Clustering with Fuzzy C-means and Common Challenges

Songyin Deng
Ulink College Guangzhou, Guangdong, 510000, China
sondeng1905@ulinkcollege.com

Abstract. Clustering algorithm is massively used in various fields such as computer vision and so do FCM. One of the purpose of this paper is to generate new insight into improvement on general clustering algorithms through this inspection of one specific clustering algorithm (FCM) help. Three common challenges of clustering (Noise problem, long operational time, and initial bias) in Fuzzy C-Means algorithm and corresponding solutions to each of these problems are introduced. Further potential of meta learning and adversarial learning to solve the clustering problems are recognized in the end.

1. Introduction
Clustering is the method of data mining that classifies the given set of unlabelled data into a number of clusters. In each cluster, the data should have high similarity with each other and there should be specific features among different clusters. Some common types of clustering algorithms are hierarchical clustering , K-Means Clustering, Fuzzy Clustering, Dynamic Clustering, Two Step Clustering, etc. Clustering today is largely involved into various fields with different tasks on pattern recognition of medical images [1], data segmentation [2], geology [3], agriculture engineering [4], man-less driving [5], drone technology[6], industrial inspection[7] and so on.

However, with the development of clustering application, some problems existing in the clustering are also found. Since the clustering algorithm is aimed to handle all the points into a number of clusters in specific purpose, it is very sensitive to the noise in dataset. The problem of noisy data not only imposes restriction on usage of clustering in different data mining tasks, and also brings uncertainty to some severe conditions like medical inspection [8]. Plus, the computational time of some original algorithms has quadratic relationship with the number of clusters. Overmuch clusters can make the time cost of some clustering algorithms be unacceptable and some algorithms are required to be improved to be more computationally economical [9, 10].

This paper will introduce more details on clustering methods and how they work. The first section of this paper will introduce the Fuzzy C-Means in details. Then the second section will inspect the contributions of other scholars in the field. In the third section, we will talk about how to solve the problems of FCM by utilizing schemes of meta learning and against learning. The final section of the paper is a summary.

2. Fuzzy C-Means algorithm
Among the many Fuzzy Clustering algorithms, the Fuzzy C-Means (FCM) algorithm is the most widely used and successful. Fuzzy C-Means was defined by Dunn [11] and developed from Hard C-Mean clustering (HCM) by J. C. Bezdek in 1981. Before FCM, the original HCM uses either 1 or 0 to identify each point’s affiliation to a single cluster in a hard way. However, FCM gives elements membership to clusters represented by \( u_{ik} \) (degree of membership of the \( i^{th} \) point to the \( k^{th} \) cluster) varied from 0
to 1, where 1 shows the point is entirely belonged to the cluster and 0 shows the point is entirely excluded from the cluster. In this way, FCM can get a more flexible result that classifies the data into clusters in a soft way. The FCM objective function is

$$J(U, V) = \sum_{i=1}^{N} \sum_{k=1}^{C} (u_{ik})^m \| y_i - v_k \|^2,$$

where the restriction function is $\sum_{i=1}^{N} u_{ik} = 1$.

In the equation, let $Y = \{ y_1, y_2, y_3, ..., y_N \}$ be a sample of $N$ observations in $R^n$ (n-dimensional Euclidean space); $c$ is the number of clusters in $Y$; $m$ is a weighting exponent and $1 \leq m \leq \infty$; $U$ is the Fuzzy C-Partition of $Y$; $v$ refers to the vectors of centers so $v_i$ represents the center of cluster $i$. Through lagrangian multiplier method, we can further derive the iterative system of equations

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \frac{||x_i - c_k||^2}{||x_i - c_k||^2}}^m,$$

$$c_j = \frac{\sum_{i=1}^{N} (u_{ik})^m x_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

In the calculation, the matrix $U$ and the number of clusters need to be artificially initialized. Then FCM needs to calculate the center of clusters by using the formula: $c_j = \frac{\sum_{i=1}^{N} (u_{ik})^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$. Continually, FCM will update the weight (membership): if the data point is closer to the center of mass $c$, the membership degree will be higher, and vice versa. The FCM algorithm will stop until the objective function value or the change of value of the function to the previous one is smaller than the threshold value [12].

