DAG-searched and Density-based Initial Centroid Location Method for Fuzzy Clustering of Big Biomedical Data

Chanpaul Jin Wang
UMass. medical school
368 Plantation St.
Worcester, MA, 01605-0002, USA

Hua Fang
UMass. medical school
368 Plantation St.
Worcester, MA, 01605-0002, USA

Honggang Wang
UMass. Dartmouth
285 Old Westport Road
North Dartmouth, MA, 02747, USA

{chanpaul.wang,HuaJulia.Fang}@umassmed.edu
Hwang1@umassd.edu

ABSTRACT

Randomly allocating initial centroids may lead to undesired steady states for fuzzy c-means (FCM) clustering. This paper proposes an alternative method to automatically search initial centroid location based on data density. Specifically, this method auto-searches points located in high-density domains as centroids using directed acycline graph (DAG) based algorithm, and then iteratively finding the optimal patterns. Compared with random initialization method, our method seems to have the potential to improve FCM accuracy for larger data size with seconds' tradeoff in computational time using published datasets.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms, Design, Experimentation, Performance, Verification.

Keywords

Initial Centroids, Fuzzy Clustering, Density, Directed Acycline Graph

1. INTRODUCTION

Fuzzy C means (FCM) is a widely studied and used clustering method in biomedical research [1, 2, 3]. Randomly selecting initial centroids is a debating feature of FCM and may lead to undesired steady states [4, 5, 6], because the randomly selected centroids may locate far from the final converged centroids at the beginning, therefore computationally it may not be efficient in situations such as when data size is large [7, 8]. Besides, the distance-based FCM may not be robust to nonspherical data [9]. Easter et al. [10] proposed a density based clustering algorithm for large spatial database with noises. Recently, Rodriguez et al. [11] proposed a fast clustering method by searching and finding data density peaks. Built upon these research findings, our project integrates the data-density idea into FCM to enhance its initial centroids location, thus improve its global convergence and robustness to data types and shapes while keeping well-documented and important FCM features [12, 13]. Specifically, we designed a directed acycline graph (DAG) based algorithm to automatically find high-density points as initial centroids, and then iteratively carries out the FCM to find optimal clusters. The rest of this paper is organized as follows: Section II presents our algorithm design; Section III evaluates the performance of our algorithm using published biomedical and simulated datasets; and Section IV concludes our work.

2. DAG-SEARCHING AND DENSITY-BASED FCM (DDF) ALGORITHM DESIGN

As demonstrated by Rodriguez et al. [11], centroids are surrounded by neighbors with lower local density and at a relatively large distance from any points with a higher local density. Based on this idea, this paper designs the DDF as follows: The core of DDF is to determine the optimal initial centroids. Assuming the dataset $X = \{X_1, X_2, ..., X_N\}$, for any data point x_i,Rodriguez et al. [11] gave the definition of its local density ρ_i and its distance δ_i from points of the higher density as Eq. (1) and Eq. (2), respectively.

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

where $\chi(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise \end{cases}$ and d_{ij} denotes the distance between the data point x_i and x_j, and d_c is a cutoff distance.

$$\delta_i = min_{j:\rho_j > \rho_i}(d_{ij}) \tag{2}$$

Given the above, we define directed neighbors as below.

Definition 1. The directed neighbor of point p is point q as follows:

$$q = argmin_{j:\rho_j > \rho_i}(d_{ij})$$

Thus, the point with the highest density has no directed neighbors. For each of remaining data points, it has only one directed neighbor. Moreover, the directed neighbor is not reversible. In other words, if the directed neighbor of Point p

BICT 2014, December 01-03, Boston, United States

DOI 10.4108/icst.bict.2014.257932

is Point q, then Point p is not the directed neighbor of Point q. We can construct a directed acycline graph, specifically, a directed-neighbor tree, to auto-search centroids rather than manually in Rodriguez et al. [11], (see Fig. 1). The tree roots at the point with the highest density, the node denotes the data point, and the weight of the edge denotes the distance between nodes.

Moreover, Easter et al. [10] defined density-reachable with several strong conditions. Yet, given the only cutoff distance d_c, we further define ρ-density-reachable as follows.

