

FACT: A New Fuzzy Adaptive Clustering Technique

Faezeh Ensan, Mohammad Hossien Yaghmaee, and Ebrahim Bagheri

Department of Computing, Faculty of engineering
Ferdowsi University of Mashhad, Mashhad, Iran
Fa_En93@stu-mail.um.ac.ir, hyaghmae@um.ac.ir,
Eb_ba63@stu-mail.um.ac.ir

Abstract. Clustering belongs to the set of mathematical problems which aim at classification of data or objects into related sets or classes. Many different pattern clustering approaches based on the pattern membership model could be used to classify objects within various classes. Different models of Crisp, Hierarchical, Overlapping and Fuzzy clustering algorithms have been developed which serve different purposes. The main deficiency that most of the algorithms face is that the number of clusters for reaching the optimal arrangement is not automatically calculated and needs user intervention. In this paper we propose a fuzzy clustering technique (FACT) which determines the number of appropriate clusters based on the pattern essence. Different experiments for algorithm evaluation were performed which show a much better performance compared with the typical widely used K-means clustering algorithm.

Keywords: Fuzzy Clustering, Unsupervised Classification, Adaptive Pattern Categorization, Fuzzy C-means.

1 Introduction

Clustering belongs to the set of mathematical problems which aim at classification and assignment of data or objects to related sets or classes. The act of classification could well be applied through supervised or unsupervised learning methods [1]. In the Supervised model, Patterns are learnt using some familiar, previously classified data. Multi-layered Perceptron - MLP, Support Vector Machine - SVM, and Decision Trees are illustrious examples of such learning algorithms. This type of learning may be called the learning by example methodology. On the other hand, and in the unsupervised method, mainly named clustering, entities are classified in homogeneous classes so that neighboring patterns are assembled in similar collections. In this approach object-class association is not previously known and clusters are formed based on some object similarity criteria.

As unsupervised learning models, such as clustering, can semi-consciously detect well separated classes amongst available data based on their intrinsic features, they have been extensively used in different scientific fields. Their use can significantly vary from Medicine and its application to disease detection, to Intrusion Detection Systems (IDS) for network activity division into two typical types of intrusive and

non-intrusive. New applications of clustering have been found in data (web) mining and adaptive systems where user characteristics modeling, session detection and etc can be achieved through modified clustering algorithms. Pattern recognition can also be an important field of clustering application.

Most clustering techniques assume a well defined distinction between the clusters so that each pattern can only belong to one cluster at a time. This supposition can neglect the natural ability of objects existing in multiple clusters. For this reason and with the aid of fuzzy logic, fuzzy clustering can be employed to overcome this weakness. The membership of a pattern in a given cluster can vary between 0 and 1. In this model one single pattern can have different degrees of membership in various clusters. A pattern belongs to the cluster where it has the highest membership value.

In this paper we aim to propose a fuzzy clustering technique which is capable of detecting the most appropriate number of clusters based on a density factor. This algorithm is completely insensitive to the initial number of employed clusters; however the initial value should always be lower than the optimal cluster number. Although a very low number of initial clusters will increase the computation time and CPU usage but it will prevent the algorithm from choosing the incorrect number of clusters. The method discovers the number of clusters by intelligently splitting capable clusters and creating new cluster centers through outlier detection.

The paper is organized in the following sections: Section 2 will introduce the proposed fuzzy clustering heuristic. Sections 3 presents experimental results obtained from the algorithm implementation and the following section will conclude and provide related works in the field.

2 FACT Heuristic

Although Fuzzy C-means algorithm shows strengths in many areas but it lacks the ability to determine the appropriate number of clusters for pattern classification and requires the user to define the correct number of clusters. Many applications of clustering like pattern recognition or intrusive data classification require the clustering algorithm to decide on the proper number of clusters, as the correct number of classes is not a priori known. In the proposed heuristic we devise an algorithm which exploits a modified version of Fuzzy C-means in which U (Membership Degree Matrix) is not randomly initialized. The other two strengths of this heuristic is that it is based on a fuzzy split-outlier detector and a Cluster Density Criterion (CDC). The fuzzy split algorithm was to some extent inspired by [8]. Some fuzzy clustering algorithms such as [9] are based on the minimization of the objective function value as their ultimate goal. This criterion serves as a great factor for the algorithms with a predefined number of clusters; however in heuristics which have an adaptive approach to cluster number assessment, this factor cannot be used. This is because the objective function will decrease with the increase of the number of clusters and hence causes further cluster splitting which results in an incorrect number of clusters (the number of clusters will most likely end up being identical to the number of available patterns). For this reason using the objective functions as the basis for successful split assessment is unreasonable. We define and apply CDC for split success comparison. FACT is comprised of 3 main steps which are further explained in the following paragraphs:

