Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v11n2p674

An enhanced fuzzy K-means clustering with application to missing data imputation

By Migdady, Al-Talib

Published: 14 October 2018

This work is copyrighted by Università del Salento, and is licensed under a Creative Commons Attribuzione - Non commerciale - Non opere derivate 3.0 Italia License.
For more information see:
http://creativecommons.org/licenses/by-nc-nd/3.0/it/

Electronic Journal of Applied Statistical Analysis
Vol. 11, Issue 02, October 2018, 674-686
DOI: 10.1285/i20705948v11n2p674

An enhanced fuzzy K-means clustering with application to missing data imputation

Hazem Migdady*[a] and Mohammad Al-Talib[b]

[a] *Amman Arab University, Mobile Computing Department, 11953 Amman, Jordan*
[b] *Yarmouk University, Department of Statistics, 21163 Irbid, Jordan*

Published: 14 October 2018

In this paper an adjustment on the Fuzzy K-means (FKM) clustering method was suggested to improve the process of clustering. Also, a novel technique for missing data imputation was proposed and it was implemented twice: (1) using FKM and (2) using the Enhanced Fuzzy K-means (EFKM) clustering. The suggested model for imputing missing data consists of three phases: (1) Input Vectors Partitioning, (2) Enhanced Fuzzy Clustering, and (3) Missing Data Imputation. The implementation and experiments showed a clear improvement in the imputation accuracy in favor of the EFKM according to the value of RMSE.

keywords: Missing Data Imputation, Cluster Analysis, Fuzzy K-means clustering, Data mining, Fuzzy sets, Fuzzy C-means.

1 Introduction

Cluster analysis is the process of grouping data points into clusters, such that similar (or of the same characteristics) data points are grouped together. It is a well known approach in statistical data analysis and data mining, and it is used in many fields such as: missing data imputation, data compression, computer graphics, machine learning, ... etc.
Many algorithms can be applied in defining clusters, but they differ in the criteria of clustering, as for example data points (or vectors) that have small euclidean distances

*Corresponding author: hmigdady@aau.edu.jo

©Università del Salento
ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

between each other.
Subdividing a data set into K pairwise disjoint clusters is known as hard K-partition of the data set. Such partition can be found based on minimizing the squared error between the empirical mean of a cluster and the data points in that cluster (Jain (2010)), this approach is called Hard K-Means Clustering (HKM). However, Bezdek et al. (1984) argue that a disadvantage of such algorithm is: there is no apparent similarity measure between the elements of clusters.
When dealing with fuzzy situations, its possible to employ the concept that suggested by Zadeh (1965) in the clustering process by assigning each element in a data set to a cluster based on a function that measures the similarity between the element and the cluster. This function, which is known as the membership function, ranges between zero and one. As the similarity between the element and a cluster increases, the degree of membership of that element with respect to that cluster approaches one. This is called fuzzy k-partitioning, or Fuzzy k-Means algorithm (FKM) when the estimated means are used to find the degree of membership, this technique was firstly suggested by Dunn (1973) and improved by Bezdek (1981).
Another clustering algorithm based on fuzzy sets is the Gaussian Mixture Model (GMM). This method is a probabilistic model assumes all data to be generated from a mixture of Gaussian distributions, such that each cluster is considered as a model with a mean and a variance. Hence, the cluster variance and the membership degree of each element are found based on the probability of the point being generated from each cluster's gaussian distribution, such that the cluster centroid is the distribution's mean.
As a fuzzy clustering method, GMM has been applied in many fields and applications. It has received enormous attention in recent literature. Although GMM is capable of achieving high accuracy, Tran and Wagner (1999) believe that it has a disadvantage in the problem of sensitivity to outliers. Chappell et al. (2010) studied the analysis of vessel encoded arterial spin labeling for vascular territory mapping and found that GMM was unable to deal with the very substantial mixing within the datasets. Baid et al. (2017) compared K-means, fuzzy K-means and GMM algorithms on segmentation of brain tumer, and found that: "K-means and GMM are more susceptible to local optima and outliers in comparison to FKM algorithm. However, FKM performs better for convex shapes than what GMM and K-means do". Wu and Yang (2002) studied HKM and FKM , and they suggested alternative measure to replace the euclidean distance in K-means clustering procedures. Also, in a similar fashion Wu et al. (2009) used a non-euclidean distance.
Jain (2010) provided an overview of clustering and summarized well-known clustering methods. Another method for data stream clustering is discussed by Aggarwal et al. (2003). Moreover, Huang and Ng (1999) applied the FKM on categorical data.
One of the most important applications of the fuzzy clustering techniques is the field of missing data imputation. Actually, the problem of missing values is common and has a crucial effect on many areas of research and applications. Handling missing values can be divided into three main categories: (1) ignoring and discarding data (2) parameter estimation, and (3) imputation. FKM was widely used and employed with missing data imputation techniques. For instance, Li et al. (2004) introduced a technique for

