Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 2533-2541

Research Article

Fuzzy C-Means Clustering Algorithm Using Initial Centroid Selection Process In Data Mining Techniques

[1]G.Sivabharathi , [2]Dr. K.Chitra Ph.D

[1]Assistant Professor, Mangayarkarasi College of Arts and Science for Women, Madurai ,

[2]Assistant Professor, Government Arts College, Melur

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract

Data mining technique is the discovering process of the large data sets for the pattern of making the group conjunction of machine learning, database design, and statistical reports to be analyzed. Data mining involves the steps are regression; summarization, Clustering, and association of large datasets through the various kinds of terminology should be used to form the bunching of datasets to group the datasets. The clustering method is a crucial way to calculate the distance between the centroids. The Clustering is the process of grouping the data, which is used to calculate the objects that are similar in characteristics and group together. This clustering method is used to choose the cluster centers and the centroids, which is calculate the distance among the objects. In this paper, we focus on the centroids to form the gap between the objects at a minimum requirement of clusters. The defined initial centroids are compared with the randomly selected initial centroids. By this way, the centroids distance calculated through the cluster formation of data. The initial clusters are augmenting the centroids, which is depends on the minimum distance. This initial centroid selection process enhanced the range among the objects. The overall selection of the centroids produces the better quality and the minimum range, Square Error, Precision and Recall calculation which is based on the threshold value. These clustering algorithm and the initial selection process used in the telecommunication and image segmentation based medical industry analysis. In fuzzy c means Clustering, considers each object with a member of initial clusters to the degree of nearness distance through the simulation of MAT LAB.

Keywords: Data mining, initial centroid selection, clustering algorithm, fuzzy c-means clustering algorithm, MAT LAB Simulation.

I. INTRODUCTION

Fuzzy c-means clustering algorithm is the role to established between the clusters with a grouping of the objects. **Anter, A. M et al. (2019)** [1] Stated this clustering algorithm is the unsupervised machine learning algorithm, which enables the information from the datasets through each center. The speed of the centroid process, which makes the distance between the clustering center to each group. **Arora, J.et al. (2019)** [2] illustrate the low interclass and high intra-class, which is in proper cluster formation. To evaluate the right distance metrics in Clustering, we have to use the high intra-class similarity. The distance can be squared to adverse the progressively higher weight on objects that are central to the initial centroids. **Busa, S. et al. (2019)** [3] reviewed the accuracy of fuzzy c-means (FCM) clustering algorithm makes the grouping strategy in the initial centroids from the datasets. The dataset was having a place with each cluster to gathered into n groups with each data point with a specific degree. **Motwani, M.et al. (2019)** [4] stated the calculation of distance metric between each data point and the position of the centroids of initial centroids are depends on the datasets. The maximum distance is analyzed as the initial centroid of the first initial centroids are clustered. The rest of the data points have remained for initial centroids. The selecting of each data point and its probability is proportional to

the square of the distance between this point and its nearest initial centroid. **Mahajan, M. et al. (2019**) [5] reported the centroids and sum of lengths of objects to the centroid of their cluster, which establish the ratio of maximum distance Initially to calculates their distance from other centroids and allocates datasets to their nearby centroids. To decide the dataset whether to keep or move to the next level of the centroid, which depends on the distance functions. leads to reducing the time and increasing the efficiency of the traditional fuzzy c means clustering algorithm.

Problem Identification

From this paper, the distance calculation and the initial centroid selection process are not yet appropriately done due to the reason for the random selection of the dataset, which is the factor to identify the exact cluster grouping using a k-means clustering algorithm.

Objective

- To propose the algorithm for fuzzy c-means clustering approach.
- To propose the distance calculation, square error, and precision level during the Clustering of the dataset in the initial centroid selection process.

Organization

The works in the proposed system and it's working methods discussed in sections III and IV.

The performance analyses with its comparison shown in section V.

