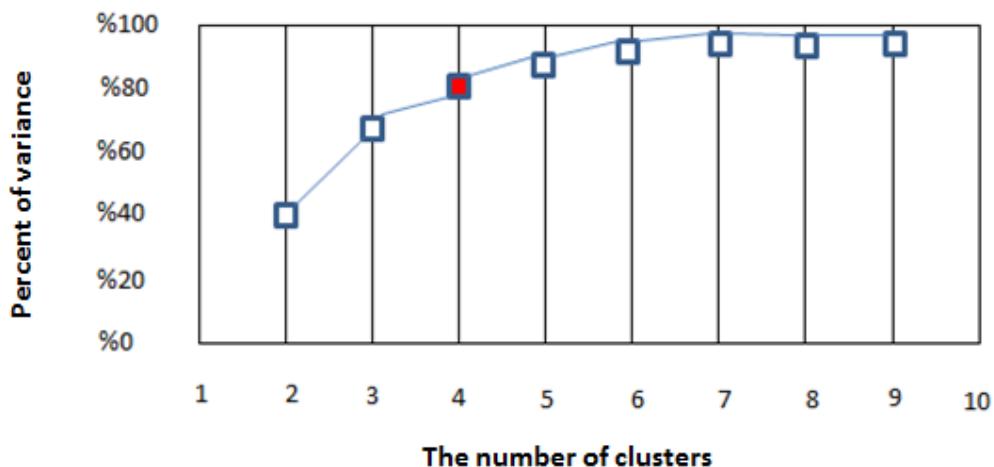


Fig. 2: Elbow method

#### 5. Enhanced K-Means Algorithm using Davis-Bouldin Index method

DBIK-Means algorithm is proposed in [7] in order to determine the number of clusters by the Davis-Bouldin Index (DBI).  $R_{i,j}$  parameter is a measure for suitable clustering schema. In this method,  $M_{i,j}$  parameter is the distance between  $i^{th}$  and  $j^{th}$  clusters centers and  $S_i$  parameter is proposed as sum of the average distances from every point to the center of cluster  $i$  and the DBI[8] is defined as  $S_i$  and  $M_{i,j}$  ratio as follows [7]:

$$R_{i,j} \geq 0$$

$$\text{then } R_{i,j} = R_{j,i}$$

$$\text{When } S_i = S_k \text{ and } M_{i,j} \leq M_{i,k}$$

then  $R_{i,j} \geq R_{i,k}$   
**When  $S_i = S_k$  and  $M_{i,j} \leq M_{i,k}$**   
 then  $R_{i,j} \geq R_{i,k}$

For clearer separation for clusters properties are used according to the equations below:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (2)$$

$$D = \max_{i \neq j} R_{i,j} \quad (3)$$

$N$  parameter indicates some clusters:

$$DB \equiv \frac{1}{N} \sum_{i=1}^N D_i \quad (4)$$

## 6. Enhanced K-Means algorithm using Silhouette method

Silhouette method is suggested in [9] to determine the optimum number of clusters in the K-Means algorithm. The Silhouette [10] is defined as equation 5:

$$S(i) = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (5)$$

$a(i)$  is average distance between  $i$  and other points of the cluster in terms of which  $i$  belongs to the cluster and  $b(i)$  is the minimum average distance  $i$  and all of the points in each cluster that  $i$  does not belong to [9]. Different values obtained are from -1 to 1 in this method as follows:

- Zero value: an entity can be related to other clusters
- Value is close to -1: an entity is misclassified
- Value is close to 1:  $S(i)$  is well clustered

Finally, the best number of clusters is obtained by the largest average of the Silhouette.

## 7. Enhanced K-Means algorithm using Calinski Harabasz method

Calinski Harabasz (CH) is introduced in [11] for determination of  $k$  (the number of the clusters) value. This method is based on the relationship between two matrixes as between cluster scatter matrix (BCSM) and within cluster scatter matrix (WCSM). As shown in equation 6, the trace BCSM is explained as sum of the distances squares between each cluster and the centroid of the data set and the trace WCSM is introduced as sum of distances squares between center and the existing pointes in each cluster [11]. It is worth noting that  $\frac{N-K}{K-1}$  factor which is defined as  $k$  value increases in this equation.

$$CH = \frac{\text{trace}(BCSM)}{\text{trace}(WCSM)} * \frac{N-K}{K-1} \quad (6)$$

## 8. Evaluation

As shown in Table 1, the clustering algorithms are compared on the basis of parameters such as convergence speed, data set size and accuracy. Convergence is an important factor in the aforementioned algorithms to show that an algorithm has not changed from one iteration to the next after every iteration. It is worth noting that the EM algorithm includes faster convergence speed than the methods and takes less time for execution. The accuracy parameter is used to represent the clustering quality where the EM algorithm shows higher accuracy than other methods. The data set size is addressed based on soft clustering and hard clustering issues. In hard

clustering, each point belongs to a cluster. However, in soft clustering each point can belong to several clusters with different probabilities. As shown in Table 1, the EM is used for large data sizes unlike the K-Means family because the EM has been proposed as soft clustering and the K-Means is hard clustering in data mining. In regard to the analysis of existing parameters, the EM algorithm is better than other methods.

TABLE 1: Comparison of Clustering Algorithms

Algorithms	Accuracy	Convergence speed	data set size
K-Means	Low	Very low	Small
EM	High	High	Large
Elbow K-Means	Low	Low	Small
DBIK-Means	Moderate	Moderate	Small
Silhouette K-Means	Moderate	Moderate	Small
CHK-Means	Low	Moderate	Small

## 9. Conclusion

In this paper, the performance of the EM and the K-Means family clustering algorithms was compared according to parameters such as accuracy, convergence speed and data set size. The results indicated that the EM algorithm performs better compared to the K-Means family.

## 10. References

- [1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, no. 14, pp. 281–297, 1967.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.
- [3] C. E. Rasmussen, "The infinite Gaussian mixture model," *Adv. Neural Inf. Process. Syst.*, pp. 554–560, 2000.
- [4] H. Shokrzadeh, S. Khorsandi, and A. T. Haghighat, "Optimized query-driven appointment routing based on Expectation- Maximization in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 35, no. 6, pp. 1749–1761, 2012.
- [5] K. Singh, D. Malik, and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal," *IJCEM Int. J. Comput. Eng. Manag. ISSN*, vol. 12, no. April, pp. 2230–7893, 2011.
- [6] P. Bholowalia, and A.Kumar“EBK-Means : A Clustering Technique based on Elbow Method and K-Means in WSN,” vol. 105, no. 9, pp. 17–24, 2014.
- [7] E. Rabiaa, B. Noura, and C. Adnene, "Improvements in LEACH based on K-means and Gauss algorithms," *Procedia Comput. Sci.*, vol. 73, no. Awict, pp. 460–467, 2015.
- [8] D. L. Davies. and D .W. Bouldin., "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no.2, pp. 224–227, 1979.
- [9] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 2321–7782, 2013.
- [10] P. J. Rousseeuw, "Silhouettes: A Graphic Aid to the Interpretation and Validation of Cluster Analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
- [11] J. Baarsch and M. E. Celebi, "Investigation of Internal Validity Measures for K-Means Clustering," *Proceeding Internatinal MultiConference Eng. Comput. Sicientists*, vol. I, 2012.