HHS Public Access

Author manuscript

Int Conf Comput Netw Commun. Author manuscript; available in PMC 2019 March 21.

Published in final edited form as:

Int Conf Comput Netw Commun. 2018 March ; 2018: 912–916. doi:10.1109/ICCNC.2018.8390419.

eFCM: An Enhanced Fuzzy C-Means Algorithm for Longitudinal Intervention Data

Venkata Sukumar Gurugubelli[1,3], **Zhouzhou Li**[2,3], **Honggang Wang**[2], and **Hua Fang**[1,3]

[1]Department of Computer and Information Science, University of Massachusetts – Dartmouth, Dartmouth, MA, 02747

[2]Department of Electrical and Computer Engineering, University of Massachusetts – Dartmouth, Dartmouth, MA, 02747

[3]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA

Abstract

Clustering methods become increasingly important in analyzing heterogeneity of treatment effects, especially in longitudinal behavioral intervention studies. Methods such as K-means and Fuzzy C-means (FCM) have been widely endorsed to identify distinct groups of different types of data. Build upon our MIFuzzy [1], our goal is to concurrently handle multiple methodological issues in studying high dimensional longitudinal intervention data with missing values. Particularly, this paper focuses on the initialization issue of FCM and proposes a new initialization method to overcome the local optimal problem and decrease the convergence time in handling high-dimensional data with missing values for overlapping clusters. Based on the idea of K-means++ [9], we proposed an enhanced Fuzzy C-means clustering (eFCM) and incorporated it into our MIFuzzy. This method was evaluated using real longitudinal intervention data, classic and generic datasets. Compared to conventional FCM, our findings indicate eFCM can improve computational efficiency and avoid the local optimization.

I. Introduction

Longitudinal behavioral intervention data are complex, typically high-dimensional and highly heterogeneous with missing values. For such data, we developed multiple imputation based fuzzy clustering (MIFuzzy) and its evaluation methods [1], [5], [6], [11], [12], [14], [16], [23].

MIFuzzy is a soft clustering method. Different from hard clustering techniques such as K-means, K-means++ and spectral clustering [2], [9], [3], which decides if a point belongs to a cluster or not, soft clustering techniques such as FCM determine the degree of a data point belonging to different clusters [4]. In medical and especially longitudinal studies, overlapping clusters are common; therefore, soft clustering has gained increasing popularity in this field. Our previous studies indicate, MIFuzzy, which integrates FCM multiple imputation models and validation methods seems to consistently outperform other conventional methods such as K-means, hierarchical, probability-based and SOM-based neural networks for incomplete longitudinal intervention data [11]. However, the literature

indicates that initialization of a clustering algorithm may affect its accuracy and convergence time, especially in big data. K-means++ [9] is a method that uses special seeding of initial centroids as an enhancement to guarantee a globally optimal solution and to improve convergence speed in standard K-means. Motivated by this idea, we proposed eFCM for original longitudinal intervention data to overcome the local optimal solution and improve convergence speed for FCM.

Our experiments showed that in terms of convergence eFCM can reduce the running time up to 30 seconds, compared to the standard FCM. Our eFCM also can overcome local optimization for datasets with overlapping clusters. When incorporating eFCM in MIFuzzy, eFCM was able to identify the optimal number of clusters in high dimensional longitudinal intervention data in a shorter time while maintaining the clustering accuracy. Besides, eFCM has performed better on synthetic datasets where different degrees of overlapping exist.

The following sections are organized as follows: Section II discusses the motivation and drawbacks of existing methods. Section III, proposes our new approach, eFCM to solve the issues addressed in Section II. Section IV evaluates eFCM on longitudinal intervention data, and synthetic datasets. The last section concludes our findings.

II. Literature Review

It is possible that K-means and FCM produce local optimal solution because they both use random initialization of cluster centroids. Due to the heterogeneous nature, missingness, and high-dimensionality of longitudinal data, K-means is not the best option for identifying overlapping clusters. Although FCM can process overlapping clusters, FCM cannot deal with missing data and is prone to local optimization, due to the random centroid initialization.[5], [12]

A variant of K-means, called K-means++, overcomes the local maximal solution in the standard K-means by using the special seeding of the initial clusters to achieve the global optimization[9]. The idea behind K-means++ is to place the initial cluster centroids as further as apart from each other so that the algorithm will not only be able to converge faster but also, guarantees to avoid local maximal solution. However, these K-means based algorithms exist in longitudinal intervention data.