missing values imputation with combination with fuzzy K-means clustering technique as a hot-deck imputation approach. Moreover, Salleh and Samat (2017) imputed missing data based on Fuzzy approach in heart disease classification. Also, Zhang et al. (2016) proposed a new clustering approach for missing data using FKM. Whereas, Tang et al. (2014) introduced a hybrid approach integrating FKM based imputation method with the Genetic Algorithm developed for missing traffic volume data estimation.
The outline of this article is as follows: in section 1 we introduced a background on clustering analysis and a variety of clustering techniques, in addition to the problem of missing data imputation. Section 2 introduces basic notions and preliminaries in addition to an overview of our basic argument. In section 3 we present the proposed modification on FKM, after that, in section 4 we introduce a model for missing data imputation. Then, section 5 presents implementation and experiments. Finally, In section 6 we set up our conclusions and remarks.

2 Preliminary and Notations

Since it was firstly proposed, FKM is considered as a tolerant technique for uncertainty and imprecision which makes it more convenient for problems of real life than what hard K-means clustering is, since it (i.e., FKM) employs the concepts of fuzzy sets theory proposed by Zadeh (1965). According to Sarkar and Leong (2001), FKM is unlike conventional K-means clustering, since FKM assigns each pattern to all clusters with different degrees of membership. Hence, FKM was used in a wide range of applications in various fields of research, such as Intrusion Detection Systems Gharehchopogh et al. (2012), Image processing Rahmani et al. (2014), web mining, consumers behavior, market segmentation Li and Lewis (2016),and many others.
The FKM is an optimization technique that employs the euclidean distance measure between input vectors and clusters centroids to calculate the degree of membership for a vector belonging to a specific cluster, taking into account that each vector belongs to all clusters with different non-zero degrees of membership, where each vector shares its membership with all clusters such that those degrees are sum up to one.
Since the degrees of membership (found based on the euclidean distance) for a vector with respect to all clusters are nonzero membership degrees (unless the vector under consideration is identical to any of the clusters centroids, then the degrees of membership will be zero for all clusters except the cluster that has an identical centroid with that vector), this would affect the efficiency of FKM especially when it is applied in some fields such as missing data imputation. Such argument could be explained in the following reasoning:
For a given data set of n objects; $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ where each object is s-dimensional vector. Let x_{ij} ($1 \leq i \leq n$ and $1 \leq j \leq s$) denote the value of attribute j in vector $\boldsymbol{x}_i$. Let $\boldsymbol{V} = \{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, ..., \boldsymbol{\nu}_K\}$ be the set of K clusters where $\boldsymbol{\nu}_k$ is s-dimensional vector denotes the centroid of the k^{th} cluster.