3. Common problems in FCM and relative researches

Since in every step of the iteration the algorithm operates on the whole dataset, FCM sometimes cannot meet the task of search for structure of data. Furthermore, because FCM always takes all the points in given data to calculate the new cluster, the result is very sensitive to the outliers in the data.

Although Fuzzy C-Means algorithm can be efficient when dealing with small scale of data, the number of iterations will have a significant increase when data and dimension in the calculation increase. Thus, FCM has a long computational time in dealing with large scale of data.

As the membership matrix and number of clusters must be artificially initialized at first, Fuzzy C-Means is largely influenced by the initial guess of the algorithm. Also, because the Fuzzy clustering targets are nonconvex, iteration in Fuzzy C-Means algorithm with the wrong initial condition would run the risk of getting into local extremum and cannot get the optimal solution.

3.1. Noise

After FCM was derived and employed in many different fields, the problem of noise caused troubles in many applications and is noteworthy. The noise in the clustering refers to the outliers of the data set in the clustering which have significant numerical difference from most of the other data or have a large spatial distance with areas with high density of data points. Scholars have been working on solving the noise problem and the followings are some of different solutions to noise problem. For example, FCM is popularly employed in image segmentation tasks. Although FCM has shown its superiority in addressing some images with normal or predictable pattern, it is found inapposite to deal with images with more complex and inhomogeneous features such as images of human brains with tumor. Shan Shen proposed a Magnetic Resonance Images (MRI) fuzzy segmentation method which utilizes the neighborhood attraction to improve the segmentation performance [13]. Besides, there are many algorithms designed to improve on FCM’s performance on noise. The noise cluster algorithm which assigns the noisy data points into the noisy class was proposed earlier in 1991 by Dave [14]. T.N. Pappas introduced the spatial constraint by using Gibbs random field model in 1992 [15]. Later demonstration showed that the algorithm is competent in preserving the significant features as well as smoothing
outliers. With the idea of spatial constraint, Weiling Cai, Songcan Chen and Daoqiang Zhao proposed another algorithm, fast generalized Fuzzy C-Means (FGFCM) algorithms, which combines local spatial and gray information together to ensure both noise-immunity and detail-preserving for image [16]; Y.A. Tolias and S.M. Panas proposed the usage of a Fuzzy Rule-Based System to eliminate noise effects by imposing spatial constraints [17]. A new Fuzzy Relational data Clustering (FRC) algorithm was developed from Relational Fuzzy C-Means algorithm by Sumit Sen and R.N. Dave to reduce noise [18]. However, the defect of this algorithm is that there is no restriction on the relational data. These data which are used in objective function should be derived from Euclidean measure of distance. Otherwise, the computational time will increase due to inefficiency of non-Euclidean measurement. The Fuzzy Segmentation algorithm for Two-Dimensional (2-D AFCM) and Three-Dimensional (3-D AFCM) Magnetic Resonance Images were proposed in respectively 1998 and 1999 by D. L. Pham and J. L. Prince to reduce noise in medical images. In 2-D AFCM, the multiplier field is introduced to adjust the centroids to the images by amending the objective function of Fuzzy C-Means algorithm. 3-D AFCM is an extension of 2-D AFCM and it adjusts the image intensities by modeling the intensity inhomogeneities as a gain field [19, 20]. PCM, which was proposed by R. Krishnapuram and J. M. Keller in 1993 was a primitive one in improving FCM with noise problem and Fuzzy Possibilistic C-Means algorithm (FPCM) was improved by taking ideas from both FCM and PCM in 1997 and the Possibilistic Fuzzy C-Means algorithm (PFCM) was later developed in 2005 [21-23].

