British Journal of Mathematics & Computer Science
2(2): 62-71, 2012

SCIENCEDOMAIN *international*
www.sciencedomain.org

An Exploratory Study of K-Means and Expectation Maximization Algorithms

Adigun Abimbola Adebisi[1], Omidiora Elijah Olusayo[1*] and Olabiyisi Stephen Olatunde[1]

[1] *Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria.*

Research Article

Received 20th December 2011
Accepted 26th January 2012
Online Ready 23rd March 2012

Abstract

In this paper, K-Means and Expectation-Maximization algorithms are part of the commonly employed methods in clustering of data in relational databases. Experiments conducted with both clustering algorithms revealed that both algorithms have been found to be characterized with shortcomings. The parameters considered in evaluating the results of findings are the number of iterations (no distinct convergence, 1), the computation time (not defined, 3.2s) and the memory space (not defined, 1.1MB) consumed at the point of convergence of both K-means and Expectation-Maximization algorithms respectively. The results obtained revealed that Expectation-Maximization algorithm's quick and premature convergence cannot be said to have guaranteed optimality of results while K-means was found not to guarantee convergence. Though reasonable conclusion could be drawn from results obtained with Expectation-Maximization algorithm, its premature convergence may raise some questions of doubt with regards to reliability of results obtained.

Keywords: K-Means, Expectation Maximization, Clustering, Student database.

1 Introduction

The K-Means algorithm is a very popular algorithm for data clustering because of its simplicity. Originally developed for and applied to the task of vector quantization, k-means has been used in a wide assortment of applications. It has been proven to be a good approach to classify data. K-Means (KM) clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait - often proximity according to some defined distance measure (MacQueen, 1967). Intuitively, patterns within a valid cluster are more similar to each other than they are to a

*Corresponding author: Email: omidiorasayo@yahoo.co.uk;

pattern belonging to a different cluster. The K-Means algorithm can also be viewed as an unsupervised classification (Vojtˇech and Vaclav, 2004).

Although K-means has the great advantage of being easy to implement, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success (Fayyad et al., 1998).

Also, it has been shown that, with K-Means, there is no guarantee for optimal clustering, since the convergence depends on the initial number of clusters selected. In addition, K-means is not considered as the best choice for clustering due to its time performance and requirements. K-means typically requires that clusters be spherical, that the data be free of noise and that its operation be properly initialized. This makes it inefficient for major industrial clustering problems (Tapas et al., 2002).

Expectation Maximization (EM) is a model based approach to solving clustering problems. It is an iterative algorithm that is used in problems where data is incomplete or considered incomplete. Unlike distance based or hard membership algorithms (such as K-Means), EM is known to be an appropriate optimization algorithm for constructing proper statistical models of the data. EM is widely used in applications such as computer vision, speech processing and pattern recognition. EM aims at finding clusters such that maximum likelihood of each clusters parameters is obtained.

In EM, each observation belongs to each cluster with a certain probability. EM clusters data, in a manner different than K-means. EM starts with an initial estimate for the missing variables and iterates to find the maximum likelihood (ML) for these variables. Maximum likelihood methods estimate the parameters by values that maximize the sample's probability for an event.

EM is typically used with mixture models. Unlike in K-means, in clustering via EM, the number of clusters that are desired are predetermined. It is initialized with values for unknown (hidden) variables. Since EM uses maximum likelihood, it most likely converges to local maxima, around the initial values. Hence selection of initial values is critical for EM. However, the EM algorithm works well on clustering data when the number of clusters is known (Sara, et al., 2006).

In this paper, an evaluation of k-means and Expectation Maximization algorithms would be carried out with the employment of the educational database of students admitted through Joint Admission and Matriculated Board (JAMB)- an accredited University admission body, and Pre-degree (PDS)- a self conducted University administered entrance examination.