As mentioned earlier, missing data are common in longitudinal trial data. Our MIFuzzy handles missing values under three data missing mechanisms: completely at random, missing at random and missing not at random[11], [12]. (1) select the intervention attributes from data, which will be used by the algorithm. (2) perform MIfuzzy clustering on each of the imputed datasets. (3) validate optimal clusters by using multiple validation indices such as Xie-beni(XB) [7] and visualization.[10]

It is evident that in MIFuzzy performs the knowledge mining by incorporating the soft clustering techniques such as FCM and the fuzzy logic theory [12]. Like K-means, we chose an initial number of clusters and we randomly assign each point to a cluster by the degree of its cluster membership. The algorithm is terminated if the change of parameters between two iterations is no more than the given sensitivity threshold. The fundamental difference

between K-means and FCM is the addition of membership values and the fuzzifier. When the value of fuzzifier is 1, the resulting clustering would be equivalent to the K-means algorithm. Since FCM is also a heuristic algorithm, it prones to local optimal solution similar to K-means.

In the following sections, we introduce and illustrate our proposed enhanced version of FCM eFCM, which can be incorporated in MIFuzzy to handle longitudinal intervention data. Also, we compare the performance of eFCM with FCM on synthetic and real longitudinal data sets.

III. Methods

Our goal is to overcome the possible local optimal and improve the computational efficiency of standard FCM algorithm, which we have mentioned earlier by adapting the initialization scheme of K-means++. We also incorporate eFCM in MIFuzzy to handle longitudinal intervention data with missing values.

The objective function of eFCM can be formulated as below:

$$E(C) = \sum_{i=1}^{N} \sum_{j=1}^{C} (u_{ij})^m d_{ij}^2 \quad (1)$$

$$where, u_{ij} = \frac{1}{\sum_{k=1}^{C} \frac{d_{ij}}{d_{ik}}^{\frac{2}{m-1}}} \quad (2)$$

$$V_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (3)$$

Where, C is the number of clusters, d_{ij} is the distance from i^{th} point to j^{th} centroid, u_{ij} is membership value of i^{th} point to the j^{th} cluster and m is the fuzzifier.

The steps involved in eFCM algorithm can be described as below:

1. Initialize membership matrix $U = [u_{ij}]$ with initial value $U^{(0)}$
2. InitializeCentroids (Algorithm 2):
   a. Sample a point uniformly from the dataset as the first centroid.
   b. Sample the next centroid from the dataset with probability proportional to the squared distance of the point to its nearest centroid.

c. Repeat above step until we have number of centroids equal to the required number of clusters.

3. Calculate the cluster centroid matrix $V^{(r)} = [v_j]$ with $U^{(r)}$, using equations (2), (3)
4. Calculate $U^{(r+1)}$
5. If $||U^{(r+1)} - U^{(r)}|| < \varepsilon$, then converge otherwise, go to step 3. ($0 < \varepsilon < 1$ is predefined threshold)

Unlike the standard FCM algorithm, our proposed initialization approach does not choose all the initial centroids at random Instead, one of the centroids is chosen uniformly at random from the sample space. Remaining centroids are chosen one at a time, based on assigned probabilities with respect to the squared distance of the point to its nearest centroid. This process will ensure that all the centroids will stay farther apart from each other.

Our eFCM was further incorporated into MIFuzzy to extend its capability to handle longitudinal intervention data. Algorithm 1 shows the description of eFCM for clustering data with missing values. Same as MIFuzzy, given the terminating cluster using the rule of thumb, the square root of (N/2) [1], [6], [22], eFCM performs multiple imputations on incomplete longitudinal data. Each imputed data set is clustered using eFCM. The optimal number of clusters is selected using average XB [7], [1] across all imputed datasets.

