International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Fuzzy C-Mean Technique for Accessing Large Database of Banking Sector

Banshidhar Choudhary[1] and Prof. Vipin Saxena[2]

Submitted: 07/05/2023 **Revised**: 16/07/2023 **Accepted**: 08/08/2023

Abstract: The Fuzzy C-Means (FCM) technique has gained significant attention in the field of data analysis and clustering due to its ability to handle complex and ambiguous data sets. In this paper, FCM technique is applied to banking data using Python and R programming languages. The objective of this study is to explore the potential of FCM in clustering of the banking data and to evaluate its performance in comparison to K-Means clustering algorithm. The paper begins by providing an overview of the FCM algorithm and its underlying principle. Thereafter, the process of preprocessing and preparing the banking data for analysis is done, further FCM algorithm in both python and R have implemented after utilizing the respective libraries and packages. The computed results are compared with widely used clustering algorithms, such as K-Means and Hierarchical clustering.

***Keywords**: Fuzzy C-Means, K-Means, Banking Data, Python and R Language*

1. Introduction

In the past few years, the banking sector has experienced a remarkable surge in the volume and intricacy of available data. The ability to extract meaningful insights from vast amount of data has become crucial for banks and financial institutions to enhance the decision-making process, under customer behaviour and manage risks effectively. Clustering techniques provide a powerful means to analyse and extract patterns from such data, enabling banks to gain valuable insights into customer segmentation, fraud detection and personalized marketing strategies. Among the various clustering techniques, the FCM algorithm has emerged as a prominent method due to its ability to handle ambiguity and uncertainty inherent in banking data. The utilization of fuzzy logic principal in FCM allows for the assignment of membership values to data points, indicating the degree to which each point belongs to specific cluster centres. This approach offers greater flexibility in representing the data structure and facilitates a more comprehensive understanding of the connections between customers, transaction and risks.

The objective of the present paper is to investigate the application of FCM technique for clustering the banking data, leveraging the capabilities of Python and R programming languages. Python and R are widely used in data analysis and have extensive libraries and packages that facilitate the implementation of FCM and other clustering algorithms. The objective is to offer researchers and practitioners a comprehensive understanding of FCM in the banking domain by leveraging the capability of both programming languages.

This study focuses on three key aspects i.e. Pre-processing and preparation of banking data, implementation of FCM algorithm in Python, R, and comparative analysis with traditional clustering algorithm such as K-Means and hierarchical clustering. The pre-processing phase involves data cleaning, transformation and feature engineering to ensure the data is suitable for clustering analysis thereafter employed the FCM algorithm, adjusting the degree of fuzziness parameter, to obtain fuzzy partitions and evaluated the clustering results. The comparison with other clustering techniques aims to assess the performance of FCM in terms of cluster quality, interpretability and computational efficiency. It is observed that FCM will demonstrate advantages in capturing the inherent fuzziness and complexity of banking data, leading to more accurate and informative clusters compared to traditional approaches.

The results of the present work will make a valuable contribution to the existing knowledge base regarding data analysis in the banking sector, specially focusing on the application of FCM for customer segmentation, fraud detection and risk management. Furthermore, by providing implementation examples in both Python and R, the main aim is to enable the potential of FCM technique to own banking datasets. The comparative analysis with traditional clustering algorithms will provide a comprehensive evaluation of FCM' effectiveness and highlights its advantages in handling the complex and uncertain nature of banking data.

Department of Computer Science
Babasaheb Bhimrao Ambedkar University, Lucknow, India, 226025
[1]ORCID:0009-0000-3405-3790, [2]ORCID:0000-0003-1035-1704
[1]banshidharphd@gmail.com, [2]profvipinsaxena@gmail.com