2 K-Means Algorithm

```
Initialize $k$ prototypes $(w_1,\ldots, w_k)$ such that $w_j = i_l, j \in \{1,\ldots,k\}, l \in \{1,\ldots,n\}$
Each cluster $C_j$ is associated with prototype $w_j$

Repeat
        For each input vector $i_l$, where $l \in \{1,\ldots,n\}$,
        Do
                Assign $i_l$ to the cluster $C_{J^*}$ with nearest prototype $w_{j^*}$
```

```
            (i.e., | $i_l$ - $w_{j*}$| ≤ | $i_l$ - $w_j$|, $j \in \{1,\ldots,k\}$ )

For each cluster C$_j$, where $j \in \{1,\ldots,k\}$, do
            Update the prototype $w_j$ to be the centroid of all samples currently in C$_j$,
            so that $w_j = \sum_{i_j \in C_j} i_l / |C_j|$
    Compute the error function:
            $E = \sum_{j=1}^{k} \sum_{i_l \in C_j} | i_j - w_j |^2$

Until E does not change significantly or cluster membership no longer changes.
```

In the k-means algorithm described above, (Mehrotra, Mohan and Ranka,1996; Kaufman and Rousseeuw, 1990; Dubes and Jain, 1988), the number of clusters is an input parameter into the algorithm.

Let the *k* prototypes ($w_1,\ldots, w_k$) be initialized to one of the *n* input patterns ($i_{1,\ldots,}i_n$). Therefore, $w_j = i_l, j \in \{1,\ldots,k\}, l \in \{1,\ldots,n\}$

C$_j$ is the jth cluster whose value is a disjoint subset of input patterns.

The quality of the clustering is determined by the performance function of the algorithm which is given as;

$$Perf_{KM} = \sum_{j=1}^{k} \sum_{i_l \in C_j} | i_j - w_j |^2$$

3 Expectation Maximization (EM) Clustering Algorithm

The inputs to this algorithm are the data set (x), the total number of clusters (M), the accepted error to converge (e) and the maximum number of iterations.

The algorithm can be subdivided into two stages, namely the initialization stage and the iterative stage which consists of two steps, expectation step (E-step) and maximization step (M-step) executed iteratively until some form of convergence is reached. The E-Step estimates the probability of each point belonging to each cluster, followed by the M-step which re-estimates the parameter vector of the probability distribution of each class. The algorithm finishes when the distribution parameters converge or reach the maximum number of iterations.

Initialization

Each class j, of M classes (or clusters), is constituted by a parameter vector (θ), composed by the mean (μ_j) and by the covariance matrix (P_j), which represents the features of the Gaussian probability distribution (Normal) used to characterize the observed and unobserved entities of the data set x.

$$\theta(t) = \mu_j(t), P_j(t),\ j = 1 \ldots M$$

On the initial instance (t = 0), the implementation can generate randomly the initial values of mean (μ_j) and of the covariance matrix (P_j).

The EM algorithm aims to approximate the parameter vector (θ) of the real distribution. Another alternative offered by MCLUST (Model-based Clustering) is to initialize EM with the clusters obtained by a hierarchical clustering technique.

E-Step

This step is responsible for estimating the probability of each element belonging to each cluster ($P(C_j \mid x_k)$). Each element is composed of an attribute vector (x_k). The relevance degree of the points of each cluster is given by the likelihood of each element attribute in comparison with the attributes of the other elements of cluster C_j (Sara *et al.*, 2006).

$$P(C_j|x) = \frac{|\sum_j(t)|^{-\frac{1}{2}}\ exp^{n_j}\ P_j(t)}{\sum_{k=1}^{M}\ |\sum_j(t)|^{-\frac{1}{2}}\ exp^{n_j}\ P_k(t)}$$

M-Step

This step is responsible for the estimation of the parameters of the probability distribution of each class for the next step. First, compute the mean (μ_j) of classes j obtained through the mean of all points in function of the relevance degree of each point.

$$\mu_j(t+1) = \frac{\sum_{k=1}^{N}\ P(C_j|x_k)\ x_k}{\sum_{k=1}^{N}\ P(C_j|x_k)}$$

To compute the covariance matrix for the next iteration, the Bayes Theorem is applied, which implies that P(A I B) = P(B I A) * P(A)P(B), based on the conditional probabilities of the class occurrence (Sara et al., 2006).

$$\sum_j(t+1) = \frac{\sum_{k=1}^{N}\ P(C_j|x_k)\ (x_k - \mu_j(t))\ (x_k - \mu_j(t))}{\sum_{k=1}^{N}\ P(C_j|x_k)}$$

The probability of occurrence of each class is computed through the mean of probabilities (C_j) in function of the relevance degree of each point from the class.

$$P_j\ (t+1) = \frac{1}{N}\sum_{k=1}^{N} P\left(C_j|x_k\right)$$

The attributes represent the parameter vector, θ, that characterize the probability distribution of each class that will be used in the next algorithm iteration.