3.2. Computational time

Computational time usually refers to the time between input of the data and output of the result. The computational time is proportional to the square of the number of cluster and this makes FCM have a long computational time when clustering the large dataset. In addition, under the same condition, there is trade-offs between the relatively short computational time and precision of the clustering result instead of achieving both two. Earlier in 1986, Robert L. Cannon, Jitendra V. Dave and James C. Bezdekk recognized the efficiency problem linked with Fuzzy C-Means Algorithm and they proposed an Approximate Fuzzy C-Means (AFCM) in which the core is to replace the necessary “exact” variates in the equation with integer-valued or real-valued estimates [24]. Besides, a novel approach of imposing spatial constraints in image segmentation problems shows comparable effectiveness as well as less computational time [17]. Mohamed S. Kamel and Shokri Z. Selim proposed their two improved algorithms which show superiority in faster processing speed. The original FCM alternates from computing the memberships values and the centers of clusters in a hill-climbing way while the new algorithms start computing one as soon as the other is found rather than delaying until all of one part is done [25]. Another pSFCM was proposed by Ming Chuan Hung and Don Lin Yang. This algorithm clusters on a simplified original dataset and by improving the initial value, it shortens the convergence time as well as keeps the quality of the clustering result [26]. Extensible Fast Fuzzy C-Means (eFFCM) aims to reduce computational complexity of clustering on especially large images simplifies the image data by selecting “featured vectors” in a small range of gray values as a representation of each homogenous segment with several hypotheses on selected vectors and corresponding part of the image. [27]. Another clustering algorithm, bit reduction by Fuzzy C-Means (brFCM), is conceived to shrink the precision of data input to speed up the computational time while the precision of the output is not affected by the data reduction [28]. In 2002, J.K. Kolen and T. Hutcheson presented an implementation of FCM by reducing the data structure of cluster centers with large data size. This implementation significantly reduces the clustering time as the runtime of the new method is reduced to be linear with the number of clusters [10]. In 2007, the Fast Generalized Fuzzy C-Means (FGFCM) was presented on the basis of FCM_S, and the new algorithm improves disadvantages of FCM_S. Originally, the time of FCM_S dependent on size of image while the FGFCM introduces the incorporation of both local spatial and gray information so that the segmenting time is rather depend on number of gray-levels $q$ than the size $N$ which is usually much larger than $q$. The computational complexity given in the literature is quantified as $O(qcI_2)$ for FGFCM and the old one is $O(NcI_1)$, where $c$ is the number of the clusters, $I_1$ and $I_2$ are the numbers of iterations [29].
3.3. Initial bias
Initialization is an important part in Fuzzy C-Means algorithm and it involves initial guess of cluster centers and number of clusters. Without any guide, people would usually set the initial conditions by empiricism. However, an improper number of clusters would lead to higher time cost and worse fuzzy clustering result and an improper initial cluster centers would lead to greater cost of time and danger to sink into local minimum.

In 1993, the Fuzzy Generalized K-Nearest Neighbor Rule is proposed on the basis of Nearest Neighbor Rule (NNR). This new rule assigns each point to a sample point instead of a cluster center so that the Fuzzy Generalized K-Nearest Neighbor algorithm avoids the initial guess of cluster number and cluster centers [30]. A new algorithm called Fuzzy Local Information C-Means (FLICM) was developed from FCM by Stelios Krinidis an Vassilios Chatzis and the main characteristic of the new algorithm is utilizing spatial similarity and grey information of data points. The FLICM casts aside the restraint of the initial empirically setting parameters [31]. The Global Fuzzy C-means Clustering algorithm (GFCM) was proposed to solve the problem of sensitivity of FCM to initial guess and it employs a deterministic approach to find data points which minimize the objective function of the FCM as set of initial seeds. The GFCM is independent of initial guess and it has been presented to have better clustering results than FCM [32]. In 1989, the idea of Unsupervised Optimal Fuzzy Clustering which was derived from an incorporation of the Fuzzy K-Means algorithm and Fuzzy Maximum likelihood estimation is able to carry out fuzzy clustering without initialization of number of clusters [33]. In 2009, a group of Chinese scholars proposed an improved FCM algorithm including a new and efficient select rule of initial seeds [34].