Algorithm 1

eFCM Algorithm

```
procedure eFCM(X, n)                    ▷ Where, X: Incomplete Data, n: Imputation Times
  EM(X)                                 ▷ Expectation Maximization
  EM Posterior
  for i = 1 : n do
    Build Markov Chain
    Between imputation steps
    Interior Imputation Steps
    Impute X to get Y_i
    for cluster = 2 : maxCluster do     ▷ Find Best
Clusters number
  for m = 1 : 10 do
    Initialize U^(0)
    InitializeCentroids(Y_i, ClusterNumber)
    for r = 1 : MaxIteration do
      Calculate V^(r)
      Calculate U^(r)
      if max|U^(r) − U^(r−1)| < ε then
        Break
    Update Xb Matrix and Lx Matrix
  Per Xb matrix, extract best cluster no. and fuzzifier
```

return *Lx, Cluster, Fuzzifier* ▷ Output

End procedure

Algorithm 2

InitializeCentroids

procedure *InitializeCentroids*(Y_i, *ClusterNumber*)

V ← sample a point uniformly from Y_i

while |V|< ClusterNumber **do**

Sample x ∈ Y_i with probability prop. to $D^2(x)$

V ← $V \cup x$

Return Initial Centroids

IV. Evaluation

Our eFCM is evaluated on a longitudinal intervention dataset, called TDTA [19], also on the IRIS [20], one synthetic dataset and S-sets, which contains four synthetic datasets[13] with varying complexity in terms of spatial distribution.

TDTA data were collected from a longitudinal culturally tailored smoking cessation intervention for 109 Asian American smokers. It contains three culturally adaptive response patterns identified by MIFuzzy [19]. For this intervention, researchers have used three components: Cognitive behavioral therapy, cultural tailoring, and nicotine replacement therapy. The first two components were measured by scores on Perceived Risks and Benefits, Family and Peer Norms, and Self-efficacy scales. Each scale has four repeated measurements, total 20 attributes, of which only Perceived Benefits and Family Norms were as input for our algorithm.[16] To evaluate if eFCM can generate the same patterns as MIFuzzy, the proposed eFCM algorithm was performed on TDTA data set.

We incorporated eFCM into MIFuzzy in MATLAB. Our code uses the Mean of XB index across all multiply imputed datasets to select the optimal number of clusters using the elbow value of XB index. Same as MIFuzzy, eFCM identify three optimal clusters. In terms of computational efficiency, the average runtime of MIFuzzy is 0.027 seconds while eFCM takes 0.013 seconds to converge.

A generic data set, IRIS, was also used to evaluate eFCM in comparison to FCM. IRIS dataset contains four attributes (sepal length, sepal width, petal length and petal width) and 150 observations, which divided into 3 clusters. Due to the overlapping clusters and multiple attributes in IRIS data, hard clustering methods(K-means, K-means++) seem not to identify clear cluster boundaries as well as soft clustering methods, FCM and eFCM. The results show that the misclassification rate of FCM is 0.11333(88.66% accuracy), while eFCM has 0.10667(89.33% accuracy). Additionally, the accuracy of K-means and K-means++ were below 81% on IRIS dataset.

We also simulated a synthetic dataset with 998 observations and two attributes, with induced overlap to test the performance of eFCM regarding its speed and ability to avoid the local optimal. While FCM takes 0.2793 seconds(67 iterations), eFCM takes only 0.1344 seconds (7 iterations) to converge to the global optimal. Figure 2 shows the final centroid location of K-means, K-means++, FCM, eFCM. Same as above, Figure 2 shows the clustering result and the centroids in comparison with K-means++, FCM, eFCM. In the figure, 'o', 'Δ','+' indicate final centroids of K-means, K-means++, FCM, eFCM respectively. From Figure 2, resulting centroid of FCM does not seem accurate as its centroids do not always in the center of clusters. However, eFCM seems to overcome the local optimal with the centroids around the centers in each cluster.