2. Related Work

From the literature, it is observed that limited research papers are available for applications of the fuzzy concepts for the large database, however some of the important papers are described here. Kaymak et al. [1] proposed extension of the objective function for dealing the issue of the fuzzy based clustering. The extensions are like the prototype, which are extended to hyper volumes and cluster measuring by assessing the similarity among clusters during optimization. Tsekouras[2] also proposed a model for fuzzy clustering based algorithm, which incorporates unsupervised learning with an iterative process in the framework and based on the use of the weighted fuzzy C-Means. The algorithm is successfully applied to three test cases, where the produced fuzzy c-means prove to be very accurate as well as compact in size. Yuet al. [3] have investigated an evolutionary fuzzy neural network using fuzzy logic Neural Network (NNs) and Genetic algorithm (GAs) for financial prediction with hybrid data input sets from different domain and the simulated results indicate that hybrid iterative evolutionary learning is better than the previous training algorithm. Popovicet al. [4]demonstrated a model based on fuzzy methods for Churn prediction in retail banking, four different prediction models called as prediction engines which were developed and the prediction engine using these sums performed best in churn prediction applied in both balanced and non-balanced test cases. Authors [5] have analysed financial markets based on fuzzy C-Means and pointed out further quantify key factor of securities business wave using FCM. Martin et al. [6] proposed hybrid model for bankruptcy using FCM, clustering, Multivariate Adaptive Regression Splines (MARS) and GA. The performance of existing model can be improved by selecting the best feature dynamically which depends on the nature of firms.Aguiar[7] focussed on two techniques based on fuzzy sets, a clustering algorithm and the fuzzy transform shown incorporate intrinsically the heuristics and anchoring of the behavioural finance theory. Zhangand Havens [8] proposed Streaming Kernal Fuzzy C-Means (stKFCM) algorithm, which reduced both computational complexity and space complexity significantly and pointed out the solution of two problems of big data like volume and real time application. Hakki et al. [9] focused on classification of the deposits and participation banks of turkey regarding soundness for FCM clustering method which relies on fuzzy logic. Costea[10] applied a technique of fuzzy logic namely FCM clustering and artificial intelligence algorithm for evaluation combatively the financial performance of Non-banking Financial Institutions (NFIs) in Romania. Balamuruganand Mathiazhagan[11] studied the fraud detection of credit cards using fuzzy logic and K-means which are developed and found that FCM produced a better result comparing to the other data mining techniques. Further, Behera and Panigrahi[12]proposed a FCM clustering algorithm which is applied to find out the normal usages pattern of credit cards user based on past activity. Once a transaction is found to be suspicious, neurone work, based learning mechanism is applied to detect whether it is false alarming or not .Fahad [13] focused on the work of an extensive empirical evaluation of the three significant aspects like Genetic K-Means, bisecting K-Means, FCM and Genetic C-Means in which Genetic K-means performance is best. Amirkhani et al. [15] proposed a new Type 2 FCM where authors used density concept and improved accuracy comparatively with previous existing methods .Saeed and Hani[16] developed a system known as Intelligent Type-2 Fuzzy Logic System (FLSs) which can detect debit cards frauds using real world datasets extracted from financial institution in Sudan. Wuet al. [17] proposed a hybrid fuzzy clustering algorithm, which combines the Crow Search Algorithm (CSA), and a fireworks algorithm and the performance is evaluated using eight well-known UCI benchmarks. The experimental analysis concluded better than other techniques. Jainet al. [18] developed a Credit Cards Fraud detection (CFD) system using FCM and comparison of performance of K-Means and C-Means clustering methods and found FCM gives better results such as false alarm rate, balanced classification rate, fraud catching rate and Mathews correlation coefficients is used for skewed data .Kaminskyi et al. [19] focused on extended review machine learning application in customer banking in which different types of regressions, fuzzy clustering, neural network, and principal component are used. This includes several applications for scoring models and fuzzy clustering application. All applications were realised on real data from the Ukrainian banking industry.

3. Proposed Work

To implement FCM in the banking sector for credit cards, the following steps are considered.

1. Fuzzy_C_Means ( ) Randomly selects C cluster centers from the given data points;
2. For each data point, the fuzzy membership (μ_{ij})to each cluster centre is computed based on the distance between the data point and the cluster centre, using the formula $= 1/\sum_1^c \ 1/(d_{ij}/d_{ik})^{\ 2/mr}$ ….(1), where c is the no of cluter centre, d_{ij}represents the Euclidean distance between i^{th}data to j^{th} cluster, d_{ik}represents the Euclidean distance between ith and jth cluster , m is fuzziness index, r
3. Fuzzy centers (v_j) are calculated by taking the weighted average of the data points, considering

fuzzy memberships. The formula for computing $v_j$is $(\Sigma^{n}_{i=1} (\mu_{ij})_m x_i) / \Sigma (\mu_{ij})^m)$ ….. (2)

4. Steps 2 and 3 are repeated iteratively until a termination criterion is achieved. The termination criterion is based on the minimum difference ($||U^{(k+1)}-V^k|| < \beta$) between the updated membership matrix and the cluster centre, where k represents the iteration steps and β is the termination threshold the termination criterion between [0, 1].

The algorithm aims to minimise the objective function (j), which quantifies the clustering quality. The objective function is calculated by summing the weighted squared Euclidean distance between the data points and corresponding cluster centres. The entire steps are represented through following flow diagram:

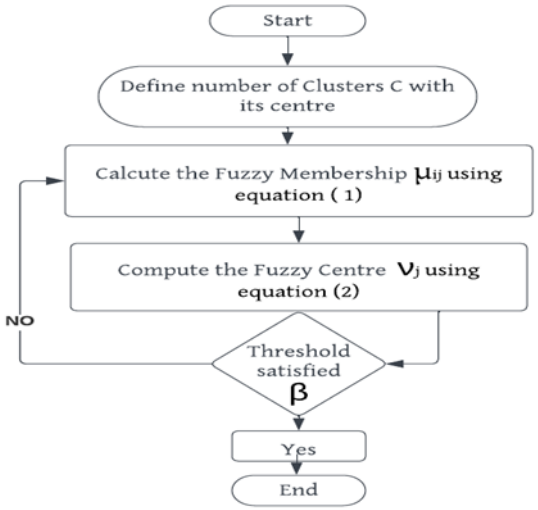


Fig 1. Flow Diagram of Fuzzy C-Means Clustering.

4. Results and Discussion

The proposed algorithm is implemented over the sets of banking industry. The dataset is related to debit cards data from different banks, a data set was formed which is consisting of data points with attributes such as ID, Age, Card1, and Card2and Card3. The purpose of clustering was to group the data points using FCM with Python language. The FCM results for Card1 vs Age, Card2 vs Age, and Card3vs Age are shown in figure 2(a), (b) and (c) , respectively. The FCM algorithm was applied with k=3 (indicating three clusters) and a fuzziness parameter of m=1.5.