Convergence Test

After each iteration is performed a convergence test which verifies if the difference of the attributes vector of an iteration to the previous iteration is smaller than an acceptable error tolerance, given by parameter. Some implementations use the difference between the averages of class distribution as the convergence criterion (Sara et al., 2006).

```
If ($\| \theta (t + 1) - \theta (t) \| <$ Q)
   stop
else
   call E-Step
end;
```

The algorithm has the property of, at each step, estimating a new attribute vector that has the maximum local likelihood, not necessarily the global, which reduces its complexity. However, depending on the dispersion of the data and on its volume, the algorithm can stop due the maximum number of iterations defined.

4 Performance Function of Expectation Maximization

Unlike K-Means in which only the centers are to be estimated, the EM algorithm estimates the centers, the co-variance matrices, $\sum_k$ and the mixing probabilities, *p(m$_k$)*.

The performance function of the EM algorithm is (Zhang, et al., 1999).

$$Perf_{EM}(X,M,\Sigma,p) = -\log\left\{\prod_{x\in S}\left[\sum_{k=1}^{K} p_k \cdot \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot EXP(-(x-m_k)\Sigma_k^{-1}(x-m_k)^T)\right]\right\}.$$

where the vector $p = (p_1, p_2,\ldots\ldots, p_K)$ is the mixing probability. *EM* algorithm is a recursive algorithm with the following two steps:

E-Step

Estimating “the percentage of *x* belonging to the *k*th cluster” (Zhang et al.,1999)

$$p(m_k \mid x) = p(x \mid m_k) \cdot p(m_k) \Big/ \sum_{x\in S} p(x \mid m_k) \cdot p(m_k),$$

where $p(x|m)$ is the prior probability with Gaussian distribution, and $p(m_k)$ is the mixing probability,

$$p(x \mid m_k) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma_k)}} \cdot EXP(-(x-m_k)\Sigma_k^{-1}(x-m_k)^T)$$

M-Step

With the fuzzy membership function from the E-Step, find the new center locations, new co-variance matrices and new mixing probabilities that maximize the performance function (Zhang et at., 1999).

$$m_k = \frac{\sum_{x\in S} p(m_k \mid x) \cdot x}{\sum_{x\in S} p(m_k \mid x)}, \quad \Sigma_k = \frac{\sum_{x\in S} p(m_k \mid x) \cdot (x-m_k)^T (x-m_k)}{\sum_{x\in S} p(m_k \mid x)}, \quad p(m_k) = \frac{1}{|S|}\sum_{x\in S} p(m_k \mid x).$$

5 Materials and Methods

The datasets used for the implementation of the two algorithms were obtained from the database of the Faculty of Engineering and Technology, Ladoke Akintola University of Technology, Ogbomoso, Nigeria in West Africa. The Student's mode of admission and final CGPA fields were then isolated for the mining. This was done to find the relationships between their mode of admission and the final CGPA. The admission modes considered here were the Pre-Degree and JAMB admission modes. However, for prepossessing of the data, the PDS mode of admission was represented by integer 1 while that of the JAMB admission mode was represented by the integer 2. The two algorithms were implemented using MATLAB tool.

Note that the EM algorithm automatically produced three clusters and on any number of re-runs, it produced exactly the same clusters which make it a very rigid algorithm.

6 Results and Discussion

6.1 Results Obtained

The performance metrics considered include the number of iterations, computation time and system memory usage at convergence.

The summary of the results obtained are as shown in Tables 1, 2 and 3 using the three performance metrics.

Table 1. Table showing number of iterations at convergence for the algorithms

Algorithm	Number of Iterations
K-Means	No distinct convergence (∞)
Expectation-maximization	1

Table 2. Table showing computation time at convergence for the algorithms

Algorithm	Computation time
K-Means	Undefined (∞)
Expectation-maximization	3.2s

Table 3. Table showing space requirements at convergence for the algorithms

Algorithm	System memory usage
K-Means	∞
Expectation-maximization	1.1MB

6.2 Discussion

6.2.1 Expectation Maximization

The expectation-maximization algorithm converged in a single run giving three clusters. Figure 1 shows the clusters with their three centroids. Two of the clusters were for students admitted through PDS with the eclipse demarcating the clusters and the third cluster for students admitted through JAMB.