eFCM was tested on manually picked S-sets of Clustering Datasets [13]. S-sets contain four datasets s1, s2, s3, s4 each with, 5000 vectors and 15 clusters. These 4 data sets have similar cluster shapes with different spatial distribution. These datasets will help evaluate how eFCM can overcome the local optimal issue with FCM. From Figures 3–6 even when the clusters were drawn closer, from s1 to s4, eFCM was able to successfully identify the clusters when compared to FCM, K-means and K-means++ in terms of the centroid location. As mentioned earlier, XB index is used to identify the optimal number of clusters, which is widely used in fuzzy clustering [7], [19]. The smallest value of XB indicates the best cluster number. From II, for all the datasets eFCM seems to perform better than FCM. Also, II summarizes the run time for both FCM and eFCM, and XB values for Synthetic data set, s1, s2, s3, s4. Especially in the case of the Synthetic data set, given the number of identified four clusters, it significantly reduces the XB value, ie., the ratio of between and within cluster variance, and reduced the convergence time by half. In the case of the s1 dataset where the clusters were well separated, FCM took a longer time to converge because of its random initialization, while eFCM reduces the runtime by four folds. For s2, s3, s4 datasets, even when there is an increase in overlap, eFCM were able to converge faster than FCM in all the scenarios.

V. Conclusion and Future work

In this paper, we discuss the underlying local optimization issue that the FCM algorithm is potentially facing. The solution to this problem could be similar to that K-means algorithm exploited: carefully selecting the initial centroids. We incorporated the centroid initialization idea of K-means++ algorithm into conventional FCM to solve the local optimization issue and exploit our MIFuzzy to handle complex longitudinal intervention data. Our proposed eFCM identified the same optimal number of clusters as MIFuzzy and seems to achieve better computational efficiency. Compared to FCM, eFCM produce better clustering results regarding the number of optimal clusters and computational speed. Similar to K-means++, eFCM seems to solve the local optimal problem but better than K-means++, while handling overlapping clusters. Although limited by the data features used in this paper, the eFCM shows the potential to be fully incorporated into MIFuzzy and could handle complex longitudinal data with better computational speed, potentially more important in processing big heterogeneous data.

Acknowledgments

This research was supported by NIH grant RO1DA033323-01A1, 1UL1RR031982-01 Pilot Project to Dr. Fang

References

1. Fang H. MIFuzzy Clustering for Incomplete Longitudinal Data in Smart Health. Smart Health (Amst). 1–2:50–65.Jun.2017
2. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 28(2):129–137.1982;
3. von Luxburg U. A Tutorial on Spectral Clustering. Statistics and Computing. 17(4):395–416.2007;
4. Dunn JC. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics. 3(3):32–57.1973;
5. Fang H, Espy KA, Rizzo ML, Stopp C, Wiebe SA, Stroup WW. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. Int J Inf Technol Decis Mak. 8(3):491–513.Sep 1; 2009 [PubMed: 20336179]
6. Fang H, Dukic V, Pickett KE, Wakschlag L, Espy KA. Detecting graded exposure effects: A report on an east boston pregnancy cohort. Nicotine Tob Res. 14(9):1115–20.Sep; 2012 [PubMed: 22266824]
7. Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 13(8):841–847.1991;
8. Bora DJ, Gupta AK. Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. International Journal of Computer Science and Information Technologies (IJCSIT). 5(2):2501–2506.2014;
9. Arthur, D, Vassilvitskii, S. Technical Report. Stanford InfoLab; 2006. K-means++: the advantages of careful seeding. Available: http://ilpubs.stanford.edu:8090/778/
10. Winkler, R, Klawonn, F, Kruse, R. Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organization. Springer; Berlin, Heidelberg: 2012. Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets.
11. Zhang Z, Fang H, Wang H. Multiple Imputation based Clustering Validation (MIV) for Big Longitudinal Trial Data with Missing Values in eHealth. J Med Syst. 40(6):146.Jun.2016 [PubMed: 27126063]
12. Zhang Z, Fang H, Wang H. A New MI-Based Visualization Aided Validation Index for Mining Big Longitudinal Web Trial Data. IEEE Access. 4:2272–2280.2016; [PubMed: 27482473]
13. Fränti P, Virmajoki O. Iterative shrinking method for clustering problems. Pattern Recognition. 39(5):761–775.2006;
14. Fang H, Brooks GP, Rizzo ML, Espy KA, Barcikowski RS. Power of models in longitudinal study: Findings from a Full-crossed Simulation Design. J Exp Educ. 77(3):215–254.Apr 1; 2009 [PubMed: 19946462]
15. Fang H, et al. Using propensity score modeling to minimize the influence of confounding risks related to prenatal tobacco exposure. Nicotine Tob Res. 12(12):1211–9.Dec; 2010 [PubMed: 21030468]
16. Fang H, Zhang Z. An enhanced visualization method to aid behavioral trajectory pattern recognition infrastructure for big longitudinal data. IEEE Transactions on Big Data. :1–1.2017
17. Fang H, Zhang Z, Wang CJ, Daneshmand M, Wang C, Wang H. A survey of big data research. IEEE Netw. 29(5):6–9.Sep-Oct;2015 [PubMed: 26504265]
18. Kim SS, Fang H, Difranza J, Ziedonis DM, Ma GX. Gender differences in the fagerström test for nicotine dependence in Korean Americans. J Smok Cessat. 7(1):1–6.Aug 1; 2012 [PubMed: 22936953]
19. Kim SS, et al. Acculturation, Depression, and Smoking Cessation: a trajectory pattern recognition approach. Tob Induc Dis. 15:33.2017; [PubMed: 28747857]
20. Anderson E. The Species Problem in Iris. Annals of the Missouri Botanical Garden. 23(3):1936.