II. RELATED WORKS

Mohebian, R. et al. (2019) [6] Stated the degrees are updated when the centers referred to each cluster and their datasets. The distance between each value and related cluster centers based on the cost of intra-class function. The membership's degree is equal to the distances that produced the collection of cluster centers with their corresponding cluster set. **Hashemzadeh, M. et al. (2019)** [7] established a better understanding of the complicated relationships between the clusters, provide a better understanding of initialization and the samples in the fuzzy membership matrix. The policy is to prevent the formation of poor-quality clusters systematically, which is to make the algorithms independent of a random initialization and the system are make the groups of approaches of the cluster structure. **Srinivas, B.et al. (2019)** [8] reviewed the K-means algorithm that is belonging to each group instead of belonging to only one group in the situation. The each point has a likelihood parameters like Segmented area, Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR)are attained in Fuzzy c means clustering,. **Bilenia, A. et al. (2019)** [9] stated the fuzzy c-means algorithm provides the information in the image better performance. Since the local neighborhood pixels are belonging to make the spatial information and high correlation with the needs to be accounted for and characterizing the segmentation method. **Ren, T. et al. (2019)** [10] established the initialization of the cluster center set that are strongly dependent on the initial clustering center set of initial centroid at the best performance. The data set is to be clustered as in the input, which has make the features, and it makes a matrix U of c rows and n columns in the output to be noted.

Lakshmi, M. A. et al. (2019) [11] considered the initial centroids and k are requires in K-means clustering algorithm, which is the number of clusters as input, which selected as random of the initial centroid. The results produced as random output. **Sangaiah, A. K. et al. (2019**) [12] stated the clustering algorithm used for better yield than the other clustering method. To assign the each document are to set labeled training data with the nearest cluster with

the help of better initial centroids. **Dutta, S. et al. (2019)** [13] reviewed the representation of the dataset, which is independently and make the features which are sufficient to enhance the distance. The Clustering is a natural way of information retrieval at a well-known problem stated. This also enables the reduced dimension of the dataset and reduces the clustering time effectively through the way summarization of the dataset task. **Aljarah, I et al. (2020)** [14] stated the classification and regression, which is the part of Clustering. The supervised learning method which considers as the data labels are available are in the centroid of initial selection. The clustering algorithm is tested with the different characteristics to evaluate its performance through the real data sets. **Zhang, C. et al. (2019)** [15] stated the application of the clustering algorithm provides the initial conditions and restrictions of Clustering.. The clustering analysis in data mining works on highly efficient when the cluster mass data analysis enhanced the performance, clustering analysis, classification, hierarchy, density, and model clustering algorithms.

III. PROBLEM STATEMENT

The various research and data mining techniques used to enhance the distance between the clusters and the distance among the objects. The data point in the cluster made the quality of the best efficiency and the error through the clustering algorithm. This problem established every one data point to the random selection of clusters compared to the given dataset. Although the issue in the Clustering of the initial centroids selection process made the random selection of initial centroids instead of making the way each centroid has the distance among the objects in the cluster dataset. These causes are affected by the initial centroid selection process to detriment the distance calculation among the clusters.

IV. PROPOSED METHODOLOGY

In the proposed method, the fuzzy c-means clustering algorithm used to calculate the distance between the initial centroid selection process. In which, the parameter used in the centroid calculation between the clusters will depend on the number of groups used (K) and the number of centroids used to calculate the distance. As the fuzzy c-means (FCM) algorithm, where the Clustering based on the intensities of the cluster dataset, which is very sensitive to the closet of the initial centroid, and the addition of relationship between the clusters have formed the distance through the initialization and the redefined dataset among the surface distances. Figure 4.1 shows the Flow diagram of the fuzzy c-means clustering algorithm with its details.