Let $d(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$ be the euclidean distance between vectors $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$ such that,

$$d(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) = \begin{cases} 0, & \text{if } \boldsymbol{x}_{i_1} \text{ and } \boldsymbol{x}_{i_2} \text{ are equal} \\ ||\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2}||, & \text{otherwise} \end{cases} \tag{1}$$

According to Sarkar and Leong (2001) in fuzzy K-means clustering technique; the distance between a data object and a cluster centroid is used to define the degree of membership function δ such that $\delta_k(\boldsymbol{x}_i)$ is the degree of a given data object $\boldsymbol{x}_i$ as a member in the cluster k, taking into account that $d(\boldsymbol{\nu}_k, \boldsymbol{x}_i) = 0$ iff $\boldsymbol{x}_i = \boldsymbol{\nu}_k$.
Fuzzy K-means clustering technique applied the concept of fuzzy sets for the clustering process. In FKM, the degree of membership $\delta_k(\boldsymbol{x}_i)$ between any vector $\boldsymbol{x}_i$ and any centroid $\boldsymbol{\nu}_k$ was defined according to the euclidean distance $d(\boldsymbol{\nu}_k, \boldsymbol{x}_i)$ as

$$\delta_k(\boldsymbol{x}_i) = \begin{cases} 1, & d(\boldsymbol{\nu}_k, \boldsymbol{x}_i) = 0 \\ \delta_k(\boldsymbol{x}_i), & d(\boldsymbol{\nu}_k, \boldsymbol{x}_i) \neq 0 \end{cases} \tag{2}$$

where, $\delta_k(\boldsymbol{x}_i)$ can be equated using equation 4 (See section 3.).
Moreover, FKM defines $\boldsymbol{M}$ to be $n \times K$ matrix such that each element in $\boldsymbol{M}$ is $\delta_k(\boldsymbol{x}_i), i = 1, ..., n; k = 1, ..., K$, to be in the i^{th} row and the k^{th} column in $\boldsymbol{M}$.
Hence,

$$\text{if } d(\boldsymbol{\nu}_k, \boldsymbol{x}_i) \neq 0, \forall k, \text{ then } \delta_k(\boldsymbol{x}_i) \neq 0, \forall\, i; k. \tag{3}$$

Note that; according to Singpurwalla and Booker (2004) and Sarkar and Leong (2001), $\sum_{k=1}^{K} \delta_k(\boldsymbol{x}_i) = 1$. So without loss of generality, assume that $\delta_k(\boldsymbol{x}_i)$ is categorized into two sets of membership degrees as follows:

$$\begin{aligned} \Delta_k^1(\boldsymbol{x}_i) &= \{\text{all membership degrees less than some small value } \epsilon\} \\ &= \{\delta_k(\boldsymbol{x}_i) | \delta_k(\boldsymbol{x}_i) < \epsilon, \text{ for some } i \text{ and small value } \epsilon\} \end{aligned}$$

and

$$\Delta_k^2(\boldsymbol{x}_i) = \{\delta_k(\boldsymbol{x}_i) | \delta_k(\boldsymbol{x}_i) \notin \Delta_k^1(\boldsymbol{x}_i)\},$$

hence, we suggest to categorize the set of vectors in each cluster into two sets as follows:

$$I_1 = \{\boldsymbol{x}_i | \delta_k(\boldsymbol{x}_i) \in \Delta_k^1(\boldsymbol{x}_i); \text{ for every vector } \boldsymbol{x}_i \text{ in cluster } k\}$$

and

$$I_2 = \{\boldsymbol{x}_i | \boldsymbol{x}_i \notin I_1\}.$$

So, in such situation when the values in $\Delta_k^2(\boldsymbol{x}_i)$ are large enough, applying the results of FKM to some applications (missing data imputation for instance), then it is obvious that the imputation is toward I_2. But having $\Delta_k^1(\boldsymbol{x}_i)$ with nonzero values makes I_1 affect the imputation process, and this would reduce the accuracy of the imputation, especially when $\Delta_k^1(\boldsymbol{x}_i)$ is small enough in values. Such situation is a result of applying the distance measure in the process of fuzzy clustering optimization for the membership function, recall equation (2), since the distance between any two nonidentical vectors will be non-zero and this will produce a non-zero degree of membership.

3 Enhanced FKM

In this section, we will explain the suggested adjustment of FKM clustering algorithm. The adjustment involves adding two more steps to the original FKM, especially when it is applied for missing data imputation. The suggested steps are: (1) Neutralization and (2) Normalization.