Table 1. Data Set of Debit Cards of Bank.

	ID	Age	Card1	Card2	Card3
0	101	52	73	64	66
1	102	32	61	70	89
2	103	44	42	74	91
3	104	38	67	73	80
4	105	35	80	86	84
5	106	28	59	71	89
6	107	25	65	59	67
7	108	43	84	69	91
8	109	36	65	68	72
9	110	27	84	58	54
10	111	45	77	73	81
11	112	42	48	70	80
12	113	30	82	96	86
13	114	47	82	70	76
14	115	53	94	68	79
15	116	55	66	70	57
16	117	50	70	72	72
17	118	48	61	61	85
18	119	26	76	82	75
19	120	33	73	68	66
20	121	39	79	83	65
21	122	41	87	57	100
22	123	54	101	60	76
23	124	29	65	79	71
24	125	40	75	62	78

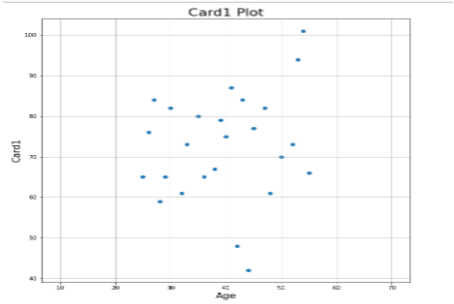


Fig 2 (a).. Card1vs Age plot

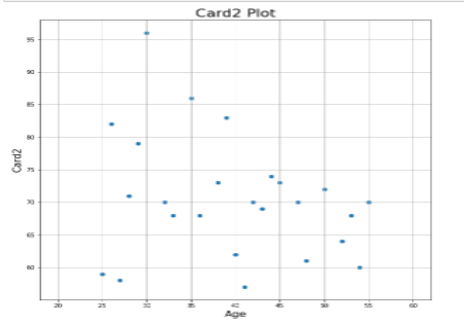


Fig 2 (b). Card2 vs Age

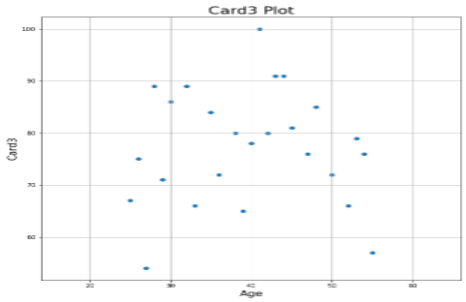


Fig 2(c). Card3 vs Age

Further second iteration cluster centre is computed based on distance of data point to cluster centre and membership matrix is obtained and is shown in the figure 3 as Representation of Membership Matrix.

```
Cluster Centers:
[[ 0.83682153  1.59147108 -0.81767348  0.01916381]
 [-1.2734103  -0.63206262  0.6652121  -0.31979548]
 [-0.28830022 -0.21941479 -0.61215247  0.47702242]]
------------------------------
Membership Matrix:
[[8.73720595e-01 1.02739817e-01 2.35395888e-02]
 [1.59824772e-03 9.95566918e-01 2.83483426e-03]
 [5.20330061e-02 8.82385003e-01 6.55819906e-02]
 [2.24585014e-02 9.42164427e-01 3.53770717e-02]
 [3.57399039e-04 2.39883700e-04 9.99402717e-01]
 [9.50054590e-03 9.66965646e-01 2.35338082e-02]
 [9.53723194e-02 8.31492933e-01 7.31347474e-02]
 [9.95558464e-01 1.39608544e-03 3.04545033e-03]
 [2.49919481e-03 9.95500601e-01 2.00020388e-03]
 [6.45047210e-01 1.81969596e-01 1.72983194e-01]
 [8.57717385e-01 5.96230344e-02 8.26595808e-02]
 [2.14377872e-02 9.56018625e-01 2.25435879e-02]
 [1.51887035e-02 1.39616146e-02 9.70849682e-01]
 [9.95561997e-01 1.71856513e-03 2.71943756e-03]
 [9.45781560e-01 1.76861067e-02 3.65323331e-02]
 [4.10593554e-01 5.06624805e-01 8.27816408e-02]
 [5.22806915e-01 3.83547006e-01 9.36460793e-02]
 [1.15973356e-01 8.59973282e-01 2.40533622e-02]
 [6.43134128e-03 1.16543958e-02 9.81914263e-01]
 [2.41481507e-01 5.82946416e-01 1.75572077e-01]
 [1.38432700e-02 6.55786025e-03 9.79598870e-01]
 [9.69091020e-01 1.66814620e-02 1.42275183e-02]
 [8.92638564e-01 4.19553456e-02 6.54060906e-02]
 [3.17170408e-02 5.67099104e-01 4.01183856e-01]
 [8.39599138e-01 1.33433464e-01 2.69673986e-02]]
```

Fig 3. Representation of Membership Matrix

On the basis of basis of above membership matrix and cluster centre we get the final cluster centre vector is obtained and represented in the figure 4.

```
[[39.45454545454545, 76.63636363636364, 72.81818181818181],
 [40.285714285714285, 70.28571428571429, 68.57142857142857],
 [39.42857142857143, 68.71428571428571, 68.85714285714286]]
```

Fig 4. Representetion of Final Cluster Centre

On the the basis of above cluster centre, the initial random cluster of Card1 vs Age is shown in the figure 5 and final cluster is obtained and shown in the figure 6 for Cards. All the Cards have been clustered and shown finally in the figure 7.which is given in figure 9 and figure 10.