Step 1 – Initialization

The existing version of the Fuzzy C-means is applied to the set of available patterns by setting the initial cluster number and m to 2. The outputs of this step are the preliminary values for U and CDC.

Step 2 – Outlier Detection

a) Cluster Member Assignment

Every pattern in the fuzzy clustering algorithm has a membership degree in all available clusters. The process of pattern to cluster assignment is done through allocating the pattern to the cluster in which it has the highest membership degree. Matrix M, $[m_{ij}]_{c \times n}$ is defined as follows:

$$M_{ij} = \begin{cases} U_{ij}, & \text{if } \text{Max}_{i=1}^c U_{ij} = U_{ij} \\ 0, & \text{else} \end{cases} \tag{6}$$

b) Local Outlier Detection

In this sub step the candidates in each cluster to be the final outliers over all of the patterns are selected. This process selects the pattern with the lowest non-zero membership value in vector M_i where i shows the current cluster (7).

$$\text{Candidate}_i = \text{Min}_{j=1}^n (M_{ij}) \quad \text{where } M_{ij} \neq 0 \tag{7}$$

c) Final Outlier Selection and Splitting

The pattern with the lowest value in the Candidate vector (OP) is selected as the ultimate outlier. The coordinates of OP are used as the basis for the center of a new cluster. Let $OP = \{op_1, op_2 \dots op_r\}$ be the outlier point, the new cluster center will be calculated using (Eq.8):

$$\text{Center (c+1)} = OP + \lambda \tag{8}$$

Where $\lambda = (\lambda_1, \lambda_2 \dots \lambda_r) \sim 0$.

Having calculated the value of the new cluster center, the previous composition of pattern classifications can be altered and rearranged based on c+ 1 cluster. Matrix U is updated using (Eq. 4) where the upper bound of k is c+1. The modified version of fuzzy C-means is now tuned using the calculated U and c+1 number of clusters and is used to create the new cluster composition. After having split the cluster formation into a new arrangement, the CDC will be updated (Eq. 9). The value obtained from the division of the new CDC to the former CDC is multiplied by a coefficient, α , which is between 0 and 1. To show that splitting has improved the clustering, θ_{t+1} should be larger than θ_t and thus the splitting procedure is confirmed and stabilized. The value for α is usually set to 0.2. The θ is named the Feedback Control Parameter (FCP) which controls the system behavior using a feedback from the prior iteration.

$$CDC_{t+1} = \sum_{i=1}^{c+1} \sum_{j=1}^n M_{ij} \tag{9}$$

$$\theta_{t+1} = \alpha \left(\frac{CDC_{t+1}}{CDC_t} \right) + (1 - \alpha) \theta_t \quad (10)$$

If the splitting has been unsuccessful 2.c is repeated with the next pattern in the Candidate vector.

Step 3 – Test

If none of the patterns available in the Candidate vector can serve as a successful splitting point for improving the current cluster arrangement, the algorithm will terminate with the current composition on hand else it will increase the number of clusters by one unit and resume algorithm execution from 2.a.

3 Experimental Results

Four main pattern sets were used in the first set of experiments. Wisconsin Breast Cancer Databases containing 699 patterns were cleaned to be used in the comparison procedure. The patterns were 9 dimensional data with 2 main classes (malignant and benign). Pima Indians Diabetes Database was the second pattern set used which included 768 patterns with 8 attributes for each pattern. The training was done to test positive or negative diabetes tests. The patterns were initially obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The third pattern set was the Liver-disorders Database from the BUPA Medical Research Ltd. This pattern set consisted of 345 patterns each with 7 numeric-valued attributes. The Statlog Project Heart Disease Database is made up of 270 patterns which are used to classify normal and abnormal patients using 13 different traits. The four pattern sets were taken from the UCI Machine Learning Repository at [10]. The pattern sets were divided into two parts for train and test purposes. The exact division is shown in table 1.