1. Neutralization: Equate each element of the set $\Delta_k^1(\boldsymbol{x}_i)$ to zero, to rule out its effect on the optimization process, and later on the imputation procedure.

2. Normalization: Normalize the values of $\Delta_k^2(\boldsymbol{x}_i)$, to maintain their summation equals to 1.

Recall that, according to Dunn (1973) the membership value of the vector $\boldsymbol{x}_i$ belonging to the k^{th} cluster, is

$$\delta_k(\boldsymbol{x}_i) = \frac{d(\boldsymbol{\nu}_k, \boldsymbol{x}_i)^{\frac{-2}{q-1}}}{\sum_{l=1}^{K} d(\boldsymbol{\nu}_l, \boldsymbol{x}_i)^{\frac{-2}{q-1}}}, \tag{4}$$

where q is a constant known as the index of fuzziness or the fuzzifier that controls the amount of fuzziness. Note that, $\sum_{l=1}^{K} \delta_l(\boldsymbol{x}_i) = 1, \forall i = 1, ..., n_r$.

Also, Bezdek (1981) applies the following equation to find the centroid of cluster k

$$\boldsymbol{\nu}_k = \frac{\sum_{i=1}^{n_r} \delta_k(\boldsymbol{x}_i) \cdot \boldsymbol{x}_i}{\sum_{i=1}^{n_r} \delta_k(\boldsymbol{x}_i)}. \tag{5}$$

Based on our argument, at each iteration, the former two equations will be implemented according to the following constraints:

- Check if $\delta_k(\boldsymbol{x}_i)$ is less than a specified value ϵ to determine if $\delta_k(\boldsymbol{x}_i)$ belongs to $\Delta_k^2(\boldsymbol{x}_i)$ or not. Hence,

$$\delta_k(\boldsymbol{x}_i) = \begin{cases} 0, & \delta_k(\boldsymbol{x}_i) \in \Delta_k^1(\boldsymbol{x}_i) \\ \delta_k(\boldsymbol{x}_i), & \delta_k(\boldsymbol{x}_i) \in \Delta_k^2(\boldsymbol{x}_i) \end{cases}. \tag{6}$$

- Apply the equation below to normalize the values of Δ_k^2such that$\sum_{l=1}^{K} \delta_l(\boldsymbol{x}_i) = 1$ as follows:

$$\delta_k^*(\boldsymbol{x}_i) = \frac{\delta_k(\boldsymbol{x}_i)}{\sum_{l=1}^{K} \delta_l(\boldsymbol{x}_i)}, \qquad \forall k = 1, ..., K. \tag{7}$$

Equations 4 and 5 are optimized iteratively until the stopping criterion is satisfied. Figure 1 below illustrates the suggested algorithm for EFKM.

The algorithm in Figure 1 applies two more steps (steps 7, 8) in comparison with FKM. Step 7 is the neutralization, while step 8 is the normalization according to equations 6 and 7, respectively.

The neutralization process (step 7) in the algorithm compares the degree of membership for each input vector with respect to each cluster with small positive integer ϵ such that if that degree of membership is less than ϵ, then the degree of membership will be replaced with zero and this will rule out the value of the corresponding attribute of that vector, and reduce its contribution in the imputation process.

The value of ϵ is neither specified nor random, it is rather an application dependent value. After that, step 8 (Normalization) modifies all the nonzero degrees of membership for each input vector to sum up to one.

```
INPUT:
        − q, K, ε
        − small positive integer α
        − maximum number of iteration Y
        − membership degrees matrix M_{n×K}
        − set of input vectors X
ALGORITHM:
        step 1: y = 0
        step 2: randomly initialize M^y
        step 3: y = y + 1
        step 4: for k = 1 ... k
                    calculate v_k
                end for
        step 5: find the set Z_k = {k|1 ≤ k ≤ K: d(v_k, x_i) = 0}
        step 6: if Z_k = ∅ then
                    calculate δ_k(x_i)
                else
                    δ_k(x_i) = 0 ∀k − Z_k
                    Σ_{k∈Z_k} δ_k(x_i) = 1
                endif
        step 7: for k = 1 ... K
                  for i = 1 ... n
                    if δ_k(x_i) < ε then
                      δ_k(x_i) = 0
                    endif
                  endfor
                endfor
        step 8: for i = 1 ... n
                  if Σ_{k=1}^{K} δ_k(x_i) < 1 then
                    for k = 1 ... K
                      δ_k(x_i) = δ_k(x_i) / Σ_{l=1}^{K} δ_l(x_i)
                    endfor
                  endif
                endfor
        step 9: M^{y+1} = δ_k(x_i) ∀k ∀i
        if ||M^y − M^{y+1}|| > α or y < Y then
            repeat steps 3 through 9
        end if
```