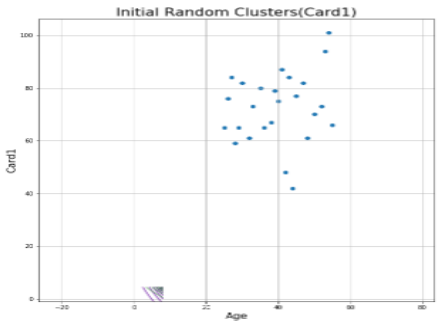


Fig 5. Representation of Initial Random Cluster

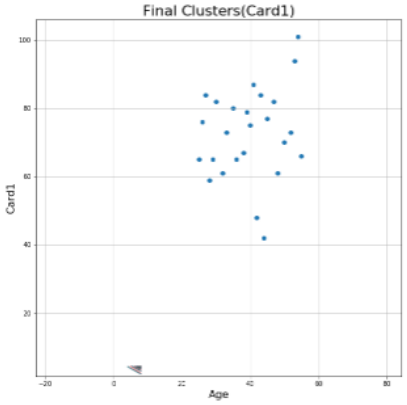


Fig 6. Representation of Final Cluster of Card1

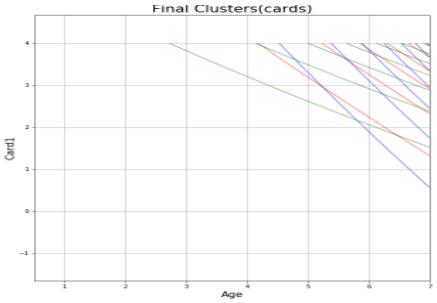


Fig 7. Representation of Final Clusters of All Cards

On the other hand, FCM is a soft clustering algorithm that assigns a membership degree to each data point for each cluster, indicating the degree of belongingness to each cluster. This allows data points to belong to multiple clusters simultaneously. FCM considers the fuzziness or uncertainty in the data and provides a more flexible, and nuanced clustering result.

From the results, it is observed that K-Means produces crisp, distinct clusters where each data point is assigned exclusively to one cluster. FCM, on the other hand, provides a more probabilistic view of cluster membership, allowing for overlapping or fuzzy boundaries between clusters. This can be advantageous when dealing with data that may exhibit inherent ambiguity or when data points belong to multiple clusters simultaneously.

The choice between K-Means and FCM depends on the specific characteristics of the dataset and the desired

interpretation of the clustering results. K-Means is often preferred when clear, non- overlapping clusters are expected, while FCM is suitable when there is a need to capture the degree of membership or uncertainty in cluster assignments.

A comparison between K-Means and FCM is represented in the figure 8

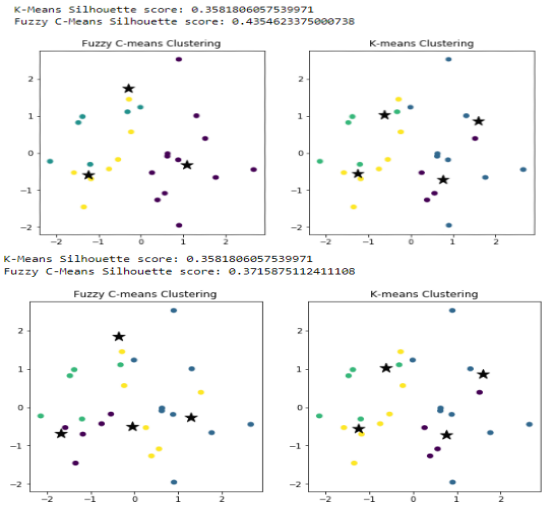


Fig 8.A Comparison between K-Means and Fuzzy C-Means Clustering

```
>
> setwd("F:/lab")
> df=read.csv("bdcard.csv")
> str(df)
'data.frame':   25 obs. of  5 variables:
 $ ID   : int  101 102 103 104 105 106 107 108 109 110 ...
 $ Age  : int  52 32 44 38 35 28 25 43 36 27 ...
 $ Card1: int  73 61 42 67 80 59 65 84 65 84 ...
 $ Card2: int  64 70 74 73 86 71 59 69 68 58 ...
 $ Card3: int  66 89 91 80 84 89 67 91 72 54 ...
> head(df)
    ID Age Card1 Card2 Card3
1 101  52    73    64    66
2 102  32    61    70    89
3 103  44    42    74    91
4 104  38    67    73    80
5 105  35    80    86    84
6 106  28    59    71    89
> df
    ID Age Card1 Card2 Card3
1  101  52    73    64    66
2  102  32    61    70    89
3  103  44    42    74    91
4  104  38    67    73    80
5  105  35    80    86    84
6  106  28    59    71    89
7  107  25    65    59    67
8  108  43    84    69    91
9  109  36    65    68    72
10 110  27    84    58    54
11 111  45    77    73    81
12 112  42    48    70    80
13 113  30    82    96    86
14 114  47    82    70    76
15 115  53    94    68    79
16 116  55    66    70    57
17 117  50    70    72    72
18 118  48    61    61    85
19 119  26    76    82    75
20 120  33    73    68    66
21 121  39    79    83    65
22 122  41    87    57   100
23 123  54   101    60    76
24 124  29    65    79    71
25 125  40    75    62    78
```

Fig 9. Data Frame of Credit Card Data

Fuzzy C-Means (FCM) differs from K-Means as a soft clustering algorithm, allowing data points to have varying degrees of membership in multiple clusters. FCM is calculating the membership degree for each data point, resulting in a probability distribution over clusters. This flexibility enables FCM to capture overlapping clusters and outliers more effectively.

