A Modified Fuzzy K-means Clustering using Expectation Maximization

Sara Nasser, Rawan Alkhaldi, Gregory Vert
Department of Computer Science and Engineering, 171, University of Nevada Reno, Reno NV 89557, USA

Abstract--K-means is a popular clustering algorithm that requires a huge initial set to start the clustering. K-means is an unsupervised clustering method which does not guarantee convergence. Numerous improvements to K-means have been done to make its performance better. Expectation Maximization is a statistical technique for maximum likelihood estimation using mixture models. It searches for a local maxima and generally converges very well. The proposed algorithm combines these two algorithms to generate optimum clusters which do not require a huge value of K and each cluster attains a more natural shape and guarantee convergence. The paper compares the new method with Fuzzy K-means on benchmark iris data.

I. INTRODUCTION

Information or knowledge can be conceptualized as data that represents some meaning. Given a set of objects to be categorized, they are certain attributes of those objects that distinguish them from others, thus forming smaller object groups. Clustering aims at finding smaller similar groups, from a larger collection of items. Computer-assisted analysis must partition objects into groups, and must provide an interpretation for this partitioning [2]. Clustering data has been core to many scientific and engineering problems. Due it's wide applicability in areas of pattern recognition, computer vision and numerous other fields, simple, fast and efficient methods are desirable. Many clustering methods exist to partition a data set by some natural measure of similarity [1]. This similarity measure places similar objects close to one another forming a group, thus several clusters related to objects are formed. An ideal clustering algorithm is one that classifies data such that samples that belong to a cluster are close to each other while samples from different clusters are further away from each other.

Many algorithms for clustering are available. A popular algorithm is the K-means where, based on a given number of clusters the algorithm iterates to find best clusters for the objects. Another well used approach is Expectation Maximization algorithm (EM) [15]. The Expectation Maximization algorithm is the most frequently used technique for estimating class conditional probability density functions (PDF) in both univariate and multivariate cases [23].

This paper discusses both the methods for clustering and presents a new algorithm which is a fusion of fuzzy K-means and EM. The approach desires to come up with a better clustering algorithm.

Section 2 discusses the importance of clustering, its problems and earlier approaches. In section 3, the new approach is presented. Section 4 shows the results obtained and in Section 5 these results are compared with other approaches and analyzed. Finally, Section 6 presents the future work.

II. PREVIOUS WORK

A lot of algorithms have been developed that suit specific domains. K-means clustering is a simple and fast approach to cluster data. The algorithm starts with a large number of seeds (initial prototypes) for the potential clusters. The samples are assigned to each cluster based on its distance from the seed. The centroid is computed for each set and the data points are reassigned. The algorithm runs until it converges or until desired number of clusters is obtained. Though the algorithm does not guarantee convergence, in practice it often converges.

K-means has been widely used in pattern recognition problems. Several variations and improvements to the original algorithm have been done. K-means algorithm by MacQueen [3] is widely used for its simplicity. Another variation of K-means was proposed by Forgy [5]; this algorithm has been shown to converge to a local minimum [6]. Elsewhere it has been showed that there is no guarantee for optimal clustering, since the convergence depends on the initial seeds selected [7]. A large number of seeds can generally lead to an optimal solution, but again this cannot be guaranteed. Improvements to the K-means algorithm were made which dealt with some problems in the simple K-means [7, 8]. K-means however is not considered as the best choice for clustering due it 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 [18].

Fuzzy Logic formularizes an intuitive theory based on human reason of approximation. It differs from the traditional logic methods where crisp or exact results are expected. The concept of fuzzy logic was first put forth by Zadeh [17]. Fuzzy Logic is used in problems where the results can be approximate rather than exact. Hence, the principles of fuzzy logic suit well to clustering problems. The results are determined by some degree of closeness to true or to false. Clustering problems generally measure some kind of closeness between similar objects. Fuzzy Logic has been widely used in various fields to provide flexibility to classical algorithm, due to its applicability to problems that do not require hard solutions. Earlier well known approach to classify data using fuzzy classification is the fuzzy c-means [26]. An improvement of K-means using the fuzzy logic theory was done by Looney [7] in which the concept of fuzziness has been used to improve K-means. Another improvement of fuzzy K-means with crisp regions was done by Watanabe [8]. Fuzzy K-means improves the basic K-means in finding good centers for clusters. An improved version of Fuzzy K-means was proposed and used [25]. This method is divided into steps, where K-means is performed at the first step and a fuzzy maximum likelihood is estimated at the second step. Then performance is measured and the number of clusters increased until convergence. This method combined fuzzy methods with statistical techniques to give optimal clusters.