Figure 1: EFKM Algorithm

4 Imputation of Missing Data using Enhanced FKM

A common method for missing data imputation based on FKM technique was suggested by Li et al. (2004). This method consists of two phases: (1) the fuzzy clustering, which applies the FKM method to generate fuzzy clusters (i.e. fuzzy sets) and (2) the imputation process that estimates the missing values based on the fuzzy clustering in the former phase.

In this paper, the suggested technique for missing data imputation follows similar approach as in Li et al. (2004), taking into consideration (according to our earlier argument) that: having any defect in the former phase would adversely affect the later one.

In details, our proposed method of missing data imputation has three main phases:

(1) Input Vectors Partitioning:

We define the reference vector to be the one with no missing value(s), and the non-reference vector as the vector with missing value(s). Categorizing the vectors as a reference or a non-reference is a partition of the original data set. Hence, $m = n + n_{nr}$ where n denotes the number of reference vectors and n_{nr} denotes the number of non-reference vectors.

(2) Enhanced Fuzzy Clustering:

In this phase, EFKM is used to cluster the reference vectors into K clusters such that each cluster is compact and far from others, so the values of $\delta_k(\boldsymbol{x}_i)$ and $\boldsymbol{\nu}_k$ are optimized.

(3) Missing Data Imputation:

When it comes down to the phase of imputing missing data, let x^*_{ij} be the missing value of attribute j in the non-reference vector $\boldsymbol{x}^*_i$, and let x_{ij} represent the value of the j^{th} attribute of reference vector i in a specified cluster. Define D_j to be

$$D_j = \{x_{ij} | \boldsymbol{x}_i \text{ is a reference vector}\};$$

the set of all values of x_{ij}, with cardinality n_D.

Now, the value of the missing attribute x^*_{ij} is more likely to be one of the values of set D_j in any specified cluster. Hence, we define π_{j^*} as the probability that x^*_{ij} equal x_{ij}, i.e.,

$$\pi_{j^*} = P(x^*_{ij} = x_{ij}) = \frac{\text{frequency of } x_{ij}}{n_D}$$

it can also be considered as the relative frequency of x_{ij}.

Having these expressions acquired, we propose applying the following formula:

$$x^*_{ij} = \sum_{k=1}^{K} \delta_k(\boldsymbol{x}^*_i) \Big[\sum_{x_{ij} \in D_j} \pi_{j^*} \cdot x_{ij} \Big], \quad \forall \text{ non-reference vector } \boldsymbol{x}^*_i \tag{8}$$

The imputation is formed based on the specific attributes that belong to each cluster.

In this phase, to impute the missing value for each non-reference vector we need to find the degree of membership of that vector ($\delta_k(\boldsymbol{x}^*_i)$) with respect to each cluster optimized from the previous phase. Hence, The first thing to do is to find the euclidean distance between the non-reference vector under consideration and each cluster. But as mentioned earlier, a non-reference vector holds missing value(s), thus the question here is how to find the euclidean distance between a cluster centroid and a non-reference vector?