To compare the performance of K-Means and FCM on a specific dataset, metrics such as silhouette score or the Dunn index can be used to evaluate the clustering quality in terms of cluster compactness and separation. In the Python, the silhouette score for K-Means and FCM are obtained as 0.220910 and 0.2169, respectively. A comparison between K-Means and FCM, including scatter plots, is shown in figure 8.

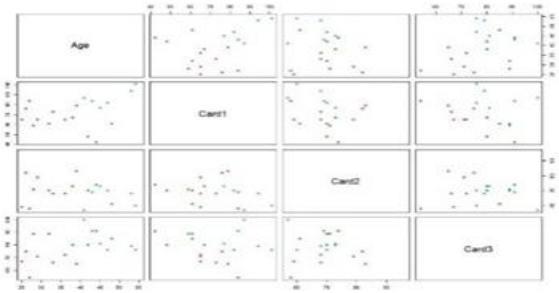
Using R language, the dataset of credit card data is represented as a data frame, which include the information such as the ID, Age, Card1, Card2 and Card3. The data frame structure head and the complete data frame itself are shown in the figure 9 with pairing of cards illustrated in the figure 10.

```
> pairs(df, col=df[,5])
> cor(df[,1:4])
              ID        Age      Card1       Card2
ID     1.0000000  0.1260897  0.38652838 -0.06922310
Age    0.1260897  1.0000000  0.19494594 -0.26947380
Card1  0.3865284  0.1949459  1.00000000 -0.06961954
Card2 -0.0692231 -0.2694738 -0.06961954  1.00000000
> require(psych)
```

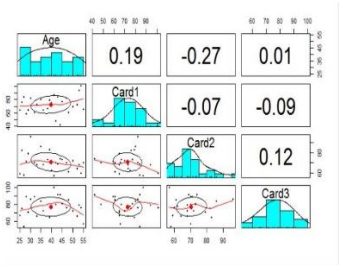
Fig 10. Pairing of Data of Credit Cards

On the basis of above as shown in the figure 9 and figure 10 a graph of attibutes of the data set is plotted and the in the figure 11 with pairing of cards as Graph of Cards and in figure 15 shown as Graph of pairing of Cards.

Based on the visual representations in Fig. 9 and Fig. 10, we plot a graph showing the characteristics of the dataset. The graph is shown in Fig. 11. In Fig. 15, we show another graph with the pairing of cards highlighted for clarity.



(a) Scattered Attributes



(b) Pairing of Cards

Fig 11. Scattered Attribtes of Cards and Pairing.

To perform fuzzy C-Means clustering, we have chosen the value of K as 3 and the fuzziness parameter m as 2. Using these parameters, the clustering result is depicted in figure 12. The data points are grouprd into three clusters like cluster1, cluster 2, and cluster3.

The summary of the Fuzzy C-Means clustering is obtaind from the "res.fcm" object which includes various information such as the crisp cluster vector, inttial cluster prototype, the distance between the intitial and final cluster prototypes. These details provide insights into the clustering process and the evolution of the cluster prototype.

```
> res.fcm <- fcm(x, centers=3)
> as.data.frame(res.fcm$u)[1:6,]
   Cluster 1  Cluster 2 Cluster 3
1 0.42768575 0.38796049 0.1843538
2 0.09269929 0.06237011 0.8449306
3 0.20982626 0.16388438 0.6262894
4 0.26470231 0.10956690 0.6257308
5 0.34079074 0.38823788 0.2709714
6 0.14514834 0.09063896 0.7642127
> res.fcm$v
              Age    Card1    Card2    Card3
Cluster 1 36.31748 72.34114 70.92645 70.25922
Cluster 2 45.33041 82.70637 69.18430 79.37595
Cluster 3 36.94124 60.95890 70.74791 83.54046
> summary(res.fcm)
Summary for 'res.fcm'

Number of data objects:  25

Number of clusters:  3

Crisp clustering vector:
 [1] 1 3 3 3 2 3 1 2 1 1 2 3 2 2 2 1 1 3 1 1 1 2 2 1 2

Initial cluster prototypes:
          Age Card1 Card2 Card3
Cluster 1  36    65    68    72
Cluster 2  53    94    68    79
Cluster 3  42    48    70    80

Final cluster prototypes:
              Age    Card1    Card2    Card3
Cluster 1 36.31748 72.34114 70.92645 70.25922
Cluster 2 45.33041 82.70637 69.18430 79.37595
Cluster 3 36.94124 60.95890 70.74791 83.54046

Distance between the final cluster prototypes
          Cluster 1 Cluster 2
Cluster 2  274.8207
Cluster 3  306.3678  563.1187

Difference between the initial and final cluster prototypes
                Age     Card1     Card2      Card3
Cluster 1  0.317479   7.34114 2.9264506 -1.7407833
Cluster 2 -7.669590 -11.29363 1.1842968  0.3759486
Cluster 3 -5.058765  12.95890 0.7479088  3.5404590
```