21. Zhang Z, Fang H. Multiple-vs non-or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. IEEE Int Conf Connect Health Appl Syst Eng Technol. 2016:219–228.Jun.2016 [PubMed: 29034067]

22. Fang H, Johnson C, Stopp C, Espy KA. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicol Teratol. 33(1):155–65.Jan-Feb;2011 [PubMed: 21256430]

23. Fang H, Rizzo ML, Wang H, Espy KA, Wang Z. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm - Pattern recognition. Pattern Recognit. 43(4): 1393–1401.2010; [PubMed: 20300543]

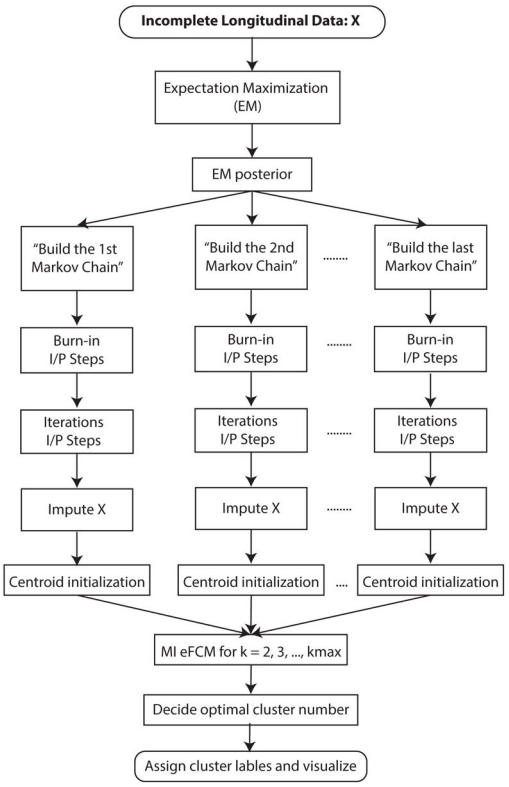

Figure 1.
eFCM built in MIFuzzy

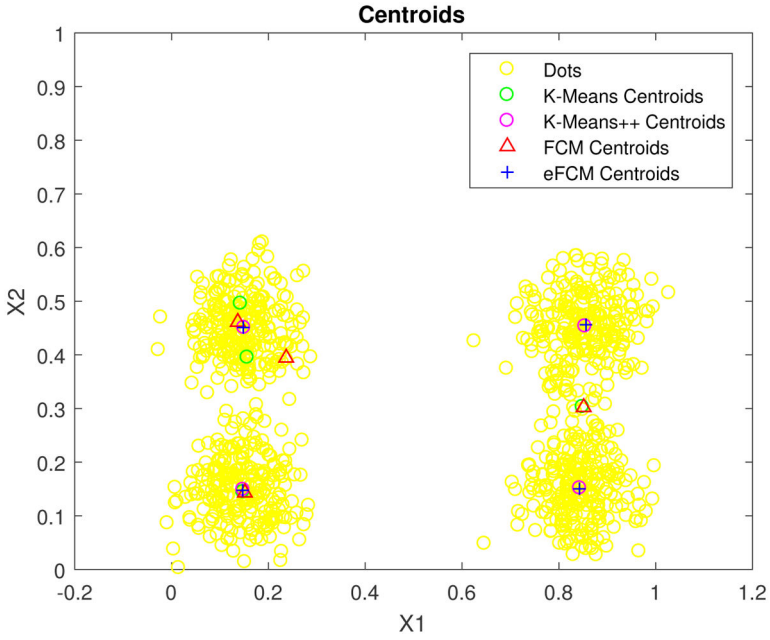