To proceed in this process, any missing value must be filled temporally with a convenient value. Many techniques are used to fill x^*_{ij} with a measure of center such as the average, the median, the weighted average ... etc. All these measures seem to work just fine, but filling the missing value by the average, for instance, of the corresponding attribute j in the reference vectors,

$$\frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad \forall j.$$

causes a level of bias to the centroids with shorter euclidean distance to that average which results inaccurate imputation especially for the non-reference vectors that do not belong to the clusters of those centroids.
On the other hand, Sarkar and Leong (2001) suggested replacing $(x^*_{ij} - \nu_{kj})^2$ by:

$$\frac{\sum_{i=1}^{n} \left(\delta_k(\boldsymbol{x}_i) \cdot (x_{ij} - \nu_{kj})^2\right)}{\sum_{i=1}^{n} \delta_k(\boldsymbol{x}_i)},$$

they stated that this substituted value will be the same for all missing values for a specific attribute, which raise the same argument: clusters with closer centroids to this value will have greater effect in the imputation process.
Our modification in this manner comes from neutralizing the missing attributes while calculating the membership value for the non-reference vector by filling the missing value x^*_{ij} with the value of the corresponding attribute in the centroid under consideration.
Then, the imputation of x^*_{ij} is as follows:

- We fill the missing value x^*_{ij} by the value of the j^{th} attribute in the centroid under consideration, i.e.,

$$x^*_{ij} = \nu_{kj} \tag{9}$$

- As the non-reference vector $\boldsymbol{x}^*_i$ is now complete, we apply equation 4 to find its membership value $\delta_k(\boldsymbol{x}^*_i)$. After that, constraints 6 and 7 are applied if necessary.
- Once the value $\delta_k(\boldsymbol{x}^*_i)$ is obtained with respect to all clusters, we apply equation 8 for every attribute j with missing value in $\boldsymbol{x}^*_i$.

Finally, this phase (i.e., the imputation phase) will be repeated for all non-reference vectors with respect to all optimized clusters.
Figure 2 below illustrates the entire model of missing data imputation process using EFKM.
According to Figure 2, the EFKM is applied to prepare the clusters by optimizing the centroids to be used later to find the membership degrees of the non-reference vectors with respect to each cluster. Those degrees are the milestone of the missing data imputation which is the last phase of the proposed model. The final output of this model is the imputed value(s) of each missing attribute(s) for every non-reference vector.

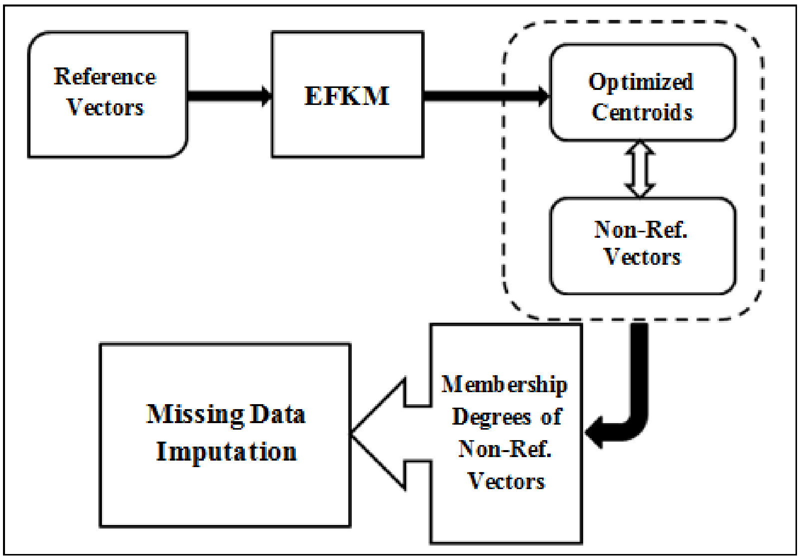

Figure 2: EFKM based Missing Data Imputation

5 Experiments and Analysis

We used two types of data sets to test our model, which are: Iris Plants Dataset and Wine Recognition Data. These datasets do not hold any non-reference vectors, so, in the partitioning phase, the algorithm picks some vectors as non-reference vectors at random basis. Such action allow us to make a reasonable evaluation since the desired value is known versus the actual value. In all experiments, the number of clusters was $K = 3$, index of fuzziness $q = 2$, the rate of missing attributes was vary, and $\epsilon = 0.005$ (taking into account that ϵ is determined according to the area of application as mentioned earlier).