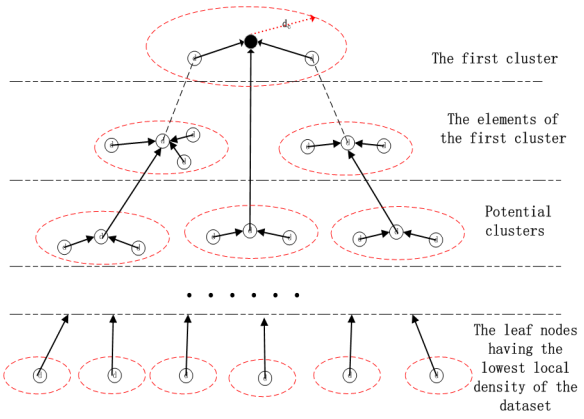


Figure 1: Illustration of the directed acycline graph: directed neighbor tree

Definition 2. (directly ρ-density-reachable) Point p is directly $\rho(\rho > 0)$ density-reachable from Point q, if $d_{pq} < d_c$, $\rho_p > \rho$, $\rho_q > \rho$.

Definition 3. (ρ-density-reachable) Point p is directly $\rho(\rho > 0)$ density-reachable from Point q, if there is a chain of points $p_1, ..., p_n$, $p_1 = q$, $p_2 = p$ such that p_i is directly $\rho(\rho > 0)$ density-reachable from p_{i+1},$\rho < min(\rho_{p_1}, ..., \rho_{p_n})$.

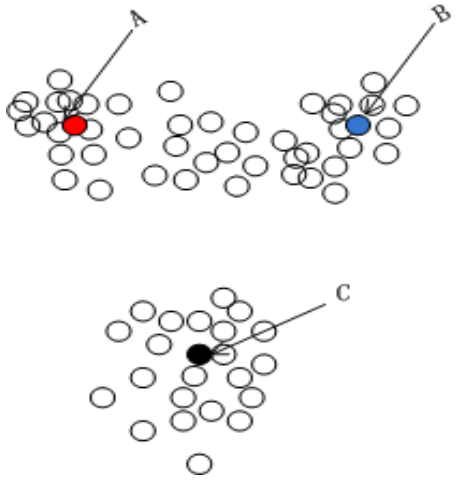


Figure 2: Example of "faked" sub-trees

Based on the above definitions, we designed a prune-converge algorithm to automatically determine the optimal initial centroids. First, at the pruning stage, we prune the directed-neighbor tree based on the weight of the edge, namely pruning all the edges with the weight larger than the cutoff distance . Then, we obtain partial sub-trees, which represent all local high-density domains and are $\rho(\rho > 0)$ density-reachable. As aforementioned, each sub-tree roots at the point with the highest density, and these points are candidates of our aimed centroids. However, special cases can occur as shown in Fig. 2. Point A has the highest density, and also is the directed neighbor of Point B and C. According to our pruning rule, Point A, B and C are the sub-roots. In fact, Point A and B should be in the same sub-tree as illustrated in Fig 2. This special case demonstrates that there could be "faked" sub-trees, and we need to merge them to nearby sub-trees. At the converging stage, we merge these "faked" sub-trees, and identify the true sub-trees. Thus, the purpose of the converging stage is two-folds: 1) finding the points directly $\rho(\rho > 0)$ density-reachable from the points of other sub-trees, and merge the sub-trees having large $\rho(\rho > 0)$ density-reachable, 2) finding the sub-trees having low local density, and excluding them from the sub-tree set.

Assuming the sub-trees as $T = \left\{\bar{T}_1, \bar{T}_2, ..., \bar{T}_S\right\}$, $\bar{T}_i \subset X$, and the points directly ρ density-reachable from other sub-trees as $\bar{T}^{'} = \{x_k\} \subset \bar{T}_i$, $\bar{T}^{'} = \bar{T}_1^{'} \cup \bar{T}_2^{'} \cup ...\bar{T}_S^{'}$, the number of points in $\bar{T}^{'}$ as D, we define the converging density ρ_{conv} as follows:

$$\rho_{conv} = \frac{1}{D} \sum_{k \in \bar{T}^{'}} \rho_k$$

Then, for any two sub-trees, if there exists ρ_{conv} density-reachable points, we merge them into a new sub-tree. After merging the sub-trees, we further exclude sub-trees with low local density. Assuming the roots of the sub-trees as $TR = \left\{tr_1, tr_2, ..., tr_S^{'}\right\}$, $tr_i \in X$, where $S^{'}$ denotes the number of sub-trees after merging, we define the average local density of sub-trees as follows:

$$\rho_{av} = \frac{1}{S^{'}} \sum_{k \in TR} \rho_k$$

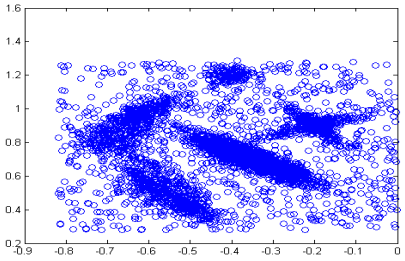
Then, for each sub-tree, if the local density of its root $\rho_{root} \geq \rho_{av}$, we identify it as an initial fuzzy centroid. The new sub-root is the point with the highest local density of the new sub-tree. These converged sub-trees are important for the fuzzy clustering. First, these sub-roots have large local density, close to the final converged fuzzy centroids. Starting from these initial sub-roots may be computationally efficient. Given a fuzzy clustering number c, we can select the c highest local-density roots of these sub-trees as the initial centroids, and their corresponding sub-trees as the c initial clusters. We then merge the remaining sub-trees into the initial c clusters according to the directed neighbors of their corresponding roots. Then, we preliminarily classify the dataset into c clusters. Assuming the initial clusters as $C = \{C_1, c_2, ..., C_c\}$, the number of data points in each cluster as $n_1, n_2, ..., n_c$,$n_1 + n_2 + ... + n_c = N$, we then iteratively implemented the FCM until reaching the steady states.

3. EVALUATION

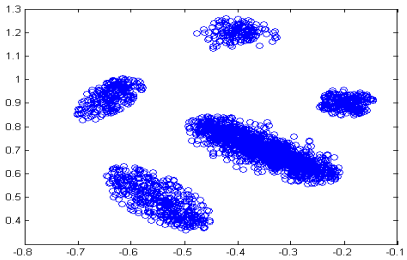
We evaluated the DDF based on the published datasets from Ref. [11], and two UCI datasets: Breast-Cancer and Iris datasets. The dataset from Ref. [11] is unlabeled, and includes noisy data as shown in Fig. 3(a). We applied the method in Ref. [11] to label these data, and removed the noisy data as shown in Fig. 3(b). Three sets of data (see Table 1) were randomly sampled from the labeled data (N_{total} = 2535) according to their cluster density ($C1_n$ = 1456,$C2_n$ = 246,$C3_n$ = 246,$C4_n$ = 431,$C5_n$ = 156). Specifically, we proportionally sampled from each cluster of the original data according to the ratios of 0.574, 0.097, 0.097, 0.170 and 0.062.

Table 1: Description of Published Datasets

			Ref. [11]		
	Breast-Cancer	iris	a	b	c
#Num. clusters	2	3	5	5	5
Size	699	150	1500	2000	2500



(a) The original unlabeled dataset



(b) labeled dataset

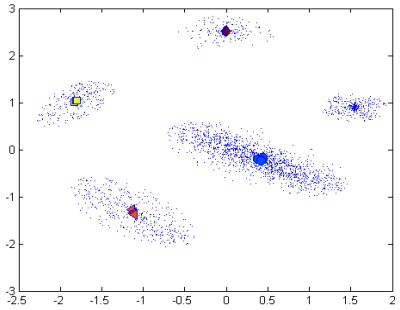
Figure 3: The origination data distribution from Ref. [11]

First, we evaluate the convergence of DDF and FCM. As shown in Tab. 2, DDF needs less iterations than FCM to converge. In particular, DDF seems to converge much more quickly than FCM in clustering big dataset. As shown in Fig. 4, with the DDF, the cluster centeroids of labeled data c from Ref. [11] are always distributed near the highest-density area of each cluster, yet the cluster centroids of FCM change seriously. Especially, DDF needs 9 iteration times much less than 44 iteration times of FCM in clustering the labeled data c from Ref. [11], which has 2500 data.

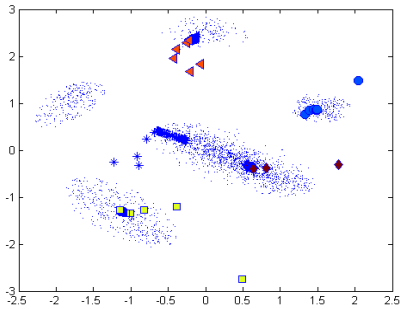
Moreover, we further demonstrate the clustering accuracy and computational time for classic FCM and our DDF as Table 3 and Table 4, respectively. The performance evaluation was conducted on Windows XP platform and regular PC desktop with Intel(R) 2.4GHZ CPU and 3G memory.