4.1 FLOW DIAGRAM

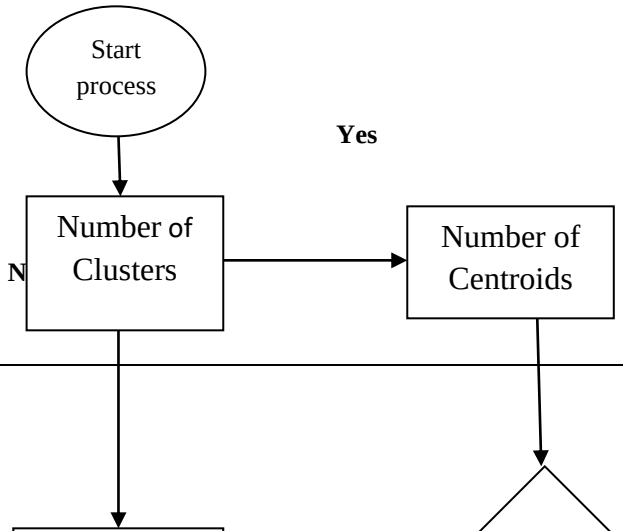

Fig 1 Fuzzy c-means algorithm Flow diagram

The Clustering of the initial centroid selection process enhanced the distance among the clusters in the dataset. The calculation of dataset between the object and cluster enhanced representatives. Euclidean, cosine, correlation, and city block distances are the methods to evaluate the distance among their performance in the algorithm It denotes the calculation of expected distances functions through the initial centroid selection at a grouping of dataset on different datasets to be enumerated.

4.2 Clustering algorithm Process

There are many techniques in the clustering algorithm, which enable the improved centers and the centroids to calculate the minimum distance for the initial centroid process. From this fuzzy c-means algorithm, which enumerates the factors like clusters, numbers of centroids used. The calculation that ensembles the process of whether it keeps to the next level like as (YES) or move to the calculation of the initial centroid (NO) depends on the computation of the dataset. If the cluster and the centroid are correct, they are move to get the unique data to access the initial centroid selection process. This method continuously going to obtain factors like minimum distance, error, accuracy, precision is to be calculated. From this method, the parameters easily computed through the redefined centroid parameters with the comparative study of combined clustering datasets. The machine-level algorithm based on the datasets and these clustering algorithms based on the centroids are optimum levels. The performance of the clustering algorithm

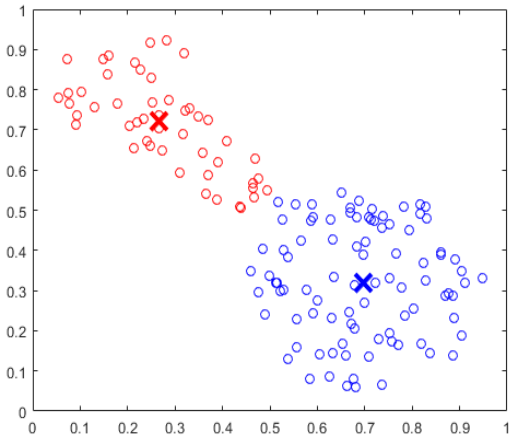

Fig. 2 clustering of datasets

Figure 2 illustrates the Clustering of datasets

4.3 Importance of Choosing Initial Centroids

Usually, we use the k-means algorithm, which enables and selects the initial centroids randomly. So, it made the problem of inappropriate centroid selection that can cause inappropriate efficacy. When we are constructing a cluster because the cluster's accuracy and quality firmly depend on the initial centroid and that selected the increase in computational time requirements. Each data point assigned to the closest group. So, we represent the fuzzy c-means algorithm to access the improved distance calculation at a minimum time.

4.4 Distance Calculation

For the initial centroid selection process, the fuzzy c-means algorithm enables the equation to implement the number of clusters used to find the initial centroid selection process.

$D(p_i, c_j)$ - $\int f_i(y) d(x, c_j) dx$ ……. (1)

The above equation which denotes the parameters like as

p – Denotes the datasets point

c – Number of clusters used

d - Distance calculation between the point y, from$f_{i\,(y)}$. The cluster representation is cj and the equation taken as integration over the uncertainty region like pdf integrals. To calculate the minimum distance between the two objects as a and b. The higher probability that a and b fall in the same cluster of the dataset.

The equations to calculate the minimum distance between the two objects are

D (t_a, t_b) = $(\sum_{i=1}^{m} |w_{t,a} - w_{t,b}|2)1/2$……. (2)

Where,

t - Denotes the time (1 2,3….)

We use the Term Frequency Inverse Document Frequency (TFIDF), the total number of frequency, to enable the initial centroid distance.