Yang HongLei and Peng JunHuan proposed an improved FCM which gets the initial conditions by estimation rather than guess. This is achieved by a modified FCM algorithm with the density estimation of the first principle components. [35]. Genetic Fuzzy C-Means algorithms utilizes a set of three genetic operators which would respectively initially choose random cluster centers, introduce mutation on the raw genetic material and create initial genomes for algorithm to find near optimal clusters. [36].

A method to estimate the number of clusters is presented by A.K. Jain using bootstrap technique [37]. This technique finds the most proper member of clusters by checking for the most stable Value of K which is the number of clusters over the bootstrap samples. A modified FCM called the psFCM algorithm can find and initialize the accurate cluster center by using a simplified set of the original complete dataset [38]. An algorithm which was derived specifically for segmentation of color images uses a novel Hierarchical Approach to determine the precise initial conditions [39]. The cluster number as well as the initial cluster centers are obtained by an alternative use of splitting and merging technique of the color images in the algorithm.

4. Discussion and Conclusion
At the beginning, the article introduces clustering technology including the most popular methods used nowadays and their main usage. Three common problems of clustering algorithms are raised.

The second part of the article describes the derivation of FCM and the operation of FCM with iterative formulas. In terms of FCM, the causes of three problems of FCM are described in detail. Then the articles respectively enumerate the researches on solving these three problems. Twelve articles aimed to solve noise problem are listed and introduced respectively. Basically, noise clustering is a typical solution to the noise problem while there are also other ideas of utilizing spatial constraint of the data, utilizing gray information and so on. Eight articles aimed to solve computational complexity are listed and introduced respectively. In conclusion, there are two core ideas to shorten the operation time: to simplify the raw data input or to modify the iterative steps of FCM. Ten articles aimed to solve problems from initialization are listed and introduced respectively. Methods of improving initialization steps are mainly based on the three following ideas: to avoid the initialized step by replacing FCM with an alternative clustering algorithm or by some modification on the initialized step, to mitigate the influence initialization has on the final clustering result or to estimate the number of clusters with computer rather than empiricism.
Based on the thinking and summary of the above methods, two possible ways to solve the problems are proposed as follow. To combine adversarial learning with FCM algorithm can be an effective way to solve the noise problem. By adding noise whose number matches the size of original data set, we could obtain a more robust algorithm after the attack of these fake noise data. Meta learning provides a novel and promising direction to deal with clustering problems. One general meta learning scheme on improving the performance of algorithms is to search through and select the reduced space in which subsets of algorithms with similar data-mining performance are arranged in order. Automatically generating update to algorithms used in datamining and selecting befittingly biased distribution of data used, meta-learning method is competent in dealing with clustering problems, such as outlier problem, related to unpredictable and large dataset. For example, applying meta learning to algorithm which aim to speed up the FCM by cutting down size of database and trying to maintain the originality of the data structure can be a potentially effective way to improve computational efficiency. Further, it may be feasible to apply meta learning to different models of adversarial learning to try finding the most suitable adversarial learning model for a specific task; it may be possible to integrate adversarial learning with other algorithms used to solve clustering problems and add them together into the subsets of the reduced space under meta learning condition. The above potential ways to solve problems in FCM as well as clustering algorithm need to be verified by more future testing on real cases.