Table 2: Iterations of DDF and FCM (m=2)

			Ref. [11]		
	Breast-Cancer	iris	a	b	c
FCM	8	9	20	20	44
DDF	8	7	8	7	9



(a) DDF



(b) FCM

Figure 4: The centroid variations of DDF and FCM for Data c from Ref. [11]

Table 3: Accuracy of DDF and FCM on different dataset

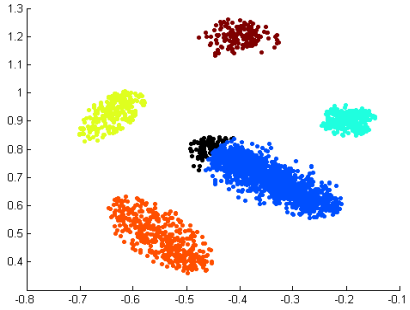
			Ref. [11]		
	Breast-Cancer	iris	a	b	c
FCM	94.85%	84%	95.67%	95.85%	67.53%
DDF	95.57%	84%	96.14%	96.40%	96.40%

As shown in Table 3, the FCM and our DDF were comparable for small to moderately large datasets (Iris, Breast Cancer, and two sets of Ref. [11]). Interestingly, DDF has significantly higher accuracy than FCM for the largest dataset of 2500 from Ref. [11], which was also visually displayed in Figure 5. Table 4 indicates that there is a trade off in terms of computational cost for a higher accuracy. However, the 8 second difference may be trivial.

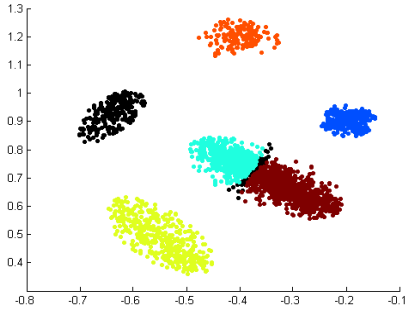
4. CONCLUSIONS

Table 4: Accuracy of DDF and FCM on different dataset

			Ref. [11]		
	Breast-Cancer	iris	a	b	c
FCM	20.0601s	0.0303s	1.1230s	1.2329s	1.1750s
DDF	1.8045s	0.3641s	7.2601s	7.0749s	9.3036s



(a) DDF



(b) FCM

Figure 5: DDF and FCM clustering results on Data c from Ref. [11]

To prevent potential issues of randomly allocating initial centroids for fuzzy clustering, this study provided an automated initialization method based on data density using directed acycline graph based algorithm (DDF). The performance evaluation indicates DDF has the potential to achieve higher accuracy and faster convergence in terms of the number of iterations when the data size is relatively large. Although there is a tradeoff in overall computational cost, the seconds' difference may not be an issue in more high performance computing environment. In the future, the robustness of DDF need to be warranted on more large and longitudinal biomedical datasets.

5. ACKNOWLEDGMENTS

This research was supported by NIH grant R01 DA033323-01A1, 1UL1RR031982-01 Pilot Project to Dr. Fang.

6. REFERENCES

[1] Fang H. and Johnson C. et al. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. *Neurotoxicol Teratol*, 33(1):155–165, Jan-Feb 2011.

[2] Fang H. and Dukic V. et al. Detecting graded exposure effects: a report on an east boston pregnancy cohort. *Nicotine Tob*, 14(9):1115–1120, 2012.

[3] M.C. Clark, L.O. Hall, and D. B. Goldgof et al. MRI segmentation using fuzzy clustering techniques. *IEEE Engineering in Medicine and Biology Magazine*, 13(5):730–742, August 2002.

[4] Kaiqi Zou and Zhiping Wang et al. An new initialization method for fuzzy c-means algorithm. *Fuzzy Optim Decis Making*, 7(4):409–416, October 2008.

[5] F. Hoppner and F Klawonn. A contribution to convergence theory o fuzzy c-means and derivatives. *IEEE transactions on fuzzy systems*, 11(5):682–694, October 2003.

[6] L. Rutkowski and J. Siekmann et al.(Eds). *Artificial intelligence and soft computing-ICAISC 2004*. Springer-Verlag, Berlin Heidelberg, 2004.

[7] J.C. Bezdek and Richard J. Hathaway et al. Convergence theory for fuzzy c -means: Counterexamples and repairs. *J. Cybernet.*, 3:58–72, 1974.

[8] Richard J Hathaway. Local convergence of the fuzzy c-means algorithms. *Pattern Recognition*, 19(6):477–480, May 1986.

[9] Zhong dong Wu and Wei xin Xie. Fuzzy c-means clustering algorithm based on kernel method. *Computational International Conference on Intelligence and Multimedia Applications*, pages 49–54, September 2003.

[10] Martin Ester and Hans-Peter Kriegel et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd International Conference on KDD*, pages 226–231, 1996.

[11] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):226–231, 2014.

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

[13] M.S. Yang. Convergence properties of the generalized fuzzy c-means clustering algorithms. *Computers and Mathematics with Applications*, 25(12):3–11, June 1993.