EM is a model based approach to solve clustering problems. It is an iterative algorithm that is used in problems where data is incomplete or considered incomplete. EM is widely used in applications such as computer vision, speech processing and pattern recognition [20, 21, 22]. EM clusters data, in a manner different than K-means. 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 [19]. EM aims at finding clusters such that maximum likelihood of each clusters parameters is obtained. 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. Several techniques have been adopted to overcome these problems some of them are discussed [23]. Few techniques deal with the selection of initial components based on some criteria. The expectation maximization algorithm is an iterative technique with three major steps. The first step is initializing the hidden variables. The second step estimates the unobserved variables with respect to the known variables. In the third step we compute the maximum likelihood for the unobserved data and then finally check for the stop condition.

The *EM* algorithm extends the basic approach to clustering in two important ways. Instead of assigning cases or observations to clusters to maximize the differences in means for continuous variables, the *EM* clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions [9]. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters. Unlike the classic implementation of K-means clustering, the general *EM* algorithm can be applied to both continuous and categorical variables (note that the classic K-means algorithm can also be modified to accommodate categorical variables) [4].

The K-means clustering algorithm has been proven to be a good approach to classify data. But K-means does not assure the best representation or fit for the data in the model. K-means algorithm uses distances from the centers of clusters to determine which sample belongs to which class. The EM algorithm works well on clustering data when the number of clusters is known. In EM, each observation belongs to each cluster with a certain probability. Our approach combines the two above method to come up with a new method for better clustering. The initial clusters centers are found using K-means algorithm. These give us centers that are widely spread within the data. EM takes these centers as it initial variables and iterates to find the local maxima. Hence, we get clusters that are distributed well using K-means and clusters that are compact using EM. Iris data, which is a well known benchmark for classification problems, is used to test the algorithm.

III. Approach

A. Fuzzy K-means

The K-means clustering algorithms are the simplest methods of clustering data. The K-means algorithm presented by Forgy [5] uses a set of unlabeled feature vectors and classifies them into k classes, where k is given by the user. From the set of feature vectors k of them are randomly selected as initial seeds. The feature vectors are assigned to the closest seeds depending on its distance from it. The mean of features belonging to a class is taken as the new center. The features are reassigned; this process is repeated until convergence.

Due to its simple method of using feature vectors as seeds and arithmetic mean as center for the clusters, K-means algorithm suffers from drawbacks. An improvement to this approach was to start with a huge random population of seeds [8]. This has been shown to find better seeds, since the initial seeds are more than K, and are distributed in the data set. Even though this was an improvement on the simple K-means, it still lacked in finding better centers, since mean does not always represent the center of a given data. A modified K-means [7] was used which uses weighted fuzzy average instead of mean to get new cluster centers. Let $\{x_1,\ldots, x_P\}$ be a set of P real numbers. The number of iteration is given as r. The weighted fuzzy average (WFA) is given by

$$\mu^{(r+1)} = \sum_{p=1,P} w_p^{(r)} x_p\,, \quad r=0,1,\ldots \qquad (1)$$

An initial mean is taken and a Gaussian is centered over the mean and weight w_p is obtained for x_P. Feature vectors are assigned to each seed, empty or small sets are eliminated. Cluster centers are replaced with weighted fuzzy averages and feature vectors are reassigned. This process is repeated until convergence.

B. Expectation Maximization

The main steps of the EM algorithm are shown below. We follow the procedure that was shown earlier [10]. The steps for our implementation of EM are as follows. We have to initialize with a guess for mean and standard deviation. The EM algorithm then searches for a ML hypothesis through the following iterative scheme.

- Initialization step: initialize the hypothesis $\theta^0=(\mu^0_1, \mu^0_2, \ldots, \mu^0_K)$

$$\theta_k^0 = \mu_k^0 \tag{2}$$

Where: K is the current number of Gaussians. σ is the standard deviation, θ^0 is the estimate at 0^{th} iteration, μ is the mean.