References
[1] C.H. Li, P.C. Yuen, “Regularized color clustering in medical image database”
[2] Klaas Klasing, Dirk Wollherr, Martin Buss, “A clustering method for efficient segmentation of 3D laser data”
[3] Dan Gill, John C. Tipper, “The Adequacy of Non-Metric Data in Geology: Tests Using a Divisive-Omnithetic Clustering Technique”
[4] A.Tagarakis, V. Liakos, S. Fountas, S. Koundouras, T. A. Gemtos, “Management zones delineation using fuzzy clustering techniques in grapevines”
[5] R. Ajaykumar, Gupta, Arpit, Merchant, Prof S N,“Automated Lane Detection by K-means Clustering: A Machine Learning Approach”
[6] Stefano Iellamo, Janne J. Lehtomaki, Zaheer Khan, “Placement of 5G Drone Base Stations by Data Field Clustering”
[7] J.G Campbell, C Fraley, F Murtagh, A.E Raftery, “Linear flaw detection in woven textiles using model-based clustering”
[8] Ping Wang, HongLei Wang, “A Modified FCM Algorithm for MRI Brain Image Segmentation”
[9] M. A. Deshmukh, Ram Meghe, R. A. Gulhane, “Importance of Clustering in Data Mining,” (2016)
[10] J.F. Kolen, T. Hutcheson, “Reducing the time complexity of the fuzzy c-means algorithm”
[11] J.C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters”
[12] James C. Bezdek, Robert Ehrlich, William Full, “FCM: the Fuzzy C-Means clustering algorithm”
[13] Shan Shen, W. Sandham, M. Granat, A. Sterr, “MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization”
[14] Rajesh N. Dave, “Characterization and detection of noise in clustering”
[15] T.N. Pappas, “An adaptive clustering algorithm for image segmentation”
[16] Weiling Cai, Songcan Chen, Daoqiang Zhang, “Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation”
[17] Y.A. Tolias, S.M. Panas, “On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system”
[18] Sumit Sen, R.N. Dave, “Clustering of relational data containing noise and outliers”
[19] Dzung L. Pham, Jerry L. Prince, “An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities”
[20] Dzung L. Pham, Jerry L. Prince, “An Adaptive Fuzzy Segmentation Algorithm for Three-Dimensional Magnetic Resonance Images”
[21] R. Krishnapuram, J.M. Keller, “A possibilistic approach to clustering”
[22] N.R. Pal, K. Pal, J.C. Bezdek, “A mixed c-means clustering model”
[23] M.R. Pal, K. Pal, J.M. Keller, J.C. Bezdek, “A possibilistic fuzzy c-means clustering algorithm”
[24] Robert L. Cannon, Jitendra V. Dave, James C. Bezdek, “Efficient Implementation of the Fuzzy c-Means Clustering Algorithms”
[25] Mohamed S. Kamel, Shokri Z. Selim, “New algorithms for solving the fuzzy clustering problem”
[26] Ming Chuan Hung, Don Lin Yang, “An efficient Fuzzy C-Means clustering algorithm”
[27] N.R. Pal, J.C. Bezdek, “Complexity reduction for "large image" processing”
[28] S. Eschrich, Jingwei Ke, L.O. Hall, D.B. Goldgof, “Fast fuzzy clustering of infrared images”
[29] Weiling Cai, Songcan Chen, Daoqiang Zhang, “Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation”
[30] M.-S. Yang, “A survey of fuzzy clustering”
[31] Stelios Krinidis, Vassilios Chatzis, “A Robust Fuzzy Local Information C-Means Clustering Algorithm”
[32] Weina Wang, Yunjie Zhang, Yi Li, Xiaona Zhang, “The Global Fuzzy C-Means Clustering Algorithm”
[33] I Gath, AB Geva, “Unsupervised optimal fuzzy clustering”
[34] Zhang Huizhe, Wang Jian, “Improved Fuzzy C-Means Clustering Algorithm Based on Selecting Initial Clustering Centers”
[35] Yang HongLei, Peng JunHuan, Xia BaiRu, Zhang DingXuan, “Remote Sensing Classification Using Fuzzy C-means Clustering with Spatial Constraints Based on Markov Random Field”
[36] M.A. Egan, M. Krishnamoorthy, K. Rajan, “Comparative study of a genetic fuzzy c-means algorithm and a validity guided fuzzy c-means algorithm for locating clusters in noisy data”
[37] A.K. Jain, J.V. Moreau, “Bootstrap technique in cluster analysis”
[38] Ming-Chuan Hung, Don-Lin Yang, “An efficient Fuzzy C-Means clustering algorithm”
[39] Khang Siang Tan, Wei Hong Lim, Nor Ashidi Mat Isa, “Novel initialization scheme for Fuzzy C-Means algorithm on color image segmentation”