The implementation process was performed through two streams:

(1) The FKM clustering is implemented to generate the fuzzy clusters. After that the proposed imputation formula (eq. 8) is performed.

(2) The EFKM clustering is implemented, and the optimized clusters will be used to make the imputation of the non-reference vectors by applying equation 8.

We used the Root Mean Squared Error (RMSE) to evaluate the imputation accuracy using EFKM in comparison with the accuracy of FKM based imputation.

Figure 3, Figure 4 below illustrate the results of implementing the streams mentioned earlier on the Iris Plants Dataset and Wine Recognition Data, respectively. In either cases the Root Mean Squared Error (RMSE) was used to evaluate the imputation accuracy. When Iris Dataset is used, the experiments were performed under two rates of missing attributes: 20% and 40%. In both rates the RMSE was in favor of the EFKM based imputation. Similarly, Figure 4 illustrates the experiments over Wine Recognition Data. These experiments were applied under two rates of missing attributes: 21% and

42%. Also, In both rates the RMSE was in favor of the EFKM based imputation.

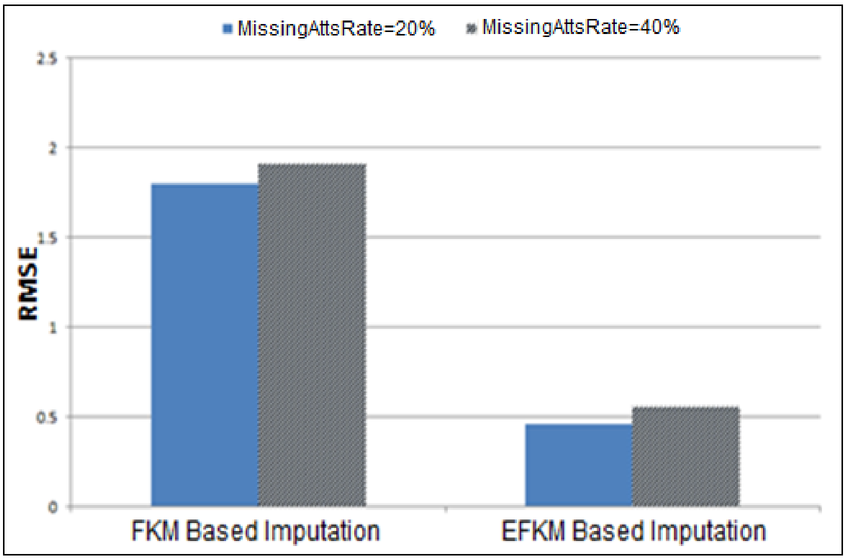

Figure 3: Experiments over Iris Plants Dataset

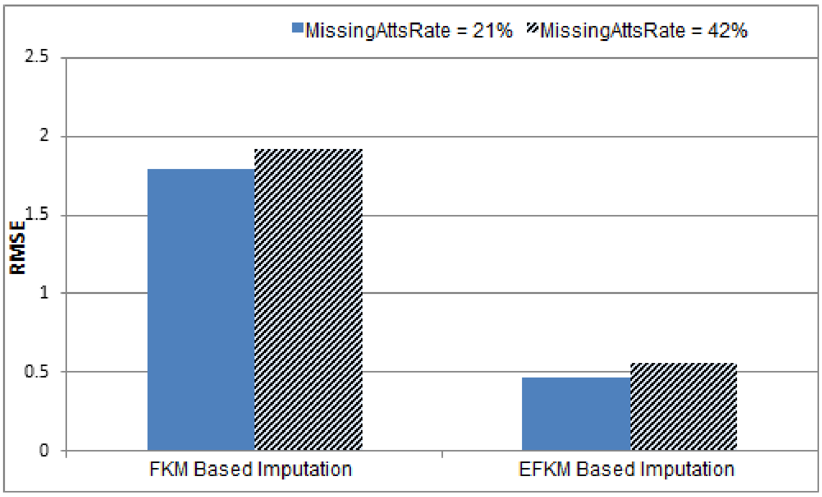

Figure 4: Experiments over Wine Recognition Data

6 Conclusion

In this paper, we developed a novel technique for missing data imputation based on a modified Fuzzy K-means clustering (EFKM). The modification of the FKM was by adding two steps to the optimization process and to the imputation phase, which are: Neutralization and Normalization. Such modification enhanced the process of clustering and consequently the imputation phase. The missing data imputation technique uses the cardinality of the set of all attribute values that appear in a specific cluster when the degree of membership of the corresponding reference vectors are non-zero, such approach sufficiently improves the imputation since it grids off any value with small degree of membership. Moreover, in this paper we suggest a technique to fill the missing values temporarily with an appropriate value to find the euclidean distance with optimized centroids, which ruled out any level of bias that may affect the imputation. Finally, the experiments showed that the proposed method (EFKM based imputation) sufficiently outperforms the FKM based imputation according to the measure of RMSE.

References

Aggarwal, C. C., Philip, S. Y., Han, J., and Wang, J. (2003). -a framework for clustering evolving data streams. In *Proceedings 2003 VLDB Conference*, pages 81–92. Elsevier.

Baid, U., Talbar, S., and Talbar, S. (2017). Comparative study of k-means, gaussian mixture model, fuzzy c-means algorithms for brain tumor segmentation.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York.