Table 1. The number of Train and Test patterns used in each pattern set

Pattern Set	Patterns	Train Patterns			Test patterns		
		<i>Sum</i>	<i>First class</i>	<i>Second class</i>	<i>Sum</i>	<i>First class</i>	<i>Second class</i>
Wisconsin Breast Cancer	683	400	303	197	183	141	42
Pima Indians Diabetes	768	500	318	182	268	182	86
Liver-disorders	345	250	110	140	95	35	60
Statlog Project Heart Disease Database	270	220	124	96	50	26	23

For the sake of clarity and to show K-means' dependency to the number of clusters, different variations of cluster numbers were created for k-means performance evaluation ranging from 2 to 34 clusters. Figure 1 depicts K-means performance under different circumstances. As it can be inferred from the diagram, different number of clusters employed, can significantly affect the final outcome and hence be evaluated as a negative impact on the overall algorithm performance.

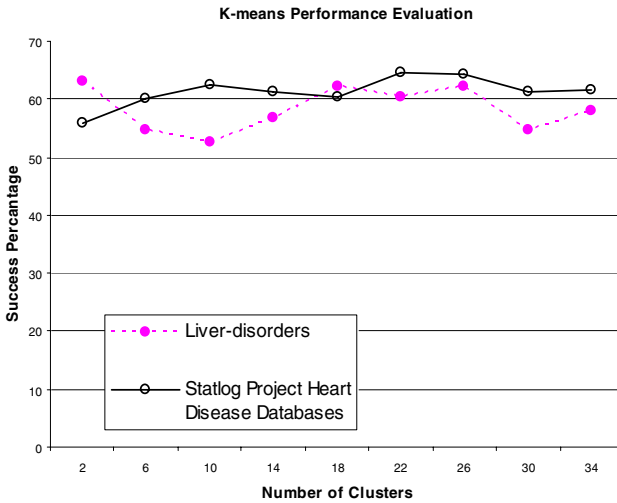


Fig. 1. K-means Performance Evaluation

The proposed heuristic was examined under several different criteria and compared with the K-means algorithm. As the K-means algorithm clusters data using a predefined number of classes and this is different from what the proposed heuristic does, the appropriate comparison model should have been devised. In our analysis, FACT was first applied to the pattern sets for clustering. This step detected the available classes (c) in the pattern sets. K-means was then executed with three different initial cluster numbers of $c-1$, c and $c+1$. Although the evaluations were initially done based upon the number of FACT detected classes but the optimal number of classes known from the omniscient were also applied to the K-means algorithm to compare the optimal success ratio of both heuristics. As K-means provides different clustering compositions each time it is run due to its sensitivity to the initial state; it was executed 100 runs on every pattern set with the specified cluster number and the average results were used. Table 2 compares the performance of both heuristics based on the percentage of correct pattern classification. As it can be clearly seen, the FACT algorithm outperforms the K-means algorithm in 3 of the pattern bases and reaches optimality in the Liver-disorders pattern set. The important point is that due to the differences in the essence of the algorithms the number of appropriate clusters for the FACT might differ from the optimal number of clusters for K-means, but even with the selection of the best number of clusters in K-means, FACT still shows better performance.

The next experiment was done on pattern sets which were statistically created to form well-defined class boundaries. Each pattern set had colonies of patterns consisting of 200 objects. 4, 5 and 7 colonies had been integrated into the pattern sets forming pattern sets with 800 (PS1), 1000 (PS2) and 1400 (PS3) patterns. To compare the performance of each clustering model, the Quadratic Error (QE) factor introduced in [11] was used. Let $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ be the members of cluster I and cc_i be the centroid of the i^{th} cluster, QE is defined as the average of the mean squared distances of each pattern to the cluster centroids as shown in (Eq.11).

$$QE = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{p_l \in C_i} \|p_l - cc_i\|^2}{|C_i|} \tag{11}$$

Table3 compares the values for the QE factor obtained from different heuristics. The smaller the value for the Quadratic Error is, the higher the inter object relationship in one cluster would be. FACT related QE values show much better performance for the proposed algorithm. Figure 2 shows one of the devised pattern sets depicting the cluster centers chosen using each algorithm which reveals a better centroid placement strategy in the FACT algorithm.

