International Journal of Computational Intelligence Systems, Vol.3, No. 3 (September, 2010), 274-279

Improved Fuzzy Art Method for Initializing K-means

Sevinc Ilhan[*]
Computer Engineering Department, Kocaeli University, Umuttepe Campus
Kocaeli, Turkey[†]

Nevcihan Duru
Computer Engineering Department, Kocaeli University, Umuttepe Campus
Kocaeli, Turkey
E-mail: nduru@kocaeli.edu.tr
www.kocaeli.edu.tr

Esref Adali
Computer Engineering Department, Istanbul Technical University, Ayazaga Campus
Istanbul, Turkey
E-mail: adali@itu.edu.tr
www.itu.edu.tr

Received: 06-07-2009
Accepted: 01-12-2009

Abstract

The K-means algorithm is quite sensitive to the cluster centers selected initially and can perform different clusterings depending on these initialization conditions. Within the scope of this study, a new method based on the Fuzzy ART algorithm which is called Improved Fuzzy ART (IFART) is used in the determination of initial cluster centers. By using IFART, better quality clusters are achieved than Fuzzy ART do and also IFART is as good as Fuzzy ART about capable of fast clustering and capability on large scaled data clustering. Consequently, it is observed that, with the proposed method, the clustering operation is completed in fewer steps, that it is performed in a more stable manner by fixing the initialization points and that it is completed with a smaller error margin compared with the conventional K-means.

Keywords: Clustering, K-means clustering, initial center determination, Improved Fuzzy ART method.

1. Introduction

Clustering is one of the important tools of knowledge discovery. In clustering process, the similar data are grouped with different unsupervised algorithms. Cluster analysis construct good cluster when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity).[1] Clustering techniques have been studied extensively in the areas of data mining, machine learning, pattern recognition, image classification, document retrievals, biological sciences etc.

The most well known algorithm for clustering is K-means developed by Mc Queen in 1967. The simplicity of K-means made this algorithm used in various fields. On the other hand, K-means can cluster huge data and also outliers quickly. K-means is a partitioning algorithm that divides data into K groups. By iterative such partitioning, K-means minimizes the sum of

[*]Computer Engineering Department, Kocaeli University, Umuttepe Campus, Kocaeli, Turkey
[†]Turkey, Kocaeli University, E-mail: silhan@kocaeli.edu.tr.

distance from each data to its clusters. In K-means algorithm, each cluster can be represented by its center. So before the iterative procedure, we initialize its K cluster centers firstly. Then we can continuously update them through the iterative procedure. It is noted that the initial cluster centers will directly affect the results of clustering. So how to select good initial clustering centers is an important issue for K-means algorithm.[2] Conventional K-means generates initial cluster centers randomly. When initial starting points close to the final solution, K-means has high possibility to find out the cluster center. Otherwise, it will lead to incorrect clustering results.[3] Briefly, the performance of K-means strongly depends on the initial guess of partition.

In the literature several methods proposed to solve the cluster initialization problem for K-means. A recursive method for initializing the means by running K clustering problems is discussed by Duda and Hart.[4] A variant of this method consists of taking data and then randomly perturbing it K times.[5] Bradley and Fayyad[6] proposed an algorithm that refines initial points by analyzing probability of data density. Shehroz and Ahmad[7] proposed Cluster Center Initialization Algorithm (CCIA) to solve cluster initialization problem. Su and Dy[8] proposed a deterministic initialization method for K-means based divisive hierarchical approach. Kohei and Barakbah[9] proposed a hierarchical K-means algorithm as a new approach to determine the centers initialization for K-means.

The developed methods are approximately categorized into three groups which are random sampling methods, distance optimization methods and density estimation methods, respectively.[10] Random sampling methods perhaps are the most widely used methods which usually initialize the clustering centers either by using randomly selected input samples, or random parameters non-heuristically generated from the inputs.[2] The main problem with random methods is that do not guarantee obtaining the optimal solution.

In this paper, Improved Fuzzy ART (IFART) is proposed as a deterministic initialization method for K-means and demonstrated that why K-means initialized with IFART is a favorable method.

This paper is organized as follows: In section 2, the basic theory of k-means algorithm is described. In Section 3, the application of IFART is described. In section 4, the clustering error estimation index for deciding the valid clustering is described. Then, experimental results are reported in Section 5. Finally, in Section 6 the conclusions are drawn.