Figure 2.
Centroids for K-means, K-means++, FCM, eFCM for synthetic dataset

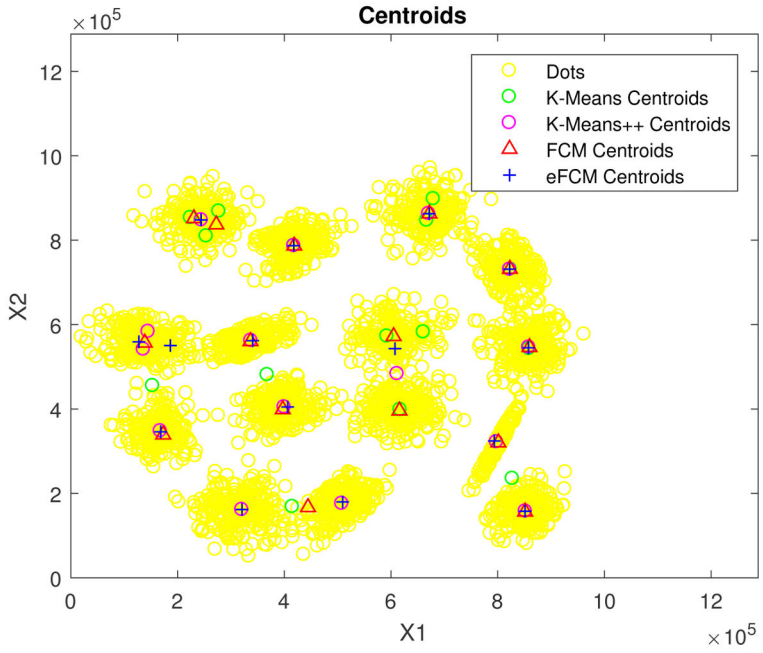

Figure 3.
Centroids for K-means, K-means++, FCM, eFCM for s1 dataset

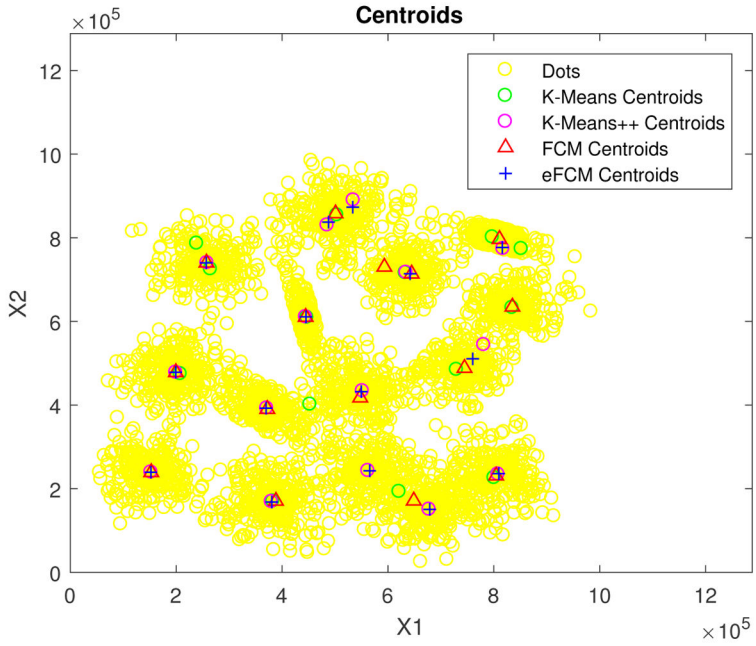

Figure 4.
Centroids for K-means, K-means++, FCM, eFCM for s2 dataset

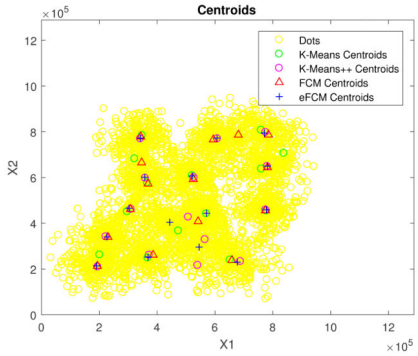

Figure 5.
Centroids for K-means, K-means++, FCM, eFCM for s3 dataset

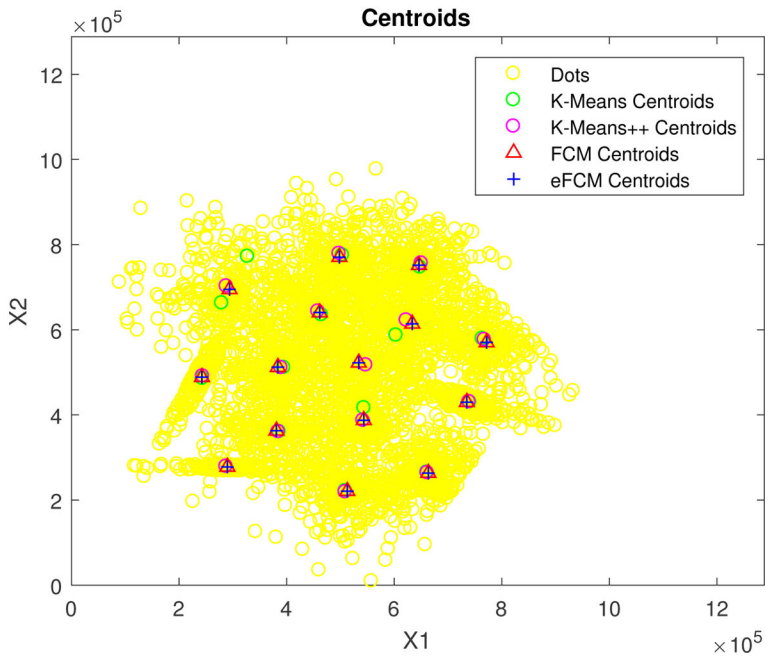

Figure 6.
Centroids for K-means, K-means++, FCM, eFCM for s4 dataset

Table I

Notation Table

C	Number of Clusters
N	Number of Observations
P	Number of dimensions
X	Observations in $R^{N\times P}$ as $\{x_1, x_2, x_3, \ldots, x_N\}$
U	Membership matrix
u_{ij}	Membership value of i^{th} point in j^{th} cluster
d_{ij}	Distance from i^{th} point to j^{th} Centroid
V	Vector of Centroids $v_1, v_2, \ldots, v_C$
i	index of i^{th} point
j, k	Index of j^{th} or k^{th} Centroid
r	Iteration
m	Fuziffier
ε	Iteration termination criterion

Table II

Evaluation Table

Dataset	*FCM**	*eFCM**	FCM(XB)	eFCM(XB)
Synthetic	0.2793	0.1344	6033	76
s1	3.0592	0.6291	82586	19183
s2	1.2689	0.6165	90009	1793
s3	1.3714	0.3067	65861	28246
s4	0.4912	0.4366	24978	24345

*Columns indicates the running time of algorithm in seconds.