Fig 12. Representation of Fuzzy C-Means Clustering

However, Root Mean Squared Deviation (RMSD) value is also computed as 12.38438 and Mean Absolute Deviation (MAD) value is 73.54047alongwith membership degree matrix of all data points of data set, Descriptive statistics for the membership degree by clusters, Dunns fuzziness coefficient is shown in figure 13 with Dunn's fuzziness performance value

```
Root Mean Squared Deviations (RMSD): 12.38438
Mean Absolute Deviation (MAD): 73.54047

Membership degrees matrix (top and bottom 5 rows):
   Cluster 1  Cluster 2 Cluster 3
1 0.42768575 0.38796049 0.1843538
2 0.09269929 0.06237011 0.8449306
3 0.20982626 0.16388438 0.6262894
4 0.26470231 0.10956690 0.6257308
5 0.34079074 0.38823788 0.2709714
...
   Cluster 1 Cluster 2 Cluster 3
21 0.5644671 0.2813455 0.1541875
22 0.2329980 0.5015659 0.2654360
23 0.2399931 0.6106292 0.1493777
24 0.5555134 0.1288299 0.3156567
25 0.3781230 0.4287585 0.1931185

Descriptive statistics for the membership degrees by clusters
         Size       Min        Q1      Mean    Median        Q3       Max
Cluster 1   10 0.4276858 0.4695384 0.5477156 0.5190719 0.5622287 0.8549918
Cluster 2    9 0.3485612 0.4287585 0.5933905 0.6106292 0.7249141 0.9181437
Cluster 3    6 0.5687800 0.6258704 0.6860803 0.6564138 0.7447941 0.8449306

Dunn's Fuzziness Coefficients:
dunn_coeff normalized
 0.4765397  0.2148095

Within cluster sum of squares by cluster:
       1       2       3
2672.6000 2804.6667  946.8333
(between_SS / total_SS =  31.4%)

Available components:
 [1] "u"          "v"          "v0"         "d"          "x"          "cluster"    "csize"
 [8] "sumsqrs"    "k"          "m"          "iter"       "best.start" "func.val"   "comp.time"
[15] "inpargs"    "algorithm"  "call"
> res1.fcm <- fcm(x, centers=3, nstart=5)
> res1.fcm$func.val
[1] 3621.014 3621.014 3621.014 3621.014 3621.014
> res1.fcm$iter
[1] 99 96 92 85 95
> summary(res1.fcm)
Summary for 'res1.fcm'

Number of data objects:  25

Number of clusters:  3

Crisp clustering vector:
 [1] 3 1 1 1 2 1 3 2 3 3 2 1 2 2 2 3 3 1 3 3 3 2 2 3 2
```

Fig 13. Dunns Fuzziness Performance Value

In the process of Fuzzy C-means clustering, there are number of iterations which are taking for clustering of scaled data and shown in figure 14 as iteration table along with error of each iteration.

```
> # Fuzzy C-means clustering
> fc <- cmeans(scaled_data, centers = 3, m = 1.5, verbose = TRUE)
Iteration:   1, Error:  2.6847646325
Iteration:   2, Error:  2.5731216796
Iteration:   3, Error:  2.5032259199
Iteration:   4, Error:  2.4525658309
Iteration:   5, Error:  2.4312364998
Iteration:   6, Error:  2.4252983396
Iteration:   7, Error:  2.4234588799
Iteration:   8, Error:  2.4227374384
Iteration:   9, Error:  2.4224048362
Iteration:  10, Error:  2.4222341746
Iteration:  11, Error:  2.4221395909
Iteration:  12, Error:  2.4220841978
Iteration:  13, Error:  2.4220505161
Iteration:  14, Error:  2.4220295327
Iteration:  15, Error:  2.4220162603
Iteration:  16, Error:  2.4220077866
Iteration:  17, Error:  2.4220023457
Iteration:  18, Error:  2.4219988398
Iteration:  19, Error:  2.4219965757
Iteration:  20, Error:  2.4219951114
Iteration:  21, Error:  2.4219941636
Iteration:  22, Error:  2.4219935496
Iteration:  23, Error:  2.4219931517
Iteration:  24, Error:  2.4219928937
Iteration:  25, Error:  2.4219927265
Iteration:  26, Error:  2.4219926180
Iteration:  27, Error:  2.4219925476
Iteration:  28, Error:  2.4219925020
Iteration:  29 converged, Error:  2.4219924723
>
> # Print the cluster centers
> print(fc$centers)
         ID        Age      Card1       Card2       Card3
1  0.3794379  0.7978562  0.5847767 -0.419885788  0.07424373
2 -0.8816051 -0.3381211 -0.8051518 -0.007137754  0.37804605
3  0.7040956 -0.7469525  0.1562531  0.747007595 -0.53110434
>
```

Fig 14. Fuzzy C-Means Iteration.