2. K-means Algorithm

The steps of K-means algorithm are as follows:

(i) Choose K input data points (vectors) to initialize K clusters.
(ii) For each input vector, find the closest center, and assign that input vector to the corresponding cluster. Euclidean distance can be used to express the distance.
(iii) Update the cluster centers in each cluster using the mean of the input vectors assigned to that cluster.
(iv) Repeat steps (ii) and (iii) until no more change in the value of the means.

3. The Proposed Algorithm: IFART

3.1. *Fuzzy Adaptive Resonance Theory (Fuzzy ART)*

Fuzzy ART neural network was introduced by Carpenter et al. (1991) and it is unsupervised category learning and pattern recognition network which incorporates computation from fuzzy set theory into ART based neural network. It's capable of rapid stable clustering of continuous input patterns and more effective for large scaled data clustering. It requires less processing time from other algorithms for clustering purpose. Fuzzy ART is a pure winner-takes-all

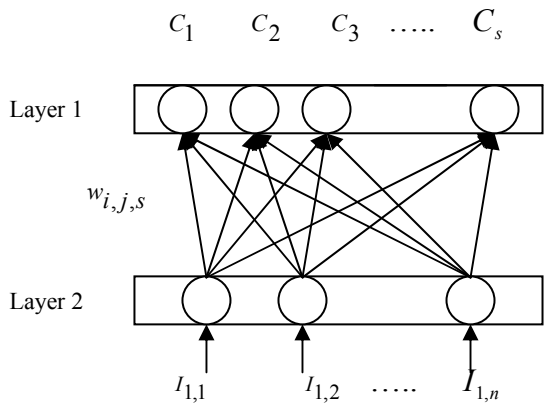

Fig. 1. The Fuzzy ART model.

architecture able to instance output nodes whenever necessary and to handle both binary and analogue patterns.[12] The Fuzzy ART model is shown in Fig. 1.

3.2. *Improved Fuzzy ART (IFART)*

The Fuzzy ART algorithm is able to create very different clustering results depending on the sequence of the data. On examining the cluster results, it was observed that effective and valid clustering could not be achieved, and that in some cases the clusters were even overlapping

After clustering is performed with Fuzzy ART, the membership degree of each input data to each cluster formed with Fuzzy ART is calculated. This calculation is made based on the cluster centers. The calculated values are stored in a membership matrix in which the rows represent the input data and the columns represent the clusters created. Each input data is transferred to the cluster to which it has the maximum membership according to the membership matrix. Thus, a moving operation is performed on the elements of the clusters formed with Fuzzy ART.

The step by step explanation of the IFART method is as follows:

Step 1. Each input value $I_{i,j}$ is normalized as in (1).

$$NI_{i,j} = \frac{I_{i,j} - \min(j)}{\max(j) - \min(j)} \quad (1)$$

i is candidate input number, j is the input evaluation criterion number, $NI_{i,j}$ represents the normalized input value. n is number of attributes.

Step 2. Initialize all the parameters: The necessary values for the choice (α), vigilance (ρ), and learning rate (β) parameters are assigned.

Step 3. Initially all weights are taken 1 and the number of the cluster is set as 1.

For all i,j $w_{i,j,s}(0) = 1$ and s=1.

Step 4. Input vector which normalized in the range [0, 1] is designated to network.

Step 5. Choice function $T_{i,s}$ is defined with the Eq. (2).

$$T_{i,s}(NI) = \frac{\sum_{j=1}^{n}\left(NI_{i,j} \wedge w_{i,j,s}\right)}{\alpha + \sum_{j=1}^{n} w_{i,j,s}} \quad (2)$$

Where, "$\wedge$" is fuzzy "AND" operator and (x $\wedge$ y) = min (x, y). α acts as a tie breaker when multiple prototype vectors are subsets of the input patterns and favors larger magnitude prototypes.