In the fuzzy c-means clustering algorithm, the initial centroid selection process enumerates the distance between the clusters at an equation of

$d_{x,v}=\sqrt{(x_1 \ \_ v_i)^2 + (x_2 - v_i)^2}$ …….. $\sqrt{(x_n)^2 - (v_i)^2}$

…… (3)

By using this equation, we compare the two datasets with the calculation of the minimum distance. The datasets in the clusters form the centroid randomly and compare to the random selection. To analyze the minimum distance among the nearest way of centroids. This formula used to find the strength of the relativity between the clusters. This fuzzy c-means clustering algorithm attains the best quality of centroid formation, and the performance of the process enhanced. The initial clustering centers connected to the objects that are obtained using the Simulation experiment. This method enhanced the better accuracy and clustering quality in comparison with another clustering algorithm. Where the distances calculated through the given number of clusters and number of centroid used to enable the information. The clusters are' K seems to the data point. The threshold is the limits obtained using the parameters like as μ,β.σ corresponding values.

4.5 Steps to involve in initial centroid selection

Step 1:Calculate the range of the clusters i=0,1,2,.....N-1 and j=0,1,2,..... N-1 as initial centroid used for the sample of 1,2,3 datasets to be calculated.

Step 2: To establish the center of the cluster formation.

Step 3: Find the total mean value. Fix 1 is its first centroid data point.

Step 4: Find the minimum distance between the initial centroids.

Step : To measure the Threshold value

Step 6: To find the precision, recall and square error of the function

4.6 Threshold Calculation

Let's take μ described as the significance level. When the probability of fuzzy nearness degree probability denoted as P(d) < μ. By comparing the equation as P(d(x) > (μ − bα)) < μ. Lets take the values for β = 0.05, and b = 1 using the normal distribution characteristic. The threshold T is set in two cases:

(1) If β − bα > μ, the threshold is = β−bα

(2) If β − bα < μ, the threshold is = β+bα

The background model described as the fuzzy nearness degree distribution is concentrated. From this time, the threshold values obtained to enable the working condition as possible.

4.7 Pseudocode for the fuzzy c-means clustering algorithm

1) Initially, to fix **the number of Clusters (c)** used for grouping, Fuzzy degree for optimization (F), and error (€).

2) To set the cluster center as $\mathbf{C_I = 0}$ for initialized for random process.

3) **n = 1**

4) **Where**, $\left|(c_i^{(n)} - c_i^{(n-1)})\right|$>€

Using the cluster center, the membership matrix calculation is, $U^{(n)}$ by:

$$U_{ij}^{(n)} = \frac{1}{\Sigma_{l=1}^{c}\left\{\frac{d(x_j \ , \ c_j^{(n-1)})}{d(x_j \ , \ c_l^{(n-1)})}\right\}^{\frac{2}{m-1}}}$$

To use the matrix of membership $U^{(N)}$, Cluster center is c_i^n is described as,

$\frac{\sum_{j=1}^{N}(u_{IJ}^{(n)})^{m} \; x_j}{\sum_{j=1}^{N}(U_{ij}^{(n)})^{m}}$, where n = n+1.

5) **Return** cluster centers c_i and membership degree U_{ij}.

V. PERFORMANCE ANALYSIS

The performance analysis of the proposed fuzzy c-means clustering algorithm enhanced the following parameters are Error, Distance; Precision is to calculated through the grouping of the initial centroid selection process. The results are performed and the existing method and the proposed method comparison are like a Partitioning Clustering Algorithm (PCM) and Fuzzy C-means Clustering Algorithm (FCM). The comparison which analyzed the performance of KM, PCM and FCM along with the working efficacy and the clusters formed the initial centroids process is enabled in the MAT LAB.

The factors are involved in FCM are calculated and displayed as follows with K-means and PCM.