- Expectation step: estimate the expected values of the hidden variables z_{ij} (mean and standard deviation) using the current hypothesis $\theta^t=(\mu^t_1, \mu^t_2, \ldots, \mu^t_K)$

$$E(z_{ik}) = \frac{\exp\left[-\frac{(x_i-\mu_k^t)^2}{2\sigma^2}\right]}{\sum_{j=1}^{K}\exp\left[-\frac{(x_i-\mu_j^t)^2}{2\sigma^2}\right]} \tag{3}$$

Where: t is the number of the iteration, $E(z_{ik})$ is the expected value for the hidden variables (namely mean and standard deviation), k is the dimension, σ is the standard deviation.

- Maximization step: provides a new estimate of the parameters.

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik})x_i}{\sum_{i=1}^{n} E(z_{ik})} \tag{4}$$

- Convergence step: if $\| \theta^{t+1} - \theta^{t} \| < e$, stop (finish iteration); otherwise, go to step 2.

The hidden variables are the parameters of the model. In our case we use mixtures of Gaussians; hence our hidden variables are the mean and standard deviation for each Gaussian distribution. We start with an initial estimate of those parameters and iteratively run the algorithm to find the maximum likelihood for our estimates. For convergence we run it number of times so that the values stop increasing.

The reason we are using EM is to fit the data better, so that clusters are compact and far from other clusters, since we initially estimate the parameters and iterate to find the maximum likelihood for those parameters. EM uses the Maximum likelihood, in which it assumes that the parameters are fixed; the best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed. In most cases the observed data could be the samples that are used for training.

C. Fuzzy K-Means Expectation Maximization (FKEM) Algorithm

We use the approach similar to the one presented by Looney [7] to obtain the initial clusters using weighted fuzzy averages, since this method works better than using simple average. We start with the weighted fuzzy K-means averaging algorithm to classify the data into the number of clusters desired, based on its features. The weighted fuzzy K-means algorithm given described above is combined with EM. A large number K of uniformly distributed random seed vectors for the cluster centers are selected. Then we eliminate any seed vectors that are too close to other seed vectors and reduce K (the number of clusters) accordingly. That is done by computing the distance between all the clusters, and eliminating the clusters that their distances are less than ∈(a value that is selected experimentally).

Assigning each of the feature vectors to the nearest random seed vector, is the next step, and it can be achieved by computing the distance between each feature vector and all other seed vectors. Then the feature vector will be assigned to the seed vector such that the distance between them is the shortest. Also, each time an assignment happens the number of feature vectors assigned to that seed vector will be incremented. All seed vectors that are the centers of empty clusters, or have fewer vectors that selected *p* vectors, are eliminated and K is reduced.

Each cluster is then given a new prototype with the current K, and that would be the weighted fuzzy average (WFA) of each class, by initially taking the sample mean $\mu^{(0)}$ and variance σ^2 to start the process. Then center a Gaussian over the current approximate WFA $\mu^{(r)}$ and iterate as follows:

$$w_p^{(r)} = \frac{\exp\left[-(x_p-\mu^{(r)})^2/2\sigma^2\right]}{\sum_{(m=1,P)}\exp\left[-(x_m-\mu^{(r)})^2/2\sigma^2\right]} \tag{5}$$

$$\mu^{(r+1)} = \sum_{(p=1,P)} w_p^{(r)} x_p, \; r = 0,1,2,.....$$

In this step, we calculate the WFA for each cluster to be the new class prototype. The next step is to compute the Maximum Likelihood estimation for the current K clusters using EM as described above and get new centers for each of the clusters. Then each of the feature vectors should be assigned to the class with the nearest weighted fuzzy average. After that, every two clusters whose prototypes are closest are merged, the average of the two prototypes will be used as the new prototype (seed) and K will be reduced accordingly. Next, empty clusters are eliminated, and the number of clusters K is reduced. This process is repeated until we reach the desired number of clusters.

IV. RESULTS

In this section we present the results obtained by the proposed algorithm FKEM. The algorithm is tested on IRIS data. Finally the section compares the results with Fuzzy K-means.