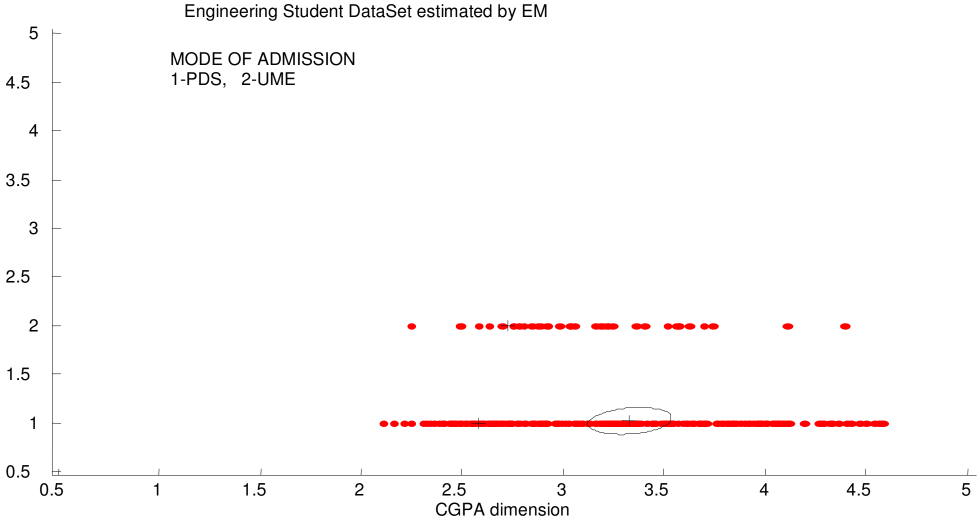

Fig. 1. Graph of Mode of Admission against CGPA

From virtual observation of the clusters, the cluster for students admitted through JAMB has its centroid on approximately 3.25 (CGPA). However, the clusters for students admitted through PDS have their centroids placed on 3.1 (CGPA) and approximately 4.1 (CGPA). Taking a linear average of the CGPA on PDS mode of admission revealed a CGPA of 3.6, it can be concluded that students that were admitted through PDS on the average performed better than those admitted through JAMB. A possible explanation being that the PDS students would have undergone a thorough pre-university academic training for a year immediately before admission and on the other hand, it is not so for their JAMB counterparts, some of which would have stayed at home for sometime expecting a better JAMB result for their admission into the University.

However, another probable explanation is that the PDS programme affords the students the opportunity of being taught some of the 100 Level courses curriculum so that most of what they are taught in 100 Level becomes more of revision, which gives them a better edge over their JAMB counterparts. The results of the experiments carried out revealed that PDS mode of admission should rather be encouraged by the management than JAMB mode.

6.2.2 K-Means

The K-Means algorithm was supplied with five (5) as input for the number of desired clusters to be produced. The number being supplied depicts the number of the classes of degree in the University: first class, second class (upper division), second class (lower division), third class and pass degree. The K-Means algorithm produced five (5) clusters for each re-run of the algorithm.

The K-Means algorithm was run a hundred and two times. It was observed that the algorithm did not show a distinct convergence even though similar pattern clusters were repeating themselves at irregular intervals. Interestingly, some particular patterns of clusters were persistent and the first cluster could be observed in Figure 2 with thirty-six (36) occurrences in one hundred and two (102) runs of the algorithm (35.3%).