Square Error

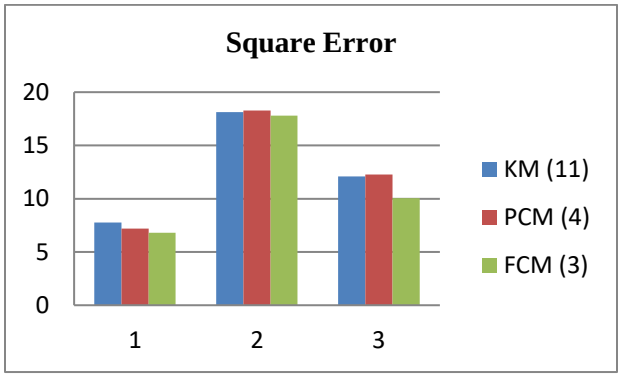

Fig.3 Square error comparison

Figure 3 represents the square error comparison of KM, PCM, FCM. The Recall and Precision values are calculated through TP, TN, FP, FN values. These are inter dependent class between each other in the machine learning method.

Distance

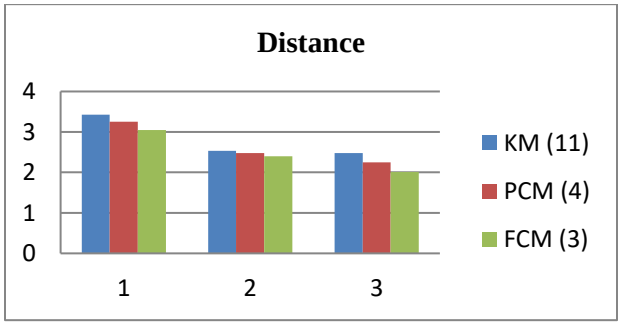

Fig. 4 Distance comparison

Figure 4 represents the Distance comparison of KM, PCM, and FCM. Where, the Recall, Precision methods are calculated through the positive and negative values depends on the threshold values of the initial centroid cluster selection process. It follows as

Recall = $\frac{TP}{TP+FN}$; Precision = $\frac{TP}{TP+FP}$

Where,

TP – True Positive; TN – True Negative

FN – False Negative; FP – False Positive

Precision

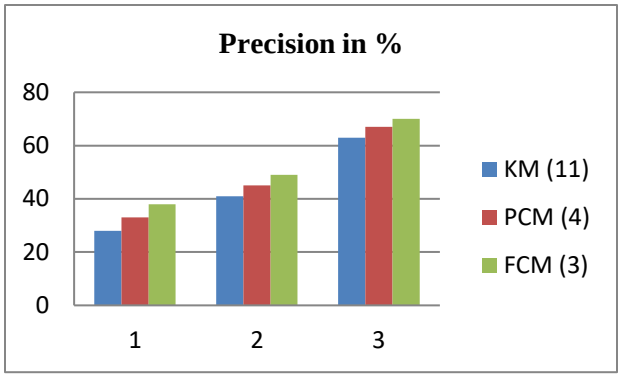

Fig 5 % of Precision comparison

Figure 5 represents the precision comparison of KM, PCM, and FCM.

Recall

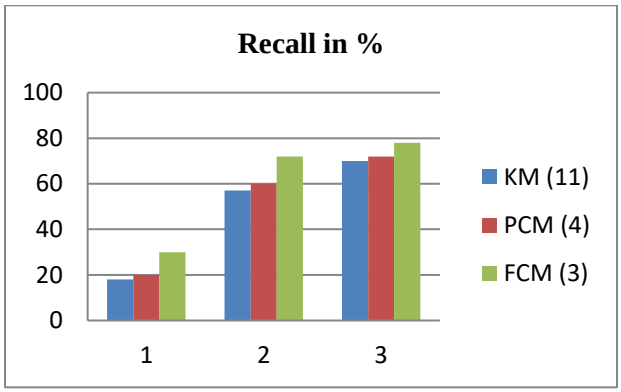

Fig 6 % of Recall comparison

Figure 6 illustrates the % of Recall comparison in KM, PCM, and FCM. The performance analysis gives the best results in the fuzzy c-means clustering algorithm in the initial centroid selection process.