For the comparison of K-Means and Fuzzy C-Means by the implementation in R language by considering Age vs ID is shown in the figure 15 as comparison between K-Means and Fuzzy C-Means (Age vs ID) whereas by considering Card1 vs Age is shown in figure 16 as comparison between K-means and Fuzzy C-means (Card1 vs Age)

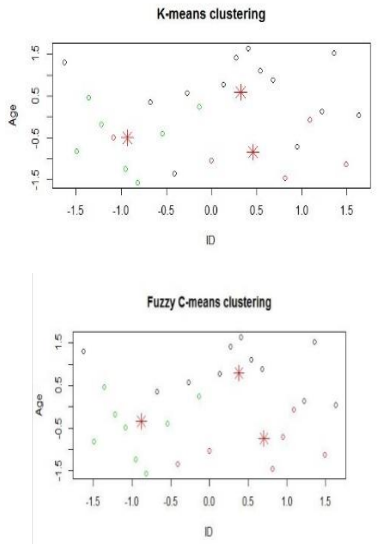


Fig 15. Comparison between K-Means and Fuzzy C-Means (Age vs ID)

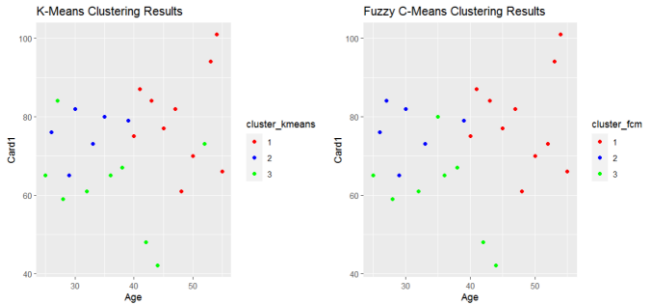


Fig 16. Comparison between K-Means and Fuzzy C-Means (Card1 vs Age)

Table 2 presents the results of applying different clustering techniques, namely K-Means and Fuzzy C-Means, on the bdcard.csv dataset using both R and Python languages. The table includes the fuzziness membership parameter (m), Means silhouette scores and Fuzzy C-Means silhouette scores for each combination of parameters.

The results show that when initializing C=2 and fuzziness parameter (m)=1.5 using R language, the K-Means silhouette scores is 0.2669444 and the Fuzzy C-Means silhouette score is 0.256082. This suggests that Fuzzy C-Means optimized for both K=3 and K=4. The table also provides similar information for additional combination of parameters, such as increasing the number of the clusters and Fuzzy C-Means silhouette scores for each combinations of parameters, such as increasing the number of clusters (C) to 3, 4, 5 and 6n along with the corresponding K-Means and Fuzzy C-Means silhouette scores.

Similarly, using the Python programming language, the data is analysed using Principal Component Analysis (PCA) and the results are presented for different numbers of clusters (C) ranging from 2 to 6. This allows for a comparison of the performance of K-Means and Fuzzy C-Means with varying clusters sizes.

Overall, the table provides a comprehensive overview of the clustering results, allowing for the evaluation and comparison of different clustering techniques and parameter combinations.

Table 2. Silhouette Score for K-Means and Fuzzy C-Means Clustering

S. No.	Data Set	Language	C(No of Cluster)	m(Fuzziness)/ PCA	K-Means Silhouette Score	Fuzzy C-Mean Silhouette Score	Conclusion
1	BDCARD.CSV	R	2	1.5	.2668444	**.**256082	FCM optimize at **K=3** and **K=4**
2			3		.2703347	**.3143806**	
3			4		.2470962	**.2886012**	
4			5		.3517374	.3491137 2	
5			6		.3873015	.3873015	
6	BDCARD.CSV	Python	2	PCA	..3721873	.3595967 5	FCM optimize at **K=4**
7			3		.4444176 5	.4354623 3	
8			4		.3581806 0	**.3715875 1**	
9			5		.509	.495	
10			6		.511	.510	

It is concluded that Fuzzy C-Means optimization is achieved at K=4, Silhouette at K=4 , Silhouette Score Measure, the quality of cluster , with values closer to 1 indicating well –separated clusters and values closer to -1 indicating poorly separated clusters.

In summary, the data demonstrated the performance of K-Means and Fuzzy C-Means cluster technique on the bdcard.csv data satisfying diff parameters and language and it evaluates the cluster quality through Silhouette Scores and identifies the optimal number of cluster for each technique.

5. Conclusions

This paper highlights the effectiveness of Fuzzy C-Means (FCM) technique in clustering of the banking data and proves to be a valuable tool for analysing large and complex datasets, offering advantages over traditional clustering algorithms. By employing FCM, bank and financial institutions can gain insights into customer segmentation, risk assessment and decision making processes. The implementation of FCM in both Python and R programming language provides researchers and practitioners with flexibility to apply the technique using preferred platform. Overall, this study contributes to the growing body of knowledge on data analysis in the banking sector and encourages further exploration of FCM in other domain as well.

References

[1] Kaymak, U., and Setnes, M. (2000), "Extended Fuzzy Clustering Algorithms". In ERIM Report Series Research in Management (Vol. ERS-2000-51-LIS, pp. 1-24). https://research.tue.nl/en/publications/12ada79b-e531-4c16-867b-5ebc4fe92163

[2] Tsekouras, G. E. (2005), "On the Use of the Weighted Fuzzy C-Means in Fuzzy Modeling".In Advances in Engineering Software (Vol.36, Issue 5, pp.287300). https://doi.org/10.1016/j.advengsoft.2004.12.001

[3] Yu, L., and Zhang, Y. Q. (2005), "Evolutionary Fuzzy Neural Networks for Hybrid Financial Prediction". In IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews (Vol.35, Issue 2, pp. 244–249). https://doi.org/10.1109/TSMCC.2004.841902