IRIS data first used by Fisher is considered to be a well-known benchmark for classification problems [24], and has been used in several classification tests [13, 14, 16]. The iris data set is a well known data set used for demonstrating the performance of classification algorithms [14]. IRIS data is a standardized data that has 150 feature vectors and can be described to be noisy and non-separable. IRIS data broadly represents two classes (Setosa, Versicolor, and Virginica) of flowers, in which Setosa is in one class and Versicolor and Virginica in the second class. Each sub-class has 50 feature vectors, and each vector has four features: sepal length, sepal width, petal length, petal width. The data needs to be classified in three classes with 50 samples in each class. Each sample has four features. The data needs to be classified into two clusters.

Analyzing the data is a very important step that should be done before any implementation work, and in this case it is important to determine which attributes or variables need to be used as determinants of the different classes. The parameters for our problem are the mean for the centers (equation 5). We start with an initial estimate of those parameters and iteratively run the algorithm to find the maximum likelihood for our estimates. This is an unsupervised method and hence the data is not trained. We use only three features for classification since only three of them can classify the IRIS data.

A Fuzzy K-means was used to classify the data and a comparison of results obtained from Fuzzy K-means and the algorithm proposed in the paper is given. The Initial seeds in both the cases were generated randomly from the feature set.

Table 1. Comparison of results of FKEM with other clustering methods

Method	**# Initial Seeds (K)**	**# Points Classified Incorrectly**
Fuzzy K-means	150	8
FCFM	150	0
FKEM	150	0
Fuzzy K-means	50	15
FKEM	50	0
Fuzzy K-means	30	29
FKEM	30	0
FKEM	10	0

Number of feature vectors that were incorrectly classified is given in the third column. The table shows results obtained from a Fuzzy K-means clustering and FKEM. Results are also compared with a Fuzzy Clustering and Fuzzy Merging (FCFM) approach to clustering [12]. FCFM is an interactive approach that asks users to make choices such as merging clusters, elimination of small clusters, etc.

V. CONCLUSIONS

The paper reviews the problems with simple K-means, and suggests improvements to the method. We desire to obtain optimal clusters with a method that is stable. K-means is considered to be one of the simple/fast methods to cluster data. Expectation Maximization is used to fit data better when the distribution or model of the data is known. When these two are combined we will get a clustering method that not only fits the number of clusters but also tries to make them compact and more meaningful. Statistical techniques have been combined before with fuzzy logic theory and have shown to yield good results. K-means can find seeds in the global space, whereas EM finds local maxima. Using the approach to get the maximum likelihood of the seeds found by K-means, we start with a big set and get local maxima around each one of them, thus increasing the chances of finding the best centers. Using the new approach, the Gaussian for the clusters fits the data better, since their parameters (mean and standard deviation) are computed based on the data and not randomly. Using EM along with Fuzzy K-means will make the classification process slower. Since the new method finds results for small value of K selected initially we can argue that we reduce the number of iterations overall.

From the results obtained we can conclude that Fuzzy K-means was not able to cluster data correctly when the initial value of K was small. If these K seeds are not good (far from the optimal centers, not scattered well), Fuzzy K-means may not find optimal clusters, since the algorithm uses the average (WFA) of these initial K seeds. If these K seeds are scattered well enough the algorithm may perform well, but this cannot be guaranteed. Hence, if good seed centers are not chosen, Fuzzy K-means will not perform very well. Since we only use the centers chosen initially and get centroids for the data. If the correct random seeds are not chosen, EM can still perform to make the K-means work correctly, since EM will iterate to find best centers for the given data. This suggests that Fuzzy K-means along with EM gives a better clustering, than Fuzzy K-means. K-means has been used widely to initialize seeds for EM, we combine these methods and hence can get both good initial clusters. With K-means convergence is not guaranteed and EM guarantees elegant converges.

VI. FUTURE WORK

Both EM and K-means require the number of clusters to be known. It is ideal to have an approach that determines the number of clusters based on the data. EM uses fixed number of mixtures (Gaussians) to represent data. We would like to expand this idea to use weighted fuzzy average in EM to determine the number of mixtures/clusters. Another extension would be to use validation methods to find the right number of clusters. Using a log likelihood function could be a possible cluster validation method.