VI. CONCLUSION

The paper discussed the data mining techniques based initial centroid selection process by using the fuzzy c-means algorithm. This algorithm should follow the clustering techniques to find the defined clustering formation through the grouping of the initial centroids. This method enumerated the better quality of performance, and yields attained in the market analysis and industrial analysis via telecommunication field and the image segmentation kind of medical fields. From this fuzzy c-means clustering algorithm to calculate the various parameters like minimum distance, Squared Error, Recall and Precision parameters attained with the activities of the clustering methods. The data mining techniques enabled by using this fuzzy c-means clustering algorithm.

References:

1. Anter, A. M., Hassenian, A. E., & Oliva, D. (2019). An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural. *Expert Systems with Applications*, *118*, 340-354.

2. Arora, J., Khatter, K., & Tushir, M. (2019). Fuzzy c-means clustering strategies: A review of distance measures. In *Software Engineering* (pp. 153-162). Springer, Singapore.
3. Busa, S., Vangala, N. S., Grandhe, P., & Balaji, V. (2019). Automatic Brain Tumor Detection Using Fast Fuzzy C-Means Algorithm. In *Innovations in Computer Science and Engineering* (pp. 249-254). Springer, Singapore.
4. Motwani, M., Arora, N., & Gupta, A. (2019). A study on initial centroids selection for partitional clustering algorithms. In *Software Engineering* (pp. 211-220). Springer, Singapore.
5. Mahajan, M., Kumar, S., & Pant, B. (2019). A Novel Cluster Based Algorithm for Outlier Detection. In *Computing, Communication and Signal Processing* (pp. 449-456). Springer, Singapore.
6. Mohebian, R., Riahi, M. A., & Kadkhodaie, A. (2019). Characterization of hydraulic flow units from seismic attributes and well data based on a new fuzzy procedure using ANFIS and FCM algorithms, example from an Iranian carbonate reservoir. *Carbonates and Evaporites*, *34*(2), 349-358.
7. Hashemzadeh, M., Oskouei, A. G., & Farajzadeh, N. (2019). New fuzzy C-means clustering method based on feature-weight and cluster-weight learning. *Applied Soft Computing*, *78*, 324-345.
8. Srinivas, B., & Rao, G. S. (2019). Performance evaluation of fuzzy C means segmentation and support vector machine classification for MRI brain tumor. In *Soft computing for problem solving* (pp. 355-367). Springer, Singapore.
9. Bilenia, A., Sharma, D., Raj, H., Raman, R., & Bhattacharya, M. (2019). Brain tumor segmentation with skull stripping and modified fuzzy C-means. In *Information and Communication Technology for Intelligent Systems* (pp. 229-237). Springer, Singapore.
10. Ren, T., Wang, H., Feng, H., Xu, C., Liu, G., & Ding, P. (2019). Study on the improved fuzzy clustering algorithm and its application in brain image segmentation. *Applied Soft Computing*, *81*, 105503.
11. Lakshmi, M. A., Daniel, G. V., & Rao, D. S. (2019). Initial Centroids for K-Means Using Nearest Neighbors and Feature Means. In *Soft Computing and Signal Processing* (pp. 27-34). Springer, Singapore.
12. Sangaiah, A. K., Fakhry, A. E., Abdel-Basset, M., & El-henawy, I. (2019). Arabic text clustering using improved clustering algorithms with dimensionality reduction. *Cluster Computing*, *22*(2), 4535-4549.
13. Dutta, S., Ghatak, S., Das, A. K., Gupta, M., & Dasgupta, S. (2019). Feature selection-based clustering on micro-blogging data. In *Computational Intelligence in Data Mining* (pp. 885-895). Springer, Singapore.
14. Aljarah, I., Mafarja, M., Heidari, A. A., Faris, H., & Mirjalili, S. (2020). Multi-verse optimizer: theory, literature review, and application in data clustering. In *Nature-Inspired Optimizers* (pp. 123-141). Springer, Cham.
15. Zhang, C., Hao, L., & Fan, L. (2019). Optimization and improvement of data mining algorithm based on efficient incremental kernel fuzzy Clustering for large data. *Cluster Computing*, *22*(2), 3001-3010.
16. Hussain, S. F., & Haris, M. (2019). A k-means based co-clustering (kCC) algorithm for sparse, high dimensional data. *Expert Systems with Applications*, *118*, 20-34.