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203.

Chappell, M. A., Okell, T. W., Jezzard, P., and Woolrich, M. W. (2010). A general framework for the analysis of vessel encoded arterial spin labeling for vascular territory mapping. *Magnetic resonance in medicine*, 64(5):1529–1539.

Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J. Cybern.*, 3(3):32–57.

Gharehchopogh, F., Jabbari, N., and Ghaffari, A. Z. (2012). Evaluation of fuzzy k-means and k-means clustering algorithms in intrusion detection systems. *International Journal of Scientific Technology Research (IJSTR), ISSN 2277-8616*, 1(11):66–72.

Huang, Z. and Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Li, D., Deogun, J., Spaulding, W., and Shuart, B. (2004). Towards missing data imputation: A study of fuzzy k-means clustring menthod. *Proceedings of 4th international conference of rough sets and current trends in computing (RSCTC)*, pages 573–579.

Li, J. and Lewis, H. (2016). Fuzzy clustering algorithmsreview of the applications. *Smart Cloud (SmartCloud), IEEE International Conference on. IEEE.*

Rahmani, M. K. I., Pal, N., and Arora, K. (2014). Clustering of image data using k-means and fuzzy k-means. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 5(7).

Salleh, M. and Samat, N. (2017). An imputation for missing data features based on fuzzy swarm approach in heart disease classification. *ICSI 2017: Advances in Swarm Intelligence*, pages 285–292.

Sarkar, M. and Leong, T. (2001). Fuzzy k-means clustering with missing values. *Proceedings of the American Medical Informatics Association, Annual Symposium*, pages 588–592.

Singpurwalla, N. D. and Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets. *Journal of the American Statistical Association*, 99(467):867–877.

Tang, J., Wang, H., Wang, Y., Liu, X., and Liu, F. (2014). Hybrid prediction approach based on weekly similarities of traffic flow for different temporal scales. *Transp. Res.*

Rec. J. Transp. Res. Board, (2443):21–31.

Tran, D. and Wagner, M. (1999). A robust clustering approach to fuzzy gaussian mixture models for speaker identification. In *Knowledge-Based Intelligent Information Engineering Systems, 1999. Third International Conference*, pages 337–340. IEEE.

Wu, B., Wang, L., and Xu, C. (2009). Possibilistic clustering using non-euclidean distance. In *Control and Decision Conference, 2009. CCDC'09. Chinese*, pages 938–940. IEEE.

Wu, K.-L. and Yang, M.-S. (2002). Alternative c-means clustering algorithms. *Pattern recognition*, 35(10):2267–2278.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zhang, L., Lu, W., and Liu, X. (2016). Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. *Knowl-Based Syst*, 99:51–70.