Step 6. Select the best matching exemplar with Eq. (3).

$$T^{*} = \max\left\{T_{i,s}, s=1,2,.....,m\right\} \quad (3)$$

Step 7. Matching test determines the appropriate cluster for the input. Matching function is computed as in (4).

$$M_{i,s}\left(T^{*}\right) = \frac{\sum_{j=1}^{n}\left(NI_{i,j} \wedge w_{i,j,s}\right)}{\sum_{j=1}^{n} NI_{ij}} \quad (4)$$

If $M_{i,s} \geq \rho$ then $T_{i,s}$ is passing the test. So the i^{th} input is added to existing cluster C_S and go to step 9. $M_{i,s} < \rho$ then $T_{i,s}$ not passing the test then go to step 8.

Step 8. Set the choice function value as $T_{i,s}$ = -1 and go to step 6. In this way, matching test continues for all of the $T_{i,s}$ values.

If none of $T_{i,s}$ pass the matching test a new cluster is created for existing input. So the i^{th} input is added to the new cluster C_{s+1}. Go to step 4 and continue with the next input.

Step 9. Update best matching exemplar (learning law).

$$W_{i,j,s}^{new} = \beta\left(NI_{i,j} \wedge W_{i,j,s}^{old}\right) + (1-\beta)W_{i,j,s}^{old} \quad (5)$$

Step 10. The algorithm continues with the next input at step 4. Stop if all data is allocated to s different clusters .

Step 11. Cluster centers are found for calculating membership degrees. Center of cluster s is calculated as in (6).

$$V_S = \frac{\sum_{i=1}^{count} I_{i,s}}{count} \quad (6)$$

$I_{i,s}$: (i=1, 2,…, count) elements of cluster s.

Membership degree of data instance i to cluster s, is defined as in (7).

$$\mu_{i,s} = \frac{\left[\frac{1}{d\left(I_i, C_s\right)}\right]^{1/(q-1)}}{\sum_{k=2}^{K}\left[\frac{1}{d\left(I_i, C_k\right)}\right]^{1/(q-1)}} \quad (7)$$

K is cluster number, q is a constant and chosen as 2. The Euclidean distance between two points is defined as in (8)

$$d\left(I_i, C_k\right) = \sqrt{\sum_{p=1}^{n}\left[A_p\left(I_i\right) - A_p\left(V_k\right)\right]^2} \quad (8)$$

A_p is p^{th} attribute of input pattern (p=1, 2…, n). n is number of attributes. V_k is center of cluster k.

Step 12. According to max membership degree the data instances are moved between clusters.

To perform the above mentioned IFART method is a computer program coded in MATLAB 7.1.

In the following section we introduce the definition of the clustering error estimation index.

4. Clustering Error Estimation Index

The main concern of data partitioning is how to correctly divide the data points into clusters. There are a number of indices proposed in literature to assess the performances of data clustering. The main ideas are twofold: (1) data points within the same cluster should locate as close as possible, and (2) data points in different clusters should be as apart as possible. Based on the two concepts, a variety of the cluster validity indices are proposed.[13] Many criteria have been developed for determining cluster validity[14,15] all of which have a common goal to find the clustering which results in compact clusters that are well separated.[16]

In this paper, an error estimation index (e) to measure the quality of clusters is used and defined as in (9).

$$e = \frac{D_{\text{intra}}}{D_{\text{inter}}} \quad (9)$$

Here, D_{intra} defines the average intra cluster distance index; D_{inter} defines the average inter cluster distance index. These two parameters are explained as in (10) and (11).[17]

$$D_{\text{intra}} = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{I_i \in C_k}\frac{d\left(I_i, C_k\right)}{count\left(C_k\right)}\right) \quad (10)$$

$$D_{\text{inter}} = \frac{2}{K(K-1)}\sum_{\substack{k=1 \\ k'=k+1}}^{K}\left(d\left(C_k, C_{k'}\right)\right) \quad (11)$$

Here, K is the number of clusters which are obtained, count (C_k) is the number of data instances which are classified to cluster k. C_k is cluster k. and $C_{k'}$ is cluster k'. With D_{intra}, the distance of each cluster's elements to the cluster center; with D_{inter}, the distance between cluster centers is defined.

A better clustering algorithm produces clusters with higher internal compactness degree (less D_{intra}) and lower similarity degree among clusters (larger D_{inter}). It can be also declared, a better clustering will be resulted in a smaller error estimation index (e).

5. Experimental Results

In this section, the performance of IFART is compared with that of the classical initialization methods (random seed) based on the following criteria: First, quality: The quality of the clustering is quantified using clustering error estimation index. Secondly, speed: The speed of K-means is evaluated through the number of iterations needed for updating the cluster centers. And thirdly, stability: the randomly initialized K-means algorithm was run with 100 different initialization points, and the clusters formed for each initialization were analyzed.