[4] Popović, D., Banka, Z.,and Dalbelo Bašić, B. (2009), " Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm". In Informatics (Vol. 33, Issue 2, pp.243-247). www.zaba.hr

[5] Zhao,Y.,Che,W.,Zhao,Q.,Gan,J.(2011) , " Research Financial Market Based on Fuzzy C-Means Clustering". In International Conference on Computer Science and Network Technology (pp. 812-815). https://doi.org/10.1109/ICCSNT.2011.6182087

[6] Martin, A., Gayathri, V., Saranya, G., Gayathri, P., and Prasanna Venkatesan. (2011), "A Hybrid Model for Bankruptcy Prediction Using Genetic Algorithm, Fuzzy C-Means and Mars". In International Journal on Soft Computing (Vol.2, Issue 1, pp.12–24). https://doi.org/10.5121/ijsc.2011.2102

[7] Aguiar, R. A. (2012), "Fuzzy Logic and Behavioral Finance: An Approach Using Fuzzy C-Means Algorithm". In International Journal of Latest Trends in Finance and Economic Sciences (Vol. 2, Issue 1, pp.1-7). http://excelingtech.co.uk/

[8] Zhang, Z., and Havens, T. C. (2013), "Scalable Approximation of Kernel Fuzzy C-Means". In 2013 IEEE International Conference on Big Data (pp.161–168). https://doi.org/10.1109/BigData.2013.6691749

[9] GÖKGÖZ, İ. H., Altinel, F., and İlker, K. O. Ç. (2013). "Classification of Turkish Commercial Banks under Fuzzy C-Means Clustering". In BDDK Bankacılık ve Finansal Piyasalar Dergisi (Vol. 7, Issue 2, pp.13-36).

[10] Costea, A. (2014), "Applying Fuzzy Logic and Machine Learning Techniques in Financial Performance Predictions". In Procedia Economics and Finance (Vol. 10, pp. 4–9). https://doi.org/10.1016/s2212-5671(14)00271-8

[11] Balamurugan, M., and Mathiazhagan, P. (2015), "Credit Card Transaction Fraud Detection System Using Fuzzy Logic and K-Means Algorithm". In International Journal of Innovative Research in Technology (Vol.2 , Issue 3, PP. 171-176)

[12] Behera, T. K., and Panigrahi, S. (2015), "Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering and Neural Network", In 2015 Second International Conference on Advances in Computing and Communication Engineering (pp. 494–499). https://doi.org/10.1109/ICACCE.2015.33

[13] Fahad, S. K. A. (n.d.), "Accuracy of Textual Document Clustering with Semantic Approach Natural Language Processing with Semantic by the help of WorldNet". In The accuracy of Clustering is assured by F-Measure. Lambert Academic Publishing (LAP).

[14] Amirkhani, A. H., Mahmoodabadi, N., and Moridi, Z. (2018), "A New Clustering Algorithm Using Interval Type-II FCM". In Astra Salvensis (Vol. 6, pp. 123-135).

[15] Atiyah, J. M., Hussein, H. H., Mohammed, E. A., and Hassen, O. A. (2019), "A Modified System of a Cryptosystem Based on Fuzzy Logic". In Journal of Advanced Research in Dynamical and Control Systems (Vol. 11, Issue 2).

[16] Saeed Khalil Saeed and Hani Hagras. (2019), "A Fraud-Detection Fuzzy Logic Based System for the Sudanese Financial Sector". In SUST Journal of Engineering and Computer Science (Vol. 20, Issue 1, pp.17-30). https://core.ac.uk/display/323246006?utm_sour

ce=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1

[17] Wu, Z.-X., Huang, K.-W., and Yang, C.-S. (2020), " A Fuzzy Crow Search Algorithm for Solving Data Clustering Problem".In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)(Vol. 12144 LNAI , 782–791). https://doi.org/10.1007/978-3-030-55789-8_67

[18] Jain, A., Purwar, A., and Yadav, D. (2021), "Credit Card Fraud Detection Using K-Means and Fuzzy C-Means". In Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies (pp. 216–240). https://doi.org/10.4018/978-1-7998-6870-5.ch016

[19] Kaminskyi, A., Nehrey, M., and Zomchak, L. (2021), "Machine learning methods application for consumer banking". In SHS Web of Conferences (Vol.107, pp.12001). https://doi.org/10.1051/shsconf/202110712001

[20] Rayavarapu, S. M. ., Prashanthi, T. S. ., Kumar, G. S. ., Lavanya, Y. L. ., & Rao, G. S. . (2023). A Generative Adversarial Network Based Approach for Synthesis of Deep Fake Electrocardiograms . International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 223–227. https://doi.org/10.17762/ijritcc.v11i3.6340

[21] Kanna, D. R. K. ., Muda, I. ., & Ramachandran, D. S. . (2022). Handwritten Tamil Word Pre-Processing and Segmentation Based on NLP Using Deep Learning Techniques. Research Journal of Computer Systems and Engineering, 3(1), 35–42. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/39

[22] Anand, R., Khan, B., Nassa, V.K., Pandey, D., Dhabliya, D., Pandey, B.K., Dadheech, P. Hybrid convolutional neural network (CNN) for Kennedy Space Center hyperspectral image (2023) Aerospace Systems, 6 (1), pp. 71-78.