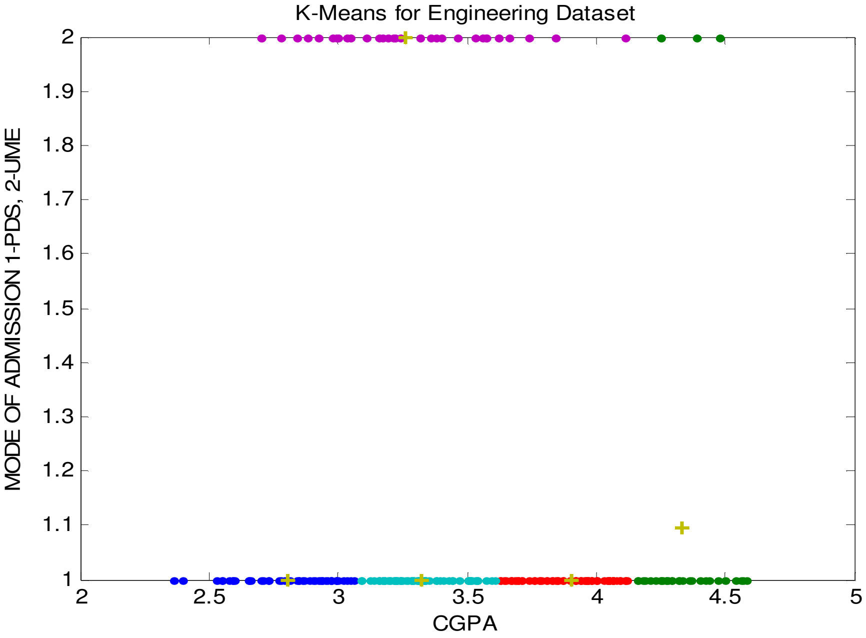

Fig. 2. Graph of mode of admission against CGPA

Another pattern of clusters with a closer persistence to the one observed above was twenty-two (22) occurrences in one hundred and two runs (102) runs of the algorithm (21.6%). This cluster pattern is shown in Figure 3.

Picking the cluster pattern with the highest number of occurrence (Figure 2), it could be assumed to be the optimum set of clusters even though there is no distinct convergence of the cluster patterns. As a result, it would be very difficult to draw any conclusion from K-Means clusters (Figures 2 and 3).

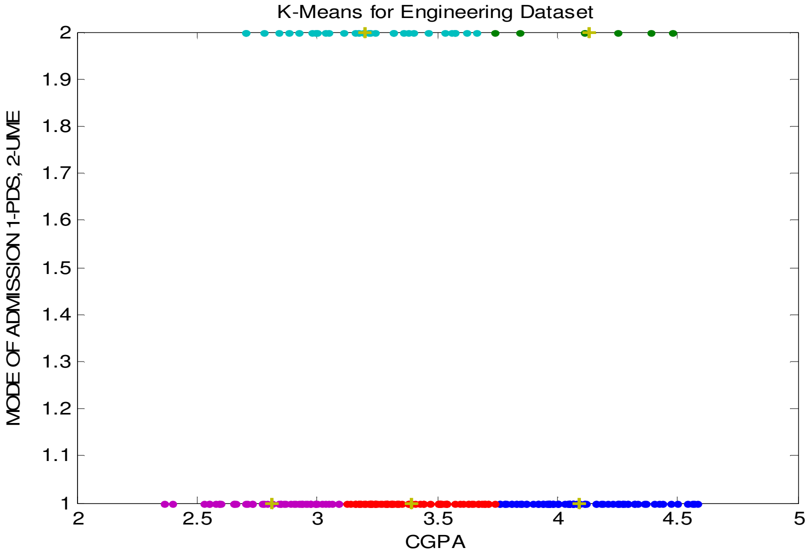

Fig. 3. Graph of mode of admission against CGPA

7 Conclusion

The two clustering algorithms considered in this study are K-means and Expectation-Maximization algorithms. After the evaluation of the two algorithms, K-means was found not to guarantee convergence while Expectation-Maximization's quick convergence doesn't guarantee optimality of results because of its single run characteristics. In view of this, it could be inferred that both algorithms are not efficient enough for the clustering problem considered in this study, hence there arises a need for an algorithm that could both guarantee convergence and optimality of results.

Competing Interests

Authors have declared that no competing interests exist.

References

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining,* edition.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1998). AAAI Press, Menlo Park, California; 1-30.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In LeCam, L. M. and Neyman, J. editors, Proceedings of the Fifth Berkeley

Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA; 1; 281-297.

Sara, N., Rawan, A., Gregory, V. (2006). A modified Fuzzy k-means clustering using Expectation Maximization: IEEE International Conference on Fuzzy Systems, Vancouver BC; 231-235.

Tapas, K., David, M.M., Nathan S.N., Christine, D.P., Ruth, S., Angela, Y.W. (2002). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 7.

Vojtˇech, F., Vaclav, H. (2004). Statistical Pattern Recognition (STPRTool) Toolbox for MATLAB: Research Reports of Centre for Machine Perception (CMP), Czech Technical University, Prague. Retrieved, 16th October 2007 from (www.cmp.felk.cvut.cz).

Kaufman, P., Rousseeuw, J. (1990). Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.

Mehrotra, K., Mohan, C., Ranka, S. (1996). Elements of Artificial Neural Networks. MIT Press.

Dubes, R.C., Jain, A.K. (1988). Algorithms for Clustering Data. Prentice Hall.

Zhang, B., Hsu, M., Dayal, U. (1999). K-Harmonic Means – A Data Clustering Algorithm, Software Technology Laboratory, HP Laboratories, Palo Alto.

© 2012 Adebisi et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.