Two different initialization schemes are compared on three real datasets and four synthetic datasets. Real datasets: Iris, Ruspini, Heart-Disease-Cleveland (HDC), Haberman's Survival (HS), Letter Recognition (LR) are taken from the UCI Machine Learning Repository[18] shown in Table 1. The synthetic datasets: Web Logs (WL), Documents_Sim (DS), Mars, and Image Extraction (IE) are taken from the databases prepared by Pei and Zaiane[19] at Canada's Alberta University, Department of Computer Sciences.

Table 1. Features of Real Datasets

Dataset	Number of Samples	Number of Attributes	Number of Clusters
Iris	150	4	3
Ruspini	75	2	4
HDC	303	14	3
HS	306	4	2
LR	20000	16	4

Conventional K-means and proposed K-means methods were performed on synthetic datasets by choosing K=3. The number of clusters for real datasets are been shown as in Table 1. The error estimation indexes related to the clustering performed are shown in Table 2. The values in the table were obtained by randomly selecting the initialization points for the conventional K-means algorithm and running for 100 different initialization points. The minimum, maximum and average clustering errors occurring as a result of these runs are shown in Table 2, columns 1, 2 and 3 respectively. The last column of the table shows the clustering error belonging to the K-means algorithm initialized with IFART.

Table 2. Error estimation indexes for all datasets

Dataset	Conventional K-means			Proposed K-means
	Min.	Max.	Average	
WL	0.0471	0.1400	0.0713	0.0538
DS	0.0593	0.1053	0.0832	0.0702
Mars	0.0778	0.1259	0.1020	0.0786
IE	0.0578	0.1318	0.0798	0.0672
Iris	0.0410	0.0964	0.0613	0.0455
Ruspini	0.0151	0.0450	0.0224	0.0192
HDC	0.0975	0.1850	0.1387	0.1381
HS	0.4809	0.6502	0.4968	0.4809
LR	0.1172	0.1685	0.1394	0.1227

As seen from the error estimations in Table 2, the K-means algorithm initialized with the IFART algorithm performs clustering with a smaller error margin in all datasets compared to the conventional K-means algorithm.

In this study, for K-means algorithm, the number of times the cluster centers were updated was indicated as the step number of algorithm, and the number of steps required by the algorithms was observed as a separate evaluation criterion. Table 3 shows the number of steps of conventional K-means and the alternative K-means.

Table 3. Step number of algorithms for all datasets

Dataset	Conventional K-means			Proposed K-means
	Min.	Max.	Average ($\cong$)	
WL	6	57	24	21
DS	9	72	27	15
Mars	6	63	31	39
IE	6	57	24	24
Iris	6	42	20	9
Ruspini	4	48	12	4
HDC	12	87	39	24
HS	2	44	15	14
LR	60	564	255	364

On analyzing the results in Table 2, it is observed that in Web Logs (WL), Documents_Sim (DS), Mars, and Image Extraction (IE) datasets, the K-means algorithm initialized with the proposed IFART method completed the clustering in fewer steps than the average number of steps of the conventional K-means algorithm. For the Mars dataset, it is observed that it is completed in more steps than K-means' average number of steps. At this situation, another point which needs to be taken into consideration is the possibility that K-means can complete the clustering in any number of steps within the range [6, 63], depending on the randomly initialized points. In this respect, the IFART initialized K-means algorithm is a more stable algorithm compared to the conventional K-means algorithm.

6. Conclusions

It is accepted that the K-means algorithm suffers from initial cluster centers. Our main purpose is to optimize the initial centers for K-means algorithm. Therefore, in this paper, the initialization cluster centers of the K-means algorithm were determined using the proposed Improved Fuzzy ART method. In this method, the data is first clustered using the Fuzzy ART algorithm. Then, the membership degree of each data object in each cluster is calculated. The element with the maximum membership degree is determined as the center of the cluster. The K-means algorithm is initialized with these cluster centers. K-means algorithms run with randomly selected cluster centers and run with cluster centers determined using IFART were applied to four real datasets and four synthetic datasets. In the application, the conventional K-means algorithm was run for 100 different initialization points. The results were analyzed in terms of the number of times the algorithms updated the cluster centers and the error margins in clustering.