REFERENCES

[1] M.S Aldenderfer, R.K. Blashfield, Cluster Analysis, Sage Publications, Beverly Hills, USA, 1984.

[2] M. J. A. Berry, G. Linoff, Data Mining Techniques- for Marketing, Sales and Customer Support. John Wiley & Sons, NY, USA, 1997.

[3] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Probability Statistics, University of California Press, Berkeley, 1967, pp. 281–297.

[4] P. S. Bradley, Usama M. Fayyad , Refining Initial Points for K-Means Clustering, Proc. 15th International Conf. on Machine Learning, 1998.

[5] E. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics, 21, 1965, 768–776.

[6] S. Z. Selim, M. A. Ismail, k-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Analysis Machine Intelligence, 6, 1984, 81–87.

[7] C. G. Looney, Interactive clustering and merging with a new fuzzy expected value, Pattern Recognition Lett., vol. 35, 2002, 187–197.

[8] N. Watanabe, T. Imaizumi, Fuzzy k-means clustering with crisp regions, The 10th IEEE International Conference on Fuzzy Systems, Melbourne, 2001, pp.199-202.

[9] The statistic homepage, http://www.statsoftinc.com/textbook/stathome.html, January 17, 2005.

[10] The Expectation Maximization Algorithm, http://www.cs.unr.edu/~bebis/MathMethods/EM/lecture.pdf, September 25, 2004.

[11] Hyungsuck Cho, Opto-Mechatronic Systems Handbook: Techniques and Applications, CRC Press, KAIST, Taejeon, South Korea, 2002

[12] Programs and Data for Downloading, http://ultima.cs.unr.edu/programs.htm, October 4, 2004

[13] Z. Ghahramani, M. Jordan. Supervised learning from incomplete data via an em algorithm. Advances in Neural Information Processing Systems, vol. 6, 1994, 120–127.

[14] S Gunn, Support Vector Machines for Classification and Regression, Technical Report. Image, Speech and Intelligent Systems Group University of Southhampton, 1998.

[15] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of Royal Statistics Society, vol. B-39, 1977.

[16] H. Surmann, A. Kanstein, K. Gosser, Self-Organizing and Genetic Algorithms for an Automatic Design of Fuzzy Control and Decision Systems, EUFIT 93 First European Congress on Fuzzy and Intelligent Technologies, Aachen, Vol. 2, 1993, pp. 1097-1104

[17] L. A. Zadeh, Fuzzy logic and approximate reasoning. Synthese, vol. 30, 1975, 407--428.

[18] Vladimir Estivill-Castro, Jianhua Yang , A Fast and Robust General Purpose Clustering Algorithm, Pacific Rim International Conference on Artificial Intelligence, 2000, pp: 208-218.

[19] P. Bradley, U. Fayyad, C. Reina, Scaling EM (Expectation Maximization) Clustering to Large Databases, Microsoft Research Report, MSR-TR-98-35,Aug 1998.

[20] Wael Abd-Almageed, Christopher E. Smith, Mixture Models for Dynamix Statistical Pressure Snakes, in Proc. IEEE International Conference on Pattern Recognition, August 2002, pp. 721–724.

[21] A.V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, C. Mao, and K. Murphy, A coupled HMM for audio-visual speech recognition, in International Conference on Acoustics, Speech and Signal Processing (CASSP'02), Orlando, FL, USA, May 2002, 13-17.

[22] A. Abu-Naser, N.P. Galatsanos, M.N. Wernick, D. Schonfeld, Object recognition based on impulse restoration with use of the expectation-maximization algorithm, Journal of the Optical Society of America A (Optics, Image Science and Vision), vol. 15, no. 9, September 1998, pp. 2327 – 40.

[23] Wael Abd-Almageed, Aly El-Osery, Christopher E. Smith, Non-Parametric Expectation Maximization: A Learning Automata Approach, IEEE Conference on Systems, Man and Cybernetics, Washington DC, 2003.

[24] Fisher,R., The use of multiple measurements in taxonomic problems. Ann. Eugenics, **7**, 1936, 179–188.

[25] I. Gath, and A.B. Geva, Unsupervised Optimal Fuzzy Clustering *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, pp. 773-781, 1989.

[26] Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.