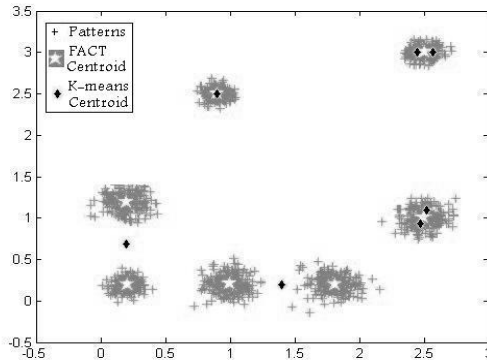


Fig. 2. PS3 and Centroid Placement Strategy

Table 2. The comparison of the two heuristics based on the correct classifications rate. 1 and 2 show the number of detected clusters using the FACT algorithm and the success percentage achieved, respectively. The number of clusters used to reach the best success percentage and the success percentage achieved in K-means are also shown in 3 and 4.

Pattern Set	FACT		K-means			
	Number of Clusters ¹	Success Percentage ²	Number of Clusters	Success Percentage	Optimal Number of Clusters ³	Optimal Success Percentage ⁴
Wisconsin Breast Cancer	8	100.00	7	98.57	2	100.00
			8	98.36		
			9	98.91		
Pima Indians Diabetes	68	71.31	67	67.67	14	69.96
			68	65.10		
			69	66.10		
Liver-disorders	2	63.16	2	63.16	2	63.16
			3	63.16		
			4	53.68		
Statlog Project Databases	9	70.00	8	59.50	18	64.66
			9	60.66		
			10	60.66		

Table 3. QE-based Comparison

Pattern Set	Number of Clusters	Quadratic Error of FACT	Quadratic Error of K-means
PS1	4	0.005934	0.006044
PS2	5	0.0064	0.0306
PS3	7	0.0075	0.0321

4 Conclusion

In this paper we have proposed a fuzzy adaptive clustering algorithm which modifies the well known fuzzy C-means. The Fuzzy C-means algorithm is altered so that it is initialized based on the Membership Degree Matrix from the previous iteration. The number of pattern classes used in the clustering process is adaptively calculated. Comparisons done between the typical K-means algorithm and the proposed heuristic demonstrate a better performance concerning the correct clustering percentage and the Quadratic Error factor. It is notable that the outstanding feature of the FACT algorithm is that it detects the correct number of pattern classes adaptively.

References

1. Everitt, B.S., Landau, S., Leese, M., Cluster Analysis, London: New York, Halsted Press, 1993.
2. Hansen, P., Mladenovic, N., J-Means: a new local search heuristic for minimum sum-of-squares clustering, *Pattern Recognition* 34 (2), 2001, pp.405–413.
3. MacQueen, J., Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol.2, 1967, pp.281-297.
4. Sanjiv, K.B., Adaptive K-Means Clustering, FLAIRS Conference 2004
5. A survey of recent advances in hierarchical clustering algorithms *The Computer Journal*, Volume 26, Issue 4, 1983, pp. 354-359.
6. Barthélemy, J.P., Brucker, F., NP-hard approximation problems in overlapping clustering, *Journal of Classification* 18, 2001, pp.159-183.
7. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
8. Guan, Y., Ghorbani, A., and Belacel, N., K-means+: An autonomous clustering algorithm, in submission.
9. Belacel, N., Hansen, P., and Mladenovic, N., Fuzzy J-means: A new heuristic for fuzzy clustering, *Pattern Recognition* 35, 2002, pp.2193–2200.
10. <http://www.ics.uci.edu/~mllearn/MLSummary.html>
11. Bacao, F., Lobo, V., Painho, M., Self-Organizing Maps as efficient initialization procedures and substitutes for k-means clustering, *International Conference on Computational Science*, 2005