The number of times the cluster centers were updated was indicated as the number of steps.

The conventional K-means algorithm is able to perform clustering with a very low or very high error depending on the initialization points selected. The same applies to the number of times the algorithm updates the cluster centers. The K-means algorithm initialized using the proposed method performs a more stable clustering operation as the initialization points are predetermined. In addition, it performs clustering with a smaller error margin than the average error margin of the conventional K-means algorithm. Furthermore, it is completed in fewer steps compared to the average number of steps of the conventional K-means algorithm.

The experiment results confirm that with the additional time allocated to the IFART algorithm at initialization, the K-means algorithm became a more stable and faster algorithm which runs with fewer errors.

7. References

1. V.E. Castro, *Why So Many Clustering Algorithms a Position Paper*, ACM SIGKDD Explorations Newsletter, **4**(1) (2002) 65–75.
2. S-Z. Yang and S-W. Luo, *A novel algorithm for initializing clustering centers,* Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou (2005) 5579–5583.
3. Y.M. Cheung, *k*-Means: A new generalized k-means clustering algorithm*, Pattern Recognition Lett. 24 (2003) 2883–2893.
4. R.O. Duda and P.E. Hart, *Pattern classification and scene analysis,* (John Wiley and Sons, NY, 1973).
5. B. Thiesson, C. Meck, D. Chickering and D. Heckerman, *Learning mixtures of Bayesian networks*, (Microsoft Research Technical Report TR-97-30, Redmond, WA. 1997).
6. P.S. Bradley and U.M. Fayyad, *Refining initial points for k-means clustering,* In: Sharlik, J. (Ed.), Proc. 15th Internat. Conf. on Machine Learning (ICML'98). Morgan Kaufmann, San Francisco, CA, (1998) 91–99.
7. S.K. Shehroz and A. Ahmad, *Cluster Center Initialization Algorithm for K-means Clustering*, Pattern Recognition Letters, 25 (2004) 1293–1302.
8. S. Ting and D. Jennifer, *A Deterministic Method for Initializing K-means Clustering*, IEEE International Conference on Tools with Artificial Intelligence, (2004) 784–786.
9. K. Arai and A.R. Barakbah, *Hierarchical K-means: An Algorithm for Centroids Initialization for K-means*, Reports of the Faculty of Science and Engineering, Saga University, **36**(1) (2007) 25–31
10. J. He, M. Lan, C-L. Tan, S-Y Sung and H-B Low, *Initialization of cluster refinement algorithms: A review and comparative study*, Proceedings of International Joint Conference on Neural Networks, (2004).
11. G.A. Carpenter, S. Grossberg and D.B. Rose, *Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system*, Neural Networks, 4 (1991.) 759–771.
12. P.K. Simpson, *Fuzzy Min-Max neural networks-Part 2: clustering*, IEEE Trans. Fuzzy Systems **1**(1) (1993). 32–45.
13. L. Sun, T. Lin, H. Huang, B. Liao and J. Pan, *An optimized approach on applying genetic algorithm to adaptive cluster validity index*, Proceedings of the Third International Conference on International Information Hiding and Multimedia Signal Processing 02, (2007) 582–585.
14. Y. Xu, G. Richard, and A. Brereton, *A comparative study of cluster validation indices applied to genotyping data*, Chemometrics and Intelligent Laboratory Systems, 78, (2005) 30–40.
15. K.L. Wu and M.S. Yang, *A cluster validity index for fuzzy clustering*, Pattern Recognition Letters, 26, (2005) 1275–1291.
16. M. El-Melegy, E.A. Zanaty, W.M. Abd-Elhafiez and A. Farag, *On cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation***,** Image Processing, IEEE International Conference on Volume 6, (2007) VI - 5 - VI – 8.
17. C. Chen and L. Wang, *An efficient and applicable clustering algorithm using Fuzzy ART*, IEEE Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China. (2006) 3178-3182.
18. C. J. Merz, P. Murphy and D. Aha, *UCI repository of machine learning databases*, (1996).
19. Y. Pei and O. Zaiane, *A synthetic data generator for clustering and outlier analysis*, Department of Computing Science, University of Alberta, Edmonton, AB, Canada. (2006).